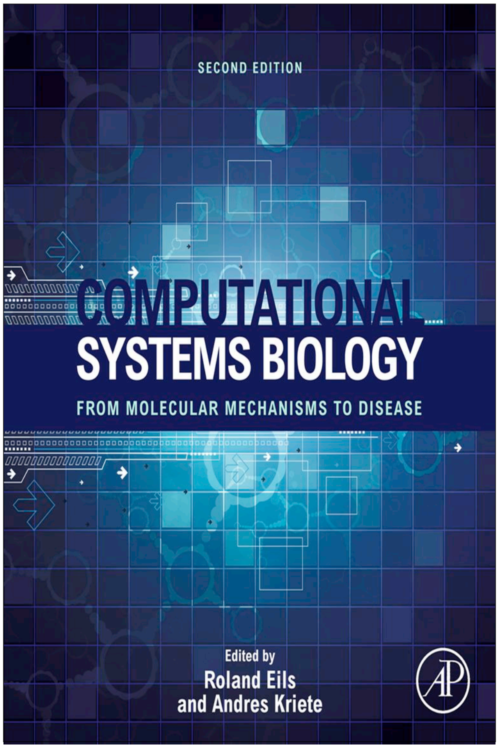


From Bahram Parvin et al., Molecular Correlates of Morphometric Subtypes in Glioblastoma Multiforme. In:
Andres Kriete and Roland Eils, editor, Computational Systems Biology. Chennai: Inc., 2014, p. 423-454.
ISBN: 978-0-12-405926-9
© Copyright 2014 Elsevier Inc.
Academic Press.

CHAPTER

18

Molecular Correlates of Morphometric Subtypes in Glioblastoma Multiforme

Hang Chang[a], *Ju Han*[a], *Gerald V. Fontenay*[a], *Cemal C. Bilgin*[a], *Nandita Nayak*[a], *Alexander Borowski*[b], *Paul Spellman*[c], *Bahram Parvin*[a]

[a]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[b]Center for Comparative Medicine, University of California, Davis, CA, USA

[c]Department of Biomedical Engineering, Oregon Health Sciences University, Portland, Oregon, USA

CONTENTS

http://dx.doi.org/10.1016/B978-0-12-405926-9.00018-6

Abstract

Integrated analysis of tissue histology with the genome-wide array and clinical data has the potential to generate hypotheses as well as be prognostic. However, due to the inherent technical and biological variations, automated analysis of whole mount tissue sections is impeded in very large datasets, such as The Cancer Genome Atlas (TCGA), where tissue sections are collected from different laboratories. We aim to characterize tumor architecture from hematoxylin and eosin (H&E) stained tissue sections, through the delineation of nuclear regions on a cell-by-cell basis. Such a representation can then be utilized to derive intrinsic morphometric subtypes across a large cohort for prediction and molecular association. Our approach has been validated on manually annotated samples, and then applied to a Glioblastoma Multiforme (GBM) cohort of 377 whole slide images from 146 patients. Further bioinformatics analysis, based on the multidimensional representation of the nuclear features and their organization, has identified (i) statistically significant morphometric sub types; (ii) whether each subtype can be predictive or not; and (iii) that the molecular correlates of predictive subtypes are consistent with the literature. The net result is the realization of the concept of pathway pathology through analysis of a large cohort of whole slide images.

1 INTRODUCTION

The interaction between underlying molecular defects and environmental factors can be captured by tumor histopathology; thus, we hypothesize that quantification of histology sections on a cell-by-cell basis in terms of morphological features and organization, leads to a new systems biology approach for the characterization and identification of molecular markers of tumor composition. As opposed to genome-wide array data, the large-scale quantitative characterization of tumor morphology from standard hematoxylin and eosin (H&E) stained tissue sections can offer alternative views for subtyping and survival analysis. Furthermore, computed morphometric indices can be tested against outcome. Meanwhile, derived representations (e.g., meta-features), from cellular level quantitative analysis, can also be utilized to probe for tumor heterogeneity and its underlying molecular basis. To answer the questions about morphometric indices that are predictive of the outcome, we have developed a computational pipeline to process large cohorts of whole mount tissue sections, which have been collected through The Cancer Genome Atlas (TCGA).

The computational pipeline consists of advanced algorithms for nuclear delineation (Chang et al. 2013-b) and tissue classification (classifying tissue into different component, e.g., tumor, necrosis), which are implemented efficiently to operate in a high performance computing environment on decomposed tissue blocks of 1k-by-1k pixels. The net result is a multidimensional representation of the tissue block that captures features at both nuclear level (e.g., size, shape, cellular density) and patch level (e.g., necrosis ratio). To tackle the large amount of technical and biological variations, expert annotated images with corresponding image-derived features (e.g., shape, intensity) are used for the construction of prior knowledge based on the Gaussian Mixture Model (GMM). In this context, the key contribution is to (i) utilize a dictionary of images for the characterization of technical variations and biological heterogeneity, and (ii) label each nucleus in the context of tumor histopathology, subjecting to spatial continuity with a graphcut formulation (Demir 2009; Latson et al. 2003). Nuclear architecture and organization are then constructed per patient as equal probability histograms that are normalized across all patients. A total of 377 GBM tumor sections from 146 patients are initially include a morphometric index, is examined for subtyping at patient level. Organization of the book chapter is as follows. Section 2 reviews prior literature. Section 3 summarizes nuclear segmentation and patch-based classification. Section 4 presents subtyping analysis and the characterization of heterogeneity. Section 5 describes the soft ware architecture. Section 6 concludes the chapter.

2 BACKGROUND

Histology sections are typically visualized with H&E stains that label DNA and protein contents, in various shades of color. Generally, these sections are rich in content since various cell types, cell states and health, cellular secretion, and cellular organization can be characterized by a trained pathologist with the caveat of inter and intra-observer variations (Dalton et al. 2000). Several reviews for the application and analysis of H&E sections can be found in Chang et al. (2011-a, 2013-b), Gurcan et al. (2009), Demir (2009). From our perspective, the trend and direction of the research community focuses on the following three key concepts.

The first key concept involves tumor grading through either rough or accurate nuclear segmentation (Latson et al. 2003) followed by cellular organization characterization (Doyle et al. 2011) and classification. In some cases, tumor grading has been associated with progression, recurrence, and invasion carcinoma (e.g., breast DCIS), where outcome is highly dependent on mixed grading (e.g., presence of more than one grade) and tumor heterogeneity. Since mixed grading appears to be present in 50% of patients (Miller et al. 2001), it introduces significant challenges to the pathologists. A recent study indicates that DCIS recurrence (Axelrod et al. 2008) in patients with more than one nuclear grade can be predicted through detailed segmentation and multivariate representation of nuclear features from H&E stained sections. In this study, nuclear regions were manually segmented from H&E stained samples, and each nucleus was profiled with a multidimensional representation. The significance of this particular study is that it has been repeated quantitatively to indicate prognostic outcome. In other related studies, nuclear features have also been shown to contribute to diagnosis and prognosis values for carcinoma of the colorectal mucosa (Verhest et al. 1990), prostate (Veltri et al. 2004), and breast (Mommers et al. 2001).

The second concept focuses on the patch-based (e.g., region-based) analysis of tissue sections based on features either engineered by human (Bhagavatula et al. 2010; Kong et al. 2010; Han et al. 2011; Kothari et al. 2012) or generated through automatic feature learning (Le et al. 2012; Huang et al. 2011). Automatic feature learning, in its simplest form, is based on independent component analysis (ICA), which computes kernels corresponding to oriented edge detectors. Another example is the independent subspace analysis (ISA), which learns invariant kernels from the data through non-linear mapping (Le et al. 2012). Yet, one of the shortcomings of ISA is that it lacks the ability to reconstruct original data, which can be attributed to its strict feed forward nature. However, this ability can be offered by some other techniques, such as Restricted Boltzmann Machines (RBM) (Hinton 2006) and Predictive Sparse Decomposition (PSD) (Kavukcuoglu 2008).

The last key concept suggests utilizing the detection of specific cells in the autoimmune system (e.g., lymphocytes) as a prognostic tool for breast cancer (Fatakdawala et al. 2010). As part of the adaptive immune response, the presence of lymphocytes has been correlated with nodal metastasis and HER2-positive breast cancer, GBM, and ovarian cancer (Zhang et al. 2003).

3 MORPHOMETRIC REPRESENTATION

Morphometric representation is in the context of nuclear segmentation and classification of tumor histopathology. Each component is summarized below.

3.1 Nuclear segmentation

The main barriers in nuclear segmentation are biological heterogeneity (e.g., cell type) and technical variations (e.g., fixation), which are visible in the dataset provided by TCGA. Present techniques have focused on thresholding with a subsequent morphological operation (Phukpattaranont et al. 2007; Ballaro et al. 2008; Petushi et al. 2006); fuzzy techniques (Latson et al. 2003; Land et al. 2008); geodesic active contour models (Fatakdawala et al. 2010; Glotsos et al. 2004); color separation followed by optimum thresholding and learning (Chang et al. 2009; Cosatto et al. 2008); hierarchical self-organizing map (Datar et al. 2008); and spectral clustering (Doyle et al. 2008). Several examples are given below. In Bunyak et al. (2011), multiphase level sets (Nath et al. 2006; Chang and Parvin 2010) were used for nuclear segmentation based on seeds detected through iterative radial voting (Parvin et al. 2007). In Al-Kofahi et al. (2010), the input image was initially classified into foreground and background regions with graph cut. The seeds were then collected from the foreground regions via a constrained multiscale *Laplacian of Gaussian (LoG)* filter and final segmentation was generated by coupling the classification along with seeds within graph cut framework. Similarly, in Kong et al. (2011), color texture extracted from the most discriminant color space was used to binarize the normalized input image into foreground and background regions; this was followed by an iterative operation, based on concave points and radial-symmetry, to split touching nuclei. Recently, a spatially constrained expectation maximization algorithm (Monaco et al. 2012) was proposed to address the "color nonstandardness" in histological sections in the HSV color space; however, the evidence shown in Section 3.1.7 B (MRGC vs MRGC-CF) indicates that strict incorporation of color and spatial information will not be sufficient. Another related work (Kothari et al. 2011) was built upon a consensus concept, where the labels were determined by multiple classifiers constructed from different reference images; we will refer to this method as MCV (multiclassifier voting), for short, in the rest of the book chapter. Although, MCV provides a better handler for the variation in the data; however, it is still possible to have noisy and erroneous classification (as shown in Figure 18.7), which is due to the lack of local statistical information and smoothness constraint.

In summary, the techniques above are often specific to small datasets that originate from a single laboratory, and ignore both the cellular heterogeneity (e.g., variation in chromatin patterns) and the technical variations manifested in both nuclear and background signals. As shown in Figure 18.1, our goal is to enable the processing of whole mount tissue sections, from multiple laboratories, to construct a large database of morphometric features, and to enable subtyping and genomic association.

Figure 18.2 shows the details of the proposed approach, where several key observations are leveraged for classifying nuclear regions: (i) global variations, across a large cohort of tissue

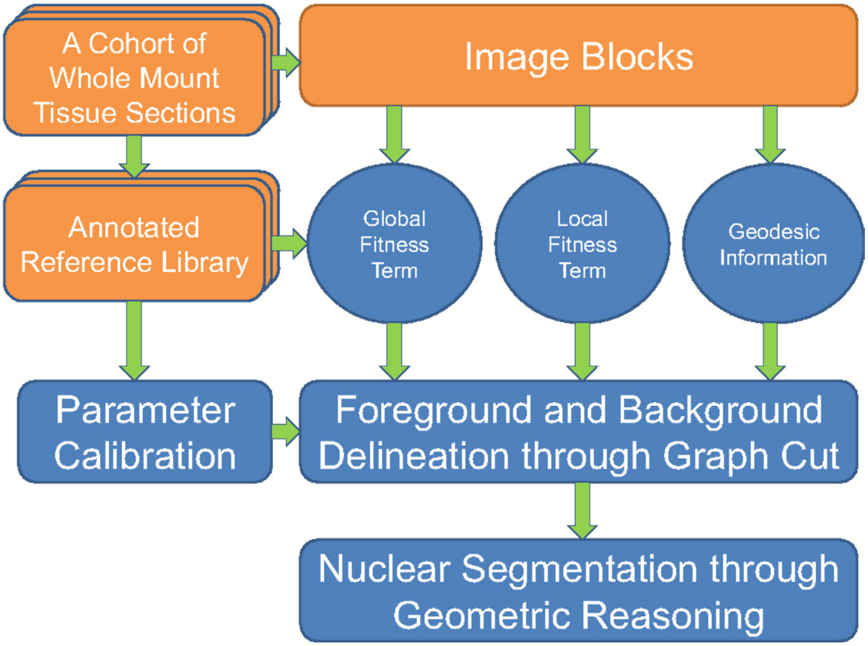

A Cohort of Whole Mount Tissue Sections
Image Blocks
Annotated Reference Library
Global Fitness Term
Local Fitness Term
Geodesic Information
Parameter Calibration
Foreground and Background Delineation through Graph Cut
Nuclear Segmentation through Geometric Reasoning

FIGURE 18.1 Work flow in nuclear segmentation for a cohort of whole mount tissue sections.

