Computer-aided evaluation of protein expression in pathological tissue images

Elisa Ficarra, Enrico Macii
Politecnico di Torino
Dep. of Control and Computer Engineering
C.so Duca degli Abruzzi 24, 10129 Torino, Italy
elisa.ficarra@polito.it, enrico.macii@polito.it

Giovanni De Micheli
Ecole Polythecnique Federale de Lausanne (EPFL)
Integrated Systems Laboratory (LSI)
IN-F Ecublens, 1015 Lausanne, Switzerland
giovanni.demicheli@epfl.ch

Luca Benini
University of Bologna
Dep. of Electronics, Computer Science and Systems
V.le Risorgimento 2, 40136 Bologna, Italy
lbenini@deis.unibo.it

Abstract

This work presents the first fully-automated computer-aided analysis approach to the quantification of the expression of receptors for the non-small cell lung carcinoma. This immunohistochemical analysis is usually performed by pathologists via visual inspection of tissue samples images. Our techniques streamlines this error-prone and time-consuming process, thereby facilitating analysis and diagnosis. Experimental results on several real-life datasets demonstrate the high quantitative precision of our approach.

1 Introduction

Direct monitoring the activity of proteins involved in the genesis and development multi-factorial genetic pathologies is a very useful diagnostic tool. It leads to classify the pathology in a more accurate way, through its particular genetic alterations, and to create new opportunities for early diagnosis as well as to provide information in future strategies for therapy.

The *EGFR/erb-B* family of receptors plays an important role for *non-small cell lung carcinoma* (NSCLC) development. Quantifying and classifying the *EGFR* expression and activity in NSCLC with special regard to the assessment of the prevalence of somatic *EGFR* mutations, as well as to ligand-receptor interactions, could lead to new insights into the modulation of *EGFR* in individual lung carcinomas. Thus, it is important to extract these information by using methodologies that give quantifiable, standardized and precise measurements [1].

An approach for monitoring and quantifying the activity of proteins is to analyze their localization and the intensity of their activity in pathological tissues by using, for example, images of the tissue where the localization of proteins, as well as their ligands, is highlighted by fluorescent-marked antibodies that can detect and link the target proteins. The antibodies are marked with a particular stain. The protein activity intensity is related to the intensity of the stains. This procedure is called *immunohistochemistry* (IHC). Figure 1.a shows an example of immunohistochemical image of lung cancer tissue.

What is interesting to extract from these images is not a specific coloured area, that is almost the standard procedure with this kind of images [2][3]. Rather, the focus is cell by cell localization of the coloured areas in particular cellular regions (i.e. membranes or cytoplasm or nuclei). Similarly, the quantification of the percentages of coloured areas at the location of interest is important because it relates to the activity of specific receptors. In other words, it is important

Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)
0-7695-2517-1/06 $20.00 © 2006 IEEE

to quantify if the proteins have or not a *membrane* activity (or *cytoplasm* or *nucleus* one), how much of that membrane is positive for the specific protein activity and, vice versa, if it is not active.

This type of analysis aims at characterizing each pathological cell, and in average the whole tissue, by performing a standardized quantitative and qualitative measurement of protein activations. Moreover, if this information is accurate and objective it can be correlated with the genetic expression data on same immunohistochemical tissue in order to better define a group of potential candidates to protein family-inhibiting therapy.

In this paper we describe a fully-automated procedure that provides standardized measures of protein activities, and related ligands, involved in the development of a pathology. This goal is reached *i)* by identifying different cellular regions, *ii)* quantifying the percentage of active areas with respect to each whole region, *iii)* classifying protein reactions according to the specific region and *iv)* quantifying the intensity of the protein activity. These analyses have traditionally been performed directly by pathologists in a very subjective and time-consuming way. The major contribution of this research is to provide an automated, fast and precise means for performing immunohistochemical image analysis. To the best of our knowledge the methodology presented in this paper is the first completely automated approach to this purpose.

Much previous work in biomedical image processing focused on automated methods for segmentation of nuclei and cells [4][5][6][7]. Classical approaches, such as active contours or watersheds, are not effective when the objects to be identified lack specific geometrical features or gradient variations. Unfortunately, these critical conditions are very common in the images targeted by our work. For example, the methods in [4][6] aim at detecting and segmenting cells in the blood using specific active contours algorithms. Both of these approaches are based on the evaluation of gradient magnitude along cell boundary and shape-based segmentation. Cancer tissue cells are characterized by not-predictable variations in shape that lead to a non-trivial determination of an effective approach based on shape-based segmentation. Moreover, in immunohistochemical cancer tissue images cells are not well separated and, in addition, they are usually not characterized by variations gradient magnitude.

