

BioSig: An Imaging Bioinformatic System for Studying Phenomics

The authors describe a system that quantifies and catalogs cellular responses under a variety of experimental conditions and maps them to image collections to create a distributed informatics architecture that facilitates sharing database content among multiple researchers.

Bahram
Parvin
Qing Yang
Gerald
Fontenay
Mary Helen
Barcellos-
Hoff
Lawrence Berkeley
National
Laboratory

Functional genomics—understanding how a genome is expressed to produce myriad cell phenotypes—presents the core challenge of the postgenomic era. Using genomic information to understand the biology of complex organisms requires comprehensive knowledge of the dynamics of phenotype generation and maintenance. A phenotype results from selective expression of the genome, creating a history of the cell and its response to the extracellular environment. Defining cell *phenomes* requires tracking the kinetics and quantities of multiple constituent proteins, their cellular context, and their morphological features in large populations. Such studies should also include responses to stimuli for use in generating and testing functional models.

Several thousand antibodies and reagents exist for differentiating between 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, many are involved in signaling pathways. The complexity of the potential events, the potential for multiple modifications affecting protein function, and the lack of information regarding where and when a protein actively participates in signaling have prevented researchers from understanding these pathways.

Inherent biological variability and genomic instability are additional factors that support the requirement for large-population analysis.

To overcome these obstacles, we have developed the BioSig imaging bioinformatic system for characterizing phenomics. Our system provides a data model for capturing experimental annotations and variables, computational techniques for summarizing large numbers of images, and a distributed architecture that facilitates distant collaboration.

CELLULAR SIGNALING

Signaling between cells and their extracellular microenvironment profoundly affects cell phenotype.¹ These interactions are the fundamental prerequisites for cell cycle control, DNA replication, transcription, metabolism, and signal transduction. A cell's ultimate decision to proliferate, differentiate, or die is a response to integrated signals from the extracellular matrix, cell membrane, growth factors, and hormones.

To understand how ionizing radiation alters tissue homeostasis, we can study the effect of low-dose radiation on the cellular microenvironment, intercell communication, and the underlying mechanisms. In turn, we can use this information to more accurately predict more complex multicellular biological responses following exposure to ionizing radiation.²

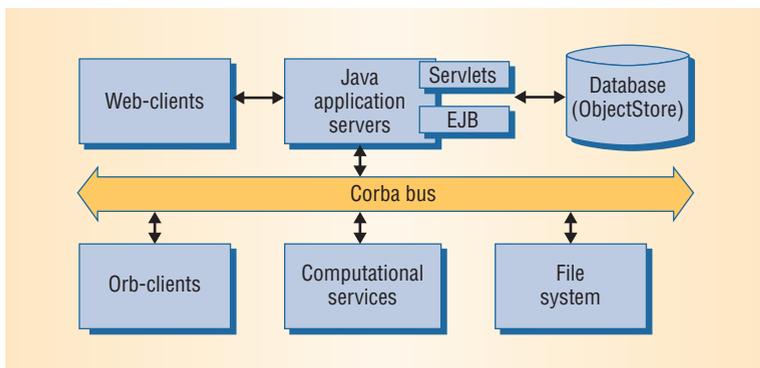


Figure 1. Informatics system architecture. The framework supports Web-based access for distant collaboration and hides detailed manipulation of the database from end users.

Recent studies have shown that certain intracellular signaling pathways link via the cell adhesion system.³ A cell uses adhesion to attach itself to the extracellular matrix via integral membrane receptors. 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.⁴ Digital microscopy can provide a detailed analysis of multidimensional responses, such as time and space. However, this method of analysis is labor-intensive, it lacks quantitative tools, and it cannot index and access information.

A typical 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—for example, an implant or radiation. Within each treatment population, mice are sacrificed at one hour, four hours, and eight hours following treatment. Tissues are then collected and sectioned, and coverslips are prepared for subsequent staining and imaging. The same experiment is then repeated for genetically altered mice for comparative analysis. Even such a simple study can generate a large volume of images and annotation data for studying cause and effect in the context of biological heterogeneity.

INFORMATICS

Phenotyping offers many degrees of freedom for relating a particular quantitative result to where a sample was obtained, how it was conditioned and treated, and so on. The BioSig informatic framework maintains these relations so that researchers can compare experimental results for validation, exploratory analysis, and hypothesis testing. These relations encode a mapping between quantitative results to images and experimental annotations.

