JMLR: Workshop and Conference Proceedings 11 (2010) 88–95 Workshop on Applications of Pattern Analysis

Unsupervised Nuclei Segmentation for HSSB1 Immunohistochemistry Lung Cancer Analysis

Ching-Wei Wang CWEIWANG@IEEE.ORG
Graduate Institute of Biomedical Engineering, National Taiwan University of Science and Technology, Taiwan

Editors: Tom Diethe, Nello Cristianini, John Shawe-Taylor

Abstract

HSSB1 has been shown to be critical in genomic stability and DNA damages, but the exact mechanism of how hSSB1 protects genome stability remains unclear. Immunohistochemistry is a typical method for studying archived tissues, and accurate and reliable segmentation of nuclei is an important step for the quantitative IHC image analysis of nuclear malignancy. This paper presents an automated unsupervised entropy-based image segmentation method for nuclei detection of IHC HSSB1 lung carcinoma tissue slides. The technique is demonstrated to perform well in preliminary experimental results and be useful in quantifying the percentage or expression level of positively stained nuclei. Furthermore, the extracted nuclei architecture information is beneficial for cancer subtype classification, and a novel approach is proposed for cancer subtype classification based on nuclei architecture information.

1. Introduction

HSSB1 is a recently discovered protein Richard et al. (2008) and has been shown to be critical in maintaining genomic stability and responding to DNA damages; HSSB1 participates in efficient homologous recombination-dependent repair of DNA double strand breaks and ataxia telangiectasia-mutated dependent signaling pathways Li et al. (2009). However, the exact mechanism of how hSSB1 protects genome stability remains unclear. Protein related biomarkers are commonly evaluated using immunohistochemistry (IHC) while nucleotide sequence related biomarkers can be evaluated in tissues using techniques such as fluorescent in-situ hybridization (FISH). The distribution of proteins within cells and tissues can be identified using IHC. Quantitative analysis of digitized IHC-stained tissue sections is increasingly used in research studies and clinical practice, and quantitative IHC techniques have often yielded clinically important information regarding patient diagnosis, prognosis, or both; the scores are used to quantify protein expression, stratifying patients and further identify effective biomarkers, which is important for therapeutic purpose Avninder et al. (2008).

Computer-assisted image analysis of IHC has been shown to reduce the variation in analysis of staining levels Seidal et al. (2001). A variety of studies have been published exploring the use of image analysis and machine vision for tissue analysis and biomarker measurement Brey et al. (2003); Camp et al. (2002). Camp et al. Camp et al. (2002) have proposed a system called AQUA for quantification of biomarker expression based on

FISH where specific fluorescent stains can be used for cell compartmentalization to detect nuclei, cytoplasm and membranes Bast et al. (2005). Robust automated approaches for IHC quantification are still under-developed and require the empirical evaluation of algorithms which can both measure the intensity and distribution of biomarker, but also do this within the architectural components of the tissue sample that are relevant to the study.

Accurate segmentation of nuclei is an important step for the quantitative IHC image analysis of nuclear malignancy; among the most useful features for cytological applications have been measures of nuclear size, pleomorphism and chromatin appearance Wu et al. (2004). To evaluate and analyze the properties of nuclei, segmentation of nuclear regions are needed. However, accurate segmentation of nuclei is often difficult because of the heterogeneous cellular staining and nuclear overlapping. The simplest approach for segmenting nuclei is a global thresholding, which is adjusted manually or determined by the measurement of the image histogram. Such method works well in high-contrast-feature tissue images such as applications to measure oestrogen and progesterone receptor levels in breast cancer Rexhepaj et al. (2008), but is not suitable for tissue images with varying image features (in Fig.1(a), some nuclei appear distinct but in the highlighted region, the contrast of image features is low). Mao et al.Mao et al. (2006) presented a supervised learning image segmentation method for P53 IHC images by separating two classes of image pixels (background and nuclei) from color image, using the learnt transformation formula from the dataset of background and nuclei pixels of the studied images, and thresholding the extracted nuclear image pixels using otsu clustering. Adaptive local thresholding techniques, which utilize local content information and automatically separates image pixels into different classes, produce significantly better results in comparison to the global thresholding method; we tested two commonly adopted techniques (Otsu clustering Otsu (1979) and K-means clustering Lloyd (1982)), showing that the unsupervised local thresholding methods are still not sufficient for nuclear segmentation (Fig.1(b,c,d)).

