A GRAPH-BASED METHOD FOR DETECTING CHARACTERISTIC PHENOTYPES FROM BIOMEDICAL IMAGES

Wei Wang[1], *Cheng Chen*[1], *Tao Peng*[1], *Dejan Slepčev*[2], *John A. Ozolek*[3], *Gustavo K. Rohde*[1]

[1]Center for Bioimage Informatics, Department of Biomedical Engineering
[2]Department of Mathematical Sciences
Carnegie Mellon University, Pittsburgh, PA. 15213
[3] Department of Pathology, Children's Hospital of Pittsburgh, Pittsburgh, PA, 15201

ABSTRACT

We propose a novel method for detecting characteristic informative phenotype patterns from biomedical images. By building a metric space quantifying the difference between images, we learn the distributions of different classes, and then detect the characteristic regions using graph partition. We show that the detected regions are statistically significant. Our approach can also be used for designing differentiating features for specific data set. We apply our method to a digital pathology problem and successfully detect two characteristic phenotypes pertaining to normal liver and hepatoblastoma nuclei. In addition to digital pathology, our method can be applied to other biomedical problems for detecting characteristic phenotypes (e.g. location proteomics, genetic screens, cell mechanics, etc.).

Index Terms— Phenotype detection, Optimal transportation distance, Graph partition, Cancer classification

1. INTRODUCTION

Exploring the information contained in large sets of biomedical images is crucial for understanding the biological mechanisms and functions. Detecting characteristic phenotypes of different populations is an important problem because it can be used to explore the relationship between the genotype and the phenotype in image-based screening experiments. This information is valuable in developing therapies to treat diseases related to genotype changes such as premature diseases [1] and cancer [2]. Characteristic phenotypes can also be used to assist clinicians for confident diagnosis. As the human brain has significant limitations in analyzing large amounts of complex information, we hypothesize that in many cases information contained in large amount of image data can be more efficiently mined through computational approaches.

For decades many methods have been proposed to automatically identify and classify phenotypes from biomedical images [3, 2, 4, 5]. The overwhelming majority of computational approaches follow a standard feature-based procedure where an image can be represented by a set of numerical features. However, the feature-based procedure that reduces each image to a set of pre-defined features results in compression of information. In this context, information from the digital image, which may have diagnostic or biological significance, is discarded. Moreover, it is difficult to interpret the results found by feature-based methods which rely on a combination of features with different units.

We propose a method for detecting characteristic phenotypes in image data over mathematical geometric spaces, which can not only detect phenotypes but also provide a way to interpret them. These calculations are accompanied by statistical significance assessments based on p-values and classification accuracies. Although our method can be used more generally, in this paper we focus on detecting characteristic phenotypes based on nuclear structure obtained from histopathology images. We describe algorithms based on optimal transportation distance (OT) for quantifying nuclear structure in histopathology images. We note that the optimal transportation distance metric has been used in the past for different image analysis problems [6, 7]. We believe OT metric is especially suitable for analyzing (modeling) the nuclear structure because it measures quantities at the heart of the problem (chromatin mass concentration and their locations).

2. DATA ACQUISITION, IMAGE SEGMENTATION AND PRE-PROCESSING

Our data consist of images from five cases of fetal-type hepatoblastoma (HB) as well as normal liver tissues (NL) taken from the Childrens Hospital of Pittsburgh. Hepatoblastoma is the most common liver tumor in infancy and can be difficult to diagnose. Slides from the representative tissue blocks were sectioned at 5 microns and stained using a popular Feulgen technique which stains DNA (chromatin) magenta only.

A semi-automatic method has been developed to segment large quantities of nuclei from Feulgen stained histological images [5], in which nuclei are roughly segmented by graph cut [8] and refined by a level set method [9]. Finally, the pathologist (JAO) reviews all the segmented nuclei, and removes those incorrectly segmented (about 40%). Then nuclei samples were converted to grayscale by selecting the green channel from the RGB images and normalized so that the sum of their intensity values is 1 to guarantee the OT metric meaningful [6, 7]. In total, we extracted 461 hepatoblastoma (HB), 396 normal (NL) nuclei from this dataset.

