Computational Immunohistochemistry: Recipes for Standardization of Immunostaining

Nuri Murat Arar[1,3], Pushpak Pati[2,3(✉)], Aditya Kashyap[3], Anna Fomitcheva Khartchenko[3], Orcun Goksel[2], Govind V. Kaigala[3], and Maria Gabrani[3]

[1] Signal Processing Laboratory (LTS5), EPFL, Lausanne, Switzerland
[2] Computer-Assisted Applications in Medicine, ETH Zürich, Zürich, Switzerland
[3] IBM Zürich Research Lab, Zürich, Switzerland
patip@student.ethz.ch

Abstract. Cancer diagnosis and personalized cancer treatment are heavily based on the visual assessment of immunohistochemically-stained tissue specimens. The precision of this assessment depends critically on the quality of immunostaining, which is governed by a number of parameters used in the staining process. Tuning of the staining-process parameters is mostly based on pathologists' qualitative assessment, which incurs inter- and intra-observer variability. The lack of standardization in staining across pathology labs leads to poor reproducibility and consequently to uncertainty in diagnosis and treatment selection. In this paper, we propose a methodology to address this issue through a quantitative evaluation of the staining quality by using visual computing and machine learning techniques on immunohistochemically-stained tissue images. This enables a statistical analysis of the sensitivity of the staining quality to the process parameters and thereby provides an optimal operating range for obtaining high-quality immunostains. We evaluate the proposed methodology on HER2-stained breast cancer tissues and demonstrate its use to define guidelines to optimize and standardize immunostaining.

1 Introduction

Immunohistochemistry (IHC) is an invaluable tool for cancer diagnosis, treatment selection, and research, owing to rapidly obtainable tissue profiles. It is widely used with different biomarkers for the identification of prognosticators of cancer progression. IHC localizes specific proteins (antigens) in tissues by exposing them to the corresponding antibodies. The antigen-antibody binding reaction generates a visual signal, whose intensity is a function of a number of parameters, including the antigen density on tissue, antibody concentration, residence (incubation) time, and tissue-preprocessing methods. Assessment of the generated visual signal conveys vital information for a patient. If the signal originates from a 'low-quality' stain, the results are unreliable – with potential dire consequences, such as false diagnosis and/or ineffective treatment. Therefore, in

© Springer International Publishing AG 2017
M. Descoteaux et al. (Eds.): MICCAI 2017, Part II, LNCS 10434, pp. 48–55, 2017.
DOI: 10.1007/978-3-319-66185-8_6

cancer diagnosis and treatment the quality of the visual signal plays as significant role as its assessment. Although immunopathology has been extensively studied and used for decades, more emphasis has been given to the antigen-antibody reactions and the assessment of the resulting signal, whereas the signal quality has generally been neglected.

Pathology laboratories use different staining process parameters to achieve a 'high-quality' stain, defined by manual assessments over parameter grids. Such manual effort with trial-and-error experiments is tedious and tissue exhaustive. Consequently, the lack of parametric standardization and reproducibility of the staining quality remain major concerns in IHC. According to *NorqiQC*'s statistics, about 20% of breast cancer IHC stains and about 30% of general cancer IHC stains have been assessed as *insufficient for diagnostic use* [1]. Indeed, tuning the parameters of the staining process is one of the crucial elements for improving the staining quality. However, the effect of the process parameters on the quality is difficult to deconvolve, since measurements on the exact same tissue location with different parameters cannot be acquired. Hence, strategies for automatic analysis of staining quality sensitivity to process parameters and for contextual quantitative analysis by using only a limited amount of tissue samples are important towards improving standardization in IHC staining. These standardized tests will enable novel avenues of tissue evaluation relevant to pathologists.

Several works in literature have performed quantitative analysis of IHC-stained tissues. Most of these (e.g. [2,3]) emphasize the quantification of biomarker expressions. To a large extent, the results show agreement between image-analysis based methods and pathologists' visual examination. However, there are only few studies that focus on the assessment of staining quality. For instance, [4] proposed quality indicators (i.e., signal intensity, tissue integrity, image integrity) for IHC-stained tissues to quantify the staining quality around predefined thresholds. In [5], a reference-based technique was described in which some quality indicators are computed for both a test specimen and a reference specimen, prepared at a standardized laboratory. Subsequently, a relative quality measure is computed using the cumulative distance between these indicators. This reference-based approach was validated in [6] using membrane connectivity as the quality measure. The drawback of the aforementioned approaches is that they quantify the staining quality either with respect to a user-defined threshold or a reference specimen, which involves an expert's intervention. To the best of our knowledge, there are no efforts in the literature that analyze the sensitivity of the staining quality to IHC process parameters with the aim of optimizing the staining quality for process parameters. In this work, we propose a novel, automated, principled method to assess the sensitivity of the staining quality to IHC process parameters – a major step towards the standardization of immunostaining.

