# TISSUE SEGMENTATION AND CLASSIFICATION USING GRAPH-BASED UNSUPERVISED CLUSTERING

*Daniel Margolis*[†‡], *Alberto Santamaria-Pang*[†], *Jens Rittscher*[†]

[†]GE Global Research, One Research Circle, Niskayuna, NY, USA
[‡]Dept. of Systems Science and Industrial Engineering, Binghamton University, Binghamton, NY, USA

## ABSTRACT

Automated segmentation and quantification of cellular and subcellular components in multiplexed images has allowed for a combination of both spatial and protein expression information to become available for analysis. However, performing analyses across multiple patients and tissue types continues to be a challenge, as well as the greater challenge of tissue classification itself. We propose a model of tissues as interconnected networks of epithelial cells whose connectivity is determined by their size, specific expression levels, and proximity to other cells. These Biomarker Enhanced Tissue Networks (BETN) reflect both the individual nature of the cells and the complex cell to cell relationships within the tissue. Performing a simple analysis of such tissue networks managed to successfully classify epithelial cells from stromal cells across multiple patients and tissue types. Further experiments show that significant information about the structure and nature of tissues can also be extracted through analysis of the networks, which will hopefully move towards the eventual goal of true tissue classification.

***Index Terms***— Tissue Classification, Epithelial Segmentation, Multiplexed Imaging, Biological System Modeling, Biomedical Image Processing

## 1. INTRODUCTION

### 1.1. Background

Molecular pathology plays a strong role in both clinical decision making and biomedical research. Advances in imaging technologies have led to increasingly improved quality of data available, while improved computational techniques allow clinicians and researchers to perform their tasks cheaper, faster, and more accurately. [1] One such technology, multiplexed imaging, offers a radically enhanced level of image analysis by combining the information content of several individual fluorescent stains of a tissue to form a single, multidimensional image. This has proved to be a major breakthrough towards the task of fully automating cell and subcellular segmentation and quantification [2]. The resulting information about the cells and their protein expression levels provides a potential to go beyond the simple quantification of cells and actually move towards classifying individual tissues and tissue regions. Such tissue analysis could allow for a far greater understanding of the underlying pathology of a particular sample, including identification of specific cancer types for personalized treatment or automated discovery of specialized tissue regions such as necrotic or hyperplastic tissue.

### 1.2. Challenges

The most serious challenge in developing a method for analyzing multiplexed tissue images is the difficulty in reproducing results across patient and tissue types. Unfortunately, the underlying nature of the tissues has a large degree of variance between individuals and especially between tissues from different organs, regions, and with different conditions. Added to this is are problems caused by a sometimes inconsistent imaging technology which is prone to artifacts, issues with resolution, and images with little to no tissue. Usually, algorithms have to be specifically developed or fine tuned for each new tissue type, and extensive system training may need to be done before successful analysis of multiple patients can be performed. An underlying model of the tissue which is patient invariant, tissue-type invariant, condition invariant, and provides a mechanism to handle artifacts would be a significant breakthrough in improving automated segmentation and quantification and lead to a base platform from which to build more complex and complete models of various tissues.

### 1.3. Prior Work

Methods utilizing clustering in biomedical imaging has been attempted previously with some success. Usually, clustering is utilized as a method for feature reduction or selection, rather than as a method for feature generation or directly attempting segmentation and quantification of cells. A formal model to combine biomarker expression data with spatial data for these purposes has not been attempted. Furthermore, the bio-inspired nature of the model, seeking to mimic tissues as actual cells with relationships between them, differs greatly from purely statistical unsupervised clustering that has been utilized previously.

## 2. METHODS

### 2.1. Cell Segmentation

Our method assumes basic segmentation at the cell level to build up graph nodes. To that end we use NaKATPase which is a membrane marker that is expressed in the overall tissue architecture. Let $I_M$ be a membrane marker and $I_T$ the image after the ridge enhancement filter proposed by Frangi [3]. The image $I_T$ provides maximum value at the center of the membrane while providing very low value in the non-membrane region, and therefore can be used as a basic "membrane enhancement filter". A basic cell detection is performed by applying a watershed algorithm[4] to the weighted probability image $I_W = I_M * I_T$.

