DOI:10.1145/2960404

Cell-graph construction methods are best served when physics-driven and data-driven paradigms are joined.

BY BÜLENT YENER

Cell-Graphs: Image-Driven Modeling of Structure-Function Relationship

THE STRUCTURE-FUNCTION RELATIONSHIP is fundamental to our understanding of biological systems at all levels, and drives most, if not all, techniques for detecting, diagnosing, and treating a disease. The predominant means of collecting structure/function data in biomedicine is reductionist and has thus led to a proliferation of complex data (for example, gene expression arrays, digital images) that captures only a fraction of the structure/function relationship. Gene sequence and expression data illustrates the structure and activities of individual genes but does not explain how these genes collaborate to control cellular and tissue-scale functions. As a result, despite the

abundance of molecular details known about wound healing, for example, it is virtually impossible to accurately predict the final functional state of a healing wound.³⁶ This illustrates a need to build models that represent the structural organization at the organ, tissue, cellular, and molecular levels. Furthermore, such models must capture relationships between these scales and relate them to the underlying functional state.

Data-driven network/graph analysis is primed to decipher cellular interactions in the intricate relationship between protein-protein interactions, genetic changes, metabolic pathways, and chemical secretions, which comprise cellular events. When extended to the organ level, the key challenge would be to link the local and global structural properties of tissues to the overall morphology and function of a tissue. Only a systems-level understanding of the various cellular processes encompassing multiple biological levels will take into account the multidimensional complexity of these processes. If the principles governing biological organization on a morphological, spectral, local, and global scale can be deduced, the correlation between structural and molecular signaling within the tissue can be understood and applied to inform and accelerate studies of organ development and tissue regeneration.

» key insights

- **Structural and spatial patterns of cell organizations in a tissue are not random but associated with the underlying functional state. Cell-graphs combine techniques from image analysis, graph theory, data mining, and machine learning to identify such patterns to predict underlying functional state. Thus, understanding structure-function relationships can be used to predict malfunctioning when the patterns start changing.**
- **Advances in tissue staining and image processing permit capturing multichannel, multiscale information which in turn can be used by the state of the art machine learning algorithms to model structure-function relationships.**

The *cell-graph* technique^{11,12,20} aims to learn structure-function relationship by modeling structural organization of a tissue/organ sample using graph theory. Its main hypothesis is that cells in a tissue/organ organize to perform a specific function. For example, the spatial distribution and interaction of cells in a salivary gland tissue is different than that of a brain tissue since they perform very different functions. Thus, if one can understand tissue organization then one can successfully predict the corresponding function. The cell-graph technique deploys image processing, feature extraction and selec-

tion, and machine learning algorithms to establish a quantitative relationship between structure and function.

As more sophisticated staining techniques that provide information about different biological scales are deployed (as will be discussed), image-driven modeling with cell-graphs provides a multiscale approach to modeling complex biological systems, as a complementary one to physics-based continuous models and methods (for example, finite element method, fine difference method). While quite successful in various engineering applications, these methods operate under computational scales

that capture the macroscale behavior of a system by smaller (micro) scale constitutive relations. For example, the Car-Parrinello Molecular Dynamics (CPMD) model employs a “microscopic” model to formulate the “constitutive relations” based on a force field between the nuclei, with a “macroscopic” model that uses mechanics for the dynamics of the nuclei.⁷ However, complex biological systems have different scales, including molecular, cellular, tissue, and organ levels, than the computational ones. Furthermore, physics-based techniques are parametric and do not leverage the massive amounts of data available due

to advances in data acquisition, such as high throughput medical imaging techniques, which has been recognized as an important research direction (see <https://datascience.nih.gov/bd2k>).

In this article, I will illustrate various cell-graph construction methods for different applications, explain the graph features used in tissue classification, and suggest how to combine physics-driven and data-driven paradigms toward a multiscale modeling for better prediction. The discussion starts from a simple graph model and progresses toward more sophisticated ones as a function of staining, and imaging techniques. This review includes both static data and time-evolving dynamic data as well.

Image-driven native tissue modeling.

Several different approaches are used to extract features at the cellular and tissue levels to distinguish and classify distinct (mal)functional states such as tumor types in cancer. The features are used to quantify the information carried in the sample, and to distinguish the diseased structures from the healthy and damaged ones.

The first approach makes use of *morphology* to quantify the size and shape of a cell or its nucleus.⁴⁰ At the cellular level, such features are used to classify a nucleus as belonging to a healthy or diseased cell. At the tissue level, the statistics of these features over the tissue are exploited in the classification of a tissue as diseased or not.

The second approach employs *intensity*, or the distribution of the color values of pixels to define features.⁴³ Such features may include the mean, standard deviation, skewness, and kurtosis of the red, green, and blue components, as well as the difference between the red and blue components, and the proportion of the blue component in RGB color space. Given this approach directly derives features from the intensity values, these features are more sensitive to the noise that arises from stain artifacts and image acquiring conditions.

The third approach exploits the *textural descriptors* as its features and considers spatial dependency of the intensity values to quantify the smoothness, regularity or coarseness of the image. The two most popular models to compute these textural descriptors are those that use run-length matrices¹⁸

Figure 1. Different tissue types and states as well as their representations as cell-graphs.

Cell-graphs are constructed from H&E stained human tissue samples and provide precise metrics to capture the spatial organization of tissues. Using these graph metrics, machine learning algorithms can predict the functional state of underlying tissue samples. As shown in the figure, cell-graphs and corresponding functional states of different tissue types quite different.