The BioSig informatic system consists of three components: a data model, a presentation manager, and a query manager. We decoupled these subsystems for ease of development, testing, and maintenance. The data model captures experimental

variables and their relationships and maps sample preparation to images and their corresponding quantitative representations. Our object-oriented data model allows bidirectional tracking between experimental annotation and computed representation. The presentation manager provides two distinct features:

- mapping between the data model and the user interface, and
- display functionality in terms of text, plots, and images.

These features avoid hardwiring the user interface in favor of a more flexible run-time model. The query manager maps high-level user queries to the Java objects that implement the data model. This approach simplifies data manipulation and hides details of the database from end users. Figure 1 shows the informatics system's architecture, which consists of

- Java application servers providing Web-accessibility and database access;
- computational services for automated image analysis;
- a file system for storing raw data; and
- a Corba bus.

We have decided not to provide a direct Corba interface to the database at this point because current Corba-to-OO database interfaces are weak and lack vendor support. The database supports some metadata computations, but BioSig's computational services perform all image-analysis operations.

Data model

Figure 2 shows the BioSig object-oriented data model, which provides navigational links between experimental variables, images, and quantitative analysis. In the actual implementation, each link often has cardinality greater than one and provides bidirectional tracking of information from any endpoint. The data model captures

- laboratory notebook information, such as sources of cells and animals and the type of treatments;
- experimental variables, such as duration of exposure and antibodies used;
- computed morphological features and cellular classifications; and
- protein colocalization features in each subcellular region.

The data model, which supports both tissue and cell-cultured studies, has been developed through repeated interviews with experimentalists who research different aspects of phenotyping that involve both static (fixed samples) and dynamic (live cells) experiments. For sensitivity and variational analysis, an *in vivo* study often consists of many animals.

In a typical experiment, researchers prepare tissue sections from an organ at a specific thickness, then they stain the sections with antibodies to localize a protein that is detected using secondary antibodies labeled with a fluorochrome such as immunofluorescence, a common tool for studying protein localization. Finally, they digitize interesting regions of each section using specific appropriate wavelengths to excite and capture emission of the fluorochromes. Biosig then archives the resulting images and annotations for subsequent postprocessing and viewing, which includes automated scaling of images, overlaying multispectral images, and generating thumbnail images.

We represent the data model as an XML document, and software converts this representation into the Java code that the object-oriented database requires. In addition, we have augmented each data object with a property object, which consists of name-value pairs that can be designated dynamically with general collection classes. These capabilities let the data model evolve and accommodate new features without requiring changes in the database access layers. To support this evolutionary model, we have developed an interface that lets researchers add new properties, specify their value types, and choose either to add them to instances on a predicate basis or apply them globally.

Presentation manager

The Biosig presentation manager supports two features: browsing the database and visualizing the result of a query function. Users browse the database using XSD, a predefined XML schema that captures annotation data, images, and corresponding high-level features. The presentation manager uses the data model and the corresponding XSL style sheets to construct a view into the database. This design bypasses hardwiring a GUI in favor of a more flexible and dynamically generated user interface. In general, such a mapping can create a complex implementation, access control, and caching into the database. However, we have simplified the presentation module to allow browsing and updating one layer at a time. A layer refers to navigation between an object and other objects that

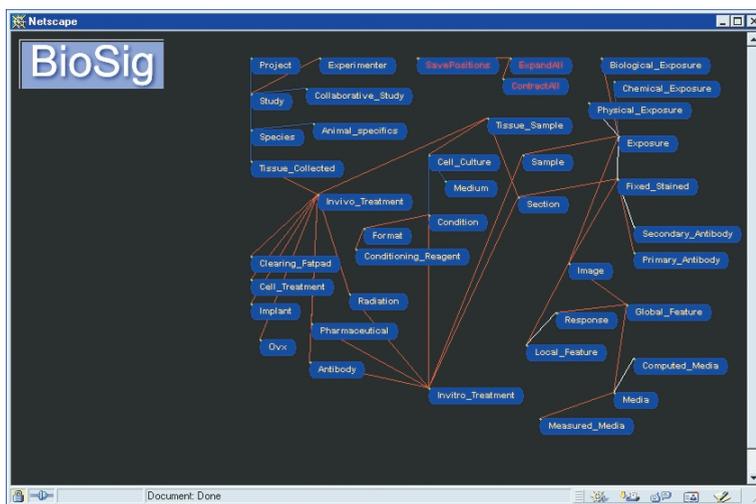


Figure 2. Coarse representation of the data model, shown as a graph. The user can click on each object to view its content in more detail.

are linked through association, aggregation, and inheritance.

