BioSig: An Imaging Bioinformatics System for Phenotypic Analysis

B. Parvin, Q. Yang, G. Fontenay, and M.H. Barcellos-Hoff
Lawrence Berkeley National Laboratory

***Abstract*— Organisms express their genomes in a cell-specific manner, resulting in a variety of cellular phenotypes or phenomes. Mapping cell phenomes under a variety of experimental conditions is necessary in order to understand the responses of organisms to stimuli. Representing such data requires an integrated view of experimental and informatic protocols. The proposed system, named BioSig, provides the foundation for cataloging cellular responses as a function of specific conditioning, treatment, staining, etc. for either fixed tissue or living cell studies. A data model has been developed to capture experimental variables and map them to image collections and their computed representation. This representation is hierarchical and spans across sample tissues, cells, and organelles, which are imaged with light microscopy. At each layer, content is represented with an attributed graph, which contains information about cellular morphology, protein localization, and cellular organization in tissue or cell culture. The web-based multilayer informatics architecture uses the data model to provide guided workflow access for content exploration.**

***Index Terms*— Imaging bioinfomatics, cell segmentation, phenotypic analysis**

I. Introduction

The challenge of the post-genomic era is functional genomics, i.e., understanding how the genome is expressed to produce myriad cell phenotypes. To use genomic information to understand the biology of complex organisms, one must understand the dynamics of phenotype generation and maintenance. A phenotype is the result of selective expression of the genome. It is an expression of the history of the cell and its response to the extracellular environment. In order to define cell "phenomes," one would track the kinetics and quantities of multiple constituent proteins, their cellular context and morphological features in large populations. Such studies should also include responses to stimuli so that functional models can be generated and tested. This paper focuses on an imaging bioinformatic system that targets mapping cell phenomics [1], [2].

Signaling between cells and their extracellular microenvironment has a profound impact on cell phenotype [3]. These interactions are the fundamental prerequisites for control of cell cycle, DNA replication, transcription, metabolism, and signal transduction. The ultimate decision of a cell to proliferate, differentiate or die is the response to integrated signals from the extracellular matrix, cell membrane, growth factors and hormones. Our current aim is to understand how ionizing radiation alters tissue homeostasis. This is achieved by studying the effect of low-dose radiation on the cellular microenvironment, inter-cell communication, and the underlying mechanisms. In turn, this information can then be used to more accurately predict more complex multicellular biological responses following exposure to different types of inhibitors.

Several thousand antibodies and reagents exist for differentiating a cell's specific protein components. Some antibodies can additionally discriminate between functional variants of a protein caused by modifications such as phosphorylation status, protein conformation and complex formation. Of the intracellular proteins, a large number are involved in signaling pathways. These pathways are currently not well understood, due to the complexity of the potential events, the potential for multiple modifications affecting protein function, and lack of information regarding where and when a protein is actively participating in signaling. Inherent biological variability and genomic instability are additional factors that support the requirement for large population analysis. The BioSig informatics approach to microscopy and quantitative image analysis has been used to build a more detailed picture of the signaling that occurs between cells, as a result of an exogenous stimulus such as radiation, or as a consequence of endogenous programs leading to biological functions. For example, recent studies have shown that certain intracellular signaling pathways are linked via the cell adhesion system [4]. Cell adhesion is how a cell attaches itself via integral membrane receptors to the extracellular matrix. Experimentally manipulating extracellular matrix receptors affects cell shape, alters the response of cells to new stimuli, and modifies multicellular organization as a function of time [5], [6]. Detailed analysis of these multidimensional responses (e.g., time and space) can be achieved using digital microscopy but is hampered by labor intensive methods, a lack of quantitative tools, and the inability to index and access information through a Web interface.

A significant aspect of a phenotypic study is that changes in shape, response, and organization are heterogeneous and cell-specific in tissue. Given the need for a large sample size (number of images) and complex hierarchical representation, it is necessary to maintain a detailed data model for managing data and information. The data model can then be used as a guided workflow for user-based annotation and browsing the database. It can also be used to construct a visual interface for querying multiple targets along with positional references and morphological features. The end

Research was funded by the Low Dose Radiation Program of the Life Sciences Division, by the Medical Sciences Division, and by the Mathematical, Information, and Computational Sciences Division of the U. S. Department of Energy under Contract No. DE-AC03-76SF00098 with the University of California. The publication number is LBNL-51202. E-mail: parvin@media.lbl.gov. Web site is at http://vision.lbl.gov.

results can then be visualized in terms of plots and collage of images with sensitivity measures. Our research has three novel components: (1) development of a novel set of algorithms for capturing cellular morphology, protein expression, and cellular organization in tissue; (2) development of a data model that couples immunohistochemistry with images, instrument configuration, and multi-layered quantitative representation, and (3) development of a distributed imaging bioinformatics system that couples the data model with a Web-based visual interface.

The organization of this paper is as follows. Section 2 provides a brief overview of the system architecture and database interaction. Section 3 outlines various components of the informatic system. Section 4 provides the details of the image analysis algorithms. Section 5 outlines the details of specific phenotypic studies. Section 6 concludes the paper.

II. Architecture

The system architecture is shown in Figure 1. BioSig contains a flat file mechanism for storing raw image data; however, compressed forms of these images, along with their computed and user defined annotations, are preserved in the database. The system consists of a secure Web server that constructs a view into the database through an object model layer and an object oriented (OO) database for storage and retrieval. BioSig uses a browser to access the Web server and the database. The database supports some computational functionalities on feature-based representation of raw images; however, all image analysis operations are performed by the computation service.

BioSig currently supports five classes of operations in order to construct the object hierarchies and provide access to the database. These include creation and validation of content, transformation, communication, security, and storage. These operational classes, with the partial exceptions of security and storage, are implemented through a component-based architecture in which processing and communication tasks are generally divided into the smallest partitions of server resources, called servlets. Servlets can coexist on a single computing platform or on disparate ones. The servlet platforms maintain computing resources such that they allow scaling for an increased load in communication from distant Web browsers and other interoperable networked applications. The servlets are intentionally small to allow for extensibility. Several servlets allow for creation of database hierarchies through the Web. These servlets leverage modern markup techniques and provide validation against the schema that constrains both the structure of the data hierarchy and the individual content of each element.

