Self Organizing Maps for Class Discovery in the Quantitative Colocalization Analysis Feature Space

Pablo Rivas-Perea, Jose Gerardo Rosiles and Wei Qian

***Abstract*— Quantitative colocalization analysis in fluorescent microscopy imaging is a promising procedure used to perform functional protein analysis. Images acquired are degraded, and the features extracted are affected by this degradation. Moreover, the classification of the data becomes uncertain. In this paper, we address an application of SOM to a clustering problem formulated via feature extraction from multichannel fluorescence microscopy. First we describe the features that are extracted. Second, we use the PCA/KLT to uncorrelate the data in the hyperplane; and Third, SOM network is trained to find and visualize the clusters (classes) in the data. The SOM model shows the existence of two classes, implying it is possible to design a classifier that distinguishes between images with colocalized structures and without them. We provide quantitative proof of the liability and discriminant capabilities of the feature space.**

I. Introduction

SUBCELLULAR colocalization analysis is a fluorescence microscopy imaging technique aimed at understanding the functional relationship between molecules in a cell. Typically this technique has been applied to measure the spatial interaction between two proteins that have been fluorescently labeled. This interaction is indicative of the functions that proteins play in the biology of a cell. However, as stated by a recent tutorial review, the actual meaning of a colocalization measure is a source of confusion and contention when microscopy images are used [1]. The typical method used in confocal microscopy is to first image two labeled proteins responding to different wavelengths (typically green and red) and then combining the two color planes into a single image, as shown in Fig. 1. Subjective analysis tries to assess spatial colocalization by visually identifying yellow colored structures in the image (i.e., pixels with large green and red intensity overlap each other). Then the biologist makes an experience-based assessment of the protein interaction which can lead to the typical error and bias found in human visual interpretation. Several qualitative approaches based on global statistics have been proposed over the years (see Section II), however they typically consist of a single coefficient/index (e.g., a correlation coefficient) which may be difficult to interpret in all situations [1] and only indicate the presence of a colocalization event without further quantitative analysis [2].

Pablo Rivas-Perea, Jose Gerardo Rosiles and Wei Qian are with the Department of Electrical and Computer Engineering, The University of Texas, El Paso, Texas (email: privas@miners.utep.edu {grosiles, wqian}@utep.edu).

This is work was partially supported by the National Council for Science and Technology (CONACyT), Mexico, under grant 193324 / 303732.

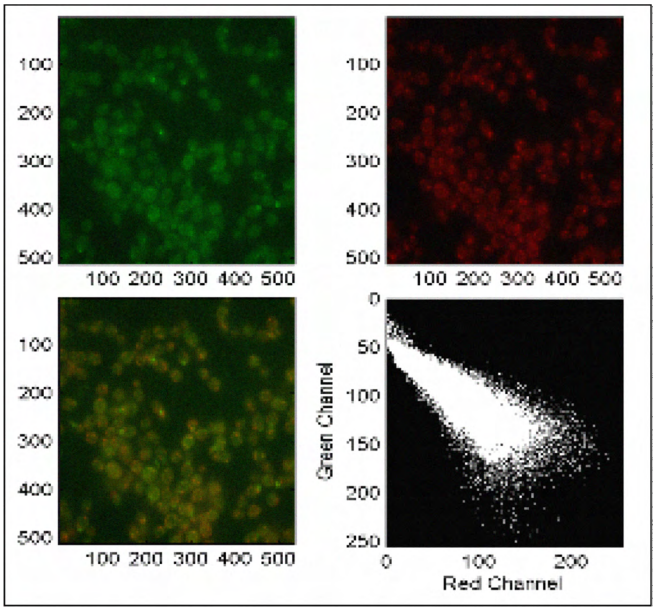

Fig. 1. From top left, we have the green channel. From top right we have the red channel, and bottom left, the twho joint channels. At bottom right we have the scatter diagram, showing the distribution of the pixles accross the two channels.

Recent work by Costes, *et al.*, [2] reports a statistical method that removes the effect of random color overlap which leads to visual inspection bias. Looking beyond statistical analysis, Bolte and Cordelieres [1] presented a colocalization method that extracts subcellular structures based on image segmentation techniques. What can be concluded from the extensive discussion in these two works is that neither visual inspection, nor global statistical analysis are sufficient to faithfully assess colocalization from images. Hence, we can affirm that the biological and biomedical communities have a real need for the creation of automated quantitative image analysis methods that discriminate images into two classes: colocalized and not-colocalized.

This paper is an attempt to formally address the problem of colocalization from the machine learning (ML) perspective. Our aim in this paper is two fold. First we define a set of features that will allow automatic binary classification (colocalized or not-colocalized) of images. Second, we evaluate the discriminative quality of these features using self organizing maps as a tool for class discovery. As features, we have selected the global colocalization coefficients which have been extensively used in the literature with the caveats

978-1-4244-3553-1/09/$25.00 ©2009 IEEE

mentioned above. Nonetheless, they have been informative in many works, and are valid from a purely statistical perspective. Hence they can be considered as a starting point in our analysis. To the best of our knowledge this is the first work using ML techniques to address the colocalization problem.