To address these issues, we developed a novel deterministic fully automated approach for the quantification of protein activities and localization of molecular activities in tissue images. We focused on lung cancer pathologies where cells are characterized by unpredictable variations in shape. This method help in estimating in a quantitative way the modulation of specific protein families in individual pathologies and, thus, to better define appropriate therapies.

2 Method: Membranes detection and parameters extraction

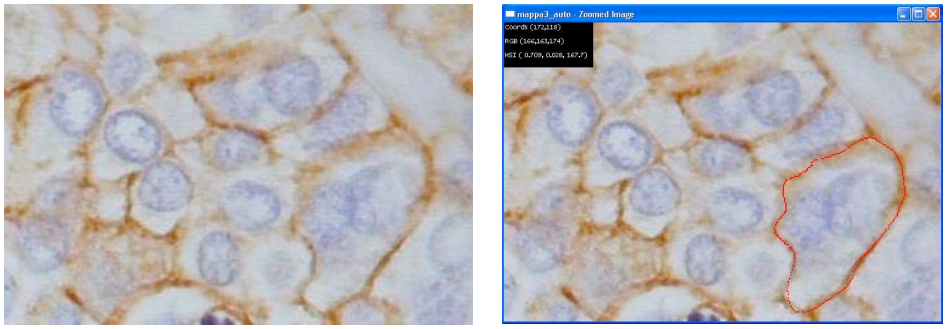
Immunohistochemical lung cancer tissue images are characterized by a blue stain as background colour and a brown stain where a receptor of the *EGFR* family is detected. We focus here on quantification of membrane receptor activity. Cell membrane segmentation is a hard problem because those membranes that are negative to the *EGFR* family of receptors, are generally not visible. In other words, they are not characterized by gradient magnitude variation. It is also possible that a cell has only some parts of its membrane positive to receptor activity.

The automated procedure is composed by several sequential steps, as outlined in the following subsections. In this work, our description concentrates on the steps we customized.

2.1 Virtual cell membrane detection

To reconstruct the cell membrane locations we first detected nucleus membranes using standard morphological segmentation approaches. For each nucleus, we detected seeds applying noise filtering, colour filtering to detect nucleus regions, artifacts removing, filling of connected components and boundaries detection. These first steps lead an approximate detection of nucleus boundaries. We used these nucleus boundaries as initial curves for the final detection of nucleus membranes. We completed the detection of nucleus membranes by applying the active contour algorithm presented on [8]. This algorithm was found very useful for nucleus membranes detection. Further details on seed detection and active contours are beyond the scope of this paper because they are obtained and implemented using standard approaches. The interested readers are directed to [8] and [6].

After detecting nucleus membranes, we implemented a procedure for *virtual cell membrane* detection. This is an important step in our approach. In fact, to perform membrane cell segmentation, we use virtual membranes as part of final-detected cell membranes in those regions that are negative to the *EGFR* family of receptors and that are as a consequence not characterized by gradient magnitude variation. Virtual cell membranes are computed as set of connected points equidistant from closest nucleus membranes. Since our analysis concerns cells in tissues, the assumption that cellular membranes are equidistant from closest nucleus boundaries is reasonable as first order approximation.



(a) (b)

Figure 1. a: example of lung cancer tissue immunohistochemical image; b: example of membranes detection, see big cell in the bottom-right part of the image

2.2 Color Filtering

To select the region that are positive to receptor activations, we filtered the image on Hue-Saturation-Intensity (HSI) colour space. We chose the HSI space because the stains we used are well defined in (HSI) space. In particular, looking at several Hue histograms of the tissue images, we noticed well-separated bi-modal value distributions. To separate the two distributions there are several standard thresholding algorithms that can be successfully employed, such as [9] [10] [11]. As expression of receptor activity we chose brown pixels with hue components minor than a threshold automatically computed by using *Ridler thresholding* as detailed in [12].

