Current methods for histological and tissue-based biomarker analysis

Hematoxylin-Eosin (H&E) staining of thin (5–7 micron) tissue sections has been used by pathologists for well over one hundred years and is widely accepted as the foundation of disease classification. Hematoxylin stains cell nuclei blue, while eosin, as a counter-stain, stains cytoplasm and connective tissue pink. Due to its long history, as well as low cost, fast preparation, and easy image acquisition, there is a strong belief that H&E will continue to be the common practice for the foreseeable future [1, 2].

Pathologists typically make diagnostic decisions from H&E stained tissue sections based on the attributes of cell size, shape, texture and color contrast of various fine features as viewed under a microscope [3]. Although a pathologist is well trained to decipher fine differences in tissue features, the analysis is inherently subjective, and use of objective quantitative analysis is limited in current clinical practice. With the advancement of digital microscopy, high quality microscopic images of specimens are becoming digitally available in large quantities [4]. Digital technology will likely lend itself to quantitative objective analysis of H&E stained tissue sections [5, 6]. However, the complex interpretations that are learned through years of training and that rely on
the complexities of the human mind will be difficult to fully recapitulate, even with current and future computational power.

One area in clinical pathology practice that lends itself more readily to quantitative image analysis is immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH), collectively known as “molecular pathology”. The advent of IHC in the 1970s enabled the visualization of specific protein biomarkers using antibodies tagged with chromogenic substrates. This technique co-exists with and augments H&E, and is becoming increasingly favored for prognostic purposes. Routine clinical tests include detection of estrogen and progesterone receptors (ER, PR) by immunohistochemistry (IHC) and determination of the ERBB2 (HER-2) receptor level based on IHC or gene copy number by fluorescent in situ hybridization (FISH). These tests are critically important for determination of appropriate therapy. ER positive [ER(+)] patients are offered anti-estrogen therapy and patients with HER-2 overexpression or amplification are offered Herceptin [7]. The only systemic alternative currently available for ER-negative [ER(-)] patients is chemotherapy.

Like H&E, IHC succumbs to the limitation of subjective quantitation and is typically analyzed in a semi-quantitative manner by visual inspection. This limitation is inherent to chromogenic staining methods [3]. For this reason, image analysis methods that can objectively assess intensity of chromogenic substrates are gaining traction in clinical practice and several standard IHC tests have recently become FDA approved [8].

More recently, immunofluorescent methods that lend themselves to more sensitive and linearly quantitative techniques are beginning to emerge [9]. The clinical value of automated fluorescence-based image analysis of protein biomarkers has been demonstrated in breast cancer [9] and lung cancer [10]. Although such methods are not yet approved for routine clinical use, it is likely that they will lead to new diagnostic approaches in clinical pathology.

**Multiplexing**

Gene and protein arrays are commonly used for measuring multiple targets (multiplexing) at the molecular level. However, gene expression may not represent actual protein expression, nor does it provide information on the cellular localization within the context of the tissue specimen. Multiplexing directly on tissue or cells, without
the need for extraction and dispersion on chips, preserves the spatial integrity of proteins while still allowing to assess multiple interactions between those proteins. This provides an entirely new way of examining biomarkers and could shed light on previously un-examined relationships between spatial location and protein-protein interactions.

For example, prediction of disease outcome or therapeutic response will likely require the analysis of multiple components from several biological pathways. In oncology, many cancer types are very heterogeneous in phenotype, and require complex screening. Breast cancer is a typical example whereby a single patient may be screened for expression of different hormone receptors (estrogen and progesterone), keratin profile to determine tumor subtype (luminal or basal), and over-expression of the oncogene Her2 for which an antibody therapy exists as an adjuvant therapy [11]. Through multiplexing analysis, it will become increasingly possible to provide customized medicine.

