Bioinformatics, 32(1), 2016, 122–129
doi: 10.1093/bioinformatics/btv542
Advance Access Publication Date: 11 September 2015
Original Paper

OXFORD

Bioimage informatics

CD30 cell graphs of Hodgkin lymphoma are not scale-free—an image analysis approach

Hendrik Schäfer[1,†], Tim Schäfer[1,†], Jörg Ackermann[1], Norbert Dichter[1], Claudia Döring[2], Sylvia Hartmann[2], Martin-Leo Hansmann[2] and Ina Koch[1,*]

[1]Department of Molecular Bioinformatics, Institute of Computer Science, Cluster of Excellence Macromolecular Complexes, Johann Wolfgang Goethe-University Frankfurt am Main, Robert-Mayer-Str. 11-15, 60325 Frankfurt am Main and [2]Dr. Senckenbergisches Institut für Pathologie, Universitätsklinikum Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Robert Murphy

Received on May 7, 2015; revised on September 7, 2015; accepted on September 8, 2015

Abstract

Motivation: Hodgkin lymphoma (HL) is a type of B-cell lymphoma. To diagnose the subtypes, biopsies are taken and immunostained. The slides are scanned to produce high-resolution digital whole slide images (WSI). Pathologists manually inspect the spatial distribution of cells, but little is known on the statistical properties of cell distributions in WSIs. Such properties would give valuable information for the construction of theoretical models that describe the invasion of malignant cells in the lymph node and the intercellular interactions.
Results: In this work, we define and discuss HL cell graphs. We identify $CD30^{+}$ cells in HL WSIs, bringing together the fields of digital imaging and network analysis. We define special graphs based on the positions of the immunostained cells. We present an automatic analysis of complete WSIs to determine significant morphological and immunohistochemical features of HL cells and their spatial distribution in the lymph node tissue under three different medical conditions: lymphadenitis (LA) and two types of HL. We analyze the vertex degree distributions of CD30 cell graphs and compare them to a null model. CD30 cell graphs show higher vertex degrees than expected by a random unit disk graph, suggesting clustering of the cells. We found that a gamma distribution is suitable to model the vertex degree distributions of CD30 cell graphs, meaning that they are not scale-free. Moreover, we compare the graphs for LA and two subtypes of HL. LA and classical HL showed different vertex degree distributions. The vertex degree distributions of the two HL subtypes NScHL and mixed cellularity HL (MXcHL) were similar.
Availability and implementation: The CellProfiler pipeline used for cell detection is available at https://sourceforge.net/projects/cellgraphs/.
Contact: ina.koch@bioinformatik.uni-frankfurt.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Hodgkin lymphoma (HL) is one of the most common types of lymphoid malignancies and one of the most common cancer types in adolescents and younger adults (Liu *et al.*, 2014). HL can be divided into classical HL (cHL) and the rare nodular lymphocyte predominant HL forms.

In cHL, the tumor cells are called Hodgkin and Reed-Sternberg (HRS) cells. HRS cells are often large and multinucleated with a diameter of up to 100 μm. The mononucleated forms are smaller with a diameter of about 20–30 μm (Drexler *et al.*, 1986; Rengstl *et al.*, 2013). HRS cells usually originate from B cells (Küppers *et al.*, 1994, 1995). In HL, tumor cells are distributed in the lymphoid tissue and make up about 1% of the infiltrating cells in a background of reactive bystander cells such as T-cells, B-cells, histiocytes and others. The HRS cells are known to express CD30 and usually CD15 (Falini *et al.*, 1987). They also produce various chemokines and cytokines which shape the microenvironment according to their needs (Liu *et al.*, 2014). A review of HL biology can be found in Küppers *et al.* (2012).

HRS cells can easily be detected by an antibody which is specific for CD30. It is used for diagnosis of the subtypes. CD30 (Stein *et al.*, 1985) is a member of the tumor necrosis factor receptor family (Dürkop *et al.*, 1992) and a marker for HL (Al-Shamkhani, 2004; Chiarle *et al.*, 1999). Malignant lymphomas are diagnostically defined by morphology. Histological samples are visualized by light microscopy and are interpreted by a trained pathologist according to the rules of the World Health Organization (Swerdlow *et al.*, 2008) in the context of experience in the diagnostic field. NScHL and MXcHL are differentiated by the presence of sclerotic bands in NScHL, which confine at least one nodular compartment containing typical HRS cells. In the absence of these sclerotic bands, a case is assigned to the MXcHL category. CD30 is present not only in lymphoid neoplasms but also on virus-infected lymphocytes and special types of activated T cells (Horie and Watanabe, 1998). Thus, the HRS cells form a subset of the $CD30^+$ cells. This explains that $CD30^+$ cells are also found in lymphadenitis (LA). LA with abundant $CD30^+$ blasts and HL can be hard to distinguish.

Because of the spread of digital pathology, public databases of HL whole slide images (WSIs) like the Cancer Digital Slide Archive (Gutman *et al.*, 2013), as well as internal databases at hospitals and institutes like the *Dr. Senckenbergisches Institut für Pathologie*, are currently emerging. For some example WSIs, see Supplementary Figures S2–S4 in the Supplementary Material. Here, we were interested in the distribution of $CD30^+$ cells in WSIs of HL biopsies to detect significant patterns of HRS cells in the tissue. We wanted to explore the biological behavior of $CD30^+$ cells under neoplastic (HL) and reactive (LA) conditions. We focused on HL because the clinical co-authors provided the data and have more than 30 years of experience with HL.