2.3 Cellular membrane detection

The detection of cellular membranes is done in two steps. Beforehand, we perform membrane segmentation in the brown areas one cell at a time and we connect them with the virtual cell membrane in those regions that are not characterized by receptor reaction. To this purpose, we developed an ad-hoc procedure, as described later in this section. The second step consists of a customized fitting procedure of the detected membrane points to complete the cellular membrane segmentation.

- *Connecting reactive membranes with virtual ones: Scanning procedure*: To connect brown areas with the virtual membrane in those regions where there was not receptor reaction, the area across the virtual membrane is dilated in order to be able to reach, if they exist, brown regions of the cell. The level of dilation is an input parameter and it depends on image resolution. We set this value to 18 (pixels) for images with a resolution of about 3nm. Then, we scan the dilated area with a scan line having one end on the center of the nucleus and the other one on the external border of dilated area.

  At each step, the points of the membrane are computed as weighted barycentre B of brown pixels among the scan line, as shown in Equation 1

$$B = \frac{\sum_j c_j I_j j}{\sum_j c_j I_j} \tag{1}$$

  where j is the coordinate on the scan-line. This coordinate is 0 on the virtual membrane, negative in the inner part of the dilated area and positive in the outer part. Ij is the value of the pixel jth and cj is a coefficient for barycentre computation. The coefficient cj is 1 for pixels on scan line negative coordinate while for positive coordinates the coefficient has a negative parabolic trend as function of coordinate j. In this way, when a brown region branches off, the scanning procedure is forced to choose as points belonging to the membrane those pixels that lie on the path closest to the nucleus. Moreover, we assigned to pixels of the scan-line the value of 1 if they belong or precede to the virtual membrane pixels. This has been done when there are not brown pixels in the scan-line, to choose as points belonging to membrane those pixels that are close to the virtual membrane. Finally, we set to 0 the pixels that are neither brown nor virtual membrane ones.

- *Fitting and complete membranes detection*: To complete the detection of cellular membranes, we implemented an iterative fitting procedure in which *outlier* pixels are deleted at each step. We defined outliers pixels the pixels located far away from the fitting line more than three-times the standard deviation. An example of membrane detection is shown in Fig. 1.b

2.4 Clinical parameter computation

We quantify the activity of membrane *EGFR* receptors through the computation of percentage of active areas with respect to each whole membrane region. Then, the final parameter is the average value of all single-cell parameters on the image.

3 Experimental Results

We tested the algorithm on three data sets. All of them were real lung cancer tissue immunohistochemical images. The three data sets present positive reactions at the *EGF-R* receptor activation. These reactions are localized in the cellular membranes. The three data sets differ because of different levels of positivity intensity.

For each data set, we first localized each cellular membrane in the image, as described in Sec. 2. Afterwards, we computed for each cell the percentage of area characterized by positive activation of receptor *EGF-R* with respect to the whole cellular membrane surface. At the end, we computed the final parameter as average value of all single-cell parameters on the image. This final parameter is the clinical parameter that characterizes the percentage of receptors that is active in the lung cancer tissue.

In order to evaluate the performance of our approach, positive protein reaction parameters have also been computed on membranes drawn manually by pathologists for taking advantage of knowledge and skills of experts in that field. Manual analysis has been performed on all the three data sets. These manual measurements were thus compared with the positive protein reaction parameters computed through our fully automated approach.

Results are reported as follow. For each data set, we compute the average error and the root mean square error (RMSE) incurred by our automated approach with respect to manual-trace measurements. We then computed the coefficient of correlation between each set of automated results and the correspondent manual-trace measurements. Finally, we performed a linear regression between automated manual-trace results to evaluate the level of confidence of the regression coefficient through the *Student t-test*.

We first evaluate the correlation between the automated and the manual-trace measurements on the first immunohistochemical lung cancer tissue image. Our analysis shows that these two sets of measurements are highly correlated, with a coefficient of correlation of 0.98. We then computed a linear regression of automated measures on manual-trace ones. We performed the *Student t-test* under the null hypothesis on the regression coefficients in order to estimate the confidence level of this regression. As a result, we rejected this hypothesis at significance level less than 1% obtaining a coefficient of the regression line of 0.96 with a *region of acceptance of the hypothesis* of the range -0.109 to 0.109. Thus, the two sets of measures are highly correlated with a confidence level greater of 99%. Figure 2 shows results obtained for *EGF-R* protein activation measurements on the first immunohistochemical lung cancer tissue image. The figure shows the automated measurements versus the manual-trace ones as well as the regression line.