**Fluorescence microscopy**

An important advantage of immunofluorescent techniques for tissue-based biomarker interrogation is the ability to multiplex proteins in a single tissue section. Fluorescence lends itself to multiplexing for two main reasons, 1) there is a wide range of fluorophore dyes and/or nanocrystals with non-overlapping emission spectra; 2) sequential staining and imaging is possible with methods such as photobleaching. Detection agents can take the form of small organic molecules (such as Cye and Alexa dyes, FITC, Rhodamine, etc), nanocrystals (“quantum dots”) [12], proteins with inherent fluorescent properties [13], and genetic tags that have been coupled with fluorochromes [14]. These molecules are frequently functionalized to couple efficiently with other biological agents such as antibodies or nucleic acids. The main forms of multiplexing include 3-4 “channels” through standard fluorescence microscopy, 5-6 channel analysis with quantum dots, “spectral bar-coding” in which a finite number of dyes are mixed in combination and precise spectral characteristics are determined on a per-pixel basis, and repeated use of a few fluorescent channels through regulating the dyes fluorescent properties such as photo-bleaching or antibody denaturation/stripping. A common requirement of all these methods is the ability to digitally reconstruct the patterns seen
for the different biomarkers. This is accomplished through either digital “layering” in which multiple images are overlapped on top of one another [15, 16], or spectral unmixing in which numerous agents each with unique spectral properties are detected simultaneously [17].

**Fluorescent dyes**

There are numerous fluorescent dyes available for use as labeling agents to tag detection agents in biological samples [14]. These dyes each have unique spectral properties, in particular differing wavelengths of light for excitation and the resulting fluorescent emission. Detection is accomplished by the use of band pass and dichroic filters allowing specific wavelengths of light to pass such that for a given filter combination a single fluorochrome is seen [18]. It is typical to examine blue, green, orange, and red fluorochromes in this manner whereby the spectrum of each dye is distinct that there is no cross-talk between the different fluorochromes. Using combinations of dyes together allows a higher level of multiplexing to be obtained through the generation of unique spectral signatures. These methods have been utilized in human genetic analysis in a process known as spectral karyotyping to detect each individual human chromosome and to determine rearrangements or fragmentation in cytogenetic analysis [19].

One potential drawback to fluorescent dyes is the impermanence of signal and loss of activity with time. A second potential limitation, which has been overcome to a large degree with advances in CCD camera technology, is the limits of detection of low signal abundance. A third challenge is the auto-fluorescence generated by the tissues. The longer the exposure required to detect the signal, the more auto-fluorescence is also accumulated in the resultant image. Thus the amount of light generated specifically by the fluorochromes must be a certain degree above the noise in the system, which largely stems from tissue auto-fluorescence. Several hardware applications such as laser scanning confocal microscopy and the use of structured light with multiple image acquisition (such as the Zeiss Apotome) allow a mechanical mechanism for removing the auto-fluorescence [20]. Through the use of spectral acquisition and deconvolution, the analysis of individual pixels based on wavelengths of
light passing through the detector (such as liquid crystal tunable filters) can be used in numerous multiplexing platforms for signal detection [17].

**Quantum Dots**

Quantum dots have been used for multispectral analysis of biological specimens [21, 22]. These are nanostructures containing a photo-activatable metal core (such as CdSe/ZnS) surrounded by a shell. Fluorescence can be generated from these molecules by excitation in the near UV range which produces a fluorescence with very distinct peak spectral output spanning as low as 10nm allowing for spectral separation between the different nanostructures. Through coupling of the nanospheres to avidin and binding to biotinylated antibodies, five distinct molecules were visualized along with nuclear staining with 4′6-DiAmidino-2-PhenylIndole (DAPI) [23]. Quantum dots have also been used for fluorescent in situ hybridization (FISH) analysis for multiplexed analysis of cellular mRNAs and cytogenetic analysis of human chromosomes [24-27] The main limitation to the level of multiplexing achievable with quantum dots are the size of the nanoparticles and full application of the use of these particles for molecular analysis will be accomplished through ongoing development.