2 Proposed Methodology

Our methodology includes an automated signal and noise segmentation algorithm, followed by a novel no-reference staining quality metric learning

technique. Accordingly, it has three main components: (i) image segmentation and representation, (ii) staining quality metric learning, and (iii) sensitivity analysis to process parameters. The components are validated through several experiments in which HER2 is selected as the biomarker of interest owing to its high clinical relevance in breast cancer diagnosis. HER2 overexpression provides insights for diagnosis and hints for a targeted therapy. As it is a transmembrane receptor, the quantification of overexpression can be modelled as 'peaks' (cell membranes) versus 'valleys' (cell's cytoplasm and stroma) detection. This is used as a guiding principle and a starting point for the tissue-based staining-quality evaluations introduced in this work.

2.1 Image Segmentation and Representation

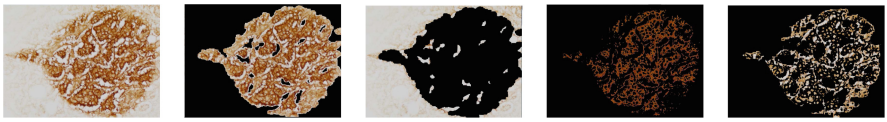
Staining quality is directly proportional to the signal-to-noise ratio, where the signal is the staining of membrane (*foreground*), and the noise is the staining of cytoplasm, nucleus, and stroma inside footprint (*background*). Therefore, our methodology starts with the detection and separation of signal and noise in IHC-stained tissue images. We propose a fully-automatic segmentation algorithm that deconvolves an image into four regions: (i) *footprint*, (ii) *off-footprint*, (iii) *foreground*, and (iv) *background*, as illustrated in Fig. 1 for a HER2-stained tissue.

We begin by finding and delineating the localized *footprint*, where a vertical microfluidic probe [7] is applied, using a combination of Otsu thresholding and the non-parametric marker-based Watershed algorithm; cf. Fig. 1(b–c). Subsequently, we segment the *foreground* within the *footprint*, as it is the region of interest, using a global thresholding. A robust threshold is determined from a 16-bin intensity histogram within the *footprint*, and the value is set as the mean of the most frequent and the maximum intensity values. Then, we extract the *background* by taking the difference of the *foreground* and the dilated foreground mask. We preserve the connectivity of the *background* by performing a morphological closing operation. Thereon, we subtract the foreground mask from the *background* to ensure that there are no remaining *foreground* pixels in it.

Global and local features are extracted to define representative signatures for the images. Global features include intensity features that are extracted from the individual segmented regions. Local features are extracted from patches within the *foreground* to capture local structural and morphological information around cells. Local features include the gray level co-occurrence matrix-based texture features, rotation and scale invariant Gabor wavelet features, and Haralick features from the dual-tree complex wavelet transform. Finally to construct a fixed-dimensional signature per image, we compute the feature-wise mean across local patches and then concatenate them to global intensity features.

2.2 Staining Quality Metric (SQM) Learning

The definition of high-quality staining varies for different tissue types (TT) and protein expressions [8], as depicted by the 'peaks' and 'valleys' guiding principle in Fig. 2. A protein overexpression exhibits distinct peaks and valleys, with



(a) (b) (c) (d) (e)

Fig. 1. Segmentation outputs of a sample HER2-stained tissue: (a) input image, (b) *footprint*, (c) *off-footprint*, (d) *foreground*, and (e) *background*.

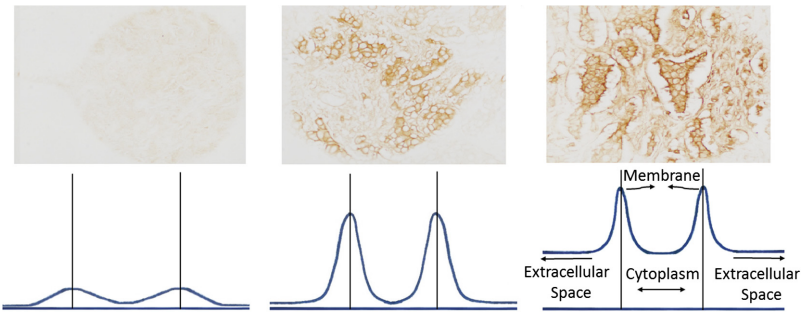


Fig. 2. 'High-quality' HER2-stained tissues for HT (left), PT (middle), MT (right), and the respective intensity profiles of the cross-sectional view of a cell.