Nuclei images are pre-processed as in our previous works [10, 11] to eliminate uninteresting variations due to arbitrary rotation, translation, and coordinate inversions of each nucleus. The procedure includes normalization by the center of mass, rotation by major axis reorientation, and coordinate "flips" set up within a least squares minimization problem.

3. METHOD

3.1. Optimal transportation distance

Here we describe the optimal transportation (OT) metric used for quantifying image differences and building image metric space. We

first do it in a continuum setting, and then apply it to discrete representations of the images considered. Let Ω represent the domain (e.g. the unit square $[0,1]^2$) over which images are defined. Let us consider probability measures (nonnegative measures of mass 1) I_0 and I_1 on Ω. Consider $\Pi(I_0, I_1)$, the set of all *couplings* between I_0 and I_1. That is consider the set of all probability measures on $\Omega \times \Omega$ with the first marginal I_0 and the second marginal I_1. More precisely, if $\mu \in \Pi(I_0, I_1)$ then for any measurable set $A \subset \Omega$ we have $\mu(A \times \Omega) = I_0(A)$ and $\mu(\Omega \times A) = I_1(A)$. Each pairing describes a *transportation plan*, that is $\mu(A_0 \times A_1)$ is telling one how much "mass" originally in the set A_0 is being transported into the set A_1.

Let $C : \Omega \times \Omega \to [0, \infty)$ be the *cost function*. That is $C(x, y)$ is the "cost" of transporting unit mass located at x to location y. We consider costs which are continuous and symmetric ($C(x, y) = C(y, x)$). The optimal transportation distance, also known as the Kantorovich-Wasserstein metric, is defined as

$$d(I_0, I_1) = \inf_{\mu \in \Pi(I_0, I_1)} \int_{\Omega \times \Omega} C(x, y) d\mu \tag{1}$$

It is well known that the above infimum is attained and that the distance defined is indeed a metric (satisfying the positivity, the symmetry, and the triangle inequality). The most frequently used cost is the quadratic cost, which we use as well: $C(x, y) = |x - y|^2$.

In our application, each nuclear structure is represented in a gray level digital image (of size 192 x 192 pixels). Each image I can be represented as $I = \sum_{i=1}^{M} v_i \delta_{x_i}$, where δ_{x_i} is a Dirac delta function at pixel location x_i, M is the number of pixels in image I, and v_i are the pixel intensity values. To accelerate the computation, we use a point mass approximation to the model the chromatin distribution of each nucleus. In specific, we use Lloyds weighted K-means algorithm [12] to adjust the position and weights of a set of $N < M$ particle masses to approximate the total intensity distribution of each nuclei. In all of the computations in this paper, $N \leq 600$. In the discrete setting, the problem reduces to finding

$$\min \sum_{i=1}^{N_p} \sum_{j=1}^{N_q} C(x_i, y_j) f_{i,j} \tag{2}$$

with N_p and N_q the number of masses chosen for representing images I_0 and I_1, subject to the constraints $\sum_{i=1}^{N_q} f_{i,j} = I_1(y_j)$ and $\sum_{j=1}^{N_p} f_{i,j} = I_0(x_i)$, $f_{i,j} \geq 0$. We utilize Matlab's implementation of a variation of Mehrotras dual interior point method [7] to solve the linear programming problem. The geodesic interpolation between I_0, and I_1 can be approximated by $\alpha \sum_{i=1}^{N_q} f_{i,j} = I_\alpha(y_j)$, $\alpha \in [0, 1]$.