**Photobleaching**

Fluorescent dyes are inherently unstable to photo bleaching and can be inactivated by exposure to proper wavelengths to photo-excite the fluorescent dyes rendering them incapable of further fluorescence. However, for some dyes, their susceptibility to photobleaching has enabled multiplexing. Using this approach, a dye-labeled antibody can be localized to the sample of interest, imaged to record the localization pattern, photo-bleached to clear the localization pattern, then re-incubated with a second dye-labeled antibody and the process is sequentially repeated. Methods for measuring multiple protein networks in tissue (the “toponome”) have relied on the use of photo bleaching to localize over 100 antigens in a single sample [28]. This method has a limitation in that the fluorescent dyes must be directly coupled to the antibodies to achieve high level multiplexing. For complex tissue samples and low abundance proteins, signal amplification may be needed to generated sufficient signal.
Image Analysis

Tissue micro-arrays (TMAs) provide the means for simultaneous analysis of tissue from large numbers of patients. Tissue microarrays contain hundreds of tissue spots (approximately 0.6 mm or greater in diameter) originating from cores of tissue from regions of interest in paraffin embedded tumor samples. The tissue cores are transferred to a recipient paraffin block, in a precisely spaced array pattern. Each block can generate up from 100-500 sections analyzed using independent IHC tests and can serve as a valuable discovery tool for tissue based biomarkers [29, 30]. However, the large amount of information in TMAs and the high dimensional nature of the data makes automated image analysis algorithms essential for high throughput segmentation of TMA images into sub-cellular compartments (i.e. cytoplasm, nucleus, membrane) and quantitation of biomarkers in these compartments [9]. In the following sections we describe algorithms for registering and segmenting multi-channel microscopic images of breast TMAs into sub-cellular compartments for automated quantitation of multiple target proteins. The multi-channel images of different biomarkers are obtained by a multi-step image acquisition procedure. These images are corrected for non-uniform illumination. A robust, multi-resolution image registration algorithm is applied to transform images into the same coordinate system. A multi-channel segmentation method is then applied to segment the registered images into sub-cellular compartments. The algorithms are general and able to handle with multiple images with arbitrary number of channels that are acquired in multiple steps. Finally, multiple target proteins are superimposed on individual compartments to calculate metrics associated with subcellular protein expressions and translocation. The overall workflow is show in Figure 1. In the following sections we describe the details of each of the preprocessing, registration, segmentation, and quantitation steps.

Image preprocessing

In the preprocessing step, raw images are smoothed by a Gaussian filter to remove noise. Non-uniform illumination is also corrected. The illumination pattern can be estimated from the images, or directly computed by using calibration targets. Most filter-cube and microscope manufacturers carry fluorescent plastic that can be used for
calibration. If the calibration images are not taken during the acquisition the illumination pattern can be estimated from a series of images. The observed image, \( I(x, y) \), can be modeled as a product of the excitation pattern, \( E(x, y) \), and the emission pattern, \( M(x, y) \). While the emission pattern captures the tissue dependent staining, the excitation pattern captures the illumination.

\[
I(x, y) = E(x, y)M(x, y)
\]  
(1)

In the logarithm domain, the equation above can be transformed to a linear form:

\[
\log(I(x, y)) = \log(E(x, y)) + \log(M(x, y))
\]  
(2)

From a set of \( N \) images, let \( I_n(x, y) \) denote ordered set of pixels. In other words, the pixels are sorted for any given \((x, y)\) location such that

\[
I_1(x, y) \leq I_2(x, y) \leq \cdots I_n(x, y) \cdots \leq I_N(x, y).
\]  
(3)

Assuming that a certain percentage \((p)\) of the image is formed from stained tissue (non-zero background), then a trimmed average of the brightest pixels can be used to estimate the excitation pattern:

\[
E_{AVE}^I(x, y) = \frac{1}{N-K+1} \sum_{n=K}^{N} \log(I_n(x, y)),
\]  
(4)

