

Automatic Tumor-Stroma Separation in Fluorescence TMAs Enables the Quantitative High-Throughput Analysis of Multiple Cancer Biomarkers

Bernd Lahrmann^{1,2}, Niels Halama^{2,3}, Hans-Peter Sinn⁴, Peter Schirmacher^{2,4}, Dirk Jaeger³, Niels Grabe^{1,2*}

1 Institute of Medical Biometry and Informatics, University Hospital Heidelberg, Heidelberg, Germany, **2** Hamamatsu Tissue Imaging and Analysis Center (TIGA), BIOQUANT, Heidelberg, Germany, **3** National Center for Tumor Diseases, University Hospital Heidelberg, Heidelberg, Germany, **4** Department of Pathology, University Hospital Heidelberg, Heidelberg, Germany

Abstract

The upcoming quantification and automation in biomarker based histological tumor evaluation will require computational methods capable of automatically identifying tumor areas and differentiating them from the stroma. As no single generally applicable tumor biomarker is available, pathology routinely uses morphological criteria as a spatial reference system. We here present and evaluate a method capable of performing the classification in immunofluorescence histological slides solely using a DAPI background stain. Due to the restriction to a single color channel this is inherently challenging. We formed cell graphs based on the topological distribution of the tissue cell nuclei and extracted the corresponding graph features. By using topological, morphological and intensity based features we could systematically quantify and compare the discrimination capability individual features contribute to the overall algorithm. We here show that when classifying fluorescence tissue slides in the DAPI channel, morphological and intensity based features clearly outpace topological ones which have been used exclusively in related previous approaches. We assembled the 15 best features to train a support vector machine based on Keratin stained tumor areas. On a test set of TMAs with 210 cores of triple negative breast cancers our classifier was able to distinguish between tumor and stroma tissue with a total overall accuracy of 88%. Our method yields first results on the discrimination capability of features groups which is essential for an automated tumor diagnostics. Also, it provides an objective spatial reference system for the multiplex analysis of biomarkers in fluorescence immunohistochemistry.

Citation: Lahrmann B, Halama N, Sinn H-P, Schirmacher P, Jaeger D, et al. (2011) Automatic Tumor-Stroma Separation in Fluorescence TMAs Enables the Quantitative High-Throughput Analysis of Multiple Cancer Biomarkers. *PLoS ONE* 6(12): e28048. doi:10.1371/journal.pone.0028048

Editor: Pierre Busson, Institute of Cancerology Gustave Roussy, France

Received: September 21, 2011; **Accepted:** October 31, 2011; **Published:** December 2, 2011

Copyright: © 2011 Lahrmann et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Funding was provided by the German Ministry for Research and Education (BMBF) in their MEDSYS and FORSYS funding programs. Grant Numbers: 0315401B (MEDSYS), 0315263 (FORSYS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: niels.grabe@bioquant.uni-heidelberg.de

Introduction

Automation in immunohistological image processing is currently an essential technological development taking place in the clinical hunt for objective biomarkers in research and diagnostics. In cancer research one of the most important but also extreme challenges is the development of methods for the automatic separation of tumor and stroma tissue [1,2]. Success here will have an enormous impact on the applicability of biomarkers in routine cancer diagnostics and therapy as well the large-scale generation of histological tissue data for research purposes. An important method routinely used in this context which we here use to illustrate the problem is the Tissue Microarray (TMA) technology, introduced in 1998 [3]. TMAs allow the simultaneous immunohistochemical analysis of several hundred tissues on a single slide [4–6]. But as generally in all fields of pathology, the manual visual scoring of TMAs is routinely based on the quantitative analysis of protein levels by pathologists or other experts is subjective, labor-intensive, is time consuming and most importantly suffers from

intra and inter-observer variability [7]. As a solution, fluorescent capable microscopic whole-slide scanners have become available recently but are still only rarely used although they will have a key role in transforming histological evaluation into objectivity. Fluorescence based staining here is essential as it overcomes the key problem of brightfield stains by the objective and automatic capturing of distinct biomarker signals [8]. Although fluorescence helps in the quantification of individual cells, it does not per se help in differentiating tumor and stroma. In fluorescence tissue slides are frequently counterstained with DAPI (4',6-diamidino-2-phenylindole) taking the role of a conventional background stain. This makes the tumor-stroma separation more complex as the primary visual information of the tissue structure is much harder to recognize in the DAPI channel than in chromogenic histology. A histological biomarker which would exclusively stain tumor tissue is not available. Instead heterogeneity of spatial protein expression patterns is inherent to cancer. An excellent example here are the aggressive triple negative breast cancer tissues which do not express the genes for the most valuable prognostic marker

like the estrogen receptor (ER), the progesterone marker (PR) and the human epidermal growth factor receptor type 2 (Her2) [9]. The absence of the expression patterns of these biomarkers disallows using any single one of them as a reference protein biomarker and renders it essential to separate the cancerous from the healthy/connective tissue by the help of objective, standardized processing algorithms based on morphological criteria. Thus, pathological evaluation routinely uses morphological criteria as a spatial reference system to determine the tumor area in cancer histology. We conclude that combining the advantages of fluorescence with automatic image acquisition and processing requires the development of algorithms for tumor-stroma separation solely from a DAPI background stain being frequently used in immunofluorescence.

Therefore, we here set out to develop such an automatic algorithm based only on the DAPI channel (Figure 1B–D). Several methods for the separation of cancerous tissue from other tissue types by morphological criteria are available in the literature. Amaral et al. [10,11] present two different methods whereby color features are used for the classification of whole TMA-cores. In [12] textural features help to separate different tissue regions on a TMA and in [13] textural features are used for the detection of pathologic regions in histological slides. But all these methods work

on chromogenic stained tissue samples where for the classification of the different tissue types the information of all 3 RGB channels was obtainable. Classifying tumor tissue only in the DAPI channel forces us to deal with less information available for the classification step compared to the previous other approaches. Only few publications deal with the classification of fluorescently stained tissues. In [14] the authors use nuclear features obtained from the DAPI-channel to distinguish whether the whole tissue is cancerous or healthy instead of classifying the different types present on the tissue. Most of the published work in biomarker research use two biomarkers for co-localization or manually segment the cancerous tissue instead of an automated manner [15–18].

Gunduz et al. [19] published a novel method for the classification of chromogenic stained brain tissue samples. They formed cell graphs based on the topological distribution of the tissue cells and extracted the corresponding graph metrics to train a classifier. The classifier was able to distinguish between cancerous and healthy tissue. A graph here is an abstract representation of objects (nodes) where pairs of these objects are linked by edges. The method was further developed in [20] and [21]. Bilgin et al. [22–23] demonstrated that they successfully analyzed breast and bone tissue samples by the help of cell graphs.

