Published in final edited form as:

Inf Process Med Imaging. 2011 ; 22: 398–410.

Identifying Nuclear Phenotypes Using Semi-supervised Metric Learning

Shantanu Singh¹, Firdaus Janoos¹, Thierry Pécot¹, Enrico Caserta², Gustavo Leone², Jens Rittscher³, and Raghu Machiraju¹

¹Dept. of Computer Science and Engg., The Ohio State University, U.S.A.

²Dept. of Molecular Genetics, The Ohio State University, U.S.A.

³General Electric Global Research Center, U.S.A.

Abstract

In systems-based approaches for studying processes such as cancer and development, identifying and characterizing individual cells within a tissue is the first step towards understanding the large-scale effects that emerge from the interactions between cells. To this end, nuclear morphology is an important phenotype to characterize the physiological and differentiated state of a cell. This study focuses on using nuclear morphology to identify cellular phenotypes in thick tissue sections imaged using 3D fluorescence microscopy. The limited label information, heterogeneous feature set describing a nucleus, and existence of sub-populations within cell-types makes this a difficult learning problem. To address these issues, a technique is presented to learn a distance metric from labeled data which is *locally adaptive* to account for heterogeneity in the data. Additionally, a label propagation technique is used to improve the quality of the learned metric by expanding the training set using unlabeled data. Results are presented on images of tumor stroma in breast cancer, where the framework is used to identify fibroblasts, macrophages and endothelial cells – three major stromal cells involved in carcinogenesis.

Keywords

Nuclear Morphometry; Metric Learning; Microscopy

1 Introduction

Recent years have seen an increasing role of quantitative imaging in systems biology research [12]. With the rapid advances made in fluorescence microscopy [10], it is now possible to image tissue sections within the 3D context of the biological microenvironment. The vast array of fluorescent proteins with distinct spectral properties and their ability to label different cell-types using transgenic animal models have opened up a whole new arena [14], where complex phenomena such as the interaction between cells in their tissue

microenvironment can be studied within a high-content, 3D spatial context. A crucial first step in these studies is identifying and characterizing the cells within the tissue. This paper focuses, in specific, on the characterization of *cell nuclei*. The morphology of the cell nucleus is frequently used as a proxy for characterizing the physiological and differentiated state of a cell. In cancer, modifications in nuclear structure have been shown to correlate with different types and stages of cancer [23]. The present work addresses the analysis of cell nuclei in the *stromal regions*¹ of the tumor microenvironment in breast cancer. This is motivated by findings [18,1] that have revealed the surprising role of stromal cells in the initiation and progression of cancer. This paper is aimed at the automatic identification of fibroblasts, macrophages and endothelial cells based on nuclear morphology and spatial positioning in 3D fluorescence microscopy images of thick tissue sections, and the characterization of the diversity of these cell-types by identifying cellular sub-populations that may exist. In these images, the ground truth information about cell-type is available through the expression of fluorescent proteins in transgenic mice. There are several challenges posed by this problem. *First*, quantitative representation of the cell nucleus requires a heterogeneous set of features corresponding to shape, texture and spatial properties. There exists no methodical way of comparing two nuclei given these features. *Second*, the number of labeled examples that can be collected in these experiments is restricted due to various factors (a) only one cell-type specific fluorescent protein is expressed per tissue section; (b) the signal quality of the transgenic fluorescence is often degraded due to background fluorescence, thus requiring an expert to validate the labeled data for each experiment; (c) stromal cells are sparsely distributed in the tissue resulting in few cells being visible in a given field of view.

To address these problems, first, an approach is proposed to learn an appropriate distance metric to compare nuclei. *Metric learning* techniques use prior information, typically the label information, to learn an optimal metric from the data for a specific task. Most techniques learn a global metric through a constrained optimization problem. In this paper, a method is proposed for learning *local metrics* in order to deal with the heterogeneity that exists within cell-types in the microenvironment. The method constructs a locally adaptive metric by extending an existing global technique by using a hierarchical Bayesian model. Next, to address the problem of small size of the training set, unlabeled data is used to enhance the quality of this metric by using label propagation on a graph constructed using the learned metric. The rest of the paper is organized as follows. Related work in cell analysis literature and metric learning is reviewed in Section 2. The metric learning framework is presented in Section 3. Results on 3D fluorescence microscopy images of nuclei in the stromal regions of mammary tissue are presented in Section 4. We conclude with a discussion in Section 5.