Open questions regarding the correlation of features in the n-dimensional hyperplane and how trivial does the clustering become in the hyperplane need to be addressed. We answer these questions using principal component analysis (PCA), to decorrelate the data in some hyperplane, and to find the clusters (classes) trough the power of Self Organizing Map (SOM) neural networks.

The paper is organized as follows. In Section II the work on colocalization statistical global analysis is reviewed, presenting the different coefficients reported in the literature. Next, in Section III we introduce the feature vector analysis. Class discovery analysis based on SOMs is discussed in Section IV. We close the paper with a summary of our analysis and a description of our future work in Section V.

II. Quantitative Colocalization Analysis

The quantitative analysis from the image processing point of view, consist of computing the spatial overlap of 2D signals in multiple spectral channels. This allows researchers to understand the mechanisms or potential of the interactions protein-to-protein with very high precision [3]-[5].

The parameters used to estimate the degree of colocalization between two channels are described in the following subsections. Such parameters (features) are categorized in two main groups: first, the features designed to work for all spatial intensity values; and second, those features designed to operate over a specific intensity value (threshold). The algorithms to compute such features are described individually in each category.

A. Features Given a Specific Region of Interest (ROI) in the Scatter Diagram

Consider two-channel digital images $x(n_1, n_2) = [R(n_1, n_2)\ G(n_1, n_2)]$. It is said that two pixels are colocalized if their respective intensities are strictly higher than thresholds k_1, k_2, and if their ratio (of intensity) is strictly higher than a given ratio r_t [6]. This can be expressed as

$$y(n_1,n_2) = \begin{cases} 1 & , \begin{array}{c} \text{if } G(n_1,n_2) > k_1, \text{ and} \\ R(n_1,n_2) > k_2, \text{ and} \\ 100\left(\frac{R(n_1,n_2)}{G(n_1,n_2)}\right) > r_t, \text{ and} \\ 100\left(\frac{G(n_1,n_2)}{R(n_1,n_2)}\right) > r_t \end{array} \\ 0 & , \quad otherwise \end{cases} \tag{1}$$

where $y(n_1, n_2)$ contains the regions (with value 1) where the pixels colocalize; $R(n_1, n_2)$ and $G(n_1, n_2)$ denote the red and green channels; k_1, k_2 are the thresholds specified by the observer; r_t is also chosen by the observer based on the a priori knowledge of the expected result.

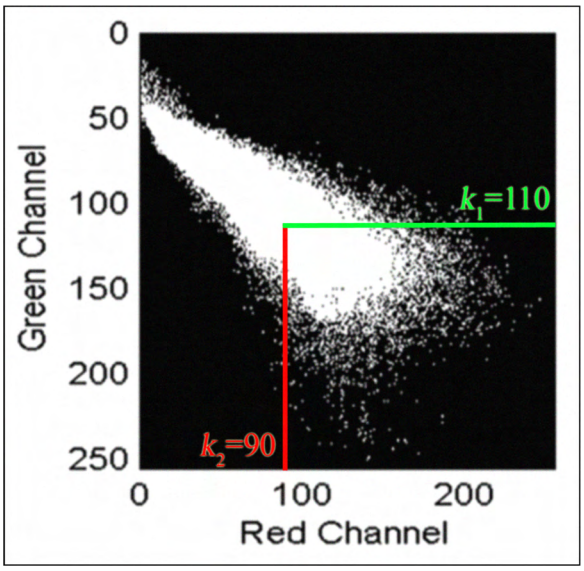

Fig. 2. From an scatter diagram, the observer can chose the threshold and the compute the colocalization for the intensities greater than those thresholds. In this example is shown the selection of the thresholds $k_1 = 110$ and $k_2 = 90$.

The thresholds k_1, and k_2 are picked by the analyst by inspection of the scatter gram [3]-[5]. Fig. 2 shows an example of the selection of thresholds from a scatter gram.

1) Colocalization Coefficients c_1, c_2 for ROI: The colocalization coefficients c_1, and c_2 describe the relationship between the colocalized pixels above the threshold k_1, k_2. [6]. More specific, it is the total number of pixels above the threshold divided by the total number of pixels in the image. This relationship can be denoted as

$$\begin{aligned} c_1 &= \frac{\#\left(y(n_1,n_2) : G(n_1,n_2) > k_1\right)}{n_1 n_2}, \\ c_2 &= \frac{\#\left(y(n_1,n_2) : R(n_1,n_2) > k_2\right)}{n_1 n_2}, \\ &\forall\ n_1,\ n_2,\ k_1,\ k_2 \end{aligned} \tag{2}$$

where c_1, c_2 have an expected result between the range [0 1], where a value of 1 means that all the pixels colocalize.