Courtesy:doi:10.1371/journal.pone.0032227

and co-occurrence matrices.²² Fractals that describe the similarity levels of different structures found in a tissue image over a range of scales have been proposed in Einstein¹⁵ and Esgiar.¹⁶ We note that none of the approaches mentioned here can model the structure-function relationship in tissue.

Prior work using graph theory to model a tissue is based on drawing a *Voronoi graph* of cells from a tissue image.^{26,42} In these studies the graph-based features are defined on the Delaunay triangulation graph or its corresponding minimum spanning tree. There are several limitations of Voronoi graphs that cell-graphs successfully remedy. First, the edge function in Voronoi graphs is fixed and dictated

by the location of vertices. The Delaunay triangulation permits the existence of edges solely between adjacent vertices. Thus, only relationships between closely located nuclei are represented. This restriction makes it impossible to generate and test different biological hypotheses for cell-to-cell interactions. Second, Voronoi graphs are restricted to planar graphs that are very limited in their structure and do not allow crossing of edges. There is no evidence to justify such a limitation in tissue structural organization. This constraint also presents difficulties with 3D images. Third, a Voronoi graph always has a single connected component (that is, the tissue is represented by a connected graph), which may not be a valid assumption for sparse tissues (those with fewer numbers of cells). Finally, the graph features are limited and mainly computed on minimum spanning trees over Voronoi graphs.

(Various modifications have been proposed to adapt Delaunay triangulations to specific biological systems by changing the triangulation technique and resulting in different neighborhood graphs.^{2,25,37} However, the feature sets constructed from these neighborhood graphs are limited and mainly based on spanning tree properties.)

The cell-graph approach generalizes graph-based approaches by allowing an arbitrary edge function between a pair of nodes based on a biological hypothesis on their pairwise relationship. In a cell-graph, cells or cell clusters of a sample tissue are the vertices. An edge is defined between a pair of cells or cell clusters based on an assumption that has a biological foundation (or hypothesis). For example, if we believe that cells that are spatially close to each other are more likely to interact (for example, signal) with each other than more distant cells, then a link can be made between them with a probability that decays exponentially with increasing Euclidean distance between them. Thus, links of a cell-graph aim to capture the biological interactions in the underlying tissue. The cell-graphs provide a precise mathematical representation of cellular organization and the extracellular matrix (ECM) that surrounds cells. If the images carry multichannel information, by applying more sophisticated stain-

ing techniques (for example, multispectral fluorescence imaging) it is possible to build cell-graphs that have different types of nodes, corresponding to different types of cells that coexist (for example, epithelial vs. fibroblast) and other ECM entities (for example, basal membrane underlying epithelial cell layers and blood vessels). With 3D images and 3D cell-graphs, such representation becomes more accurate and powerful. Cell-graphs bring the well-established principles of graph theory and provide a rich set of features defined precisely by these principles to be used as quantitative descriptor features. These features could be defined and computed locally from a single node's point of view (for example, number of its neighbors), or globally for the entire tissue sample (that is, the shortest or longest distance in the cell-graph between any two nodes). Cell-graphs can use cell level attributes such as convexity, size, physical contact, shape, and so on to define similarity metrics for establishing links between a pair of nodes.

As an introductory example for application of the cell-graphs, consider automated diagnosis of cancer from digital images (that is, digital pathology). The “gold standard” for cancer diagnosis remains the expert (qualitative) opinion of pathologists specially trained to recognize indicative morphological signatures of different tumors in histopathology slides. This process is not only time consuming but also subject to inter-observer variability. The cell-graph method can successfully assist diagnostics by automating this process. For example, consider the problem of predicting cancer for three morphologically distinct tissues (brain, breast, and bone) from histopathology images. Figure 1 shows the cell-graphs of three different human tissue samples in two different functional states: healthy and cancerous.

Methodology

The cell-graph methodology is image-driven and utilizes different technologies ranging from fluorescence microscopy to confocal microscopy. While most of the work we report is based on hematoxylin and eosin (H&E) stained image analysis, the cell-graph technique benefits from more sophis-

ticated staining techniques discussed later in this article.

Formally, let $G=(V, E)$ denote a cell-graph with V and E being the set of nodes and edges of the graph, respectively. The overall methodology is shown in Figure 2. It starts with image analysis and ends with checking the accuracy of the machine learning algorithms.

Identification of nodes for a cell-graph. Nodes of a cell-graph are associated with individual cells, thus the first step is to distinguish the cells from their background based on the color information of the pixels. Standard imaging techniques can be used for this part of the process. We note that cell-graphs do not require precise cell segmentation and morphology since determining cell locations are enough to identify node set. Cell segmentation is an active area of research and outside the scope of this work; we refer readers to many survey papers on this topic.^{21,29,44}

One earlier approach used for node identification is to have two control parameters: the size of the grid, and the threshold value.¹¹ The grid size determines the down sampling rate, that is, the resolution of the resultant image. Consequently, a node can represent a single cell, a part of a cell, or a bunch of cells, depending on the grid size. The finer the grid size, the closer a node is to a single cell. For each grid entry, the average values of pixels located in this grid entry are computed and compared against a threshold value to determine the nodes of the cell-graph. Thresholding eliminates the noise that arises from the staining artifacts and misassignment of black pixels in the color quantization step.

There are more sophisticated approaches to cell segmentation and node identification depending on the type of the tissue and how much segmentation accuracy is required. For example, cells can be identified by us-

Figure 2. Methodology for image-driven tissue modeling.