III. Informatics

To understand the practical requirements of the informatic system, consider the following. A typical *in vivo* study includes a number of genetically similar mice at different stages of their development: virgin, pregnant, lactate, and involution. In each category, mice are partitioned for treatment types (e.g., implant, radiation) that they will receive. Within each treatment population, mice are sacrificed at 1 hour, 4 hours, and 8 hours post treatment time. Tissues are then collected and sectioned, and coverslips are prepared for antibody treatment and subsequent imaging. The same experiment is then repeated for genetically altered mice for comparative analysis. It is clear that even such a simple study can generate a large number of images and annotation data to address cause and effect in the context of biological heterogeneity. A data model has been developed to capture and link laboratory notebook information, experimental variables, images, and computed annotations corresponding to the cellular organization and distribution.

Phenotyping has many degrees of freedom that should relate a particular quantitative result with (1) where a sample was obtained, (2) how it was conditioned, (3) how it was treated, etc. The informatic framework maintains these relations so that different experimental results can be compared for validation, exploratory analysis, and hypothesis testing. These relations encode a mapping between quantitative results to images and experimental annotations. The informatic system consists of three components. These include (1) data model, (2) presentation manager, and (3) query manager. These subsystems are decoupled for ease of development, testing, and maintenance. The purpose of the data model is to provide (1) an underlying structure for capturing complex data types and their relationships, and (2) a guided workflow for entering experimental variables in order to homogenize experimental protocols, e.g., concentration, incubation time, temperature and the sequence of a specific experiment. Implementation of the data model is object oriented and provides bidirectional tracking and annotation and measured feature data. The presentation manager utilizes the data model to construct a flexible graphical view of the database. Furthermore, it provides the display functionality for a particular query in terms of graphs or images. The query manager maps high-level user queries to the Java objects with the intent of simplifying and hiding detailed manipulation of the database from the end users. Each of these components is discussed in further detail.

The server-side components enable interactive views into the database content through a modular architecture. Each view of the database can be defined by a user and his role to meet requirements for customization. These views are visually expressed as a directed graph and its layout is enhanced through *GraphViz*, which is an open source ATT software project.

A. Data model

The data model, shown in Figure 2, is object-oriented and provides navigational links from the laboratory notebook, experimental variables, images, and detailed quantitative results. In the actual implementation, each link may have a cardinality of more than one, and provides bidirectional tracking of information from any end point. The significance of this data model is that it supports both fixed tissue and living cell studies. The model has been developed through examination of steps in immunohistochemistry and sample preparation. Experiments on fixed tissue often involve several animals going through

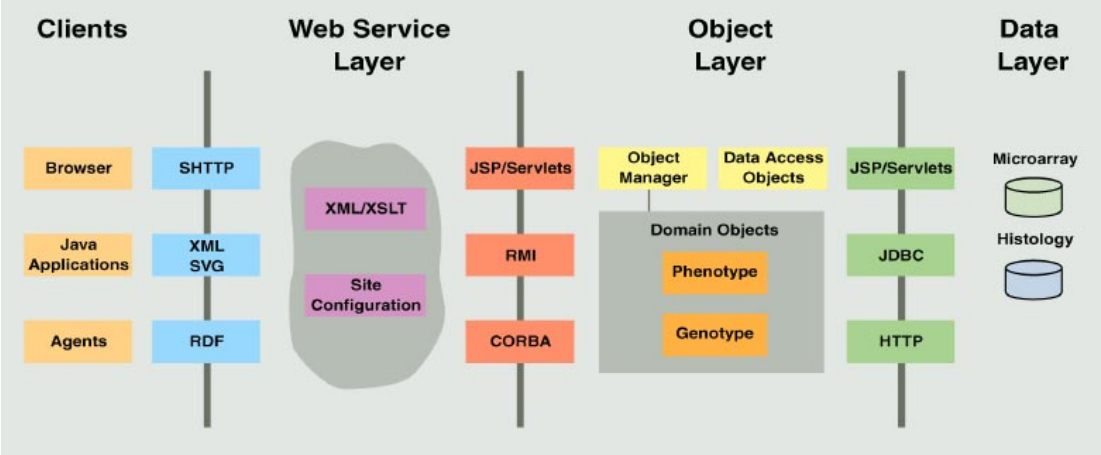


Fig. 1. Distributed imaging bioinformatics architecture for phenotypic studies is layered, uses a graphic interface, and provides an object model for improved scalability.

specific treatments, radiation, implants, or a pharmaceutical at a specific dosage and time. Tissue sections are then prepared from an organ at a specific thickness, then stained with primary and secondary antibodies labeled with a fluorochrome such as immunofluorescence, a common tool for studying protein localization. The model is represented with XML, and tools have been developed to convert the XML representation into Java code that is required by the object oriented database. In addition to the static definition of each object, a property object (name-value pair) is added for extensibility. An interface has been developed to add these new properties, specify their value types, and choose to add them to instances on a predicate basis or apply them globally. These value types include scalar data or links to instances or collections of existing and future data objects at specific layers of the hierarchy.

The model couples experimental variables (user's annotation) to feature-based representation of images, which is essentially an attributed graph. The nodes in this graph correspond to cells, and the edges correspond to the relationships between the cells. We refer to this as tissue representation which has a structure and distribution. This representation is repeated for each cell and each organelle for drilling up and down the data space.

B. Presentation manager

The presentation manager supports two functions: (1) guided exploration of the database and (2) visualization of a particular query operation. These functionalities are enhanced through the Web and scalable vector graphics (SVG), which is a W3C standard for describing two-dimensional graphics. SVG is an extensible XML-based format for interactive presentation that incorporates images, text, shapes, and video and allows for their precise layout and animation through declarative methods. SVG greatly facilitates rich presentation of data-driven graphics, and its rendering is accomplished through viewers that work as Web browser plugins or as standalone applications.

