Joint co-clustering: co-clustering of genomic and clinical bioimaging data

Elisa Ficarra[◇], Giovanni De Micheli[◇], Sungroh Yoon*, Luca Benini[‡], Enrico Macii[†]

[◇]Ecole Polythecnique Federale de Lausanne (EPFL), LSI, Station 14, 1015 Lausanne, Switzerland

*Computer Systems Laboratory, Stanford University, Stanford, CA 94305, USA

[‡]University of Bologna, DEIS, Viale Risorgimento,2 Bologna ITALY

[†] Politecnico di Torino, DAUIN, Corso Duca degli Abruzzi,24 Torino ITALY

{elisa.ficarra@polito.it, giovanni.demicheli@epfl.ch, sryoon@stanford.edu, lbenini@deis.unibo.it, enrico.macii@polito.it}

Abstract

For better understanding of genetic mechanisms underlying clinical observations, and better defining a group of potential candidates to protein family-inhibiting therapy, it is interesting to determine the correlations between genomic, clinical data and data coming from high resolution and fluorescent microscopy. We introduce a computational method, called *joint co-clustering*, that can find co-clusters or groups of genes, bioimaging parameters and clinical traits that are believed to be closely related to each other based upon given empirical information. As bioimaging parameters, we quantify the expression of *growth factor receptor EGFR/erb-B* family in *non-small cell lung carcinoma* (NSCLC) through a fully-automated computer-aided analysis approach. This immunohistochemical analysis is usually performed by pathologists via visual inspection of tissue samples images. Our fully-automated 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. The joint co-clustering method was tested with the receptor *EGFR/erb-B* family data on *non-small cell lung carcinoma* (NSCLC) tissue and identified statistically significant co-clusters of genes, receptor protein expression and clinical traits. The validation of our results with the literature suggest that the proposed method can provide biologically meaningful co-clusters of genes and traits and that it is a very promising approach to analyze large-scale biological data and to study multi-factorial genetic pathologies through their genetic alterations.

Keywords: Image processing, gene clustering, protein activities, clinics.

I. Introduction

For better understanding of genetic mechanisms underlying clinical observations, it is interesting to determine which genes and clinical traits are interrelated. In the last few years a consistent amount of research in genomics has been done concerning correlation of gene expression to multi-factorial genetic pathologies. Microarray data analysis, as well as real-time PCR, are useful techniques exploited so far to this purpose [1] [2]. Despite this effort, results obtained are strongly limited by the poor informative content provided by clustering techniques applied to gene expression data [3].

At the same time, in the field of biomedical and molecular imaging, new techniques have been shown to be effective in extracting clinical and functional biological information from images of molecules and tissues [4] [5]. By observing processes as they happen within the cell, these techniques add an important extra dimension to the understanding of cell behavior and functioning for early disease detection and drug response. In clinics, new applications of conventional imaging technologies are likely to play increasingly important roles, particularly in oncology.

These two independent sources of information, namely gene expression mining techniques and fully-automated bioimaging, can be correlated to enhance gene expression analysis and to increase the amount of confidence in the hypothesized gene expression paths. For this purpose, we developed a joint co-clustering technique able to extract clinical bioimaging parameters through a fully-automated computer-aided approach and to perform co-clustering technique between clinical bioimaging parameters and gene expression data.

Our proposed method consists of two steps. As first step, we early developed a computational method that can deterministically find all the co-clusters, between clinical traits and gene expression data, satisfying specific input parameters in an efficient manner [6]. To measure correlation between a gene and a clinical trait, existing approaches obtain a vector of the expression level of the gene over a number of samples and another vector of the value of the clinical trait over the same samples and then calculate statistical correlation between two vectors. By applying this procedure to many genes, we can identify some genes correlated to the clinical trait of interest. Proceeding one step further from prior methods that can reveal one-to-many relationships between a single trait and multiple genes (or vice versa), we developed a method that can find many-to-many relationships between genes and traits using a clustering technique called *co-clustering*. This method possesses clear advantages over heuristic methods that can provide only partial solutions and other exact algorithms that are not scalable to large-scale problems [7] [8].

