BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 8 2009, pages 1070–1075
doi:10.1093/bioinformatics/btn426

Data and text mining

Graphical methods for quantifying macromolecules through bright field imaging

Hang Chang[1,2,*], Rosa Anna DeFilippis[3], Thea D. Tlsty[3] and Bahram Parvin[1,4]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [2]Institute of Automation, Chinese Academy of Sciences, Beijing, China, [3]Department of Pathology, University of California, San Francisco and [4]Department of Electrical Engineering, University of California, Riverside, USA

Received on March 12, 2008; revised on August 8, 2008; accepted on August 10, 2008

Advance Access publication August 14, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Bright field imaging of biological samples stained with antibodies and/or special stains provides a rapid protocol for visualizing various macromolecules. However, this method of sample staining and imaging is rarely employed for direct quantitative analysis due to variations in sample fixations, ambiguities introduced by color composition and the limited dynamic range of imaging instruments. We demonstrate that, through the decomposition of color signals, staining can be scored on a cell-by-cell basis. We have applied our method to fibroblasts grown from histologically normal breast tissue biopsies obtained from two distinct populations. Initially, nuclear regions are segmented through conversion of color images into gray scale, and detection of dark elliptic features. Subsequently, the strength of staining is quantified by a color decomposition model that is optimized by a graph cut algorithm. In rare cases where nuclear signal is significantly altered as a result of sample preparation, nuclear segmentation can be validated and corrected. Finally, segmented stained patterns are associated with each nuclear region following region-based tessellation. Compared to classical non-negative matrix factorization, proposed method: (i) improves color decomposition, (ii) has a better noise immunity, (iii) is more invariant to initial conditions and (iv) has a superior computing performance.

contact: hchang@lbl.gov

1 INTRODUCTION

Macromolecules (proteins, nucleic acids, lipids and carbohydrates) can be rapidly visualized in cells and tissue via staining with antibodies and/or special stains, followed by bright field color imaging. However, the quantitative analysis of such images is often hindered by variations in sample preparations, the limited dynamic range of color cameras, and the fact that image formation is not at a specific excitation and emission frequency, which is the hallmark of fluorescence microscopy. Through consistent sample preparation, fixation and imaging, we suggest that the signals associated with a macromolecule can be decomposed in the color space, and can render a scoring value on a cell-by-cell basis. Following this protocol, protein, lipid and DNA complexes are visualized with antibodies and special stains, and then imaged with a color CCD camera attached to a microscope.The key contributions of this article are in: (i) formulating the color decomposition as a global optimization problem, (ii) representing the signal complexes, associated with protein localization, with multiple prior models and (iii) applying the proposed method to the analysis of an end point on a cell-by-cell basis. In this context, global optimization is realized through the graph cut method, multiple prior models are specified through user initialization, and signal analysis, on a cell-by-cell basis, is established through a best effort in establishing cellular boundaries. The logical flow of these various computational steps is shown in Figure 1, whereby the user first specifies regions associated with positive staining in an image, the nuclear regions are then automatically detected as a dark elliptic region (Yang and Parvin, 2003), and are later further refined following color decomposition. The morphology and position of nuclear features allow the region-based tessellation of the image, and the subsequent scoring of the signaling complex on a cell-by-cell basis.

We applied our method to fibroblasts grown from histologically normal breast tissue biopsies obtained from women from two distinct populations. The biopsies were digested in solution and the fibroblasts purified and grown *in vitro*. These fibroblasts were then grown under conditions that support adipocyte differentiation for 5–7 days before being fixed and stained with *hematoxylin* and *Oil Red O*, which stain DNA and lipids, respectively. Although hematoxylin and Oil Red O visualize nuclei in blue and lipids in red, respectively, there is still some overlap in the color space.

This article has been organized as follows: Section 2 reviews previous research in the area of color decomposition from histologically stained tissues; Section 4 demonstrates the effectiveness of our method when stains are co-localized; Section 3 provides the details of our method; Section 5 summarizes the results of our method and the application of our method to a large dataset and Section 6 concludes the article.