2) Weighted Colocalization Coefficients wc_1, wc_2: Similar to the Colocalization Coefficients described previously, the Weighted Colocalization Coefficients describe a relationship between the total sum of the colocalized pixels, and the total intensity of the pixels above the threshold k_1, k_2 [6]. This relationship is given by

$$wc_1 = \frac{G_1}{G_2}, \tag{3}$$

$$wc_2 = \frac{R_1}{R_2}, \tag{4}$$

where wc_i denotes the Weighted Correlation Coefficient of the ith channel of the original image, G_1 and R_1 describe the sum of all intensities that colocalize for green and red

channel respectively; R_2 and G_2 denote the sum of all intensities above the thresholds k_1 and k_2 respectively. Also, similar to the coefficients previously described, the expected result is between the range [0 1].

3) Mander's Colocalization Coefficients M_1, M_2: The coefficients M_1, M_2 are are applied only to a specific ROI selected by the observer using the scatter diagram. M_i is used to describe the contribution of the ith channel to the colocalization based on the intensity values. These coefficients are proportional to the amount of fluorescence of colocalizing objects in each channel of the image, relative to the total fluorescence in that channel. Such relationship is described [5] [7] [8] [9] as

$$M_1 = \frac{\sum_{n_1,n_2}(G(n_1,n_2) : R(n_1,n_2) > k_2)}{\sum_{n_1,n_2} G(n_1,n_2)} \quad (5)$$

$$M_2 = \frac{\sum_{n_1,n_2}(R(n_1,n_2) : G(n_1,n_2) > k_1)}{\sum_{n_1,n_2} R(n_1,n_2)} \quad (6)$$

The range of the coefficients is between [0 1]. The meaning can be explained with an example: $M_1 = 1.0$ and $M_2 = 0.1$ means that green channel pixels colocalize with red, in contrast, only 10% of pixels in red channel colocalize with green.

B. All Plane Features

The following coefficients can work either with or without thresholds. In this section we define them assuming no thresholds are provided.

1) Pearson's Correlation Coefficient r_P: Also known as Correlation Coefficient, the Pearson's Correlation Coefficient is widely used and accepted, especially in mean squares, and in many of the regression applications. It provides information about the relationship between the region of intensities and their distribution [5] [7] [8] [9] [10] [11]. In [12] the authors proposed an expansion of this coefficient for individual analysis. The Pearson's Correlation Coefficient is denoted by

$$r_P = \frac{\sum\left(R(n_1,n_2) - \overline{R}\right)\left(G(n_1,n_2) - \overline{G}\right)}{\sqrt[2]{\sum\left(R(n_1,n_2) - \overline{R}\right)^2 \sum\left(G(n_1,n_2) - \overline{G}\right)^2}}$$

where r_P denotes the correlation coefficient. The value of r_P is in the range $[-1\ 1]$, and expresses the level of linear correlation among the two images.

2) Overlap Coefficient oc (The Multiply Method): This coefficient indicates the overlap between the two channels; it shows a degree of colocalization [7]. As a difference from Pearson's this coefficient will not return negative values and will not average any pixel intensity, and it is not sensitive to intensity variations [5] [8]. However the authors of [9] strongly recommends that this coefficient should be used under one condition: $\frac{\sum_{n_1,n_2} G(n_1,n_2)}{\sum_{n_1,n_2} R(n_1,n_2)} \approx 1$. The coefficient is defined as

$$oc = \frac{\sum_{n_1,n_2} G(n_1,n_2)R(n_1,n_2)}{\sqrt[2]{\sum_{n_1,n_2} G(n_1,n_2)^2 \sum_{n_1,n_2} R(n_1,n_2)^2}} \quad (7)$$

where the range is between [0 1]. A value of zero means that no pixels overlap et all, while a one means that all the pixels overlap.

3) Fraction of Colocalizing Regions r_1, r_2 (Overlap Coefficients): This coefficient represent the differences between intensities in each channel [5] [7] [8]. This coefficients overcome the problems generated from a restriction in overlap coefficient oc. However, they are very sensitive to the absolute fluorescent intensity; this means that if one channel has been treated in some way (bleaching for instance) in a different amount than the other in a way that the total intensity vary, this will affect these coefficients [9]. We can denote such coefficients as follows

$$r_1 = \frac{\sum_{n_1,n_2} G(n_1,n_2)R(n_1,n_2)}{\sum_{n_1,n_2} G(n_1,n_2)^2} \quad (8)$$

$$r_2 = \frac{\sum_{n_1,n_2} R(n_1,n_2)G(n_1,n_2)}{\sum_{n_1,n_2} R(n_1,n_2)^2} \quad (9)$$

were the individual expected range may vary. The results can be interpreted as an indicator of the contribution of each antigen to the areas with colocalization.

4) Mander's Colocalization Coefficients m_1, m_2:: The coefficients m_1, m_2 are a particular case of coefficients M_1, M_2 described in a previous section with but $k_1 = 0$ and $k_2 = 0$. Coefficient m_i, is used to describe the contribution of the ith channel to the colocalization based on the intensity values. These coefficients are proportional to the amount of fluorescence of colocalizing objects in each channel of the image, relative to the total fluorescence in that channel. Such relationship is described [5] [7] [8] [9] as

$$m_1 = \frac{\sum_{n_1,n_2}(G(n_1,n_2) : R(n_1,n_2) > 0)}{\sum_{n_1,n_2} G(n_1,n_2)} \quad (10)$$

$$m_2 = \frac{\sum_{n_1,n_2}(R(n_1,n_2) : G(n_1,n_2) > 0)}{\sum_{n_1,n_2} R(n_1,n_2)} \quad (11)$$

where m_1, m_2 are in the range [0 1]. The meaning can be explained with an example: $m_1 = 1.0$ and $m_2 = 0.3$ means that green channel pixels colocalize with red, but only 30% of pixels in red channel colocalize with green.

