CELL CYCLE STAGING FROM FLUORECENCE DAPI IMAGES

Ivan Sahumbaiev[1*], *Anabela Ferro*[2,3*], *Tânia Mestre*[1], *Raquel Seruca*[2,3,4] *and J.Miguel Sanches*[1]

[1]Institute for Systems and Robotics (ISR/IST), LARSyS, IST, Universidade de Lisboa, Portugal
[2]IPATIMUP, Institute of Molecular Pathology and Immunology, University of Porto, Portugal
[3]Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal
[4]Faculdade de Medicina da Universidade do Porto, Portugal
*Authors with equal contributions; corresponding author: jmrs@tecnico.ulisboa.pt

ABSTRACT

Cell cycle staging information is very important in a wide range of biological problems. Here, we proposed a new method based on 4',6-diamidino-2-phenylindole (DAPI)-stained *in situ* fluorescence microscopy images. It is known that the intensity and size of nuclei are discriminative features for cell cycle staging. A clustering analysis, herein computed from DAPI-stained nuclei, of a population of cells is performed using a two stage classifier.

The first stage, a modified version of the *k*-means classical unsupervised method, incorporates the *a priori* information about the expected DNA amount in G_1 and G_2 cell cycle phases. The labels obtained in the first step feed a Gaussian Mixture Model (GMM) classifier that tunes the final shape and separation hyperplanes of the clusters.

The obtained results are consistent with the typical distribution of cells by the G_1, S and G_2 stages described in the literature. Further, they are molecularly validated by a Fucci system. This new image analysis method is suitable for a rapid and inexpensive estimation of cell cycle staging of biological samples, while preserving the nature of analyzed cells.

Index Terms **–** Cell Cycle Staging, DAPI, Fluorescence Microscopy Images, Unsupervised Classification

1. INTRODUCTION

In the mitotic cell cycle, resting/quiescent cells are in G_0 phase and are diploid, owning two sets of chromosomes (2N). Diploid cells that initiate the cell cycle are in G_1 phase and then proceed to an intermediate phase of synthesis (S phase) during which DNA and protein content is doubled. Upon completion of this phase, cells are at G_2 phase and are tetraploid (4N). At this point, mitosis (M phase) ensues and cells split in two new, diploid daughters cells [1]. As cells cycle through each division, the surveillance of the fidelity of this process is fundamental [2]. Thus, the determination of the cell cycle *momentum* may point out important biological cues on the physiological status of individual or subpopulations of cells. Several methods have been developed and used to quantify DNA content in biological samples. DNA content analysis progressed from highly laborious and time-consuming methods to faster and highly quantitative techniques [1]. The accessibility to DNA fluorophores that bind stoichiometrically to DNA (*e.g.* DAPI) allowed the direct quantitative estimation of specimen's DNA content; this principle has been largely explored in methods as flow cytometry, image analysis and laser scanning cytometry. Flow cytometry has been, undoubtedly, the most disseminated cell analysis method [3]. Nevertheless, due to its fluidic principle, one of the major drawbacks of flow cytometry is the need of disaggregation of cellular samples, a critical aspect for rare, invaluable biological samples. As such, computer-based fluorescence image analysis presents itself as a very reliable, accurate and cost-effective solution, allowing quantitative measurements of multi-color staining and DNA content in virtually any biological specimen [4].

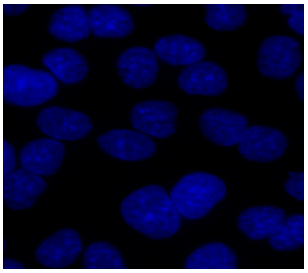


Figure 1 - *DAPI-stained nuclei fluorescence image*

In this work we propose a fluorescence image-based algorithm targeted at identification/segmentation of DAPI-stained nuclei and, subsequently, the determination of DNA content of single or sub-populations of cells, without disrupting the natural organization of the *in vitro* culture systems; assessment of cell cycle dynamics of sub-populations of cells in *fluorescence microscopy* (FM)

This work was supported by FCT projects UID/EEA/50009/2013, PTDC/BIM-ONC/0171/2012, PTDC/BIMONC/0281/2014, PTDC/BBB-IMG/0283/2014 and by FEDER funds through the Operational Programme for Competitiveness Factors (COMPETE) (PEst-C/SAU/LA0003/2013). AF is a FCT fellowship recipient (SFRH/BPD/97295/2013). We thank João Vinagre for the imaging acquisition assistance.

images was also possible. The bioimaging analysis tool relies on DAPI binding characteristics [5] and is based, in an unsupervised manner, on the quantification of the area and total intensity of DAPI-stained nuclei of acquired FM images as shown in Figure **1**.