As second step, we developed a fully-automated tissue image processing method, namely computer-aided protein quantification tool, able to extract a set of clinical parameters that give a characterization of the pathology dynamics. This tool was successfully tested on *non-small cell lung carcinoma* (NSCLC) tissue images in order to characterize and quantify, in a standardized way, the activation of the *EGFR/erb-B* protein receptor family that plays an important role in *non small cell lung carcinoma* growing [9]. This type of analysis aims at characterizing each pathological cell, and on average the whole tissue, by performing a standardized quantitative and qualitative measurement of protein activations.

This information can be treated as a clinical parameter, and can be finally correlated with the genetic expression data on same lung carcinoma tissue in order to better define a group of potential candidates to protein family-inhibiting therapy. For this purpose, we developed the proposed fully-automated *joint co-clustering* approach to find correlations between genetic data and clinical and bioimaging parameters.

The tool was tested with the epidermal growth factor receptor *EGFR/erb-B* family data set in the *non-small cell lung carcinoma* (NSCLC) tissue. The *EGFR/erb-B* family of receptors plays an important role for 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/erb-B* in individual lung carcinomas. Thus, it is important to extract these information by using methodologies that give quantifiable, standardized and precise measurements [9]. We quantified the activity of the *EGFR/erb-B* receptors in NSCLC *immunohistochemical* images of 70 patients. Subsequently, we correlated these bioimaging parameters with the expression of genes that regulate the transcription of the *EGFR/erb-*

B protein family, measured on same tissues and on the same data sets of 70 patients and other clinical traits (e.g. tumor classification, survival, follow-up etc.). Results show that there is a strong correlation between bioimaging parameters quantifying *EGFR/erb-B* protein family activations and their gene regulative expression. To justify our analyses, we present some supporting evidence of our results in the literature. Our experimental studies suggest that joint co-clustering is a very promising approach to analyze large-scale biological data and to study multi-factorial genetic pathologies through their genetic alterations. Moreover, this approach enables new opportunities for early diagnosis and provides information in future strategies for therapy.

Section II explains the computer-aided protein quantification tool. Section III explains our method to find co-clusters. Experimental results and discussions are presented in Section IV, followed by concluding remarks in Section V.

II. Computer-aided protein quantification tool: membranes detection and parameters extraction

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.

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 [10] [11]. 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 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.

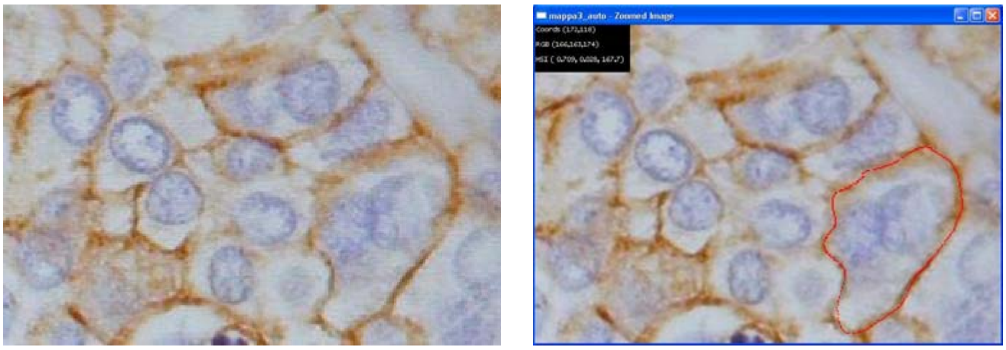
In this section we describe a fully-automated procedure that provides standardized measures of protein activities, and related ligands, involved in the development of a pathology. This goal is reached *i)* by identifying different cellular regions, *ii)* quantifying the percentage of active areas with respect to each whole region, and *iii)* 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 this kind of 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 [12] [13] [14] [15]. 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. 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.

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



(a) (b)

Fig. 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

this work, our description concentrates on the steps we customized.

A. 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 [16]. 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 [16] and [14].

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. Note that, we implemented a customized procedure for virtual cell membranes design because alternative traditional methods, such as Voronoi tessellation [17], build curves equidistant from points. Since these alternative methods fix as center of tessellation a point instead of complex shape membranes (i.e. nucleus membranes), they obtain virtual curves that have few edges and sharp angles between edges. These curves are thus poor approximation of real cell membranes.

B. 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 [18] [19] [20]. 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 [21].