### 2.2. Graph Theory

The concept of a network (or graph) model was considered to be a good approach for representing tissues. In the most general sense, a graph can be defined as some set of nodes *N* which represent objects and a set of links *L* which represent the relationship between two objects [5]. Using cells as nodes and the relationship between them as links, a graph representing all the cells could be formed which could then be analyzed to provide features for classification and other tasks. Thus, for any two cells, $A, B \in N$, a relationship between them is defined as $\{A, B\} \in L$. Furthermore, the nature of a graph based on cell relationships is inherently non-reflexive, since cells aren't related to themselves, and undirected, since the relationship between to cells is independent. Thus, for any $\{A, B\} \in L, \{B, A\} == \{A, B\}$ and for any $A \in N$, $\{A, A\} \notin L$. Additionally, while a fully connected network was theoretically possible, it would have been inconsistent with the nature of the problem. A fully connected network is one where for every $A, B \in N$, it must follow that $\{A, B\} \in L$. Biologically, a cell only interacts directly with its nearest neighbors and those cells then together form clusters known as glands. Therefore, having links between every cell, especially those in different glands, wouldn't make sense. Furthermore, the differences between cell types, represented by biomarker expression data, would not be properly utilized and represented in a single fully connected network. Thus, a network model which reflects these issues needed to be developed.

### 2.3. Biomarker Enhanced Tissue Network

The Biomarker Enhanced Tissue Network (BETN) model makes the initial assumption that tissues are not merely collections of individual cells, but are complex systems whose nature is greater than just a sum of its parts. Knowledge about the underlying biology allowed the system to make two further assumptions, which are at the core of the network:

1. The closer in *proximity* of one cell to another, the higher the likelihood that the cells are connected to each other.
2. The greater the *similarity* between cells, the higher the likelihood that the cells will connect.

Given these assumptions, we can develop an energy functional to maximize in which the energy in the system is the similarity based on biomarker expression modified by the proximity between cells. Thus, the energy between any two cells *A* and *B* would be defined as:

$$E(\{A, B\} \in L) = e^{-\frac{D}{G}} M(K(A), K(B)) \quad (1)$$

where *D* is a measure of the pixel distance between *A* and *B*, $K(x)$ is a kernel transformation function, $M(K(x), K(y))$ is a measure of similarity between two kernel products, and *G* is a"growth" factor that will be explained later.

Each cell has a large number of coupled and uncoupled features which may be used when building the BETN. The kernel transformation $K(x)$ allows all the features of a particular cell, and the relationships between those features, to be utilized when determining similarity. These kernels do not need to follow Mercer's conditions (one to one, positive semi-definite), and may be linear or nonlinear in nature. The measure of similarity $M(K(x), K(y))$ must simply be some function which returns a higher value when similarity between the kernelized cell features is greater.

The "growth" factor represents a sensitivity threshold for network formation. For any given tissue image data, there exists the potential for a fully connected and a fully unconnected ($L = \emptyset$) BETN depending on how much similarity is required to form a connection. Thus, as the sensitivity threshold increases, a fully unconnected network would slowly grow until it was fully connected. Each threshold setting therefore produces a BETN that is related to the BETN at every other threshold setting, and the way the networks relate to each other can be defined in the form of a differential equation (similar to Level Sets).

### 2.4. Epithelial Classification

Earlier attempts at epithelial classification using a Support Vector Machine [6] had shown that Keratin staining levels had worked as useful features, and they were chosen as the feature set to test the BETN. However, it should be noted that any combination of biomarker features could be chosen as desired. The kernel transformation $K(x)$ of the vector of features *x* was kept simple, with each feature being transformed into the standard score (standardized) with the Equation a. To find the similarity measure $M(K(x), K(y))$ between any two cells *x* and *y*, we first found the mean of those cells' kernelized feature vectors (Eq. b) and the overall mean kernelized feature vectors of all cells (Eq. c). These were compared (Eq. d) to get a single vector of ratios from which the similarity measure could be found by taking the mean of all the elements of this vector (Eq. e and f).

a. $K(x) = \frac{x-\mu}{\sigma}$

b. $G(x,y) = \frac{K(x)+K(y)}{2}$

c. $H = \frac{\sum_i^n K(i)}{n}$

d. $I(x,y) = \frac{G(x,y)}{H}$

e. $r = |I(x,y)|$

f. $M(K(x),K(y)) = \frac{1}{r}\sum_i^r I_i(x,y)$

This BETN used the Keratin staining level of the whole cell, as well as the levels for the membrane, cytoplasm, and nucleus specifically. These staining levels were normalized for each individual tissue sample. Cells which belonged to a cluster greater than five cells were considered to be epithelial, while those that did not were considered to be stromal or artifacts. This value of five was empirically determined with several patient and tissue types. All results were verified by trained pathologists.