2 REVIEW OF PREVIOUS WORK

Current practices in the quantitative assessment of histological samples fall into two categories: standard imaging microscopes and specialized systems. Standard imaging microscopes use a color CCD camera and non-coherent light. The images are scored with non-negative matrix factorization (NMF) (Rabinovich *et al.*, 2003), which models each dye as an additive factor in the color space, where

[*]To whom correspondence should be addressed.

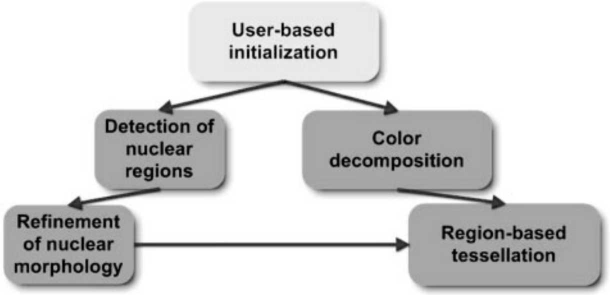
User-based initialization
Detection of nuclear regions
Color decomposition
Refinement of nuclear morphology
Region-based tessellation

Fig. 1. Computational steps in quantifying stained samples: in a single image, the user initializes the stained region associated with a signaling macromolecule. Learned parameters are subsequently used for the rest of the dataset.

the resulting image is deconvolved by color unmixing. This method is quite powerful; however, it ignores spatial relationships between nearby pixels, is sensitive to the initial condition, and may not be able to deconvolve dyes that have too much overlap in the color space. In addition to its successful application to spatial data (Lee and Seung, 1999), NMF has also been used in the analysis of gene expression data to reveal an intuitive meaning in terms of a small subset of metagenes (Gao and Church, 2005). Specialized systems (Papadakis *et al.*, 2003) leverage tunable illumination, fast hyperspectral imaging and monochromatic CCD cameras. The primary advantage of such an optical band pass is in its ability to resolve different stains whose spectra overlap when using the standard system. However, it still requires additional processing, from a stack of band pass images, for cell-by-cell analysis.

Our approach relies on a standard microscope with a color CCD camera (e.g. bright field imaging) to demonstrate that the proposed method: (i) improves color decomposition, (ii) has better noise immunity and (iii) has superior performance.

3 APPROACH

In our computational protocol, nuclear segmentation and color decomposition proceed in parallel. Although nuclear segmentation is not the focus of our article, it is an important step in constructing a system, and a method is outlined here for completeness. However, this method can be replaced with other nuclear segmentation methods. In this system, segmentation of nuclear regions is realized by detecting elliptic features (Yang and Parvin, 2003) corresponding to potential dark regions. These regions are further filtered for their intensities and shape features. At the same time, the amount of staining is characterized by a graph cut algorithm through color decomposition. In rare cases, samples may be locally corrupted with foreign materials leading to small dark patches that resemble the nuclear signature. These dark patches are further filtered following color decomposition to eliminate false association to the nuclear stain. Finally, staining is associated with each nucleus through region-based tessellation.

3.1 Nuclear segmentation

Examples of treated and positive control images are shown in Figure 2a and b, respectively. The nuclear intensities, in the color space, are quite similar to the surrounding background, thus hindering the use of traditional delineation based on the color features. However, by converting the color image into a gray-level image, distinct intensity features of the nuclear region are accentuated through a decrease in the intensity magnitude. This decrease in the intensity feature has a trough that can be detected as a dark elliptic feature. Conversion to a gray-level image is as follows:

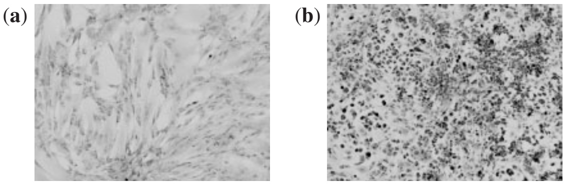

Fig. 2. Images of (**a**) human mammary fibroblasts and (**b**) mouse pre-adipocytes (positive control) grown under conditions that support adipocyte differentiation for 5–7 days before being fixed and stained with hematoxylin and Oil Red O.