The schema, shown in Figure 2a, is represented in XSD (XML schema), and the presentation manager constructs a view into the database using this representation and the corresponding style sheets (XSL) for browsing and updating. XML generation is perfomed through small, efficient servlets that target relevant content in the database. The stylesheet is compiled into bytecode in order to avoid the overhead of request-time parsing. *GraphViz* can also export its output in SVG, which can subsequently be customized through stylesheet transformations. The presentation manager can display the result of a query function in either graphics or a collage of images. The graphics include dose-response plots and scatter diagrams of computed features as a function of independent variables. Examples of the presentation manager are shown in Figures 3 and 4.

C. Query manager

The query manager provides a set of predefined operators and dynamically generated templates to assist in information visualization and hypotheses testing. These operators help to draw contrast between computed features and their corresponding annotation data, and estimate statistical measures such as analysis of variance for sensitivity analysis. The templates correspond to attributes of a set of classes in the data model. Once these fields are selected, constraints can be specified, and the query results visualized through the presentation manager. The system translates a query into a Java program that manipulates the database to retrieve required information. Through its deep fetch mechanism, the object oriented database simplifies sensitivity analysis such as analysis of variance since each computed feature has to be mapped to its source; e.g., animal or cell culture. An example of such a high-level operator includes correlation of a particular computed feature with respect to an independent variable; e.g., *correlate "organization" of an acinus between samples that have been treated with 2-Gy levels of radiation and those that have not been radiated at all.* In this case, organization is a feature that quantifies global layout of a number of epithelial cells for a cell culture colony.

The query manager also has a unique "query by feature" search mechanism in which a feature is an attribute computed

(a)

Fig. 2. Coarse representation of the BioSig data model shows close coupling between lab notebook, experimental variables, images, and feature-based representation of images. Each image is summarized in terms of tissue, cell, and organelle content.

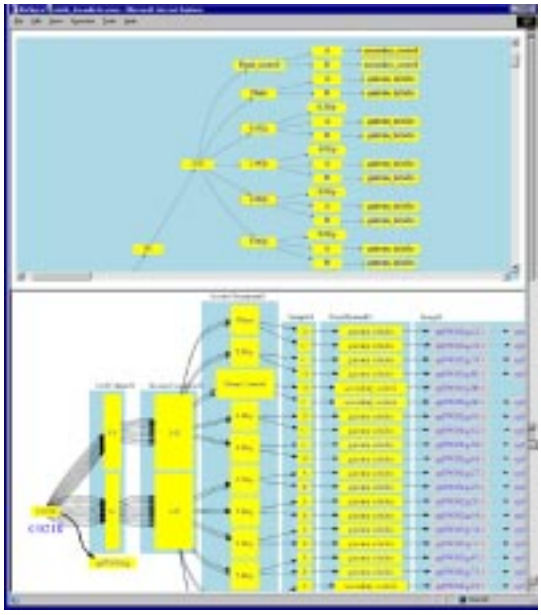


Fig. 3. Guided workflow annotation and exploration of the database content.

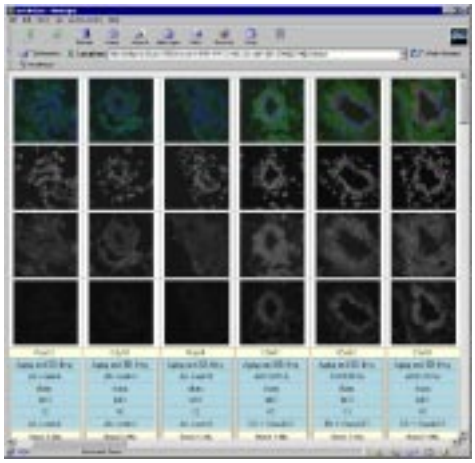


Fig. 4. Query results for a collage of images and their annotations for protein colocalization studies. Composite images are automatically generated and scaled.

from raw image data. A typical experiment can generate several hundred images that correspond to tissue or cultured cells and are stained with a particular fluorochrome. It is often of significant interest to represent a few hundred images with two or three images that are representative of the image collection. In our system, this is known as the average behavior operator, which utilizes indices corresponding to computed features (e.g., morphology or protein localization) to retrieve desired samples.

IV. Extraction of nuclei

Quantitative analysis and change detection [7] at the cellular level is an important step which can lead to a detailed understanding of protein localization as a function of microenvironment or genetic alterations. In our research, the nuclei of interest reside in a thin layer that surrounds a particular type of capillary known as a lumen. These nuclei are known as lumenal epithelial cells. During cell culture studies, a single lumenal epithelial cell divides to form a hollow sphere known as an acinus. This process often takes 10 days, when at different time points, the microenvironment is disrupted to study cell-to-cell communication. Similarly, current *in vivo* studies targets epithelial cells for normal (wild type) versus genetically altered animals (heterozygote) so that a link between changes in the microenvironment and intracellular signaling can be made as a function of genetic alteration. Furthermore, neither cellular structures nor responses are homogeneous. As a result, automatic segmentation and labeling of cells are an important aspect of any large-scale phenotypic analysis.

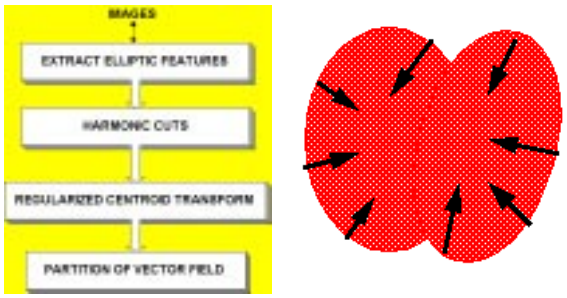
The current approach to extracting subcellular regions, e.g., nuclei, is to introduce a fluorescent dye to enable imaging and quantitative analysis. Segmentation is a hard problem since compartments may be overlapping (e.g., touching nuclei), cells have many internal structures, signal expression for each cell may not be homogeneous, and images are noisy. Furthermore, for certain studies, cells have to be classified with respect to their position and their response cataloged in time. For example, cells of interest may reside in a thin layer that surrounds a particular type of capillary. For 2D data, our previous approach [8] used both step and roof edges to partition a clump of nuclei in a way that is globally consistent. Step edges correspond to the boundaries between nuclei and background, and roof edges correspond to the boundary between neighboring (touching) nuclei. A unique feature of this system was its hyperquadric representation of each hypothesis and the use of this representation for global consistency. Global consistency was obtained through a cost function that was minimized with dynamic programming.