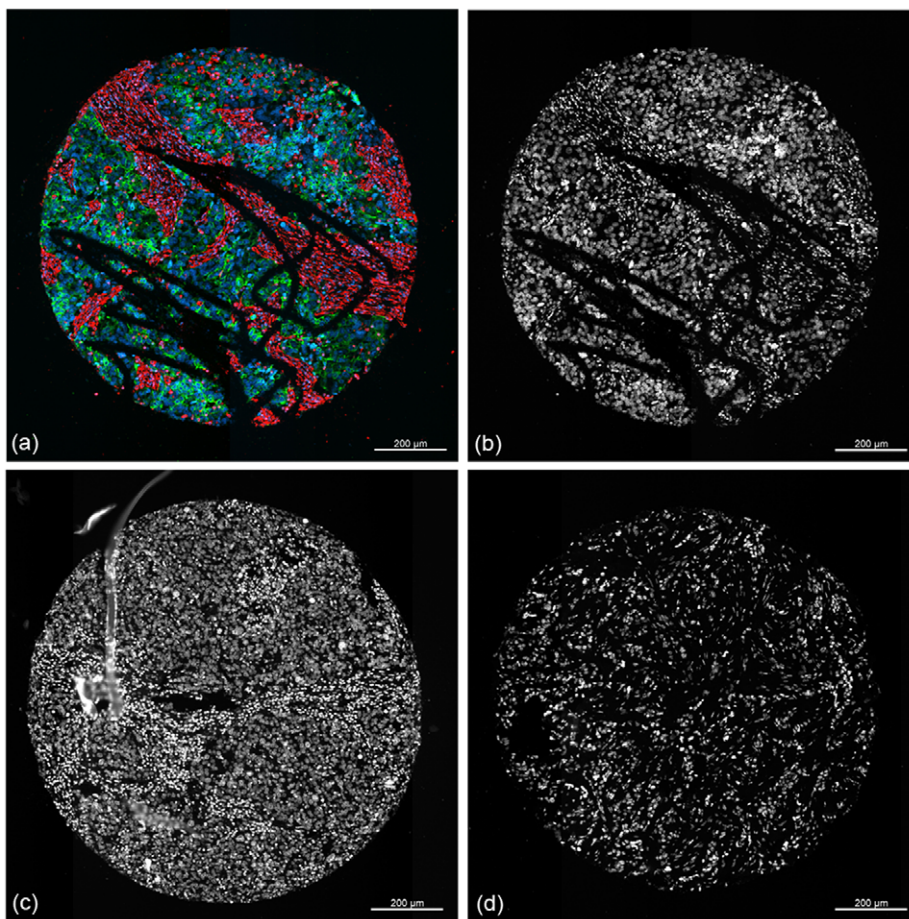


Figure 1. Microscopic image examples of different TMA-cores. (a) Representation of all 3 channel of a fluorescently stained core in RGB colorspace. Glyphs originated due to TMA's preparation. Red representing the stromal marker (Vimentin), green the tumor marker (CK19) and blue the DAPI channel highlighting the cell nuclei; (b) the DAPI channel of (a) as an intensity image: in general tumor cells are darker and tighter connected than stromal cells; (c) another DAPI image of a core with a high density of cells; (d) an example of a core with a lower density of cells **shows** the high heterogeneity among the cores. doi:10.1371/journal.pone.0028048.g001

They evaluated their method on hand-selected and non-biomarker characterized breast cancer samples.

Here we further developed this approach by developing a novel method capable of classifying fluorescently stained Tissue Microarrays. Our method uses cell graphs based on three different categories of features reflecting the properties of the cells contained in the graph (nodes) and their similarity (edges). From a potential set of features we determine those which are best capable of separating tumor and stroma tissue. Clearly, performing an accurate tumor-stroma separation is already a challenging task. Using furthermore only the DAPI-channel for this task requires an even higher performance in segmentation and classification.

As the first step we performed watershed segmentation and then we built cell graphs by linking the segmented cell nuclei under each other. The linking of the cells is based on new rules especially adapted for fluorescently stained TMAs which can consist of several different tissues types. Instead of using only topological graph metrics for the cell-graph classification, we also determine the morphological and intensity based cell features of each cell-graph. By combining all three feature types we were able to obtain a successful tissue classifier for fluorescent histological slides.

We demonstrate our method on 180 core images of TMAs from invasive triple negative breast cancer biopsies containing cancerous tissue as well as stroma (connective tissue). Our method was able to separate tumor and connective tissue that coexist on the same tissue core with a total overall accuracy of 88.80(\pm 07.73) %.

Materials and Methods

Tissue samples

The data set total consists of 210 tissue microarray core images of invasive triple negative breast cancer biopsies obtained from 6 TMAs. The tissue was obtained from the tissue bank of the National Center for Tumor Diseases (NCT) at the University Hospital Heidelberg. Obtaining tissue samples was approved by the ethics committee of the Medical Faculty Heidelberg. According to the official regulations of the University's Tissue Bank determined by said ethics committee no individual consent has to be obtained from individual patients for patient samples older than 3 years. Documentation of all procedures are handled in an ISO certified process by the NCT tissue bank. Each TMA contains two cores of 1 mm diameter from 42 different patients (total 84 cores per TMA). One core is obtained from the periphery of the tumor and the other is obtained from the tumor's center. We excluded cores from our data set if their area was below fifty percent of a regular core or if unusable. Each image is taken in a 20 fold magnification and has an average size of 2800 \times 2900 pixels. All TMAs are stained with 3 fluorescent dyes. Every TMA was stained with DAPI highlighting the cell nuclei as a

counterstain The other used antibodies (Vimentin, CK19 and CK5/6) were conjugated with Alexa Fluor[®] 488 (FITC alternative, green fluorescent dye) or Alexa Fluor[®] 594 (red dye). Figure 1A illustrates a tissue core stained with 2 different biomarkers and DAPI as counterstain. Figure 1B–D illustrates further representative examples of the DAPI channel of three different tissue-cores.

Image acquisition

Fluorescently stained TMAs were automatically imaged using the Nanozoomer HT Scan System (Hamamatsu Photonics, Hamamatsu Japan) capable of scanning whole slides. Glass slides were scanned at 20 fold magnification (resolution of 0.46 μ m/pixel). For the scanning of the glass slides, the slide scanner automatically detects the region of interest that contains the array of cores and also determines automatically a valid focal plane for scanning. The resulting virtual slides had an averaged file size of 5 GB. Single core images with an average size of 2800 \times 2900 pixels were located and extracted from the TMAs using template matching [24].

General image analysis workflow

The key concept in this manuscript is the cell graph which we use to capture the topological cell distribution in tissues as well as the spatially related local cell features for classification. The major steps in this approach are the segmentation of the cell nuclei in the DAPI channel using watershed segmentation, the construction of the cell graphs, extracting the topological and local cell features from these graphs and use them to train a classifier. Image processing algorithms were developed using Matlab[™] (Mathworks, Natick, Mass., USA) with the image processing toolbox.

Our image analysis pipeline contains the following conceptual steps (as illustrated in Figure 2):

2.1 Pre-processing: We first applied several image enhancement methods to prepare the image for the subsequent segmentation step.

2.2 Cell segmentation: A Watershed-Transformation was applied for the cell nuclei segmentation.

2.3 Cell graph generation: Based on the segmented nuclei we generated cell graphs which represent the topological distribution of the nuclei on the tissue cores. We calculated several features for every (sub-) graph and also calculated intensity and morphological base features for every single nucleus on a core.

2.4 Classification and feature selection: A Support Vector Machine was trained for classification step and the F-Score was calculated for feature selection.