the distinction decreasing with the expression level. For instance, a high-quality staining of a HER2 overexpressed *primary breast tumor* tissue (PT) and of a *lymph node metastasis* tissue (MT) would entail a high level of distinction, whereas a high-quality staining of a healthy tissue (HT) with low HER2 expression would have an absent or a low level of distinction. Therefore, we propose to learn independent TT-specific SQMs, which are classifiers evaluating the staining quality for the corresponding TTs. We design the SQM through evaluation and combination of several quality indicators (QI), as shown in Fig. 3. In this work, we define two QIs acquired via probabilistic classifiers. The first QI comprises tissue-discrimination probabilities to indicate the informativeness of the staining towards expressing the TT categories. The second QI conveys the signal-to-noise contrast level (CL) to suggest a sample's degree of agreement with the expected signal-to-noise CL for each TT category. The QIs for a sample with a feature signature $\boldsymbol{\mathcal{X}}$ can be denoted as $P(\boldsymbol{\mathcal{X}}_1 = TT_i), TT_i \in \{PT, MT, HT\}$ and $P(\boldsymbol{\mathcal{X}}_2 = CL_i), CL_i \in \{High, Low)\}$. TT labels for the training samples are obtained from the tissue provider, whereas CL labels are determined by the consensus of three experts.

SQM Learning and Quality Assessment: Given the two QIs for the samples, TT-specific SQMs are trained independently by using the samples corresponding to the respective TT categories. For example, the QIs obtained for the samples of the PT category are used to model SQM_{PT}. For training the

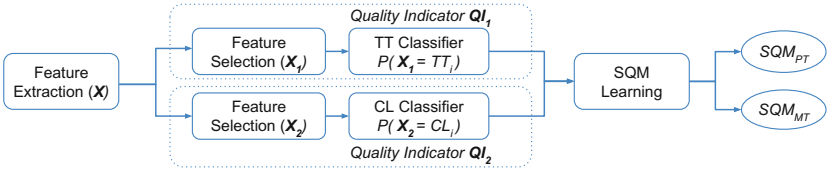


Fig. 3. Overview of the SQM learning.

SQMs, TT-specific quality labels $\in \{Acceptable, NotAcceptable\}$ are obtained from the experts. The output probability of belonging to the *Acceptable* quality class, given by a TT-specific SQM, is considered to represent the sample's overall quality value (QV) for the respective TT category. The QV of a sample belonging to a TT category is expected to reflect the underlying staining quality on a quantitative scale. For instance, considering the staining-quality expectations for a PT tissue type, if the first QI of a given sample conveys a high $P(\boldsymbol{\mathcal{X}} = PT)$ probability, and the second QI expresses a high signal-to-noise contrast level $P(\boldsymbol{\mathcal{X}} = High)$, then using the SQM_{PT} is expected to result in a high QV for the sample.

2.3 Analysis of Sensitivity to Process Parameters

In the final step of the methodology, we analyze the sensitivity of the staining quality to the variations in the IHC staining-process parameters, i.e., the antibody concentration (C) and the residence time (RT). The aim is to estimate the optimal operational range for the parameters for obtaining high-quality stains. As the quality expectations differ across TTs, we hypothesize that the optimal process parameters generating high-quality stains may also vary across TTs. Thus, we perform the sensitivity analysis independently for each TT. We compute QVs for all samples belonging to a TT category by using the respective SQM. The QVs obtained are distributed over a range of C and RT values analytically and experimentally. To obtain a comprehensive visualization and evaluation of the sensitivity of QV, we interpolate QVs at intermediate C and RT configurations, and fit a smooth 3D manifold for the QVs.

Similarly to the variational quantification approach in [9], we measure the sensitivity of QV per TT at all possible configurations over the entire C and RT parameter space. For a given point on the 3D manifold, denoted by $p_i = (RT_i, C_i, QV_i)$, we quantify the variations in 8-neighboring directions. The variations are measured by the eigenvalues of a covariance matrix at p_i, where the covariance matrix is generated from the QV differences between p_i and its 8-connected neighboring points. We consider the "maximum eigenvalue" to quantify the sensitivity at p_i, as it indicates the direction and the degree of maximum variation at that point. A higher "maximum eigenvalue" at a point conveys a high degree of variation, implying a high sensitivity of the staining quality to slight variations in parameters around that point. We project the sensitivity values onto a 2D contour map to visualize and estimate the stable operating configurations per TT. For a TT category, we combine the information from the

staining quality 3D manifold and the 2D sensitivity map to determine the optimal range for the parameters. Thereby, we aim at maximizing the acceptable staining quality and minimizing the sensitivity in parametric operation.