3.2. Estimating data distributions over geometric spaces

Assume there are W classes in the dataset. After computing the pairwise distances between all image pairs under OT metric, we can imagine each images as a data point in OT metric space, then learn the distributions of the images for different classes. The nonparametric method kernel density estimation [13] is applied to approximate probability distribution of each class W in the OT metric space. To simplify the problem, we use standard spherical Gaussian kernel with mean 0. The *pdf* is formulated as

$$\hat{p}_h(\mathbf{x} \mid W) = \frac{1}{n_W h} \sum_{j=1}^{n_W} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{d^2(\mathbf{x}, \mathbf{x_j})}{2h^2}\right) \tag{3}$$

with $d(\mathbf{x}, \mathbf{x_j})$ representing their distance in metric space and n_W representing number of data points in class W. The optimal bandwidth h is selected according to [14]. Therefore, for each data point $\mathbf{x_i}$ there are W corresponding probability distribution values, we denote them as $p(\mathbf{x_i}|W)$ under optimal bandwidth h.

3.3. Detecting characteristic regions using graph partition

Given the probability distributions for different classes, a characteristic region for a certain class is defined as a subset of sample points, which satisfies two criteria: (1) in this subset, the total probability density of points belonging to this class is higher than the other subsets; (2) points in this subset should be close to each other. In our application, we have two classes, c_0 or c_1. We first divide all the sample points into two parts, in which the probability density of sample points under one class is larger than the other, such as $p(\mathbf{x_i}|c_0) \geq p(\mathbf{x_i}|c_1)$. Then, we can formulate the detection of characteristic region in one class as a graph partition problem by representing the sample points in such part as a weighted undirected graph $G = (V, E)$, where V and E are defined as a set of nodes and edges respectively. In this graph, each node corresponds to one data point, in this part, in OT metric space, and each pair of nodes is connected by an edge $E\,(i, j)$.

Based on the definition of a characteristic region, information of both probability densities and distance will be considered and the weight function $w\,(i, j)$ is defined as:

$$w\,(i, j) = exp\left(\frac{q_i q_j}{\sigma_q}\right) \times \begin{cases} exp\left(-\frac{D_{ij}}{\sigma_D}\right) & \text{if } \sigma_D \leq r \\ 0 & \text{if } otherwise \end{cases} \tag{4}$$

where $q_i = |p(\mathbf{x_i}|c_1) - p(\mathbf{x_i}|c_0)|$. For node i supposing there are two classes: c_0, c_1; σ_q, σ_D are scaling parameters, D_{ij} is the distance between node i and j, and r is a distance threshold. The intuition of such weight function is to reflect the likelihood that two nodes belong to the same characteristic region.

The whole graph of one class can be partitioned into n disjoint sets: $V_1 \cap V_2 ... \cap V_n = 0$, $V_1 \cup V_2 ... \cup V_n = V$, each V_i can be regarded as a characteristic region when maximizing the similarity among the nodes within sets and minimizing the similarity across different sets. The normalized cut method [15] can be employed, in which maximizing similarity within the sets and minimizing similarity between different sets are proved to be identical and can be satisfied simultaneously.

For each class, we apply graph partition and choose the disjoint set which maximizes the $\frac{1}{M} \sum_{i \in V_i} \sum_{j \in V_i} w(i, j)$ as its characteristic region, where M corresponds to the number of edges in that region.

3.4. Characteristic region validation

Characteristic phenotypes or patterns pertaining to each class can be defined as regions in the metric space that are well populated by one class, but not the others. In this section, we propose a method to validate that the characteristic region detected in section 3.3 is statistically significant in the sense that these regions are significant in differentiating the data from different classes. In other words, excluding these regions would reduce the classification accuracy significantly compared with randomly exclusion of the same amount of data from the original data set. We do this validation in the following manner. Given a data set, firstly, we can build a classifier (in our case we use K-Nearest-Neighbor classifier with voting) over all samples, and establish a classification accuracy a_o corresponding

to this classifier. Secondly, we exclude data points ($R\%$ of all the data) from characteristic regions from both classes, and compute the classification accuracy a_r. Thirdly, we randomly exclude the same amount of $R\%$ points from both classes, and run this step for 20,000 times, then build a experience distribution of the accuracy $H(a)$. Finally, we quantitatively show the effect of accuracy decreasing by computing the p-value of the classification accuracy a_r.