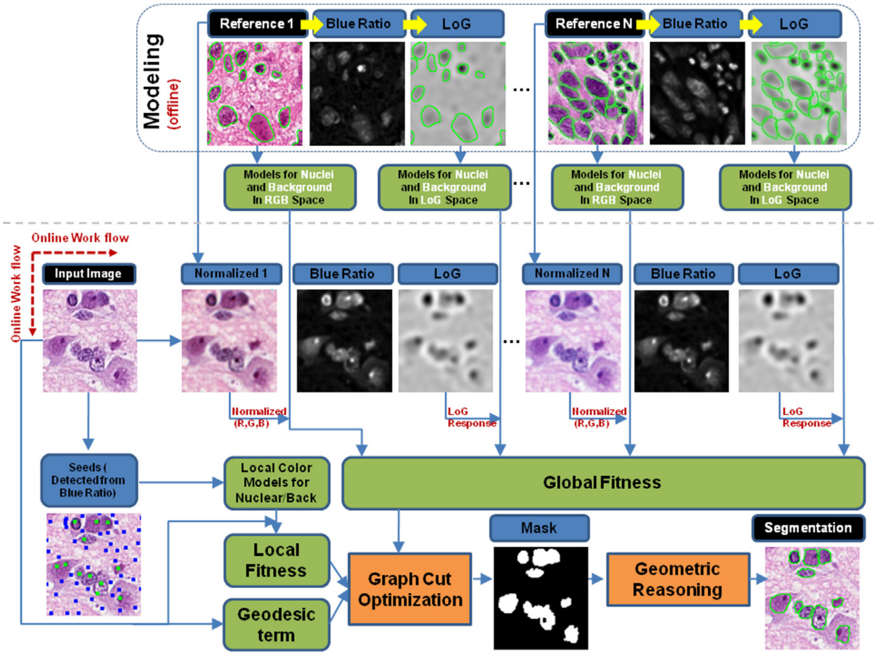


FIGURE 18.2 Steps in nuclear segmentation.

sections, can be captured by a representative set of reference images, (ii) local variations, within an image, can be captured by local foreground (nuclei)/background samples detected by *LoG* filter, and (iii) variations in image statistics, between a test image and a reference image, can be reduced by color normalization. These concepts are integrated within a graph cut framework, to separate the nuclei or clumps of nuclei from the background. Afterwards, the potential clumps of nuclei are partitioned through geometric reasoning. In the rest of this section, we summarize (a) the construction of the global prior models from a diverse set of reference images, (b) the strategy for color normalization, (c) the strategy for dimension reduction based on color transformation, (d) the details of feature extraction, (e) the multireference graph cut formalism for nuclei/background separation, and (f) the partitioning of a clump of nuclei into individual nucleus.

3.1.1 Construction and representation of priors

This step aims to capture the global variations for an entire cohort based on a well-constructed reference library. In our analysis, the target cohort consists of 377 individual tissue sections, from which a representative of N ($N = 20$) reference images of 1k-by-1k pixels at 20X have been selected by expert, based on staining and morphometric properties, in such a way that each one of them is a unique exemplar of tumor phenotypes. Therefore, we suggest that each reference image possesses a unique feature space, in terms of *RGB* and *LoG* responses, that leads to 2N feature spaces for the reference set:

$$\{F^1_{RGB_1}, F^2_{RGB_2}, \ldots, F^1_{RGB_N}, F^{N+1}_{LOG_1}, F^{N+2}_{LOG_2}, \ldots, F^{2N}_{LOG_N}\} \qquad (18.1)$$

where $F^{N+i}_{LoG_i}$ and $F^i_{RGB_i}$ and are *LoG* feature space and *RGB* feature spaces, respectively, for the i^{th} referencee image, $1 \leq i \leq N$. Subsequently, each reference image is manually segmented and processed with a *LoG* filter (please refer to Section 3.1.3 for the details on our *LoG* integration), at a fixed scale, followed by the collection of foreground (nuclei) and background statistics in both the *LoG* response and *RGB* space. Due to the distinct modes in feature spaces, we choose to capture the heterogeneities with *GMM*. Hence, the conditional probability for pixel, p, with feature, $f^k(p)$, in the k^{th} ($k \in [1, 2N]$) feature space, belonging to either nuclear region ($l = 1$) or the background region($l = 0$) can be expressed as a mixture with D component densities:

$$GMM^k_l(p) = \sum_{j=1}^{D} \tilde{p}(f^k(p)|j)p(j) \qquad (18.2)$$

Here $p(j)$ is the mixing parameter, which corresponds to the weight of component j, and $\sum_{j=1}^{D} P(j) = 1$. Each mixture component is a Gaussian with mean μ and covariance matrix Σ, in the corresponding feature space:

$$\tilde{p}(f^k(p)|j) = \frac{1}{2\pi^{\frac{3}{2}}|\sum|_j^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}(f^k(p) - \mu_j)^T \sum_j^{-1} (f^k(p) - \mu_j)\right) \qquad (18.3)$$

where $P(j)$ and $(\mu_j, \sum_j)$ for $\tilde{p}(C_p|j)$ are estimated by expectation maximization (EM) algorithm (Tomasi).

3.1.2 Color normalization

The purpose of color normalization is to reduce the variations between an input test image and a reference image in image statistics. Thus, the prior models, constructed from each reference image, can be applied. Here, we adopted the color map normalization (Kothari et al. 2011) for its effectiveness on histological images. Let

- input image, I, and reference image, Q, have K_I and K_Q unique color triplets, respectively, in terms of (R, G, B);
- $\mathbb{R}_C^{I/Q}$ be a monotonic function, which maps the intensity of a specific color channel, $C \in \{R, G, B\}$, from Image I or Q to a rank that is in the range $[0, K_I)$ or $[0, K_Q)$;
- (r_p, g_p, b_p) be the color of pixel p, in image I, and $(\mathbb{R}_R^I(r_p), \mathbb{R}_G^I(G_p), \mathbb{R}_B^I(b_p))$ be the ranks for intensities in each color channel; and
- the intensity values r_{ref}, g_{ref}, and b_{ref} in each color channel, from image Q, have ranks:

$$
\begin{aligned}
\mathbb{R}_R^Q(r_{ref}) &= \left\lfloor \frac{\mathbb{R}_R^I(r_p)}{K_I} \times K_Q + \frac{1}{2} \right\rfloor \\
\mathbb{R}_G^Q(g_{ref}) &= \left\lfloor \frac{\mathbb{R}_G^I(g_p)}{K_I} \times K_Q + \frac{1}{2} \right\rfloor \\
\mathbb{R}_B^Q(b_{ref}) &= \left\lfloor \frac{\mathbb{R}_B^I(b_p)}{K_I} \times K_Q + \frac{1}{2} \right\rfloor
\end{aligned}
\tag{18.4}
$$

As a result, the color for pixel, p: (r_p, g_p, b_p), will be normalized as $(r_{ref}, g_{ref}, b_{ref})$. Different from standard quantile normalization, which utilizes all pixel values in the image, color map normalization is based on the unique colors in the image, thereby, excludes color frequencies as a result of technical variations and tumor heterogeneity. Figure 18.2 shows some examples of color map normalization.

3.1.3 Color transformation

For more efficient integration of the LoG responses, a color transformation step is preferred to transform RGB space into a gray-level image for the accentuation of the nuclear stain and attenuation of background. Since present techniques in color decomposition (Rabinovich et al. 2003; Ruifork and Johnston 2001) are either very time-consuming or do not yield favorable outcomes, a more efficient strategy, which we refer to as blue ratio transformation, is proposed as follows: $BR(x,y) = \frac{100*B(x,y)}{1+R(x,y)+G(x,y)} \times \frac{256}{1+B(x,y)+R(x,y)+G(x,y)}$, where $B(x,y)$, $R(x,y)$, and $G(x,y)$ are the respective blue, red, and green intensities at position (x,y). In this formulation, the first and second terms accentuates nuclear stain while, the second term attenuates the background signals. Subsequently, the LoG responses are always computed at a single scale from the blue ratio image. In Figure 18.3 , we demonstrates the improvements resulting from the blue ratio transformation compared to color decomposition (Ruifork and Johnston 2001).

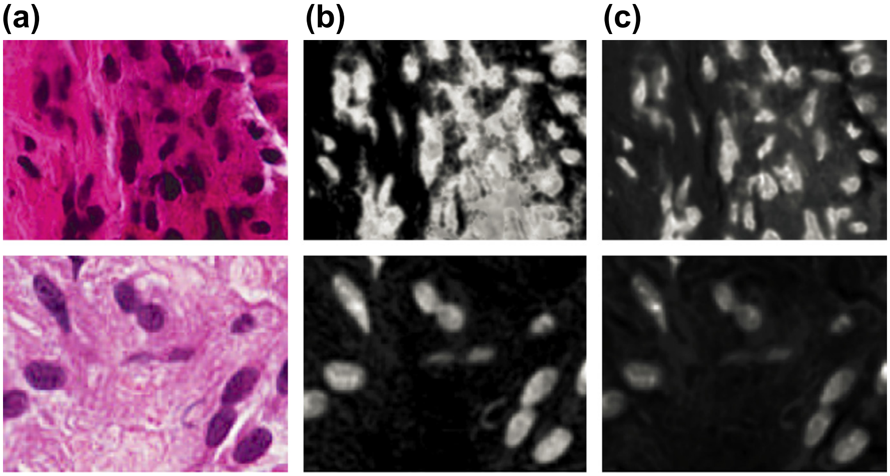


FIGURE 18.3 (a) Two diverse pinhole of tumor signatures; (b) decompositions by Ruifork and Johnston (2001); (c) blue ratio images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this book.)

3.1.4 Feature extraction

We integrate both color and scale information, where the color information is directly from the *RGB* space, and the scale information is encoded by the *LoG* response. The following are steps for feature extraction:

1. Normalize the input image against every reference image, as described in Section 3.1.2;
2. Transform each normalized image into a blue ratio image, as described in Section 3.1.3;
3. Apply the *LoG* filter to each blue ratio image, at a fixed scale; and
4. Represent each pixel, in the test image, by its *RGB* color in each of the normalized images and the *LoG* response, from each of the blue ratio images.

As a result, each pixel in the test input image, where the first N features are normalized *RGB* colors, and the last N features are *LoG* responses extracted from the blue ratio of the normalized images. All $2N$ features are assumed to be independent of each other, per the selection of reference images. The rational for our feature extraction strategy is that: (1) color information is insufficient for the delineation of nuclear regions from the background due to large variations in the data; (2) the scales of the nuclear region and background structure are typically different; and (3) the nuclear region responds well to *LoG* filter (Al-Kofahi et al. 2010).

3.1.5 Multi-reference graph cut model

Since the intrinsic and extrinsic variations in a cohort are incorporated within a graph cut framework, 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. Though the nodes and edges typically correspond to the pixels (P) and their adjacency relationships, respectively, there are special nodes known as terminals, which correspond to the set of labels to be assigned to the pixels.