$$I(x,y)=\alpha * R(x,y)+\beta * G(x,y)+\gamma * B(x,y)$$

where R, G and B are the red, green and blue channels, respectively, of the original color image, and $\alpha=0.21$, $\beta=0.72$ and $\gamma=0.07$. Unlike immunofluorescence labeling, thresholding is inadequate for this class of images. Our approach is to detect elliptic features (Yang and Parvin, 2003) for the delineation of dark regions in the image. Let the linear scale-space representation of the original image $I_0(x,y)$ at scale σ be given by:

$$I(x,y;\sigma)=I_0(x,y)*G(x,y;\sigma) \tag{1}$$

where $G(x,y;\sigma)$ is the Gaussian kernel with a SD of σ. For simplicity $I(x,y;\sigma)$ is also denoted as $I(x,y)$ below. At each point (x,y), the iso-intensity Contour is defined by:

$$I(x+\Delta x,y+\Delta y)=I(x,y) \tag{2}$$

where $(\Delta x,\Delta y)$ is the displacement vector. Expanding and truncating the above equation using Taylor's series, we have the following estimation:

$$\frac{1}{2}(\Delta x,\Delta y)H(x,y)(\Delta x,\Delta y)^T+(I_x,I_y)(\Delta x,\Delta y)^T=0 \tag{3}$$

where

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

is the Hessian matrix of $I(x,y)$. The entire image domain is divided by Equation (2) into two parts:

$$I(x+\Delta x,y+\Delta y)>I(x,y) \tag{4}$$

and

$$I(x+\Delta x,y+\Delta y)<I(x,y) \tag{5}$$

or locally

$$\frac{1}{2}(\Delta x,\Delta y)H(x,y)(\Delta x,\Delta y)^T+(I_x,I_y)(\Delta x,\Delta y)^T>0 \tag{6}$$

and

$$\frac{1}{2}(\Delta x,\Delta y)H(x,y)(\Delta x,\Delta y)^T+(I_x,I_y)(\Delta x,\Delta y)^T<0 \tag{7}$$

If $H(x,y)$ is positive definite, then the region defined by Equation (4) is locally convex. Similarly, if $H(x,y)$ is negative definite, then the region defined by Equation (5) is locally convex. To determine whether $H(x,y) > 0$ or $H(x,y) < 0$, we analyze this feature in both cases: (I) $H(x,y) > 0$. Then $I_{xx} > 0$, $I_{yy} > 0$, and hence $I_{xx} + I_{yy} > 0$, and positive Laplacian means that (x,y) is a 'dark point', i.e. a point that is darker than its neighbors; and (II) $H(x,y) < 0$. Then $I_{xx} < 0$, $I_{yy} < 0$, and hence $I_{xx} + I_{yy} < 0$, and negative Laplacian means that (x,y) is a 'bright point', i.e. a point that is brighter than its neighbors.

From a computational perspective, we have the following definition: a point is a bright (dark) elliptic feature at scale σ if the Hessian matrix of $I(x,y;\sigma)$ is negative (positive) definite at that point. The net result of applying dark elliptic feature detection is a binarized mask corresponding to foreground and background. However, very small regions may have been created as a result of inherent noise in the image, which are then removed based on a size threshold. The threshold is determined by the correct segmentation for a population of nuclear features. In rare cases, the mask corresponding to the dark elliptic features may be corrupted by foreign objects, which can be resolved and corrected after signal decomposition.

Nuclear regions provide the basis for region-based tessellation along curvilinear boundaries. Formally, let N_i correspond to the i-th $\in [0,K)$ nuclei in the image, $q \in N_i$, and p be a point in the image. Then region-based tessellation is defined by $V_i = \{p | dist(p,N_i) < dist(p,N_j), j \in \{0,1,\ldots,K-1\}$ and $j \neq i$ where $dist(p,N_i) = \min_{q \in N_i} |p-q|$. Computationally, this tessellation is computed through the application of the watershed method (Vincent and Soille, 1991) to the distance transform that is computed from binarized nuclear masks (Jan and Hsueh, 2000).