A new approach has been developed that is simpler, more robust, and is now part of our production system [9]. This system is also model-based and assumes that the projection of 3D nuclei onto a 2D image is locally quadratic. Instead of grouping step and roof edges, we initiate from a representation that corresponds to the zero crossing of the image in the local coordinate system. The zero crossing image is then filtered with geometrical and illumination constraints to reveal internal structures. These internal structures are then removed and interpolated with the corresponding boundary conditions. Each clump is then partitioned into several nuclei through a process that we call a centroid transform. The steps in the computational protocol are shown in Figure 5a. The centroid transform essentially projects each point along the contour into a localized center of mass, as shown in Figure 5b. The solution is regularized to eliminate noise and other artifacts along the contour. This is shown in Figure 6. In the remainder of this section, each step of the process is described in more detail.

A. Elliptic regions

Let $I_0(x,y)$ be the original image. In the linear (Gaussian) scale space, its representation at scale σ is given by $I(x,y;\sigma) = G * I$, where G is a 2D Gaussian. The vector field of gradient $\nabla I = (I_x, I_y)^T$ can be classified by its Jacobian or the Hessian matrix:

$$H(x,y) = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{pmatrix}$$



(a) (b)

Fig. 5. Segmentation process: (a) protocol for extracting delineating touching nuclei; and (b) evolution of centroid transform between two adjacent nuclei.

Bright elliptic regions can then be defined as the set of points satisfying the following conditions:

$$\begin{cases} I_{xx} < 0 \\ I_{yy} < 0 \\ I_{xx}I_{yy} - I_{xy}^2 > 0 \end{cases} \quad (1)$$

which means that both eigenvalues of the Hessian matrix are negative, or, in other words, $H(x,y)$ is negative definite. Similarly, a dark elliptic region can be identified by the following conditions:

$$\begin{cases} I_{xx} > 0 \\ I_{yy} > 0 \\ I_{xx}I_{yy} - I_{xy}^2 > 0 \end{cases} \quad (2)$$

This classification is deduced directly from the classic method for flow pattern classification [10]. In scale-space theory [11], $I_{xx}I_{yy} - I_{xy}^2$ is referred to as the elliptic feature. Other properties of this feature will be discussed in Section IV-C.

B. Harmonic cuts

The next step of the computational process is to remove small elliptic regions from the cell and interpolate their region. This is essentially a noise removal step; however, our data set has both random noise (CCD noise) and speckle noise (internal structures within the cell). Previous efforts in noise removal have been limited to filtering random noise [12]; however, structural details behave much like speckle noise and more advanced techniques need to be developed. To motivate our solution, let us first consider the one-dimensional interpolation problem. A one-dimensional function $I(x)$ with the region in the interval (a,b) can be interpolated with the average of the two endpoints, $\frac{I(a)+I(b)}{2}$. However, this approach breaks continuity of interpolation. A better approach is to weight the interpolation, at each point x, as a function of its distance to the boundary condition; i.e., let $I^{new}(x) = (b-x)I(a)/(b-a)+(x-a)I(b)/(b-a)$. It can be shown that this representation is equivalent to minimizing

$$\frac{1}{2}\int_a^b I_x^2 dx \quad (3)$$

subject to the boundary conditions

$$\begin{cases} I(a) = I_a \\ I(b) = I_b \end{cases} \quad (4)$$

The 2D case is more complex because the boundary is often noisy and irregular, and it is not clear whether propagating intensity based on distance transform will have desirable properties. We suggest that one way to ensure continuity is to regularize the solution by extending the 1D solution to 2D; i.e., by minimizing the following functional:

$$\frac{1}{2}\iint_D I_x^2 + I_y^2 dxdy \quad (5)$$

The Euler solution to this functional is the Laplace equation:

$$\nabla^2 I = I_{xx} + I_{yy} = 0 \quad (6)$$

Equation (6) is a two-dimensional harmonic function defined on D, and thus we call this method "harmonic cut." Harmonic functionals satisfy the Laplace equation and have many important properties [13]. The Laplace equation is a special case of the Poisson equation, which has been studied extensively.

C. Regularized Centroid Transform

At this stage of the computational process, each cell is represented with a smooth surface corresponding to each of its subcompartments. The next step of the process is to separate nuclei that are grouped together into a clump; i.e., touching one another. This is achieved using the *Regularized Centroid Transform* (RCT).

Figure 5b shows the basic idea for the RCT technique. The intent is to map vectors originating from the boundary of an ellipse to its centroid. If these vectors can be computed, then the entire boundary can be grouped together. This is true for both boundaries and their *interior* points; i.e., grouping utilizes not only the edges but also the regional information. The main issue is that centroids are unknown and that there are many centroids in the image. This is resolved by first computing a vector field that can then be used to partition touching objects.

Let $I(x,y)$ be the original intensity image. At each point (x_0,y_0), its equal-height contour is defined by

$$I(x,y) = I(x_0,y_0) \quad (7)$$

Expanding and truncating the above equation using Taylor's series, we have the following estimation:

$$I_x u + I_y v + \frac{1}{2}[I_{xx}u^2 + 2I_{xy}uv + I_{yy}v^2] = 0 \quad (8)$$

where $u = x - x_0$ and $v = y - y_0$, or in the standard form

$$\frac{1}{2}w^T H w + b^T w = 0 \quad (9)$$

where $H = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{pmatrix}_{(x_0,y_0)}$ is the Hessian matrix, $b = \begin{pmatrix} I_x \\ I_y \end{pmatrix}_{(x_0,y_0)}$ is the gradient of intensity, $w = (u,v)^T$ is the centroid in the local coordinate system. Recall that the

centroid of the quadratic curve defined by Eq. (9) satisfies the following linear constraint:

$$Hw + b = 0 \tag{10}$$

If H is non-singular, then the centroid can be determined directly; i.e.,

$$w = -H^{-1}b \tag{11}$$

However this is not always true, and in general, the zero set defined by

$$\begin{vmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{vmatrix} = I_{xx}I_{yy} - I_{xy}^2 = 0 \tag{12}$$

is non-trivial, and can be further classified into two categories:

1) uniform regions that correspond to zero intensity gradient of the image with the result that there is no information to estimate the centroid, and
2) elliptic features that occur in non-uniform regions.