C. 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.

The first step of cellular membrane detection is the *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. *I*j is the value of the pixel *j*th and *c*j is a coefficient for barycentre computation. The coefficient *c*j 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.

The second step in the detection of cellular membranes is the *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

D. Clinical parameter computation

We quantify the activity of membrane *EGFR/erb-B* 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.

III. Co-clustering method

The co-clustering approach finds groups of genes and clinical parameters that are believed to be closely related to each other based upon given empirical information. In particular, it can find many-to-many relationships between genes and traits using a clustering technique called co-clustering. Here the term co-clustering or refers to an unsupervised learning technique that performs simultaneous clustering of rows and columns in a matrix to find (possibly) overlapping submatrices covering the matrix.

More specifically, given gene expression data and clinical parameter values, we first create a matrix called *correlation matrix* that can collectively represent the degree of correlation between genes and clinical traits. Each row and column of this matrix corresponds to a gene and a clinical trait, respectively. Then, our method searches co-clusters or submatrices (with some semantics to be defined) covering the correlation matrix.

A. Definitions

Let S represent a set of clinical samples. For each sample in S, gene expression levels are measured by the DNA microarray technology of choice. Let G be the set of genes in the measurement. Clinical traits are recorded for each sample. Let T be the set of the recorded traits.

The input of the proposed approach is composed of two data matrices. One is a gene expression data matrix denoted by pair $A = (G, S)$, where, $A \in \mathbb{R}^{|G| \times |S|}$, and the element a_{ik} of the matrix A represents the expression level of gene i for sample k. The other matrix is denoted by pair $B = (T, S)$, and the element b_{jk} of the matrix B is the value of trait j for sample k. The columns of A and B are arranged in the same order. Depending upon the type of trait j, b_{jk} may be quantitative, categorical, or others.

The output is a set of co-clusters. A co-cluster is composed of a gene set $I \subseteq G$ and a trait set $J \subseteq T$ and represents a group of genes and traits closely related to each other, given the input matrices A and B. A co-cluster can formally be defined by the following series of definitions.

Definition 1: For V, a vector on $\mathbb{R}$, the range of V, denoted by $\textsc{range}(V)$, is the absolute difference between the largest and the smallest elements of V.

Definition 2: Given V and W, two real vectors of the same dimension, the *linear deviation* of V and W, denoted by $\textsc{lin-dev}(V, W)$, is defined as

$$min\{\textsc{range}(V - W), \textsc{range}(V + W)\}. \tag{2}$$

Definition 3: Given the input matrices A and B, a *correlation matrix*, denoted by C, is a matrix where the row set and the column set of C are G and T, respectively, and the element c_{ij} is the statistic indicating the degree of correlation between gene i and trait j and is defined in *significance analysis of microarrays* (SAM) [8], namely,

$$c_{ij} = \frac{r_{ij}}{s_{ij} + s_0}, \tag{3}$$

where r_{ij} is a score to measure the degree of correlation between the expression level of gene i and the value of clinical trait j, s_{ij} is the "gene-specific scatter" or the standard deviation of repeated expression measurements, and s_0 is a "fudge" factor to prevent the computed statistic from becoming too large when s_{ij} is close to zero [23]. Figure 2 shows the correlation matrix construction scheme.

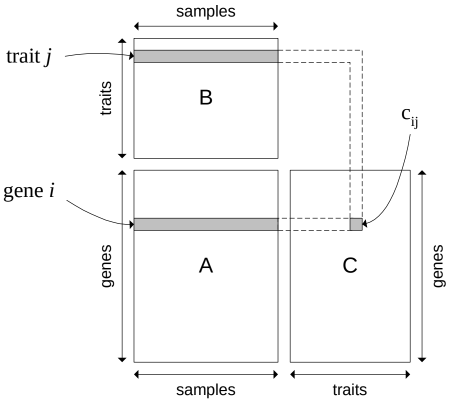


Fig. 2. Construction of the correlation matrix. A co-cluster appears as a submatrix of the correlation matrix C.

Definition 4: Given the correlation matrix $C = (G, T)$ and thresholds $\tau \geq 0$ and $\pi > 0$, a *co-cluster* is a matrix, denoted by $D = (I, J)$, satisfying the following conditions: (1) $I \subseteq G$ and $J \subseteq T$; (2) for any two column vectors V and W of size $|I|$ in D, LIN-DEV$(V, W) \leq \tau$.