Figure 3. Geometric interpretation of changes in cell-graph features.

ing eigenvalues of the Hessian matrix of the image^{17,23} with two parameters of interest: $R_b = \lambda_1 / \lambda_2$, and $S = \|\nabla^2 f\|$ where S is the Frobenius matrix norm and is used to differentiate objects of interest from the background, whereas R_b is a measure to differentiate between blob-like structures and ridge-like structures. S will be low in background pixels as the eigenvalues for pixels lacking contrast will be small. In high-contrast regions however, at least one of the eigenvalues will be high and S will be large.⁴ For further details, we refer the reader to several excellent survey papers on this topic as cited above.

Establishing edges in a cell-graph. After determining the vertex set V , an edge (u, v) between a pair of nodes u and v can be defined by making use of the biological insight and knowledge of the interaction of the cells in a spe-

cific tissue type. For example, it may be more likely that physically adjacent cells signal each other than the ones far away. Such distance-based interaction among the elements is well understood in physical systems based on energy minimization. In the absence of any other similarity measure between a pair of cells, one can adapt a simple Euclidean distance measure for defining an edge between them. Therefore, we translate the pairwise spatial relation between every two nodes to the possible existence of links in a cell-graph.

An edge (u, v) can be established probabilistically or deterministically or a combination of these two methods. For example, in probabilistic cell-graphs¹¹ the probability of creating a link between any two nodes may decay exponentially with the Euclidean distance between them employing a

probability function $P(u, v) = e^{d(u, v)/(L)}$ where $d(u, v)$ is the Euclidean distance, and L is the largest Euclidean distance between two nodes of the grid. The model parameters α and β must be chosen between 0 and 1. These parameters affect the number of the links and the connectivity of the graphs. Selecting smaller values of these parameters results in a smaller number of links. Different probability functions such as power law $P(u, v) = d(u, v)^{-\alpha}$ can also be used based on a different hypothesis. Intuitively, the closer two cells are, the more likely that they share a relationship. This probability quantifies the possibility for one of these nodes to be grown from the other, thus aiming to model the prevalence of the disease state in a tissue.

An edge (u, v) can be established deterministically if the distance $d(u, v)$ is less than a threshold (for example, two cells are physically touching each other). This edge function captures cell interaction through adhesion. When the dataset is large, cross-validation techniques can be used to identify the optimal threshold that might signify cell-cell communication. In cases when the dataset is limited in size, heuristics such as five times the average radius of a nucleus (for example, 20 microns) can be used.

Note that the presence of a link between nodes does not specify what kind of relationship exists between the nodes (cells); it simply indicates that a relationship of some sort is hypothesized to exist, and that it is dependent on the distance between cells. Surprisingly, the distance measure alone is sufficient to reveal important, diagnostic structural differences in human tissues (see sidebars 1–3 in the online appendix).

Feature extraction. After constructing the cell-graphs, the next step is to define and extract graph features to train machine learning algorithms for classification of tissue functional states. We consider two types of features to be used by classification algorithms: local features at the individual cell level, and global features at the tissue level. Table 1 (in an online appendix accompanying this article in the ACM Digital Library dl.acm.org) summarizes the graph features to capture information from different scales.⁶ By computing the distri-

bution of local features, one can obtain some of the global features. However, some other global features, such as the ratio of the size of the giant connected component over the size of the entire graph, can only be computed over the entire graph.

The spectrum of a graph, which is the set of graph eigenvalues computed from the adjacency matrix or from its Laplacian, also provides global features such as the spectral radius and Eigen exponents. The eigenvalues of the Laplacian relate to the graph invariants better than the eigenvalues of the adjacency matrix.⁸ For example, the number of eigenvalues with a value of 0 gives the number of connected components in the graph. Moreover, as the eigenvalues of the Laplacian lies in the range $[0,2]$, it is easier to compare the spectra of graphs with different sizes.

Feature selection and machine learning. Feature selection helps to overcome the problem of curse of dimensionality and may increase classification accuracy. Note the importance of graph features vary from one type of tissue to another. For example, the most important features for bone tissue classification (using f-scores for feature selection) are the *number of nodes* (fs=1.685), *giant connected ratio* (fs=1.094), *number of central points* (fs=1.607), and *clustering coefficient* (fs=1.069) (the next f-score value corresponds to *percentage of end points* which is much smaller).⁴ Interestingly, while the number of nodes (that is; cell density) is an important feature for brain tissue analysis, it is not so for breast tissue.⁵ Some features such as *clustering coefficient* and *number of central points* are important for bone tissue, as well as for breast and brain tissue. Other influential features include *average effective eccentricity*, *number of links*, and *average path length*. While there is a tremendous amount of work on it (for an online repository see <http://featureselection.asu.edu/>), feature selection is based on heuristics and different methods yield different subsets.