3.2 Signal decomposition

Signal decomposition, through color unmixing, aims to identify the signaling macromolecules that are associated with each dye. For example, Figure 2 shows how labels for nuclei and lipids are distributed in the RGB space. The main contribution of this article is in characterizing color unmixing as a segmentation problem that incorporates neighborhood information through a global optimization framework. The optimization framework is based on the graph-cut method (Boykov and Marie-Pierre, 2001), which is briefly summarized. In this context, the image is represented as a graph $G = \langle \bar{V}, \bar{E} \rangle$, where $\bar{V}$ is the set of all nodes, and $\bar{E}$ is the set of all arcs connecting adjacent nodes. Usually, nodes and edges correspond to pixels ($\mathcal{P}$) and their adjacency relationship, respectively. Additionally, there are special nodes that are known as terminals, which correspond to the set of labels that can be assigned to pixels. In the case of a graph with two terminals, terminals are referred to as the source (S) and the sink (T). The labeling problem is to assign an unique label x_p (0 for background, and 1 for foreground) for each node $p \in \bar{V}$, and the image cutout is performed by minimizing the Gibbs energy E (Geman and Geman, 1984):

$$E = \sum_{p \in \bar{V}} E_1(x_p) + \sum_{(p,q) \in \bar{E}} E_2(x_p, x_q) \tag{8}$$

where $E_1(x_p)$ is the likelihood energy, encoding the data fitness cost for assigning x_p to p, and $E_2(x_p,x_q)$ is the prior energy, denoting the cost when the labels of adjacent nodes, p and q, are x_p and x_q, respectively. The likelihood energy is computed in an eight-connected neighborhood. Figure 3 shows how a local neighborhood is partitioned through a two-terminal graph-cut segmentation.

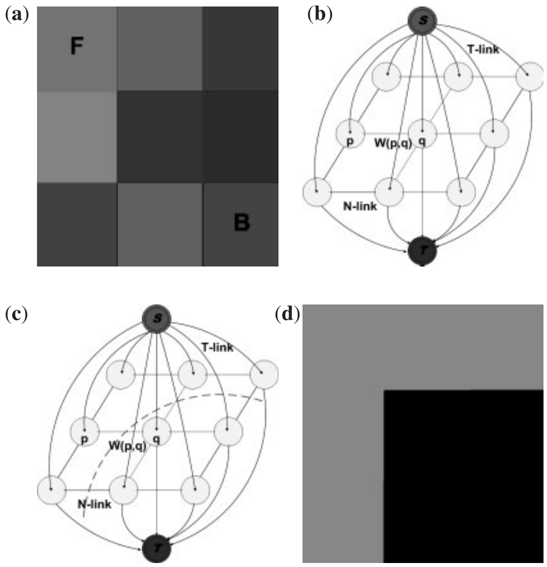

Fig. 3. An example of two-terminal (class) graph-cut segmentation: (**a**) an image grid (3 × 3), where 'F' and 'B' correspond to foreground and background seeds, respectively; (**b**) a graph constructed from image (a); (**c**) an optimum cut shown as a red line; and (**d**) a final labeling result where grid points are assigned to terminals S and T after the cut.

The optimization algorithms could be classified into two groups: Goldberg–Tarjan 'push-relabel' methods (Goldberg and Tarjan, 1988), and Ford–Fulkerson 'augmenting paths' (Ford and Fullkerson, 1962) . The details of the two methods can be found in Cook *et al.* (1998).