5) Intensity Correlation Analysis (CoAn):: Two images vary around their respective mean if their intensities vary in synchrony [9]. Therefore the *product of the differences from the mean* (PDM) will be positive for such images. However if the intensities vary asynchronously, the PDM

will be negative. This PDM is an analysis of the relationship between intensities and is denoted as

$$CoAn = \left(G(n_1, n_2) - \overline{G}\right)\left(R(n_1, n_2) - \overline{R}\right) \quad (12)$$

where $\overline{G}$ and $\overline{R}$ denote the arithmetic mean. A value of $CoAn > 0$ implies that the intensities vary synchronously, with $CoAn < 0$ implies the opposite. For instance, if the pixels in green are varying above their mean, and the respective red intensities are varying below their mean, then the $CoAn$ is negative.

6) Intensity Correlation Quotient - ICQ:: The ICQ value is based on the $CoAn$ sign [9]. The ICQ is defined as In this feature, all the positive occurrences in the $CoAn$ are counted and stored in ζ, as well as the negative occurrences are stored in ξ. So that the ICQ is the quotient denoted as

$$ICQ = \left(\frac{\zeta}{\xi}\right) - 0.5, \quad (13)$$

where ζ is the number of positive PDMs and ξ is the number of negative PDDMs. The range of ICQ falls between $[-0.5\ 0.5]$. The results can be interpreted as follows: $ICQ \approx 0$ means random discoloration; $-0.5 \leq ICQ < 0$ means segregated discoloration; $0 < ICQ \leq 0.5$ means dependent discoloration.

7) Scatter Matrix for two channels:: The scatter matrix $S(l_2, l_2)$, or also called scatter gram, is a very important tool that maps the intensity content of the two channels into a single matrix [7] [10]. In order to construct $S(l_2, l_2)$ we must take the intensity values of $G(n_1, n_2)$ and $R(n_1, n_2)$ as coordinates in $S(l_2, l_2)$. A not trivial way to represent this could be

$$S(l_2, l_2) \equiv \#\left\{G(n_1, n_2) = l_1, R(n_1, n_2) = l_2, \forall n_1, n_2\right\} \quad (14)$$

Depending on the purpose, $S(l_2, l_2)$ can be also thought as a voting mechanism.

C. Summary

All the features previously mentioned can characterize the degree of colocalization and provide unique information to the performer of the colocalization study. The content of Table I summarizes the features.

TABLE I

SUMMARY OF THE FEATURES FOR QUANTITATIVE COLOCALIZATION

Name	Variables	Range	Ref.
ROI Coloc. Coeff.	c_1, c_2	[0 1]	[6]
Weighted Coloc. C.	wc_1, wc_2	[0 1]	[6]
Mander's Coloc. C. - ROI	M_1, M_2	[0 1]	[5]
Pearson's Corr. C.	r_P	[−1 1]	[11]-[13]
Overlap C. (Mult. Meth.)	oc	[0 1]	[7]-[9]
Frac. Coloc. Reg. (O. C.)	r_1, r_2	vary	[7]-[9]
Mander's Coloc. C. - Gen.	m_1, m_2	[0 1]	[7]-[9]
Intensity Corr. Anal.	$CoAn$	vary	[9]
Intensity Corr. Quot.	ICQ	[−0.5 0.5]	[9]
Scatter Diag./Matrix	$S(n_1, n_2)$	vary	[7],[13]

III. Feature Vectors Analysis

In the implementation we are using real data images from the yeast database [14]. Yeast database is a very good alternative to test the feature extraction methodologies. It contains 547 pairs of image samples. Images are of size $[512 \times 512]$. An example of the yeast database was shown in Fig. 1.

The feature vector we are utilizing is composed of all the features found in Table I (except by the scatter gram). The feature vector can be expressed as

$$\begin{aligned}\mathbf{f} \quad = \quad & [c_1\ c_2\ wc_1\ wc_2\ M_1\ M_2\ r_P\ ... \\ & oc\ r_1\ r_2\ m_1\ m_2\ CoAn\ ICQ]^T\end{aligned} \quad (15)$$

where $\mathbf{f}$ is a column vector and T denotes the transposition operator.

In this section we address the problem of the data clusters interpretation of the vectors $\mathbf{f}$ by First, finding the subspace where the data is decorrelated using the Karhunen-Loeve Transformation (KLT), which is analog to finding the principal components. Also we utilize the SOM neural network to find clusters in the data. Finally from the clusters found with the SOM network, we analyze the liability and discriminant capabilities between the data clusters through Fisher's criterion.

