A MODEL OF THE SPATIAL MICROENVIRONMENT OF THE COLONIC CRYPT

Violeta N. Kovacheva[⋆] *David Snead*[†] *Nasir M. Rajpoot*[§,‡]

[⋆] Department of Systems Biology, University of Warwick, UK
[†] Department of Histopathology, University Hospitals Coventry and Warwickshire, UK
[§] Department of Computer Science and Engineering, Qatar University, Qatar
[‡] Department of Computer Science, University of Warwick, UK

ABSTRACT

There have been great advancements in the field of immunofluorescence imaging. The surge in development of analytical methods for such data makes it crucial to develop benchmark synthetic datasets for objectively validating these methods. We propose a model of the healthy colonic crypt microenvironments. Our model can simulate immunofluorescence image data with parameters that allow control over cellularity, cell overlap ratio, image resolution, and objective level. To the best of our knowledge, ours is the first model to simulate immunofluorescence image data at subcellular level for healthy colon tissue, where the cells have several compartments and are organized to mimic the microenvironment of tissue *in situ* rather than dispersed cells in a cultured environment. Validation of the model has been performed by comparing morphological features of the tissue structure between real and simulated images. In addition, we compare the performance of two cell counting algorithms. The simulated data could also be used to validate techniques such as image restoration, cell segmentation, and crypt segmentation.

***Index Terms*—** Immunofluorescence, colon tissue, spatial model, synthetic data

1. INTRODUCTION

Fluorescence microscopy combined with digital imaging constructs a basic platform for numerous biomedical studies in the field of cellular imaging. As studies relying on analysis of digital images become popular, the validation of such analytical tools gains significance. A common approach for validation is to compare the algorithm's results with expert-labelled data. Nevertheless, the repeatability and accuracy of manual labeling can always be questioned due to human error sources [1] and the process is very time-consuming. In addition, developing a model can provide a better understanding of the system and aid experimental design. We propose a method for simulating fluorescent images for the spatial microenvironment of healthy colon tissue. The tissue microenvironment is composed of a single layer of epithelium forming glandular structures, called crypts (as shown in Figure 1). The tall columnar epithelial cells have oval basal nuclei. Stroma fills the space between the crypts and contains several types of cells, such as lymphocytes, plasma cells and fibroblasts.

Several frameworks for synthetic fluorescence image data generation have been proposed in the literature. The simplest of these simulate populations of spots as spheres [2] or Gaussian-like 3D objects [3]. Lockett *et al.* [4] used a more complex set of shapes, such as curved spheres, discs, and dumbbells. More recently, realistic simulations have been presented. For example, Lehmussola *et al.* [5] designed a simulator called SIMCEP, which can simulate large 2D cell populations with realistic looking cytoplasm, nuclei and cell organelle. Svoboda *et al.* generated a model to simulate fully 3D image data of cell populations [6] and later of healthy colon tissue [7]. However, these models only included cell nuclei. In addition, the shape of the nuclei in the colon tissue model [7] is not very realistic and does not reflect the variety of cells phenotypes found in real tissue. On the other hand, Zhao and Murphy [8] presented a machine learning method to generate realistic cells with labeled nuclei, membranes and a protein expressed in a cell organelle. However, this approach is restricted to individual cells and only one protein of interest at a time.

In this work, we expand the SIMCEP tool in order to simulate images for the healthy colon tissue. There are different cell types, characterized by different appearance and localization of the cells. The cells are organised to mimic the colon tissue structure. Hand-marked histology data has been used in order to generate realistic chromatin texture, nuclei morphology, and crypt architecture. Our model also incorporates various cells phenotypes found in real histology data. We have developed this 2D model with the foresight of extending it to simulate multiplex fluorescence and histology images.

2. METHODS

2.1. Real data input

In order to make the model realistic, Hematoxilyn and Eosin (H&E) slides from colon cancer patients were analysed. Healthy (as graded by two pathologists) visual fields at 40×