where \( K \) is set to an integer closest to \( N(1-p) + 1 \). In our experiments, we set \( p \) to 0.1 (10%). In the above equation, the average emission pattern of the tissue is assumed to be uniform. Since the images are recovered up to a scale factor, we can drop the constant term introduced by the uniform emission pattern. This approximation holds if a large number of images are used in the averaging process. However, a large percentage of pixels (90%) are already excluded to eliminate the non-tissue pixels in the images. To overcome the limited sampling size, we approximate the log of the excitation pattern with polynomials;
The parameters $a_i$ are solved by minimizing the mean squared error \[31\]. The surface generated by the polynomial coefficients are then used to correct individual images. A sample average excitation pattern and the estimated polynomial illumination surface are shown in Figures 2a and 2b, respectively. Figures 3a and 3b show a fluorescent image before correction and after correction, respectively. Use of polynomial surface to estimate the illumination pattern is not limited to fluorescent microscopy; brightfield image can be corrected similarly. If each color channel is corrected separately, this corrects for the color temperature of the light source as well (See Figures 3c and 3d).

**Image Registration**

Image registration techniques, such as mutual information or correlation-based techniques can be used to register images from each step of a sequentially multiplexed study and align them accurately. The experiments are designed such that at each step of the sequential staining, images of the nuclei are acquired. Figure 4 shows an example sequence of images where the first image is the image of the nuclei. Similarly, the nuclei can be stained by Hematoxylin in the H&E stain. The first nuclei image is set as the reference image and each of the subsequent nuclei images are registered to the reference. Once the transformation parameters are estimated, then all the channels can be mapped onto the reference coordinate system.

Given two set of nuclei images, one being the reference image from the first step, $I_N^{(1)}(x,y)$, and the second being either from the subsequent fluorescent acquisitions, or the nuclei channel from the final H&E step, $I_N^{(k)}(x,y)$, we find a transformation $T^{k:1}$, such that the image similarity measure between $I_N^{(1)}(x,y)$ and $I_N^{(k)}(T^{k:1}(x,y))$ is maximized. Different combinations of transformations and image similarity measures have different performance computation requirements, and are suitable for different applications. In this work, we use a rigid transformation and a mutual information based image similarity measure \[32-42\]:

\[
E^i_{AVE}(x,y) = \sum_{0 \leq i, j \leq p} a_i x^i y^j.
\]
\[ S(t) = -\sum_i \sum_\kappa p(t, \kappa | t) \log \frac{p(t, \kappa | t)}{p_M(t | t)p_F(\kappa)} \]  \hspace{1cm} (6)

where \( p, p_M, \) and \( p_F \) are the joint, marginal moving, and marginal fixed probability distribution of the image intensities; \( t \) is the parameter vector of the transform; \( \iota \) and \( \kappa \) are the intensity values in the respective images. In order to improve the robustness of the algorithm, we use a multi-resolution strategy to find the transform that aligns the two images. Figure 5 illustrates the registration of a breast cancer tissue image using the DAPI channel and subsequent transformation of the CFP channel image using the registration results of the DAPI channel.

While for fluorescent acquisitions, the nuclei images are part of the experiment (DAPI acquisition in our case), computing the nuclei image from color images of the H&E needs an additional step. The red, green and blue components of the H&E image are used to generate nuclei image using the following equation;

\[ I_N^{HE}(x, y) = c \cdot \left( \frac{I_{BLUE}^{HE}(x, y)}{\sqrt{I_{RED}^{HE}(x, y) \cdot I_{GREEN}^{HE}(x, y)}} \right)^\gamma \]  \hspace{1cm} (7)

where \( c \) and \( \gamma \) are tuning parameters for contrast and gamma correction, respectively. Since the hematoxylin stains the nuclei to blue, the highest contrast is achieved when the blue channel is normalized by the geometric mean of the red and green channels. Figures 6a and 6b show an H&E image and its estimated nuclei component computed by the above equation, respectively. A two-channel fluorescent imaging of the same tissue stained with molecular biomarkers with green representing a membrane related marker (beta-catenin) and blue representing a nuclei related stain (DAPI) are shown in Figures 6c and 6d, respectively. The registered DAPI image in the H&E coordinate system is shown in Figure 6e. The registration parameters estimated from the DAPI and H&E images are then used to map the beta-catenin channel into the H&E coordinate system (shown in green color in Figure 6f).
**Image Segmentation**