2 Related Work

Nuclear Analysis

A method to compare nuclear morphologies was presented in [13] that uses a combination of large deformation metric mapping and multi-dimensional scaling to derive a metric in an unsupervised manner. A method to perform spherical mapping of the nuclear volume is

presented in [7] based on a mechanical model of deformation to produce a normalized representation of the nucleus, thereby enabling the comparison of sub-cellular structures between two cells. An expansive set of features to characterize subcellular structures in a cell is proposed in [2]; this set is frequently used in image-based cellular analysis to quantify cellular phenotypes in fluorescence microscopy images. In cell analysis literature to date, there has been no work in the direction of *learning* an optimal metric from data to compare nuclei. To the best of our knowledge, this is the first formal analysis of this problem.

Metric learning

Among the earliest pieces of work on this topic was presented by Xing et. al. [21], where a metric is learned using prior knowledge about the similarity or dissimilarity of a set of points. Neighborhood component analysis [8] was used to learn a metric optimized for nearest-neighbor based classification. An approach to estimate a Mahalanobis metric using an efficient optimization methodology has been proposed in [5]. These approaches infer a *global* metric from the data. Among local techniques, an approach to learn a point-wise metric was proposed in [20], where the method requires solving expensive optimization problems for each data point. In this paper, a local metric learning approach is proposed that extends the method presented in [5] by casting the problem in a hierarchical Bayesian framework.

3 Methods

In this section, a *hierarchical Bayesian approach* is proposed for learning a locally adaptive distance metric. Next, a graph is constructed based on this metric to improve the quality of the learned metric by expanding the size of the training set through *label propagation* [22]. Finally, the nuclear features used to represent shape, appearance and spatial neighborhood are described.

The Global Metric Model

Consider the set of nuclei in an image $\mathcal{X} \triangleq (\mathcal{X}_l, \mathcal{X}_u) \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l, \mathbf{x}_{l+1}, \dots, \mathbf{x}_n\}$ where $\mathcal{X}_l$ and $\mathcal{X}_u$ corresponds to the labeled and unlabeled set respectively, and $\mathbf{x}_i$ is the representation of the nucleus in feature-space of dimension M . For each $\mathbf{x}_i \in \mathcal{X}_l$ there exists a label $y_i \in \{1, \dots, c\}$, where c is the number of cell-types have fluorescent markers identifying them. The sets $\mathcal{S} = \{(i, j) | y_i = y_j\}$ and $\mathcal{D} = \{(i, j) | y_i \neq y_j\}$ correspond to pairs of points that have the same or different label respectively.

The distance between two nuclei i, j is given by the metric $d_{\mathbf{A}_0}(\Delta \mathbf{x}_{i,j}) \triangleq \Delta \mathbf{x}_{i,j}^T \mathbf{A}_0 \Delta \mathbf{x}_{i,j}$ with $\mathbf{A}_0$ the metric tensor and $\mathbf{x}_{i,j} = \mathbf{x}_i - \mathbf{x}_j$. This global metric tensor is assumed to be distributed according to a Wishart distribution [6] $\mathcal{W}(\mathbf{A})$ on the space of positive semi-definite (PSD) matrices, defined as:

$$p(\mathbf{A}_0 | \mathbf{A}) \propto |\mathbf{A}^{-1} \mathbf{A}_0|^{M/2} \exp \left\{ -\frac{1}{2} \text{Tr}(\mathbf{A}^{-1} \mathbf{A}_0) \right\}.$$

The PSD matrix A expresses the prior belief about the metric tensor A_0 and is typically initialized to the $M \times M$ identity matrix implying an equal and independent contribution of all the features in measuring the distance. In order that the metric on this data keep the pairs in $\mathcal{S}$ close and the points in $\mathcal{D}$ apart, the data likelihood is given by:

$$p(\Delta \mathbf{x}_{i,j} | y_i = y_j, A_0) \propto \exp \left\{ -\frac{\lambda}{2} |d_{A_0}(\Delta \mathbf{x}_{i,j}) - u|_+ \right\}$$

$$\text{and} \quad p(\Delta \mathbf{x}_{i,j} | y_i \neq y_j, A_0) \propto \exp \left\{ -\frac{\eta}{2} |l - d_{A_0}(\Delta \mathbf{x}_{i,j})|_+ \right\},$$