For graphs with two terminals, the terminals are generally referred to as the source (S) and the sink (T). The labeling problem is to assign a unique label x_p (1 for foreground, and 0 for background) for each node, $p \in \bar{V}$, which is performed by minimizing the Gibbs energy E (Geman et al. 1984):

$$E = \sum_{p \in \bar{V}} E_{fitness}(x_p) + \beta \sum_{(p,q) \in \bar{E}} E_{smoothness}(x_p, x_q) \tag{18.5}$$

where $E_{fitness}(x_p)$ encodes the data fitness cost for assigning x_p to p, and $E_{smoothness}(x_p, x_q)$ denotes the cost when the labels of adjacent nodes, p and q, are x_p and x_q, respectively; additionally, β is the weight for $E_{smoothness}$. For more information regarding the Goldberg-Tarjan "push-relabel" methods (Goldberg and Tarjan 1988), and Ford-Fulkerson "augmenting paths" (Ford and Fullkerson 1962), the two groups of algorithms utilized in the graph cut optimization, please see Cook et al. (1998).

To capture the intrinsic variations of the nuclear signature, we expressed, the data fitness term as a combination of the global property map and intrinsic local probability map, where the former captures the global variations of the cohort, and the latter captures local intrinsic image property in the absence of color map normalization. Equation 18.5 is then rewritten as

$$E = \sum_{p \in \bar{V}} (E_{gf}(x_p) + E_{lf}(x_p)) + \beta \sum_{(p,q) \in \bar{E}} E_{smoothness}(x_p, x_q) \tag{18.6}$$

where E_{gf} and E_{lf} are the global and local data fitness terms, respectively, encoding the fitness cost for assigning x_p to p. Below, we discuss each of the terms together with the optimization process.

3.1.5.1 GLOBAL FITNESS TERM

The global fitness is constructed based on manually annotated reference images. Assume that there are N reference images: Q_i, $i \in [1, N]$. Additionally, for each reference image, *GMMs* are used to model the nuclear signal and background in both *RGB* space and *LoG* response space, respectively: GMM^k_{Nuclei}, $GMM^k_{Background}$, in which $k \in [1, 2N]$, and the first N *GMMs* are for the *RGB* space, and the last N *GMMs* are for *LoG* response space.

A normalized image, U_i, is first generated for the input test image, I, with respect to every reference image, Q_i. Subsequently, color and *LoG* responses of U_i are collected to construct $2N$ features per pixel, where the first N features are from the normalized color space, and the second N features are from *LoG* responses. Let,

- p be a node corresponding to a pixel;
- $f^k(p)$ be the k^{th} feature of p;
- α be the weight of *LoG* response;
- $\boldsymbol{p}^k_i$ be the probability function of f^k being nuclei $(l = 1)$/background $(l = 0)$:

$$\boldsymbol{p}^k_l(p) = \frac{GMM^k_l(p)}{\sum_{j=0}^{1} GMM^k_j(p)} \tag{18.7}$$

- λ_i be the weight for Q_i:

$$\begin{aligned}\lambda_i &= \frac{1}{3}\sum_{c}^{C\in\{R,G,B\}} \lambda_i^C \\ \lambda_i^C &= H^C(Q_i)H^C(U_i)/(||H^C(Q_i)||||H^C(U_i)||)\end{aligned} \tag{18.8}$$

where $||\cdot||$ is L_2 norm, $H^C(\cdot)$ is the histogram function on a specific color channel $C \in \{R,G,B\}$ of an image. Intuitively, λ measures similarity between two histograms derived from Q_i and U_i. It weighs the fitness for the application of the prior model, constructed from Q_i, onto the features extracted from the normalized image U_i.

Thus, the global fitness term is then defined as

$$E_{gf}(x_p = i) = -\sum_{k=1}^{N}\lambda_k \log(\boldsymbol{p}_i^k(f^k(p))) - \alpha\sum_{k=N+1}^{2N}\lambda_{k-N}\log(\boldsymbol{p}_i^k(f^k(p))) \tag{18.9}$$

where the first and second terms integrate normalized color features, and the second integrates the *LoG* responses.

3.1.5.2 LOCAL FITNESS TERM

At the cohort level, the global fitness term is designed through the utilization of both color and *LoG* information in the feature spaces of the references. However, the incorporation of information in the original color space of the input image is also important due to the local variations for a number of reasons, i.e., local lesions, non-uniformity in the tissue sections, etc. Taking this into consideration, the local data fitness of pixel, *p*, is constructed from the foreground and background samples in the neighborhood around *p*. These samples are detected by a *LoG* filter on the blue ratio image, where positive and negative peaks of the *LoG responses* often, but not always, correspond to the background and foreground (nuclear region), respectively. The following are details of the construction of the local fitness term:

1. Samples collection: This step aims to provide local foreground and background samples for further modeling of local image statistics. Figure 18.4 gives an example of the typical positive and negative peak responses associated with the *LoG* filter. To further improve the accuracy of the detected samples we've implemented the protocol outlined below:
   - **a.** Create a blue ratio image (Section 3.1.3): In this transformed space, the preferred frequency of the background intensity always corresponds to the peak of the intensity histogram.
   - **b.** Construct distributions of the foreground and background: here, we apply the *LoG* filter on the blue ratio image, and construct distributions of the blue ratio intensity at the detected peaks corresponding to the negative and positive *LoG* responses, respectively. Accuracy of detected samples can be improved in the following step.
   - **c.** Constrain the sample selection: Three criteria are applied to improve the accuracy of detected samples: (i) the *LoG* responses must be above a minimum conservative threshold to remove noise introduced by artifacts; (ii) the intensity associated with

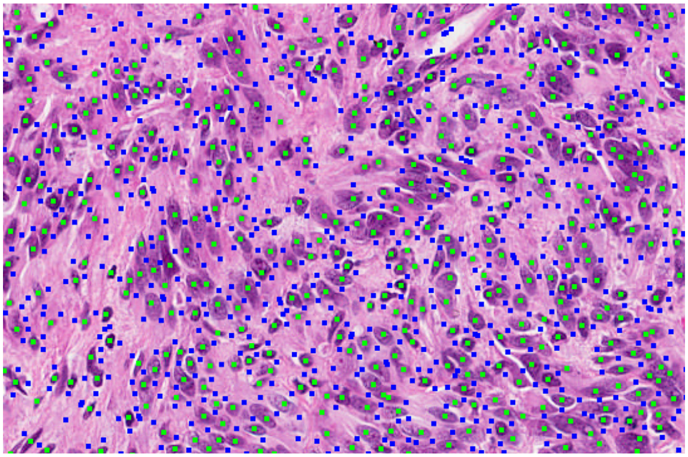
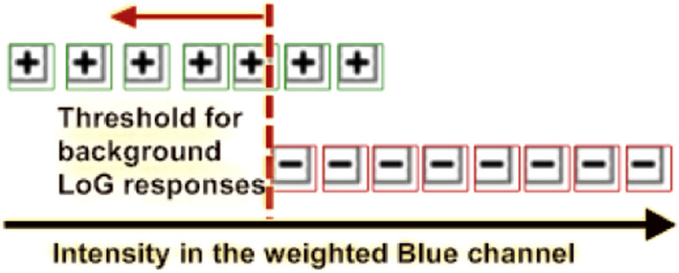


FIGURE 18.4 An example of the *LoG* response for detection of foreground (green dot) and background (blue dot) signals indicates an excellent performance on the initial estimate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this book.)



Threshold for background LoG responses
Intensity in the weighted Blue channel

FIGURE 18.5 *LoG* responses can be either positive (e.g. potential background) or negative (e.g. foreground or part of foreground) in the transformed blue ratio image. In the blue ratio image with the most negative *LoG* response, the threshold is set at the minimum intensity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this book.)

foreground samples must concur with the background peak, specified in step (a); and (iii) within a small neighborhood of $w_1 \times w_1$, the minimum blue ratio intensity, at the location of negative seeds, is set as the threshold for background peaks, as shown in Figure 18.5.

2. Local foreground and background color modeling: Foreground and background statistics, for each pixel, p, within a local neighborhood, $w_2 \times w_2$, are represented by two GMMs in the original color space, which correspond to the nuclei and background models, GMM^{Local}_{Nuclei}, *and* $GMM^{Local}_{Background}$, respectively.

Then the definition of the local fitness term is:

$$E_{lf}(x_p = i) = -\gamma \log(\boldsymbol{p}_l(f(p))) \tag{18.10}$$

where $f(p)$ refers to the RGB feature of node (pixel), p, in the original color space; γ weights the local fitness term; $\boldsymbol{p}_l$ is the probability function of f being Nuclei ($l = 1$)/background ($l = 0$), which results in

$$\boldsymbol{p}_l(p) = \frac{GMM_l^{Local}(p)}{\sum_{j=0}^{1} GMM_j^{Local}(p)} \tag{18.11}$$

3.1.5.3 SMOOTHNESS TERM

While both global and local data fitness terms encode the likelihood of a pixel being foreground/background, the smoothness term ensures the smoothness of labeling between adjacent pixels. In the graph configuration, fitness and smoothness are encoded by t-links (links between node and terminals) and n-links (links between adjacent nodes), respectively. Therefore, in order to utilize geodesic information, we follow the setup in Boykov and Kolmogorov (2003) for n-links. As a result, the max-flow/min-cut solution for the graph corresponds to a local geodesic or, in a continuous case, to a minimal surface. Given the weighted graph, $G = \langle \bar{V}, \bar{E} \rangle$, constructed in Section 3.1.5. Let,

- $\{e_k \mid 1 \leq k \leq n_G\}$ be a set of vectors for the neighborhood system, where n_G is the order of the neighborhood system, and the vectors are ordered by their corresponding angle ϕ_k with respect to the $+x$ axis, such that $0 \leq \phi_1 < \phi_2 \cdots < \phi_{nG} < \pi$. For example, when $n_G = 8$, we have $e_1 = (1, 0)$, $e_2 = (1, 1)$, $e_3 = (0, 1)$, $e_4 = (-1, 1)$, as shown in Figure 18.6(a);
- w_k be the weight for the edge between pixels: p and q, which are in the same neighborhood system, and $\vec{pq} = \pm e_k$;
- L be a line formed by the edges in the graph, as shown in Figure 18.6(c);
- C be a contour in the same 2D space where graph G is embedded, as shown in Figure 18.6(b);
- $|C|_G$ be the cut metric of C, as in

$$|C|_G = \sum_{e \in \bar{E}_C} W_e \tag{18.12}$$

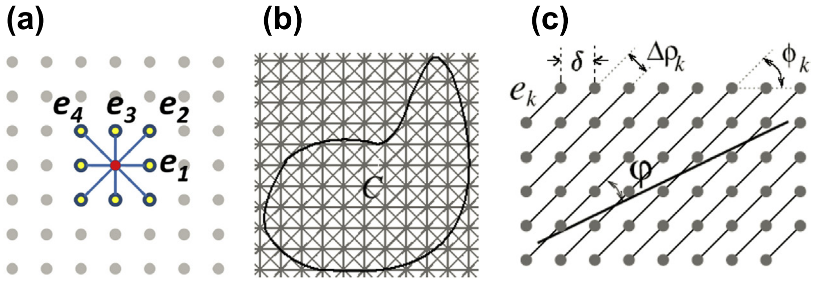


FIGURE 18.6 (a) Eight-neighborhood system: $n_G = 8$; (b) contour on eight-neighborhood 2D grid; (c) one family of lines formed by edges of the graph.