We defined CD30 cell graphs based on $CD30^+$ cell positions (See Section 3.5 for the exact definition of a cell's position.). Many biological phenomena can be represented as graphs, and graphs have also been applied to model cell distributions and cancer. In Gunduz *et al.* (2004), the authors have considered graphs based on all cells in low-resolution brain tissue samples and have applied them to classify the samples into tumor, inflammatory and healthy tissue. For automated cancer diagnosis, an augmented version of the method, using weighted graphs, has been published, representing cell clusters instead of single cells (Demir and Yener, 2005; Demir *et al.*, 2005a, b). In Oztan *et al.* (2012), features extracted from nuclei- and cytoplasm-based cell graphs have been used in combination with other descriptors to grade follicular lymphoma. Texture-based methods have been used to describe clinically relevant characteristics of tissue (Ergen and Baykara, 2014; Fatima *et al.*, 2014; Lessmann *et al.*, 2007). Texture-based methods do not explicitly determine cell positions.

While cell detection is a very common task in biology, only a few studies have tried to detect all cells in WSIs (Huang *et al.*, 2011). In Kong *et al.* (2009) and Sertel *et al.* (2009), neuroblastomas have been classified based on WSIs. To the best of our knowledge, no systematic analysis of WSIs exists for HL.

Our approach uses cell positions to create cell graphs of $CD30^+$ cells in WSIs. We use individual cells due to the low density of $CD30^+$ cells in HL. In contrast to other graph-based methods, we analyzed high-resolution images and did not rely on pre-selected regions. This gave us the opportunity to gather data on the whole lymph node in contrast to restrict the analysis to a specific region.

For reasons of clarity and comprehensibility, we divided the Section 3 into five parts. A definition of cell graphs is given in Section 3.1. Section 3.2 considers scale-freeness. Afterward, we describe a null model in Section 3.3. Section 3.4 introduces the image format and the three image sets we analyzed. We briefly describe our imaging pipeline and its implementation in Section 3.5. Section 4 presents a validation of the imaging pipeline and the analysis of the vertex degree distributions of the cell graphs.

2 Approach

We apply the concept of network theory to the spatial distribution of $CD30^+$ cells in tissue. Our contribution includes the development and validation of a new imaging pipeline, the definition and computation of cell graphs, the formulation and simulation of an appropriate null model and the analysis of the vertex degree distributions of the cell graphs.

We applied our imaging pipeline to WSIs of HL biopsies to detect $CD30^+$ cells. On the basis of the two-dimensional cell coordinates, we constructed cell graphs for a selection of exemplary images of two subtypes of HL and LA. First, we investigated whether the distribution of $CD30^+$ cells in the tissue differed from randomly distributed cells. We simulated an appropriate null model for each WSI and compared it to the observed cell distribution. Second, we compared the vertex degree distributions for two subtypes of HL and LA. Third, we discussed the functional form of the vertex degree distributions and evaluated the compatibility of the distributions, using standard models of network construction as, e.g. the scale-free Barabási–Alberts model (Albert and Barabási, 2002).

3 Methods

3.1 Cell graphs

A cell graph represents the spatial distribution of one or several cell types within the tissue. In this work, we considered the special case of $CD30^+$ cells in images of HL. We thus called the resulting cell graph the *CD30 cell graph* of an image. In a *cell graph*, $G = (V, E)$, each vertex v represents a cell at a position in space and an edge $e = (u, v)$ is created between a vertex pair u, v, if the Euclidian distance between u and v is less than an edge threshold t. We set the threshold, t, to 700 pixels, i.e. to 175 μm or, about 10 times the diameters of an average cell. Figure 1 depicts an example cell graph of an NScHL image. Here, we can clearly see the sclerotic regions of the lymph node that dissect the cell graph into several connected components.

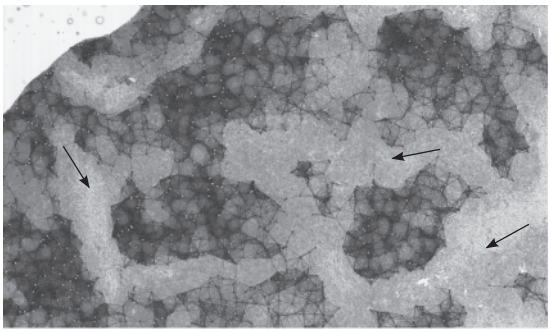

Fig. 1. A section of a stained NScHL image and the corresponding cell graph. The detected cells are visible as dots and represent the vertices of the graph. An edge is created between a pair of vertices if they are $<175\ \mu$m apart. The graph consists of several connected components, which are often separated by sclerotic lymph node regions indicated by the arrows. A larger color version of this image is available in the Supplementary Figure S12

Our definition of CD30 cell graphs applies the well-known rules of unit disk graphs to connect vertices according to their Euclidian distances. Unit disk graphs are intersection graphs of equally sized circles in the plane (Clark *et al.*, 1990) and have been applied to model, e.g. broadcast networks (Kammerlander, 1984). A generalization of unit disk graphs are geographical threshold graphs that consider, beside the distances, the sum of weights (Masuda *et al.*, 2005). A similar approach would be possible for an alternative definition of cell graphs because each vertex has properties like size and volume which could be interpreted as the weight of the vertex. Topological properties like the small-world property (Watts and Strogatz, 1998) and the scale-freeness (Barabási and Albert, 1999) have been demonstrated for geographical threshold graphs even for vertices that are uniformly distributed in space (Masuda *et al.*, 2005). The strong influence of the construction rules on topological structures of geographical threshold graphs makes an interpretation of topological structures complicated for non-homogeneously distributed cells in the tissue. The simplicity of the statistical properties of unit disk graphs motivates our definition of CD30 cell graphs.