where $u \geq 0, l \gg u, |c|_+ = \min(0, c)$ and $\lambda \geq 0, \eta \geq 0$ are non-zero constants. This form likelihood function requires that similar points be less than a distance u of each other and dissimilar point be a large distance l apart. Maximizing the posterior distribution $p(A_0 | \mathbf{x}_i, \mathbf{x}_j, y_i, y_j)$ with respect to A_0 results in the following constrained optimization problem [5]:

$$\min_{A_0} \text{LogDiv}(A, A_0) + \lambda \sum_{i,j \in \mathcal{S}} \zeta_{i,j} + \eta \sum_{i,j \in \mathcal{D}} \vartheta_{i,j}$$

$$\text{subject to} \quad d_{A_0}(\Delta \mathbf{x}_{i,j}) \leq u + \zeta_{i,j} \quad (i, j) \in \mathcal{S},$$

$$d_{A_0}(\Delta \mathbf{x}_{i,j}) \geq l - \vartheta_{i,j} \quad (i, j) \in \mathcal{D},$$

$$A_0 \succeq 0, \quad \zeta_{i,j} \geq 0, \quad \text{and} \quad \vartheta_{i,j} \geq 0.$$

where $\text{LogDiv}(A, A_0) \triangleq \text{Tr}(A^{-1}A_0) - \log |A^{-1}A_0| - d$ is the log-determinant divergence. This optimization problem can be efficiently computed using the technique of Bregman projections [9], involving rank-one updates of A_0 by projecting onto each of the constraints iteratively. The Bregman updates preserve the PSD property of A_0 in each iteration, thereby guaranteeing that the solution is a metric without explicitly handling the PSD condition.

The Local Metric Model

As shown in Section 4.1 a single global metric tensor fails to capture the heterogeneity of the differences between nuclei, suggesting the concept of an ensemble $\mathcal{A} = \{A_k, k=1 \dots K\}$ of K local metrics that vary according to nucleus type. Each local metric is characterized by the metric tensor A_k , which is again assumed to have a Wishart prior $\mathcal{W}(A_0)$. Also, for each pair of cell nuclei i, j we associate a hidden variable $z_{i,j} \in \{1, \dots, K\}$ indicating the metric active on that pair. Specifically, if $z_{i,j} = k$, then $d_{A_k}(\Delta \mathbf{x}_{i,j}) \triangleq (\Delta \mathbf{x}_{i,j})_{A_k}^\top (\Delta \mathbf{x}_{i,j})$. The marginal probability of a particular nucleus $\mathbf{x}_i$ belonging to the local support of a metric A_k

is modeled by a Gaussian distribution $p(\mathbf{x}_i|z_{i,j}=k, \mu_k, \Sigma_k) = \mathcal{N}(\mu_k, \Sigma_k)$ where $\mu \triangleq \mu_1 \dots \mu_K$, $\Sigma \triangleq \Sigma_1 \dots \Sigma_K$. Let $p(z_{i,j}=k) \triangleq \pi_k$ represent the marginal probability of a pair of nuclei belonging to the support of A_k , such that $0 \leq \pi_k \leq 1, \forall k = 1 \dots K$ and $\sum_{k=1}^K \pi_k = 1$. Therefore, $\theta \triangleq \{\mu, \Sigma, \pi\}$ are the parameters of the Gaussian mixture model (GMM) of the support of the local metric ensemble $\mathcal{A}$. As before, each local metric tensor A_k is adapted such that for all pairs of cell nuclei $(\mathbf{x}_i, \mathbf{x}_j)$ with labels y_i, y_j which belong to its local support, it tries to keep similar nuclei less than distance $u > 0$ of each other and dissimilar ones at a large distance $l \gg u$ apart as per:

$$p(\Delta \mathbf{x}_{i,j} | \mathbf{x}_i, \mathbf{x}_j, y_i=y_j, z_{i,j}=k, A_k) \propto \exp \left\{ -\frac{\lambda}{2} |d_{A_k}(\Delta \mathbf{x}_{i,j}) - u|_+ \right\}$$

$$\text{and} \quad p(\Delta \mathbf{x}_{i,j} | \mathbf{x}_i, \mathbf{x}_j, y_i \neq y_j, z_{i,j}=k, A_k) \propto \exp \left\{ -\frac{\eta}{2} |l - d_{A_k}(\Delta \mathbf{x}_{i,j})|_+ \right\}.$$