2. MULTI-LABEL UNSUPERVISED CLASSIFICATION

This section describes the method of classification of each cell within the FM in three active phases: G_1, S and G_2.

2.1. Data acquisition

The *in vitro* cultures analyzed comprise normal murine mammary gland epithelial (NMuMG)-Fucci2. These cells were fixed, permeabilized and cell's nuclei were stained with DAPI. Images were then captured with Zeiss Apotome Axiovert 200M ImagerZ1 fluorescence microscope with 40X/1.3 oil DICII (UV) VIS-IR objective **(**Carl Zeiss, Thornwood, NY). Aiming at extracting complete information on DAPI-stained nuclei, multiple images in different planes along the z-axis (60 stacks) were acquired and then merged together by projecting into a single image. The acquisition parameters (exposition time and maximum pixel intensity) were maintained constant in all experiments. FM images were then analyzed with the following cell cycle staging pipeline: 1) image pre-processing for contrast adjustment and enhancement; 2) nuclei segmentation based on the Otsu's method; 3) area and total intensity computation of each nucleus, 4) multi-label classification of each nucleus in G_1, S and G_2 classes with an unsupervised algorithm that incorporates *a priori* information about the expected amount of DNA in each stage. The core of the method is a cascade of a modified *k*-means unsupervised classifier **Error! Reference source not found.**, followed by a modified EM supervised classifier [6] fed by the labels of the first classifier.

2.1. Image pre-processing pipeline

One of the major issues faced with nuclei segmentation algorithms is the incorrect segmentation of nuclei, usually associated with considerable nuclei proximity and overlapping [7]. To minimize this difficulty, several times associated with wide blurred regions at the borders of the nuclei we applied to each FM image a contrast adjustment to enhance and sharpen the morphological borders of each nucleus. After this pre-processing operation a simple Otsu segmentation procedure is able to provide enough segmentation quality results to isolate each nucleus. Subsequently, quantitative features of each i^{th} cell (area, A_i, and total intensity, TI_i), that reflect the underlying biological changes occurring in the nucleus during the whole cell cycle (including replication of DNA content) are computed as follows

$$\mathrm{A}_i = \sum_j \pi_{ij} \tag{1}$$

$$\mathrm{TI}_i = \sum_j x_{ij}\pi_{ij} \tag{2}$$

where π_{ij} and x_{ij} are the j^{th} elements of the binary mask, π, and image intensity, x, of the i^{th} cell in the image. Finally, due to the correlation between these two features, all data was normalized by z-score value [8].

2.2. Classification/Staging

The multi-label classification strategy used here is unsupervised because in these typical problems of cell cycle from DAPI there is no information about the population of cells.

The goal is to estimate a vector of labels $\boldsymbol{c} = [c_1, c_2, \dots c_N]^T$ where $c_k \in \{G_1, S, G_2\}$ where N is the number of nuclei in the image.

A classification approach described in this paper is performed using a two-stage classifier. The first stage, a modified version of the *k*-means classical unsupervised method, incorporates the *a priori* information about the expected DNA amount in G_1 and G_2 cell cycle phases. The labels obtained in the first step initialize a Gaussian Mixture Model (GMM) classifier that tunes the final shape and separation hyperplanes of the clusters.