2.1 Preprocessing

In this first step, we applied several methods to enhance the quality of the core image for the subsequent classification. We

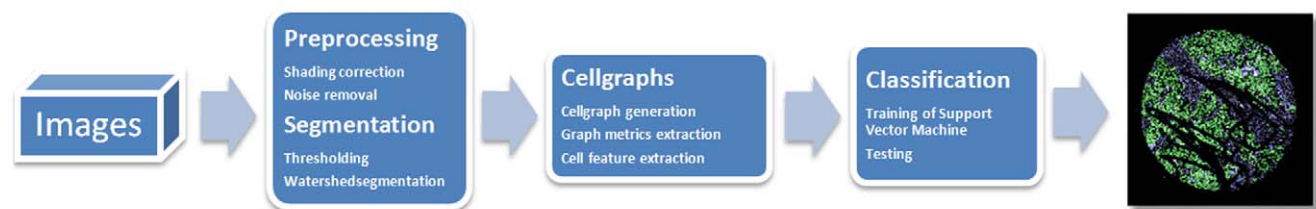


Figure 2. A Flowchart showing the single steps of our methodology. After obtaining the images, pre-processing steps enhance the image quality and watershed segmentation for the subsequent segmentation is applied. Accordingly the cell graphs are generated and features are computerized. The last step uses a SVM to classify the graphs as either tumor or stroma. doi:10.1371/journal.pone.0028048.g002

started to remove shading artefacts, which delineate the result of various optical phenomena such as lens vignetting or photo bleaching. Shading artefacts in fluorescence imaging can also be caused due to auto fluorescence of the samples or the mounting medium. Shading correction (flat field compensation) was used to compensate for lens vignetting as well as for inhomogeneity in the illumination. Shading correction was achieved by performing a black balance calibration using clear background areas. The next step in the image processing pipeline was the removal of noise and other small particles which were not suitable for later analysis. To exclude unspecific and diffuse background staining all pixels with intensity levels below a threshold of 25 were set to zero. A Median-Filter with a 3×3 kernel was used to smooth the image. The resulting image was converted into a binary image (by the use of Otsu's method [25]) in which objects with an area smaller than 150px (smaller than the size of connective tissue nucleus) are removed. Objects outside the regular core shape were removed using morphological operations like closing or opening combined with area filter. Eventually, isolated nuclei were observed inside the core. We assumed that these isolated nuclei belong to non tumorous cells and were thus excluded from the tumor tissue. To this end, we determined the smallest bounding box of the objects and expanded it by 20px in each direction. Based on this new coordinates, an image was cropped from the original binary image and the present objects in this

image were counted. If only one object was present, the object was removed while the presence of more than one object implies contact to other cells and the object remained. Furthermore, in several cores we discovered large overstained areas with maximum intensity levels. These areas, which could be caused by agglomerated connective tissue cell nuclei at the TMA preparation or to high exposure times, are not suitable for further analysis and were removed. Figure 3B shows the results of the preprocessing step.

2.2 Cell Segmentation

Automated cell segmentation in fluorescently stained TMA can be problematic for reasons that include cell overlapping or clustered cells, complex tissue structure, debris and uneven background intensity due to auto fluorescence. Another difficulty is intensity variation between the nuclei which can lead to over-segmentation of the cell nuclei. Due to these intensity variations among nuclei, we first divided the image into one image representing only objects with a brighter illumination and one representing the darker objects. We then applied the segmentation step separately on both of these images. This separation was done by calculating a threshold based on Otsu's method [25] ignoring background pixels. A segmentation algorithm that has proven to be very useful for many nuclei or cell segmentation cases is the watershed segmentation [26–28]. We applied seeded watershed

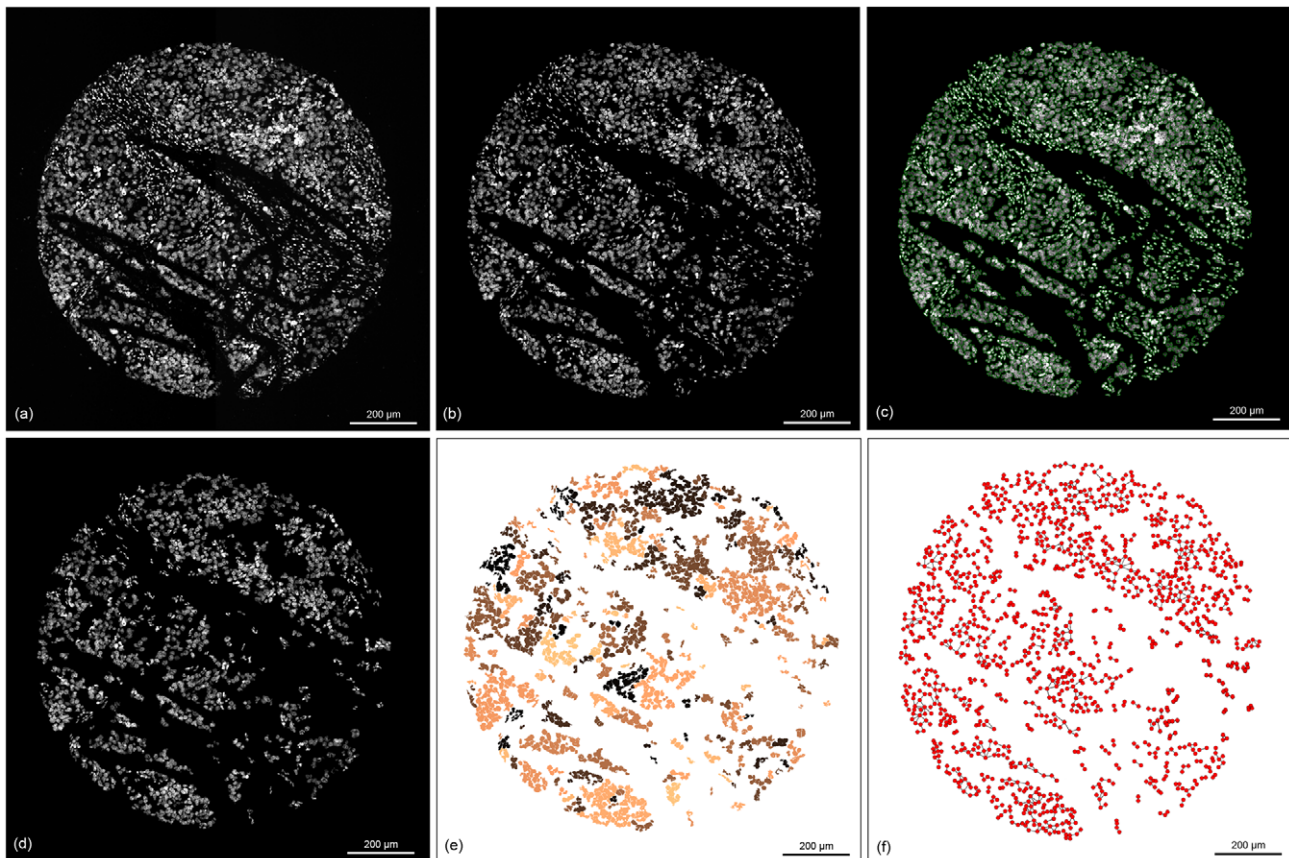


Figure 3. The different image processing steps and the graph generation steps. (a) original image of the DAPI-channel; (b) image after shading correction and noise removal; (c) result of the watershed segmentation, the segmented cells are highlighted by green contour; (d) the image after removal of single cells; (e) showing the cells which were connected via the graph generation step in the same color (cells marked with the same color belong to the same sub-graph); (f) cell graph representation of the cells. The red dots are the nodes which represent the cells, the black lines are the edges between them.

doi:10.1371/journal.pone.0028048.g003

segmentation for the segmentation. Seeded watershed segmentation means, that starting regions, which are called seeds, are given as input to the watershed segmentation. We set the seeds in an automated way using the h-maxima transform [29]. The result of this segmentation step is shown in Figure 3C.