The major limitation is that the centroids at singular points of the Hessian are not well defined. Since the basic formulation of centroid transform is ill-posed [14], a regularized formulation is implemented. Let the centroid at (x, y) be denoted by $(u(x, y), v(x, y))^T$, then the regularized model can be expressed as:

$$\begin{aligned} \min E(u, v) \quad = \quad & \tfrac{1}{2}\iint ||H \cdot (u, v)^T + b||^2 + \\ & \alpha(||\nabla u||^2 + ||\nabla v||^2)dxdy \end{aligned} \tag{13}$$

or

$$\begin{aligned} \min E(u, v) \quad = \quad & \tfrac{1}{2}\iint (I_{xx}u + I_{xy}v + I_x)^2 + \\ & (I_{xy}u + I_{yy}v + I_y)^2 + \\ & \alpha(u_x^2 + u_y^2 + v_x^2 + v_y^2)dxdy \end{aligned} \tag{14}$$

where the first and second terms are the error of estimation, the third term is the smoothness constraint, and $\alpha(> 0)$ is the weight factor. The discrete Euler-Lagrange equations of the variational problem of Equation 14 can then be expressed as:

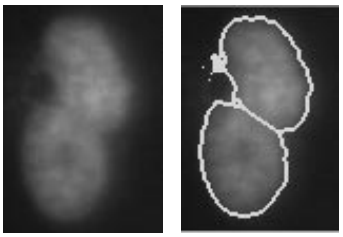
$$\begin{cases} I_{xx}(I_{xx}u + I_{xy}v + I_x) + I_{xy}(I_{xy}u + I_{yy}v + I_y) - \\ \alpha(u_{xx} + u_{yy}) = 0 \\ I_{xy}(I_{xx}u + I_{xy}v + I_x) + I_{yy}(I_{xy}u + I_{yy}v + I_y) - \\ \alpha(v_{xx} + v_{yy}) = 0 \end{cases} \tag{15}$$

D. Partitioning Vector Field

The final step of segmentation is to compute the partition of a vector field corresponding to the RCT. Consider an autonomous system of differential equations

$$\begin{cases} \frac{dx}{dt} = u(x, y) \\ \frac{dy}{dt} = v(x, y) \end{cases} \tag{16}$$

The computed vector field can be partitioned simply be migrating each point to its local centroid, as shown in Figure 5b. In this context, the RCT is a model-based watershed method. An example of segmentation results for two overlapping nuclei is shown in Figure 6.



(a) (b)

Fig. 6. Segmentation of two touching nuclei.

E. Representation and classification

Phenotyping is often multispectral for separating structural and functional information. In this context, a sample is tagged with fluorescent dye and imaged at 360 nm to reveal nuclear formation (shape and organization). Phenomics is imaged at other excitation frequencies; e.g., 490 nm and 570 nm. In our system, the structure of each nucleus is represented by an ellipse as well as hyperquadrics, and its protein expression is read and processed from other channels in the region of interest. The ellipse fit is based on estimating the parameters of polynomial $F(a, x) = ax^2 + bxy + cy^2 + dx + ey + f$ subject to the constraint that $4ac - b^2 = 1$ [15]. A 2D hyperquadric [16], [17] is a closed curve defined by

$$\sum_{i=1}^{N} |A_i x + B_i y + C_i|^{\gamma_i} = 1 \tag{17}$$

Since $\gamma_i > 0$, (17) implies that

$$|A_i x + B_i y + C_i| \leq 1 \quad \forall i = 1, 2, ..., N \tag{18}$$

which corresponds to a pair of parallel line segments for each i. These line segments define a convex polytope (for large γ) within which the hyperquadric is constrained to lie. This representation is valid across a broad range of shapes which need not be symmetric. The parameters A_i and B_i determine the slopes of the bounding lines and, along with C_i, the distance between them. γ_i determines the "squareness" of the shape.

The fitting problem is as follows. Assume that m data points $p_j = (x_j, y_j), j = 1, 2, ..., m$ from n segments ($m = \sum_{i=1}^{n} m_i$) are given. The cost function is defined as:

$$\epsilon^2 = \sum_{j=1}^{m} \frac{1}{||\nabla F_j(p_j)||^2}(1 - F_j(p_j))^2 + \lambda \sum_{i=1}^{N} Q_i \tag{19}$$

where $F_j(p_j) = \sum_{i=1}^{N} |A_i x_j + B_i y_j + C_i|^{\gamma_i}$, ∇ is the gradient operator, λ is the regularization parameter and Q_i is the constraint term [17]. The parameters A_i, B_i, C_i, γ_i are calculated by minimizing ϵ using the Levenberg-Marquart nonlinear optimization method [18] from a suitable initial guess [17]. Each nucleus in the image is further classified with respect to the position in the lumen. Figure 7a shows an example of ellipse fitting and classification of nuclei in the image. Figure 7b shows that p53 expression is (1) punctate

in the second channel, (2) heterogeneous for cells with same classification, and (3) higher in the lumenal than stromal cells (cells in the periphery of the image).

Classification of each cell in tissue is performed by representing cellular organization with an attributed graph, as shown in Figure 8. The nodes and edges in this graph correspond to cells and their relationship, respectively. The attributed graph provides the macro information about the micro anatomy where lumen can be localized and cell lines can be labeled with respect to their positions with the lumen.