Therefore for any pair of data-points, the probability is given by :

$$\begin{aligned} p(\Delta \mathbf{x}_{i,j}, \mathbf{x}_i, \mathbf{x}_j, z_{i,j}, \mathcal{A} | y_i, y_j, \theta, A_0) \\ = \left[\prod_{k=1}^K p(A_k | A_0) \right] p(z_{i,j} | \pi) p(\mathbf{x}_i | z_{i,j}, \mu, \Sigma) p(\mathbf{x}_j | z_{i,j}, \mu, \Sigma) p(\Delta \mathbf{x}_{i,j} | \mathbf{x}_i, \mathbf{x}_j, y_i, y_j, z_{i,j}, \mathcal{A}). \end{aligned}$$

Assuming pair-wise independence of the cell nuclei, the complete data log probability is

$$\ln p(\Delta \mathcal{X}, \mathcal{X}, \mathcal{Z}, \mathcal{A} | \mathcal{Y}, \theta, A_0) = \sum_{k=1}^K [\ln p(A_k | A_0) + \ln \pi_k] + \sum_{i=1}^N \left[\ln p(\mathbf{x}_i | z_i, \mu, \Sigma) + \sum_{j>i}^N p(\Delta \mathbf{x}_{i,j} | \mathbf{x}_i, \mathbf{x}_j, y_i, y_j, z_i, z_j, \mathcal{A}) \right],$$

where $\Delta \mathcal{X} \triangleq \{\Delta \mathbf{x}_{i,j}, \forall i=1 \dots N-1, i < j \leq N\}$, $\mathcal{Y} \triangleq \{y_i, \forall i=1 \dots N\}$ and, $\mathcal{Z} \triangleq \{z_{i,j}, \forall i=1 \dots N-1, i < j \leq N\}$.

Local Metric Estimation

The optimal estimate of the GMM parameters θ^* and metric tensor ensemble $\mathcal{A}^*$ is obtained using expectation maximization (EM) which involves iterating the following two steps until convergence:

$$\textbf{E-step: } \mathcal{Q}((\mathcal{A}, \theta), (\mathcal{A}, \theta)^n) = \mathbb{E}_{p(\mathcal{Z} | \Delta \mathcal{X}, \mathcal{X}, \mathcal{Y}, \mathcal{A}^n, \theta^n, A_0)} [\ln p(\Delta \mathcal{X}, \mathcal{X}, \mathcal{Z}, \mathcal{A} | \mathcal{Y}, \theta, A_0)]$$

$$\textbf{M-step: } \underset{\mathcal{A}, \theta}{\operatorname{argmax}} \mathcal{Q}((\mathcal{A}, \theta), (\mathcal{A}, \theta)^n),$$

where

$$p(\mathcal{Z} | \Delta \mathcal{X}, \mathcal{X}, \mathcal{Y}, \mathcal{A}^n, \theta^n, A_0) = \prod_{i=1}^{N-1} \prod_{j>i}^N p(z_{i,j} | \Delta \mathbf{x}_{i,j}, \mathbf{x}_i, \mathbf{x}_j, y_i, y_j, \mathcal{A}^n, \theta^n, A_0). \text{ We}$$

define the responsibility term $\gamma_{i,j}(k) \triangleq p(z_{i,j}=k | \Delta \mathbf{x}_{i,j}, \mathbf{x}_i, \mathbf{x}_j, y_i, y_j, \mathcal{A}^n, \theta^n, A_0)$, which is evaluated as:

$$\gamma_{i,j}(k) = \frac{p(\Delta \mathbf{x}_{i,j}, \mathbf{x}_i, \mathbf{x}_j, y_i, y_j, z_{i,j}=k, \mathcal{A}^n | \theta^n, A_0)}{\sum_{z_{i,j}=1}^K p(\Delta \mathbf{x}_{i,j}, \mathbf{x}_i, \mathbf{x}_j, y_i, y_j, z_{i,j}, \mathcal{A}^n | \theta^n, A_0)}.$$