2.3 Cell graph Generation

A graph is denoted as a set of objects where some pairs of objects are connected by links. The connected objects are represented by mathematical abstractions called nodes (also called vertices), and the links that connect some pairs of nodes are called edges. Formally, a graph is an ordered pair $G = (V, E)$ where V is the set of nodes and E the set of edges linking the nodes of V . In our work, each of the former segmented cell nuclei was used as a node. Figure 4 shows a conceptual representation of cell graphs.

Different approaches are presented in literature for generating cells graphs, which represent the topological behaviour of tissues or cells in different scientific questions [19], [21–23], [30]. In [19] Gunduz et al. make use of the Waxman model for the cell graph generation. Bilgin et al. [22] and Gunduz et al. [21] use a probability function for linking the cells among themselves. In their approaches the probability of linking cells decays with a growing Euclidian distance between the cells centroids. In [23], [30] cells are simply linked if the Euclidian distance between their centroids is below a specific distance. Tumor cells generally appear in clusters, accordingly they can be expected in a marginal distance of another or appearingly “touching” each other. Thereby, this “touching” of nuclei occurs because of the three dimensional structure of the histological sections. By using the nuclei centroids for distance measurements alone it is possible that cells get linked although they are more apart than typical tumor cells. In our case, we are performing a pre-classification by only building links between cell nuclei touching each other and thereby excluding solitary cells (of connective tissue origin) from the graph construction step. In our method we test if cells touch each other by the following steps. We first convert the result of the watershed segmentation into a binary image and then we dilate each of the segmented cell nuclei separately. The dilation of an (cell nucleus)-Image I with a structuring element S , denoted as $I \oplus S$, is defined as the set operation $I \oplus S = \{z | (\hat{S})_z \cap I \neq \emptyset\}$ where \hat{S} denotes the symmetric structuring element. We chose a diamond-shaped structuring element with a distance from the origin of 2. We then determine, if the cell nuclei were in very close contact (“touching” appearance) and set a link between them, if their pixel intersection was not an empty set after the dilation step:

$$E(I, J) = \begin{cases} 1 & \{x | x \in I \wedge x \in J\} \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where I and J are the particular images of two neighboring cell nuclei. In tissues, tumor cells are eventually tightly surrounded by connective tissue cells which could, after applying the above described distance rule lead to structural errors in the cell graph. Usually, the tumor cells are appearing with lower intensity levels than the connective tissue cells. Hence we link only cells, if the difference between their intensity levels is lower than a specific threshold. This threshold is dependent on variations in staining and fluorescence signal acquisition efficiency. We here empirically determined a difference of 30 intensity values as an applicable threshold for our data set. Concluding, neighboring cell nuclei with an intensity difference below this threshold are linked:

$$E(I, J) = \begin{cases} 1 & \left| \overline{AM}_I - \overline{AM}_J \right| \leq 30 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Where \overline{AM} is the arithmetic mean of the cell image intensity level, X the number of rows, Y the number of columns and $S = X * Y$. Summing up, setting a link between two cell nuclei in our model depends on the probability of touching each other and that the difference of their intensity levels is lower than a specific threshold. Figure 4D shows an example image whereby single cells are removed. Figure 4E highlights the cell nuclei, which were linked through this graph generation step in the same color. A visual graph representation of this step is shown in Figure 4F. Cells which were not connected during the graph generation process were treated in an additional step described in section “single cell classification”.

Cell Graph Features

After generating the cell graphs, we computed several features for the training of the classifier. In total we computed 22 features which can be divided into three different categories. The first 10 features, in literature usually called graph metrics [19], [23], capture the topological behavior of the graphs like the number of cells in a graph, the number of links between them or further topological relations among the cells (feature category T). The remaining 12 features capture morphological properties (feature category M) like area, shape as well as intensity based properties (feature category I) of the single cell nuclei of a graph and were

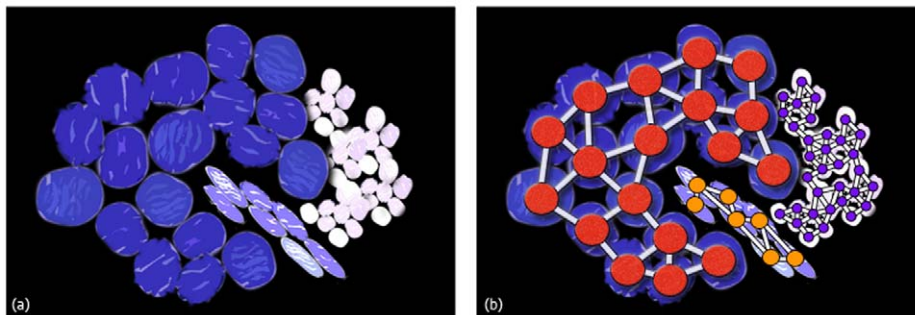


Figure 4. Conceptual representation of cell graphs. (a) Artificial sketch of 3 different cell type: tumor cells in blue, lymphocytes in white and in purple fibroblast. (b) Cell graph representation of (a). Cells are depicted as nodes and links between them represent biological relations. doi:10.1371/journal.pone.0028048.g004

chosen based on their expected suitability. The last two categories of features are first computed for each single cell nucleus and then the average is used as a feature of the corresponding graph. Keep in mind that several of these intensity based features depend on the imaging conditions like the exposure time, the concentration of the biomarker, time lag between staining and imaging due to photo bleaching and further more. It requires that those conditions remain constant across the datasets. In Table 1 the applied features and graph metrics are described in detail.

2.4 Classification and feature selection

Support vector machines (SVMs) [31] are commonly used as supervised learning methods for classification in computational biology and image processing tasks [32–34]. Starting point for the training of a SVM is a set of training data whose class membership is known:

$$D = \{(\bar{x}_i, y_i) | \bar{x}_i \in \mathbb{R}^2, y_i \in \{-1, +1\}\} \quad (3)$$

where \bar{x}_i are the feature vectors and y_i their respective class labels (tumor cells or connective tissue cells). The SVM maps these input vectors into a higher dimensional space and constructs an optimal hyper plane separating the data into two groups. By solving a quadratic programming optimization problem, the SVM calculates the normal vector \bar{w} and the bias b of the separating hyper plane $\bar{w}\bar{x} - b = 0$ which maximizes the margin between the support vectors \bar{x}_i of different classes. The width of the margin is equal to $\frac{2}{\|\bar{w}\|^2}$, thus the widest margin between the vectors is found by minimizing under the restrictions $y_i(\bar{w}\bar{x} - b) \leq 1$, requiring a

Table 1. Graph metrics used to train the classifier and their description.