Subcellular quantification of target proteins requires segmentation of the subcellular compartments, such as nuclei, membrane, and cytoplasm. The segmentation can be achieved either by segmenting each channel separately or segmenting all channels at once. Segmentation of ridge-like and blob-like structures is one of the most common segmentation tasks in medical and life sciences imaging applications. Commonly, such applications require detecting vessels [43], bronchial tree [44], bones [45, 46], nodules [47-49] in medical applications, and detecting neurons [42, 50], nuclei [51], and membrane [52] structures in microscopy applications. Partitioning of a multiple channel digital image into multiple segments (regions/compartments) is one of the most critical steps for quantifying one or more biomarkers in molecular cell biology, molecular pathology, and pharmaceutical research.

Subcellular segmentation can be achieved in most cases using a single channel. For many of the approaches in the literature, pre-processing of the images is required to reduce noise or to extract features from the images. Common pre-processing steps include thresholding, edge detection, and smoothing. A number of cellular-level segmentation tasks have been accomplished by combining only these “pre-processing” steps. In [53], median filtering and morphological operations of erosion and dilation were adapted to segment melanomas and lymphocytes. Binary thresholding was similarly combined with morphological operations and image thinning to segment cells in histology images in [54]. However, more intelligent pre-processing is needed for non-trivial applications. In [55, 56], cellular images of breast tissue were denoised using a non-linear diffusion filter followed by a directional coherence filter to enhance the boundaries of the nucleus. A non-linear illumination correction method was described earlier.

Region-based segmentation, in which the image is divided into regions based on some homogeneity criteria, is commonly used. Typically, segmentation starts from a seed pixel or region (e.g., center of the nucleus) determined automatically or specified by some user interaction, and then grows to include neighboring pixels that meet a specified intensity, texture, or shape criteria. In the connected-threshold region-growing
algorithm [57] for example, all neighboring pixels between an upper and lower threshold values are included recursively. This is after smoothing with an edge-preserving smoothing filter such as anisotropic diffusion [52], curvature flow [53], and bilateral [53] filters. In fluorescence cellular microscopy image, the seed points may be obtained by initial thresholding, using a top-hat filter [58], or by computing distance transforms, to obtain pixels in the cells, which are then grown to the cellular boundaries. In the confidence-connected region-growing approach [34, 59-62], the mean and standard deviation of the seed region, and all pixels around the region that fall within a range of the mean are recursively accepted as part of the region. When there are no more pixels to include, the mean and standard deviation of the newly obtained region is computed and the process iterated a few times. The isolated-connected algorithm [34, 35] in which two seed points are given from two different regions, e.g., nucleus and cytoplasm, can be used to segment the whole cell. The goal of the algorithm is to grow a region that is connected to the first region but not connected to the second. A binary search is used to find the optimal separating intensity value. In [63], region growing was employed after binary thresholding to segment the nuclei in FISH images. In Figure 7b, the result of membrane segmentation is overlaid on the tissue image. A membrane-stained channel image was smoothed with curvature flow filter to remove the noise while preserving the membrane markings. A non-maximal suppression algorithm was then employed to remove spurious signals and retain only those that form a continuous boundary.

**A Unified Segmentation Algorithm**

While different segmentation algorithms can be used for each of the nuclei and membrane compartments, an alternative is to use the same algorithm in different modes. For example the curvature metric derived from nuclei and membrane images can be used as metrics to classify segments using supervised or unsupervised (parametric or non-parametric) algorithms [64]. A commonly used approach to compute curvature-based metrics is to extract them from the eigenvalues of the Hessian matrix. Due to their invariance to rigid transformations, these metrics can be used for a broad class of ridge-like and blob-like structures [65]. The Hessian of an image \( I(x, y) \) is defined as
\[ H(I(x, y)) = \begin{bmatrix} \frac{\partial^2 I(x, y)}{\partial x^2} & \frac{\partial^2 I(x, y)}{\partial x \partial y} \\ \frac{\partial^2 I(x, y)}{\partial y \partial x} & \frac{\partial^2 I(x, y)}{\partial y^2} \end{bmatrix}. \] (8)

