INTEGRATED PROFILING OF CELL SURFACE PROTEIN AND NUCLEAR MARKER FOR DISCRIMINANT ANALYSIS

Ju Han, Hang Chang, Kumari Andarawewa, Paul Yaswen, Mary Helen Barcellos-Hoff, and Bahram Parvin

Lawrence Berkeley National Laboratory, Berkeley, CA 94720

ABSTRACT

Cell membrane proteins play an important role in tissue architecture and cell-cell communication. We hypothesize that segmentation and multivariate characterization of the distribution of cell membrane proteins, on a cell-cell basis, enable improved classification of treatment groups and identify important characteristics that can otherwise be hidden. We have developed a series of computational steps to (i) delineate cell membrane protein signals and associate them with specific nuclei, (ii) compute a coupled representation of the multiplexed DNA content with membrane proteins and other end points, (iii) rank computed features associated with such a multivariate representation, (iv) visualize selected features for comparative evaluation, and (v) discriminate between treatment groups in an optimal fashion. The novelty of our method is in the segmentation of the membrane signal and the multivariate representation of phenotypes on a cell-cell basis. To test the utility of the new method, the proposed computational steps were applied to images of cells that have been irradiated with different radiation qualities in the presence and absence of TGFβ. These samples are labeled for their DNA content and E-cadherin membrane protein. We demonstrate that multivariate representation of cell-cell phenotypes improves predictive and visualization capabilities among different treatment groups, and increases quantitative sensitivity of cellular responses.

***Index Terms*—** Multivariate analysis of imaging assay, Cadherin, Segmentation, Irradiation

1. INTRODUCTION

Cell surface proteins regulate cell-cell interactions and physical properties of tissues. E-cadherin is one such a calcium-dependent cell adhesion molecule that influences differentiation and tissue structure. It forms adherens junctions between epithelial cells and communicates with the actin cytoskeleton through associated intracellular proteins. As an endpoint, E-cadherin has been studied extensively, since it appears to function as a barrier to cancer. Loss of E-cadherin has been associated with (i) increased motility, (ii) cancer progression and metastasis, and (iii) increased resistance to cell death [1]. Since down-regulation of E-cadherin is an important endpoint for quantitative systems biology, we hypothesize that detailed quantitative representation of the E-cadherin signals provide important clues for understanding the effects of different biological perturbations. Furthermore, we reason that representation of E-cadherin on a cell-cell basis, coupled with morphological and structural features obtained by other imaging probes, will provide a multivariate representation that can be mined to improve predictive capability. This paper introduces a novel method for characterizing the E-cadherin signal on a cell-cell basis and demonstrates that multivariate representation of imaging data is useful for (i) characterizing heterogeneity, (ii) identifying features that are not visually obvious to human observers, (iii) reducing the number of imaging probes that are needed for differentiating phenotypes associated with different sets of experimental treatments, and (iv) visualizing multivariate representation of localization data in the same way that expression data are presented.

With the exception of [2], few studies have been published quantifying membrane signals for high content screening. Even in this case, few details of the methodology are provided. Furthermore, biological samples are limited to HeLa cells that are known to have well-behaved size and shape features. Complexities and challenges associated with quantifying cell surface protein patterns originate from (i) background variation, (ii) non-uniformity of the signal, (iii) non-uniformity in the width and strength of the signal, and (iv) non-uniformity in nuclear size and shape features. More elaborate multivariate representations have been proposed [3, 4]; however, analyses do not target cell surface proteins on a cell-cell basis.