Name	Description
Graph metrics (topological features)	
(1) Nr. of nodes	Defines the number of nodes in a graph.
(2) Nr. of edges	Total number of edges in a graph.
(3) Average degree	The average degree of the nodes of the graph. The degree of a node is defined as the number of its edges. It explains the number of neighbour nodes.
(4) Diameter	The eccentricity $e(u)$ of a graph node u in a connected graph G is the maximum graph distance between u and any other node v of G : $e(u) := \max\{d(u, v) v \in V\}$. The maximum eccentricity is the graph diameter.
(5) Radius	The minimum graph eccentricity is called the graph radius: $rad(G) := \min\{e(u) u \in V\}$
(6) Nr. of central points	Number of nodes that have eccentricity equal to the radius.
(7) Average clustering coefficient	Here, the average clustering coefficients of the nodes of the graphs are used as a global metric. The clustering coefficient C_i of a node v_i is given as: $C_i := \frac{2 \times E_i}{k \times (k-1)}$ Where k is the number of neighbours of node i and E_i is the number of existing edges between its neighbours.
(8) Nr. of end nodes.	The number of nodes with degree equal one.
(9) Percentage of end nodes.	The percentage of end nodes of a graph.
(10) Hop-plot exponent	The hop-plot exponent is computed by the slope of the hop-plot values as a function of h in log-log-scale. The hop-plot value reflects the size of a neighbourhood between nodes within a hop h . For hop h , the hop plot value is defined as the number of node pairs such that the path length between these pairs is less than or equal to h hops.
Averaged node features (morphological and intensity features)	
(11) Average area	The average area of the cells of the graph.
(12) Average eccentricity	The eccentricity is the ratio of the distance between the centroid of the ellipse and its major axis length. The value is between 0 and 1 (an ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment.). Eccentricity is given as: $e := 2 \times \sqrt{\frac{\left(\frac{x_{max}}{2}\right)^2 - \left(\frac{x_{min}}{2}\right)^2}{x_{min}}}$ Where x_{max} is the major axis length and x_{min} the minor axis length.
(13) Average equivalent diameter	The equivalent diameter specifies the diameter of a circle with the same area as the cell: $d_e = \frac{2 \times \sqrt{A}}{\sqrt{\pi}}$ Where A is the area of the cell.
(14) Average extent	The extent specifies the proportion of pixels in the smallest rectangle containing the cell that are also in the region. Computed as the cell area divided by the area of the smallest rectangle containing the cell.
(15) Average major axis length	Major axis length specifying the length of the major axis of the ellipse that has the same normalized second central moments as the cell.
(16) Average minor axis length	Minor axis length the length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the cell.
(17) Average max Intensity	Maximum intensity of the cell.
(18) Average min Intensity	Minimum intensity of the cell.
(19) Average mean intensity	Mean intensity of the cell.
(20) Average perimeter	Perimeter: the distance around the boundary of the cell.
(21) Average STD Intensity	Standard deviation of the intensity of a cell.
(22) Average median intensity	Median intensity level of a cell.

doi:10.1371/journal.pone.0028048.t001

separable data set. The hyper plane then is used as a sign function for the classification of each feature vector of the test set. The classification function returns either +1 if the test data is member of the class, or -1 if it is not. When perfect separation is not possible, a slack variable ξ_i is introduced for each vector \bar{x}_i . The constraints for computing the optimal hyperplane are then formulated as $y_i(\bar{w}\bar{x} - b) \leq 1 - \xi_i$ and the hyperplane can be found by minimizing:

$$\min \left[\frac{1}{2} w^2 + C \sum_{i=1}^N \xi_i \right] \tag{4}$$

where C is a cost parameter that determines the effect of outliers on the resulting hyper plane. The described SVM is capable to separate linear data. To create a classifier which is able to classify nonlinear data the kernel trick is applied. The key idea is to transform \bar{x}_i into a higher dimensional space to find a separating hyper plane using a kernel. This allows the algorithm to fit the maximum-margin hyper plane in a transformed feature space. Equation 4 can be rewritten as (5):

$$\text{minimize : } W(\alpha) = - \sum_{i=1}^N \alpha_i + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j k(x_i, x_j) \tag{5}$$

$$\text{subject to : } \sum_{i=1}^N y_i \alpha_i = 0, \forall_i : 0 \leq \alpha_i \leq C \tag{6}$$

where the α_i values are the Lagrange multipliers, which can be positive or negative, due to the equality constraints and $k(x_i, x_j)$ is the kernel function. In this article, we used a radial basis kernel (RBF) $k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$ which is also known as Gaussian kernel.

Feature Selection

We calculated the F-score for the selection of the features included in the SVM. Feature selection is a technique to find a subset of features by removing most unimportant and redundant features from the feature space. This technique generally helps to improve the total performance of the classifier, speeding up the learning process, allows a better representation of important features and results in a remaining feature set with maintained discriminatory power. The F-score measures the discrimination between two sets of features [35]. A higher F-score indicates to a higher discriminate feature than a feature with a lower F-score. We calculated the F-score for each feature i as described in (7) with the given training vectors $x_k, k = 1, \dots, m$:

$$F(i) \equiv \frac{(\bar{x}_i^t - \bar{x}_i)^2 + (\bar{x}_i^s - \bar{x}_i)^2}{\frac{1}{n_t - 1} \sum_{k=1}^{n_t} (x_{k,i}^t - \bar{x}_i^t)^2 + \frac{1}{n_s - 1} \sum_{k=1}^{n_s} (x_{k,i}^s - \bar{x}_i^s)^2} \tag{7}$$

where $\bar{x}_i^t, \bar{x}_i^s, \bar{x}_i$ are the mean values of the i th feature of the tumor, stroma, and whole data set. $x_{k,i}^t$ is denoted as the i th feature of the tumor instance and $x_{k,i}^s$ the i th feature of the stroma instance.

Single cell classification

Based on the two criteria for the cell graph generation (intensity and distance), it can occur that individual cells are not linked to any other cell. Thus, these cells are not included in the classification step and we treat them with an additional algorithm

in a separate single cell classification step. We first try to identify inflammatory cells (lymphocytes e.g.) and fibroblasts which are part of the stromal class. Usually, inflammatory cell nuclei appear as small roundish cell nuclei with a very high intensity compared to other cells on the core. Cell nuclei are therefore classified as inflammatory cells when: the cell nucleus intensity is higher than a specific level, a metric which calculates the roundness is higher than a threshold and the area is smaller than 500 pixels:

$$IC(I) := \begin{cases} 1 & \left(\overline{AM}_i = \frac{1}{S} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} i(x,y) \right) > 200 \wedge \frac{4 \times \pi \times S}{w} > 0.8 \wedge S < 500 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where \overline{AM}_i is the arithmetic mean intensity, $S = X \times Y$ the area and w the perimeter of a cell nucleus. Fibroblasts generally have an elliptical shape and were identified by:

$$F(I) := \begin{cases} 1 & 2 \times \sqrt{\frac{\left(\frac{x_{max}}{2}\right)^2 - \left(\frac{x_{min}}{2}\right)^2}{x_{min}}} > 0.90 \wedge S < 3000 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where x_{max} is the major- and x_{min} the minor-axis of the cell nuclei. These values are utilized to calculate the eccentricity of an ellipse. The eccentricity of a circle is 0 and an ellipse which eccentricity is 1 is a line segment. The remaining cell nuclei were classified by the use of a support vector machine. We used the 12 morphological and intensity based features already mentioned in section ‘‘Cell Graph features’’ to classify each single cell nucleus. We trained the SVM with the single cell nuclei of our training set and evaluated the algorithm separately as depicted in result section.

Results

The overall goal of our approach was to automatically classify each cell of a TMA core by the help of the generated cell graphs. The training and the classification is based only on the DAPI channel primarily staining the nuclei. Figure 5 illustrates the results of our approach on 4 different TMA-cores.