Condition (1) indicates that D is a submatrix of the correlation matrix C. Condition (2) is to require that every pair of $|I|$-dimensional column vectors from D exhibit correlation with respect to the metric LIN-DEV.

B. Algorithm overview

As detailed in [6], the proposed co-clustering algorithm consists of three steps: First, an intermediate data matrix called *correlation matrix* is constructed from the input matrices. Then, special co-clusters called *pairwise co-clusters* are found in the correlation matrix. Finally, co-clusters are derived from the pairwise co-clusters. An overview of the algorithm is presented below, and more details can be found in [6].

In the first step, each element c_{ij} of the correlation matrix is calculated using the SAM procedure. When calculating c_{ij}, we must follow a procedure for multiple comparisons, thus ensuring that too many falsely significant ones are not declared [22], [23]. To this end, the *false discovery rate* (FDR) is estimated for each c_{ij} by random permutation of the data for gene expression among the different experimental arms. The SAM procedure can be outlined as follows [8]:

1) For given j, compute statistic c_{ij} for $i = 1, 2, \ldots, |G|$, where $|G|$ is the number of genes in the gene expression matrix.
2) Compute order statistics $c_{(1)} \leq c_{(2)} \cdots \leq c_{(|G|)}$.
3) Take M sets of permutations of the vector associated with trait j. For each permutation m, compute statistics c_{ij}^{*m} and corresponding order statistics.
4) From the set of M permutations, estimate the expected order statistics by $\overline{c}_{(i)} = (1/M) \sum_m c_{(i)}^{*m}$ for $i = 1, 2, \ldots, |G|$.
5) Plot the values of $c_{(i)}$ versus the values of $\overline{c}_{(i)}$.
6) For Δ, a fixed threshold, find the first $i = i_1$ such that $c_{(i)} - \overline{c}_{(i)} > \Delta$, starting at the origin and moving up to the right. All genes past i_1 are called *significant positive*. Similarly, find *significant negative* genes. For each Δ, define the upper cut-point $cut_{up}(\Delta)$ as the smallest c_{ij} among the significant positive genes, and similarly define the lower cut-point $cut_{low}(\Delta)$.
7) For a grid of Δ values, compute the total number of significant genes (from the previous step), and the median number of falsely called genes, by computing the median number of values among each of the M sets of $c_{(i)}^{*m}$ (for $i = 1, 2, \ldots, |G|$) that fall above $cut_{up}(\Delta)$ or below $cut_{low}(\Delta)$.
8) Estimate P_0, the proportion of true null (unaffected) genes in the data set (see [8] for details).
9) The median of the number of falsely called genes from Step 6 is scaled appropriately, according to the value of P_0 (see [8] for details).
10) A value of Δ can be specified by the user and the significant genes are listed.
11) The FDR is computed as the median of the number of falsely called genes divided by the number of genes called significant.

After having computed the correlation matrix, the next step is to find a special type of co-cluster called *pairwise* *co-cluster*. A pairwise co-cluster is a co-cluster with only two traits and can therefore be represented by a submatrix (of the correlation matrix) with two columns. Pairwise co-clusters are used later as seeds to find (non-pairwise) co-clusters. To find a pairwise co-cluster in the correlation matrix $C = (G, T)$, we first select two distinct columns $v, w \in T$ and construct from them two $|G|$-dimensional column vectors $V = (c_{1v}, c_{2v}, \dots, c_{|G|v})$ and $W = (c_{1w}, c_{2w}, \dots, c_{|G|w})$. Then, we compare V and W to identify I, a set of dimensions over which V and W are correlated ($I \subseteq G$). Finally, we remove all $i \in I$ such that p-value of c_{iv} or c_{iw} is greater a given threshold. By definition, the matrix denoted by pair $(I, \{v, w\})$ represents a co-cluster, and this co-cluster with only a pair of traits is called *pairwise co-cluster*.

In the last step of our method, co-clusters are derived from pairwise co-clusters. Recall that T is the set of clinical traits or the set of column indices in the correlation matrix $C = (G, T)$. Our co-clustering method examines elements $J \in 2^T$ in such an order that efficient enumeration is possible to find a co-cluster (I, J). To this end, a data structure called *prefix tree* or *trie* [24] is employed to systematically represent the elements of the power set 2^T. Each node in the trie represents candidates for co-clusters, and using an efficient traversal method nodes are gradually merged and pruned, resulting in co-clusters in their final form.