We initialize our system with multiple models of foreground and background, specified by the user stroke, from a subset of images. An underlying feature distribution, corresponding to a user stroke, is then represented with the Gaussian mixture model, in the color space, i.e. each foreground and background signature is not represented with a single Gaussian model. Let the conditional density for a pixel feature, C_p, belonging to a multi-colored object $\mathbb{O}$ be a mixture with M component densities:

$$GMM_{\mathbb{O}}(C_p) = \sum_{j=1}^{M} p(C_p|j)P(j) \tag{9}$$

where a mixing parameter $P(j)$ corresponds to the weight of component j and where $\sum_{j=1}^{M} P(j) = 1$. Each mixture component is a Gaussian with mean μ and covariance matrix Σ. In *RGB* color space:

$$p(C_p|j) = \frac{1}{(2\pi)^{\frac{3}{2}} |\Sigma|_j^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(C_p - \mu_j)^T \Sigma_j^{-1} (C_p - \mu_j)\right) \tag{10}$$

Then the data fitness term is defined as,

$$E_1(x_p) = \begin{cases} \frac{GMM_B(C_p)}{GMM_F(C_p) + GMM_B(C_p)}, & \text{if } x_p = Foreground \\ \frac{GMM_F(C_p)}{GMM_F(C_p) + GMM_B(C_p)}, & \text{if } x_p = Background \end{cases} \tag{11}$$

in which $GMM_F(C_p)$ and $GMM_B(C_p)$ are the probabilities of feature C_p (in *RGB* space) from the foreground and background models, respectively. For example,

$$GMM_F(C_p)=\sum_{j=1}^{M}p(C_p|j)P(j) \qquad (12)$$

where $M=10$ was manually selected in our implementation to capture wide variations in staining. $P(j)$ and (μ_j, Σ_j) for $p(C_p|j)$ are estimated by EM algorithm (Tomasi, 2004). And the smoothness term is defined as,

$$E_2(x_p,x_q)=w_e(p,q)\propto\exp(-|C_p-C_q|)\cdot\frac{1}{dist(p,q)} \qquad (13)$$

where $|C_p-C_q|$ is the Euclidean distance between feature vectors of C_p and C_q in RGB space, and $dist(p,q)$ is the Euclidean distance between p and q in the image grid. Next, we construct the graph G according to Table 1 and optimizing the objective function with the graph cut algorithm (Boykov and Marie-Pierre, 2001). After decomposition, nuclear segmentation is further validated by removing nuclear candidates that partially overlap with the signaling molecule. However, such an overlap is a rare event, and it is always due to sample contamination in a small area. In cases where macromolecules of interest are localized in nuclei, then the experiment needs to be designed properly, i.e. (i) assure that the dye has sufficient color separation with nuclear labeling hematoxylin, (ii) use another dye for labeling nuclei or (iii) label for cytoplasm as a reference. IHC is a visualization and scoring protocol for pathologists, and staining is always designed properly to assure color separation.

Table 1. Edge weights for the graph construction, where $\mathbb{N}$ is the neighborhood system, and μ is the weight for smoothness

Edge	Weight	For
$p\rightarrow S$	$\frac{GMM_F(p)}{GMM_F(p)+GMM_B(p)}$	$p\in\mathcal{P}$
$p\rightarrow T$	$\frac{GMM_B(p)}{GMM_F(p)+GMM_B(p)}$	$p\in\mathcal{P}$
$w_e(p,q)$	$\mu\cdot\exp(-\lvert Cp-Cq\rvert)\cdot\frac{1}{dist(p,q)}$	$\{p,q\}\in\mathbb{N}$

3.3 Validation with synthetic data

To evaluate the performance and error rate of the system, a synthetic image has been generated that simulates the nuclear size, shape and contrast associated with complexes. Nuclear size and shape are derived from the average behavior of nuclear features, following segmentation, from a subset of images. Contrast is derived from the user-based annotation of regions corresponding to positive and negative staining from the same subset of images. Gaussian noise is added, and the signal-to-noise ratio is changed from 12 dB to 0 dB, from left to right, respectively. Subsequently, segmentation error is quantified against the ground truth, and results are summarized in Figure 4 and Table 2. This error corresponds to mismatches (e.g. differences) between the known prior mask and the computed segmentation.

3.4 Comparison with NMF

Figure 5 shows the signal decomposition results of the image in Figure 2a, based on graph cut and NMF with identical initialization. We used an improved version of NMF, based on sparseness constraints. The matlab code is available online (Hoyer, 2004), and has been rewritten in C++ for comparative analysis. The sparseness

Table 2. Comparison of segmentation errors for synthetic images of Figure 4 based on graph cut and NMF, respectively

Method	Figure 4a	Figure 4b	Figure 4c	Figure 4d
Graph cut	0.5%	1.2%	3.5%	5.4%
NMF	0.8%	4.7%	31.2%	43.3%

Errors are measured against known ground truth.