Cell segmentation step

The cell nuclei segmentation was evaluated on 3 randomly selected real core images obtained from one TMA. In total 5162 nuclei were used and ground truth was obtained from one expert who marked the over- and under-segmented cell nuclei. The watershed algorithm proposed here can correctly segment 94.1% (± 3.75) of the nuclei. Table 2 shows the detailed segmentation results and Figure 3C shows an example of this step whereby the segmented nuclei are surrounded by a green contour.

Feature selection

Feature selection simplifies and shortens training of a classifier, and frequently also improves its accuracy. For feature selection from 30 core image we first generated in total 7888 topologically disjoint cell graphs leading to use the same total number of feature vectors. This total set of features comprises 4065 feature vectors for the tumor class and 3823 for the stroma class. The feature values occur within largely varying numerical ranges. Therefore we normalized them to the range of [0,1] to enhance the learning progress.

We calculated the F-score (discriminative power of a feature) for each of the 22 features from Table 1 to determine the best feature set for the classification task. Based on the results shown in Table 3

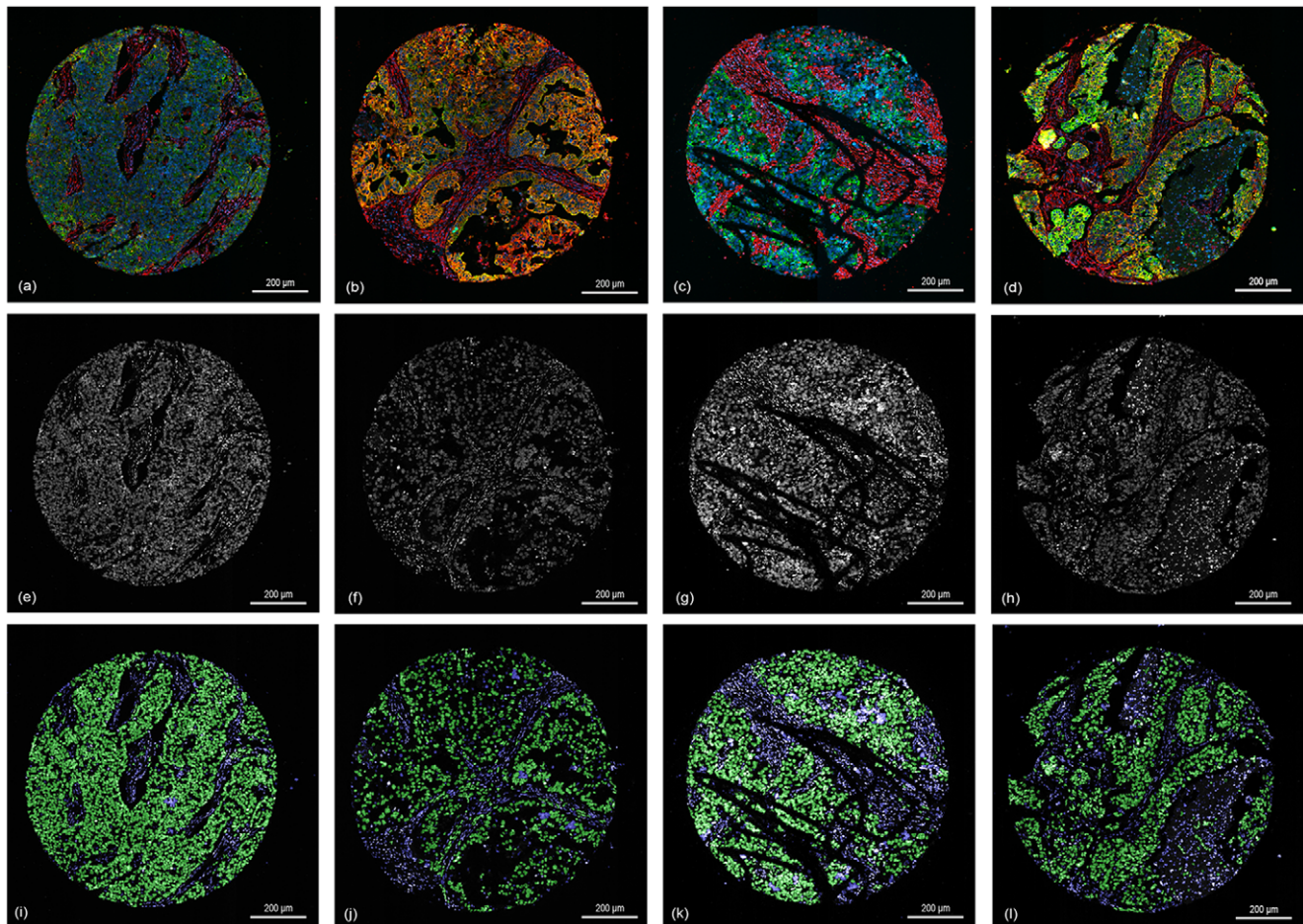


Figure 5. The results of the classification. (a–d) showing the original RGB core images; (e–h) showing the corresponding DAPI channel as an intensity image of the cores (a–d); (i–l) results of the classification step, green = cells classified as tumor cells, blue = cells classified as stroma cells. doi:10.1371/journal.pone.0028048.g005

we picked the best 15 features for the training of the support vector machine. Our results allow the comparison the discriminative power of the three studied categories (T, I, M) of features. We find that the most predictive features are the morphological and intensity based features like the standard deviation of the intensity of a cell graph, the minor axis length of the cell nuclei within a cell graph and the equivalent diameter of them. The features with the lowest performance belong to the graph metrics: the number of central points, the number of end nodes and the number of edges.

Cell-graph Classification

For the training of the SVM we used a training set of 30 cores obtained from one slide and with the above described feature set. The remaining 180 cores were used as test set and are obtained

from 5 other slides. Images used in the test set were excluded from the training set. For training the classifier we manually segmented the 30 cores into tumor and stroma according to the used biomarkers. Vimentin was used as stromal marker. Although it is partly expressed in tumor cells Keratin 5/6 and 19 allowed the distinct segmentation of the tumor. Cores where no stain for the keratins was observed were excluded from the study. We determined the total areas of both classes in the data set averaged over all cores, the area of tumor is 64.6% whereas the area of stroma is 35.4%. Table 4 shows the averaged accuracies of the training set, the 5 slides of the test set and the total averaged values of the test set. We present the overall accuracies and the producer's accuracy of the tumor and the stroma class. Averaged over the test set, we achieved a total overall accuracy of 88.80(± 0.73). For the prediction of the tumor class we achieved

Table 2. Accuracy of the watershed cell segmentation.

	Total Nuclei	Correctly segmented	Over-segmented	Under-segmented
Sum	5162	4860	272	30
Percentage	100%	94.1(± 3.75)	5.3(± 4.0)	0.6(± 0.3)

doi:10.1371/journal.pone.0028048.t002

Table 3. The F-scores of each feature in descending order.

Features	Type	F-score
Average STD Intensity	I	0.240
Average minor axis length	M	0.212
Average equivalent diameter	M	0.209
Average area	M	0.182
Average max intensity	I	0.143
Average mean intensity	I	0.108
Average perimeter	M	0.063
Average median intensity	I	0.056
Average major axis length	M	0.051
Hop-plot exponent	T	0.039
Average degree	T	0.038
Percentage of end nodes	T	0.038
Average eccentricity	T	0.038
Diameter	T	0.038
Radius	T	0.038
Average extent	M	0.038
Average min Intensity	I	0.016
Nr. of nodes	T	0.008
Average clustering coefficient	T	0.007
Nr. of edges	T	0.006
Number of end nodes	T	0.004
Number of central points	T	0.001
Average morphological features	M	0.144
Average intensity features	I	0.112
Average topological features	T	0.023

The table shows the evaluated features sorted by their decreasing value for tissue classification (F-score). For each feature it is given whether it is of morphological (M), intensity (I), and topological character (T).

doi:10.1371/journal.pone.0028048.t003

an average accuracy of $88.02(\pm 13.51)$ and for the stroma class an accuracy of $84.67(\pm 11.80)$.