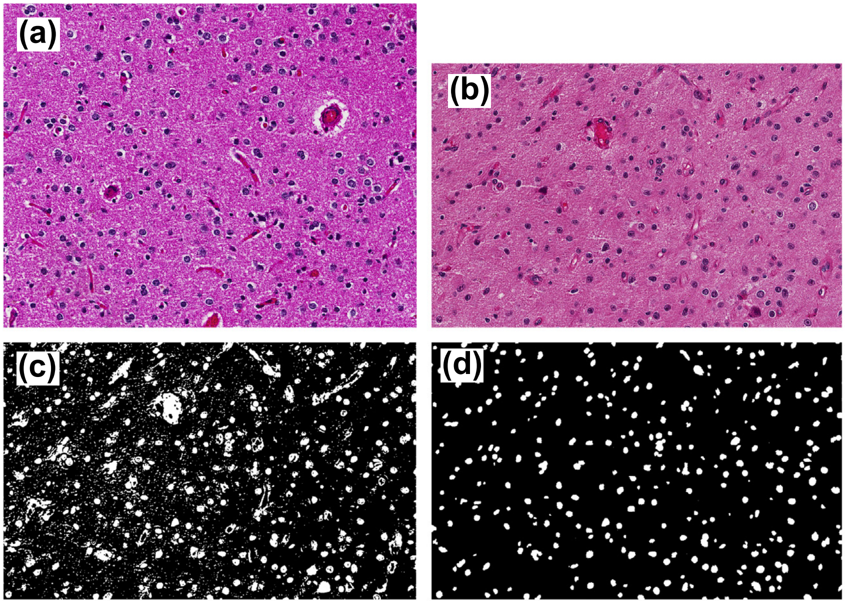


FIGURE 18.7 A comparison between MCV and MRGC (as shown in (c) and (d), respectively) based on the same reference image, as shown in (a). Even though the test image and the reference image are slightly different in color space, compared with MCV, MRGC still produces (1) more accurate classification, due to the encoding of statistics from test image's color space via local probability map; (2) less noisy classification due to the smoothness constrain.

where $\bar{E}_C$ is the set of edges intersecting contour C;

- $|C|_R$ be the Riemannian length of contour C;
- $D(p)$ be the metric(tensor), which continuously varies over point, p, in the 2D Riemannian space;

Based on Integral Geometry (Santalo 1979), the Crofton-style formula for the Riemannian length, $|C|_R$ of contour C, can be written as,

$$\int \frac{\det D(p)}{2(u_L^T \cdot D(p) \cdot u_L)^{\frac{3}{2}}} n_c dL = 2|C|_R \tag{18.13}$$

where u_L is the unit vector in the direction of the line L, and n_C is a function that specifies the number of intersections between L and C. According to Boykov and Kolmogorov (2003), the local geodesic can be approximated by the max-flow/min-cut solution ($|C|_G \rightarrow |C|_R$) with the following edge weight setting:

$$W_k(p) = \frac{\delta^2 \cdot |e_k|^2 \cdot \Delta\phi_k \cdot \det D(p)}{2 \cdot (e_k^T \cdot D(p) \cdot e_k)^{\frac{3}{2}}} \tag{18.14}$$

TABLE 18.1 Edge weights for the graph construction, where N is the neighborhood system, and β is the weight for smoothness.

Edge	Weight	For
$p \rightarrow S$	$E_{gf}(x_p = 1) + E_{lf}(x_p = 1)$	$p \in P$
$p \rightarrow T$	$E_{gf}(x_p = 0) + E_{lf}(x_p = 0)$	$p \in P$
$w_e(p, q)$	$\beta \cdot W_k(p)$	$\{p, q\} \in N, \phi_{\vec{pq}} \in \{\phi_k, \pi + \phi_k\}$

where δ is the cell-size of the grid, $\triangle\phi_k$ is the angular difference between the k^{th} and $(k+1)^{\text{th}}$ edge lines, $\triangle\phi_k = \phi_{k+1} - \phi_k$, and

$$D(p) = g(|\nabla|) \cdot \mathbf{I} + (1 - g(|\nabla|)) \cdot \mathbf{u} \cdot \mathbf{u}^{\mathrm{T}} \tag{18.15}$$

Here $\mathbf{u} = \frac{\nabla I}{|\nabla I|}$ is a unit vector in the direction of image gradient at point p, $\mathbf{I}$ is the identity matrix, and $g(x) = \exp(-\frac{x^2}{2\sigma^2})$.

3.1.5.4 OPTIMIZATION

Table 18.1 provides the details of the graph construction; the graph is further partitioned based on the max-flow/min-cut algorithm in Boykov and Kolmogorov (2004). As a result, the input test image is labeled into foreground and background. Details about the optimization can be found in Boykov and Kolmogorov (2004).

3.1.6 Nuclear mask partitioning

After the separation of foreground and background, the next step is to partition potential clumps of nuclei. Due to the fact that nuclear shape is typically convex, concavity detection and geometric reasoning (Wen et al. 2009) are applied to address the ambiguities associated with the delineation of overlapping nuclei. Details can be found in Wen et al. (2009).

3.1.7 Experimental results and discussion

In this section, we (i) discuss parameter setting, and (ii) evaluate performance of the system against previous methods.

3.1.7.1 EXPERIMENTAL DESIGN AND PARAMETER SETTING

Our experiment was carried out at 20X, where 20 reference images with size 1k-by-1k pixels were manually selected and annotated by an expert to capture the technical variations and biological heterogeneities in the target cohort. During nuclear segmentation, only the top $M = 10$ reference images with the highest weight of λ were used as a trade-off between computational complexity and performance. The number of components for *GMM* was evaluated and selected to be $D = 20$, while the parameters of *GMM* were estimated via *EM* algorithm. The other parameter were set at: $\alpha = 0.1$, $\beta = 10.0$, $\gamma = 0.1$, $w_1 = 100$, $w_2 = 100$, and $\sigma = 4.0$ (the scale for both seeds detection and *LoG* feature extraction), where w_1 was selected to minimize the seeds detection error on the annotated reference images; σ was determined based

TABLE 18.2 Comparison of average classification performance among our approach(MRGC), our previous approach (Chang et al., 2011-a), MCV approach in Kothari et al. (2011), and random forest. For MCV, only color in *RGB* space is used, which is identical to Kothari et al. (2011). For random forest, the same features are used: $\{R, G, B, LoG\}$, and the parameter settings are: $ntree = 100$, $mtry = 2$, $node = 1$.

Approach	Precision	Recall	F-Measure
MRGC-MS (Multi-Scale LoG)	0.77	0.82	0.794
MRGC	0.79	0.78	0.785
MRGC-CF (Color Feature Only)	0.72	0.83	0.771
MRGC-GF (Global Fitness Only)	0.80	0.71	0.752
Our Previous Approach	0.78	0.65	0.709
MCV	0.69	0.75	0.719
Random Forest	0.59	0.76	0.664

on the preferred nuclear size at 20X; and all other parameters were selected through cross validation.

3.1.7.2 EVALUATION

A two-fold cross validation, with optimized parameter settings, was applied to the reference images, followed by a comparison of average classification performance between our approach, MCV (Kothari et al. 2011), and random forest (Breiman 2001). As summarized in Table 18.2, our approach exhibits a superior performance.

The effectiveness of the local probability map is demonstrated with an intuitive example, as shown in Figure 18.8, where the characterization of the nuclei with low chromatin content

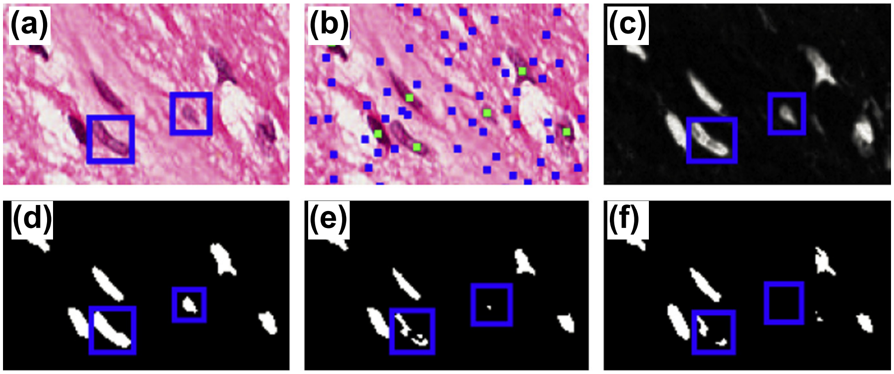


FIGURE 18.8 A comparison among our approach, MCV, and random forest. (a) Original image patch; (b) detected seeds, green: nuclei region; blue: background; (c) local nuclei probability established based on seeds; (d) classification by our approach; (e) classification by MCV; (f) classification by random forest. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this book.)

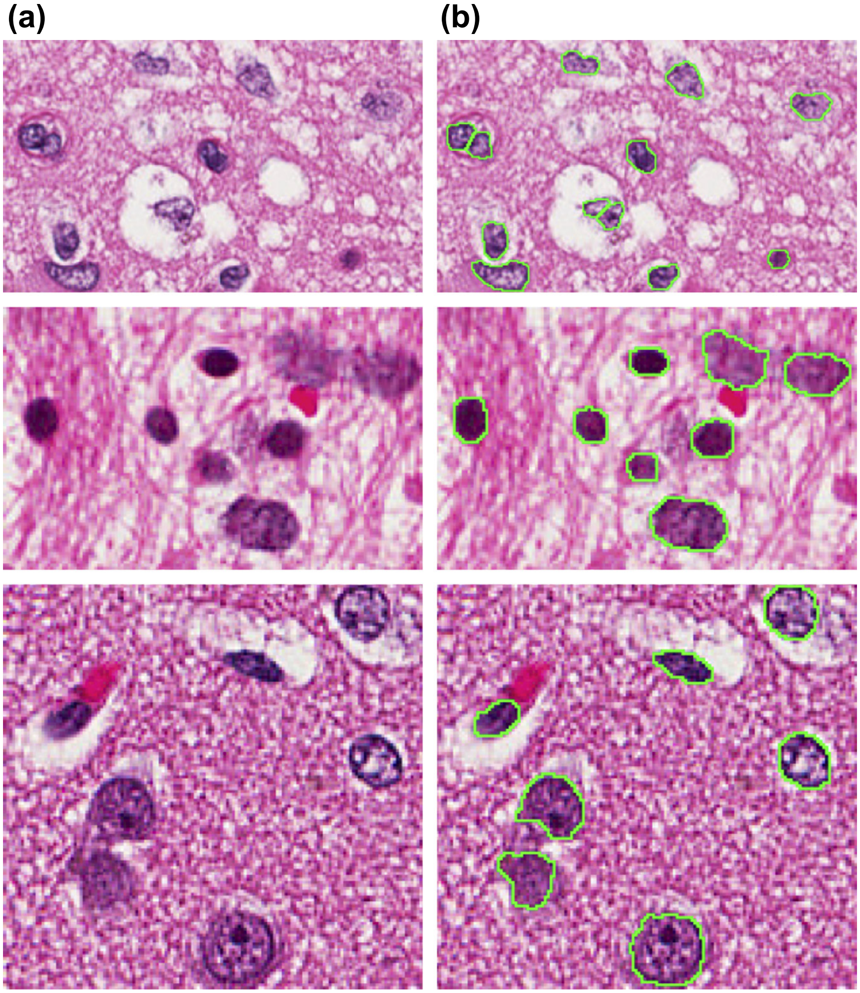


FIGURE 18.9 Segmentation on nuclei with low chromatin patterns. (a) Original image patch; (b) segmentation results.

(shown in the blue bounding boxes) is clearly improved with the help of local probability map. Figure 18.9 gives another example for further demonstration of the effectiveness of our approach on the segmentation of low chromatin nuclei.

Finally, Table 18.3 gives an object level comparison of the nuclear segmentation performance between our previous approach (Chang et al. 2011-a) and our current approach (Chang et al. 2011-a). Let,

- *MaxSize*(a, b) be the maximum size of nuclei a and b,
- *Overlap*(a, b) be the amount of overlap between nuclei, a and b.

TABLE 18.3 Comparison of average segmentation performance between our current approach (MRGC), and our previous approach (Chang et al. 2011-a), in which $precison = \frac{\#correctly\_segmented\_nuclei}{\#segmented\_nuclei}$, and $recall = \frac{\#correctly\_segmented\_nuclei}{\#manually\_segmented\_nuclei}$.

Approach	Precision	Recall	F-Measure
MRGC	0.75	0.85	0.797
Our Previous Approach	0.63	0.75	0.685

Subsequently, given any manually annotated nucleus, n_G, as ground truth, if there is one and only one segmented nucleus, n_S, that satisfies $\frac{Overlap(n_G,n_S)}{MaxSize(n_G,n_S)} > T$, then n_S is considered to be the correct segmentation for n_G. In our experiment, the threshold was set to be $T = 0.8$ empirically.