The eigenvalues \( (\lambda_1(x, y) \leq \lambda_2(x, y)) \) of the Hessian matrix can either be numerically calculated or analytically written in terms of the elements of the Hessian Matrix:

\[ \lambda_{12}(x, y) = \frac{1}{2} \left\{ \frac{\partial^2 I(x, y)}{\partial x^2} + \frac{\partial^2 I(x, y)}{\partial y^2} \pm \sqrt{\left( \frac{\partial^2 I(x, y)}{\partial x^2} - \frac{\partial^2 I(x, y)}{\partial y^2} \right)^2 + 4 \left( \frac{\partial^2 I(x, y)}{\partial x \partial y} \right)^2} \right\}. \] (9)

The eigenvalues encode the curvature information of the image, and provide useful cues for detecting ridge-like membrane structures, or blob-like nuclei structures. However, the eigenvalues are dependent on image brightness. We define the following two curvature-based features that are independent of image brightness:

\[
\theta(x, y) = \tan^{-1} \left( \frac{\lambda_1(x, y)}{\lambda_2(x, y)} \right),
\]

\[
\phi(x, y) = \tan^{-1} \left( \frac{\lambda_1(x, y)^2 + \lambda_2(x, y)^2}{I(x, y)} \right)^{1/2},
\]

and refer to them as shape index, and normalized-curvature index respectively. This is essentially the same as defining the eigenvalues in a polar coordinate system. This transformation also results in bounded features, \( -\frac{3\pi}{4} \leq \theta(x, y) \leq \frac{\pi}{4} \), and \( 0 \leq \phi(x, y) \leq \pi / 2 \).

A general likelihood function estimator that calculates the probability maps of vessel, membrane and nuclei-like structures in images can be formulated by exploring the expected values of the features. For example for bright membrane and vessel like structures, the shape index is close to \( -\pi / 2 \), whereas for blob-like nuclei structures, the shape index is close to \( -3\pi / 4 \). These constraints are used to compute the initial
foreground set for membrane and nuclei structures. Also for bright structures, it is less likely for objects to have non-negative shape index values compared to noise where non-negative values can easily occur. An initial segmentation based on the shape index and the normalized-curvature index separates the image pixels into three subsets: background, foreground, and indeterminate. Indeterminate subset comprises all the pixels that are not included in the background or foreground subsets. From these subsets, the background and foreground intensity distributions, as well as the intensity log-likelihood functions are estimated. The algorithm keeps iterating by using two out of the three features at a time to estimate the distribution of the feature that is left out. Usually three iterations are sufficient for a convergence. In the final step these log-likelihood functions are combined to determine the overall likelihood function. A probability map that represents the probability of a pixel being a foreground is calculated.

This non-parametric method is different from existing parametric approaches, because it can handle arbitrary mixtures of blob and ridge like structures. This is essential in applications such as in tissue imaging where a nuclei image in an epithelial tissue comprises both ridge- and blob-like structures. The network of membrane structures in tissue images is another example where the intersection of ridges can form structures that are partially blobs. Accurate segmentation of membrane and nuclei structures forms the base for higher level scoring and statistical analysis applications. For example, distribution of a target protein on each of the segmented compartments can be quantified and related to clinical outcomes. Scores measuring translocation of a protein between segmented compartments can reveal protein specific pathways, and response to drug therapy. Using the spatial and intensity interrelation of nuclei and membrane biomarkers, it is possible to separate the epithelial nuclei from the stromal nuclei. Figure 8 shows sample images of membrane and nuclei and their segmentation results.

**Segmentation of Cytoplasm and Epithelial Regions**
Cytoplasm can be detected either by using a specific marker, or detecting it computationally. Based on binary image morphological operations, Ding [63] describes a method that identifies the cytoplasm as the region around the nuclei. While this works
well for sparse cell images, defining these regions in tissue is more complicated. Whenever we have membrane compartment in addition to nuclei we can determine the cytoplasm as the region between membrane and nuclei.