Denoting the term $N_k \triangleq \sum_{i=1}^{N-1} \sum_{j>i}^N \gamma_{i,j}(k)$, the M-Step for the GMM parameters results in the following closed-form updates:

$$\pi_k^{(n+1)} = \frac{2N_k}{N(N-1)}, \quad \mu_k^{(n+1)} = \frac{1}{N_k} \sum_{i=1}^{N-1} \sum_{j>i}^N \gamma_{i,j}(k) \mathbf{x}_i,$$

$$\text{and } \Sigma_k^{(n+1)} = \frac{1}{N_k} \sum_{i=1}^{N-1} \sum_{j>i}^N \gamma_{i,j}(k) (\mathbf{x}_i - \mu_k^{(n+1)}) (\mathbf{x}_i - \mu_k^{(n+1)})^\top.$$

The M-Step for the metric tensor A_k is

$$A_k^{(n+1)} = \underset{A_k}{\operatorname{argmax}} \sum_{i=1}^{N-1} \sum_{j>i}^N \gamma_{i,j}(k) \ln p(\Delta \mathbf{x}_{i,j} | \mathbf{x}_i, \mathbf{x}_j, y_i, y_j, z_{i,j}=k, A_k) + \ln p(A_k | A_0),$$

which is again equivalent to:

$$\min_{A_k} \operatorname{LogDiv}(A_0, A_k) + \lambda \sum_{i,j \in \mathcal{S}} \gamma_{i,j}(k) \zeta_{i,j} + \eta \sum_{i,j \in \mathcal{D}} \gamma_{i,j}(k) \vartheta_{i,j}$$

$$\text{subject to } d_{A_0}(\Delta \mathbf{x}_{i,j}) \leq u + \zeta_{i,j} \quad (i, j) \in \mathcal{S},$$

$$d_{A_0}(\Delta \mathbf{x}_{i,j}) \geq l - \vartheta_{i,j} \quad (i, j) \in \mathcal{D},$$

$$A_0 \succeq 0, \quad \zeta_{i,j} \geq 0 \quad \text{and} \quad \vartheta_{i,j} \geq 0.$$

Inducing a Global Metric from the Local Metric Ensemble

Finally, under this local metric model, the global distance between any two cell nuclei i, j in feature space is then defined as :

$$d_{\mathcal{A}}(\Delta \mathbf{x}_{i,j}) \triangleq \sum_{k=1}^K \gamma_{i,j}(k) \Delta \mathbf{x}_{i,j}^T \mathbf{A}_k \Delta \mathbf{x}_{i,j}.$$

In general, the construction of this metric does not ensure that the triangular inequality holds. In order to construct a globally consistent metric on the data, the distance between two points is defined as the geodesic distance on the graph where the edge weights are given by $d_{\mathcal{A}}(\Delta \mathbf{x}_{i,j})$. The geodesic distance is obtained using Floyd's algorithm [4] to compute the shortest paths between all pairs of vertices on this graph.

Expanding the training set using label propagation

The training data used in learning metric described above was restricted to the labeled set $\mathcal{X}_l$. Since the quantity of labeled data available per experiment is limited, an approach is formulated to increase the training set by incrementally adding new points from the unlabeled data for which the labels can be reliably estimated. The label propagation technique [22] provides a method to learn a *label function* using both labeled and unlabeled points. This method is used to add new points to the training set, re-estimate the local metric based on this expanded set, and repeat this process until no new points can be added. The details of this process are described as follows. To begin, a weighted graph $\mathcal{G}$ is constructed

on the points in $\mathcal{X}$ represented by the $n \times n$ weight matrix $W_{ij} \triangleq \exp(-d_{\mathcal{A}}(\Delta \mathbf{x}_{i,j})/\sigma^2)$, where kernel bandwidth σ is a tunable parameter. A *label function* $f: \mathcal{X} \rightarrow \{0, 1\}^c$ maps data points to their labels (using a *1-of-c* representation for labels). The *prediction matrix* $F \in \mathbb{R}^{n \times c}$ represents the values of the label function on $\mathcal{X}$. Define a split $F = (F_u F_l)$ where the two matrices correspond to the unlabeled and labeled points respectively. The label propagation estimates F_u given W and F_l by minimizing the quadratic energy function

$E(F) = \sum_{i,j} W_{ij} \|F(\mathbf{x}_i) - F(\mathbf{x}_j)\|_2^2$ keeping F_l fixed. The minimization is obtained in a

closed solution $F_u = (D_{uu} - W_{uu})^{-1} W_{ul} F_l$ where D is the diagonal matrix $D_{ii} = \sum_{j=1}^n W_{ij}$. Here, the matrix W is split into four sub-matrices defined as $W = (W_{uu} W_{ul}; W_{lu} W_{ll})$ where W_{uu} corresponds to pairs of points from the product set $\mathcal{X}_u \times \mathcal{X}_u$. W_{ul} , W_{lu} and W_{ll} are constructed similarly. The matrix D is split similar to W . The points for which the labels are predicted with high confidence in F_u (as measured through label entropy) are added to the training set. A new metric is now learned using this updated set and the process is repeated until no new points can be added.