3.2 Vertex degree distribution and scale-freeness

Vertex degree distribution and scale-freeness are significant properties to describe large networks. The degree distribution of all vertices of a network is denoted by $p(k)$ and gives the probability that a randomly chosen vertex has the degree k. For randomly connected Erdös–Rényi networks (Erdös and Rényi, 1959), $p(k)$ follows a Poisson distribution.

Scale-free networks show highly inhomogeneous vertex degree distributions, which follow a power law $p(k) \sim k^{-\gamma}$, where γ is called the degree exponent (Barabási and Albert, 1999). A model for these graphs based on preferential attachment has been described by Albert and Barabási (2002). Biochemical networks have been found to be scale-free, e.g. please refer to the work of Jeong *et al.* (2000) and Guelzim *et al.* (2002). For most of the networks, the degree exponent lies between 2 and 3 (Steuer and Zamora-Lopez, 2008).

3.3 Null model and statistical analysis

A significant deviation from randomness is a prerequisite for the statistical analysis and biological interpretation. Randomness would mean that the WSI would give a snapshot of cells at random locations and any pattern would not be significant.

In first approximation, the positions of cells can be seen as a planar point pattern that results from a homogeneous Poisson point process (Diggle, 1983). The position of each cell is chosen randomly, each position is equally probable and the location of other cells is ignored for the placement of a cell. Unit disk graphs (Clark *et al.*, 1990; Kammerlander, 1984; Yeh *et al.*, 1984) are equivalents to cell graphs of such random planar point patterns. The probability to find k neighbors in a unit disk is given by a Poisson distribution

$$\mathrm{Pois}(k, \mu) = e^{-\mu} \frac{\mu^k}{k!} \tag{1}$$

where μ is the mean number of neighbors in the disc. Networks with Poisson distributed vertex degrees are not scale-free because moments of arbitrary order are defined for the Poisson distribution.

For a WSI with a mean density ρ of CD30$^+$ cells at random positions, a Poisson distribution with parameter $\mu = \rho\ \pi\ t^2$ would completely describe the vertex degree distribution of the cell graph. Distortions from the Poisson distribution can be expected only from the volume exclusion effect and effect of the border of the finite tissue area.

The volume exclusion effect prevents an infinite growth of the number of cells in the neighborhood. In the cell graph, the neighborhood area of a cell has the size $A = \pi t^2$ with $t = 175\ \mu$m, while the area of the profile of a CD30$^+$ cell in a WSI is typically $a = 349\ \mu\mathrm{m}^2$. Consequently, a rough estimate for the maximal number of cells that may populate the neighborhood is $A/a \approx 276$. Since $\mu \ll 276$ is true for all WSIs in our dataset, we can neglect the effect of this theoretical upper bound on the expected Poisson distribution.

Cells located closer than $175\ \mu$m to the tissue border have a reduced neighborhood area and contribute with a lower parameter $\tilde{\mu} < \mu$ to a superposition of Poisson distributions. They deform the vertex degree distribution $\mathrm{Pois}(k, \mu)$ towards smaller degrees. To estimate this deformation, we performed 10 simulations of the homogenous Poisson point process for each WSI. We counted the number of tiles that contained tissue in the WSI and computed the total size of these tiles. For each cell in the WSI, we chose a new random position in a rectangular area of identical size and computed a random cell graph of the randomly generated spatial distribution of cells. The gamma distribution

$$g(k; \alpha, \beta) = \frac{\beta^\alpha\, k^{\alpha-1}\, e^{-k\beta}}{\Gamma(\alpha)} \tag{2}$$

with vertex degree, k, rate parameter, α, and shape parameter, β, is a generalization of the exponential distribution which has been used to model populations in biology (Dennis and Patil, 1984; Engen and Lande, 1996).

3.4 Image data

The *Dr. Senckenbergisches Institut für Pathologie* provided the input images for our study. The lymph node sections were double-stained with hematoxylin and a fuchsine immunostaining, targeting CD30. This is a standard procedure used in the diagnosis of HL. The tissue samples were digitized using an Aperio ScanScope device (LeicaBiosystems, 2011). The output image format was a pyramidal format containing several resolution levels of the same field of view, see Supplementary Figure S11a. At level 0, the WSIs had a resolution of $0.25\ \mu$m per pixel and dimensions of up to $100\,000 \times 100\,000$ pixels. Performing a cell detection in the complete WSI requires a very high image quality. In contrast, for the diagnosis, only a few cells are inspected in detail by pathologists. We pre-selected and

analyzed 35 WSIs of good quality. Pathologists diagnosed and labeled each image, resulting in three separate image sets, each of which consists of at least 11 images. Two sets represent the most common subtypes of cHL, MXcHL and NScHL. The third set contains LA cases. The latter can be seen as control group, as LA is an inflammation of the lymph node and the typical immune response for many diseases, e.g. bacterial and viral infections.