The presentation manager can display the result of a query function in either text or graphics. The graphics include dose-response plots and scatter diagrams of computed features as a function of independent variables. Figure 3 shows examples of the presentation manager.

Query manager

Biosig's query manager provides a set of predefined operators and templates to assist in information visualization and hypothesis testing. These operators help draw a contrast between computed features and their corresponding annotation data, and they compute a variety of statistical measures such as analysis of variance and principal component analysis. These templates translate a query into a Java program that manipulates the database to retrieve the required information.

Through its deep fetch mechanism, the object-oriented database simplifies sensitivity testing such as analysis of variance because the system maps each computed feature to its source such as the animal or cell culture. For example, a high level template may correlate a particular computed feature with respect to an independent variable: "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, we define *organization* as a feature that quantifies the global layout of epithelial cells for a cultured colony.

The query manager also has a unique "query by feature" search mechanism in which a feature is an attribute computed from raw image data. A typical experiment can generate several hundred images that correspond to tissues or cell cultures fixed at a specific time point. Researchers are often interested in viewing two or three images that correspond to



Figure 3a. Client's view of an image collection for *in vivo* studies. The panel on the left shows navigation to a particular object in the data model. The panel in the middle shows automatic construction of composite images corresponding multispectral data sets. The panel on the right shows a collection of thumbnail images that corresponds to a specific study.

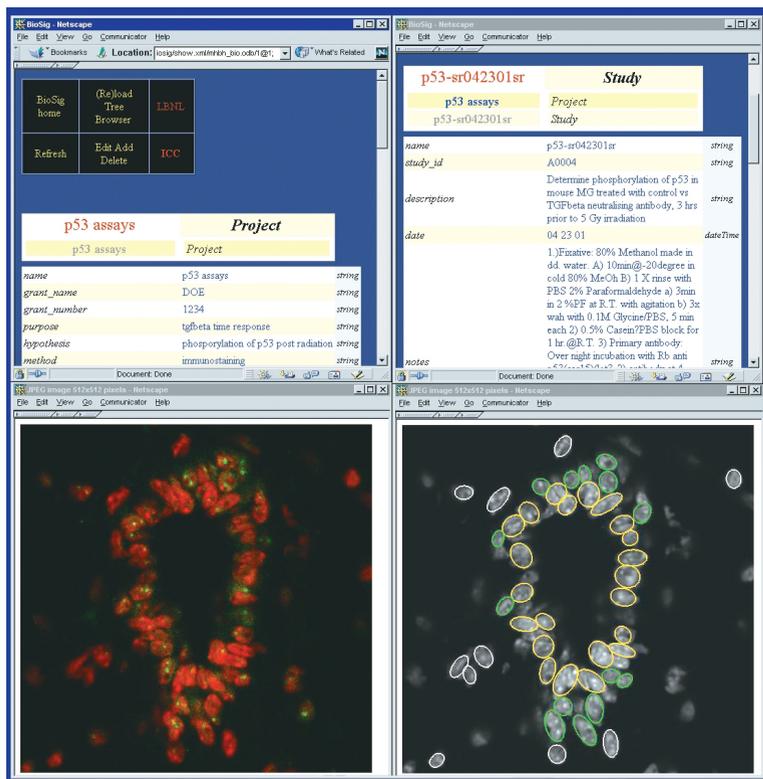


Figure 3b. Client's view of raw and processed images and their corresponding annotations for an *in vivo* study. The image on the left combines two images captured with red and green excitation frequencies. The protein expression, shown in green, is heterogeneous for cells in the immediate vicinity of the lumen, which is marked by a dark hole surrounded by cells. The image on the right shows cellular classification as a function of proximity to the lumen.

the average behavior of an image collection at each time point. We use an *average behavior* operator that utilizes indices corresponding to computed features such as morphological or protein-localization attributes to retrieve desired samples.