3.1 Nuclear Features

In the present experiments, three sets of nuclear features were used corresponding to shape, appearance and spatial neighborhood as described below. Note that the proposed metric learning approach is independent of the choice of these features – they can be chosen to suit the specific phenotyping questions pertinent to the study.

Spherical harmonic based shape features—The shape of a nucleus is modeled using a spherical harmonic representation of its surface [3]. The surface of the object is mapped on

to a sphere through a transformation that minimizes area distortion. The mapping results in a representation of the original surface through a set of three coordinate functions ($x(\phi, \theta)$, $y(\phi, \theta)$, $z(\phi, \theta)$) on the unit sphere. Each function maps a point on the sphere (given by ϕ and θ) to the corresponding coordinate on the surface. A representation of these coordinate functions is obtained in terms of their projection onto spherical harmonic basis functions.

The function $\mathbf{x}(\phi, \theta)$ is given by $x(\phi, \theta) = \sum_{l=0}^{\infty} \sum_{m=-l}^l c_{l,m}^x Y_{l,m}(\phi, \theta)$ where $Y_{l,m}(\phi, \theta)$ are the spherical harmonic basis functions of degree l and order m [3]. Similarly for $\mathbf{y}(\phi, \theta)$ and $\mathbf{z}(\phi, \theta)$. Fig. 1(c) shows the spherical harmonics reconstruction of a nucleus using $l = 5$.

Radial profile—The appearance of the nuclear volume is modeled using the radial distribution of the density of the DNA stain as a function of its distance from the nuclear surface. This feature can be efficiently computed using the distance transform on the foreground mask of the nucleus (Fig. 1b). The distances are scaled to lie in the range $[0, 1]$, binned at intervals of fixed length and the average DNA intensity for each bin is computed to give the radial profile.

Spatial distribution—The spatial characteristic of a nucleus is represented by the density of cells in the local neighborhood of the cell. The local density is computed by identifying nuclear centers in the image and computing the kernel density estimate of the distribution of points in the 3D volume. The size of the kernel bandwidth σ_d is set as a certain factor of the average nuclear diameter.

4 Experimental Results

The methods were applied to a study of the tumor microenvironment in murine breast cancer [18] that seeks to understand the role of stromal cells in the initiation and progression of cancer. Previous studies using knockout models have shown that fibroblasts, the primary constituents of the extra-cellular matrix, play a role in tumor suppression in breast cancer. Other cells in the microenvironment – most importantly the macrophages and endothelial cells – are also known to participate in the cancer process. In these experiments, we focus on identifying the different cell-types in the microenvironment of a normal mouse mammary gland based on nuclear morphology and spatial positioning within the tissue, and further, on identifying sub-populations that may exist within these cell-types.

Data collection

Tissue sections were collected from two-month wild-type mice. The cell-types are identified using fluorescent proteins that are endogenously expressed in the transgenic mouse lines. In a given specimen in our experiments, endogenous fluorescence is present in only one of the three stromal cells (macrophages, fibroblasts or endothelial cells). To identify cell nuclei, the tissue sections were stained with DAPI, a fluorescent marker for DNA. An Olympus FV1000 confocal microscope with an objective of 40x/1.3NA was used to collect images. The images were acquired at an in-plane resolution $0.31\mu\text{m}$ and axial resolution of $0.5\mu\text{m}$, and have field of view $317\mu\text{m} \times 317\mu\text{m}$ and depth of $40\text{--}70\mu\text{m}$.

Image Processing

The confocal images consist of two channels corresponding to the nuclear stain (represented in green) and the cell-specific fluorescent protein (represented in red). The channels were denoised and the nuclear channel was segmented using a standard processing pipeline [15]. Segmentation errors were manually corrected in most cases, while a small fraction of poor segmented nuclei were discarded. The fluorescence protein channel was then used to detect nuclei corresponding to the cell-type identified by the protein. The set of nuclei selected through this process were manually validated. These operations were executed on images for each cell-type and constituted the labeled examples for the study. The nuclei that were not identifiable by the fluorescent protein were used as unlabeled examples.