4. RESULTS

4.1. Identifying characteristic region

Based on the pre-processed data set introduced in section 2, we randomly select 500 nuclei (250 from HB, 250 from NL). Followed the procedure introduced in section 3.1, 3.2 and 3.3, we successfully identify the characteristic region for this data set. The two-dimensional representation (obtained using the multi-dimensional scaling technique [16], higher dimensions not shown) is shown in Figure 1(A). In this figure each point represents a nuclear structure. The solid shapes (indicated by large circles) indicate the regions in this space that have been identified as characteristic region. Although, it seems that characteristic regions still contain data from opposite class in this two-dimensional view, characteristic regions are distinct in OT metric space. The nuclear structures pertaining to each identified pattern are shown in Figure 1(B, C).

By observing the identified patterns, we can find two obvious distinctions, the area of the nuclei and the concentration of chromatin (one is more concentrated in the peripheral). The first distinction is further confirmed by the histogram of area (computed by counting the number of pixels inside each nuclei) shown in Figure 2(A). We confirm the second distinction as follows: based on the detected characteristic regions, we can find a characteristic direction by connecting two farthest nuclei in different regions under OT metric. Figure 2(B) (bottom) shows this characteristic direction as well as the histogram of the projection over this direction [17] (In this bottom figure, only the left most and right most images are real nuclear images. The intermediate ones are interpolated images according to the geodesic interpolation as described in section 3.1.). The histograms over the corresponding images indicate the relative number of nuclei in each population of data (normal vs HB) that was closest (in the OT sense) to it. The distributions of nuclei of different classes over this direction are very different, which confirm that there is a big distinction in the concentrations of chromatin in different classes. This result of finding the characteristic patterns also provide a guideline in designing more discriminant features.

4.2. Validating characteristic region

As described in section 3.4, we build a K-Nearest-Neighbor classifier with voting [5] to classify a group of M images, and the label of that group is determined by taking the majority vote in that group. We do 10-fold cross-validation to report the results, and also do cross-validation in the training set to find the minimum group size $M = 35$ when the data can be classified with accuracy $a_o = 100\%$. We exclude 25 points (10% of total) in detected characteristic regions from both classes, and compute the classification accuracy $a_r = 94.67\%$. Then, we randomly exclude the same amount of 10% points from both classes, and run this step for 20,000 times, then build a histogram of the classification accuracy $H(a)$ as shown in Figure3. The p-value corresponding to the classification accuracy a_r in $H(a)$ is 0.0036. The decrease in classification accuracy from

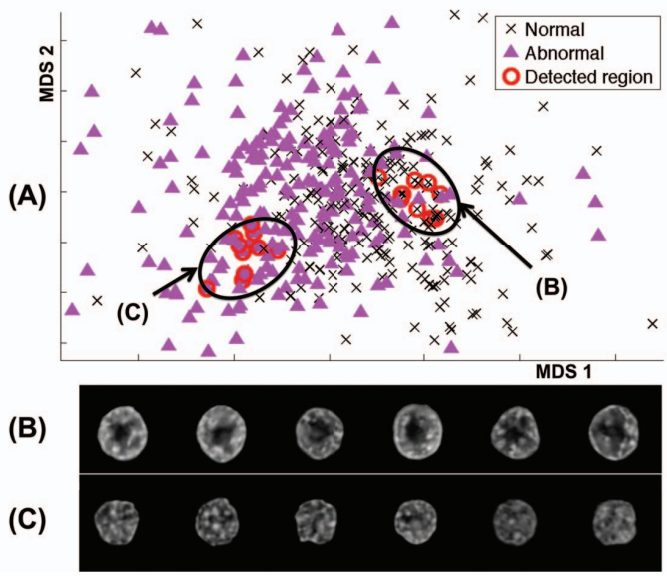

Fig. 1. Detected characteristic region. (A): Two-dimensional representation of nuclear population (Normal and Abnormal), with axis corresponding to directions computed by multi-dimensional scaling technique. The solid shapes indicates the 2 characteristic regions corresponding to NL and HB. (B,C): Nuclei samples of characteristic regions for NL and HB.