In this type of image modality the image intensity results form the binding of DAPI to A-T rich regions in DNA which is a stochastic process that is appropriately described by a normal distribution. In the 2D-space of the features, (A, TI), each class is well described by an ellipsoid because both features are strongly correlated. Thus, the population is described by a Gaussian Mixture Models (GMM), where each k^{th}component is a multivariate normal distribution

$$\mathrm{p_k}(\mathrm{x_i}|\mu_\mathrm{k}, \Sigma_\mathrm{k}) = \frac{\exp\left\{-\frac{1}{2}(\mathrm{x_i}-\mu_\mathrm{k})^\mathrm{T}\Sigma_\mathrm{k}^{-1}(\mathrm{x_i}-\mu_\mathrm{k})\right\}}{(2\pi)^{d/2}\sqrt{|\Sigma_\mathrm{k}|}} \tag{3}$$

where $x_i = (A_i,\ TI_i)^T$ is the vector of observations/features. μ_k is the mean and Σ_k is the covariance matrix that defines the geometrical features of the cloud/cluster, such as shape, volume and orientation. The covariance matrix, can be decomposed as follows [9],

$$\Sigma_\mathrm{k} = \alpha_\mathrm{k}\mathrm{D_k}\mathrm{A_k}\mathrm{D_k^T} \tag{4}$$

where $\mathrm{D_k}$ is orthogonal, $\mathrm{A_k}$ is a diagonal matrix of the eigenvalues of Σ_k and α_k is a scalar. The $\mathrm{D_k}$ matrix defines the orientation of the cloud, $\mathrm{A_k}$ defines its shape and α_k defines the volume.

The implementation of the cluster analysis via Gaussian Mixture Model (GMM) was performed through the Expectation-Maximization (EM) algorithm [11] using the probabilistic model described in (3). In this algorithm, indicator (binary) variables, z_{ik}, that are equal to 1 if the i^{th} point is classified in k^{th} cluster and 0 otherwise, are estimated in a two step iterative algorithm. z_{ik}, the indicator variables, and $\mathrm{c_i}$, the classification results, are related by $c_i = \sum_{k=1}^{3} \mathrm{k}\, \mathrm{z}_{ik}$.

In the first step latent membership weights are estimated and used in the second step of maximization of the likelihood function. Both steps alternate until convergence is

achieved. In each iteration the covariance matrix Σ_k^n is computed as follows:

$$\Sigma_k^{n+1} = \Sigma_k^n + A_k^n w_k^n \quad (5)$$

where Σ_k^n is the current estimation computed in the previous iteration, A_k^n is the diagonal matrix defined in (4) and w_k^n is the current relative amount of data in k^{th} cluster. This incremental estimation of the covariance matrix aims at dealing with the ill-conditioned nature of the covariance matrix specially when the current number of points in the corresponding cluster is small. That is the reason why the updating term in (5) depends on the relative importance of the k^{th} cluster in the current iteration. Figure 2 displays an example of the evolution in the estimation of the clusters with this incremental strategy to estimate the parameters of the clusters.

The initialization of the EM algorithm, $\boldsymbol{c^0}$, is obtained using a modified version of the classical *k*-means strategy with the Euclidean distance. Here a constraint based on the biology of the problem is imposed. It is known that the amount of DNA in G_2 is twice of the amount of DNA in G_1. Considering this, we assumed that the centroid of G_2 is twice of G_1,

$$\mu_{G_2} = 2\mu_{G_1} \quad (7)$$

where μ_{G_1}, μ_{G_2} are the means of the corresponding clusters.

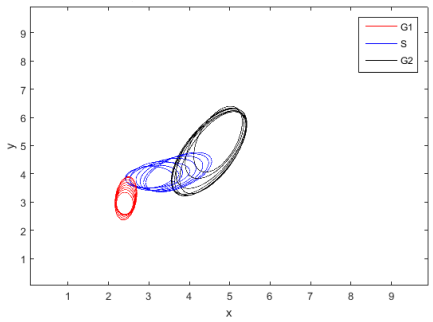


Figure 2 – *Graphic representation of the clustering analysis depicting the changes of shape, volume and orientation of the covariance matrix from the cell cycle classifier. Red cluster corresponds to cells in G_1 phase, blue cluster to cells in S phase and green cluster to cells in G_2 phase.*

The EM algorithm is highly sensitive to the amount and distribution of the input data. This sensitivity is strongly related with the determinant of matrix Σ_k involved in Equation (3) that can be ill-conditioned, especially if the number of elements in the cluster is too small, e.g. $N_k = 1$. To avoid this problem spectral decomposition [12] of Σ_k is used in its inversion during the M-step of the EM algorithm.