As mentioned in the method section, some of the cell nuclei were not linked during the cell-graph generation step with a cell-graph because they are too far from other cells or the difference between their intensities does not allow linking them. We added an optional algorithm for the classification of these cells and present the results in Table 5. By the use of this single cell classification algorithm we achieved an overall accuracy of $86.12(\pm 07.46)$, an accuracy of $86.60(\pm 13.22)$ for the tumor class and $79.75(\pm 13.27)$ for the stroma class. The use of the single cell classification

therefore reduces the total accuracy but gives the advantage to classify each cell available on the core.

Discussion

The high throughput analysis of fluorescently stained histological slides is becoming of increasing importance. To biomedical studies it offers the great advantage of objective biomarker quantification in individual color channels. Unfortunately, it in the same time results in an increased difficulty of distinguishing tumor and stroma tissue. We therefore here presented a new method capable of classifying fluorescently stained histological slides into tumor and stroma in the DAPI channel. The restriction to the DAPI channel severely reduces the available information for the image processing steps compared to the other methods based on all RGB color channels. We demonstrate that our approach is capable of separating the tumor from other tissue kinds on the same core in triple negative breast cancer as an example of highly heterogeneously expressing tumor tissue.

Technically, we make use cell graphs which have been used previously deployed for chromogenic tissue. As we did not know which kind of features might be effective for classifying fluorescence tissue we combined different kinds of features: topological features (T), morphological (M) and intensity based features (I) of the single cell nuclei related to the particular cell graph. We performed a systematic and quantitative evaluation of the value the individual features contribute to the discrimination between tumor and stroma tissue. Our results showed that the most relevant features indeed do not belong to the previously in literature used topological metrics but belong to the class of either morphological or intensity based features. The integration of morphological and intensity based features into the classification step significantly contributed to the increase in the classifiers power to discriminate tumor and stroma. The use of a broad spectrum of feature types reflects the complex visual tasks pathology perform in their daily histological assessments.

For segmentation we used the watershed algorithm. Nuclei were linked if they were close to each other and the difference of their intensity level is in a certain threshold. The linked cell nuclei then formed the cell graphs representing the topological structure of the whole tissue core. After segmentation the support vector machine was trained using the extracted features. Our evaluation on a total test set of 180 single core images yielded an overall accuracy of $88.80(\pm 07.73)$ or combined with the optional single cell classification an overall accuracy of $86.12(\pm 07.46)$.

Compared to other methods classifying cancer tissues we achieved improved results. But the direct comparison of the obtained results with previous works is difficult due to different tissue types, preparations methods and staining protocols used in evaluations. Bilgin et al. make use of cell graphs [22] for the classification of chromogenic stained breast cancer tissue samples

Table 4. The average classification accuracies.

	Training set	Test set					Average 1-5
		1	2	3	4	5	
Overall	88.47(± 06.68)	87.65(± 08.19)	90.30(± 06.44)	88.68(± 07.19)	88.76(± 06.98)	88.59(± 09.83)	88.80(± 07.73)
Tumor	89.26(± 10.20)	87.56(± 13.29)	87.83(± 12.47)	88.00(± 17.64)	88.98(± 10.01)	87.71(± 14.13)	88.02(± 13.51)
Stroma	85.14(± 10.95)	81.19(± 11.62)	91.45(± 06.21)	82.97(± 15.12)	80.02(± 12.35)	86.90(± 13.69)	84.67(± 11.80)

The table shows the accuracies of the training set and the accuracies of the slides from the test set.

doi:10.1371/journal.pone.0028048.t004

Table 5. The accuracies of the slides of the test set with the additional single node classification.

	Test set					Average 1–5
	1	2	3	4	5	
Overall	84.02(±07.83)	87.46(±6.85)	86.12(±08.23)	86.28(±05.88)	86.69(±08.51)	86.12(±07.46)
Tumor	85.62(±13.26)	86.05(±12.57)	86.31(±17.08)	88.39(±09.77)	86.64(±13.41)	86.60(±13.22)
Stroma	75.94(±13.11)	84.58(±13.63)	80.23(±14.24)	76.29(±11.81)	81.72(±13.55)	79.75(±13.27)

doi:10.1371/journal.pone.0028048.t005

and reached an overall accuracy of 79.2% on their data set. Breast cancer TMA cores stained with chromogenic dyes were processed in [12] and classification achieved an accuracy of 80% for standard H&E stains and an accuracy of 78% of IHC stained tissues. As the direct comparison of the quantitative results obtained is nearly impossible we here performed the systematic quantitative feature comparison which clearly shows the importance of utilizing all three different feature types. From this we conclude that the challenges arising from working only with DAPI signals in a single color channel can at least well be compensated for by using multiple cell graph feature categories.

The standard deviation of the achieved accuracies in our own data shows an inferior performance on some cores of the SVM. There are several reasons for the performance-loss on such cores. Some cores have a squashed boundary due to the manufacturing of the tissue cores in which cores are pressed into preformed holes. This eventually leads to a higher concentration of cells at border regions which again could lead to a misclassification of these cells due to the wrong topological organization. Another reason for difficulties in processing individual TMAs is that there are stronger intensity alterations between the slides caused by different staining conditions or different imaging conditions such as the exposure

time. By standardized experimental protocols this could be avoided in a more routine application in future.

It is expected that the method will require further adaptation of particular parameters and a general improvement when being challenged with tissues of different topological structures. We expect that the performance of our method could decrease on tissues consisting of a greater number of appearingly isolated cells. In such a case the single cell classification based could be improved using for example texture based features. Generally the here presented method could also be deployed for separating not only stroma and tumor but other cell types like lymphocytes. Therefore, the here presented method is an important improvement towards the automated evaluation of new tumor marker in clinical research and diagnostics.

Acknowledgments

We would like to thank Claudia Ernst for assembling the tissue arrays.

Author Contributions

Conceived and designed the experiments: BL NG. Performed the experiments: BL NH NG. Analyzed the data: BL. Contributed reagents/materials/analysis tools: HPS PS DJ. Wrote the paper: BL NG.