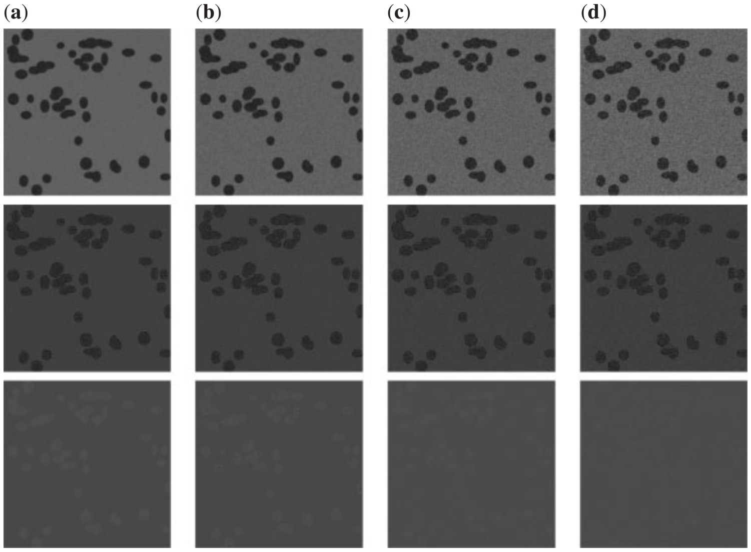

Fig. 4. Noise is added to a synthetic image at 12 dB (**a**), 7 dB (**b**), 3 dB (**c**) and 0 dB (**d**), and the segmentation results based on graph cut ($\mu=100$) and NMF are shown in the second and third rows, respectively.

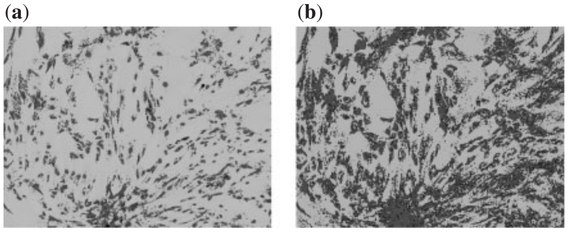

Fig. 5. A comparison of signal decomposition by graph cut ($\mu = 100$) (**a**) and NMF (**b**) indicates superior performance with the graph-cut method.

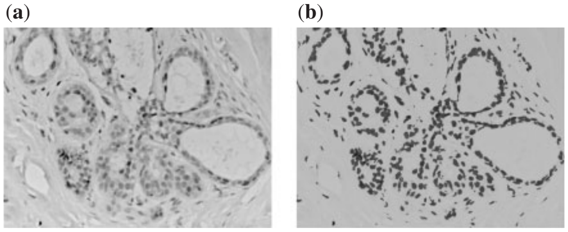

Fig. 6. Decomposition ($\mu = 100$) of color space when nuclear and antibody stains colocalize: (**a**) a bright field image of human mammary tissue stained for phosphorylated γH2AX and hematoxylin, and (**b**) its color decomposition.

constraint enables discovery of parts-based representation, and is shown to have a better quality than standard NMF. The results indicate that NMF decomposition is more noisy, mainly because it ignores neighborhood information. Other linear unmixing techniques may also produce noisy output, since regularization is often ignored. It is important to note that NMF is inherently a gradient descent method, and therefore, user-based initialization remains an integral component. Otherwise, the method converges to a wrong fixed point.

Finally, with respect to the computational complexity, the costs for graph cut and NMF, based on the same image (1280 × 960), are 10 s and 10 min, respectively.