2. TECHNICAL APPROACH

Our computational protocol avoids traditional ad-hoc steps in favor of model-based geometric methods to delineate subcellular regions, associate cell surface protein signals with particular cells, and drive a multivariate representation of each cell for further analysis. By this method, the E-cadherin signal is coupled with labeled nuclear regions so that context can be established. The protocol consists of five major steps: nuclear segmentation, E-cadherin localization, feature extraction, feature selection, and discriminant analysis, as shown Fig. 1. First, each nucleus is localized using an edge-based method, and then grouped subject to convexity and continuity. Unlike thresholding, edge-based methods have an improved immunity to non-uniformity in the fluorescent signal, thus providing a more robust and accurate delineation of nuclear boundaries. Second, the E-cadherin signal is inferred through an iterative voting method with respect to continuity. A unique aspect of this technique is in the topography of the voting kernel, which is iteratively refined and reoriented. The technique clusters and groups membrane signals along the tangential direction. It has an excellent noise immunity and is tolerant to perturbations in scale. Furthermore, the membrane signal is registered with the corresponding nuclear region by an evolving front. Third, approximately 400 features corresponding to morphology (e.g., size, aspect-ratio, bending energy of contour), structure (e.g., texture features), localization (e.g., fluorescent intensity and its derived features) and organization (e.g., relationship between cells represented as an attributed graph) are computed for each cell. These measurements are stored in a schema that captures their relationship, order, and cardinality. Fourth, a minimal subset of features from the full multivariate representation is selected and ranked to maximize class separability, where classes correspond to different treatment groups. Finally, the discriminating and predictive capability of an optimal feature subset is evaluated using the Holdout method and

The Research was supported by National Aeronautics and Space Administration Grant No. T6275W, NASA Specialized Center for Research in Radiation Health Effects, and by the Low Dose Radiation Research Program and the Medical Imaging Program, Office of Biological Effects Research, U.S. Department of Energy, Grant No. DE-FG03-01ER63240.

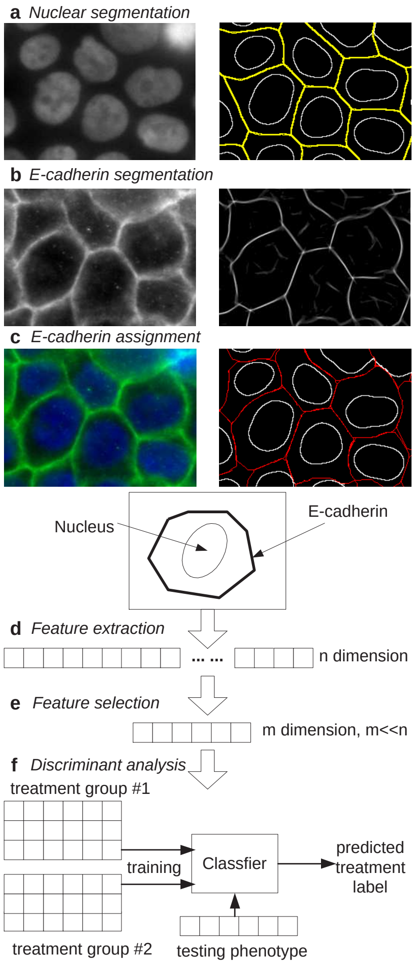


Fig. 1. Multivariate representation of nuclear and E-cadherin responses for discriminant analysis of iron and gamma irradiation.

linear discriminant analysis (LDA) classifiers.

2.1. Nuclear segmentation

Nuclear segmentation enables context for quantifying localization and other structural and morphological features on a cell-cell basis. However, as a result of sample preparation and fixation, fluorescent signals of adjacent nuclear regions overlap and form a clump. It is important to quantify the phenotypic signature at the individual cell level by partitioning a clump of cells. Initially, our computational protocol delineates isolated nuclear regions. Next, it partitions touching cells by applying a series of geometric constraints [5]. The basic idea is that nuclear geometry is almost convex, and that at the intersection of the overlapping boundaries, folds (points of maximum curvature) are formed. Thus, by grouping folds that are formed by a closed contour, a convex partition can be inferred. The technique is iterative, and has been shown to be effective in segmenting touching nuclei.