C. Remarks

To assess the degree of correlation, in Definition 2 we introduced a metric called *linear deviation*. This is not to deny the effectiveness of a conventional statistic such as the Pearson correlation coefficient [22] but to transform it to a computation-efficient form, minimizing loss in the detection power. It is possible to see the relationship between LIN-DEV and the Pearson correlation coefficient, as shown in [6]: a lower value of LIN-DEV typically corresponds to a higher level of either positive or negative correlation.

The specific definition of r_{ij} in Definition 3 varies depending upon the type of clinical trait j. For example, if clinical trait j has quantitative values then r_{ij} is defined in terms of the Pearson correlation coefficient [22] between the i-th row vector of the matrix A and the j-th row vector of the matrix B.

D. Joint co-clustering method

The joint co-clustering approach is a fully-automated framework that aims to extract receptor and protein expressions from tissue images and correlate these bioimaging parameters with other clinical traits and the gene regulative expression of same receptors and proteins evaluated on same tissues. Thus, the joint co-clustering framework consists on the co-clustering algorithm where clinical traits are obtained through the fully automated protein quantification tool.

IV. Experimental results and Discussion

Experimental results were separately obtained for the computer-aided protein quantification tool and the co-clustering to demonstrate their accuracy and robustness. Afterwords, we present experimental results of joint co-clustering method.

A. Computer-aided protein quantification results

We tested the algorithm on four data sets. All of them are composed by real lung cancer tissue immunohistochemical images. For each data set, the images show different portions of same IHC tissue. The four data sets present positive reactions at the *EGFR/erb-B* receptor activation. These reactions are localized in the cellular membranes. The four data sets differ because of different levels of positivity intensity.

For each data set, we first localized each cellular membrane in the images, as described in Sec. II. Afterwords, we computed for each cell the percentage of area characterized by positive activation of receptor *EGFR/erb-B* 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 each 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 data sets. These manual measurements were thus compared with the positive protein reaction parameters computed through our fully automated approach.

We show in this paper results on two of all data sets in order to demonstrate the accuracy and robustness of our approach. On the other data sets, we obtained similar results and performance. Details can be found in [25].

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 set. 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 3 shows results obtained for *EGF-R* protein activation measurements on the first immunohistochemical lung cancer tissue image set. 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 I. Table I 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 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

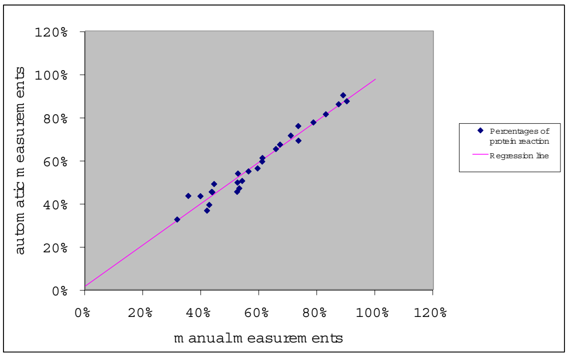


Fig. 3. Results on the first data set: the plot shows the automated procedure measurements versus the manual-trace ones and the regression line

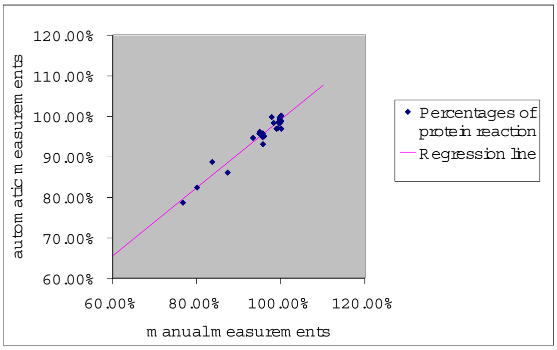


Fig. 4. Results on the second data set: the plot shows the automated procedure measurements versus the manual-trace ones and the regression line

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 4 shows these results for *EGF-R* protein activation measurements on the second immunohistochemical lung cancer tissue image set. 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 I.

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 4, 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%.

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

TABLE I
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.