The selected graph features are input to a classifier such as the Artificial Neural Networks, Bayesian Networks, or Support Vector Machines (SVM) for learning and predicting the functional state associated with the structure. The main challenge from learning perspec-

The cell-graph technique aims to learn structure-function relationship by modeling structural organization of a tissue/organ sample using graph theory.

tive is the unbalanced class representation. For example, while there is abundant data labeled as “cancer” class for almost all the tissue types we studied, much less data was available labeled as “healthy” class. Standard techniques such as under sampling and over sampling of data are applied to cope with this problem, and in addition each dataset is normalized and centered. In SVM the test data is classified by determining the side of the hyperplane they lie on in the kernel-mapped space. The radial basis function (RBF) kernel, also referred to as the Gaussian kernel (that is, $K(x_i, x_j) = \exp(-||x_i - x_j||^2 / 2\sigma^2)$), is commonly used as a kernel to map the data into an infinite dimensional Hilbert space. While there are some parameters in SVM that can be fine-tuned to increase learning accuracy, the default settings used in Matlab, LibSVM, and other packages have shown to be sufficient. In order to extend SVM for the classification of three classes, one can employ the one-against-one approach²⁴ where three two-class SVM classifiers are established for each pair of classes in the training dataset. Each sample in the test data is assigned to a class by these classifiers and the class with the majority vote is chosen as the final result. If there is equal voting for the three classes, the class that has the largest margin from the separating hyperplane is chosen. The Bayesian classifier maximizes the posterior probability, which is a function of the likelihood and prior probability with the assumption that the data points are drawn from Gaussian distributions. The KNN (K-nearest neighborhood) classifies each data point to the class that is most common among its K-nearest neighbors determined by a Euclidean distance-based difference. In this study, we test three values $K=10, 11$, and 12 , and choose the values that achieve the highest grading accuracy. Both Bayesian and KNN classifiers can readily handle multiclass classification problem.

We note that classification accuracy of different machine learning algorithms may vary on the same feature set. For example, for histopathological grading of follicular lymphoma (FL) images into one of three grades, a comparison of three classifiers (SVM, Bayesian, and KNN) show different accuracy results³⁴ (see Table 3 in the online appendix).

Finally, the cell-graph features can be used to design specialized kernels as an alternative approach to RBF or polynomial kernels for graph classification problems. Such kernel computation is based on feature-vectors constructed from different global topological attributes, as well as global label features. The main idea²⁴ is the graphs from the same class should have similar topological and label attributes. A detailed comparison on real benchmark datasets shows that our topological and label feature-based approach delivers better or competitive classification accuracy, and is also substantially faster than other graph kernels.²⁷ It is the most effective method for large unlabeled graphs.

Feature interpretation. There are three types of cell-graph features: cliqueness metrics (for example, clustering coefficient), compactness metrics (for example, number of central points), and distance metrics (for example, diameter).

It is best to explain the meanings of graph features within an application domain. Recently, the cell-graph technique was used to quantify changes in the cellular dynamics of submandibular gland (SMG) morphogenesis as a function of ROCK1-mediated signaling in this process.⁶ The laboratory analysis verified that the average diameter of the SMG increased and that the thickness decreased following inhibitor treatment, which is consistent with the overall decrease in cellular contractility (see Figure 3). Additionally, the total number of cells is also decreased with inhibitor treatment. This implies the overall compactness of the explant decreases both at the tissue and at the cellular level with ROCK inhibitor-treatment. The values for certain cell-graph features captured this observation: the clustering coefficient in the control tissues was greater than in the ROCK inhibitor-treated tissues. The clustering coefficient gives a measure of compactness of a tissue. That is, cells in the

ROCK inhibitor-treated tissues were further apart from each other and thus, had fewer edges, or links, per unit area, measurable as a decreased clustering coefficient. The average path length, which measures the average shortest path between two cells, increases with ROCK inhibition and number of connected components, which is the number of cell-linked cell clusters, decreases. Less compact tissue should have a smaller number of linked cells, an increased inter-cellular distance (that is, longer average path length) and, hence, a lower number of connected components. Cell-graph features were thus able to predict known ROCK inhibitor-induced global tissue changes.⁶

Enhanced Cell-Graph Models

This section explores how to go beyond simple cell-graphs without self-loops, multiple edges, and attributes to more complex ones.

Hierarchical cell-graphs. So far the cell-graphs discussed are used to model diffusive structural organizations such as the ones found in brain tissue samples. Other tissue types such as breast or prostate exhibit more complex structural organizations and require enhancing the cell-graph approach further⁵ (see comparison results in Sidebar 2 in the online appendix). For example, to model the lobular structure of breast tissue, a 2-phase cell-graph construction is proposed⁵ (see Figure 4). After cell segmentation, first, connected subgraphs have been built to capture the local structure of lobular/glandular architecture so that each connected component represents a lobular/glandular structure. A biologically meaningful hypothesis for this step is that within a glandular structure there is a high interaction among the cells as a function of physical contact. Second, the interactions among the connected components are modeled with the hypothesis that the likelihood of interglandular interaction through ECM may decrease as a function of the spatial distance between them. These two phases may admit different edge functions. For example, intra-glandular edges can be assigned deterministically while interglandular edges would be defined probabilistically.⁵ The hierarchical cell-graphs enable us to model and test different biological hypoth-

Figure 4. Hierarchical cell-graphs for breast tissue modeling.

Each connected component of a hierarchical cell-graph corresponds to a lobular structure, which is modeled by using simple cell-graph approach.²³ Courtesy: Conf Proc IEEE Eng Med Biol Soc. 2007;2007:5311-4.

Figure 5. ECM aware cell-graphs.

Fractured bone tissue example where fracture cells exist in the middle of the original image and stem cells repairing fracture are colored in blue. Courtesy: doi: 10.1007/s10618-009-0153-2

esizes on the interaction of cells, and glands by changing the edge function.