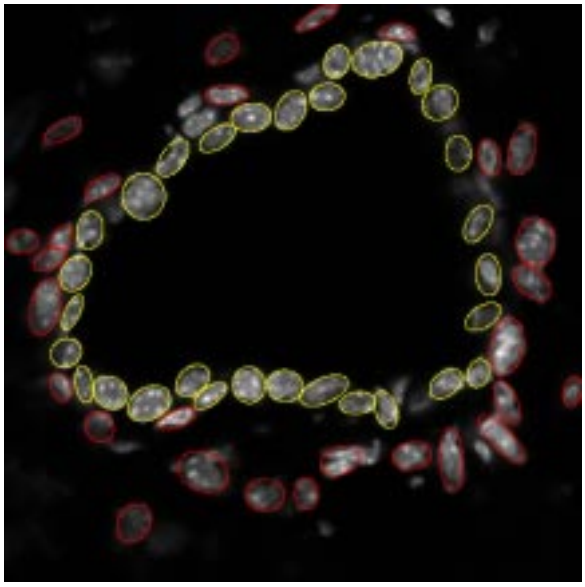
V. Applications

Examples of two applications are included here to show how BioSig can be used. The first one corresponds to cell culture studies involving cell-cell communication and adhesion for low radiation exposures. The second one provides the basis for establishing a link between extra-cellular manipulation and intra-cellular signaling for normal (wild type) versus genetically altered animals (heterozygote).

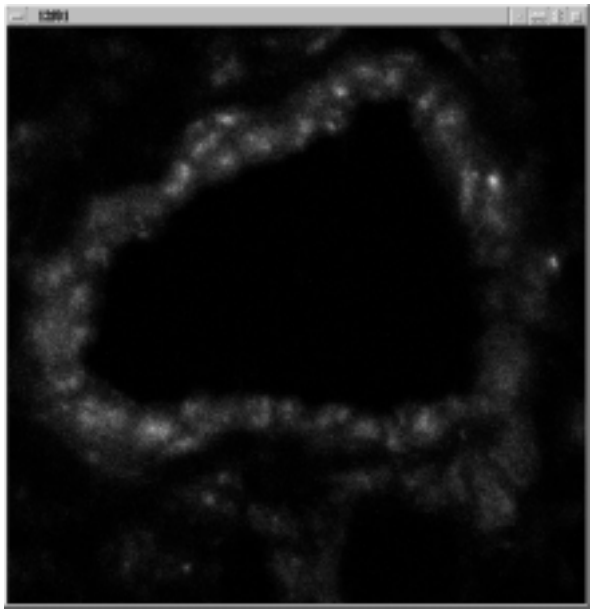
A. Cell culture studies

To determine whether low-dose radiation promotes aberrant extracellular matrix (ECM) interactions, we have utilized BioSig to examine integrin and E-cadherin localization in preneoplastic human cells surviving radiation. Integrins are a family of epithelial receptors for the ECM, while E-cadherin maintains normal cell-cell interactions and architecture. We used the HMT-3522 (S1) human breast cell line cultured within a reconstituted ECM [19]. These cells are genomicaly unstable but phenotypically normal in that they recapitulate normal mammary architecture in the form of a multicellular, three-dimensional acinus [20]. These clusters express integrins in a polarized fashion and develop an organized ECM over the course of 7 to 10 days in culture. The intent is to examine the consequences of exposing these cells to ionizing radiation and a particular protein modifier, as shown in Figure 9. Antibodies to E-cadherin, beta 1 integrin or alpha 6 integrin were detected using a green fluorescent label while nuclei were counterstained with a red fluorescent DNA dye. These were imaged using confocal fluorescence microscopy and were recorded using a 12-bit CCD camera. Cells that survived either 2 Gy or EGF showed decreased beta 1 or alpha 6 integrin localization, respectively. However, when cells were exposed to both radiation and EGF-, additional perturbations were noted. The clusters were disorganized, did not polarize the integrins at the cell surface, and failed to express E-cadherin, indicative of a lack of structural organization. An example of the untreated cells is shown in Figure 10a, which is stained for beta 1 integrin (green) with red nuclei. Comparing this sample to Figure 10b, which is a colony of cells that were irradiated and treated with EGF-, shows that the localization of beta 1 integrin is perturbed, as is the organization of the colony.

The above characteristics along with the organization of each colony were computed and stored in the database using the techniques described in section IV. A pair of segmented images from untreated and treated samples, their segmentation, and organization are shown in Figure 11. These images



(a)



(b)

Fig. 7. Segmentation and response: (a) segmentation and classification of nuclei in mammary gland shows epithelial cells in yellow and stromal cells in red; (b) P53 expression in the second channel indicates that it is expressed less in stromal cells.

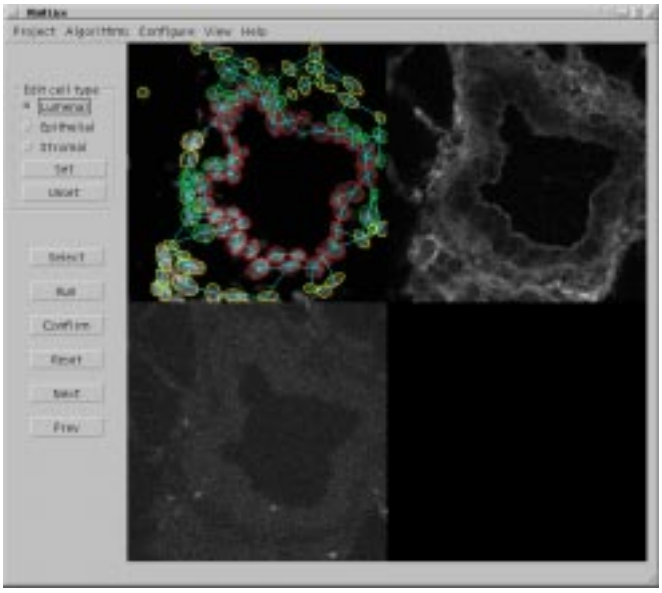


Fig. 8. Segmentation is followed by the graph representation of the tissue where protein colocalization, in specific cell lines, can be registered in the spectral stack.