A. Principal Component Analysis / KLT

Principal Component Analysis, PCA, is a very widely used technique for dimensionality reduction. The objective of PCA is to transform the representation space $\mathbf{f} \in \mathbf{F}$ into a new space $\mathbf{x} \in \mathbf{X}$, in which the data is decorrelated. The covariance matrix in this space is diagonal. The PCA method leads to find the new set of orthogonal axis to maximize the variance of the data.

The final objective for our problem is to decorrelate the data and normalize it. It is not out objective the dimensionality reduction of the problem.

The KLT is analog to PCA, however in the KLT the input vectors $\mathbf{f}$ are normalized to the interval [0,1] before applying the PCA steps.

Using the PCA we have reduced the dimensionality of the data by keeping two and three dimensions only, from the original fourteen dimensions. We want to point that the clustering is non-trivial in this problem. Such dimensionality reduction is presented here for visualization purposes only. In Fig. 3 a) are shown the two principal components, and in Fig. 3 b) are the three principal components (corresponding to the tree largest eigenvectors). Even though the data is not correlated, no clusters or classes are trivially visible using the PCA approach for two and three dimensions, thus, in the next section, a powerful method to find clusters is approached: SOM.

As we mentioned, the use of PCA in dimensionality reduction was aimed to visually show the lack of clusters. In practice for this project the PCA/KLT approach will

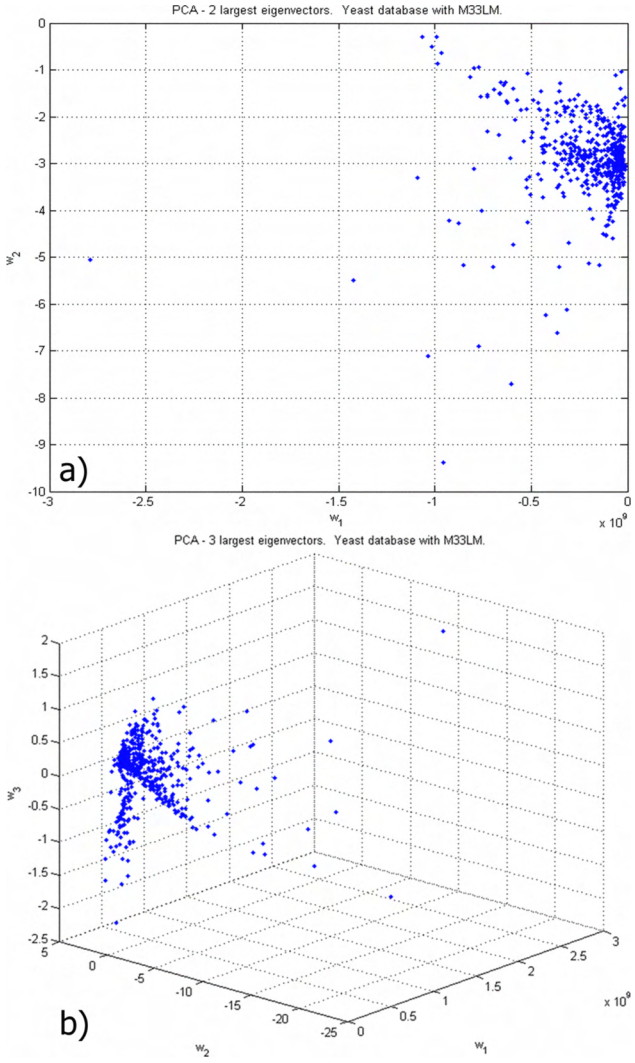

Fig. 3. In a) is a plot of the 2 largest eigenvectors using PCA for feature analysis. A plot of the 3 largest eigenvectors using PCA for feature analysis is shown in b).

be used keeping all its eigenvectors and eigenvalues. No dimensionality reduction will be performed.

IV. Self Organizing Maps Clustering

It is well known that the Self Organizing Maps (SOM) architecture is constructed by a competitive non-supervised algorithm [15]. Kohonen's algorithm tries to approximate each neuron to the input pattern and at the end of each iteration, the closest neuron to the input pattern is the winner. The weights of the network are adjusted to all the neuron's neighborhood commonly using a Gaussian function.

The basic Kohonen algorithm is defined as

$$D = E\left[f_i\left(\mathbf{x}\right)\right] \tag{16}$$

$$f_i\left(\mathbf{x}\right) = \sum_{j=1}^{M} \varphi_{cj}\left(t\right) \left\|\mathbf{x}\left(t\right) - \mathbf{w}_j\right\| \tag{17}$$

where $E\left[\cdot\right]$ denotes the expected value, $\mathbf{x} \in \mathbf{X}$ is the input vector given by the problem, $\mathbf{w}_j$ are the synaptic weights at coordinates (k_1, k_2), $\varphi_{cj}\left(t\right)$ is a neighborhood function, having the neuron map c, and $c\left(c_1, c_2\right)$ as the coordinate of the winning neuron, while M corresponds to the number of the winning neuron. The neighborhood function is a Gaussian defined as

$$\varphi_{cj}\left(t\right) = e^{-\frac{\left\|d_c - d_j\right\|^2}{2\sigma\left(t\right)^2}} \tag{18}$$

where d_c is the location of the winning neuron, d_j is the location of the jth neuron, and $\sigma\left(t\right)$ is the variance of the neighboring neurons at time t. The variance will be decreasing as time increases in order to control the neighborhood size among neurons at a given time t.