Let us denote the thresholded nuclei, and membrane sets with \( M(x,y) \), and \( N(x,y) \), respectively. Cytoplasm, denoted by \( C(x,y) \), is defined as the union of sets of small regions circumscribed by membrane alone or membrane and nuclei pixels (See green regions in Figures 8b and 8d). Only pixels that are not defined as \( M(x,y) \) or \( N(x,y) \) can be defined as \( C(x,y) \).

Epithelial and stromal regions are morphologically different. Let \( U(x,y) \) defined as

\[
U(x,y) = C(x,y) \cup M(x,y) \cup N(x,y)
\]

(12)
denote the superset union of the nuclei, cytoplasm, and membrane sets. Since the stromal nuclei are not connected through membrane structures, and are sparsely distributed, they can be detected by a connected component analysis of \( U(x,y) \). An epithelial mask, \( E(x,y) \), is generated as a union of large connected components of \( U(x,y) \). For the sample images, any connected component larger than 800 pixels is accepted as a part of the epithelial mask. The nuclei set is then separated into epithelial nuclei \( N_e(x,y) \) and stromal nuclei \( N_s(x,y) \) by masking,

\[
N_e(x,y) = N(x,y) \cdot E(x,y), \quad \text{and} \quad N_s(x,y) = N(x,y) \cdot (1 - E(x,y)).
\]

(13), (14)

Figures 8b and 8d show the separated epithelial nuclei (blue) from the stromal nuclei (grey).

**Multi-channel Segmentation Techniques**

After images are registered, points with the same coordinates in different channels correspond to the same structural location. Each pixel in the image has a vector of features, i.e., intensity values of images from different channels. These features are then used to segment the vector valued image into different sub-cellular compartments.
Many segmentation algorithms have been developed for digital microscopic images. In this work, we use one of the most widely used segmentation algorithms, \( k \)-means clustering, for its simplicity and robustness. \( k \)-means clustering algorithm divides feature space into clusters and maximizes the distances between the cluster centers. An iterative procedure is used to find these cluster centers. Each pixel is given the same label as the cluster center that is closest to it in the feature space.

Suppose \( \mathbf{I}(x,y) = [I_1(x,y), \ldots, I_c(x,y)] \) \((x,y) \in \Omega\) is a \( c \)-channel image of one tissue sample. The \( k \)-means segmentation algorithm proceed as follows:

**Step 0.** Initialize cluster centers \( \mu_1, \ldots, \mu_k \);

**Step 1.** Assign a label \( l(x,y) \) to each pixel based on its distance in the feature space to the cluster centers:

\[
l(x,y) = \arg \min_j \| \mathbf{I}(x,y) - \mu_j \|
\]

**Step 2.** Update cluster centers as:

\[
\mu_j = \frac{1}{N_j} \sum_{i \in \mathcal{J} \neq j} \mathbf{I}(x,y)
\]

where the summation is over all image pixels with label \( j \), and \( N_j \) is the total number of pixels with label \( j \).

**Step 3.** Go to step 1 if not convergent.

Convergence is achieved when the changes in clusters centers are smaller than a preset threshold. \( k \)-means segmentation result of the multiplexed images shown in Figure 4 is shown in Figure 7a.

Two rounds of images were acquired for a TMA with 60 samples including both normal breast tissues and breast cancer tissues. Each round consisted of multiple biomarkers. Figure 4 shows four images of these markers for one tissue sample. \( k \)-means segmentation algorithm is applied to segment multi-channel images into five compartments Figure 7a shows one typical segmentation result): nucleus (blue),
membrane (red), smooth muscle (violet), cytoplasm (green), and background (black). Figure 7b shows the overlay of the boundaries of the membrane class on one membrane-stained channel.