### 2.5. Network Analysis

#### *2.5.1. Degree*

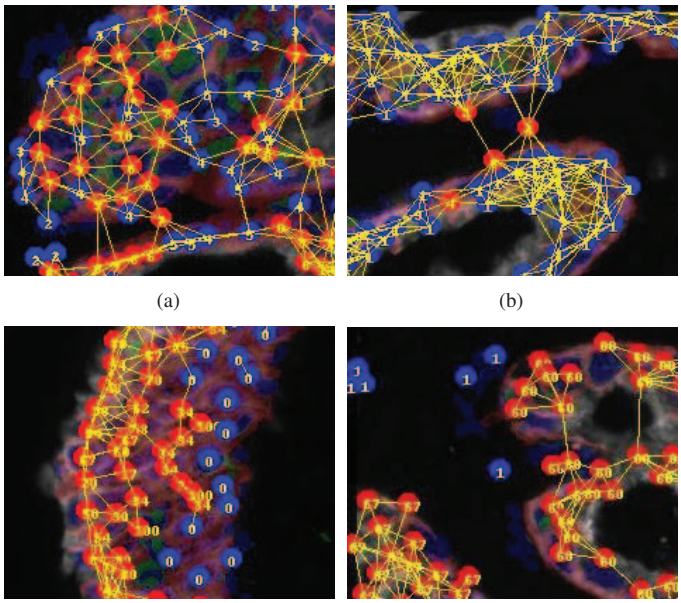
The degree of a given node *x* is the number of links in *L* which contain it. The degree $D(x)$ can be found using Equation 2. Cells with high degree are more likely to be epithelial cells, though in practice was not as useful a classifier for that as cluster size. The degree of a cell is very similar in many ways to clustering coefficient, as it provides information about the shape of the network and feedback for the threshold setting. An example of this can be seen in Figure 1(a).

$$D(x) = \left|\{y \in N : \{x,y\} \in N\}\right| \tag{2}$$

#### *2.5.2. Betweenness*

Betweenness [7] is part of centrality theory, and it is one measure of the importance of a particular node to its cluster. Specifically, the Betweenness is the number of shortest paths that use that node *s* over all the shortest paths between every node. Given the set of all shortest paths *S* (determined by Dijkstra's algorithm), the betweenness *B* can be found in (Eq. 3). High Betweenness would mean that a cell is important to the structure of a gland or important in connecting two glands into one cluster. This can be seen in Figure 1(b).

$$B(x) = \frac{|s|}{|S|} : x \in Q, Q \in S \tag{3}$$



(a) (b) (c) (d)

**Fig. 1**. Network analysis methods. Nodes with high values for a particular method are highlighted (red). (a): Degree. (b): Betweenness. (c): Clustering coefficient. (d): Connected components.

#### *2.5.3. Clustering Coefficient*

Clustering coefficient [8] is a measure of the transitivity of a node. Given some neighborhood *N* defined in Equation 4, the clustering coefficient $C(x)$ can be found by Equation 5.

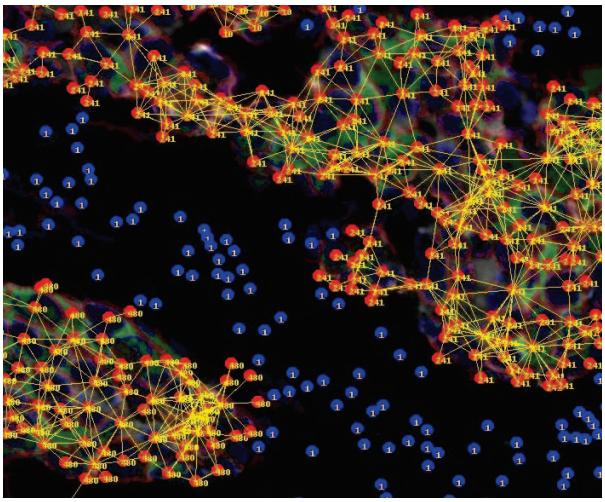
$$N(x) = \{y \in N : \{x,y\} \in L\} \tag{4}$$

$$C(x) = \frac{2|\{y,z\}|}{D(x) * (D(x)-1)} : y,z \in N(x), \{y,z\} \in L \tag{5}$$