3.5 Imaging pipeline

To detect all $CD30^+$ cells in the WSI, we developed a software solution in Java, using the Openslide library in combination with the Java Advanced Imaging API (Goode and Satyanarayanan, 2008) and an interface to CellProfiler2 (Lamprecht *et al.*, 2007). The software ran on a Linux cluster (10 nodes, 80 cores, up to 128 GB RAM/node, OpenSuSE 13.1 64 bit) to process large images in parallel. It performed preprocessing, split the large WSI into tiles and detected cells in those tiles. The positions of all cells were stored in a database. Runtimes were in the range of a few hours for the analysis of an image.

3.5.1 Generation of image tiles from the WSI

A WSI of highest resolution may exceed a size of 30 GB and is too large to fit into the memory of today's standard computers. Therefore, we covered the image by smaller tiles, which we exported to standard TIFF files. We used these files as input images for the cell detection step. We generated tiles with a core area of 1024×1024 pixels, i.e. 236×236 μm^2, plus a border of 100 pixels, i.e. 25 μm at highest resolution. Supplementary Figure S11b sketches the cover of an image by tiles. The overlapping borders ensured that we did not ignore cells that were cut by a border. Cells in the overlapping area could lead to duplicate entries for a cell. If several cells shared the same position, we kept only one of them. The processing of the tiles did not have any effect on the resulting graphs or detected cells. The splitting of the image was required for parallelization and to reduce the memory needed to open a single image.

3.5.2 Preprocessing

The input images contained stained tissue in front of a bright background. Many of them showed artifacts like air bubbles, small tissue fragments and stain residues. We identified the region of interest (ROI), i.e. the tissue sections, in the image on the second resolution level. The ROI was defined using a minimal distance to mean approach based on brightness and saturation pixel descriptors (Schäfer *et al.*, 2013). The image was segmented into tissue and brighter, less saturated background. For each tile, we checked whether it contained enough tissue pixels. Only tiles with at least 100 tissue pixels were considered to be part of the ROI. In the next step, we applied a region growth algorithm (Gonzalez and Woods, 2008) to combine adjacent tissue tiles into a single object, representing a tissue section. All tissue sections smaller than 4 mm^2 were considered as artifacts and filtered.

3.5.3 Cell detection

For the image tiles of the WSI, we initially applied a color deconvolution to split the image into the three channels hematoxylin, new fuchsine and non-specific background. We saved the new fuchsine channel as a CD30 gray-scale image.

Because of differences in the degree of staining and exposure, the thresholds used for segmentation were adapted for each WSI. We tried several methods to determine the thresholds: the MoG Global method, adaptive threshold and manual determination. The MoG Global method and the adaptive threshold method, both provided by CellProfiler2, did not perform well. Because of the inhomogeneous staining of the images and the high differences of cell counts within the image tiles, the two algorithms failed to calculate a suitable threshold. Thus, we used manual thresholds applied by the first authors (H.S. and T.S.) independently. For each pixel that had an intensity above the threshold, we applied a region growth algorithm based on a four-connected neighborhood to identify the profile of a $CD30^+$ cell.

The identified connected areas of high intensity included fragments that were too small to represent a cell. We filtered them by discarding small objects with an area of <1750 pixels (about 110 μm^2). All remaining objects were considered cells in the image. For each cell, we computed the x and y coordinates of its center of gravity and call them the cell's position. Supplementary Figures S6–S11 illustrate the steps of the pipeline.

We manually marked all cells in a set of randomly chosen image tiles. For validation purposes, we compared the cells detected by our pipeline to the results of manual cell labeling. We counted the number of true positives, TP, false positives, FP, and false negatives, FN. As quality criteria for our automated cell detection, we determined values of precision and sensitivity according to

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{and} \quad \text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

4 Results and discussion

4.1 Image properties

The total number of cells varied in the range of a few hundred cells up to more than 90 000 $CD30^+$ cells per image. The average size of the $CD30^+$ cells was 294 μm^2. Besides the total number of cells, the cell density also fluctuated significantly within our set of 35 images. The cell density was measured as number of cells per cm^2 in the ROI (see Supplementary Fig. S1). Table 1 gives values averaged over all images of the two HL subtypes, MXcHL and NScHL, and LA, respectively. The highest mean value of 104.3 cells per cm^2 was measured for MXcHL, whereas the lowest value of 28.8 cells per cm^2 was observed for LA. The mean value of 74.7 cells per cm^2 for NScHL laid between the values for MXcHL and LA. Within each subgroup of disease, the cell density varied significantly from image to image. Standard deviations of 72%, 70% and 105% were measured for the mean cell densities of MXcHL, NScHL and LA, respectively.

There are a number of effects which may be responsible for the high variety of cell densities. First, the staining quality of the images varies significantly and influences the outcome of the cell detection pipeline. Second, our input data are an arbitrarily chosen

Table 1. Total cell count and cell density averaged over the images in the medical subtypes MXcHL, NScHL and LA, respectively

	Number of cells		Cell density	
Diagnosis	Mean	SD	Mean	SD
MXcHL	19.0	15.7	104.3	75.1
NScHL	13.2	13.2	74.7	52.5
LA	3.0	2.2	28.8	30.3

The total cell counts are given in kCounts and the cell density in number of cells per cm^2.