3. RESULTS

Forty seven images with a total of 998 DAPI-stained nuclei are used for experimental data. For molecular validation proposes we compared each cell in terms of FUCCI system and DAPI classification. This system allows a colored readout of the cell cycle progression based on the expression of cell cycle-specific fluorescent markers, thus allowing the discrimination between non-proliferative (G_0/G_1) and proliferative (S and G_2/M) phases of cells [13] .

In Figure 3 an example of classification of the image in

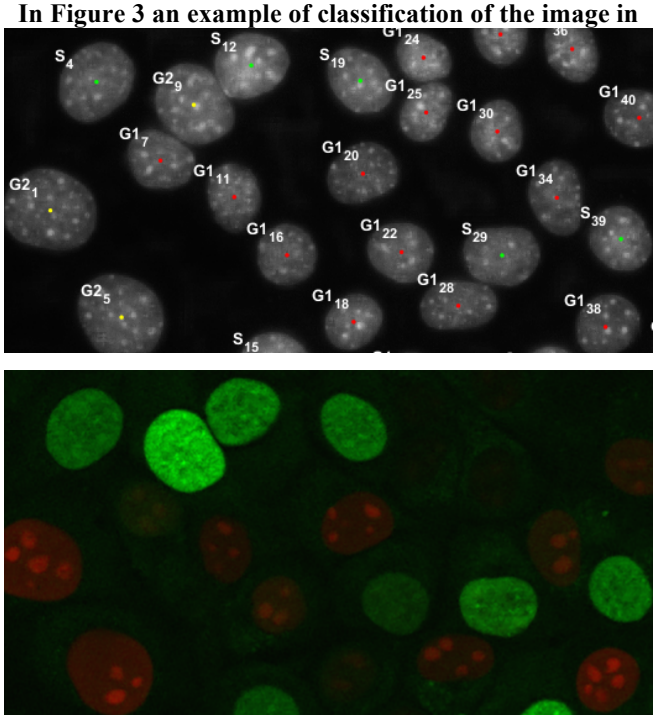


Figure 5 is displayed where each point in the feature space, $(A_i, TI_i)^T$, corresponds to a single cell. In this case 62% of nuclei were classified in G_1, 29% in S and 9.5% in G_2 as shown in the histogram displayed in Figure 4. This histogram is similar to the typical ones described in the literature [14].

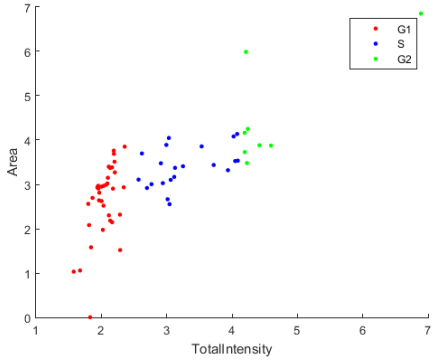


Figure 3 – *Distribution of cells after classification. Red dots correspond to cells in G_1 phase, blue dots to cells in S phase and green dots to cells in G_2 phase.*

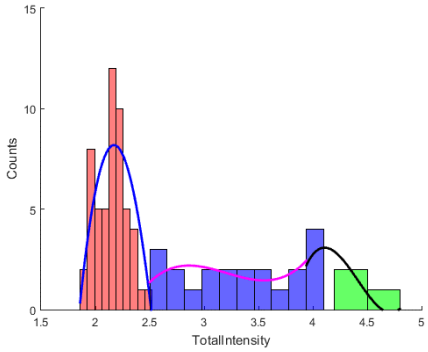
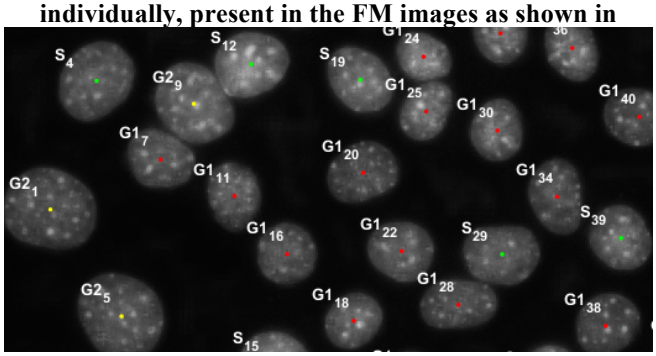