References

- Zinchuk V, Zinchuk O, Okada T (2007) Quantitative colocalization analysis of multicolor confocal immunofluorescence microscopy images: pushing pixels to explore biological phenomena. *Acta histochemica et cytochemica* 40: 101–111. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1993886&tool=pmcentrez&rendertype=abstract>.
- Comeau JWD, Costantino S, Wiseman PW (2006) A guide to accurate fluorescence microscopy colocalization measurements. *Biophysical journal* 91: 4611–22. Available: <http://www.ncbi.nlm.nih.gov/pubmed/17012312>.
- Kallioniemi OP, Wagner U, Kononen J, Sauter G (2001) Tissue microarray technology for high-throughput molecular profiling of cancer. *Human molecular genetics* 10: 657–62. Available: <http://www.ncbi.nlm.nih.gov/pubmed/11257096>.
- Simon R, Sauter G (2004) Tissue microarray (TMA) applications: implications for molecular medicine. *Expert Reviews in Molecular Medicine* 5: 1–12. Available: <http://www.ncbi.nlm.nih.gov/pubmed/14987401>. Accessed 4 Nov 2010.
- Simon R, Mirlacher M, Sauter G (2005) Tissue microarrays. *Methods in molecular medicine* 114: 257–68. Available: <http://www.ncbi.nlm.nih.gov/pubmed/16156109>. Accessed 4 Nov 2010.
- Bubendorf L, Nocito A, Moch H, Sauter G (2001) Tissue microarray (TMA) technology: miniaturized pathology archives for high-throughput in situ studies. *Journal of Pathology* 195: 72–79.
- Fuchs TJ, Wild P, Moch H, Buhmann J (2008) Computational Pathology Analysis of Tissue Microarrays Predicts Survival of Renal Clear Cell Carcinoma Patients. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*: 1–8.
- Vonesch J-luc, Unser M, Vonesch C, Aguet F (2006) The Colored Revolution of Bioimaging. *IEEE Signal Processing Magazine* 23: 20–31.
- Foulkes WD, Smith IE, Reis-Filho JS (2010) Triple-negative breast cancer: current state of the art. *New England journal of medicine* 363: 1938–1948. Available: <http://www.ncbi.nlm.nih.gov/pubmed/21388048>.
- Amaral T, Mckenna S, Robertson K, Thompson A (2008) Classification of breast tissue-microarray spots using texture histograms. In: *Medical Image Understanding and Analysis*. Dundee, United Kingdom. pp 999–1002.
- Amaral T, Mckenna S, Robertson K, Thompson A (2008) Classification of Breast-Tissue Microarray Spots using Colour and Local Invariants. *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. pp 999–1002.
- Wang C-W (2011) Robust automated tumour segmentation on histological and immunohistochemical tissue images. *PLoS one* 6: e15818. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3046129&tool=pmcentrez&rendertype=abstract>. Accessed 17 Mar 2011.
- Karacali B, Tözener A (2007) Automated detection of regions of interest for tissue microarray experiments: an image texture analysis. *BMC Medical Imaging* 20: 1–20. doi:10.1186/1471-2342-7-2.
- Padfield D, Chen Y, Roysam B, Cline H, Lin G, et al. (2006) Cancer Tissue Classification Using Nuclear Feature Measurements From DAPI-Stained Images. *Proceedings of 1st Workshop on Microscopic Image Analysis with Applications in Biology*. pp 86–92.
- Dolled-Filhart M, Rydén L, Cregger M, Jirstrom K, Harigopal M, et al. (2006) Classification of breast cancer using genetic algorithms and tissue microarrays. *Clinical cancer research* 12: 6459–68. Available: <http://www.ncbi.nlm.nih.gov/pubmed/17085660>. Accessed 21 Mar 2011.
- Camp RL, Chung GG, Rimm DL (2002) Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nature Medicine*;doi:10.1038/nm791.
- Teverovskiy M, Vengrenyuk Y, Tabesh A, Sapir M, Fogarasi S, et al. (2008) Automated localization and quantification of protein multiplexes via multispectral fluorescence imaging. *5th IEEE Int Symp Biomed Imag: From Nano to Macro*. pp 300–303. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4540992>.
- Tabesh A, Vengrenyuk Y, Teverovskiy M, Khan FM, Sapir M, et al. (2009) Robust tumor morphometry in multispectral fluorescence microscopy. *Medical Imaging 2009: Computer-Aided Diagnosis* 7260: 726015–726015-9. Available: <http://link.aip.org/link/PSISDG/v7260/i1/p726015/s1&Agg=doi>. Accessed 11 Jan 2011.
- Gunduz C, Yener B, Gultekin SH (2004) The cell graphs of cancer. *Bioinformatics (Oxford, England)* 20 Suppl 1: i145–51. Available: <http://www.ncbi.nlm.nih.gov/pubmed/15262793>.

20. Demir C, Gultekin SH, Yener B (2005) Augmented cell-graphs for automated cancer diagnosis. *Bioinformatics* (Oxford, England) 21 Suppl 2: ii7–12. Available: <http://www.ncbi.nlm.nih.gov/pubmed/16204128>.
21. Gunduz-Demir C (2007) Mathematical modeling of the malignancy of cancer using graph evolution. *Mathematical biosciences* 209: 514–27. Available: <http://www.ncbi.nlm.nih.gov/pubmed/17462676>.
22. Bilgin C, Demir C, Nagi C, Yener B (2007) Cell-graph mining for breast tissue modeling and classification. *IEEE Engineering in Medicine and Biology Society 2007*: 5311–4. Available: <http://www.ncbi.nlm.nih.gov/pubmed/18003206>.
23. Bilgin CC, Bullough P, Plopper G, Yener B (2010) ECM-aware cell-graph mining for bone tissue modeling and classification. *Data Mining and Knowledge Discovery*; Available: <http://www.springerlink.com/index/9624818545253XN8.pdf>. Accessed 8 Jul 2010.
24. Lahrman B, Halama N, Westphal K, Ernst C, Elsawaf Z, et al. (2010) Robust gridding of TMAs after whole-slide imaging using template matching. *Cytometry. Part A: the journal of the International Society for Analytical Cytology*. Available: <http://www.ncbi.nlm.nih.gov/pubmed/20662092>. Accessed 13 Oct 2010.
25. Otsu N (1979) A threshold selection method from grey level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9: 62–66.
26. Beucher S (1992) The watershed transformation applied to image segmentation. *Scanning Microscopy International* 6: 299–314.
27. Wahlby C, Sintorn I-M, Erlandsson F, Borgefors G, Bengtsson E (2004) Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. *Journal of Microscopy* 215: 67–76. Available: <http://doi.wiley.com/10.1111/j.0022-2720.2004.01338.x>.
28. Lin G, Adiga U, Olson K, Guzowski JF, Barnes CA, et al. (2003) A Hybrid 3D Watershed Algorithm Incorporating Gradient Cues and Object Models for Automatic Segmentation of Nuclei in Confocal Image Stacks. *Development* 36: 23–36. doi:10.1002/cyto.a.10079.
29. Soille P (1999) *Morphological Image Analysis: Principles and Applications* Springer Verlag. pp 170–171.
30. Bilgin CC, Lund AW, Can A, Plopper GE, Yener B (2010) Quantification of three-dimensional cell-mediated collagen remodeling using graph theory. *PLoS one* 5: Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2948014&tool=pmcentrez&rendertype=abstract>. Accessed 24 Oct 2010.
31. Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20: 273–297. Available: <http://www.springerlink.com/index/10.1007/BF00994018>.
32. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America* 97: 262–7. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=26651&tool=pmcentrez&rendertype=abstract>.
33. Matsuda S, Vert J-philippe, Saigo H, Ueda N, Toh H, et al. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Science*. pp 2804–2813. doi:10.1110/ps.051597405.
34. Mourão-Miranda J, Bokde ALW, Born C, Hampel H, Stetter M (2005) Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *Neuro Image* 28: 980–95. Available: <http://www.ncbi.nlm.nih.gov/pubmed/16275139>.
35. Chen Y-wei, Lin C-jen (2006) Combining SVMs with Various Feature Selection Strategies. In: *Feature extraction, foundations and applications* Springer. 315:323 p.