2D cross-section and planar view on a spatial distribution of cells in the complex 3D structure of a lymph node. Third, the progression of disease is unknown. Since we work with human data, each image represents the state when HL or LA was diagnosed for the patient. Information on the duration and progression of the disease is lacking. Fourth, each image shows the distribution of $CD30^+$ cells in a single lymph node only.

The high variety of the cell densities may have an impact on the results. The number of high-quality images available for the study does not allow a further separation of the images into groups of similar cell density.

4.2 Validation of the imaging pipeline

We validated the image processing outcome for three images, one of each image set. Table 2 lists the overall values for cell counts, precision and sensitivity, respectively. For all three images, the total precision is 0.84 and the total sensitivity is 0.95. These values are satisfactorily high when compared with benchmark values of 61% for precision and sensitivity in the AMIDA13 challenge where the harder task has been to detect mitosis events (Veta *et al.*, 2015).

Figure 2 illustrates the results for the image of the medical subtype NscHL. Each column in Figure 2 shows the fraction of FN, FP and TP separately for various tiles of the image. For each randomly chosen tile, its relative coordinates in the WSI are encoded on the *x*-axes. The results are very heterogeneous. The fraction of TP is high, i.e. about 80%, for some tiles, but drops to 15–30% for other tiles. By visual inspection of the tiles with low fraction of TP, we observed a low degree of staining (see Supplementary Fig. S5). The pipeline applied an overall intensity threshold for a WSI and thus failed to detect cells with very low staining. It is worth mentioning that detecting all cells in a WSI is much more complex than detecting cells in small, preselected image regions. Apart from technical challenges due to the image size and differences in staining between the WSIs, the varying degree of staining within a single image in combination with the large differences in cell counts made it difficult to find thresholds for segmenting the images automatically. The $CD30^+$ cells we are interested in are also very variable in shape and size, and they cluster very tightly in some areas. This makes object-based detection difficult, as previously reported for macrophages in breast cancer (Krüger *et al.*, 2013). Using manually selected thresholds, our pipeline was able to reach good precision and sensitivity for the majority of image tiles.

4.3 CD30 cell graph properties

The number of edges in the graphs ranged from some hundreds to several millions of edges. For all images, the mean cluster coefficient was 0.67, with a standard deviation of 0.05. The mean distance of a cell to its closest neighbor differed for the medical subtypes. The mean distances were $61 \pm 10\,\mu m$, $43 \pm 12\,\mu m$ and $38 \pm 6\,\mu m$ for LA, MXcHL and NScHL, respectively.

Table 2. Validation results of cell detection

	Cell count				
Diagnosis	TP	FP	FN	Precision	Sensitivity
MXcHL	952	189	12	0.83	0.98
NScHL	686	155	81	0.81	0.89
LA	921	137	18	0.87	0.98
All	2559	481	111	0.84	0.95

4.4 CD30 cell graphs are not random

Figure 3 shows exemplary vertex degree distributions $p(k)$ for the cell graph of an image of LA (part a), NScHL (part b) and MXcHL (part c). Each of these distributions deviates significantly from the Poisson distribution $\mathrm{Pois}(k, \mu)$ with parameter μ chosen according to the mean density of $CD30^+$ cells. The error bars in Figure 3 show the effect of the finite tissue area on the Poisson distribution. The simulated vertex degree distributions turn out to be indistinguishable from the Poisson distribution. A sufficient proof for the non-randomness of a cell graph is the significant deviation of the vertex degree distribution from the corresponding Poisson distribution that we observed for all images.

4.5 $CD30^+$ cells cluster in the tissue

For each WSI, the average vertex degree of the CD30 cell graphs is higher than expected from a homogeneous density of $CD30^+$ cells in the tissue (Fig. 3). The high vertex degree suggests that $CD30^+$ cells form clusters in the tissue. A spatial clustering is in accordance with our expectation from the visual inspection of the images. Looking at the distribution of $CD30^+$ cells in different subtypes, cells in images of NScHL showed a pronounced spatial clustering, while they were more homogeneously distributed in images of LA. For MXcHL, the cells appeared to be significantly clustered for some images but more evenly distributed for other images.

Several possible causes for the clustering of the cell came to mind. First, clustering could be related to cell division in combination with relatively immobile cells: a few $CD30^+$ cells appear at or migrate into a lymph node. Once in place, they become immobile. When the $CD30^+$ cells divide, the resulting daughter cells stay close to each other, leading to a clustering in the area. Second, another possibility would be attraction between mobile cells or attraction of mobile cells to certain areas in the lymph node: the $CD30^+$ cells move through the lymph node tissue and form clusters in regions they are attracted to. Further investigations, targeting the mobility of the malignant cells and cell–cell communication, are necessary to understand the underlying process.