Figure 4 – *Histogram of the distribution of cells after classification. Red histogram and blue curve correspond to cells in G_1 phase. Blue histogram and red curve to cells in S phase. Green histogram and black curve to cells in G_2 phase.*

The red cluster corresponds to the G_1, in which the areas and total intensity of the nuclei are smaller than in the other phases because in this phase the amount of chromosomes is N, corresponding to the smaller amount of DNA among the all phases. The blue cluster corresponds to the S phase, where values of area are almost constant, whereas an increase of total intensity is observed, which is typical of replicative nuclei where the DNA is replicating toward G_2. The green cluster corresponds to G_2 where the DNA already has already replicated and the number of chromosomes is 2N and the amount of DNA is double of G_1. Furthermore, the segmentation algorithm revealed a high efficiency and throughput; the average time of analysis for one FM image is approximately ten seconds, with 97% of recognized cells.

The designed method allows the operator to retrieve, from the clustering diagram, each nuclei analyzed, individually, present in the FM images as shown in



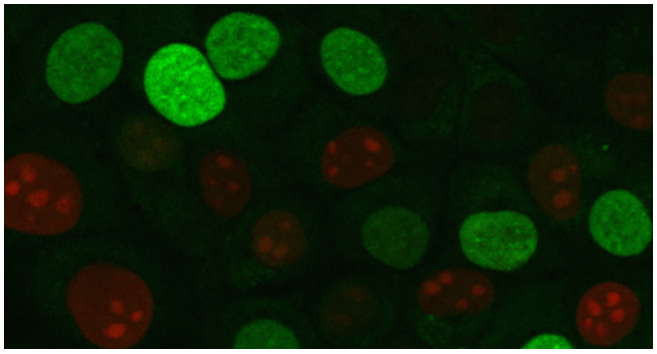


Figure 5.

Finally, the comparison between the algorithm herein described and Fucci system shows that 89% of G_1 Fucci cells and 86% of Fucci cells at $S/G_2/M$ phase were correctly classified by the new procedure based only in DAPI-stained nuclei. Moreover, this algorithm correctly assigned 93% of Fucci cells transitioning from G_1 to S phase, as either in G_1 or S phase. Despite the limitations of Fucci system, our analysis biomaging pipeline using DAPI showed an overall accuracy of 94.5% for cell cycle stage classification.

4. CONCLUSIONS

In this work, a new unsupervised algorithm for cell cycle phase determination based on DAPI-stained FM images is presented. The first step of the proposed methodology consists on the processing of FM images, aiming to extract the area and total intensity of each DAPI-stained nucleus. The second step consists on the classification algorithm. This algorithm is based on a modified EM algorithm, where each cluster is modeled by a covariance matrix with a set of geometrical parameters. Also, an updating rule for the covariance matrix was pursued in order to define the cluster according to the number of points. To avoid random initialization, a modified *k*-means strategy was applied. This stage allowed the incorporation of prior biological background regarding the cell cycle. Moreover, we have connected the sensitivity of the EM algorithm to the input data by applying a SVD approach in the calculation of the inverse covariance matrix, which allowed us to increase the algorithm's classification and improve its stability on cell cycle phasing. Remarkably, the obtained results are consistent with the typical distribution of cells by the G_1, S and G_2 stages extensively described in the literature. Interestingly, our new bioimaging tool allows, also, the possibility to perform cell cycle dynamics analysis of cell populations, as well as of single cells. The application of low cytotoxic, cell membrane-permeant fluorescent dyes (*e.g.*, vybrant® dye cycle™ stains) turs this new cell cycle classifier suitable for *in vivo* applications.