The expected value $E\left[f_i\left(\mathbf{x}\right)\right]$ is similar to the k-means algorithm in the sense that, if we remove the neighboring function $\varphi_{cj}\left(t\right)$ then, the remaining equations are similar to the mathematical formulation of the k-means algorithm [16]. Thus, the SOM can be also defined as a setting M number of cluster center (according to M neurons) and organizing them in a SOM lattice array. Then, using $\min\left(E\left[f_i\left(\mathbf{x}\right)\right]\right)$, these clusters are being updated until a stop criterion is reached. The formation of M micro-clusters will be merged with M clusters to gain the final result.

To perform the SOM analysis, we start by extracting a feature vector according to Equation (15). For the yeast data base a total of 547 training vectors are obtained. As described in the previous subsection, the feature vectors are processed using the PCA/KLT approach. This will allow us to have uncorrelated data. The dimensionality is kept as in the original problem. In addition, the data is normalized to have a dynamic range in the $[0\ 1]$ interval. We found that the resulting reference vectors had similar dynamic ranges. This represented an advantage since the numerical accuracy was improved in all of our experiments.

From our past successful implementations on SOM [15] and from numerous applications in the literature, we know that an appropriate size of the SOM map is ten times the size of the feature vector. Therefore in our case, since we have fourteen features, the size of the map should be 140. For convenience, a map of size $[12 \times 12]$ was chosen (since $[11 \times 11] = 121$, and would be under the recommended size). To initialize the weights of the SOM we used the well known approach that suggests to initialize the weights $\{\mathbf{w}_j\left(0\right)\}_{j=1}^{M}$ from the available set of feature vectors $\{\mathbf{x}_i\left(0\right)\}_{i=1}^{N}$ in a random manner. The advantage of this approach is that the initial map will be in the range of the final map.

The implementation of our cluster analysis method is summarized in Fig. 4. In this algorithm the microscopic images have to go through a restoration process to remove distortions introduced by the acquisition process. In this case, we use the M33LM algorithm proposed by the authors in [17].

Fig. 5 shows the hexagonal topology created in this problem and the corresponding connections with the neighboring

Algorithm DISCOVER CLUSTERS

Step 1. Perform image restoration in the images using the M33LM algorithm.
Step 2. Construct the feature vector from the restored images, in the form of Equation 18 using the coefficients in Table I.
Step 3. Perform PCA/KLT and project the data using the transformation matrix given by PCA/KLT.
Step 4. Construct a SOM of size of [12 × 12].
Step 5. Select the *Hexagonal* topology for the map.
Step 6. Train the SOM network with the 547 feature vectors.
Step 7. Compute Fisher Criterion.

Fig. 4. Algorithm to perform cluster discovery. A summary of the main steps.

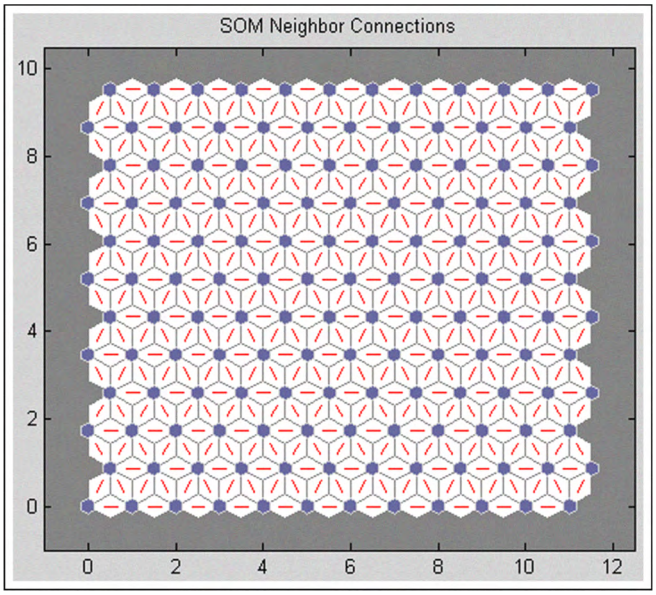

Fig. 5. The neighbor neurons and its corresponding connections in the SOM.

neurons. As the training is performed, the weighted connections between neurons (synapses) change approximating the input pattern. In Fig. 6 are shown the synapses after training was finished; as can be seen there is a strong separation of the neurons indicated by darker colors confirming the presence two classes. Another useful figure that tells us how many features are associated exactly with each neuron (hits) is presented in Fig. 7. It is best if the hits are fairly and evenly distributed across the neurons. In this case, the data is concentrated a little more in the lower-left neurons, but overall the distribution is fairly few and even [18]. The fact that only a few neurons are exactly hitting the input pattern, it is very good, since it is a signal of the good generalization capabilities of the network.