3.2 Patch-based analysis

The flip side of morphometric analysis is to quantify composition of each tissue section in terms of distinct histopathology, such as tumor or necrosis regions. This allows each nucleus can be tracked to its specific compartment. It is our suggestion that, compared to human engineered features (Lowe 1999; Dalal and Triggs 2005), unsupervised feature learning is more tolerant to batch effect (e.g., technical variations associated with sample preparation) and can learn pertinent features without user intervention. In our case, features are learned using PSD [74]. It contains both a feed-forward stage (e.g., encoding) and a feed backward stage (e.g., decoding), where the decoding step reconstructs the original patch through a sparse activation of an over-complete dictionary, and the encoding step efficiently produces the sparse code directly from the the original patch. The learned features are then summarized utilizing some pooling strategies for improved robustness of the representation, which will eventually be used towards the construction of classifier. Figure 18.10 indicates the overall recognition framework of our approach.

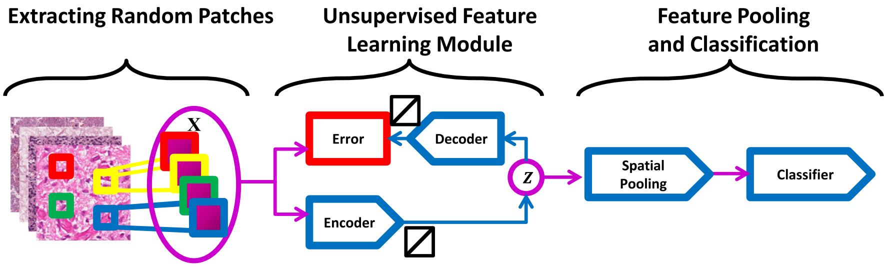

Extracting Random Patches
Unsupervised Feature Learning Module
Feature Pooling and Classification
X
Error
Decoder
Z
Encoder
Spatial Pooling
Classifier

FIGURE 18.10 Illustration of recognition framework including the encoder, decoder, and pooling.

3.2.1 Unsupervised feature learning

Randomly selecting from the cohort, the sparse auto encoder takes a set of vectorized image patches, X, as input, with the objective is to constructing a sparse representation, Z, for each input, X, with a highly efficient feed forward fashion. Meanwhile, the feedback mechanism minimizes the reconstruction error of the original signal based on an over-complete dictionary, D. The objective function is as follows:

$$F(X) = ||WX - Z||_F^2 + \lambda||Z||_1 + ||DZ - X||_F^2 \qquad (18.16)$$

where $X \in \mathcal{R}^n$, $Z \in \mathcal{R}^k$, the dictionary is $D \in \mathcal{R}^{n\times k}$, and the encoder is $W \in \mathcal{R}^{k\times n}$. The first and last terms are for encoding and decoding, respectively, whereas the second term denotes the sparse constraint with λ as the weighting parameter for sparsity, i.e., sparsity is increased with a higher value of λ.

In our experiment, to achieve the best performance in classification, λ is set to be 0.3 empirically.

The optimal D, W, and Z are learned through the minimization of $F(X)$, which is iterative by fixing one set of parameters while optimizing others and vice versa, i.e., iterate over steps (2) and (3) below:

1. Randomly initialize D and W.
2. Fix D and W and minimize Equation 18.16 with respect to Z, where Z for each input vector is estimated via the gradient descent method.
3. Fix Z and estimate D and W, then approximate D *and* W through stochastic gradient descent algorithm. This is used to improve the scalability of the optimization, which is necessary due to the large scale of training samples.

Figure 18.11 shows a few examples of the dictionary elements, computed from GBM dataset, which captures color and texture information in the cohort. Generally, this information is difficult to obtain using hand engineered features.

3.2.2 Classification

During classification, every image, in the training dataset, is divided into non-overlapping image patches, which are further processed with the feed forward encoding operation, $Z = WX$, followed by the max-pooling strategy upon the sparse codes forming the features for training. Here, a multi-class regularized SVM is used for classification with a regularization parameter 1 and a polynomial kernel of degree 3.

3.2.3 Evaluation

We opted to curate three classes that correspond to necrosis, transition-to-necrosis, and to tumors. The pure necrotic regions are free of DNA contents. However, as an intermediate step of the dynamic process of necrosis, transition-to-necrosis regions have punctated or diffused DNA contents. Our evaluation involves a dataset containing 1400 images curated from samples scanned with 20X objective. During unsupervised feature learning, 50 patches of size 25 × 25 pixels were randomly selected for each image in the dataset. They were down sampled by a factor of 2 and normalized in the range of [0,1] in the color space before being fed into the system. The size of the dictionary was set to be 1,000 to achieve the best classification

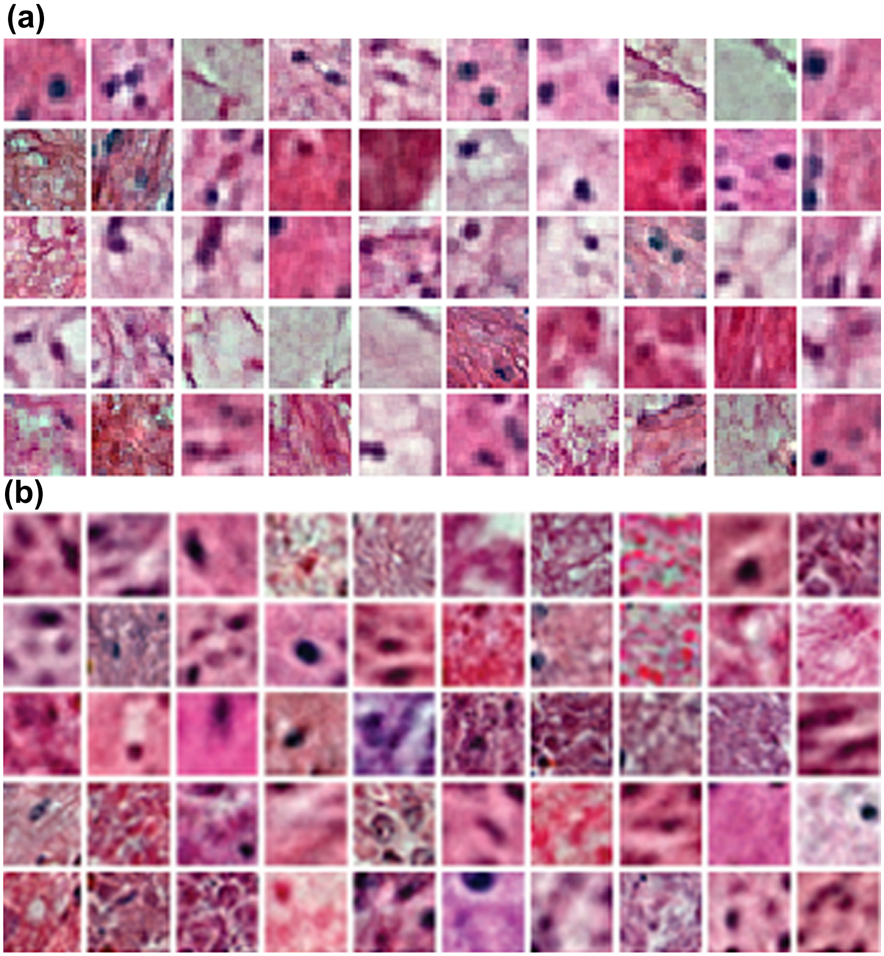


FIGURE 18.11 Representative set of computed basis function, *D*, for (a) the KIRC dataset and (b) the GBM dataset.

performance during cross-validation, and the max-pooling strategy was performed on a 4-by-4 neighboring patch of the learning step. During classification, the total number of necrosis, transition-to-necrosis, and tumor patches were 12,000, 8,000, and 16,000, respectively. From, amongst those, 4,000 patches per category were randomly selected for training, and another 4,000 per category were randomly selected for testing. We repeated the classification for 100 times, and reported the performance in Table 18.4. Figure 18.12 shows an example of the reconstruction of a heterogeneous image with the tumor region on the right and transition-to-necrosis on the left. Based on the dictionary derived above, this indicates that the transition-to-necrosis region is visually distinguishable from the tumor during reconstruction.

TABLE 18.4 Confusion matrix for classifying three different morphometric signatures in GBM.

Tissue type	Necrosis	Tumor	Transition to necrosis
Necrosis	77.6	7.7	14.6
Tumor	0.5	93.3	6.0
Transition to Necrosis	10.9	6.3	82.8

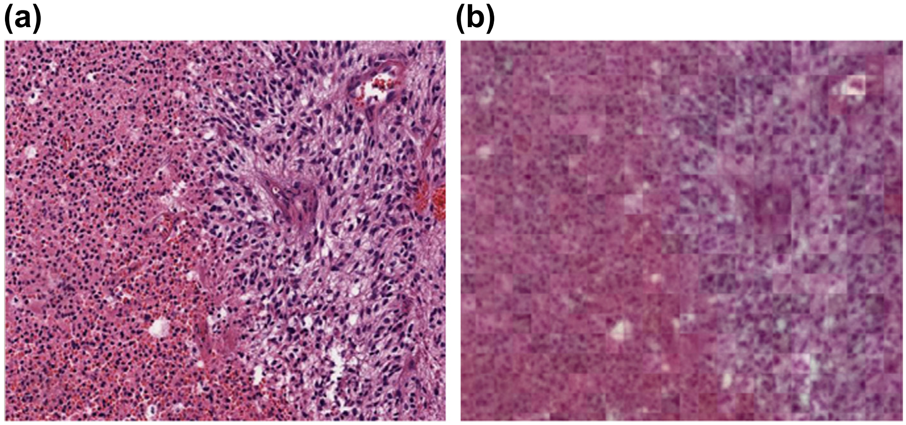


FIGURE 18.12 (a) A heterogeneous tissue section with transition to necrosis on the left and tumor on the right, and (b) its reconstruction after encoding and decoding.

Figure 18.13 shows some examples of classification on whole slide tissue sections with size 20k × 20k pixels. The examples are consistent with the evaluation and annotation from our pathologist, and further demonstrate the efficacy of our system.

4 BIOINFORMATICS ANALYSIS

Integrated analysis of tissue histology with the genome-wide array (e.g., OMIC) and clinical data has the potential to generate hypotheses as well as be prognostic. Different from typical subtyping analysis at patient level (Chang et al. 2011a, 2013b), we focus on subtypes that correspond to tumor composition, and evaluate every morphometric index or pairs of morphometric indices that are predictive of the outcome. Meanwhile, indices for morphometric heterogeneity are also derived in order to identify the molecular basis of tumor heterogeneity.

4.1 Morphometric summarization and subtyping at the block level

Morphometric summarization and subtyping on tissue patches provide a way for tissue compositional analysis. In this study, each tissue section is decomposed into non-overlapping patches (blocks) of 1k × 1k pixels. Distributions of each computed morphometric index are then

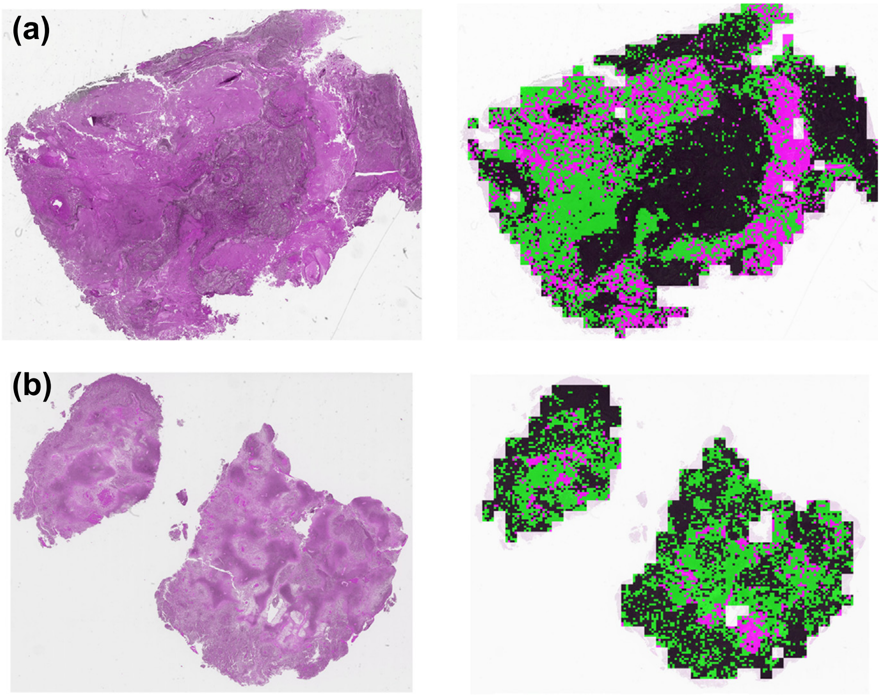


FIGURE 18.13 Two examples of classification results of a heterogeneous GBM tissue sections. The left and right images correspond to the original and classification results, respectively. Color coding is black (tumor), pink (necrosis), and green (transition to necrosis). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this book.)

constructed per block and normalized across all tissues within the same tumor type, thereby enabling morphometric subtyping at block level for a given tumor type as well as subsequent survival and compositional analysis. However, prior to the analyses, several problems will first need to be solved: (i) undesired effects on subtyping caused by background/border blocks in the tissue section; (ii) extremely large number of blocks per tissue section, e.g., 2500; (iii) imbalanced number of blocks per tissue section due to the variation of tissue size. To address the issues above, a computational pipeline with four major steps has been developed:

1. *Block filtering:* Any blocks containing background regions, where the background is detected at low resolution of the tissue section, are removed from subsequent analysis.
2. *Block Sampling:* Representative blocks for each tissue section are identified and selected as the centroids of the morphometric clusters, then they are computed through a *k*-means algorithm on all the blocks within the tissue section. As a result, the number of selected blocks is the same as the number of clusters, k, which is set to be proportional to the total number of blocks in the tissue section (e.g., 1%).
3. *Block Clustering:* Consensus clustering (Monti et al. 2003) is performed on representative blocks across different tissue sections for the identification of subtypes of a given tumor type, where the derived subtypes are further refined by removing the outliers through

the silhouette analysis (Rousseeuw 1987; R. V. et al. 2010). Eventually, the refined subtypes and their corresponding blocks are used to construct a classifier for block labeling.