Another popular approaches for nuclear detection are watershed algorithms Vincent and Soille (1991). As in practice, the Vincent-Soille watershed tends to produce an over-segmentation(Fig.1(e)), we tested a watershed algorithm (adapted from Eddins (2006)) and marker-controlled watershed method Matlab with optimized empirically-set parameters (Fig.1(g,f)). However, the watershed methods still produce many false positives and false negatives. As a result, a robust method for nuclear segmentation in challenging IHC tissue images is needed, and in this study, a simple method is presented for nuclear segmentation in IHC tissue images (an output by the proposed method is shown in Fig.1(h)).

Mao et al.Mao et al. (2006) presented a supervised learning image segmentation method for P53 IHC images by separating two classes of image pixels (background and nuclei) from color image, using the learnt transformation formula from the dataset of background and nuclei pixels of the studied images, and thresholding the extracted nuclear image pixels using otsu clustering. In this study, a generic unsupervised image segmentation method with entropy-based multistage separation technique is developed. In manufacturing IHC, DAB reacts with the targeting biomarker to give a brown coloration and Haematoxylin stains the background tissue blue. Although the dyes used are visualized as having different colors, the resulting stains actually have complex overlapping absorption spectra, and the quantification of DAB cannot be determined at a single wavelength because the optical density at such wavelength is determined by the total absorption of the multiple stains. Hence, Ruifrok and

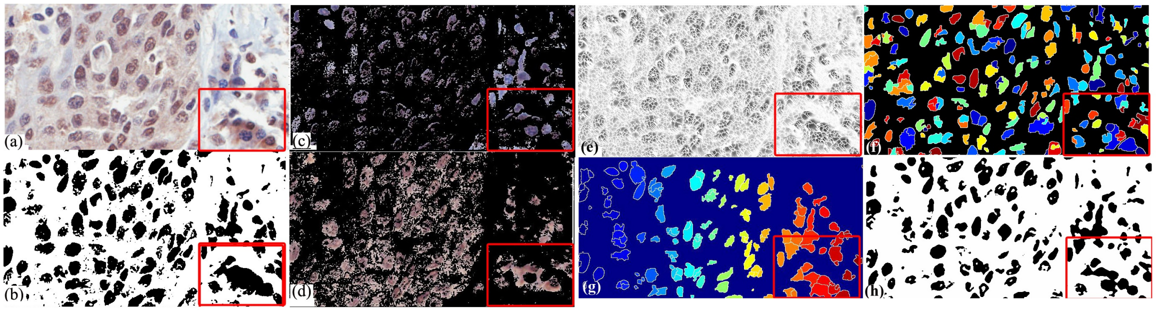

Figure 1: Performance comparison with K-means clustering, Otsu clustering, the presented method: (a) IHC Lung carcinoma image stained with HSSB1, (b) poor segmentation by Otsu unsupervised clustering, which automatically separates the image into two classes but contains a lot of false detection, (c)(d) poor segmentation by K-means clustering, which automatically separates the image into three classes here but the resulting clusters are poor in nuclear segmentation, (e) over-segmentation by Vincent-Soille watershed algorithm Vincent and Soille (1991) (f) poor segmentation result by marker-controlled watershed method Matlab (g) segmentation result with many false positives by optimized watershed transformation (adapted from Eddins (2006)) (h) improved nuclear segmentation by the proposed entropy-based method

Johnston Ruifrok and Johnston (2001) presented a color deconvolution approach to extract independent Haematoxylin and DAB stain contributions from individual IHC images using orthonormal transformation of RGB. In segmenting HSSB1 lung carcinoma tissue image, we use the independent Haematoxylin stains as the tissue architecture information and an entropy-based multistage segmentation model is applied to the tissue architecture information for identifying regions of nuclei. The experimental results show that the proposed method is promising in nuclear segmentation on IHC HSSB1 lung carcinoma tissue images.

The outline of this paper is as follows. The automated nuclei detection method is introduced in section 2, and the experimental results are displayed in section 3. The paper is concluded with future work in section 4.