NUCLEI EXTRACTION

An important step in mapping protein localization into structural components for phenotypic studies involves automatic delineation of nuclei from each image in the database, a process known as *segmentation*. In addition, for *in vivo* studies, the system also needs to classify each cell type based on its spatial location in tissue so that quantitative analysis is cell specific. Segmentation presents a difficult problem because of noise, technical variations in sample preparation, and overlapping of adjacent nuclei, which forms a clump. Noise can be random or speckled. While the imaging device often causes random noise, speckled noise corresponds to tiny localized substructures such as chromatin, which stand strongly against the diffused nuclear signature in the presence of a fluorescent dye.

Our segmentation algorithm is model-based and assumes that any projection of the 3D nuclear structure is locally quadratic along its boundary. Our system automatically extracts speckled noise with elliptic features and then uses a harmonic cuts technique to interpolate them. As Figure 4 shows, once we obtain a smooth representation of the image, we use a centroid transform technique⁵ to partition adjacent nuclei. Centroid transform is a newly developed technique that collapses the content of each nucleus into its localized center of mass (<http://vision.lbl.gov/Projects/BioSig>).

Phenotyping often involves using multispectral imaging to couple morphological features with protein localization or physiological responses. Researchers usually tag a sample with a fluorescent dye and image it to reveal the nuclear formation's shape and organization. They then image protein expression using fluorochromes that are excited at other wavelengths, which are separated using specific optical filters. Our system represents each delineated nucleus with an ellipse and a hyperquadric. A hyperquadric representation is more powerful because it captures detailed morphological features.⁶

BioSig constructs a representation of the underlying image in the form of region adjacency graph in which each node corresponds to a nucleus and its attributes—such as shape and protein expression—and each link corresponds to the relationship between neighboring nuclei. The graph-based

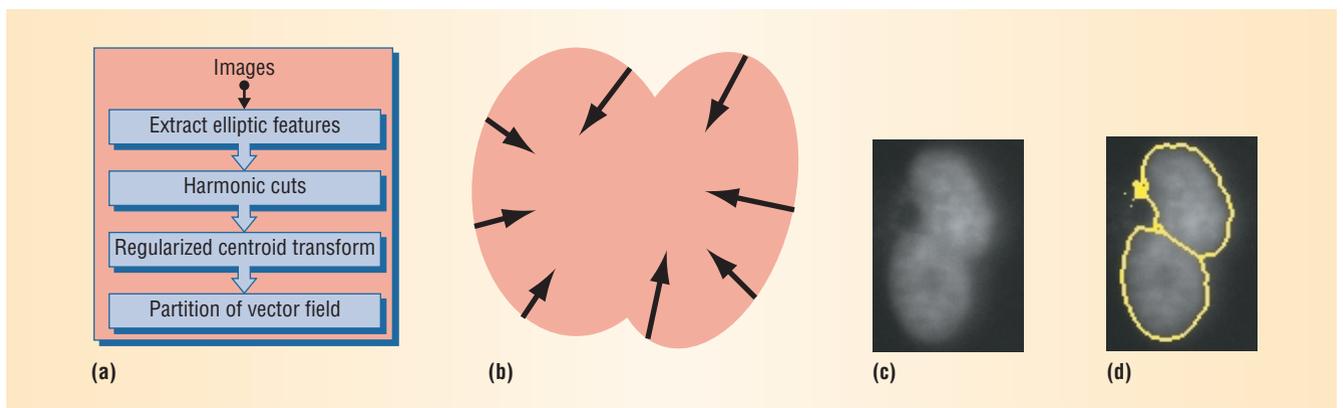


Figure 4. Detection and removal of noisy regions. The segmentation protocol detects and removes noisy regions with elliptic features, interpolates noisy regions with harmonic cuts, and separates touching compartments with centroid transform: (a) the protocol for extracting touching nuclei; (b) graphical representation for evolution of a centroid transform between two adjacent nuclei; (c) a pair of touching nuclei; and (d) the result of partitioning adjacent nuclei.

representation is information-preserving and facilitates many types of higher-level data analysis. For example, for in vivo studies, we need to classify each nucleus in the image with respect to its position in the lumen—a dark region surrounded by cells. We base this classification on locating the lumen inferred from the region-adjacency graph and labeling the cells in the lumen’s immediate vicinity as the luminal epithelial cells. We then classify the cells furthest from the lumen as stromal cells.

Figure 3b shows examples of ellipse-fitting and nuclear classification with coded colors. Note that protein expression, represented by the green color on the left in Figure 3b, is neither diffused in the nuclear region nor homogeneous in cells with the same classification. Thus, detailed quantitative analysis of textured protein expression across different cellular classifications and under a variety of experimental conditions can render new insights, a feature that BioSig supports.