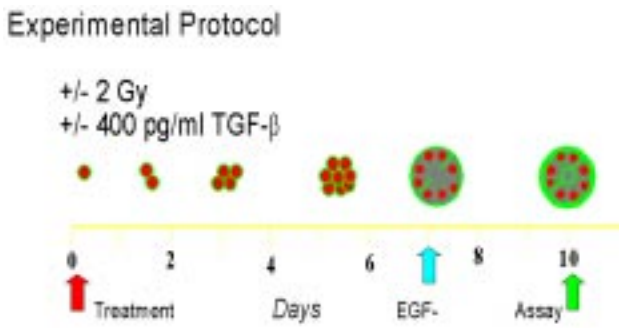
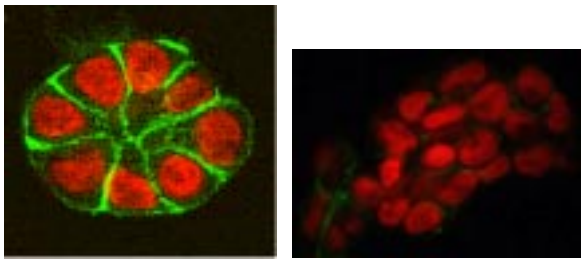


Fig. 9. Experimental protocol for *in vitro* treatment of a colony.

correspond to a feature-based representation of the "organized" and "disorganized" state of the colony in the database.

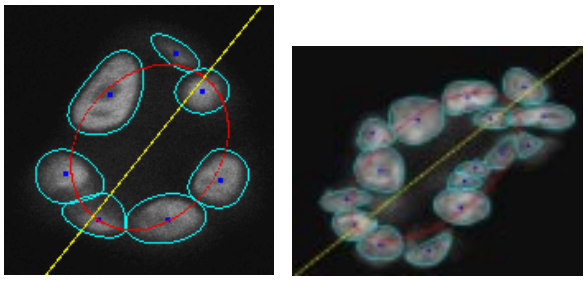
B. Tissue studies

One of the most rapid cellular responses to low-dose radiation is the activation of the transcription factor p53 (a DNA repair molecule), whose abundance and action dictates individual cellular consequences regarding proliferation, differentiation, and apoptosis. Described as the guardian of the genome by Science in 1995, p53 is one of the most rapid cellular responses to radiation. Activation of p53 allows it to bind to DNA and to transactivate target genes. A major cellular function of the p53 tumor suppressor protein is its role in promoting genome integrity. Whereas *intracellular* radiation-induced mediators of p53 stability have been the subject of intense study, little is known about the *extracellular* factors that affect the p53 response to ionizing radiation. A number of striking similarities exist between p53 and TGFβ: both regulate complex cellular decisions regarding cell fate [21],



(a) (b)

Fig. 10. Organization of a colony as a result of radiation and TGFβ treatment: (a) an untreated sample maintained its symmetry along the lumen; (b) a treated sample lost its symmetric organization.



(a) (b)

Fig. 11. Organization of a colony as a result of low-dose radiation and EGF- treatment indicates lack of symmetry around the lumen. Nuclei are segmented, represented with hyperquadrics, and symmetry is measured by fitting an ellipse; (a) an untreated sample maintains symmetry along the lumen; and (b) a treated sample loses its symmetric organization.

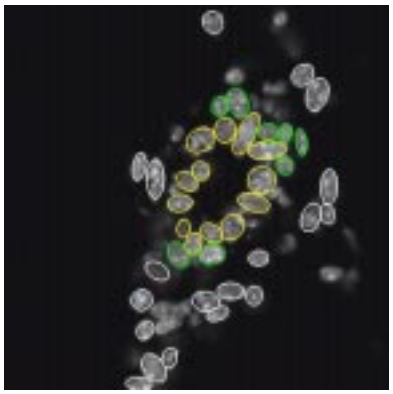
both are induced by a variety of damage and specifically ionizing radiation, and both are rapidly activated and exist in latent forms. In the present study, we used p53 antibodies that bind to a phosphorylated form of the protein that is induced upon radiation exposure. The significance of this study is that TGFβ is extracellular while p53 is intracellular.

Confocal microscopy is used to collect the distribution of p53 immunoreactivity. Segmentation technique of section IV, based on DAPI immunofluorescence, provides a discrete region of interest for p53 localization. Nuclear features such as shape, size, volume, relative location and intensity along with organization of the tissue are computed and stored in the database. These features are then used to track the level and distribution of p53 within specific tissue compartments. Perhaps as important as immunoreactive positive cells are negative cells, especially if they are restricted to certain cellular phenotypes indicating a failure to respond to radiation damage. The first result is shown in Figure 12, where BioSig provides a visual representation of p53 expression in three categories of nuclei (red for lumenal epithelial, cyan for myoepithelial, and blue for stromal cells) for a population of 54 images corresponding to wild type tissue sections. The plot provides simple visualization of a population of cells and how p53 is expressed in each cell type for all images.

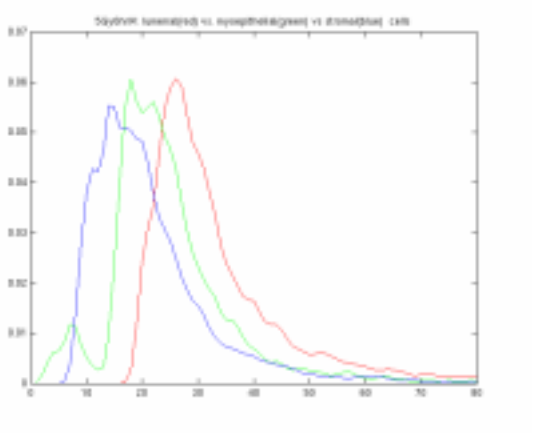
Next an experiment was designed to study the impact of TGFβ on the p53 as a result of an external exposure and different strands of mice (genetically altered). Normal mice (control animals) were exposed to low-dose radiation, tissues were collected, samples were treated with appropriate antibodies, and a large number of images were produced. Genetically altered mice, with only one copy of TGFβ (as opposed to two), were also externally exposed, etc. The protocol was repeated without any external exposure on both strands of mice. The experiment produced several thousand images that were archived in the database along with their annotations. Algorithms described in section IV were applied to these images, cells were detected and classified, and their expression was computed. The results indicate that p53, in the range where signal is being observed, is expressed less in genetically altered mice, thus, a link between extracellular condition and intracellular event is made. BioSig maps image contents to specific population response from unstructured data, allows operators to manipulate the database to retrieve a particular view of the data, and enables simple visualization of these data for population studies.