Feature Extraction

The spherical harmonics coefficients were computed based on the SPHARM-PDM method [17] using a degree of $l = 5$ resulting in a 108-dimensional shape feature. The radial profile was computed by discretizing into 10 bins, thereby resulting in a 10-dimensional appearance feature. The spatial density of nuclei was computed by first identifying nuclear centers from the segmentations and then computing the kernel density estimate using a Gaussian kernel with bandwidth of $20\mu\text{m}$. This results in a scalar-valued density feature. The dimensionality of the shape and appearance features were reduced independently using PCA, maintaining 95% of the data variance. Metric learning was performed on the resulting in a set of features.

4.1 Cell-Type Classification

Results were generated on a dataset of 984 nuclei, of which 229 were labeled, consisting of 102 macrophages, 95 fibroblasts and 32 endothelial cells. In the following, the semi-supervised extension refers to the use of the label propagation technique described in Section 3. Results on cell-type classification were obtained using k -nearest neighbors classifier. Classification rates were compared for five different metrics – the standard Euclidean metric (EUC), the global metric [5] with and without semi-supervised extension (GMET-SM and GMET-S respectively) and the proposed local metric with and without semi-supervised extension (LMET-SM and LMET-S respectively). An entropy threshold of 0.01 was used for the semi-supervised extensions in all cases. In LMET-SM and LMET-S, the number of GMMs was set to 3, equal to the number of classes in the data set. The comparative results are summarized in Fig. 2. Classification rates were averaged over 5 runs.

The rates for both GMET-S and LMET-S show an average improvement over EUC. However the error variance in both cases was high because of the small size of the training set. The variance decreases significantly in GMET-SM and LMET-SM with the use of unlabeled data that provides a more robust estimate of the metrics. In all cases, the classification accuracy for macrophages was higher than for the other two cell-types. This is explained by the fact that in many cases fibroblasts and endothelial cells have similar morphologies, making the task of discriminating between them difficult. These error rates decrease in LMET-SM, due to the local metric being able to learn the appropriate weighting of the spatial feature as shown in Fig. 3. In this figure, while the nuclei come from two

different cell-types, their morphologies are very similar. This is seen in the 2D segmentation masks as well as in the 3D visualization of the difference between the shapes.

However, endothelial cells lie in closely packed regions surrounding blood vessels while fibroblasts are typically scattered in the stroma as seen in the left panel of Fig 3. Thus the spatial neighborhood is a good feature to discriminate between these two nuclei. The learned metric was able to accurately identify this difference by appropriately weighting the spatial feature to discriminate between these samples.

4.2 Identifying Cellular Sub-populations

There is significant genetic and physiological heterogeneity observed within cell-types [11]. The problem of identifying patterns of cellular heterogeneity has been studied in *cell culture* to understand the effects of perturbations in a system [16]. In the following, we attempt to identify cellular subtypes *within the tissue sections* by using the locally adaptive metrics to induce a clustering on the set of nuclei. For each cell-type, a weighted graph was constructed from the data, with the vertices as nuclei and edge weights given by the distance under the learned metric. By employing a graph clustering method [19], subgroups within each cell-type were identified by manually choosing a scale parameter that resulted in meaningful clusters based on visual examination. The subgroups that were discovered through this process are shown in Fig. 4. For fibroblasts, it was observed that the three nuclear subgroups vary primarily in the degree of flatness of the nuclear shape. The macrophage subgroups were observed to differ in size as well as in the internal appearance. For example, the MAC1 group consists of nuclei that have low intensity regions in the interior, most likely due to the macrophages undergoing phagocytosis [11] (engulfing a foreign substance). Nuclei in groups MAC1 and MAC3 were observed to be similar in shape and size, both being significantly smaller than MAC2. The differences between the groups END1 and END2 appear to be that of elongatedness and curvedness of their shapes. The most likely explanation for this difference is the heterogeneity of cell-types that line blood vessels – endothelial cells form the inner lining of blood vessels, and a related cell-type, *pericytes*, wrap around the vessels forming an outer sheath. Since these cells are in very close proximity it is difficult to distinguish between them using fluorescent markers. The cluster analysis however is able to distinguish between these two types based on their nuclear morphology. This result is an indicator that morphological analysis of nuclei can be used to supplement traditional staining techniques or endogenous fluorescent proteins in distinguishing between cell-types.