Moreover, we computed the difference between automated and manual-trace measurements and we performed the same *Student t-test*. We found that the difference between the two typologies of measurements is not significant and the average of differences between automatic and manual measurements is of 0.773%. Finally, the RMSE of our automated measurements is 3.3% (with a confidence of 99%), as shown in the first row of Table 1. Table 1 shows, in the first column, the computed percentage of receptor activation in the lung cancer tissue. In this first data set that percentage is 58.65%.

We performed the same analysis also on the second and third data set of immunohistochemical lung cancer tissue images. On the second data set, our analysis showed that automated and manual-trace measurements were highly correlated with a coefficient of correlation of 0.97. Performing the *Student t-test* under the null hypothesis on the regression coefficients we finally rejected this hypothesis at significance level less than 1%. We obtained a coefficient of the regression line of 0.85 with a *region of acceptance of the hypothesis* of the range -0.11 to 0.11. Figure 3 shows these results for *EGF-R* protein activation measurements on the second immunohistochemical lung cancer tissue image. By performing the *Student t-test* on the difference between automated and manual-trace measurements we found that the difference between this two typologies of measurement is not significant. Moreover, the average of difference between automatic and manual measurements is of 0.25% and the RMSE of our automated measurements is about 1.6%, as shown in second row of the Table 1.

The percentage of receptors actives in this second lung cancer tissue set is 95.89%. In this data set the *EGF-R* receptor is highly active in most of the cells on the tissue. Looking at Figure 3, we notice that almost all the measurements are clustered around a very high value while only a few measures are slightly smaller. This leads to a very little dispersion of the measures. At the same time, since the significance is computed with respect to the dispersion, lower values on the data set slightly affect the slope of the regression line thus increasing the level of the significance of the test. Nevertheless, also in this case, the automated and manual-trace measurements are correlated with a confidence level greater of 99%.

On the third data set, our analysis showed that the automated and the manual-trace measurements were again

highly correlated with a coefficient of correlation of 0.976. Performing the *Student t-test* under the null hypothesis on the regression coefficients we finally rejected this hypothesis at significance level less than 1% and we obtained a coefficient of the regression line of 0.958 with a *region of acceptance of the hypothesis* of the range -0.096 to 0.096. Thus, the two sets of measures are correlated with a confidence level greater of 99%. Figure 4 shows these results for *EGF-R* protein activation measurements on the third immunohistochemical lung cancer tissue image. Moreover, the *Student t-test* on the difference between automated and manual-trace measurements showed that the difference between the two typologies of measurement is not significant and that the average error is of 0.145%. Finally, the RMSE of our automated measurements is about 2.65%, and the percentage of receptors active in the third lung cancer tissue set is 83.96%, as shown in the third row of Table 1.

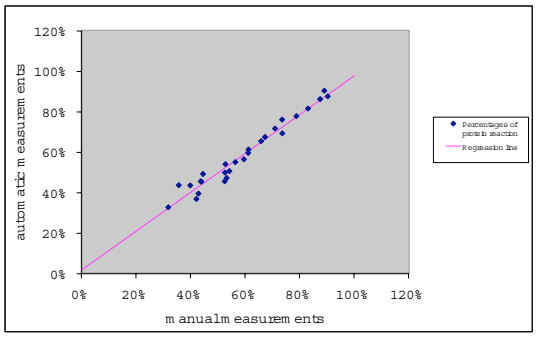


Figure 2. Results on the first data set: the plot shows the automated procedure measurements versus the manual-trace ones and the regression line

Positive Mem Reaction(%)	Average Error (%)	RMSE (%)
58.65	-0.77	3.3
95.89	-0.25	1.58
83.96	-0.145	2.65

Table 1. Results on percentage computation of receptor *EGFR* family activation on the three tissue image experimental data sets: first column shows the clinical parameter while the other ones indicate the average error and the root mean square error incurred by the automated procedure.