To study the effect of microenvironment on cell-cell communication, it is often necessary to compute the structure of a colony in 3D. We have extended our segmentation algorithm to investigate the structural integrity of a colony as a function of the number of gap junctions. Figure 5 shows one such example, where nuclear structure and gap junctions are imaged at 490 nm and 570 nm, respectively. In this figure, gap junctions correspond to impulses in the cytoplasmic region. We have developed computational tools to extract these impulses and use nuclear regions as context to filter out potential false alarms.

APPLICATIONS

Two applications show BioSig in use. The first corresponds to cell culture studies involving cell-to-cell communication and adhesion. The second application provides the basis for establishing a link between extracellular events and intracellular sig-

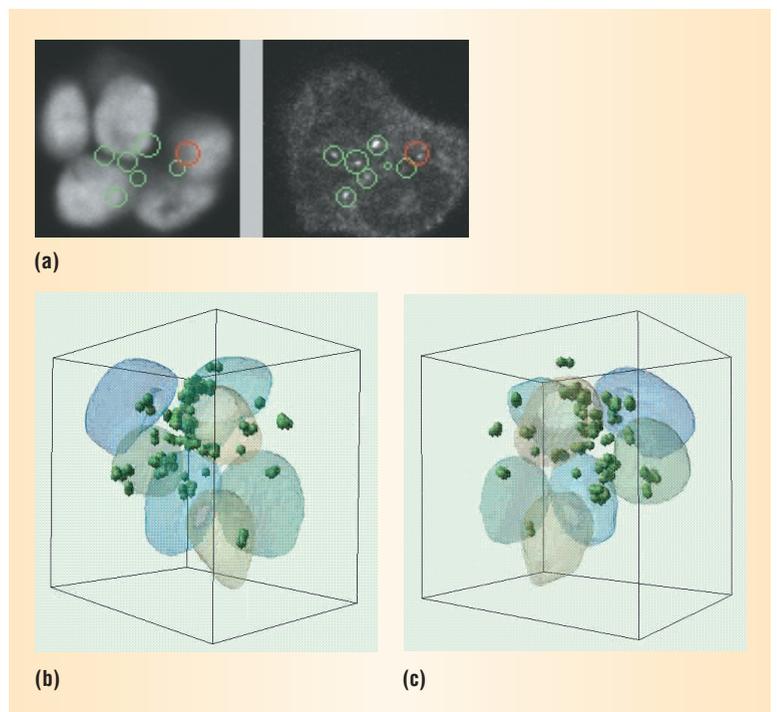


Figure 5. Extension of 2D segmentation to 3D data. (a) One slice of 3D colony observed with confocal microscope shows nuclei formation at 490 nm gap junctions corresponding to cell-cell communication at 570 nm; computed views of (b) the cultured colony and (c) their gap junction.

naling for normal (wild type) versus genetically altered animals.

In vitro studies

During cell culture studies, a single luminal epithelial cell divides to form a hollow sphere known as an acinus. This process often takes 10 days, during which, at different times, we disrupt the microenvironment to study cell-to-cell communication.

To determine whether low-dose radiation promotes aberrant extracellular matrix (ECM) interactions, we used BioSig to examine integrin and