B. Co-clustering and Joint co-clustering results

The co-clustering was first tested with the Acute Myelogenous Leukemia (AML) data set [7]. The AML data set used included two matrices. One was a gene expression data matrix with 6283 genes and 119 samples. The other was a matrix of 15 clinical parameters measured from the identical samples. We used the procedure described in Section III to produce the correlation matrix. We identified 43 co-clusters. To justify the grouping of certain genes and clinical traits by the co-clusters found from the AML data, we present some supporting evidence of co-clustered genes and traits from the literature. In addition, we show that certain Gene Ontology terms annotating genes in some co-clusters are significantly over-represented. Taken together, these experimental studies suggest that our method can find biologically meaningful co-clusters. Details on these results can be found in [6].

The joint co-clustering was tested with the epidermal growth factor receptor *EGFR/erb-B* family data set in the *non-small cell lung carcinoma* (NSCLC) tissue. The *EGFR/erb-B* family of receptors plays an important role for NSCLC development. Quantifying and classifying the *EGFR/erb-B* expression and activity in NSCLC with special regard to the assessment of the prevalence of somatic *EGFR/erb-B* mutations, as well as to ligand-receptor interactions, could lead to new insights into the modulation of *EGFR/erb-B* in individual lung carcinomas. Thus, it is important to extract these information by using methodologies that give quantifiable, standardized and precise measurements. We quantified the activity of the *EGFR/erb-B* receptors in NSCLC *immunohistochemical* images of 70 patients. Subsequently, we correlated these bioimaging parameters with the expression of genes that regulate the transcription of the *EGFR/erb-B* protein family, measured on same tissues and on the same data sets of 70 patients and other clinical traits (such as tumor type classifications, namely diagnosis, T, N, stage, size, survival, etc). Furthermore, we found supporting evidence of our results in the literature.

Note that results on *EGFR/erb-B* protein expression (i.e.quantification) have been already given in this section (see Section IV-A). As result of joint co-clustering between *EGFR/erb-B* protein expression and the regulative expression of the *EGFR/erb-B* protein transcripts, we found out significant correlations in about 83% of the studied cases. Among this percentage, we found co-clusters chracterized by up-regulation of the transcripts and over-expression of the proteins. Among tumors that did not exhibit over-expression, i.e., the tumors that showed low protein positivity or negative staining, no gene up-regulation was observed. Moreover, high-level regulation was significantly more frequent in tumors with highest staining than in tumors with medium staining. Similar results was reported in recent studies [26] [27].

The remaining 17% of studied cases presents activation of EGFR protein family (visible through image analysis) but no up-regulation of the expression of the protein transcripts. In these particular cases the 70% of the tumor was squamous cell carcinoma (SqCa), while the 30% was adenocarcinoma (AdCa). Similar findings have been reported not only in lung carcinomas but also in other tumors, such as renal, pancreatic, breast, and colon carcinomas. Although protein overexpression in these tumors probably is caused by transcriptional or post-transcriptional activation, various theories have been proposed to explain the underlying mechanisms [28] [26]. Post-translational changes as well as changes in genetic enhancer elements [29] [30] were shown to be associated with an increased *EGFR* expression. Recently, a polymorphic CA-repeat in intron 1 of *EGFR* has been shown to have an important impact on *EGFR* transcription and expression, too and seems to be a major target of *EGFR* mutations [31] [27]. In the literature it has been found that this mechanism can explain protein over-expression in about 18.7% of cases [27], that is in accordance with our results.

We found out also that trait ”diagnosis” (e.g., SqCA, AdCa, LCa) is correlated with genes erb-1, erb-2 and TGF-alpha with a FDR of 1%. Similarly, ”size” trait is correlated with erb-1 gene and ”survival” trait is correlated with the erb-2 gene, as confirmed in [32] [33].