3 Experimental Validation

For evaluating the potential of the proposed methodology, we collected stained samples of 36 cores across 19 unique patients for three different tissue types on tissue micro-arrays (TMA). To increase the accuracy of the true-positive stain and to minimize tissue usage, we applied microimmunohistochemistry (μIHC) using vertical microfluidic probe [7] for the staining. The TMA cores were stained for HER2 protein with three antibody dilutions $C = \{6.25, 12.5, 25\}$ μg/mL. Each core was patterned with eight footprints of increasing residence time $RT = \{12, ..., 289\}$ s, generating 288 IHC-stained samples in total. As first step, we performed signal-to-noise separation, followed by comprehensive feature extraction for all samples in the dataset.

SQM Learning: The first step in learning the SQMs is to generate an automatic TT classifier. For the classifier training, we selected a balanced subset of 160 samples from the complete dataset. The subset consisted of samples across all C and RT values. They contained sufficient cells and represented their TT class in terms of both poor (insufficient and over-staining) and high-quality staining. In the training phase, the optimal set of 73 features $\boldsymbol{\mathcal{X}}_1$ for the TT classification were chosen by Random Forest (RF) feature importance measure out of a complete set $\boldsymbol{\mathcal{X}}$ of 353 features. For designing the best TT classifier, we experimented with different hyperparameters, such as patch size, feature combination, and classification algorithms. The classifiers were compared using accuracies computed through a 10-fold cross-validation on the training data. The best accuracy of **0.82** was obtained using an Support Vector Machine (SVM) with RBF kernel classification algorithm, which was modeled using features extracted on patches of in size 64×64 pixel. Similarly, for the second quality indicator, we trained a signal-to-noise CL classifier on an independent subset of 77 samples across all TT categories. The samples selected clearly represented high and low contrast levels irrespective of their TT. We achieved the best classification performance of **0.95** using SVM with RBF kernel with 63 RF-selected features extracted on patches of in size 64×64 pixel. SQMs were learned for individual TTs by computing classification probability maps for the quality indicators. ROC curves and AUC values were computed for evaluating the performance of the SQMs learned. SQM_{PT} and SQM_{MT} achieved **0.83** and **0.90** AUC scores, respectively, indicating a good separability of *Acceptable* and $NotAcceptable$ staining-quality classes. In this work, we evaluated only for HER2 overexpressive PT and MT types, as they are of higher importance than HT in cancer diagnosis.

Analysis of Sensitivity to Process Parameters. Each SQM provides staining-quality scores, QVs, for the samples belonging to the corresponding

TT category. We interpolated QVs for the entire range of C and RT, and fitted 3D manifolds for each SQM. Afterwards, the eigenvalue-based variational quantification approach was used to inspect the sensitivity of the staining quality. Considering the range of both parameters, we analyzed the variation in C and RT within $\pm 1\ \mu g/mL$ and $\pm 5\,s$ respectively. Thereon, the 2D staining-quality contour maps and the sensitivity maps for each SQM were plotted.

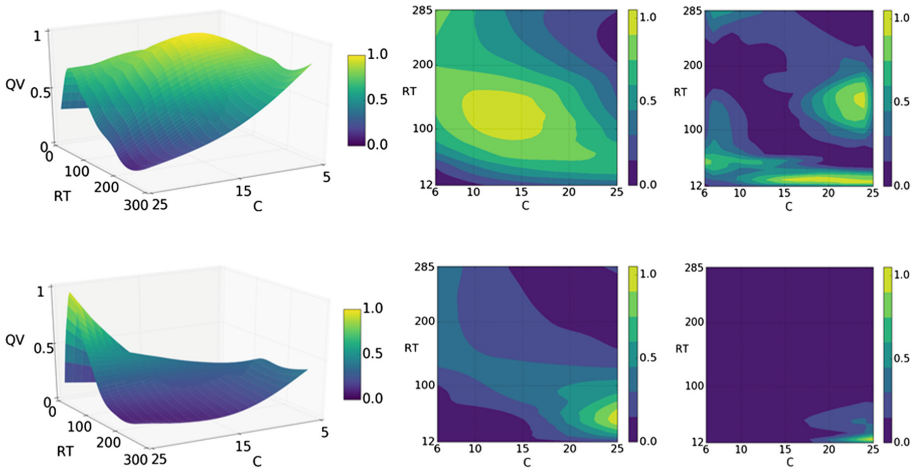


Fig. 4. 3D SQM manifolds (left), 2D staining quality maps (middle), and sensitivity maps (right) for PT (top row) and MT (bottom row).