2. Methods

2.1. Extract Tissue Architecture Information

Given an IHC image, we first separate independent DAB and Haematoxylin stain contributions by the color deconvolution approach Ruifrok and Johnston (2001). In this study, the normalized optical density (OD) matrix, M, to describe the colour system for orthonormal transformation is defined as follows:

$$M = \left(\begin{array}{ccc|c} \text{R} & \text{G} & \text{B} & \\ 0.65 & 0.704 & 0.286 & \text{Haematoxylin} \\ 0.072 & 0.99 & 0.105 & \text{Eosin} \\ 0.268 & 0.57 & 0.776 & \text{DAB} \end{array} \right) \tag{1}$$

Given C is 3×1 vector for amounts of the three stains at a particular pixel, the vector of OD levels detected at that pixel is equal to $L = CM$. Therefore, multiplication of the OD image with the inverse of OD matrix results in orthogonal representation of the stains forming the image ($C = M^{-1}L$), and hence colour de-convolution matrix is defined as:

$$D = M^{-1} = \left(\begin{array}{ccc|c} \text{R} & \text{G} & \text{B} & \\ 1.8801 & -0.0736 & -0.5952 & \text{Haema.} \\ -1.0172 & 1.1353 & -0.4826 & \text{Eosin} \\ -0.5553 & -0.1265 & 1.5733 & \text{DAB} \end{array} \right) \tag{2}$$

The extracted Haematoxylin OD image is then used as tissue architecture information and applied with the multistage entropy-based segmentation method.

2.2. Multistage Entropy-Based Segmentation of Nuclei

All statistical operations are performed on the normalized image histogram, $P = \{p_0, ..., p_{2^c-1}\}$ where the valid intensity scales from 0 to $2^c - 1$, and image entropy E(P) is calculated using discrete histogram P as follows.

$$H(A) = -\sum_{i=0}^{j} p_i \log p_i \tag{3}$$

$$H(B) = -\sum_{i=j}^{2^c-1} p_i \log p_i \tag{4}$$

$$H_j = -\log P(A) - \log P(B) - \frac{H(A)}{P(A)} - \frac{H(B)}{P(B)} \tag{5}$$

where $j \in \{0...2^c - 1\}$, $A = \{0...j\}$ and $B = \{2^c - 1...j\}$.

The entropy maximum is calculated as $\max H(P)$, which defines the cut-off point j for assigning image pixels into different classes where $H(P) = \{H_0...H_{2^c-1}\}$.

After calculating 2D image histogram entropy function, we first apply an eight stage maximum entropy function to automatically separate input image into eight layers, and then a two stage entropy function to extract potential regions of nuclei, which is then processed by morphological operations to produce final nuclear segmentation results. The algorithm is described below.

- divide histogram into four equal sub-histograms P_1, P_2, P_3, P_4, obtaining j_1, j_3, j_5 where $j \in 0...2^c - 1$
- compute maximum entropy points j_0, j_2, j_4, j_6 for the four different P intervals, where $j_0 = \arg\max H(P_1), j_2 = \arg\max H(P_2), j_4 = \arg\max H(P_3), j_6 = \arg\max H(P_4)$.

Table 1: Pixel-based quantitative performance evaluation

	Accuracy	TP rate	FP rate	FN rate	TN rate	Precision
Aver.	0.87	0.75	0.05	0.25	0.95	0.92

- use $j_0 ... j_6$ to categorize input image into eight layers
- calculate new histogram P^*
- compute $j^* = \arg\max H(P^*)$ and categorize input image into 2 categories, including nuclei and non-nuclei.
- apply the morphological operations described below

The purpose of the morphological function is both to reduce spurious false positive detection and increase low contrast true negative detection. The method re-assigns each image pixel value using the most frequent intensity level within its neighborhood. Given an image $I(X, Y)$ and neighborhood radius r, the output image $I'(X, Y)$ is formulated as follows.

$$I'(x, y) = \arg\max_I(\#I(K, L)) \tag{6}$$

where $K = \{x - r, ..., x + r\}, L = \{y - r, ..., y + r\}$, and r is empirically set as 3.