4 SIGNAL DECOMPOSITION WITH COLOCALIZED SIGNALS IN THE COLOR SPACE

In the previous example, the signal from Oil Red O is registered outside of the nuclear regions, and therefore there is little overlap between stains in the color space. In order to examine the efficacy of our proposed method for colocalized staining, we imaged samples stained for phosphorylated $\gamma H2AX$, which localizes to the nucleus. In this experiment, the user provides examples of signals associated with each stain for establishing the prior knowledge. Subsequently, energy functions were computed according to Table 1, and optimal partitions were computed. An example, shown in Figure 6, indicates postive and negative staining, where the results are examined against manual scoring. We have used three representative images from a dataset for comparative analysis. These three images that were selected to represent a diverse signature of samples, were scored by a pathologist, and then compared with automated analysis. Because the process is tedious, a minimal number of images have been selected. Results are shown in Table 3, where the second and third columns correspond to the number of positive-and negative-stained nuclei that have been misclassified when compared to the ground-truth data. The last column lists the total number of cells in each image. The average error is <5%.

Table 3. Number of misclassified nuclei in each image for positive and negative stains, followed by the total number of cells in each image

	No. of positive nuclei	No. of negative nuclei	Total no. of cells
Image 1	10	8	715
Image 2	17	12	395
Image 3	4	7	381

5 BIOLOGICAL INVESTIGATION: A CASE STUDY

We have applied our method to a dataset of 192 images of fibroblasts obtained from women from two distinct patient populations. These samples are imaged on a Nikon Eclipse TE 2000 E, which is equipped with a color camera with a spatial resolution of 1280 × 960 pixels and a dynamic range of 8 bits per channel in RGB space. The illumination power is maintained at the same level, and all images are automatically corrected for shading and non-uniformities against a blank slide. Images are processed, and the amount of lipid is quantified for each cell in each image. The net result of color decomposition is a binarized mask, shown in red in Figure 8, corresponding to positive stains, where the intensity in the red channel is aggregated on a cell-by-cell basis. In addition to color decomposition, Figure 8 indicates nuclear position, in green, and how the space between nuclear regions are partitioned through region-based tessellation, in yellow. Tessellation allows the signal complex to be associated with the corresponding nuclear region. The binarized masks provide the context for aggregating intensity features in the red channel, and associating them to each nucleus. Finally, the results are represented as two probability density functions for each population, as shown in Figure 7. The KS test between these two distributions computes a P-value of <0.0001, thus indicating that the two populations are different.

One of the concerns regarding population study is due to the impact of error in nuclear segmentation for aggregating lipid signals. Our experience indicates that the error in nuclear segmentation is less than a few percent. Therefore, given a very large population of cells, such an error will be buried as noise. It will only be an issue if the central question is to search for outliers in a study.

6 CONCLUSION AND FUTURE WORK

We have developed a novel method for signal decomposition in the color space for scoring the amount of staining associated with macromolecules. The method is based on graph cut, which has a superior performance to NMF. Subsequently, strength of staining is associated to each cell through nuclear segmentation and region-based tessellation. In addition, we have demonstrated the efficacy of our method in samples when the stains are co-localized in

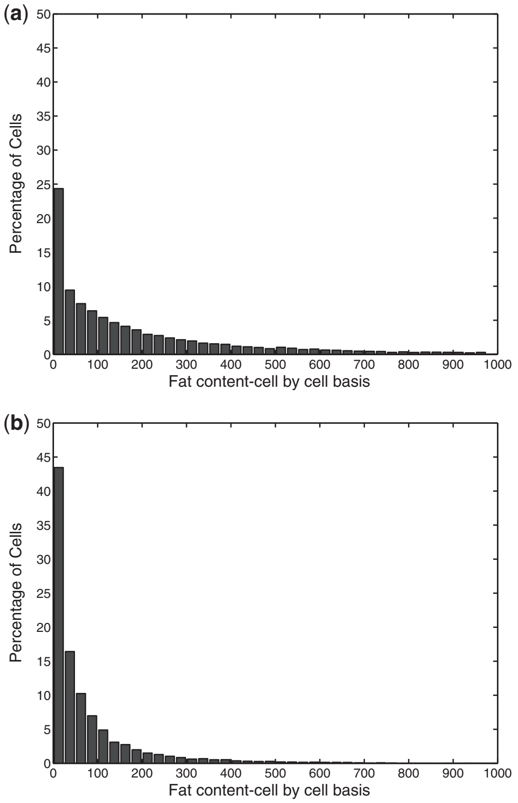

Fig. 7. Probability density functions corresponding to the fat content on a cell-by-cell basis for each of the two populations, where (**a**) corresponds to a population represented by Figure 8c and (**b**) corresponds to a population represented by Figure 8a. The KS test computes a P-value of 0.001 indicating that these two populations are different.