ECM-aware cell-graphs. In a tissue sample, there is more than one type of cells of interest, including blood cells, cancerous cells, normal cells, stem cells repairing a damaged tissue (for example, bone fracture), and so on. These cells are not only distinguishable from each other by their color and size, but also carry valuable information about the underlying functional state. ECM-aware cell-graphs take advantage of the heterogeneity of tissue samples to encode more information. After image segmentation, each cell is assigned a color-code based on the ECM composition of its surroundings. For each color code, a dedicated cell-graph is constructed and graph features are extracted. As a result, multiple cell-graphs coexist for modeling the same tissue and their combined feature set can be used for tissue classification. Figure 5 shows a tissue sample from fractured bone and corresponding cell-graphs constructed from segmentation, which illustrates that ECM-aware cell-graphs result in high accuracy in bone tissue classification problems⁴ (see Sidebar 3 online).

Similarly, Figure 6 captures the spatial organization of epithelial, and mesenchymal cells that coexist during the branching morphogenesis of submandibular gland.⁶

Cell-graphs with multiple staining. With immunohistochemical staining becoming more widely used in digital pathology practices, richer sets of biological information also become available to construct cell-graphs that are more realistic and relevant. These staining techniques indicate the expression level of various proteins and provide important information for learning the underlying functional state. For example, for the automated grading of breast cancer in 3D tissue sections, both the distribution and expression level of a lateral cell-membrane protein integrin $\alpha 3$ has been used to hypothesize the interaction between cell pairs.³⁵ The cell-graph edges have been established based on the integrin $\alpha 3$ densities between the nuclei pairs (see Figure 7). As the cancer progresses, reduced expression of integrin $\alpha 3$ is observed corresponding to the loss of interaction between the cells which is an important feature for

Figure 6. Stitched images of submandibular gland were segmented using the active contour method to define epithelial (white) vs. mesenchymal tissue (black) in control (a) and ROCK inhibitor-treated explants (d).

These masks were used to identify the epithelial nuclei (b, e) and mesenchymal nuclei (c, f). Using each nucleus as a vertex, cell-graphs were constructed for control and ROCK inhibitor-treated tissues, respectively (g, h), where zoomed regions of cell graphs corresponding to regions of the original images (shown as red boxes in a and d) are shown in detail. Epithelial tissue is represented by the blue graph and the mesenchymal tissue is represented by the red graph. We discarded the sublingual tissues and only used the submandibular gland. Courtesy: doi: 10.1371/journal.pone.0032906.g002

Figure 7. Non-invasive breast tissue sample.

Non-invasive breast tissue sample with two stains is shown in (a). Corresponding cell-graph based on 3D distance is shown in (b) and the cell-graph that considers expression levels is shown in (c). Table 4 in appendix shows the improvement in classification by capturing the underlying biology more effectively. Courtesy: doi: 10.1109/ISBI.2013.6556431.

grading. Table 4 in the online appendix illustrates that grading accuracy increases by capturing this information.

Recently, a new technique called *hyperplexed immunofluorescence technology* has allowed unlimited stains on a single specimen.¹⁹ Molecular stains are quantified in single cells and sub-cellular compartments, yielding unparalleled insights into the biology of intact tissues. Each stain reflects a different biological property—cell types, signaling processes, and so on. As an extension of ECM-aware cell-graphs, one can establish links based on spatial proximities of pairs of cells using the expression of each marker, thus building one cell graph for each marker. We expect the feature set of each cell-graph captures a different biological property. Figure 8 shows the cell-graphs for three different stains. Statistical analysis shows that feature sets obtained from these cell-graphs come from different probability distributions (see Table 5 in the online appendix).

3D tissue analysis with cell-graphs.

3D confocal imaging techniques have been a powerful tool for cell biologists and engineers providing 3D spatial information regarding the location of specific structures within cells and tissues. Some of the work discussed previously used 3D modeling.^{6,35} To capture the 3rd dimension, one needs to stitch z-stack sequences with some overlapping. This process requires defining a depth parameter k and then ensuring some overlap between stack $i-1$, stack i , and stack $i+1$. For example, in Oztan et al.³⁵ z-stack sequences of 8 slices deep with 25% overlap with the preceding and following z-stack sequences

are constructed. It is straightforward to extend 2D Euclidean distance to 3D. Consider two vertices: $u = (x_u, y_u, z_u)$ and $v = (x_v, y_v, z_v)$ then the distance between them in 3D would be

$$\sqrt{(x_u - x_v)^2 + (y_u - y_v)^2 + (z_u - z_v)^2}$$

Similarly, different distance metrics can be adapted such as the Lp distance:

$$L_p = \left(\sum_{i=1}^3 |x_i - y_i|^p \right)^{1/p}$$

The features defined on 2D cell-graphs can be used in the 3D case with additional computational demand.

3D cell-graphs have been constructed to model a 3D cellular environment and quantify type I collagen remodeling and fibrillogenesis with respect to mesenchymal stem cell organization over time.³ In that work, an initial result on how to integrate a physics-based mechanical model³⁰ with the cell-graph approach is also shown. The results are verified on multiple experiments and provide the first quantitative support to the hypothesis that continuity between extracellular and intracellular environments is required for stem cell fate determination (see Figure 9 in the online appendix).

Time Series of Tissue Evolution

Up to this point, the discussion on cell-graphs was confined to static histology samples. Here, we are interested in modeling the evolution of time dependent cell and tissue growth.

Spatiotemporal cell differentiation.