3. Results

We stained HSSB1 Richard et al. (2008) on a TMA with 150 tissue cores (image dimension = 107248 × 67918 = 20.4Gb), and the types of lung cancer that were contained in the arrays are two major sub-types of non-small cell lung cancers, including adenocarcinoma and squamous carcinoma. The TMA slide was scanned using Aperio Scanscope CS2 (Aperio Technologies Inc. San Diego USA), at 40× objective magnification. Nine different tissue cores were randomly selected for evaluation; the image size of individual tissue cores is around 2896 × 2756 and the nuclear areas of each tissue core were manually marked to produce ground truth data. A quantitative performance evaluation was conducted by comparing the ground truth data and the system output (Table 1). The system achieves 92% precision and 75% recall rates and has been demonstrated to be promising in nuclear cell detection on HSSB1 lung tissue images. In applications, the extracted regions of nuclei can then be used to analyze nuclei malignancy; it can be used to quantify the percentage or intensity levels of positively stained nuclei as shown in Fig.2.

4. Conclusion and Future Work

We have presented an unsupervised entropy-based system to detect nuclei in HSSB1 IHC lung tissue slides. The method has shown to perform well in image segmentation in the experiments. Furthermore, the extracted nuclei information is demonstrated to be useful in

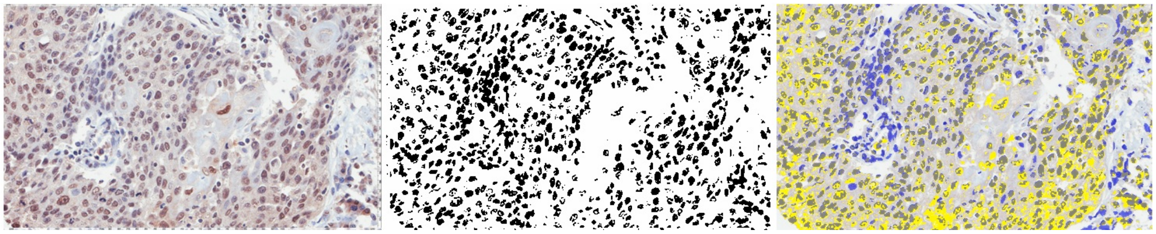

Figure 2: Extracted nuclei (middle) can be used for quantitative IHC analysis such as the percentage or intensity levels of positively stained nuclei (right), where gray areas are positively nuclei and blue areas are negatively stained nuclei.

quantitative IHC. In addition, apart from analyzing nuclei activity, we would like to extend the method for cancer subtypes classification in lung cancer. We plan to utilize the identified nuclei architecture information for automated classification of cancer subtypes. Fig.3 shows two types of lung carcinomas, including adenocarcinoma and squamous carcinoma. The characteristic histologic feature of Adenocarcinoma is glandular structure (Fig.3(a)), where nuclei form snake-like shapes. On the other hand, the characteristic feature of squamous carcinoma(Fig.3(b)) is sheet-like structure. Hence, after obtaining nuclei architecture information (Fig.3(c,d)) by the proposed method, scene abstraction function can be applied (Fig.3(e,f)) to remove isolated or small islands of nuclei. In future work, we plan to utilize the extracted connecting components (Fig.3(g,h)) as patterns to distinguish the two non-small cell lung cancers by detecting large regions of connecting components as glandular structures and recognizing the tissue slides as adenocarcinoma cases.

References

S Avninder, K Ylaya, and SM Hewitt. Tissue microarray: a simple technology that has revolutionized research in pathology. *J Postgrad Med*, 54(2):158–62, 2008.

RC Jr Bast, H Lilja, N Urban, DL Rimm, H Fritsche, J Gray, R Veltri, G Klee, A Allen, N Kim, S Gutman, MA Rubin, and A. Hruszkewycz. Translational crossroads for biomarkers. *Clin Cancer Res*, 11(17):6103–8, 2005.

EM Brey, Z Lalani, C Johnston, M Wong, LV McIntire, PJ Duke, and CW Patrick. Automated selection of dab-labeled tissue for immunohistochemical quantification. *J Histochem Cytochem*, 51(5):575–84, 2003. ISSN 0022-1554.

RL Camp, GG Chung, and DL Rimm. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nature Medicine*, 8(11):1323–7, 2002.