2.2. Segmentation of E-cadherin signals

There are two critical steps in segmentation of the membrane-bound protein. In the first step, the membrane-bound protein is accentuated and enhanced. In the second step, the membrane-bound protein is assigned to individual nuclear features. The first step utilizes iterative tangential voting to remove noise, enhance signal, and complete perceptual gaps. The second step leverages evolving fronts, in the context of the nuclear region, for the assignment process.

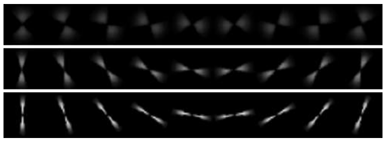


Fig. 2. A sample of kernel topography: Oriented kernels for inference of continuity are bidirectional, and their energy dissipates as a function of distance. Initially, the energy is dispersed (top row), but becomes more focused (bottom row).

2.2.1. Iterative tangential voting

The membrane signals correspond to the negative curvature maxima at a given scale within the image space. But curvature features are noisy and may suffer from undesirable artifacts. The process is initiated by voting with a Gaussian kernel at each image feature point. Let $F(x_o, y_o)$ be the curvature feature at location (x_o, y_o) in the image. Let (x_n, y_n) be a point in the neighborhood of (x_o, y_o) that can be influenced with a kernel applied at position (x_o, y_o). The initial voted image is then represented as

$$V(x_n, y_n) = \sum_{(x_n,y_n)\in Neighbor(x_o,y_o)} \{F(x_o, y_o) * G_{(x_o,y_o)}(\sigma)\}$$

The refinement of the voted image is iterative, involving application of a more focused kernel at the next iteration along the direction $\alpha = \arctan\{(V_{yy} - K_{max})/V_{xy}\}$, where V_{yy} and V_{xy} are the local derivatives of the voted image, and K_{max} is the maximum curvature computed from the Hessian of the voted image. The shape of the kernels, shown in Fig. 2, indicates whether the energy distribution of the kernel is focused or dispersed. Initially, the energy is dispersed; however, at each consecutive iteration, the energy becomes more focused and at the same time the kernel orientation is redirected along the direction of maximum response. Details of iterative tangential voting can be found in [6].

2.2.2. Evolving fronts

The next step of the computational process is to design an initial condition, and define additional constraints for robust segmentation. The initial condition is derived from the region of the space identified by the segmented DNA stain, presented earlier. This is based on region-based Voronoi tessellation of the nuclear mask, which generates a curvilinear partition of the image space, shown in Fig. 1a. Let K be the total number of nuclear region, N_i, in the image. The Voronoi region is defined by $V_i = \{p|dist(p, N_i) < dist(p, N_j), j \in \{0, 1, \ldots, K-1\}, j \neq i\}$, in which $dist(p, N_i)$ is the distance between point p and the nuclear region N_i, $dist(p, N_i) = \min_{q\in N_i} |p - q|$.

Initiating from the Voronoi region, assignment of the cell surface protein is computed by optimizing an evolving front where external forces are defined by the gradient vector field [7]:

$$E = \int_0^1 \frac{1}{2}(\alpha|X'(s)|^2 + \beta|X''(s)|^2) + E_{ext}(X(s))ds \qquad (1)$$

where, $X(s) = [x(s), y(s)], s \in [0, 1]$, is the curve representation. The first and second terms ensure smoothness through stretching and bending. The third term attracts the curve towards a derived representation of the cell surface protein marker, which is a function of the voted image. The evolving front corresponds to

$$X(s,t) = \alpha X''(s,t) - \beta X'''(s,t) - \nabla E_{ext} \qquad (2)$$

where, $\nabla E_{ext} = -V$ and $V(x,y) = (u(x,y), v(x,y))$ is the gradient vector flow that minimizes the energy functional

$$\epsilon = \int\int \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + |\nabla f|^2 |V - \nabla f|^2 dxdy \quad (3)$$

where $f(x,y)$ is a skeletonized representation of the voted image.