Recently, in vitro (3D hydrogel models) evolutions of 11 different cell-lines from different tissues that develop solid tu-

mors (epithelial, connective, and neural) were studied in order to determine which structural properties (captured by graph metrics) dominate the differentiation.²⁸ The cell-lines include MCF10A: Precancerous Human, Breast Epithelial; AU565: Human Breast Cancer HER2+/ER-; MCF7: Human Breast Cancer HER2-/ER+; MDA-MB231: Human Breast Cancer HER2+/ER2+; hDFB: Human Dermal Fibroblasts; NHA: Normal Human Astrocytes; U118MG: Human Glioblastoma; NHOst: Normal Human Osteoblast; MG63: Human Osteosarcoma; RWPE-1: Non-tumorigenic Human Prostate; DU145: Human Prostate Carcinoma.²⁸ Although there are limitations to in vitro studies, the cell lines used in this study represent a range of tissue types allowing one to directly compare the structural profiles of various functional states through analysis of cell-graph metrics as follows.

The structural features of underlying tissue samples are calculated on each cell-graph using $G_i = (V_i(t), E_i(t))$, where $V_i(t)$ and $E_i(t)$ represent the list of vertices and nodes at time point t and i represents the index for the cell line. As a result, time series of tissue evolution can be represented by a 3rd order tensor with the modes: *features* $\times$ *time* $\times$ *cell-line* whose dimensions are I, J , and K , respectively.³⁰ An entry T_{ijk} in this data cube corresponds to the value of metric i at time point j for cell-line k where $i = 1, \dots, 20$; $j = 1, \dots, 6$; and $k = 1, \dots, 11$.

Two common models in multi-way data analysis are Tucker3 and Parallel Factor Analysis (PARAFAC).¹ A Tucker3 model with orthogonality constraints on component matrices is a generalization of SVD from matrices to high-order datasets and is also called Higher-Order Singular Value Decomposition (HOSVD)¹⁰ or multilinear SVD. Using a Tucker3, a 3-way tensor $T \in \mathbb{R}^{I \times J \times K}$ is modeled as follows:

$$T_{ijk} = \left(\sum_{r=1}^R \sum_{q=1}^Q \sum_{p=1}^P G_{pqr} A_{ip} B_{jq} C_{kr} \right) = E_{ijk}$$

where P, Q and R indicate the number of components extracted from first, second and third mode ($P \leq I$, $Q \leq J$, and $R \leq K$), respectively. $A \in \mathbb{R}^{I \times P}$, $B \in \mathbb{R}^{J \times Q}$, and $C \in \mathbb{R}^{K \times R}$ are the component matrices. $G \in \mathbb{R}^{P \times Q \times R}$ is the core tensor and $E \in \mathbb{R}^{I \times J \times K}$ represents the error term.⁵

Three-mode tensor analysis in feature mode for outliers identified six features such as; average degree; clus-

Figure 8. Cell-graphs overlaid on images of (a) E-Cadherin (b) Pan-Keratin (c) Keratin 15.

The cell-graphs reveal distinct spatial patterns showing orthogonal information can be obtained.

tering coefficient; number of central points; number of connected components; standard deviation of edge lengths; and number of isolated points that capture the *compactness*, *clustering*, and *spatial uniformity* of the 3D architectural changes for each cell type throughout the time course.²⁸ Importantly, four of these metrics are also the discriminative features for our histopathology data from the previous studies reviewed earlier.

Dynamic cell-graphs. Spatiotemporal development of tissues/organs requires modeling of cell-to-cell interactions over time and has proven to be difficult. For example, while the branching processes in developing organs (lungs, pancreas, kidneys, salivary, and mammary glands) have been studied in detail, we are still far from comprehending the integrated process.⁹ Computational modeling of morphogenesis starts with mathematical models for understanding the fundamental properties of cell clusters.^{14,41} These theories were followed by continuum, physics-based models, which considered a tissue to be composed of cells and ECM and described the stress forces between these two structures.^{32,33} Such models found a wide area of applications including modeling of epithelial morphogenesis in 3D breast culture acini,³⁸ as well as lung³¹ and kidney branching morphogenesis.³⁹ However, these models are data agnostic and focus on optimization of the model parameters for the best outcome.

Advanced imaging techniques provide a vast amount of image data that motivates data-driven modeling approaches. Data-driven techniques based on cell-tracking (identifying the same cell over different time points) are computationally challenging at the organ level. The cell-graphs provide a scalable alternative by tracking the graph properties instead of individual cells. For example, dynamic cell-graphs have been constructed to model the growth and cleft formation in SMG branching morphogenesis.⁴⁵ The model takes the initial gland morphology and nuclei locations from an initial image along with basic biological control parameters such as epithelial growth factor (EGF) concentration as input. The EGF concentration levels determine the mitosis and cleft deepening rates. At each

Data-driven techniques based on cell-tracking are computationally challenging at the organ level. Cell-graphs provide a scalable alternative by tracking the graph properties instead of individual cells.

iteration of the algorithm, the cells are divided into two populations based on their distance from the gland boundary, namely internal (I) and periphery (P). Subsets I_0 and P_0 of I and P , respectively are chosen to undergo a proliferation attempt. Cells in P_0 that successfully undergo mitosis create new cells (or vertices) V_0 that are added to V .