The quality of the clusters generated by the SOM can be assessed by the Fisher Ratio. The Fisher's criterion computes the separation between classes and the inner reliability of the classes at the same time. A feature should be more discriminative when the Fisher's ratio is higher. Good

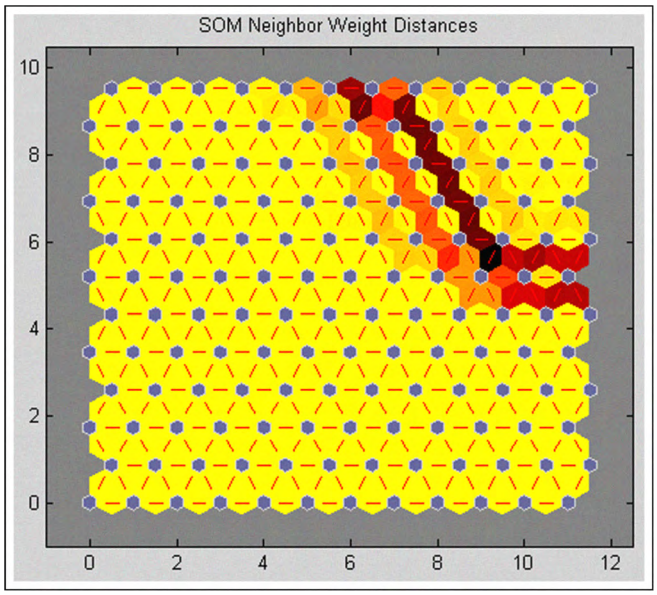

Fig. 6. The distance between the weights in the SOM. Blue hexagons represent the neurons; red lines connect neighboring neurons; colors in the regions containing the red lines indicate the distances between neurons. The darker colors represent larger distances. The lighter colors represent smaller distances. A group of light segments appear in the lower-left region, and at the top-right there is other group shown as darker segments. This grouping indicates that the network has clustered the data into two groups (classes).

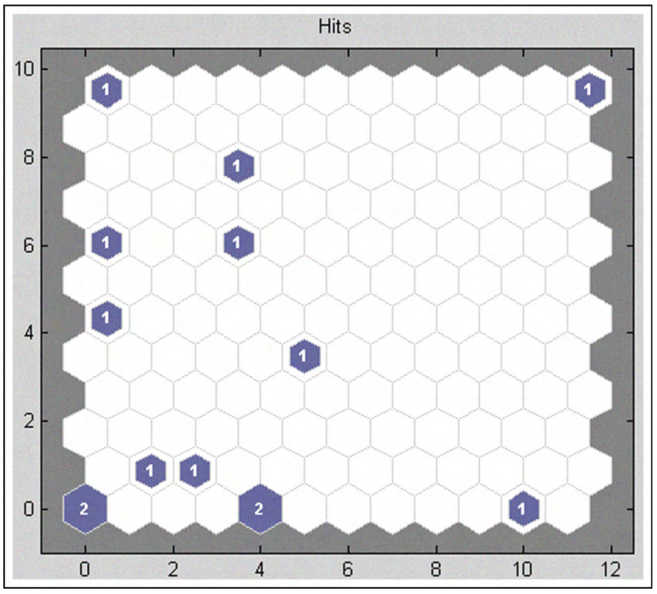

Fig. 7. Hits Plot, shows where the neurons match exactly the input pattern. As desired, neurons are hitting fairly and evenly across the map. The network has good generalization since only few neurons fit exactly the input pattern.

features must have high class mean values as well as high reliability [19]. The generalized Fisher ratio is given by

$$F = \frac{\frac{1}{N}\sum_{j=1}^{N}(m_j - \overline{m})^2}{\frac{1}{NP}\sum_{j=1}^{N}\sum_{i=1}^{P}(x_{ij} - m_j)^2}, \quad (19)$$

where

$$\overline{m} = \frac{1}{N}\sum_{j=1}^{N} m_j \quad (20)$$

is the mean of all the means. At this point, we know how many classes are because of the use of the SOM. There are two classes distributed in the hyperplane. Therefore, from the labeled data we can quantify the decorrelation of the data. After the implementation, we have concluded that the two classes found by the SOM, are well defined, reliable and discriminant to each other according to the results shown in Table II.

TABLE II
FISHER RATIO RESULTS.

Sum of Fisher Ratio	1224.8
Mean Fisher Ratio	2.2391

V. CONCLUSIONS

We have presented 14 coefficients that exist in the literature for colocalization analysis through fluorescence microscopy. Most of the coefficients assume images with zero background. We used our algorithm M33LM [17] for image restoration, before feature extraction.

After the features were extracted with the restored images, a feature vector analysis was performed in order to verify the structure and discriminant properties of the data. We have implemented the PCA/KLT to decorrelate the data in the hyperplane. Then, we presented an application of the SOM for non-trivial class discovery. After feature vector projection into the KLT domain, a SOM network was constructed with a map size of 12×12 and a *hexagonal* topology. As a result, two classes were found.