Finally, we identified co-clusters of erb-1 and erb-2 proteins. These co-clusters consisted of the 54% of the studied cases and were characterized, in particular, by erb-1 and erb-2 protein expression, their genetic regulation and SqCA/AdCA type classification of tumor. As results, it can be seen that the over-expression of erb-1 (EGFR) impacts the SqCa tumors more than AdCa ones (56% of cases vs. 32%) and vice versa for the expression of erb-2 protein and the AdCa tumors (63% vs. 26%). Evidence of our results was found in literature [34] [35]. We found out also co-clusters that presented over-expression and gene up-regulation for erb-1 protein and did not exhibit gene up-regulation nor over-expression for erb-2. These co-clusters were characterized by a percentage of SqCa tumors higher than those of AdCa ones (67% vs. 0%). Vice versa was found for co-clusters that presented over-expression and gene up-regulation for erb-2 protein and did not exhibit gene up-regulation nor over-expression for erb-1 (100% of AdCa tumors). SqCA and AdCA tumors were found also in co-clusters characterized by either gene up-regulation nor over-epression for both erb-1 and erb-2 (67% of AdCa and 27% SqCa). We found supporting evidence also for these last analyses in the literature [27].

V. Conclusions

We presented a fully-automated framework for finding co-clusters of genes and clinical traits using microarray data and bioimaging and clinical parameter information.

We first quantified the expression of receptors in carcinoma tissue images by using our fully-automated protein quantification tool. This immunohistochemical analysis (IHC) 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.

Data coming from IHC images can be treated as a clinical parameters, and can be finally correlated with the genetic expression data on same lung carcinoma tissue (and same set of patients) in order to better define a group of potential candidates to protein family-inhibiting therapy. For this purpose, we developed the proposed fully-automated joint co-clustering approach. An intermediate data matrix called correlation matrix was computed from microarray data and bioimaging and clinical parameter information by means of a statistical method. We then modeled a co-cluster by a submatrix of the correlation matrix with some semantics and aimed at finding statistically significant co-clusters.

In order to validate our approach, we found supporting evidence of our analysis in the literature. Results show that there is a strong correlation between bioimaging parameters quantifying *EGFR/erb-B* protein family activations and their gene regulative expression measured on same tissues. These preliminary results show that the joint co-clustering is a very promising approach to analyze large-scale biological data and to study multi-factorial genetic pathologies through their genetic alterations. Moreover, this approach enables new opportunities for early diagnosis and provides information in future strategies for therapy.

VI. Acknowledgments

We wish to acknowledge the Swiss Institute for Experimental Cancer Research (ISREC) in Lausanne, Switzerland and the Department of Pathology at the S. Luigi Hospital, Turin, Italy for their helpful and valuable advices and discussions.

References

[1] S. Saviozzi, G. Iazzetti, E. Caserta, A. Guffanti, R.A. Calogero, Microarray data analysis and mining, Methods Mol Med. 94 (2004) 67-90.

[2] S. Yoon, C. Nardini, L. Benini, and G. De Micheli, Enhanced pClustering and its applications to gene expression data, Proc. of IEEE Bioinformatics and Bioengineering (2004) 275-282.

[3] O.G. Troyanskaya, Putting microarrays in a context: Integrated analysis of diverse biological data, Briefings in Bioinformatics 6(1) (2005) 34-43.

[4] A. Hengerer, A. Wunder, D.J. Wagenaar, A.H. Vija, M. Shah, J. Grimm, From Genomics to Clinical Molecular Imaging, Proceedings of the IEEE 93(4) (2005) 819-828.

[5] W. Chen, M. Reiss, D.J. Foran, Unsupervised tissue microarray analysis for cancer research and diagnostics, IEEE Transactions on Information Technology in Biomedicine 8(2) (2004) 89-96.

[6] S. Yoon, L. Benini, G. De Micheli, Finding Co-Clusters of Genes and Clinical Parameters, Proc. of the 27th Annual International Conference of the IEEE EMBS (2005) 4799-4802.

[7] L. Bullinger et al., Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia, N. Engl. J. Med. 350(16) (2004) 16051616.

[8] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci USA, 98(9) (2001) 51165121.

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

[10] E.M. Brey et al. Automated Selection of DAB-labeled Tissue for Immunohistochemical Quantification, The Journal of Histochemistry and Cytochemistry 51(5) (2003) 575-584.

[11] A. Ruifrok, R. Katz, D. Johnston, Comparison of Quantification of Histochemical Staining by Hue-Saturation-Intensity (HSI) Transformation and Color Deconvolution, Applied Immunohistochemistry and Molecular Morphology 11(1) (2004) 85-91.

[12] 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 9(3) (2005) 475-486.

[13] D.P. Mukherjee, N. Ray, S.T. Acton, Level Set Analysis for Leukocyte Detection and Tracking, IEEE Transaction on Image Processing 13(4) (2004) 562-572.