S Eddins. Cell segmentation. http://blogs.mathworks.com/steve/2006/06/02/cell-segmentation/, 6 2006.

Y. Li, E. Bolderson, R. Kumar, P. Muniandy, Y. Xue, D. Richard, M. Seidman, T. Pandita, K. Khanna, and W. Wang. hssb1 and hssb2 form similar multiprotein complexes that

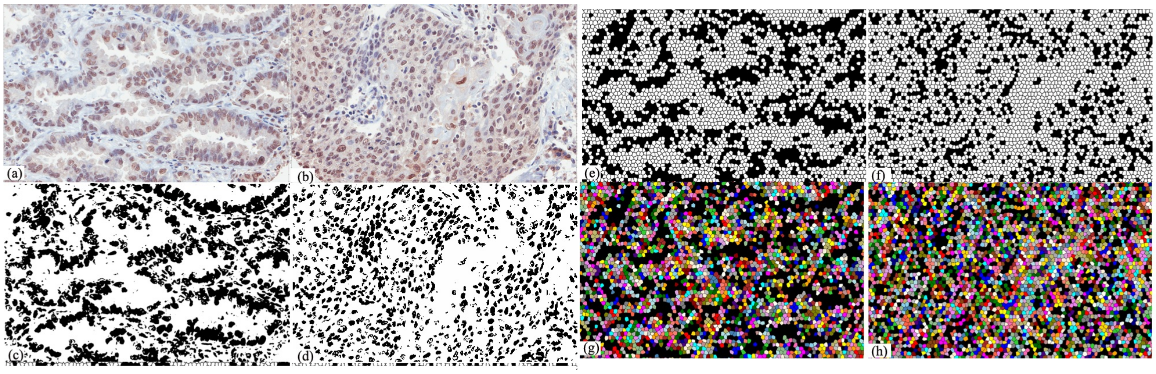

Figure 3: Cancer subtype classification based on nuclear layout patterns: (a) adenocarcinoma tissue image with snake-like glandular structure, (b) squamous carcinoma tissue image with sheet-like structure, (c,d) extract nuclei architecture, (e,f) scene abstraction to remove isolated or small islands of nuclei, (g,h) detect connecting components and look for large regions of connecting components as glandular structure for pattern recognition of adenocarcinoma.

participate in dna damage response. *The Journal of Biological Chemistry*, 284:23525–23531, 2009.

S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

K. Mao, P. Zhao, and P.-H. Tan. Supervised learning-based cell image segmentation for p53 immunohistochemistry. *IEEE Trans. Biomedical Engineering*, 53(6):1153–1163, 2006.

Matlab. Image processing toolbox 7.0: Marker-controlled watershed segmentation.

N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9:62–66, March 1979. minimize inter class variance.

E. Rexhepaj, D.J. Brennan, P. Holloway, E.W. Kay, A.H. McCann, G. Landberg, M. Duffy, K. Jirstrom, and W. Gallagher. Quantification of histochemical staining by color deconvolution. *Breast Cancer Research*, 10(5), 2008.

D. J. Richard, E Bolderson, L Cubeddu, RI Wadsworth, K Savage, GG Sharma, ML Nicolette, S Tsvetanov, MJ McIlwraith, RK Pandita, S Takeda, RT Hay, J Gautier, SC West, TT Paull, TK Pandita, MF White, and KK Khanna. Single-stranded dna-binding protein hssb1 is critical for genomic stability. *Nature*, 453:677–681, 2008.

AC Ruifrok and DA. Johnston. Quantification of histochemical staining by color deconvolution. *Anal Quant Cytol Histol*, 23:291–299, 2001.

T. Seidal, A. J. Balaton, and H. Battifora. Interpretation and quantification of immunostains. *The American journal of surgical pathology*, 25(9):1204–1207, September 2001. ISSN 0147-5185. URL `http://view.ncbi.nlm.nih.gov/pubmed/11688582`.

L Vincent and P Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans.Pattern Anal.Mach.Intell*, 13(6):583–598, 1991.

H.-S. Wu, L. Deligdisch, and J. Gil. Segmentation of microscopic nuclear image - a review. *Recent Res. Devel. Electronics*, 2, 2004.