In the case of NScHL, the high degree of clustering could also be influenced by the sclerotic bands that are typical for this subtype of HL. The overall cell density is very low in the sclerotic regions, which are devoid of $CD30^+$ cells, and these cells cluster even more

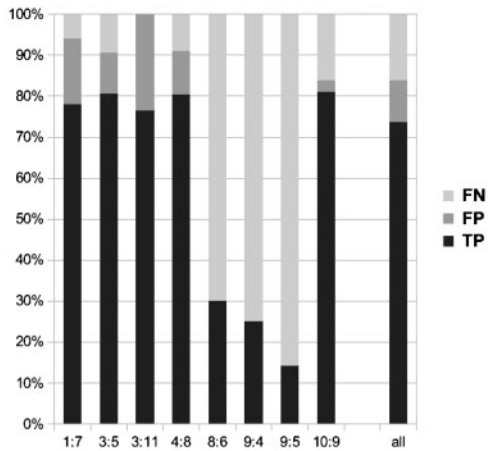

Fig. 2. Validation of the image pipeline for eight randomly chosen tiles of an example NscHL image. On the *x* axis, the tile coordinates are listed. The *y* axis shows the validation results in %. TP, true positives; FN, false negatives; FP, false positives

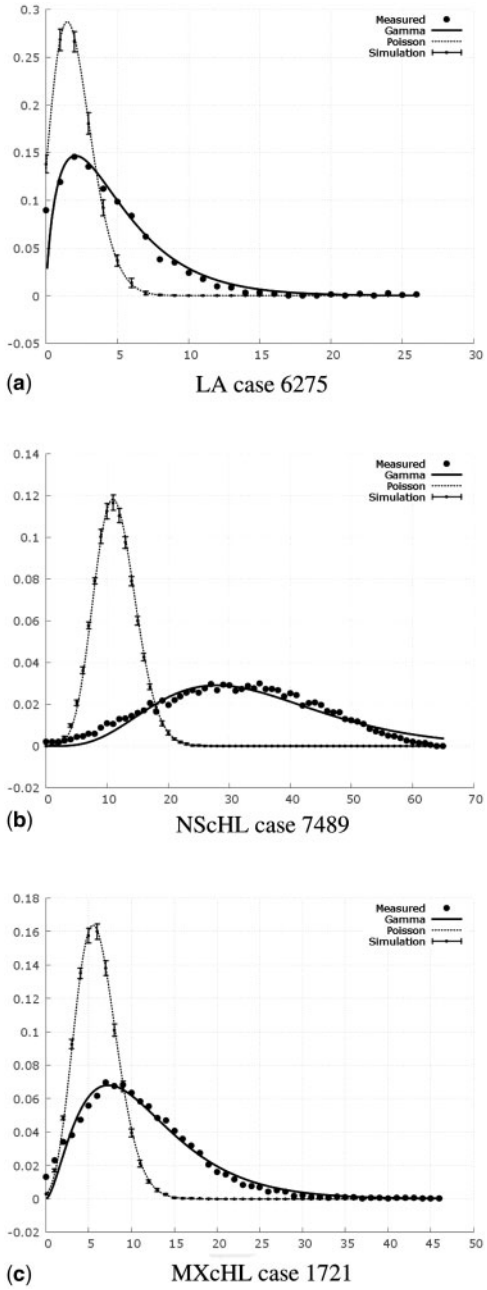

Fig. 3. Exemplary vertex degree distributions $p(k)$ for the cell graph of an image of lymphadenitis (**a**), nodular sclerosis (**b**) and mixed cellularity (**c**). The filled circles depict the observed degree distribution. The distribution deviates significantly from the corresponding Poisson distribution (null model, dotted line) but is much better described by the gamma distribution (solid line). We simulated unit disk graphs according to the null model. The error bars illustrate the mean and standard deviation of 10 simulations

in the rest of the tissue. For further investigations, mitosis markers like Ki-67 (Gerdes *et al.*, 1983) could be used.

4.6 CD30 cell graphs are not scale-free

We found that a gamma distribution best describes the vertex degree distributions, see e.g. Fig. 3. Moments of arbitrary degrees exist for this two-parameter exponential family and hence, the CD30 cell graphs are not scale-free. In contrast to biochemical networks which are scale-free, cell graph do not exhibit that property.

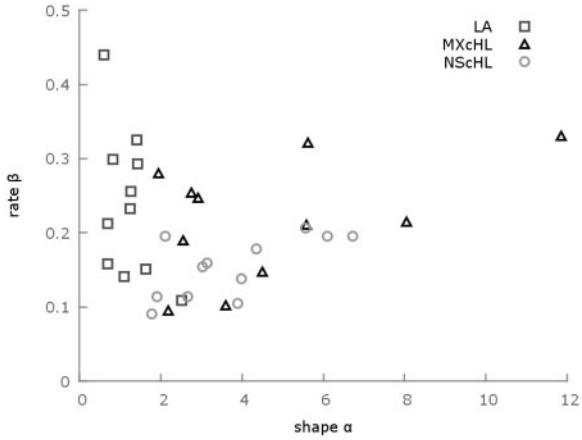

Fig. 4. The rate parameter, β, and shape parameter, α, of the gamma distribution fitted to observed vertex degree distributions. Each data point represents a WSI of MXcHL, LA or NScHL

4.7 CD30 cell graphs show significant differences between the disease types

In CD30 cell graphs, high probabilities for a large vertex degree are typical for NScHL, while low values are typical for LA. This is not a surprising result because most LA images show a considerably lower density of $CD30^+$ cells than NScHL and MXcHL. It is, however, notable that NScHL graphs often exhibited high vertex degrees even though their mean cell density is comparably low. This could be related to the inhomogeneous distribution of $CD30^+$ cells in NScHL with dense cell clusters, see e.g. Fig. 1.

In Figure 4, the rate and shape parameters of the gamma distribution are plotted for all images. Shape parameter values between 0 and 2 are typical for LA cases. The rate parameter lies below 0.21 for all NScHL cases. Note that LA could be separated from the cHL cases quite well.