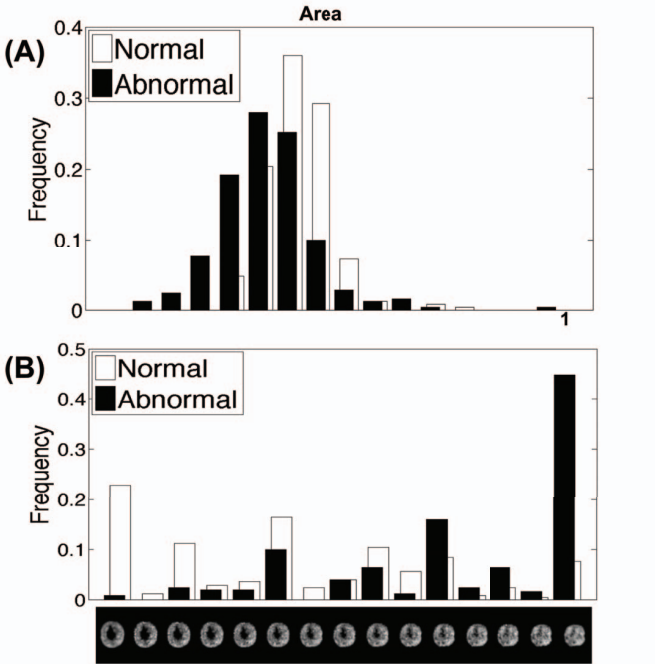

Fig. 2. Histogram of individual features. (A) shows the histogram of the area feature with the largest normalized to 1. (B) shows the histogram of the projection over the characteristic direction, detected by our method. This direction (bottom figure) shows variation in where chromatin is positioned within the nucleus.

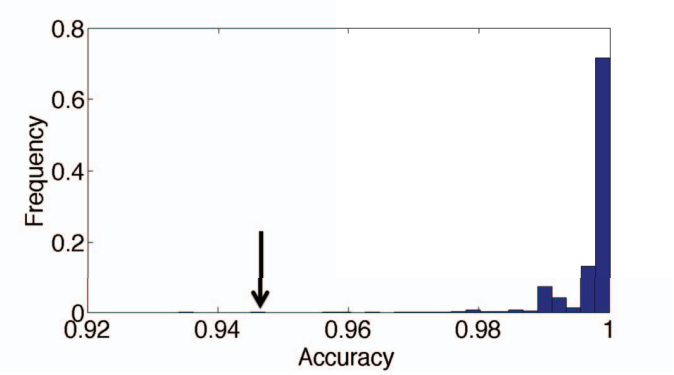

Fig. 3. Histogram of the accuracy by randomly excluding the same amount of points from both classes and running for 20,000 times.

100% as well as p-value computation show the characteristic region we detect is significant for differentiating these two classes.

4.3. Classification based on detected pattern

Another method for estimating the significance of the regions detected is to examine if data points from these regions alone can be used to correctly classify the data. Given an unknown group of nuclei, we can compute its histogram of projection over the direction shown in Figure 2(B) (bottom), and then compare the result histogram with histogram shown in Figure 2(B) and assign this unknown group the same label with the more similar one in distribution. Followed this idea, we can successfully classify the unknown group of nuclei with 100% accuracy whenever 50 or more nuclei are in that group.