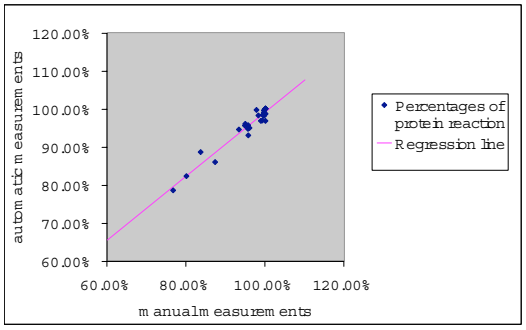


Figure 3. Results on the second data set: the plot shows the automated procedure measurements versus the manual-trace ones and the regression line

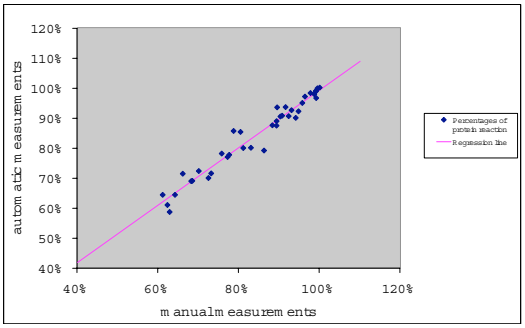


Figure 4. Results on the third data set: the plot shows the automated procedure measurements versus the manual-trace ones and the regression line

4 Conclusions

We presented a fully-automated computer-aided analysis approach to the quantification of the expression of receptors in carcinoma tissue images. This immunohistochemical analysis is usually performed by pathologists via visual inspection of tissue samples images. Our techniques streamlines this error-prone and time-consuming process, thereby facilitating analysis and diagnosis. In particular, our method leads to classify protein reactions according to a specific cell region and to quantify the percentage and the intensity of this protein activity. The effectiveness of the proposed method has been tested using immunohistochemical non-small cell lung carcinoma tissue images. Results of comparison with manual-trace method on several real-life datasets demonstrate the high quantitative precision of our approach.

As future work we want to correlate the clinical data coming from lung cancer tissue images analysis with gene expression data on same immunohistochemical tissue in order to better define a group of potential candidates to protein family-inhibiting therapy.

References

[1] T.K.Taneja, SK.Sharma "Markers of small cell lung cancer" *World Journal of Surgical Oncology*, Vol(2):10, 2004

[2] E.M.Brey et al. "Automated Selection of DAB-labeled Tissue for Immunohistochemical Quantification" *The Journal of Histochemistry and Cytochemistry*, Vol51(5):575-584, 2003

[3] A.Riufrok, R.Katz, D.Johnston "Comparison of Quantification of Histochemical Staining by Hue-Saturation-Intensity (HSI) Transformation and Color Deconvolution" *Applied Immunohistochemistry and Molecular Morphology*, Vol(11):1, 2004

[4] L.Yang, P.Meer, D.J.Foran "Unsupervised Segmentation Based on Robust Estimation and Color Active Contour Models" *IEEE Transactions on Information Technology in Biomedicine*, Vol(9):3, 2005

[5] D.P.Mukherjee, N.Ray, S.T.Acton "Level Set Analysis for Leukocyte Detection and Tracking" *IEEE Transaction on Image Processing*, Vol(13):4, 2004

[6] A.Elmoataz, S.Schupp, R.Clouard, P.Herlin, D.Bloyet "Using active contours and mathematical morphology tools for quantification of immunohistochemical images" *Signal Processing*, Vol(71):215-226, 1998

[7] N.Malpica et al. "Applying Watershed Algorithms to the Segmentation of Clustered Nuclei" *Cytometry*, Vol(28):289-297, 1997

[8] M. Jacob, T. Blu, M. Unser "Efficient Energies and Algorithms for Parametric Snakes" *IEEE Transactions on Image Processing*, Vol. 13(9):1231-1244, 2004

[9] T. W. Ridler, S. Calvard, "Picture thresholding using an iterative selection method", *IEEE Trans. on Systems, Man, and Cybernetics*, 8(8): 630-632, August 1978

[10] J. Kittler, J. Illingworth, C. Y. Suen "Minimum error thresholding", *Pattern Recognition* 19: 41-47, 1986

[11] N. Otsu "A threshold selection method from gray level histograms", *IEEE Trans. on Systems, Man, and Cybernetics*, SMC-9: 62-62, 1979

[12] E. Ficarra, L. Benini, E. Macii, G. Zuccheri "Automated DNA Fragments Recognition and Sizing through AFM Image Processing" *IEEE Transactions on Information Technology in Biomedicine* Vol.9(4):508-517, 2005