5. REFERENCES

[1] G. Cooper. "The cell: A molecular approach. 2nd edition." *Sunderland (MA): Sinauer Associates. The Eukaryotic Cell Cycle* (2000). Available from: "http://www.ncbi.nlm.nih.gov/books/NBK9876/", accessed on 28.03.2015.

[2] Green, Douglas R., and Gerard I. Evan. "A matter of life and death." *Cancer cell* 1.1 (2002): 19-30.
[3] Godfrey, W. L., et al. "Complementarity of Flow Cytometry and Fluorescence Microscopy." *Microscopy and Microanalysis* 11.S02 (2005): 246-247.
[4] Lichtman, Jeff W., and José-Angel Conchello. "Fluorescence microscopy." *Nature methods* 2.12 (2005): 910-919.
[5] Neidle, Stephen. "DNA minor-groove recognition by small molecules." *Natural product reports* 18.3 (2001): 291-309.
[6] McLachlan, G.J. "The EM algorithm and extensions." John Wiley & Sons, Inc. 2008
[7] Wählby, Carolina, et al.,Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. *Journal of Microscopy* 215.1 (2004): 67-76.
[8] G. Guojun, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*. Vol. 20. Siam, 2007.
[9] A. Jain, A. Bajpai, and M. Rohila, Efficient Clustering Technique for Information Retrieval in Data Mining, *Int. Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, Volume 2, Issue 6* (2012).
[10] J. Banfield, and Adrian E. Raftery. "Model-based Gaussian and non-Gaussian clustering." *Biometrics* (1993): 803-821.
[11] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the royal statistical society. Series B (methodological)* (1977): 1-38.
[12] Moon, Todd K, and Wynn C. Stirling. Mathematical Methods and Algorithms for Signal Processing. Upper Saddle River, NJ: Prentice Hall, 2000.
[13] Sakaue-Sawano et al., Visualizing spatiotemporal dynamics of multicellular cell-cycle progression, Cell. 2008 Feb 8; 132(3):487-98
[14] Cell cycle staging of individual cells by fluorescence microscopy. Vassilis Roukos, Gianluca Pegoraro, Ty C Voss & Tom Misteli.Nature Protocols 10, 334–348 (2015)

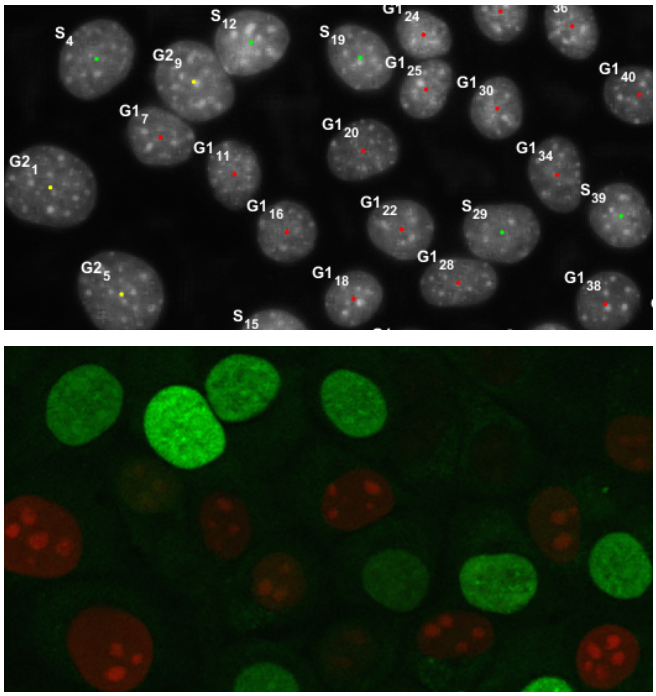


Figure 5 – *The upper image shows nuclei classified according to the output of the proposed algorithm. Red dots correspond to cells in G_1, blue dots correspond to cells in S phase and green dots are cells in G_2. The lower image shows the same nuclei stained with the Fucci system. Cells in G_0/G_1 are red and cells in S and G_2/M (proliferative phases) are green or yellow.*