5 Conclusion

In this article, we describe the first systematic, graph-based exploration of stained cancer tissue, here, of CD30-stained tissues of HL. Our contributions are (i) the automated analysis of WSIs, using an own image pipeline, (ii) the definition of a graph-based data structure, the CD30 cell graphs, to abstract from images, which enables for computational analysis, (iii) the statistical analysis of the cell graphs which shows that they differ significantly from randomly generated cell graphs, (iv) the analysis of vertex degree distributions in the cell graphs which follow a gamma distribution and are thus not scale-free, which is unusual for biological networks and (v) the application to HL subtypes which shows unexpected results.

We performed an image analysis on a dataset of HL images to detect $CD30^+$ cells. MXcHL cases had the highest average cell density, followed by NScHL. The LA cases showed considerably lower $CD30^+$ cell densities than the two forms of cHL.

We defined a new data structure, the cell graph, which describes the spatial neighborhood of cells. We analyzed the vertex degree distributions of CD30 cell graphs. The comparisons with Poisson distributions indicate that the distributions in cell graphs differ significantly from those of simulated unit disk graphs. The vertex degree distributions of cell graphs follow a gamma distribution for which moments of arbitrary degree exist. This finding rules out the hypothesis of scale-free cell graph networks.

Scale-free network models, e.g. the Barabasi–Albert model, are not suitable to describe the complex, cooperative system of cells in the HL tissue. This means that no key regulators exist in the network of HL cells. The lack of key regulators could mean that the distribution of CD30 cells is mainly influenced by external factors, e.g. the lymph node structure or the microenvironment. If the CD30 cells shape the network topology, they all show a similar behavior.

We found that the $CD30^+$ cells cluster in the tissue, and the clustering is usually high for NScHL and low for LA. The reason is unknown. We discuss several possible scenarios: high proliferation and rather immobile cells, or mobile cells which are attracted to each other or to certain lymph node regions. For NScHL, the sclerotic bands may increase the clustering of $CD30^+$ cells in the non-sclerotic lymph node tissue.

We demonstrate that cell graphs enable the analysis of HL tissue properties. The investigation of the vertex degree distributions in cell graphs already gave new insights. The cell graph concept can be extended by data on locations of other cell types from the tumor micro-environment and blood vessels to help in better understanding how tumor cells spread through the lymph system.

Findings of this investigation provide objective parameters for $CD30^+$ cells in HL. The method can be extended for different immuno cell types, their shapes and localizations as well as their interactions. These data can be used for a deeper understanding of the biological meaning of interactions between $CD30^+$ tumor cells and possibly also in the future to assess the impact of antitumor drugs, help with the prognosis and develop new therapies.

Funding

This research was supported by the DFG project NE1438/4-1: Mature T-cell Lymphomas—Mechanisms of Perturbed Clonal T-Cell Homeostasis. T.S. was partly supported by a grant from Friedrich Naumann Foundation for Freedom.

Conflict of Interest: none declared.

References

Al-Shamkhani,A. (2004) The role of CD30 in the pathogenesis of haematopoietic malignancies. *Curr. Opin. Pharmacol.*, **4**, 355–359.

Albert,R. and Barabási,A.-L. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**, 47–97.

Barabási,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Chiarle,R. *et al.* (1999) CD30 in normal and neoplastic cells. *Clin. Immunol.*, **90**, 157–164.

Clark,B.N. *et al.* (1990) Unit disk graphs. *Discrete Math.*, **86**, 165–177.

Demir,C. and Yener,B. (2005) Automated cancer diagnosis based on histopathological images: a systematic survey. *Technical report*. Rensselaer Polytechnic Institute.

Demir,C. *et al.* (2005a) Augmented cell-graphs for automated cancer diagnosis. *Bioinformatics*, **21**, ii7–ii12.

Demir,C. *et al.* (2005b) Spectral analysis of cell-graphs for automated cancer diagnosis. In: Peter Fritzson,L. (ed.) *Proceedings of the Conference on Modeling and Simulation in Biology, Medicine and Biomedical Engineering*, The Programming Environments Laboratory, Department of Computer and Information Science, Linköping University, pp. 153–160.

Dennis,B. and Patil,G.P. (1984) The gamma distribution and weighted multimodal gamma distributions as models of population abundance. *Math. Biosci.*, **68**, 187–212.

Diggle,P.J. (1983) *Statistical Analysis of Spatial Point Patterns*. Academic Press, London, UK.

Drexler,H.G. *et al.* (1986) Hodgkins disease-derived cell lines HDLM-2 and L-428: comparison of morphology, immunological and isoenzyme profiles. *Leuk. Res.*, **10**, 487–500.

Dürkop,H. *et al.* (1992) Molecular cloning and expression of a new member of the nerve growth factor receptor family that is characteristic for Hodgkin's disease. *Cell*, **68**, 421–427.

Engen,S. and Lande,R. (1996) Population dynamic models generating species abundance distributions of the gamma type. *J. Theor. Biol.*, **178**, 325–331.

Erdös,P. and Rényi,A. (1959) On random graphs. *Publicationes Mathematicae*, **6**, 290–297.

Ergen,B. and Baykara,M. (2014) Texture based feature extraction methods for content based medical image retrieval systems. *Bio-Med. Mater. Eng.*, **24**, 3055–3062.