4. *Block labeling*: Each non-representative block is assigned to a subtype through the nearest-neighbor classifier, which built from previous step.

4.2 Integrated analysis at the patient level

The block level subtyping, which enables a compositional representation for each patient, indicates the percentage of each subtype (i.e. the percentage of the blocks belonging to each subtype) at the patient level. Subsequently, this allows each subtype to be correlated with genomic data or clinical covariates for integrated analysis. In the field of multivariate survival analysis, one possible way to explore the relationship between the compositional covariates and survival distribution is to utilize the following parametric model (Fox 2002):

$$h\,(t) = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \tag{18.17}$$

where $h(t)$ is the hazard function, $X_i(i \in [1,k])$ are the covariates, and α is a constant representing the log-baseline hazard. Without specifying the baseline hazard function $\alpha(t) = \log h0(t)$, the following Cox proportional hazards (PH) model can be estimated by the partial likelihood method,

$$h\,(t) = h0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \tag{18.18}$$

To further explore the relationship between compositional covariates and survival distribution in the presence of important clinical covariates (e.g., age at initial pathologic diagnosis), the Cox PH model can be rewritten as:

$$h\,(t) = h0(t)\exp(\beta_1 C_1 + \beta_2 C_2 + \cdots + \beta_{N-1} C_{N-1} + \beta_N Age) \tag{18.19}$$

Here, C_i is the percentage of the i^{th} subtype derived at the patient level. Due to the linear correlation among all the compositional covariates, $\sum_{i=1}^{N} C_i = 1$, only $N - 1$ of them are included in the model. From amongst these, the ones with small p-values (those under a certain threshold) are identified as statistically significant predictors of survival distribution.

Consequently, the identified histological predictor can now be used to infer the best correlated molecular candidates through correlation analysis. To this end, we adopted Pearson's product-moment correlation coefficient for this study. The significance of computed correlation between the histological predictor and expression values of each probe set for all available patients were further assessed by a two-tailed *t-test* with *n-2* degrees of freedom (*n* is the number of patients), where *p*-values for all probe sets were computed and corrected for multiple testing using a false discovery rate (FDR) Hochberg 1995.

4.3 Clustering results

Our representation for each block is a concatenated vector, from two 25-bin equal probability histograms, for nuclear size and cellularity, respectively. During the block filtering

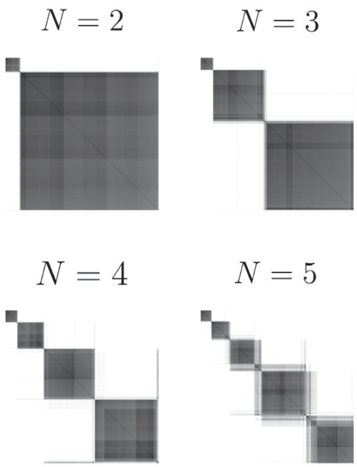


FIGURE 18.14 Consensus clustering matrix of 146 TCGA patients with GBM for cluster number $N = 2$ to $N = 5$.

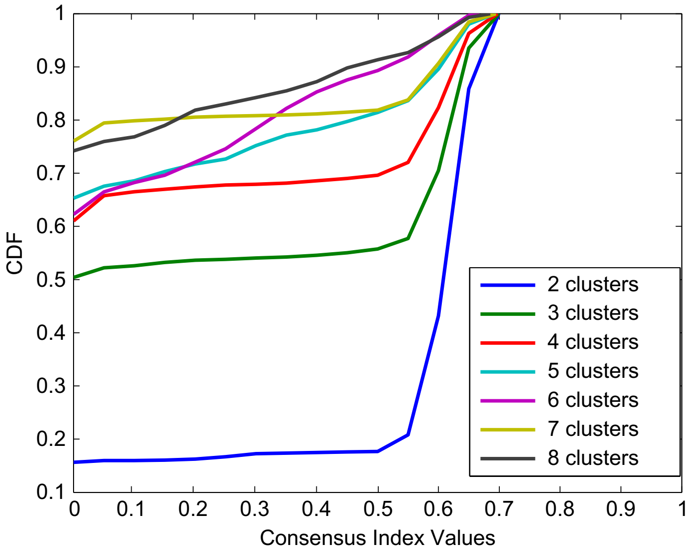


FIGURE 18.15 Consensus clustering CDF for cluster number $N = 2$ to $N = 8$.

step, a total of 162,510 tissue blocks (non-background/non-border) were identified, among which 1,582 representative ones were selected for consensus clustering. In the consensus clustering step, the k-means algorithm, with squared Euclidean distance as the distance metric, was repeated for 200 iterations with a sampling rate of 0.8. As shown in Figures 18.14 and 18.15, respectively, the derived consensus matrix and CDFs (cumulative density function)

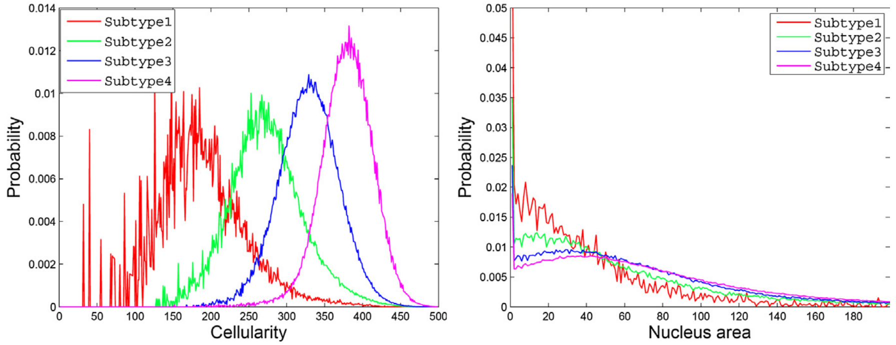


FIGURE 18.16 Average equal-bin-width histograms of cellularity and nuclear size for each block-level subtype ($N = 4$).

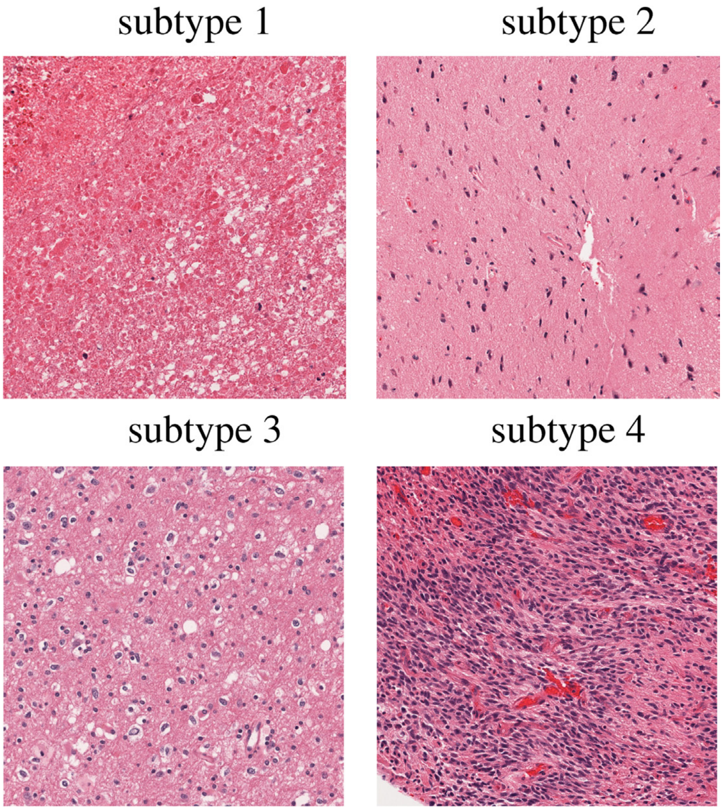


FIGURE 18.17 Representative blocks for each morphometric subtype. Each block is of 1000-by-1000 pixels at 20× resolution.

reveal four robust clusters (clustering stability significantly decrease for $N > 4$). Among all the representative blocks in the four subtypes (clusters) identified above, 1535 of them with positive silhouette value were retained as training samples for the construction of classifier. The average equal-bin-width histograms for each base block, from each subtype, are shown in Figure 18.16. Figure 18.17 gives examples of representative blocks from each one of the four subtypes, which exhibit significantly different signatures in cellularity (i.e., subtypes 1 to 4 correspond to extremely low, mid and high cellularity, respectively) and nuclear size (i.e., subtypes 1 to 4 exhibits a monotonically increasing trend in nuclear size).

4.4 Survival analysis and genomic association

Block labeling leads to a compositional representation at patient level, indicating the percentage of each subtype per patient. In the presence of age, the relationship between the patient level tumor composition and survival distribution, is then modeled as independent prognostic factors in Equation 18.19.

The implementation of survival analysis is based on the R survival package. Due to the linear dependence among all four compositional covariates, as mentioned in Section 4.2, we evaluate all possible combinations of the three covariates for survival analysis. The results, summarized in Table 18.5, indicates that both age and C_1 (Subtype1 composition) have consistently high hazard ratio with p-values < 0.1 (i.e., these covariates are negatively correlated with survival). As shown in Figures 18.16 and 18.17, subtype1 reveals a necrotic-like signature of small nuclear size and extremely low cellularity. Consistent with previous literature, this result indicates a negative correlation between the extent of necrosis and survival in GBM (Pierallini et al. 1998). Figure 18.18 shows a heat map of 48 probe sets that are significantly correlated with the subtype1 composition, with FDR adjusted p-value < 0.02. These probe sets were mapped into genes for further analysis.

Pathway and subnetwork enrichment analysis (see Figure 18.19) are then performed on the identified genes determined to have a significant correlation to the subtype1 composition. Pathway enrichment revealed STAT3, which is known to be a master regulator in GBM (Liu et al. 2010; Rahman et al. 2002), while the subnetwork enrichment identified AGT, PKC, PDGF, CEBPA, and TNF as the major hubs.

Temozolomide (TMZ), as a part of treatment for the patients in this cohort, interferes with DNA replication through methylation. However, some tumor cells are able to

TABLE 18.5 Multivariate survival analysis results by fitting the Cox PH model.