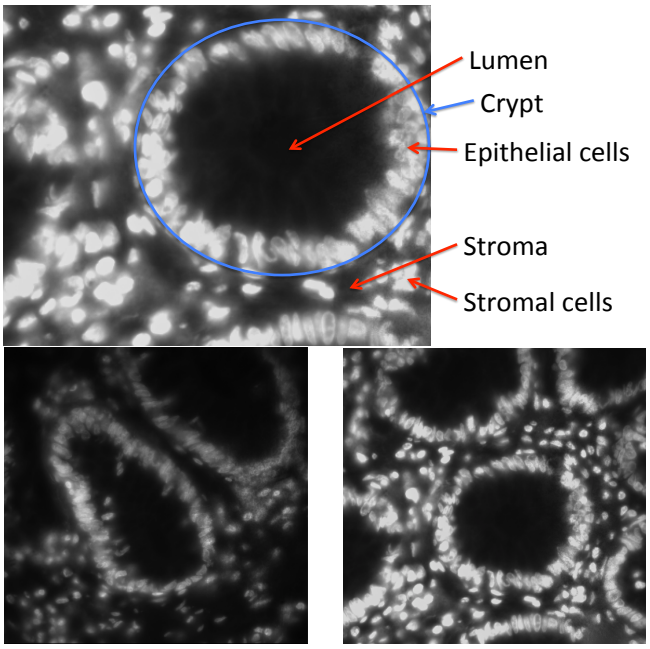


Fig. 1. Immunofluorescent images with nuclear marker DAPI depicting the structure of healthy colon tissue.

magnification were selected. Individual nuclei in each image were hand-marked as epithelial or stromal. Size and 13 Haralick texture features [9] were extracted for each nucleus. Affinity Propagation [10] was used to phenotype the nuclei according to their 13 texture features. For each of the clusters found, the mean and standard deviation of the length of the major axis and the ratio between the minor and major axis were obtained. In addition, we calculated the frequency with which nuclei belonging to each phenotype are found to be epithelial or stromal, and incorporate the phenotype frequency into our model as described in Section 2.3.

In addition to this, visual fields at 20× magnification were selected for analysis of crypt sizes. In these, 585 healthy crypts were hand-marked. We calculated the mean and standard deviation of the minor axis and the ratio between minor and major axes for each group. These were then incorporated into the model.

2.2. Tissue structure

Given an image resolution and magnification level, we first determine an appropriate radius of the cells, r, of 6 μm, and a radius of the crypts corresponding to the mean length on the minor axis, μ_b, found from the H&E images. These parameter values are used to determine the number of crypts and cells to be simulated in the image. The number of crypts, N_c in an $i_h \times i_w$ image is determined by the following:

$$N_c = \lfloor i_h/(3b) \rfloor \lfloor i_w/(3b) \rfloor . \quad (1)$$

Crypts are simulated as elliptical structures. For each crypt, the minor axis b is set to $\mu_b + U(-\sigma_b, \sigma_b)$, where σ_b is the standard deviation of the minor axis found in the H&E images, and $U(x_1, x_2)$ is a number uniformly drawn from the range $[x_1, x_2]$. To determine the length of major axis, a, we use the mean ratio between the minor and major axes, $\mu_e = \mu_b/\mu_a$. Then a is given by $b/(\mu_e + U(-\sigma_e, \sigma_e))$, where σ_e is the standard deviation of the ratio. The degree of rotation of the major axis, ϕ, of the crypts is chosen at random. The crypt outline is then computed as follows,

$$R(\theta) = \frac{ab\sqrt{2}}{\sqrt{(b^2 - a^2)\cos(2\theta - 2\phi) + a^2 + b^2}} + u \quad (2)$$

where $R(\theta)$ is the polar radius, $\theta \in [0, 2\pi]$ is the polar angle and $u = 0.1U(-0.6, 1)$ slightly deforms the crypts.

Then, the crypt centres, $\mathbf{c} = (x_c, y_c)$, are selected so that the crypts don't overlap. The epithelial cells are placed at a random location along the crypt edge. Once the cells are placed, they are rotated so they point towards the crypt centre and their nuclei are displaced closer to the edge of the crypt. The stromal cells are placed uniformly in the space outside the crypts. Stromal cells are rotated in a direction given by $\phi + U(-\pi/6, \pi/6)$ to reflect the structure of the stromal tissue that can be observed in histology images.

The maximum amount of cell overlap is also controlled by the parameter L_{max} [5]. The relative amount of overlap, L_{ij}, that is caused on the region of pixels R_i defined by one simulated cell by the region of pixels R_j of another cell is measured by

$$L_{ij} = \frac{|R_i \cap R_j|}{|R_i|}, i \neq j \quad (3)$$

where $|\cdot|$ is the cardinality of a set. Then, for example, setting $L_{max} = 1$ doesn't pose any restrictions on overlap, whereas $L_{max} = 0$ doesn't allow overlap. Overlap can be controlled either on the cytoplasm or nuclei regions. When a cell is placed randomly, if the overlap criterion is not satisfied, a new set of coordinates is chosen.

