Detection of malignant mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens:

Supplementary Material

Akif Burak Tosun[1,*], Oleksandr Yergiyev[2], Soheil Kolouri[1], Jan F. Silverman[2], Gustavo K. Rohde[1,3,4]

[1]Dept. of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA;

[2]Dept. of Pathology and Laboratory Medicine, Allegheny General Hospital, Pittsburgh, PA 15212 USA;

[3]Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA;

[4]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Headline: **Mesothelioma detection from effusion fluid**

*Corresponding author contact;

E-mail: tosun@cmu.edu;

Phone: +1-412-268-8379;

Address: HH-C119, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.

ABSTRACT

Mesothelioma is a form of cancer generally caused from previous exposure to asbestos. Although it was considered a rare neoplasm in the past, its incidence is increasing worldwide due to extensive use of asbestos. In the current practice of medicine, the gold standard for diagnosing mesothelioma is through a pleural biopsy with subsequent histologic examination of the tissue. The diagnostic tissue should demonstrate the invasion by the tumor and is obtained through thoracoscopy or open thoracotomy, both being highly invasive surgical operations. On the other hand, thoracocenthesis, which is removal of effusion fluid from the pleural space, is a far less invasive procedure that can provide material for cytological examination. In this study, we aim at detecting and classifying malignant mesothelioma based on the nuclear chromatin distribution from digital images of mesothelial cells in effusion cytology specimens. Accordingly, a computerized method is developed to determine whether a set of nuclei belonging to a patient is benign or malignant. The quantification of chromatin distribution is done by using the optimal transport-based linear embedding for segmented nuclei in combination with the modified Fisher discriminant analysis. Classification is then performed through a k-nearest neighborhood approach and a basic voting strategy. Our experiments on 34 different human cases result in 100% accurate predictions computed with blind cross validation. Experimental comparisons also show that the new method can significantly outperform standard numerical feature-type methods in terms of agreement with the clinical diagnosis gold standard. According to our results, we conclude that nuclear structure of mesothelial cells alone may contain enough information to separate malignant mesothelioma from benign mesothelial proliferations.

Key terms: Mesothelioma, chromatin distribution, cancer detection, nuclear structure, cytology, optimal transport.

INTRODUCTION

In this supplementary material, we first provide detailed information about particle approximation step which is modelling the chromatin distribution of each nucleus. Subsequently, in order to give a better understanding we provide detailed information about our linear optimal transport (LOT) framework.

PARTICLE APPROXIMATION

Before applying our LOT approach to a given a set of images I_1; I_2; ..; I_K, we first compute the particle approximation for each image, which is a weighted combination of 'particles' with mass (intensity) m_i and location x_i:

$$\mu = \sum_{i=1}^{N_\mu} m_i \delta_{x_i} \quad (1)$$

where, N_μ is number of particles that are used to represent the image μ, m_i is the pixel intensity and δ_{x_i} is a Dirac delta function at pixel location x_i. Originally, an image can be represented in this form when N_μ is equal to number of pixels in that image. However, given optimal-transport is generally an $O(N_\mu^3)$ problem, using all the pixels in an image will be computationally expensive. Instead, we chose to approximate the images to reduce the cost of linear programming solution.

In our application, each segmented nuclear structure is represented in a 300 x 300 pixels gray level digital image. We used a point mass approximation to model the intensity distribution of each nucleus image. Note that, as described in subsection "Comparing nuclear chromatin using transport-based morphometry", the luminance component of segmented RGB nuclei images was extracted and intensities were normalized so that the pixel brightness (higher intensity) indicates the amount of locally concentrated chromatin (the brighter the pixel, the more chromatin). Specifically, we use Lloyd's weighted K-means algorithm (28) to adjust the position and weights of a set of $N \ll M$ particle masses to approximate the total intensity distribution of each nucleus, where M is the number of pixels in the image. In order to keep the balance between accuracy (a good approximation of the images in the mean squared error sense) and speed, the number of particles was chosen to be $N = 800$ in this particular study. The problem has now been reduced to finding the OT distance between two images I_1 and I_2 :

$$I_1 = \sum_{i=1}^{N_p} p_i \delta_{x_i} \quad , \; I_2 = \sum_{j=1}^{N_q} q_j \delta_{y_j}$$

with $N_p$and $N_q$being the number of particles chosen. An illustration of particles approximated for a nucleus is given in Figure 3 (Step 1).

LINEAR OPTIMAL TRANSPORT

Here, we describe the optimal transportation metric used for quantifying and classifying nuclear structure. We first do it in a general setting, and then apply it to discrete representations of the images considered.

Let π represent the domain (e.g. the unit square $[0,1]^2$) over which images are defined. Let us consider probability measures P_1 and P_2 on π. Recall that probability measures are nonnegative and that the measure of the whole set π is 1: $P_1(\pi) = P_2(\pi)$. In the context of images, the measure of a set is the sum of intensities over all pixels in the set. In addition, we will often refer to the measure of a set as its "*mass*" when discussing optimal transport.