	Covariates in the Cox PH model							
	C1 + C2 + C3 + Age		**C1 + C2 + C4 + Age**		**C1 + C3 + C4 + Age**		**C2 + C3 + C4 + Age**	
	Hazard ratio	***p-value***	***Hazard ratio***	***p-value***	***Hazard ratio***	***p-value***	***Hazard ratio***	***p-value***
C1	1.0184	**0.0652**	1.0168	**0.0856**	1.030	**0.0631**	NA	NA
C2	0.9885	0.2342	0.9869	0.2771	NA	NA	0.9706	**0.0631**
C3	1.0016	0.7303	NA	NA	1.013	0.2771	0.9834	**0.0856**
C4	NA	NA	0.9984	0.7303	1.012	0.2342	0.9819	**0.0652**
Age	1.0283	**7.37e-5**	1.0283	**7.37e-5**	1.028	**7.37e-5**	1.0283	**7.37e-5**

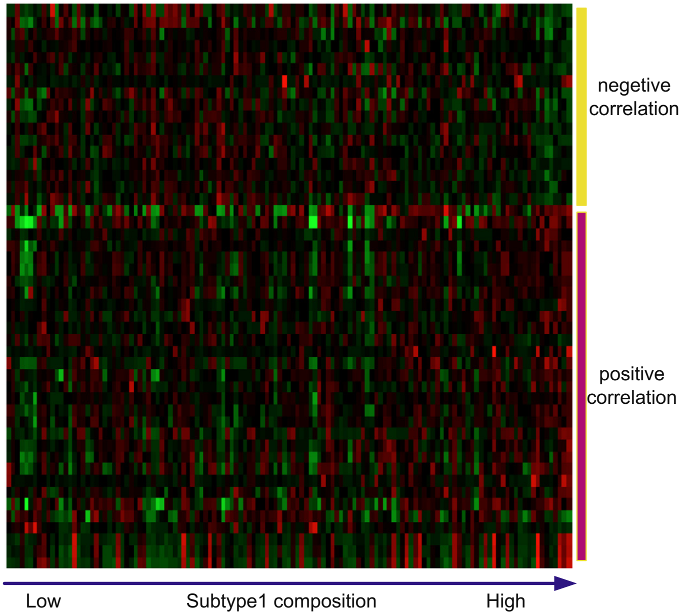


FIGURE 18.18 Heatmap of top 48 probesets (rows) that best correlate with the subtype1 composition, with FDR adjusted *p*-value < 0.02.

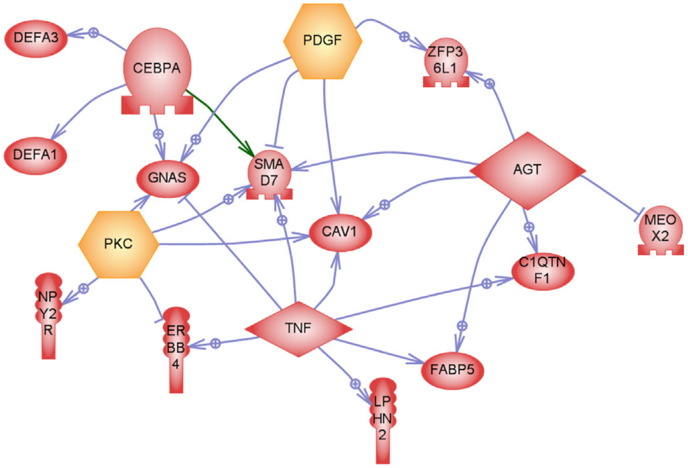


FIGURE 18.19 Subnetwork enrichment analysis for Subtype 1 reveals AGT, PDGF, PKC, TNF, and CEBPA as dominant regulators with *p*-value of less than 0.05.

repair the damage through the expression of AGT. In GBM, AGT maintains normal function of vasculature (Kakinuma et al. 1997), and cellular concentration of AGT enzyme is a primary determinant of the cytotoxicity of TMZ (Stupp et al. 2001) *in vitro*. Whereas PKC (Protein Kinase C) is well established in cancer signaling and therapy, as it is involved in proliferation, migration, and malignant transformation (Kazanietz 2010), and its isozyme has been suggested for chemotherapeutic targets in GBM (Martin and JHussanini 2005). TNF, on the other hand, refers to a group of cytokines that induce proliferation, inflammation, and apoptosis, depending upon the adaptor proteins. TNF is part of the anti-tumor strategy in which human glioma cell lines express its proteins. Manipulation of these proteins has shown to induce apoptosis in glioma cells (Chen et al. 1997). Other hubs are highly ranked in the TCGA gene tracker.

5 COMPUTATIONAL PIPELINE

The significance of our computational pipeline is its capacity of large-scale data analysis, which meets the TCGA requirements on data processing. As shown in Figure 18.20, the pipeline has the following four components: (I) consistency maintaining between the local and remote registries, (II) tissue section visualization, (III) data processing and feature importing, and (IV) data summarization through normalization. Details of each component are as follows:

I. Consistency of images between a local registry and TCGA registry (at the National Cancer Institute (NCI)) is constantly maintained. Images newly imported into TCGA's registry are synchronized with the local registry for processing. At present,

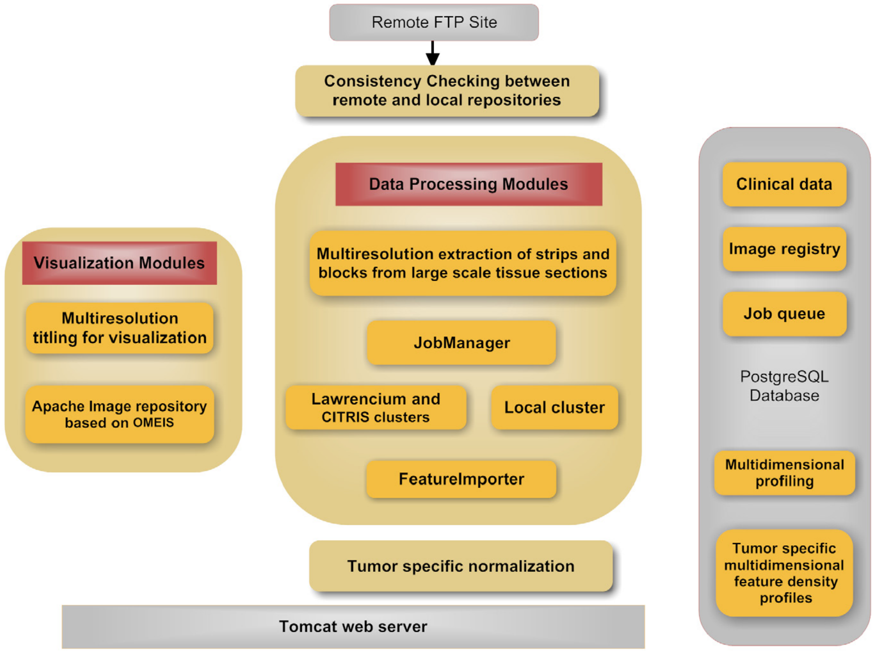

Remote FTP Site
Consistency Checking between remote and local repositories
Visualization Modules
Multiresolution titling for visualization
Apache Image repository based on OMEIS
Data Processing Modules
Multiresolution extraction of strips and blocks from large scale tissue sections
JobManager
Lawrencium and CITRIS clusters
Local cluster
FeatureImporter
Tumor specific normalization
Tomcat web server
Clinical data
Image registry
Job queue
PostgreSQL Database
Multidimensional profiling
Tumor specific multidimensional feature density profiles

FIGURE 18.20 Computational pipeline consists of four modules: downloads images from the NIH repository. Each image is partitioned into strips of (1k-by-number of columns), stored in the OMEIS image server. Each strip is partitioned into blocks of 1k-by-1k pixels, where each block is submitted to one of the two clusters at Berkeley lab. Computed representations are then imported into a PostgreSQL database.

both frozen sections and those from paraffin embedded blocks are provided; both types of images are registered and displayed through our system, but only paraffin embedded blocks are analyzed. Each image is partitioned and stored as strips of 1k-by-number of columns on the OME image server (Goldberg et al. 2005; Parvin et al. 2000).

II. Visualization of each whole mount tissue section is provided with a GoogleMaps™-like interface, which is realized through tiling and the utilization of Flash technology. Prior to visualization, each tissue section is partitioned and stored as tiles of 256-by-256 pixels at various resolutions. When a user interaction (e.g., drag and zoom) with the interface in an effort to view the tissue section in a browser, the required tiles are immediately downloaded from the server and assembled for display. All data and images are publicly available on the following website: http://tcga.lbl.gov.

III. During data processing, each strip is partitioned into 1k-by-1k blocks, and submitted for cluster computing, where the block size is optimized with respect to processing time and wait time in the queue. Currently, it takes four days to process the entire GBM data set, which consists of 344 tissue sections. To further reduce the computational cost, a multithread implementation of the computational methods has been developed to utilize the multiple CPU cores of each node in the cluster. After processing, extracted features are then imported back into BioSig (Han et al. 2010; Monti et al. 2003), an imaging bioinformatics system, for (1) quality control by overlaying original image with the representation (e.g., nuclear segmentation) and (2) further bioinformatics analysis.

IV. The feature representation summarization is performed through procedural programming. BioSig is built upon PostgreSQL (PG), which enables the transparent transformation of SQL queries, through its server programming interface (SPI), for high performance applications. As a critical component, it increases the flexibility for feature manipulation, and bioinformatics analysis, which directly leads to an increased productivity by saving the reprocessing step when alternative representations need to be tested. For a specific tumor type (GBM in our case), each computed feature (e.g., nuclear size, cellularity, texture) is normalized through the following four steps before further analysis is conducted (e.g., subtyping, genomic association): (i) construct a density distribution for each feature per tissue; (ii) construct a global distribution for each feature per tumor type by combining all the feature distributions per tissue within the same tumor type; (iii) Re-bin the global distribution so that each bin has a similar population of cells of a given feature value; and (iv) Re-map the local density distributions to computed global bins of equal weight. All the steps above are implemented through SPI, and it enables the comparison of morphometric factor, in context, through its distribution function, which further improves the computational efficiency by avoiding classical clustering operations on an extremely large number of cells in the tissue section. When there are multiple tissue sections per patient, an average distribution is computed and archived. Furthermore, all of the data, for each tissue section, are downloadable and is visualizable.

6 CONCLUSION

This chapter introduced the concept of tumor heterogeneity in the context of nuclear morphology and organization. However, nuclear morphology is a complex segmentation problem subject to batch effect and biological variations. We proposed a dictionary based approach that captures intrinsic diversities in tumor signature for nuclear segmentation. Consequently, subtyping based on cellularity (e.g., rate of proliferation) was shown to be one of the better morphometric indices that correlate with the outcome. By performing subtyping, at the block level instead of a cohort of while slide images, we were able to identify four clusters that ultimately form the basis of representation for each histology section. Additionally, the molecular correlates of morphometric subtypes revealed molecular markers that are consistent with the literature for targeted therapy. The end result of the realization of one of the concepts in pathway pathology for revealing outcome based on computed morphometric indices and molecular correlates of computed indices.

References

Al-Kofahi, Y., Lassoued, W., Lee, W., and Roysam, B. (2010). Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Trans. Biomed. Eng.* **57**(4):841–852.

Axelrod, D., Miller, N., Lickley, H., Qian, J., Christens-Barry, W., Yuan, Y., Fu, Y., and Chapman, J. (2008). Effect of quantitative nuclear features on recurrence of ductal carcinoma in situ (dcis) of breast. *Cancer Inform.* **4**:99–109.

Ballaro, B., Florena, A., Franco, V., Tegolo, D., Tripodo, C., and Valenti, C. (2008). An automated image analysis methodology for classifying megakaryocytes in chronic myeloproliferative disorders. *Med. Image Anal.* **12**:703–712.

Bhagavatula, R., Fickus, M., Kelly, W., Guo, C., Ozolek, J., Castro, C., and Kovacevic, J. (2010). Automatic identification and delineation of germ layer components in H&E stained images of teratomas derived from human and nonhuman primate embryonic stem cells. In *ISBI*, pp. 1041–1044.

Boykov, Y., and Kolmogorov, V. (2003). Computing geodesics and minimal surfaces via graph cuts. In *Proceedings of IEEE ICCV*, vol. 1, pp. 26–33.

Boykov, Y., and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. PAMI* **26**(9):1124–1137.

Breiman, L. (2001). Random forests. *Mach. Learn.* **45**(1):5–32.