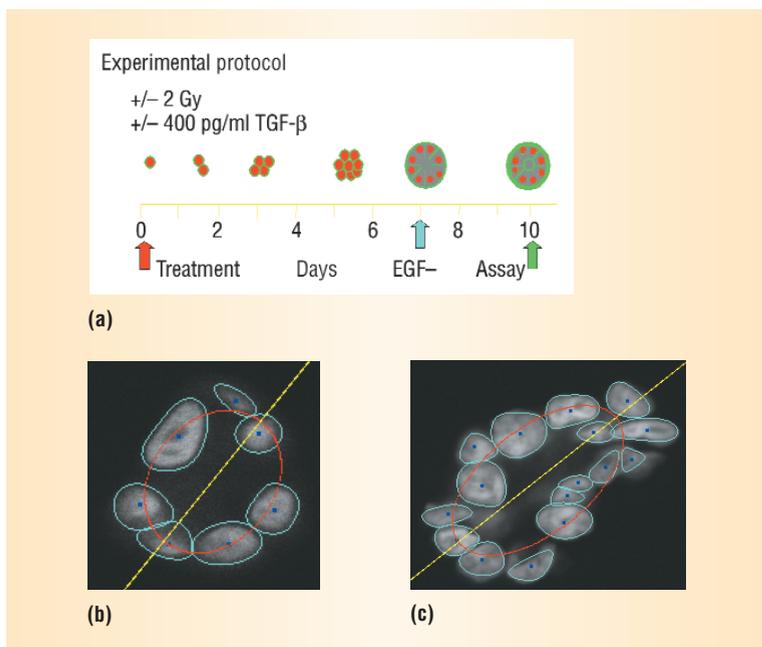


Figure 6. Colony organization resulting from low-dose radiation and TGFβ treatment, which indicates a lack of symmetry around the lumen. Nuclei are segmented—represented with hyperquadrics—and symmetry is measured by fitting an ellipse to all nuclei: (a) experimental protocol; (b) an untreated sample maintains symmetry along the lumen; and (c) a treated sample loses its symmetric organization.

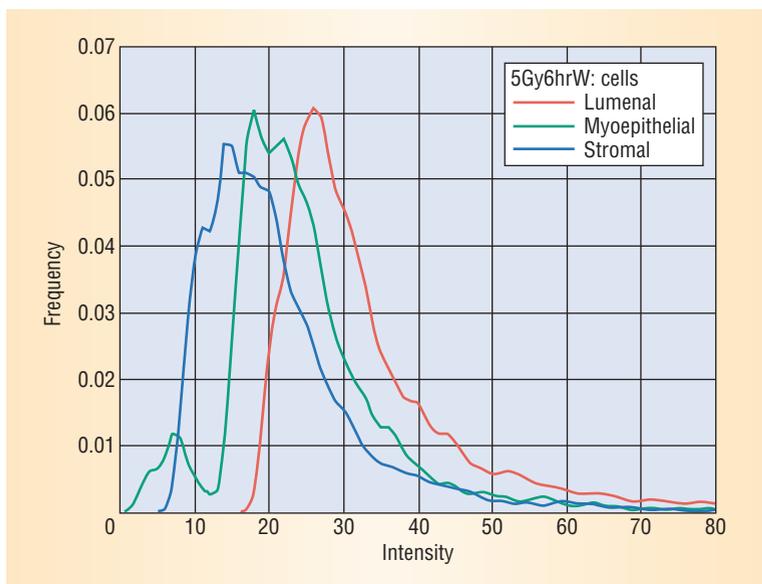


Figure 7. Population studies for p53. The studies show probability density functions for p53 response in each cell type, with red denoting luminal, green denoting myoepithelial, and blue denoting stromal.

E-cadherin localization in preneoplastic human cells that survive radiation, as Figure 6 shows. Integrins are a family of epithelial receptors for the ECM, while E-cadherin maintains normal cell-to-cell interactions and architecture.

We used the HMT-3522 (S1) human breast-cell line cultured within a reconstituted ECM. These

genomically unstable cells are phenotypically normal in that they recapitulate normal mammary architecture in the form of a multicellular, 3D acinus.⁷ These clusters express integrins in a polarized fashion and develop an organized ECM over the course of seven to 10 days in culture. As Figure 6a shows, this study examines the consequences of exposing these cells to ionizing radiation and TGFβ, a known protein modifier.

Using a green fluorescent label, we detected antibodies to E-cadherin, beta 1 integrin, or alpha 6 integrin, while we counterstained nuclei with a red fluorescent DNA dye. Cells that survived either 2 Gy or TGFβ showed decreased beta-1 or alpha-6 integrin localization, respectively. As Figure 6c shows, however, when we exposed cells to both radiation and TGFβ, additional perturbations occurred, disorganizing the clusters, halting integrin polarization at the cell surface, and blocking E-cadherin expression—all indicating a lack of structural organization. BioSig enables mapping experimental variables, such as radiation dosage and presence of growth factors, to biological end points, such as overexpression or underexpression of integrins, which are computed from multispectral images.

In vivo studies

One of the most rapid cellular responses to low-dose radiation is the activation of the transcription factor p53, which is involved in orchestrating DNA repair, and whose abundance and action dictate individual cellular consequences regarding proliferation, differentiation, and apoptosis. Described as the “guardian of the genome,” p53 is one of the most rapid cellular responses to radiation. Protein modifications such as phosphorylation dictate the activation of p53, allowing it to bind to DNA and to transactivate target genes.⁸ A major cellular function of the p53 tumor suppressor protein is its role in promoting genome integrity.