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

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

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

[17] F. Aurenhammer, R. Klein, Voronoi Diagrams, Handbook of Computational Geometry (Ed. J.-R. Sack and J. Urrutia) Ch. 5 (2000) 201-290.

[18] T. W. Ridler, S. Calvard, Picture thresholding using an iterative selection method, IEEE Transactions on Systems, Man, and Cybernetics 8(8) (1978) 630-632.

[19] J. Kittler, J. Illingworth, C. Y. Suen, Minimum error thresholding, Pattern Recognition 19 (1986) 41-47.

[20] N. Otsu, A threshold selection method from gray level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1979) 62-62.

[21] 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 9(4) (2005) 508-517.

[22] B. Rosner, Fundamentals of Biostatistics, Duxbury Edit. 5th edition (2000).

[23] S. Draghici, Data Analysis Tools for DNA Microarrays, Chapman & Hall/CRC (2003).

[24] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, Data Structures and Algorithms, Reading, Massachusetts: Addison-Wesley, 1983.

[25] E. Ficarra, E. Macii, L. Benini, G. De Micheli, Computer-aided evaluation of protein expression in pathological tissue images, Proc. of IEEE International Symposium on Computer-Based Medical Systems (CBMS 2006) (2006) 413-418.

[26] S. Suzuki, Y. Dobashi, H. Sakurai, K. Nishikawa, M. Hanawa, A. Ooi, Protein overexpression and gene amplification of epidermal growth factor receptor in nonsmall cell lung carcinomas, Cancer 103(6) (2006) 1265-73.

[27] C. Kersting, N. Tidow, H. Schmidt, et al., Gene dosage PCR and fluorescence in situ hybridization reveal low frequency of egfr amplifications despite protein overexpression in invasive breast carcinoma, Lab Invest. 84 (2004) 582-587.

[28] J. Amann, S. Kalyankrishna, P.P. Massion, et al., Aberrant Epidermal Growth Factor Receptor Signaling and Enhanced Sensitivity to EGFR Inhibitors in Lung Cancer, Cancer Research 65 (2005) 226-235.

[29] F. Gebhardt, H. Burger, B. Brandt, Modulation of EGFR gene transcription by secondary structures, a polymorphic repetitive sequence and mutationsa link between genetics and epigenetics, Histol Histopathol 15 (2000) 929936.

[30] J.M. McInerney, M.A. Wilson, K.J. Strand, et al. A strong intronic enhancer element of the EGFR gene is preferentially active in high EGFR expressing breast cancer cells, J. Cell Biochem. 4 (2001) 538549.

[31] N. Tidow, A. Boecker, H. Schmidt, et al., Distinct amplification of an untranslated regulatory sequence in the egfr gene contributes to early steps in breast cancer development, Cancer Res. 6 (2003) 11721178.

[32] Z. Zheng, G. Bepler, A. Cantor, E. B. Haura, Small Tumor Size and Limited Smoking History Predicts Activated Epidermal Growth Factor Receptor in Early-Stage Non-small Cell Lung Cancer, Chest 128 (2005) 308-316.

[33] J. Brabender, K. D. Danenberg, R. Metzger, P. M. Schneider, J. M. Park, D. Salonga, A. H. Hlscher, P. V. Danenberg, Epidermal Growth Factor receptor and HER2-neu mRNA expression in Non-Small Cell Lung Cancer is correlated with Survival, Clinical Cancer Research 7 (2001) 1850-1855.

[34] Y. Zhou, S. Li, Y.P. Hu, J. Wang, J. Hauser, A.N. Conway, M.A. Vinci, L. Humphrey, E. Zborowska, J.K.V. Willson, M.G. Brattain, Blockade of EGFR and ErbB2 by the Novel Dual EGFR and ErbB2 Tyrosine Kinase Inhibitor GW572016 Sensitizes Human Colon Carcinoma GEO Cells to Apoptosis, Cancer Res. 66 (2006) 404-411.

[35] S. Dacic, M. Flanagan, K. Cieply, S. Ramalingam, J. Luketich, C. Belani, S.A. Yousem, Significance of EGFR protein expression and gene amplification in non-small cell lung carcinoma, Am. J. Clin. Pathol. 125(6) (2006) 860-865.