Bunyak, F., Hafiane, A., and Palanippan, K. (2011). Histopathology tissue segmentation by combining fuzzy clustering with multiphase vector level set. *Adv. Exp. Med. Biol.* **696**:413–424.

Chang, H., and Parvin, B. (2010). Multiphase level set for automated delineation of membranebound macromolecules. In *ISBI*, pp. 165–168.

Chang, H., Defilippis, R., Tlsty, T., and Parvin, B. (2009). Graphical methods for quantifying macromolecules through bright field imaging. *Bioinformatics* **25**(8):1070–1075.

Chang, H., Fontenay, G., Han, J., Cong, G., Baehner, F., Gray, J., Spellman, P., and Parvin, B. (2011-a). Morphometric analysis of TCGA Glioblastoma Multiforme. *J. BMC Bioinform.* **12**(1).

Chang, H., Han, J., Borowsky, A., Loss, L., Gray, J., Spellman, P., and Parvin, B. (2013-b). Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association. *IEEE Trans Med Imaging.*

Chen, T., Hinton, D., Sippy, B., and Hoffman, F. (1997). Soluble TNF-alpha receptors are constitutively shed and down regulate adhesion molecule expression in malignant gliomas. *Neuropathology* **56**:541–550.

Cook, W. J., Cunningham, W. H., Pulleyblank, W. R., and Schrijver, A. (1998). *Combinatorial Optimization*. John Wiley & Sons.

Cosatto, E., Miller, M., Graf, H., and Meyer, J. (2008). Grading nuclear pleomorphism on histological micrographs. In *International Conference on Pattern Recognition*, pp. 1–4.

Dalal, N., and Triggs, B. (2005). Histograms of oriented gradient for human detection. In *Proc. of CVPR*, pp. 886–893. Berlin: Springer.

Dalton, L., Pinder, S., Elston, C., Ellis, I., Page, D., Dupont, W., and Blamey, R. (2000). Histological grading of breast cancer: linkage of patient outcome with level of pathologist agreements. *Mod. Pathol.* **13**(7):730–735.

Datar, M., Padfield, D., and Cline, H. (2008). Color and texture based segmentation of molecular pathology images using HSOMs. In *ISBI*, pp. 292–295.

Demir, C., and Yener, B. (2009). Automated cancer diagnosis based on histopathological images: a systematic survey.

Doyle, S., Agner, S., Madabhushi, A., Feldman, M., and Tomaszewski, J. (2008). Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *ISBI*, pp. 496–499.

Doyle, S., Feldman, M., Tomaszewski, J., Shih, N., and Madabhushu, A. (2011). Cascaded multiclass pairwise classifier (CASCAMPA) for normal, cancerous, and cancer confounder classes in prostate histology. In *ISBI*, pp. 715–718.

Fatakdawala, H., Xu, J., Basavanhally, A., Bhanot, G., Ganesan, S., Feldman, F., Tomaszewski, J., and Madabhushi, A. (2010). Expectation-maximization-driven geodesic active contours with overlap resolution (EMaGACOR): application to lymphocyte segmentation on breast cancer histopathology. *IEEE Trans. Biomed. Eng.* **57**(7):1676–1690.

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

Fox, J. (2002). Cox proportional-hazard regression for survival data. In (R. Fox (Ed.), *An R and S-PLUS Companion to Applied Regression*. Thousand Oaks, CA: Sage.

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

Glotsos, D., Spyridonos, P., Cavouras, D., Ravazoula, P., Dadioti, P., and Nikiforidis, G. (2004). Automated segmentation of routinely hematoxylin-eosin stained microscopic images by combining support vector machine, clustering, and active contour models. *Anal. Quant. Cytol. Histol.* **26**(6):331–340.

Goldberg, A. V., and Tarjan, R. E. (1988). A new approach to maximum-flow problem. *J. Assoc. Comput. Mach.* **35**(4):921–940.

Goldberg, I., Allan, C., Burel, J., Creager, A., Falconi, H., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P., and Swedlow, J. (2005). The Open Microscopy Environment (OME) data model and xml files: open tools for informatics and quantitative analysis in biological images. *Genome Biol.* **6**(5):R4.

Gurcan, M., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., and Bulent, Y. (2009). Histopathological image analysis: a review. *IEEE Trans. Biomed. Eng.* **2**:147–171.

Han, J., Chang, H., Andrarwewa, K., Yaswen, P., Barcellos-Hoff, M., and Parvin, B. (2010). Multidimensional profiling of cell surface proteins and nuclear markers. *IEEE Trans. Comput. Biol. Bioinform.* **7**(1):80–90.

Han, J., Chang, H., Loss, L., Zhang, K., Baehner, F., Gray, J., Spellman, P., and Parvin, B. (2011). Comparison of sparse coding and kernel methods for histopathological classification of glioblastoma multiforme. In *Proc. ISBI*, pp. 711–714.

Hinton, G. (2006). Reducing the dimensionality of data with neural networks. *Science* **313**:504–507.

Hochberg, Y. B. Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**:289–300.

Huang, C., Veillard, A., Lomeine, N., Racoceanu, D., and Roux, L. (2011). Time efficient sparse analysis of histopathological whole slide images. *Comput. Med. Imaging Graph.* **35**(7–8):579–591.

Kakinuma, Y., Hama, H., Syugiyama, F., Goto, K., Murakami, K., and Fukamizu, A. (1997). Antiapoptotic action of angiotensin fragments to neuronal cells from angiotensinogen knock-out mice. *Neurosci. Lett.* **232**:167–170.

Kazanietz, M. (2010). *Protein Kinase C in Cancer Signaling and Therapy*. Humana Press.

Kong, J., Cooper, L., Sharma, A., Kurk, T., Brat, D., and Saltz, J. (2010). Texture based image recognition in microscopy images of diffuse gliomas with multi-class gentle boosting mechanism. In *ICASSAP*, pp. 457–460.

Kong, H., Gurcan, M., and Belkacem-Boussaid, K. (2011). Partitioning histopathological images: an integrated framework for supervised color-texture segmentation and cell splitting. *IEEE Trans. Med. Imaging* **30**(9):1661–1677.

Kothari, S., Phan, J. H., Moffitt, R. A., Stokes, T. H., Hassberger, S. E., Chaudry, Q., Young, A. N., and Wang, M. D. (2011). Automatic batch-invariant color segmentation of histological cancer images. In *ISBI, IEEE*, pp. 657–660.

Kothari, S., Phan, J., Osunkoya, A., and Wang, M. (2012). Biological interpretation of morphological patterns in histopathological whole slide images. In *ACM Conference on Bioinformatics, Computational Biology and Biomedicine.*

Land, W., McKee, D., Zhukov, T., Song, D., and Qian, W. (2008). A kernelized fuzzy support vector machine CAD system for the diagnostic of lung cancer from tissue. *Int. J. Funct. Inform. Personal. Med.* **1**(1):26–52.

Latson, L., Sebek, N., and Powell, K. (2003). Automated cell nuclear segmentation in color images of hematoxylin and eosin-stained breast biopsy. *Anal. Quant. Cytol. Histol.* **26**(6):321–331.

Le, Q., Han, J., Gray, J., Spellman, P., Borowsky, A., and Parvin, B. (2012). Learning invariant features from tumor signature. In *ISBI*, pp. 302–305.

Liu, Y., Li, C., and Lin, J. (2010). Stat3 as a therapeutic target for glioblastoma. *Anticancer Agents Med. Chem.* **10**(7):512–519.

Lowe, D. (1999). Distinctive image features from local scale-invariant features. In *ICCV*, pp. 1150–1157.

Martin, P., and JHussanini, I. (2005). PKC eta as a therapeutic target in glioblastoma multiforme. *Expert Opin. Ther. Targets* **9**(2):299–313.

Miller, J. C., and Fish, N. E. (2001). In situ duct carcinoma of the breast: clinical and histopathologic factors and association with recurrent carcinoma. *Breast J.* **7**:292–302.

Mommers, E., Poulin, N., Sangulin, J., Meiher, C., Baak, J., and van Diest, P. (2001). Nuclear cytometric changes in breast carcinogenesis. *J. Pathol.* **193**(1):33–39.

Monaco, J., Hipp, J., Lucas, D., Smith, S., Balis, U., and Madabhushi, A. (2012). Image segmentation with implicit color standardization using spatially constrained expectation maximization: detection of nuclei. In *Medical Image Computing and Computed-assisted Intervention-MICCAI*, pp. 365–372.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering – a resampling-based method for class discovery and visualization of gene expression microarray data. In *Mach. Learn. Funct. Genomics Special Issue*, pp. 91–118.

Nath, S., Palaniappan, K., and Bunyak, F. (2006). Cell segmentation using coupled level sets and graph-vertex. In *Medical Image Computing and Computed-assisted Intervention-MICCAI*, pp. 101–108.

Parvin, B., Cong, G., Fontenay, G., Taylor, J., Henshall, R., and Barcellos-Hoff, M. (2000). Biosig: a bioinformatic system for studying the mechanism of inter-cell signaling. In *IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pp. 281–288.

Parvin, B., Yang, Q., Han, J., Chang, H., Rydberg, B., and Barcellos-Hoff, M. H. (2007). Iterative voting for inference of structural saliency and characterization of subcellular events. *IEEE Trans. Image Process.* **16**(3):615–623.

Petushi, S., Garcia, F., Haber, M., Katsinis, C., and Tozeren, A. (2006). Large-scale computations on histology images reveal grade-differentiation parameters for breast cancer. *BMC Med. Imaging* **6**(14):1070–1075.

Phukpattaranont, P., and Boonyaphiphat, P. (2007). Color based segmentation of nuclear stained breast cancer cell images. *ECTI Trans. Electr. Eng. Commun.* **5**(2):158–164.

Pierallini, A., Bonamini, M., Pantano, P., Palmeggiani, F., Raguso, M., Osti, M., Anaveri, G., and Bozzao, L. (1998). Radiological assessment of necrosis in glioblastoma: variability and prognostic value. *Neuroradiology* **40**(3):150–153.

R.V. et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**(1):98–110.

Rabinovich, A., Agarwal, S., Laris, C., Price, J. H., and Belongie, S. (2003). Unsupervised color decomposition of histologically stained tissue samples. In *NIPS*, pp. 667–674.

Rahman, S., Harbor, P., Chernova, O., Barnett, G., Vogelbaum, M., and Haque, S. (2002). Inhibition of constitutively active Stat3 suppresses proliferation and induces apoptosis in glioblastoma multiforme cells. *Oncogene* **21**(55):8404–8413.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**:53–65.

Ruifork, A., and Johnston, D. (2001). Quantification of histochemical staining by color decomposition. *Anal. Quant. Cytol. Histol.* **23**(4):291–299.

Santalo, L. A. (1979). *Integral Geometry and Geometric Probability*. Addison-Wesley.

Stupp, R., Gander, M., Leyvraz, S., and Newland, E. (2001). Current and future development in the use of temozolomide for the treatment of brain tumours. *Lancet Oncol.* **2**:552–560.

Tomasi, C. Estimating Gaussian Mixture Densities with EM—A Tutorial, www.cs.duke.edu/courses/spring04/cps196.1/handouts/EM/tomasiEM.pdf.

Veltri, R., Khan, M., Miller, M., Epstein, J., Mangold, L., Walsh, P., and Partin, A. (2004). Ability to predict metastasis based on pathology findings and alterations in nuclear structure of normal appearing and cancer peripheral zone epithelium in the prostate. *Clin. Cancer Res.* **10**:3465–3473.

Verhest, A., Kiss, R., d'Olne, D., Larsimont, D., Salman, I., de Launoit, Y., Fourneau, C., Pastells, J., and Pector, J. (1990). Characterization of human colorectal mucosa, polyps, and cancers by means of computerized mophonuclear image analysis. *Cancer* **65**:2047–2054.

Wen, Q., Chang, H., and Parvin, B. (2009). A Delaunay triangulation approach for segmenting clumps of nuclei. In *ISBI*, pp. 9–12.

Zhang, L., Conejo-Garcia, J., Katsaros, P., Gimotty, P., Massobrio, M., Regnani, G., Makrigiannakis, A., Gray, H., Schlienger, K., Liebman, M., Rubin, S., and Coukos, G. (2003). Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. *N. Engl. J. Med.* **348**(3):203–213.