VI. Conclusion

In the post-genome-sequencing era, quantitative imaging of complex biological materials is a critical problem. Currently, sequential measurements obtained with different microscopy techniques preclude detailed analysis of multidimensional responses (e.g., time and space). Quantification of spatial and temporal concurrent behavior of multiple markers in large populations of multicellular aggregates is hampered by labor-intensive methods, a lack of quantitative tools, and the inability to index information. Ideally one would track the kinetics and quantities of multiple target proteins, their cellular context, and morphological features in three dimensions using large populations. The BioSig informatics approach to microscopy and quantitative image analysis has been used to build a more detailed picture of the signaling that occurs between cells, as a result of an exogenous stimulus such as radiation, or as a consequence of endogenous programs leading to biological functions.



(a)



(b)

Fig. 12. Population studies for p53: (a) segmentation and classification of a group of cells around lumen (yellow: lumenal-epithelial cells, green: myo-epithelial cells, and white: stromal cells); (b) probability density functions for response of p53 in each cell type (red: lumenal-epithelial, green: myo-epithelial, blue: stromal).

References

[1] B. Parvin, Q. Yang, G. Fonteny, and M. Barcellos-Hoff, "Biosig: An imaging bioinformatic system for studying phenomics," *IEEE Computer*, vol. 35, pp. 65–71, 2002.

[2] B. Parvin and D. Callahan, "Biosig: An informatics framework for representing the physiological responses of living cells," *BioSilico*, vol. 1, no. 1, pp. 42–46, 2003.

[3] C. Roskelley, A. Srebrow, and M. Bissell, "A hierarchy of ecm-mediated signalling regulates tissue-specific gene expression," *Current Opinion in Cell Biology*, vol. 7, no. 5, pp. 736–747, 1995.

[4] F. Wang, V. Weaver, O. Petersen, C. Larabell, S. Dedhar, P. Briand, R. Lupu, and M. Bissel, "Reciprocal interactions between beta 1-integrin and epidermal growth factor receptor in three-dimensional basement membrane breast cultures: A different perspective in epithelial biology," *Proceedings of the National Academy of Sciences of United States of America*, vol. 95, no. 25, pp. 14 821–14 826, 1998.

[5] F. Giancotti and E. Ruoslahti, "Integrin signaling," *Science*, vol. 285, pp. 1028–1032, 1999.

[6] A. Maniotis, C. Chen, and D. Ingber, "Demonstration of mechanical connections between integrins, cytoskeletal filaments, and nucleoplasm that stabilize nuclear structure," *Proceedings of the National Academy of Sciences of United States of America*, vol. 94, pp. 849–854, 97.

[7] N. Bourbakis, D. Kavraki, X. Yuan, and M. Goljan, "Recording changes in biological in vivo cells by using the l-g methodology," in *International Conference on Information Intelligence and Systems*, 1999, pp. 56–63.

[8] G. Cong and B. Parvin, "Model-based segmentation of nuclei," *Pattern Recognition*, vol. 33, no. 8, pp. 1383–1393, 2000.

[9] Q. Yang and B. Parvin, "Harmonic cut and regularized centroid transform for localization of subceullar structures," *IEEE Transaction on Biomedical Engineering*, vol. 50, no. 4, pp. 469–476, 2003.

[10] A. R. Rao and R. C. Jain, "Computerized flow field analysis: Oriented texture fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 693–709, 1992.

[11] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," *Journal of Applied Statistics*, pp. 225–270, 1994.

[12] P. Perona and J. Malik, "Scale space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 629–640, 1990.

[13] L. Alfors, *Complex Analysis*. McGraw-Hill, 1966.

[14] A. Tikhonov, "The regularization of ill-posed problems," *Dokl. Akad. Nauk.*, vol. SSR 153, no. 1, pp. 49–52, 1963.

[15] A. Fitzgibbon, M. Pilu, and R. Fisher, "Direct least square fitting of ellipses," in *Proceedings of the International Conference on Pattern Recognition*, 1996, pp. 253–257.

[16] A. Hanson, "Hyperquadrics: smoothly deformable shapes with convex polyhedral bounds," *Computer Vision, Graphics, and Image Processing*, vol. 44, pp. 191–210, 1988.

[17] S. Kumar, S. Han, D. Goldgof, and K. Boeyer, "On recovering hyperquadrics from range data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 11, pp. 1079–1083, 1995.
[18] W. Press, S. Teukollsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*. Cambridge Unversity Press, 1992.
[19] O. Briand, P. abd Petersen and V. Deurs, "A new diploid nontumorigenic human breast epithelial cell line isolated and propagated in chemically defined medium," *In Vitro Cell Development Biology*, vol. 23, pp. 181–188, 1987.
[20] V. Weaver, A. Fischer, O. Petersen, and M. Bissel, "The importance of the microenvironment in breast cancer progression: recapitulation of mammary tumorigenesis using a unique human mammary epithelial cell model and a three-dimensional culture assay," *Biochemical Cell Biology*, vol. 74, no. 12, pp. 833–51, 1996.
[21] A. Levine, "P53, the cellular gatekeeper for growth and division," *Cell*, vol. 88, pp. 323–331, 1997.

Bahram Parvin is a staff scientist in Computing Sciences at Lawrence Berkeley National Laboratory (LBNL). His area of research includes computer vision, feature-based representation of spatio-temporal data, and the design of intelligent systems for knowledge discovery. Parvin received his M.S. and Ph.D. in Electrical Engineering from Purdue and the University of Southern California, respectively. He is a senior member of IEEE and has served as a member of program and organizing committee in IEEE Conferences on computer vision.

Qing Yang is a computer scientist in the Computing Sciences at Lawrence Berkeley National Laboratory. His research interests include image processing, computer vision and bioinformatics. He received his M.S. and Ph.D. in computer science from the Institute of Automation, Chinese Academy of Sciences.

Gerald Fontenay is a computer scientist in Computing Sciences at LBNL. His research interests include visual interfaces for exploration of scientific data, distributed system architecture, and database development. Fontenay received his BS in computer science from San Francisco State University.

Mary Helen Barcellos-Hoff is a staff scientist in the Life Sciences Division at LBNL. Her research uses quantitative microscopy to study how ionizing radiation affects tissue microenvironments, cell interactions, and the development of breast cancer. Barcellos-Hoff received a PhD in experimental pathology from the University of California, San Francisco.