Once the number and size of crypts has been determined and the crypts have been placed, we calculate the number of cells, N that will be placed in the image. Firstly, an estimate of the area of a stromal cell, A is calculated:

$$A = \pi((2 - 0.7L_{max})r)^2. \quad (4)$$

Here the factor of r accounts for the effect of overlap and doesn't go below 1 as stromal cells are generally sparse. The area covered by stroma, A_s is found by counting the pixels outside the outlines of the crypts. Then the number of stromal cells is given by $N_s = \frac{\nu_s A_s}{A}$, where $\nu_s \in [0, 1]$ is a user-defined parameter for the cellularity (density) of stromal cells.

Similarly, the number of epithelial cells is determined by

$$N_e = \frac{\nu_e P}{2(1.4 - L_{max})r}, \quad (5)$$

where P is the perimeter of the crypts in the image, $\nu_e \in [0, 1]$ is a user-defined parameter for the cellularity of epithelial cells, and the factor in the denominator accounts for the effects of overlap. The overlap factor is smaller than the one for stromal cells because epithelial cells are more tightly packed. Then the final number of cells is given by $N = N_s + N_e$.

2.3. Single cell

Each of the N cells is constructed separately. Before a cell is synthesized, it is randomly assigned to one of the phenotypes found in the real data with probability equal to the frequency of the phenotypes in H&E tissues.

Two types of shapes are included in the simulation. The cytoplasm for stromal cells, cell nuclei and cell organelle are generated using a parametric model proposed in [5]. In this case the random shapes are initialized as a circle parameterized as $(x(\theta), y(\theta)) = (cos\theta, sin\theta)$, where $\theta \in [0, 2\pi]$ is the polar angle. The angle θ is sampled at k ($k = 10$) equidistant points to generate a regular polygon (Figure 2 (a)). Then a random polygon is created by randomising the spatial locations of the vertices as follows:

$$\begin{aligned} x_i(\theta_i) &= [U(-\alpha, \alpha) + cos(\theta_i + U(-\beta, \beta))], \\ y_i(\theta_i) &= [U(-\alpha, \alpha) + sin(\theta_i + U(-\beta, \beta))] \end{aligned} \quad (6)$$

for $i = 1, ..., k$, where α controls the randomness of the circle radius and β controls the randomness of sampling. Then we obtain the means, μ_l and μ_w, and standard deviations, σ_l and σ_w, for the nuclei major and minor axes, respectively, from the H&E data phenotypes. These are used to obtain the sizes for the modelled nuclei as

$$\begin{aligned} \mu_l^n &= \mu_l + U(-\sigma_l, \sigma_l), \\ \mu_w^n &= \mu_w + U(-\sigma_w, \sigma_w). \end{aligned} \quad (7)$$

Then, the size of the modelled cell cytoplasm is chosen to be

$$\begin{aligned} \mu_l^c &= U(2.5, 2.9)\mu_l^n, \\ \mu_w^c &= U(2.5, 2.9)\mu_w^n. \end{aligned} \quad (8)$$

The user can choose the sizes μ_l^o and μ_w^o of the cell organelles. The polygons are scaled with the respective value as

$$\begin{aligned} \hat{x}_i(\theta_i) &= x_i(\theta_i)\mu_l^{n/c/o}, \\ \hat{y}_i(\theta_i) &= y_i(\theta_i)\mu_w^{n/c/o}. \end{aligned} \quad (9)$$

Finally, the vertices are interpolated using cubic splines (Figure 2 (b)).

The cytoplasm of epithelial cells is generated starting from the polygon shown in Figure 2 (c). The set of original coordinates $\{(x_i, y_i), i = 1, ..., k\}$ are then randomised and scaled

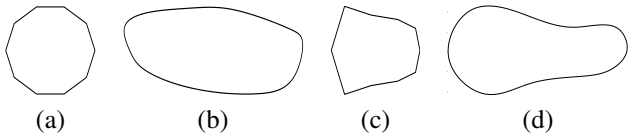


Fig. 2. Examples of cell shapes. Figures (a) and (c) show polygons without any randomness for the stromal and epithelial cells. Figures (b) and (d) show the shapes with dislocated vertices after spline interpolation. Here $\alpha = 0.2, \beta = 0.05$.

$$\begin{aligned} \hat{x}_i &= \mu_l^c(x_i + U(-\alpha/2, \alpha/2)), \\ \hat{y}_i &= \mu_w^c(y_i + U(-\alpha/2, \alpha/2)). \end{aligned} \quad (10)$$

As before, cubic splines are used to interpolate between the vertices. An example of the shape obtained can be seen in Figure 2 (d).