Cells with identical topology and growth are permitted only at the gland boundary, where a hypothesized “nutrient medium” provided by the mesenchyme is accessible. In the dynamic cell-graph model, this similarity is enforced via the local structural (graph) properties of cell-graphs that maintain consistency in the topology of the SMG throughout the development stages. When first created, potential daughter vertices are placed outside the initial gland boundary in a region within 20° of the surface normal from the parent vertex at a minimum distance of one cell diameter, but less than the specified maximum edge length. Some parameter K of possible candidate daughter vertices satisfying these spatial and angular constraints are chosen, and the daughter vertex with the closest local cell-graph features to the parent vertex is selected as the optimal daughter vertex. These local structural features assess the spatial uniformity (clustering coefficient), connectedness (degree, closeness centrality, betweenness centrality), and compactness (edge length statistics) of the cell-graph. New edges, E_0 , are also constructed based on the distances from the new cells to existing cells in G . Bud outgrowth is modeled by the annexation of new nodes into the gland boundary.³³ However, the main difficulty with this approach is to derive the smooth shape formation from the cell-to-cell interactions as discussed in Dhulekar et al.¹³

Conclusion

This article explored various cell-graph constructions to model the information encoded in the image of a complex structure like the human brain or bone tissue. The main assumption of the cell-graph approach is that cells in a tissue organize in a certain way to perform a specific function; thus, understanding structure would predict the function or malfunction.

Graph theory implementation pro-

vides a rich and rigorous set of features that are almost an order of magnitude greater than prior methods, which were limited to a few features. Furthermore, it facilitates new kernels and data mining algorithms such as subgraph mining. These features can then be used to train machine learning algorithms for predicting the functional class label, given the feature set for test data. We note that identifying graph metrics that help to predict long-term functionality by linking engineered tissue structure to function is an important step toward optimizing biomaterials for the purposes of regenerative medicine.

Any interdisciplinary work requires strong collaboration between biomedical experts and computational scientists, given that interpretability of the results is crucial. It is particularly important to understand and relate the computational feature space back to the original problem domain to advance the knowledge there. Some of the cell-graph features are not intuitive while still useful for classification and prediction (while they remain effective, interpretability of features poses a specific challenge to the convoluted features obtained by deep learning algorithms).

Finally, we note that a coupling of continuum physics-based models with discrete data-driven models such as the cell-graphs may provide more accurate prediction as the complexity of the underlying problem increases (for example, organ morphogenesis). In Metzger et al.³¹ an initial attempt for such model combining is reported by replacing the “springs”³² with weighted cell-graph edges where weights are calculated directly from images of collagen fibers. However, much work needs to be done in this direction since as reported,⁴⁵ the cell-graphs are agnostic to the physical laws that govern the underlying structural organization, and are sufficient to predict complex shape formation and need to be coupled with techniques such as the level set method.

Additional background information, literature, and figures appear in an online appendix available with this article in the ACM Digital Library (<http://dl.acm.org/citation.cfm?id=2960404&picked=formats>).