Falini,B. *et al.* (1987) Expression of lymphoid-associated antigens on Hodgkin's and Reed-Sternberg cells of Hodgkin's disease. An immunocytochemical study on lymph node cytospins using monoclonal antibodies. *Histopathology*, **11**, 1229–1242.

Fatima,K. *et al.* (2014) A new texture and shape based technique for improving meningioma classification. *Microsc. Res. Tech.*, **77**, 862–873.

Gerdes,J. *et al.* (1983) Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. *Int. J. Cancer*, **31**, 1320.

Gonzalez,R.C. and Woods,R.E. (2008) *Digital Image Processing*, 3rd edn. Prentice Hall International, Upper Saddle River, New Jersey.

Goode,A. and Satyanarayanan,M. (2008) A vendor-neutral library and viewer for whole-slide images. *Technical report*, Computer Science Department, Carnegie Mellon University.

Guelzim,N. *et al.* (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.*, **31**, 60–63.

Gunduz,C. *et al.* (2004) The cell graphs of cancer. *Bioinformatics*, **20**, i145–i151.

Gutman,D.A. *et al.* (2013) Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J. Am. Med. Inform. Assoc.*, **20**, 1091–1098.

Horie,R. and Watanabe,T. (1998) CD30: expression and function in health and disease. *Seminars in Immunology*, **10**, 457–470.

Huang,C.-H. *et al.* (2011) Time-efficient sparse analysis of histopathological whole slide images. *Comput. Med. Imaging Graph.*, **35**, 579–591.

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kammerlander,K. (1984) C 900—an advanced mobile radio telephone system with optimum frequency utilization. *IEEE J. Selected Areas Commun.*, **2**, 589–597.

Kong,J. *et al.* (2009) Computer-aided evaluation of neuroblastoma on whole-slide histology images: classifying grade of neuroblastic differentiation. *Pattern Recognit.*, **42**, 1080–1092.

Krüger,J.M. *et al.* (2013) Combat or surveillance? Evaluation of the heterogeneous inflammatory breast cancer microenvironment. *J. Pathol.*, **229**, 569–578.

Küppers,R. *et al.* (1994) Hodgkin disease: Hodgkin and Reed-Sternberg cells picked from histological sections show clonal immunoglobin gene rearrangements and appear to be derived from B cells at various stages of development. *Proc. Natl. Acad. Sci. USA*, **91**, 10962–10966.

Küppers,R. *et al.* (1995) Hodgkin's disease: clonal Ig gene rearrangements in Hodgkin and Reed-Sternberg cells picked from histological sections. *Ann. N. Y. Acad. Sci.*, **764**, 523–524.

Küppers,R. *et al.* (2012) Hodgkin lymphoma. *J. Clin. Invest.*, **122**, 3439–3447.

Lamprecht,M.R. *et al.* (2007) CellProfiler: free, versatile software for automated biological image analysis. *Biotechniques*, **42**, 71.

LeicaBiosystems (2011) Digital Pathology—Aperio: Leica Biosystems. Online. http://www.leicabiosystems.com/pathology-imaging/aperio-digital-pathology/ (24 July 2015, date last accessed).

Lessmann,B. *et al.* (2007) A method for linking computed image features to histological semantics in neuropathology. *J. Biomed. Inform.*, **40**, 631–641.

Liu,Y. *et al.* (2014) The microenvironment in classical Hodgkin lymphoma: an actively shaped and essential tumor component. *Semin. Cancer Biol.*, **24**, 15–22.

Masuda,N. *et al*. (2005) Geographical threshold graphs with small-world and scale-free properties. *Phys. Rev. E*, **71**, 036108.

Oztan,B. *et al*. (2012) Follicular lymphoma grading using cell-graphs and multi-scale feature analysis. *Proc. SPIE*, **8315**, 831516.

Rengstl,B. *et al*. (2013) Incomplete cytokinesis and re-fusion of small mononucleated Hodgkin cells lead to giant multinucleated Reed-Sternberg cells. *Proc. Natl. Acad. Sci. USA*, **110**, 20729–20734.

Schäfer,T. *et al*. (2013) Image database analysis of Hodgkin lymphoma. *Comput. Biol. Chem.*, **46**, 1–7.

Sertel,O. *et al*. (2009) A combined computerized classification system for whole-slide neuroblastoma histology: Model-based structural features. In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2009)*, pp. 7–18.

Stein,H.M. *et al*. (1985) The expression of the Hodgkin's disease associated antigen Ki-1 in reactive and neoplastic lymphoid tissue: evidence that Reed-Sternberg cells and histiocytic malignancies are derived from activated lymphoid cells. *Blood*, **66**, 848–858.

Steuer,R. and Zamora-Lopez,G. (2008) Global network properties. In: Junker,B.H. and Schreiber,F. (eds.) *Analysis of Biological Networks*. John Wiley & Sons, New Jersey, pp. 31–59.

Swerdlow,S.H. *et al*. (2008) *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissue*. World Health Organization, Geneva.

Veta,M. *et al*. (2015) Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med. Image Anal.*, **20**, 237–248.

Watts,D. and Strogatz,S. (1998) Collective dynamics of small-world networks. *Nature*, **393**, 440–442.

Yeh,Y.-S. *et al*. (1984) Outage probability in mobile telephony with directive antennas and macrodiversity. *IEEE Trans. Vehicular Technol.*, **33**, 123–127.