We designed an experiment to examine the distribution of p53 nuclear immunoreactivity using optical microscopy. We imported experimental variables, images, and their corresponding quantitative analysis into BioSig. We then queried specific features to track the level and distribution of p53 within specific tissue compartments. In the result shown in Figure 7, BioSig provides a visual representation of p53 expression in three categories of nuclei—red for luminal epithelial, cyan for myoepithelial, and blue for stromal cells—for a population of images from wild-type tissue sections.

In another experiment, we studied p53 expression as a result of radiation exposure in genetically

altered—transgenic—strains of mice. We exposed normal mice—the control animals—and transgenic mice to radiation, collected tissues, treated the samples with antibodies, generated a large volume of images and annotation data, and stored the images and data in a database. Our studies demonstrated that less p53 is localized in the luminal epithelial cells of genetically altered mice, demonstrating a link between the mouse genetic background and an intracellular response to radiation.

We are using the BioSig informatics approach to microscopy and quantitative image analysis 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. We have posted the details of our data model, database usage, and methods for importing legacy data into a database on the Web, and we will soon facilitate public access to certain aspects of this database. ■

Acknowledgements

Our research was funded in part by the Low-Dose Radiation Program of the Life Sciences Division; Medical Sciences Division; the Mathematical, Information, and Computing Sciences Division; and the Director of Lawrence Berkeley National Laboratory of the US Department of Energy under contract number DE-AC03-76SF00098 with the University of California. LBNL publication number LBNL-47456.

References

1. C.D. Roskelley, A. Srebrow, and M.J. Bissell, "A Hierarchy of ECM-Mediated Signaling Regulates Tissue-Specific Gene Expression," *Current Opinions in Cell Biology*, vol. 7, no. 5, 1995, pp. 736-747.
2. M.H. Barcellos-Hoff, "How Do Tissues Respond to Damage at the Cellular Level? The Role of Cytokines in Irradiated Tissues," *Radiation Research*, vol. 150, 1998, pp. 109-120.
3. F. Wang et al., "Reciprocal Interactions between Beta 1-Integrin and Epidermal Growth Factor Receptor in Three-Dimensional Basement Membrane Breast Cultures: A Different Perspective in Epithelial Biology," *Proc. Nat'l Academy of Sciences of United States of America*, vol. 95, no. 25, 1998, pp. 14,821-14,826.
4. F.G. Giancotti and E. Ruoslahti, "Integrin Signaling," *Science*, vol. 285, 1999, pp. 1028-1032.
5. Q. Yang and B. Parvin, "Harmonic Cuts and Regularized Centroid Transform for Localization of Subcellular Structures," *Proc. Int'l Conf. Pattern Recognition*, IEEE Press, Piscataway, N.J., 2002.
6. B. Parvin et al., "BioSig: A Bioinformatic System for Studying the Mechanism of Inter-Cell Signaling," *Proc. IEEE Int'l Symp. Bio-Informatics and Biomedical Engineering*, IEEE Press, Piscataway, N.J., 2000, pp. 281-288.
7. V.M. Weaver et al., "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, 1996, pp. 833-851.
8. A.J. Levine, "P53, the Cellular Gatekeeper for Growth and Division," *Cell*, vol. 88, 1997, pp. 323-331.

Babram Parvin is a staff scientist in Computing Sciences at Lawrence Berkeley National Laboratory (LBNL). He conducts research on computer vision, feature-based representation of time varying scientific images, and distributed informatics for visual servoing. Parvin received a PhD in electrical engineering from the University of Southern California. He is a senior member of the IEEE. Contact him at parvin@media.lbl.gov.

Qing Yang is a computer scientist in Computing Sciences at LBNL. His research interests include image processing, pattern recognition, and bioinformatics. Yang received a PhD in computer science from the Institute of Automation, Chinese Academy of Sciences. He is a member of the IEEE. Contact him at qyang@media.lbl.gov.

Gerald Fontenay is a computer scientist in Computing Sciences at LBNL. His research interests include database development for scientific data, distributed computing, Web-based applications, and object-oriented design. Fontenay is completing a BS in computer science from San Francisco State University. Contact him at fontenay@media.lbl.gov.

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. Contact her at MHBarecellos-Hoff@lbl.gov.