The significance of this paper relies on the fact that an automatic classification method can be derived from this work. Moreover, to the best of our knowledge this is the first approach to solve the colocalization problem from an objective machine learning point of view. Furthermore, subjectivity in the classification can be minimized with this approach. The SOM provided with two group of neurons that represent two patterns in the feature induced hyperplane: colocalization and no-colocalization.

The development of such a classifier that provides meaningful results to researchers in Biological Sciences is the subject of our current work.

REFERENCES

[1] S. Bolte and F. P. Cordelieres, "A guided tour into subcellular colocalization analysis in light microscopy," *Journal of Microscopy*, vol. 224, no. 3, pp. 213-232, December 2006. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2818.2006.01706.x

[2] S.V. Costes, D. Daelemans, E.H. Cho, Z. Dobbin, G. Pavlakis, S. Lockett, "Automatic and quantitative measurement of protein-protein colocalization in live cells. *Biophysical Journal*. 2004 Jun;86(6):3993-4003.

[3] Q. Wu, F. Merchant and K. R. Castleman, Microscope Image Processing. Amsterdam ; Boston: Elsevier/Academic Press, 2008, pp. 548.

[4] P. M. Lukas Landmann , "Colocalization analysis yields superior results after image restoration," *Microsc. Res. Tech.*, vol. 64, pp. 103-112, 2004.

[5] MediaCybernetics. (2002, 22/03/2002). Application note 1: Colocalization of fluorescent probes. [On Line]. 2008(November/01), pp. 20. Available: http://www.spectraservices.com/Merchant2/pdf/AppInfo-CoLocFluorProbes.pdf

[6] Z. Wang. (2006, 26/05/2006). How to quantify the colocalization in your images? [On Line]. 2008(November/01), pp. 7. Available: http://dbc.bio.uci.edu/pdf_documents/Colocalization.pdf

[7] V. Zinchuk, O. Zinchuk and T. Okada, "Quantitative colocalization analysis of multicolor confocal immunofluorescence microscopy images: pushing pixels to explore biological phenomena," *Acta Histochem. Cytochem.*, vol. 40, pp. 101-111, Aug 30. 2007.

[8] CoLocalization Research Software. (2008, 02/11/2008). CoLocalizer pro user guide. ver. 2.5. [On Line]. 2008(November/01), pp. 67. Available: http://homepage.mac.com/colocalizerpro/Resources/CoLocalizer Pro User Guide.pdf

[9] McMaster Biophotonics Facility. (2006, 30/11/2006). Colocalization. [On Line]. 2008(November/01), pp. 20. Available: http://www.macbiophotonics.ca/PDF/MBF_colocalisation.pdf

[10] Indiana Center for Biological Microscopy. (2007, 10/05/2007). Analysis of colocalization using metamorph. [On Line]. 2008(November/01), pp. 7. Available: http://www.nephrology.iupui.edu/imaging/tutorials/Analysis of Colocalization using Metamorph.pdf

[11] J. Adler, F. Bergholm, S. Pagakis and I. Parmryd, "Noise and Colocalization in Fluorescence Microscopy: Solving a Problem," *Microscopy and Analysis*, vol. 22, Sep/2008. 2008.

[12] L. F. Agnati, K. Fuxe, M. Torvinen, S. Genedani, R. Franco, S. Watson, G. G. Nussdorfer, G. Leo and D. Guidolin, "New Methods to Evaluate Colocalization of Fluorophores in Immunocytochemical Preparations as Exemplified by a Study on A2A and D2 Receptors in Chinese Hamster Ovary Cells," *J. Histochem. Cytochem.*, vol. 53, pp. 941-953, August 1. 2005.

[13] J. Adler and I. Parmryd, "Letter to the Editor," *J. Microsc.*, vol. 227, pp. 83-83, 2007.

[14] R. Howson, W. K. Huh, S. Ghaemmaghami, J. V. Falvo, K. Bower, A. Belle, N. Dephoure, D. D. Wykoff, J. S. Weissman and E. K. O'Shea, "Construction, verification and experimental use of two epitope-tagged collections of budding yeast strains," *Comp. Funct. Genomics*, vol. 6, pp. 2-16, Feb. 2005.

[15] M. I. Chacon and P. Rivas-Perea, "Performance Analysis of the Feedforward and SOM Neural Networks in the Face Recognition Problem," *Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007. IEEE Symposium on,* pp. 313-318, 2007.

[16] L. Qiang, Y. Jin-Shou, "Fuzzy Self-Organizing Map Neural Network Using Kernel PCA and the Application", *Advances in Natural Computation*, Springer Lecture Notes in Computer Science. Vol 3610/2005. pp 81-90.

[17] P. Rivas-Perea, J.G. Rosiles, and W. Qian, "Image Restoration for Quantitative Colocalization: Performance Analysis and Response Of Colocalization Coefficients," *Biomedical Imaging: Nano to Macro, 2009. 6th IEEE International Symposium on , Under Review*, June 2009.

[18] H. Demuth, Neural Network Toolbox. User Guide. ,v6 ed.U.S.A.: The Math Works Inc., 2008, pp. 315.

[19] Maravall, D. Reconocimiento de Formas y Visión Artificial. RA-MA, 1993.