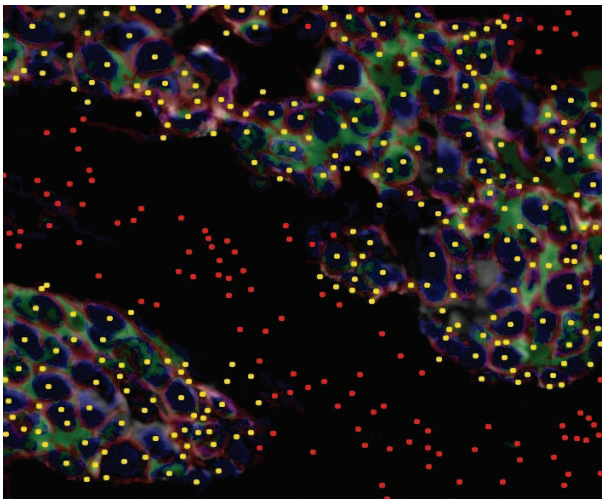
Cells with high clustering coefficients are therefore more likely to be tightly clustered, while lower ones are more likely to be on the outside of a gland or cluster. They are especially important as a tissue level measure, where they define how tight a cluster is, glandular shape, and reflect greatly as a measure of the threshold setting. When used as a classifier in this network, they were able to separate certain glands from others without any other features, showing that they could have great use as a differentiating feature of tissues. An example can be seen in Figure 1(c).

#### *2.5.4. Connected Components*

The number of components connected to a node directly or indirectly is a measure of the size of the subgraph to which a node belongs. Using a depth-first or breadth-first search, it can be recursively determined for every node $y \in N$ whether there is any path that can be formed by links in *L* that exists



(a)



(b)

**Fig. 2**. Results. (a): BETN Network of epithelial (red) cells. (b): Classification of epithelial (yellow) cells based on BETN.

between $y$ and $x$. The number of nodes in $N$ for which a path exists to $x$, in addition to itself, is the number of connected components. An example of this can be seen in Figure 1(d).

## 3. RESULTS

The classification of epithelial cells required no graph analysis beyond the number of connected components. This basic property was found to be almost linearly correlated with a cell being epithelial in nature. Furthermore, if the tissue sample contains multiple, distinct glands the clusters were often gland specific and could be used for segmentation and classification of glands. Figures 2(a) and 2(b) show an example of this classification with and without the BETN overlaid.

## 4. CONCLUSIONS

Network theory, utilizing an formal model which includes both biomarker expression and spatial data for determining connectivity, provides a potential to perform classification and segmentation tasks on a much wider scope as well as derive significant features about the overall tissue. As an unsupervised method, the BETN model does not require training, provides networks across a range of sensitivities, and can combine spatial and biomarker features in any configuration to provide any number of cluster based features for differentiating tissues and discovering knowledge. This capability will constitute a step-change in the analysis of tissue images at both the cellular and tissue level.

## 5. REFERENCES


[1] Thomas J. Fuchs and Joachim M. Buhmann, "Computational pathology: Challenges and promises for tissue analysis," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7-8, pp. 515 – 530, 2011.

[2] Hisataka Kobayashi, Michelle R Longmire, Mikako Ogawa, Peter L Choyke, and Satomi Kawamoto, "Multiplexed imaging in cancer diagnosis: applications and future advances," *The Lancet Oncology*, vol. 11, no. 6, pp. 589 – 595, 2010.

[3] Alejandro F. Frangi, Wiro J. Niessen, Koen L. Vincken, and Max A. Viergever, "Muliscale vessel enhancement filtering," in *MICCAI*, 1998, pp. 130–137.

[4] Luc Vincent and Pierre Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, pp. 583–598, June 1991.

[5] C. Berge, *Théorie des graphes et ses applications*, Dunod Paris, 1958.

[6] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[7] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[8] D.J. Watts and S.H. Strogatz, "Collective dynamics of small-worldnetworks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.