References

- Acar, E. and Yener, B. Unsupervised multiway data analysis: A literature survey. *IEEE Transactions on Knowledge and Data Engineering* 21, 1 (2009), 6–20.
- Albert, R., Schindewolf, T., Baumann, I. and Harms, H. Three-dimensional image processing for morphometric analysis of epithelium sections. *Cytometry* (1992), 13:759–765.
- Bilgin, C.C. et al. Quantification of three-dimensional cell-mediated collagen remodeling using graph theory. *PLoS One* 5, 9 (2010), e12783.
- Bilgin, C.C., Bullough, P., Plopper, G.E., and Yener, B. ECM-aware cell-graph mining for bone tissue modeling and classification. *Data Min. Knowl. Discov.* 20, 3 (May 2010), 416–438.
- Bilgin, C., Demir, C., Nagi, C. and Yener, B. Cell-graph mining for breast tissue modeling and analysis. In *Proc. of IEEE EMBC* (2007).
- Bilgin, C.C., Ray, S., Baydil, B., Daley, W.P., Larsen, M. and Yener, B. Multiscale feature analysis of salivary gland branching morphogenesis. *PLoS One* 7, 3 (2012).
- Car, R. and Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Physical Review Letters* 55, 22 (1985), 2471–2474.
- Chung, F.R.K. Spectral graph theory. Conference Board of the Mathematical Sciences, *American Mathematical Society* 92 (1997). Providence, RI.
- Davies J. *Branching Morphogenesis*. Springer-Verlag, 2004.
- de Lathauwer, de Moor, L.B. and Vandewalle, J. A multilinear singular value decomposition. *SIAM J. Matrix Analysis and Apps* 21, 4 (2000), 1253–1278.
- Demir C., Gultekin, S.H. and Yener, B. Augmented cell-graphs for automated cancer diagnosis. *Bioinformatics (Suppl 2)* 21, (2005), ii7–ii12.
- Demir C., Gultekin, S.H. and Yener, B. Learning the topological properties of brain tumors. *IEEE/ACM Trans. Computational Biology and Bioinformatics* 2, 3 (2005), 262–270.
- Dhulekar, N., Oztan, B. and Yener B. Model coupling for predicting a developmental patterning process. In *Proc. of SPIE* 2016.
- Eden, M. A two-dimensional growth process. In *4th Berkeley Symposium on Mathematical Statistics and Probability* (1961), 223–239.
- Einstein, A.J., Wu, H.S., Sanchez, M. and Gil, J. Fractal characterization of chromatin appearance for diagnosis in breast cytology. *Journal of Pathology* 185, 4 (1998), 366–381.
- Esgiar, A.N., Naguib, R.N.G., Sharif, B.S., Bennett, M.K., Murray, A. Fractal analysis in the detection of colonic cancer images. *IEEE Trans. Information Technology in Biomedicine* 6, 1 (2002), 54–58.
- Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A. Multiscale vessel enhancement filtering. *Lecture Notes in Computer Science* (1998), 130–137.
- Galloway M.M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*. (1975) 4:172–179.
- Gerdes et. Al. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *PNAS* 110, 29 (2013), 11982–11987.
- Gunduz, C., Yener, B., and Gultekin, S.H. The cell graphs of cancer. *Bioinformatics* 20 (2004), 1145–1151.
- Gurcan, M.N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N. and Yener, B. *Histopathological Image Analysis: A Review*. 2009.
- Haralick R.M. Statistical and structural approaches to texture. In *Proc. of IEEE*. 67, 5 (1979), 786–804.
- Hladuvka, J., Konig, A. and Grotler, E. Exploiting eigenvalues of the Hessian matrix for volume decimation. In *Proceeding of the 9th International Conference in Central Europe on Computer Graphics, Visualization, and Computer Vision* (2001), 124–129.
- Hsu, C. and Lin, C. A comparison of methods for multiclass support vector machines. *IEEE Trans. on Neural Networks* 13, 2 (2002), 415–425.
- Jaromczyk J.W. and Toussaint G.T. (1992). Relative neighborhood graphs and their relatives. In *Proc. IEEE* 80 (1992), 1502–1517.
- Keenan S.J., Diamond, J., McCluggage, W.G., Bharucha, H. Thompson, D., Bartels, B.H. and Hamilton, P.W. An automated machine vision system for the histological grading of cervical intraepithelial neoplasia. *J. Pathol.* 192, 3 (2000), 351–362.
- Li, G., Semerci, M., Yener, B. and Zaki, M.J. Effective graph classification based on topological and label attributes. *ASA Data Science J. Statistical Analysis and Data Mining* 5, 4 (2012), 265–283.
- McKeen-Polizzotti, L. et al. Quantitative metric profiles capture three-dimensional temporospatial architecture to discriminate cellular functional states. *BMC Medical Imaging* 11.1 (2011), 1.
- Meijering, E. Cell segmentation: 50 years down the road. *IEEE Signal Processing Magazine* 29, 5 (Sept. 2012), 140–145.
- Meineke, F., Potten, C. and Loeffler, M. Cell migration and organization in the intestinal crypt using a lattice-free model. *Cell Proliferation* 34 (2001), 253–266.
- Metzger, R., Klein, O., Martin, G. and Krasnow M. The branching programme of mouse lung development. *Nature* 453 (2008), 745–750.
- Murray, J.D. and Oster, G.F. Generation of biological pattern and form. *Math Med Biol.* 1 (1984), 51–75.
- Oster, G.F., Murray, J.D., and Harris, A.K. Mechanical aspects of mesenchymal morphogenesis. *J. Embryol Exp Morphol* 78 (1983), 83–125.
- Oztan, B., Kong, H., Gurcan, M.N. and Yener, B. Follicular lymphoma grading using cell-graphs and multi-scale feature analysis. *SPIE Medical Imaging*. International Society for Optics and Photonics, 831516–831516.
- Oztan, B., Shubert, K.R., Bjornsson, C.S., Plopper, G.E. and Yener, B. Biologically-driven cell-graphs for breast tissue grading. In *Proceedings of IEEE 10th International Symposium on Biomedical Imaging (Apr. 2013)*, 137–140.
- Plopper, G., Larsen, M. and Yener, B. *Image-enhanced Systems Biology: A Multiscale, Multidimensional Approach to Modeling and Controlling Stem Cell Function in Computational Biology of Embryonic Stem Cells*. Ming Zhan, ed. Bentham Science Publishers, 2012, 71–87.
- Raymond, E., Raphael, M., Grimaud, M., Vincent, L., Binet, J.L., Meyer, F. Germinal center analysis with the tools of mathematical morphology on graphs. *Cytometry* 14 (1993), 848–861.
- Rejniak, K.A. An immersed boundary framework for modeling the growth of individual cells: An application to early tumour development. *J. Theor Biol* 247, 1 (2007), 186–204.
- Srivathsan, A., Menshikau, D., Michos, O. and Iber, D. Dynamic image-based modelling of kidney branching morphogenesis. *Computational Methods in Systems Biology. Lecture Notes in Computer Science* 8130 (2013), 106–119. Springer, Berlin Heidelberg.
- Street W.N., Wolberg, W.H. and Mangasarian, O.L. Nuclear feature extraction for breast tumor diagnosis. *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*. San Jose, CA, 1905:861–870.
- Turing A.M. The chemical basis of morphogenesis. *Philos Trans R Soc Lond B Biol Sci* 237, 641 (1952), 37–72.
- Weyn, B. et al. Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis. *Cytometry* (1999), 35:23–29.
- Wiltgen M., Gerger, A. and Smolle, J. Tissue counter analysis of healthy common nevi and malignant melanoma. *Int J Med Inform.* 69(1), 17–28, 2003.
- Xing, F. and Yang, L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. *IEEE Rev Biomed Eng.* (Jan. 6, 2016).
- Yener, B., Dhulekar, N., Ray, S., Yuan, D., Oztan, B., Baskaran, A. and Larsen, M. Prediction of growth factor dependent cleft formation during branching morphogenesis using a dynamic graph-based growth model. *IEEE/ACM Trans. Computational Biology and Bioinformatics*.

Bülent Yener (yener@cs.rpi.edu) is a professor in the Department of Computer Science and in the Department of Electrical, Computer and Systems Engineering at Rensselaer Polytechnic Institute, Troy, NY. He is the founding director of Data Science Research Center at RPI as well as co-director of Pervasive Computing and Networking Center.

Copyright held by owner/author.

Watch the author discuss his work in this exclusive *Communications* video. <http://cacm.acm.org/videos/cell-graphs>