The texture for the cytoplasm and organelle is generated using a well-known procedural model [12] for texture synthesis. As the nuclei texture is an important factor when grading a tumor, a more sophisticated method was adopted for synthesizing it. In particular, we used a parametric model using wavelets [11]. The model is applied to the texture of the nucleus found to be the cluster centroid by Affinity Propagation in order to generate a large texture image. Firstly, the H&E image is converted into grey-scale. Since the method requires a rectangular image as input, the areas outside the nucleus are filled by reflecting at the boundary. For each pixel outside, a pixel equal distance away from the boundary is found and their values are equated. When a nucleus of the given phenotype is being synthesized, a random part of the texture image is selected and used as the texture.

2.4. Measurement error

The final step of the simulation degrades the ideal images constructed in the previous sections. This resembles the degradation caused by the real measurement system. Firstly, uneven illumination, I_s, is simulated by adding a second degree parabolic polynomial in the image. The center of the simulated illumination source can be inputted. The energy of the illumination source is controlled by a parameter E_s. The autofluorescence effect, I_a with energy E_a, is simulated as a spatially slowly changing random texture [12]. In addition, convolution with a 2D Gaussian, G, is used to simulate the point spread function. Finally, we add zero mean Gaussian noise, N_g with variance σ_g to approximate the CCD detector noise. Hence, the simulated image degraded by the acquisition system, $\hat{I}$, obtained from an ideal image I is given by:

$$\hat{I} = [(I + E_s I_s + E_a I_a) * G] + N_g, \quad (11)$$

where $*$ is the convolution operator.

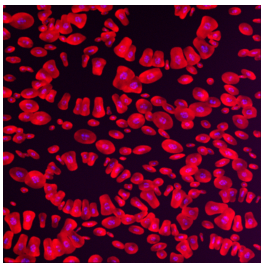
a)  b) 

Fig. 3. Example of simulated images. The images are 1000×1000 pixels, at magnification $40\times$, with $L_{max} = 0.2$ in the cytoplasm and $\nu_e = \nu_s = 1$. (a) shows the cytoplasm in red, and nuclei in blue. (b) displays only the nuclei.

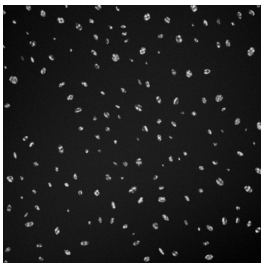
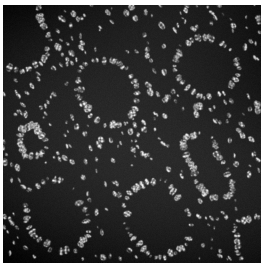
a)  b) 

Fig. 4. Example of nuclei channel of simulated images. The images are 1000×1000 pixels, at magnification $40\times$. Parameters are (a) $L_{max} = 0$ in the cytoplasm, $\nu_e = \nu_s = 0.7$, and (b) $L_{max} = 0.8$, $\nu_e = \nu_s = 1$.

Table 1. Mean and standard deviation (shown in brackets) of crypt morphological features for real and synthetic data. Measurements are in pixels for a 20x images

	Area	Major	Minor	Ratio	Solidity
Real	14171 (11244)	162 (78.7)	105 (28.9)	0.695 (0.158)	0.979 (0.017)
Synthetic	14088 (6756)	159 (49.5)	110 (23.6)	0.718 (0.13)	0.988 (0.024)

3. RESULTS AND DISCUSSION

Figure 3 shows an example of a synthesized image. By visual inspection, we can see that the tissue structure closely resembles that of real data, shown in Figure 1. The most distinguishing characteristic of the colon microenvironment is the crypt structure. Hence, in order to validate the model, we compared morphological features of the synthesized crypts with those calculated from the hand-marked histology images. The results are shown in Table 1. One clear difference is that in Figure 3 (b) nuclei regions are much more easily distinguishable. However, this is due to the fact that the example is synthesized with a small amount of overlap between cytoplasmic regions being allowed. Depending on the purpose for image synthesis, one may require to have fewer, easily separable cells (Figure 4 (a)), or more crowded and overlapping cells (Figure 4 (b)). Varying these parameters could be essential, for example, when testing cell segmentation and counting algorithms. The results from cell counting experiments using ImageJ [13] and CellProfiler [14] are shown in Table 2. Cell counting was done on 30 simulated samples with the same parameters as the example shown in Figure 3. Cell counting was done both on the non-overlapping nuclei regions and on the cytoplasmic regions where overlap of 0.2 was allowed. We can see that ImageJ tended to clump cells together if they overlap, whereas CellProfiler over-segmented them. However, further parameter optimization may give better results.