A sample example is selected from our dataset to demonstrate the behavior of iterative tangential voting. The results are shown in Fig. 1. In addition to iterative tangential voting, segmentation through evolving fronts is also demonstrated.

2.3. Multivariate representation of cellular features

Phenotypic signatures are computed from every imaging probe that labels an organelle or localization of a specific protein. In this case, three distinct feature sets of morphology, structure, and fluorescence signal are extracted from each marker. For example, in the case of a marker associated with the nuclear region, morphological features of shape (e.g., area, aspect ratio, axis), bending energy computed from curvature of bounding contour are also extracted at multiple scales. Structural features correspond to textural attributes that are detected from first, second, and third order derivatives of oriented Gaussian filters [8]. These oriented filters capture responses of inherent image features at multiple scales:

$$\begin{aligned}
G_1^\theta &= G_x \cos\theta + G_y \sin\theta \\
G_2^\theta &= G_{xx} \cos 2\theta + 2G_{xy} \sin\theta \cos\theta + G_{yy} \sin 2\theta \\
G_3^\theta &= G_{xxx} \cos 3\theta + 3G_{xxy} \sin\theta \cos 2\theta \\
&\quad +3G_{xyy} \sin 2\theta \cos\theta + G_{yyy} \sin 3\theta,
\end{aligned}$$

where $G_x = \partial G(x,y)/\partial x$, $G_y = \partial G(x,y)/\partial y$, and G(x,y) is a 2-D Gaussian function. Finally, the fluorescence signal is quantified at global and local scales. While global representation relies on average signal within the organelle of interest, local representation characterizes how the fluorescence signal is spatially distributed within the nuclear mask. An example of this representation is the change in the fluorescence signal along the radial direction originating from the center of the mass. Since the texture feature vector is rather large, its dimensionality is reduced through principal component analysis (PCA) for subsequent analysis. We opted not to apply the PCA to the entire representation since the physical meaning of the feature set will be lost during the projection operation. A total of 324 texture features are computed and, through PCA dimensionality reduction, 8 projected features that account for 99% of the total variance are retained for further analysis.

2.4. Feature selection and discriminant analysis

Feature selection ranks the feature sets based on some measure of class separability. Here, the class separability is defined as the ratio of the determinant of mix-class scatter and the determinant of within-class scatter matrices. This measure will take a large value when samples are well clustered around their class means and the clusters of different classes are well separated. The Holdout method is used to evaluate performance of discriminant analysis. In this method, half of the data is randomly selected for training a classifier, and the other half is used for testing. Classification is based on linear discriminant analysis, and the process of sample selection, training, and classification is repeated to assure that the classification performance is not compromised by a specific set of training samples.

Table 1. Experimental variables

Sham			
TGFβ			
Iron irradiation	0.1Gy + TGFβ	Gamma irradiation	0.03Gy + TGFβ
	0.2Gy + TGFβ		0.1Gy + TGFβ
	0.5Gy + TGFβ		0.4Gy + TGFβ
	1Gy + TGFβ		1Gy + TGFβ
	1Gy		2Gy + TGFβ
			2Gy

Table 2. Top-ranked feature combinations for discriminating different cellular phenotypes among all treatment groups. PC stands for principal component.

	Features	Discriminating power
1-feature	(1) Mean E-cadherin signal	2.1832
	(2) Total E-cadherin signal	1.7691
	(3) Nuclear texture PC #2	1.5306
	(4) Variance of E-cadherin signal	1.4724
	(5) Nuclear size	1.2398
	(6) Nuclear texture PC #1	1.2312
2-feature	(1) + (3)	3.3555
	(1) + (4)	2.8346
	(2) + (3)	2.6733
	(1) + (2)	2.6170
	(1) + (6)	2.5661
3-feature	(1) + (3) + (4)	4.2184
	(1) + (3) + (6)	4.0437

3. EXPERIMENTAL RESULTS

The study involved a multifactorial experiment in which radiation of two different qualities (iron, gamma) were applied separately in combination with TGFβ to cultured MCF10A cells. The radiation qualities are varied to examine whether this parameter influences cellular responses independently of toxicity. TGFβ belongs to a family of cytokines and modulates cellular responses to radiation [9]. It was added to mimic an effect of stromal cells on radiation response in tissues. The data set used for this analysis consisted of a total of 13 treatment groups with 20 to 80 images in each group and up to 6000 cells per group. Table 1 summarizes the different treatment groups.