the same subcellular regions. Our method has been applied to a dataset of fibroblasts derived from the two patient populations that are different when placed under adipocyte differentiation conditions. Our continued research will focus on other end points for these primary cells in order to visualize and quantify a more comprehensive representation of active macromolecules.

ACKNOWLEDGEMENTS

Authors thank Dr Colleen Fordyce and Dr Kurt Thorn of UCSF for cell-culture support and imaging expertise.

Funding: National Cancer Institute (1P01CA107584-01A1); US Department of Energy, Office of Biological and Environmental Research (DE-AC03 SF0098).

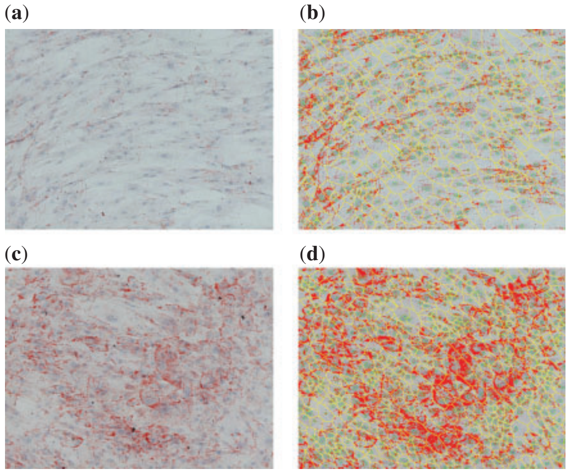

Fig. 8. Segmentation and color decomposition from two images in the dataset indicate how region-based tessellation enables quantifying signal macromolecules on a cell-by-cell basis. ($\mu = 100$, $\sigma = 2.5$)

Conflict of Interest: none declared.

REFERENCES

Boykov,Y. and Marie-Pierre,J. (2001) Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proceedings of IEEE ICCV*. pp. 105–112.

Cook,W.J. *et al*. (1998) *Combinatorial Optimization*. John Wiley & Sons.

Ford,L. and Fullkerson,D. (1962) *Flows in Networks*. Princeton University Press.

Gao,Y. and Church,G. (2005) Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, **21**, 3970–3975.

Geman,S. and Geman,D. (1984) Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE Trans. PAMI*, **6**, 721–741.

Goldberg,A.V. and Tarjan,R.E. (1988) A new approach to maximum-flow problem. *J. Assoc. Comput. Mach.*, **35**, 921–940.

Hoyer,P. (2004) Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, **5**, 1457–1469.

Jan,S. and Hsueh,Y. (2000) Primitive spatial relations based on SKIZ. *Images Vis. comput.*, 597–605.

Lee,D. and Seung,H. (1999) Learning the parts of object by non-negative matrix factorization. *Nature*, **401**, 788–791.

Papadakis,A. *et al*. (2003) A novel spectral microscope system: application in quantitative pathology. *IEEE Trans. Biomed. Eng.*, **50**, 207–217.

Rabinovich,A. *et al*. (2003) Unsupervised color decomposition of histologically stained tissue samples. *Arch. Pathol. Lab. Med.*

Tomasi,C. (2004) Estimating Gaussian mixture densities with EM - a tutorial. Available at *www.cs.duke.edu/courses/spring04/cps196.1/handouts/EM/tomasiEM.pdf*.

Vincent,L. and Soille,P. (1991) Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.*, **13**, 583–598.

Yang,Q. and Parvin,B. (2003) Harmonic cut and regularized centroid transform for localization of subcellular structures. *IEEE Trans. Biomed. Eng.*, **50**, 469–476.