Let $c: \pi \times \pi \to [0, \infty)$ be the cost function. That is $c(x, y)$ is the "cost" of transporting unit mass located at x to the location at y. The optimal transportation distance measures the least possible total cost of transporting all of the mass from P_1 to P_2. To make this precise, for any measurable set $A \subset \pi$ we have $\mu(A \times \pi) = P_1(A)$ and $\mu(\pi \times A) = P_2(A)$, where μ is a coupling within the set of all couplings between P_1 and P_2 . Note that, the set of all couplings ($\prod(P_1, P_2)$) is the set of all probability measures on with the first marginal P_1 and the second marginal P_2. Each coupling describes a transportation plan $\mu(A_0 \times A_1)$, which tells the amount of "*mass*" that is originally in set A_0 transported into set A_1.

We consider optimal transportation with quadratic cost $c(x, y) = |x - y|^2$. The optimal transportation distance, also known as the Kantorovich–Wasserstein distance, is then defined by

$$d(P_1, P_2) = \left(\inf_{\mu \in \prod(P_1, P_2)} \int_{\pi \times \pi} |x - y|^2 d\mu\right)^{1/2} \quad (2)$$

It is well known that the above infimum is attained and that the distance defined is indeed a metric. For the quadratic cost, the space of probability measures is endowed with a structure of a Riemannian manifold. This Riemannian manifold structure is needed to be able to consider paths and in particular the shortest path (i.e., geodesics) connecting any two probability measures, which, in our case, two images of nuclei in the space of images.

We start by explaining how the general framework applies to particle measures (such as particle approximations obtained above). A particle probability measure, μ, which approximates the image is given as $\mu = \sum_{i=1}^{N} m_i \delta_{x_i}$ where $x_i \in \pi$, $m_i \in (0,1]$, $\sum_{i=1}^{N} m_i$=1.

We note that for $A \subset \pi$, $\mu(A) = \sum_{i:x_i \in A} m_i$. An integral with respect to measure μ is $\int_A f(x) d\mu(x) = \sum_{i:x_i \in A} f(x_i) m_i$.

For obtaining the OT distance, let I_0 and I_1 be two images in our image set. Then, their particle approximated definitions are given as $I_0 = \sum_{i=1}^{N_{I_0}} m_i \delta_{x_i}$ and $I_1 = \sum_{j=1}^{N_{I_1}} p_j \delta_{y_j}$. The set of couplings between these two images $\prod(I_0, I_1)$ is then given by a set of $N_{I_0} \times N_{I_1}$ matrices:

$$\prod(I_0, I_1) = \{ \sum_{i=1}^{N_{I_0}} \sum_{j=1}^{N_{I_1}} f_{i,j} \delta_{x_i, y_j} \quad :$$

$$f_{i,j} \geq 0 \quad \text{for } i = 1, \dots, N_{I_0}, \quad j = 1, \dots, N_{I_1} ,$$

$$\sum_{j=1}^{N_{I_1}} f_{i,j} = m_i \quad \text{for } i = 1, \dots, N_{I_1}, \text{ and}$$

$$\sum_{i=1}^{N_{I_0}} f_{i,j} = p_j \quad \text{for } j = 1, \dots, N_{I_0}\}.$$

Since it is clear from the context we will make no distinction between measures in $\prod(I_0, I_1)$ and matrices $f = [f_{i,j}]$ that satisfy the conditions above. The optimal transportation distance between I_0 and I_1 defined in (2) is the solution to the following linear programing problem:

$$d^2(I_0, I_1) = \min_{f \in \prod(I_0, I_1)} \sum_{i=1}^{N_{I_0}} \sum_{j=1}^{N_{I_1}} \left| x_i - y_j \right|^2 f_{i,j} \tag{3}$$

For illustration of transportation plan, see Step 2 in Figure 3 in main manuscript. The minimization is performed utilizing the linear programming approach described in Wang et al. (25). In above formulation of optimal transport, particles from the template image, I_0, can split and redistribute over many particles in the target image, I_1. To avoid this particle splitting effect, as suggested by Wang et al. (25), we define the centroid of particle $q_k \delta_{x_k}$ in image I_1 to be,

$$a^k = \sum_{j=1}^{N} f_{k,j} y_j / q_k \tag{4}$$

Then the linear embedding of I_1 is obtained by applying the discrete transportation plan between the reference I_0 and I_1 to the coordinates y_j via

$$t_n = (\sqrt{q_1} a_n^1 \dots \sqrt{q_N} a_n^N)^T \tag{5}.$$

We denote t_1 to be the linear optimal transport (LOT) embedding of I_1. This embedding has dimensions $t_1 \in \mathbb{R}^{N \times 2}$ for two dimensional images. Hence, the embedding is interpretable in the sense that any point in this space can be visualized by simply plotting the vector coordinates (each in $\mathbb{R}^2$) in the image space π.