Table 2. Cell counting results for ImageJ and CellProfiler with counting based on non-overlapping nuclei or cytoplasm regions with $L_{max} = 0.3$. Mean and standard deviation (in brackets) are shown normalised by the ground truth.

ImageJ Nuclei	0.997 (0.004)
ImageJ Cytoplasm	0.703 (0.034)
CellProfiler Nuclei	1.218 (0.164)
CellProfiler Cytoplasm	1.113 (0.143)

4. CONCLUSIONS

Modern high-throughput imaging methods have raised the need for automated analytical frameworks. However, validation of such methods has been challenging since ground truth information in cell biological research is often missing, and verification using manual methods introduces variable results. Hence, simulation is a valuable tool when trying to develop, validate, and compare analytical methods.

We presented a model for simulating healthy colonic crypt architecture. The framework has several parameters, which allow control over the tissue appearance. Detailed analysis of hand-marked H&E images has enabled us to make the model realistic by learning parameters to generate realistic cell phenotypes, chromatin texture, nuclei morphology, and crypt architecture. To the best of our knowledge, ours is the first model to simulate immunofluorescence image data at subcellular level, where the cells have several compartments and are organized to mimic the microenvironment of tissue *in situ* rather than dispersed cells in a cultured environment. The synthesized data could be used to validate techniques such as image restoration, cell and crypt segmentation.

In the future, the simulation framework will be extended to also include colon cancer simulation. We are also looking into developing models for different protein expressions in the cells. In addition, when considering tissue structures, one should take into account regions outside the cells as this is where many of the challenges that face image processing algorithms come from. Hence, we need to consider simulation of the extra-cellular matrix and the lumen texture.

5. REFERENCES

[1] D. Webb, M.A. Hamilton, G.J. Harkin, S. Lawrence, A.K. Camper, and Z. Lewandowski, "Assessing technician effects when extracting quantities from microscope images," *Journal of microbiological methods*, vol. 53, no. 1, pp. 97–106, 2003.

[2] A.M. Grigoryan, G. Hostetter, O. Kallioniemi, and E.R. Dougherty, "Simulation toolbox for 3d-fish spot-counting algorithms," *Real-Time Imaging*, vol. 8, no. 3, pp. 203–212, 2002.

[3] E.M.M. Manders, R. Hoebe, J. Strackee, A.M. Vossepoel, and J.A. Aten, "Largest contour segmentation: a tool for the localization of spots in confocal images," *Cytometry*, vol. 23, no. 1, pp. 15–21, 1996.

[4] S.J. Lockett, D. Sudar, C.T. Thompson, D. Pinkel, and J.W. Gray, "Efficient, interactive, and three-dimensional segmentation of cell nuclei in thick tissue sections," *Cytometry*, vol. 31, no. 4, pp. 275–286, 1998.

[5] A. Lehmussola, P. Ruusuvuori, J. Selinummi, H. Huttunen, and O. Yli-Harja, "Computational framework for simulating fluorescence microscope images with cell populations," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 7, pp. 1010–1016, 2007.

[6] D. Svoboda, M. Kozubek, and S. Stejskal, "Generation of digital phantoms of cell nuclei and simulation of image formation in 3d image cytometry," *Cytometry part A*, vol. 75, no. 6, pp. 494–509, 2009.

[7] D. Svoboda, O. Homola, and S. Stejskal, "Generation of 3d digital phantoms of colon tissue," in *Image Analysis and Recognition*, pp. 31–39. Springer, 2011.

[8] T. Zhao and R.F. Murphy, "Automated learning of generative models for subcellular location: building blocks for systems biology," *Cytometry Part A*, vol. 71, no. 12, pp. 978–990, 2007.

[9] R.M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

[10] B.J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[11] J. Portilla and E.P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[12] K. Perlin, "An image synthesizer," *ACM Siggraph Computer Graphics*, vol. 19, no. 3, pp. 287–296, 1985.

[13] M.D Abràmoff, P.J. Magalhães, and S.J. Ram, "Image processing with imagej," *Biophotonics international*, vol. 11, no. 7, pp. 36–43, 2004.

[14] A.E. Carpenter, T.R. Jones, M.R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D.A. Guertin, J.H. Chang, R.A. Lindquist, J. Moffat, P. Golland, and D.M. Sabatini, "Cellprofiler: image analysis software for identifying and quantifying cell phenotypes," *Genome biology*, vol. 7, no. 10, pp. R100, 2006.