3.1. Evaluation of phenotypic features

Segmentation of the nuclear and E-cadherin signals enables a multivariate representation of the phenotypic signature on a cell-cell basis. These measurements are then aggregated and ranked for comparing different treatment groups. Feature selection identifies an optimum subset of features for classification between different treatment groups, as shown in Table 2. Fisher discriminant ratio is used to measure the feature discriminating power.

3.2. Quantitative comparison of phenotypic variability

Fig. 3 shows loss of E-cadherin in irradiated samples, which is accompanied by an increase in the peakedness (kurtosis) of the distribution. The relationship between the loss of E-cadherin and heterogeneity of the membrane signal is expected; however, Fig. 3 also indicates that at equivalent radiation doses in the presence of TGFβ, loss of heterogeneity with Fe is higher than with γ radiation. This is an observation that can only be quantified through detailed analy-

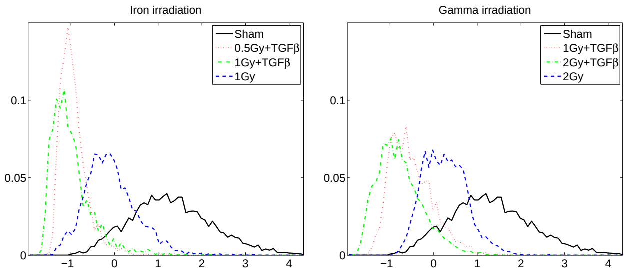
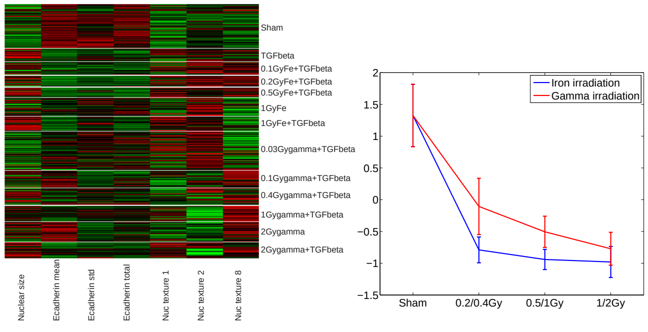


Fig. 3. Distribution and dose response of E-cadherin signal on a cell-cell basis: Samples were treated with indicated doses of iron and gamma irradiation to examine relative biological effects on E-cadherin expression. Iron and gamma irradiation at equal toxicity dosage is shown with the same color in the two figures. Above probability density functions are normalized with $\mathcal{N}(0, 1)$.



(a) (b)

Fig. 4. (a) Heat map of top 7 features with respect to the 13 treatment groups on a cell-cell basis. (b) Dose response of E-cadherin on a cell-cell basis indicates a sharper drop in the membrane protein in low dosage as a result of Fe irradiation. The error-bars correspond to the standard deviation of the signal at each dosage.

sis on a cell-cell basis, and appears to be dependent on the presence of TGFβ. While this analysis is based on quantitation of a labeled probe, the heat map in Fig. 4(a) indicates that size and structural features (e.g., texture) of the nuclear region increase with Fe radiation and TGFβ treatment when compared to of γ radiation. Through cell-cell multivariate analysis, heterogeneity, hidden variables (e.g., size, texture) can be identified and visualized.

3.3. Classification of treatment groups

Table 3 summarizes classification accuracy between using single or multiple features of treated and control groups. In most cases, a single feature is sufficient for discrimination; however, in the absence of TGFβ and presence of a high dosage of γ, additional features can contribute to an improved classification.