5. CONCLUSION AND DISCUSSION

We proposed a novel method for detecting characteristic phenotypes patterns from biomedical images. We also describe two methods that can be used to determine the importance of the phenotype patterns relative to the task of image classification, together with a p value computation that provides an estimate of the statistical significance of the results. These methods were applied to the task of determining what characteristic nuclear structures allow one to distinguish between NL and HB. Two distinct types of information were established: size (area), and chromatin distribution (uniform vs. concentrated along the border of the nucleus). While only this application was investigated in this paper, we believe the methods presented could also be applied to other biomedical problems such as genetic screens, location proteomic studies, and others.

6. ACKNOWLEDGEMENT

We wish to thank Drs. Ann B. Lee and Robert. F. Murphy, from Carnegie Mellon University, for useful discussions. This work was partially supported by NIH grant GM088816 and GM075205.

7. REFERENCES

[1] K. N. Dahl, P. Scaffidi, M.F. Islam, A.G. Yodh, K.L. Wilson, , and Tom Misteli, "Distinct structural and mechanical properties of the nuclear lamina in hutchinson–gilford progeria syndrome," *Proc Natl Acad Sci U S A*, vol. 103, no. 27, pp. 10271–10276, 2006.

[2] A Ruiz, M Ujaldon, J.A Andrades, J. Becerra, Kun Huang, T. Pan, and J. Saltz, "The gpu on biomedical image processing for color and phenotype analysis," *Proceedings of IEEE BIBE*, pp. 1124–1128, 2007.

[3] 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 Biol.*, vol. 7, no. 10, 2006.

[4] N. Hamilton, R. Pantelic, K. Hanson, J.L. Fink, S. Karunaratne, and R.D. Teasdale, "Automated subcellular phenotype classification: An introduction and recent results," *Proceedings of the workshop on Intelligent systems for bioinformatics*, pp. 67 – 72, 2006.

[5] W. Wang, J.A. Ozolek, and G.K. Rohde, "Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images," *Cytometry A*, 2010, In Press.

[6] S. Haker, L. Zhu, A. Tennembaum, and S. Angenent, "Optimal mass transport for registration and warping," *Intern. J. Comp. Vis.*, vol. 60, no. 3, pp. 225–240, 2004.

[7] Y. Rubner, C. Tomassi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Intern. J. Comp. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[8] Y. Boyokov and G. Funka-Lea, "Graph cuts and efficient n-d image segmentation," *Intern. J. Comp. Vis.*, vol. 70, no. 2, pp. 109–131, 2006.

[9] C. Li, R. Huang, Z. Ding, C. Gatenby, D. Metaxas, and J. Gore, "A variational level set approach to segmentation and bias correction of images with intensity inhomogeneity," *Int Conf Med Image Comput Comput Assist Interv*, vol. 11, no. Pt 2, pp. 1083–91, 2008.

[10] G.K. Rohde, A.J.S. Ribeiro, K.N. Dahl, and R.F. Murphy, "Deformation-based nuclear morphometry: capturing nuclear shape variation in hela cells," *Cytometry A*, vol. 73, no. 4, pp. 341–50, Apr 2008.

[11] G.K. Rohde, W. Wang, T. Peng, and R.F. Murphy, "Deformation-based nonlinear dimension reduction: applications to nuclear morphometry," *Proc. IEEE Int. Symp. Biomed. Imaging*, 2008.

[12] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[13] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, 2005.

[14] T. Peng, W. Wang, G.K. Rohde, and R.F. Murphy, "Instance-based generative biological shape modeling," *Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging*, pp. 690–693, 2009.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transaction of Pattern Analysis and Machine Learning*, vol. 22, no. 8, August 2000.

[16] T. Cox and M. Cox, *Multidimensional Scaling*, Chapman and Hall, 1994.

[17] W. Wang, J.A. Ozolek, D. Slepcev, Ann Lee, C. Chen, and G.K. Rohde, "An optimal transportation approach for nuclear structure-based pathology," preprint.