Figure 4 displays the SQM manifolds, the staining-quality and sensitivity contour maps for both the PT and MT tissue categories. For PT, the SQM manifold shows that high-quality staining can be obtained when operating in the range of $9 < C < 17\,\mu g/mL$ and $85 < RT < 160\,s$. It also illustrates the variation in the staining quality, i.e., that the staining quality is low for low-end and high-end C and RT values. These observations align with the concepts of insufficient and over-staining, respectively, which cause a decay in staining quality. Therefore, observations from the quality map can aid in reducing false negative and false positive staining. In addition, the sensitivity map indicates the stability of the staining quality for a given parameter configuration. It shows that the staining quality is slightly sensitive towards the lower end and the upper end of the aforementioned range of C values. Combining the knowledge from both the maps, an operational range of $11 < C < 15\,\mu g/mL$ and $85 < RT < 130\,s$ can be selected for generating stable and high-quality stains. In a similar analysis for MT, it can be observed that the high staining-quality range is highly unstable. Thus, a more stable operating range should be selected from the sensitivity map, with a small compromise in the staining quality. To verify the robustness of the optimal parameters computed, 95% staining-quality confidence ellipses were evaluated for 500 bootstrap datasets, which resulted in consistent performance and behavior across all datasets.

4 Conclusions

In this work, we have proposed a framework that uses visual computing and machine learning techniques to address a prevalent challenge in IHC. We first devised an automatic methodology for the quantification of the staining quality in IHC-stained images. Then we introduced a tool for standardizing the IHC process parameters via automatic determination of the operating bounds. The proposed framework was applied to HER2-stained breast cancer tissues, with promising results achieved in the experiments conducted. The computationally extracted results were validated against subjective expert opinions, with the staining behavior found in line with underlying biology of the tissue type. Furthermore, a quantitative evaluation of these results led to the development of SQMs, reducing subjectivity and uncertainty of such staining, while defining operational parameters that lead to high-quality stains. Inclusion of further quality indicators and availability of more stained tissue specimens will undoubtedly refine our system. These promising results show the potential of our approach towards the standardization of immunostaining using automatic process-parameter optimization in IHC, but most importantly, for reducing uncertainty in cancer diagnosis and treatment selection.

References

1. Vyberg, M., Nielsen, S.: Proficiency testing in immunohistochemistry - experiences from Nordic Immunohistochemical Quality Control (NordiQC). Virchows Arch. **468**, 19–29 (2016)
2. Brügmann, A., Eld, M., Lelkaitis, G., Nielsen, S., Grunkin, M., Hansen, J.D., Foged, N.T., Vyberg, M.: Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. Breast Cancer Res. Treat. **132**, 41–49 (2012)
3. Laurinaviciene, A., Dasevicius, D., Ostapenko, V., Jarmalaite, S., Lazutka, J., Laurinavicius, A.: Membrane connectivity estimated by digital image analysis of HER2 immunohistochemistry is concordant with visual scoring and fluorescence in situ hybridization results: algorithm evaluation on breast cancer tissue microarrays. Diagn. Pathol. **6**, 87 (2011)
4. Pinard, R., Tedeschi, G.R., Wang, D., Williams, C.: Methods and system for validating sample images for quantitative immunoassays. US Patent 8160348 (2009)
5. Grunkin, M., Hansen, J.D.: Assessment of staining quality. International Patent WO2015135550 (2015)
6. Brügmann, A., Grunkin, M., Nielsen, S., Jensen, V., Heikkila, P., Gaspar, V., Vyberg, M.: Image analysis of breast cancer HER2 protein expression used in assessment of staining quality. Virchows Arch. **465**, S20 (2014)
7. Kaigala, G.V., Lovchik, R.D., Drechsler, U., Delamarche, E.: A vertical microfluidic probe. Langmuir **27**, 5686–5693 (2011)
8. Hoda, S.A., Brogi, E., Koerner, F.C., Rosen, P.P.: Rosen's Breast Pathology. Wolters Kluwer, UK (2014)
9. Seguin, B., Saab, H., Gabrani, M., Estellers, V.: Estimating pattern sensitivity to the printing process for varying dose/focus conditions for RET development in the sub-22nm era. In: Proceedings of SPIE, vol. 9050 (2014)