Using the same strategy, we evaluated the classification accuracy for Fe and γ irradiation. Results are shown in Table 4. Again combining representations based on quantification of the labeled probe and computed textured features results in an improved predictor for separating treatment groups. It is interesting to note that at 1Gy of Fe versus 2Gy of γ, variation of the E-cadherin signal per cell is also an important indicator for improving classification accuracy. Furthermore, the dose response curves are significantly different, as shown in Fig. 4(b). This quantitative insight can only be revealed through cell-cell analysis. Global florescence analysis hides the differences in the dose-response curves.

Table 3. Classification accuracy of sham versus irradiated samples computed through linear discriminant analysis.

Number of features	1	2	3
Sham vs. 0.5GyFe+TGFβ	90.16%	91.90%	92.75%
Sham vs. 1GyFe+TGFβ	90.58%	93.03%	94.48%
Sham vs. 1GyFe	79.30%	82.57%	85.12%
Sham vs. 1Gyγ+TGFβ	85.96%	93.01%	93.41%
Sham vs. 2Gyγ+TGFβ	89.61%	91.86%	92.71%
Sham vs. 2Gyγ	74.88%	88.18%	90.08%

Table 4. Classification accuracy of Fe versus γ irradiation is computed through linear discriminant analysis.

Number of features	1	2	3
0.5GyFe+TGFβ vs. 1Gyγ+TGFβ	72.73%	89.42%	89.49%
1GyFe+TGFβ vs. 2Gyγ+TGFβ	59.36%	71.02%	71.80%
1GyFe vs. 2Gyγ	82.29%	92.87%	97.39%

4. CONCLUSIONS

Our computational protocol generates a coupled multivariate representation of the spatial features for each cell and the membrane bound proteins exhibiting continuous fluorescent signals along cell surface boundaries. Experimental results demonstrate that multivariate representation of cell-cell phenotypes improves predictive and visualization capabilities among different treatment groups, and increases quantitative sensitivity of cellular responses.

5. REFERENCES

[1] U. Cavallaro and G. Christofori, "Cell adhesion and signalling: implications for tumour progression," *Nat Rev Cancer*, vol. 11, no. 12, pp. 118–32, 2004.

[2] N.L. Prigozhina and L. Zhong et al., "Plasma membrane assays and three-compartment image cytometry for high content screening," *Assay Drug Dev Technol*, vol. 11, no. 12, pp. 29–48, 2007.

[3] L.H. Loo, L.F. Wu, and S.J. Altschuler, "Image-based multivariate profiling of drug responses from single cells," *Nature Method*, vol. 4, no. 5, pp. 445–453, 2007.

[4] M.R. Lamprecht, D.M. Sabatini, and A.E. Capenter, "Cellprofiler: free, versatile software for automated biological image analysis," *Biotechniques*, vol. 42, pp. 71–75, 2007.

[5] S. Raman and C.A. Maxwell et al., "Geometric approach to segmentation and protein localization in cell culture assays," *J Microscopy*, vol. 225, no. 1, pp. 22–30, 2007.

[6] H. Chang and K.L. Andarawewa et.al, "Perceptual grouping of membrane signals in cell-based assays," in *Proc IEEE Int Symp on Biomedical Imaging*, 2007, pp. 532–535.

[7] C. Xu and J.L. Prince, "Gradient vector flow: A new external force for snakes," in *Proc CVPR*, 1997, pp. 66–71.

[8] R.A. Young, R.M. Lesperance, and W.W. Meyer, "The gaussian derivative model for spatial-temporal vision: I. cortical model," *Spatial Vision*, vol. 14, pp. 261–319, 2001.

[9] M.H. Barcellos-Hoff, "Radiation-induced changes in transforming growth factor e1 and subsequent extracellular matrix reorganization in irradiated murine mammary gland," *Cancer Res*, vol. 53, pp. 3880–3886, 1993.