Quantitation of Subcellular Biomarkers
A number of steps are required for quantitation of subcellular biomarkers. The membrane compartment of tumor cells can be identified by pan-cadherin staining and 4',6-Diamidino-2-phenylindole (DAPI) is used to identify nuclei. A nonlinear non-parametric mapping function with multiple inputs, including the local geometry of the pixel distributions as well as the intensity values, maps each of these two channels to probability values indicating the likelihood of pan-cadherin pixels being membrane, and likelihood of DAPI pixels being nuclei. To define cytoplasm, a definite decision for each pixel can be determined by thresholding the probability maps at 50% rate. For example, a 50% probability of a pixel on a DAPI image implies that the pixel is equally likely to be background and nucleus.

The distribution of biomarkers in each of these regions can be represented by a probability distribution function (PDF). For example the PDF of the biomarker on the membrane corresponds to its weighted empirical distribution, where the membrane probability map determines the weights. We denote the mean and the standard deviation of the biomarker distribution on each of the regions as $\mu_R$, and $\sigma_R$, respectively, where $R$ can be any of the nuclei, membrane, cytoplasm or extra cellular matrix (ECM) regions. ECM is defined as all the non-background pixels not classified as nuclei, membrane or cytoplasm. Compartmental biomarker distribution can then be related to disease outcome or response to therapy.

Summary
Quantitation of multiple tissue based biomarkers requires several sequential steps including tissue staining with target specific antibodies labeled with fluorescent reagents, image capture, preprocessing, registration, segmentation and subcellular quantitation. This provides the flexiblity to quantify biomarkers in more than one cellular compartment, thus maximizing the amount of data present in a tissue image, and
enabling more extensive analysis of the role of biomarkers in predicting response to therapy and patient survival.

**Acknowledgment**

The authors would like to thank the significant contributions of Maximilian Seel, Harvey Cline, Melinda Larsen to this work.
Figure 1: Schematic representation of image processing procedures. In this paper, we focus on the steps in the shaded box.
Figure 2: (a) Trimmed average image of one channel; (b) A third order polynomial approximation of the illumination pattern.
Figure 3: Fluorescent image showing DAPI staining; (a) before correction, (b) after illumination correction. A bright field DAB staining image before illumination correction, (c) before correction, (d) after illumination correction.
Figure 4: Four channels of one tissue sample. The top row shows images from the first imaging round with a nucleus-stain (left) and a membrane-stain (right). The bottom row shows images from the second imaging round with cyan fluorescent protein (CFP) stain (left) and smooth muscle actin stain (right).
Figure 5: Overlay of DAPI images before registration (a) and after registration (b) and overlay of CFP images before transformation (c) and after transformation (d). The color channels of the images are from subsequent steps of the sequential staining. The reference image is in red component, and the subsequent step is in blue channel, and the green is set to average of the two images. Any color shifts are due to misregistration.
Figure 6: (a) A three-channel (red, green, blue) color image of an H&E stained breast tissue section. (b) The nuclei component of the H&E color image. (c) A two-channel fluorescent imaging of the same tissue stained with molecular biomarkers. Green and blue colors are used to visualize the beta-catenin, and DAPI images, respectively. (d) DAPI channel of the molecular markers. (e) Registered DAPI image in the H&E coordinate system. (f) The beta-catenin image (shown in green color) superimposed with the H&E image.
Figure 7: a) K-means segmentation result of the multiplexed images shown in Figure 4. b) Curvature flow smoothing based membrane detection from Figure 4 overlaid on the membrane stained image.
Figure 8: (a) Raw image intensities (Breast cancer tissue), Red-Membrane, Blue-Nuclei, Green-Estrogen Receptor biomarkers. (b) Detected Compartments; Red-Membrane, Blue-Epithelial Nuclei, Green-Cytoplasm. (c) Raw image Intensities (Colon Tissue), Red-Membrane, Blue-Nuclei, Green-cMet markers. (d) Detected Compartments; Red-Membrane, Blue-Epithelial Nuclei, Gray-Stromal Nuclei, Green-Cytoplasm. Pink regions are excluded from quantitation. Both the background and the Extra Cellular Matrix (ECM) are presented with black color.
References


60. Melonakos, J., et al., *A Probabilistic Model for Haustral Curvatures with Applications to Colon CAD*.