5 Conclusions

In this paper, an approach was presented for analyzing nuclear phenotypes to identify cell-types in 3D images of tissue sections. This problem poses several challenges – (a) a diverse set of features is required to describe the nuclear phenotype (b) training data available per experiment is limited, and (c) the existence of cellular subpopulations results in phenotypic heterogeneity within cell-types. An approach of constructing a *locally adaptive* distance metric learning framework was adopted to address this problem, which provided a principled way of comparing a pair of nuclei by learning the structure of the metric locally

in a supervised framework. In addition, a label propagation scheme was employed that incrementally expands the training set by selecting previously unlabeled points which can be confidently labeled, thereby circumventing the constraint of limited training examples. In addition to demonstrating results on the problem of cell-type identification, the learned metric was also used to identify cellular *subtypes* using a clustering strategy. Through the latter, we demonstrated the existence of phenotypic heterogeneity of nuclei within cell-types.

Acknowledgments

This work was supported by National Institutes of Health Grants UL1RR025755 and 5 P01 CA 097189-05.

References

1. Bissell MJ, Radisky D. Putting tumours in context. *Nature Reviews. Cancer*. 2001; 1(1):46–54.
2. Boland MV, Markey MK, Murphy RF. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*. 1998; 33(3):366–375. [PubMed: 9822349]
3. Brechbühler C, Gerig G, Kibler O. Parametrization of closed surfaces for 3-d shape description. *Computer Vision and Image Understanding*. 1995; 61(2):154–170.
4. Cormen, TH.; Leiserson, CE.; Rivest, RL.; Stein, C. *Introduction to Algorithms*. The MIT Press; New York: 2001.
5. Davis, JV.; Kulis, B.; Jain, P.; Sra, S.; Dhillon, IS. Information-theoretic metric learning. *Proceedings of the 24th International Conference on Machine Learning, ICML 2007*; 2007. p. 209–216.
6. Eaton, ML. *Lecture notes-monograph series*. Vol. 53. Institute of Mathematical Statistics; 1983. *Multivariate statistics: a vector space approach*.
7. Gladilin E, Goetze S, Mateos-Langerak J, Van Driel R, Eils R, Rohr K. Shape normalization of 3D cell nuclei using elastic spherical mapping. *Journal of Microscopy*. 2008; 231(Pt 1):105–114. [PubMed: 18638194]
8. Goldberger J, Roweis S, Hinton G. Neighbourhood components analysis. *Advances in Neural Information Processing Systems*. 2004
9. Kulis, B.; Sustik, M.; Dhillon, I. Learning low-rank kernel matrices. *Proceedings of the 23rd International Conference on Machine Learning*; New York: ACM; 2006. p. 505–512.
10. Lichtman JW. Fluorescence microscopy. *Nature Methods*. 2005; 2(12)
11. Lodish, H.; Berk, A.; Kaiser, CA.; Krieger, M.; Scott, MP.; Bretscher, A.; Ploegh, H.; Matsudaira, P. *Molecular Cell Biology*. 6th edn. W.H. Freeman; New York: 2007.
12. Megason SG, Fraser SE. Imaging in Systems biology. *Cell*. 2007; 130(5):784–795. [PubMed: 17803903]
13. Rohde GK, Ribeiro AJS, Dahl KN, Murphy RF. Deformation-based nuclear morphometry: capturing nuclear shape variation in HeLa cells. *Cytometry. Part A: the Journal of the International Society for Analytical Cytology*. 2008; 73(4):341–350.
14. Shaner NC, Steinbach PA, Tsien RY. A guide to choosing fluorescent proteins. *Nature Methods*. 2005; 2(12):905–909. [PubMed: 16299475]
15. Singh, S.; Raman, S.; Caserta, E.; Leone, G.; Ostrowski, M.; Rittscher, J.; Machiraju, R. Analysis of Spatial Variation of Nuclear Morphology in Tissue Microenvironments. 2010 7th IEEE International Symposium on Biomedical Imaging: From Nano to Macro; Los Alamitos: IEEE; 2010.
16. Slack MD, Martinez ED, Wu LF, Altschuler SJ. Characterizing heterogeneous cellular responses to perturbations. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(49):19306–19311. [PubMed: 19052231]

17. Styner M, Oguz I, Xu S, Levitt JJ, Shenton ME, Gerig G. Framework for the Statistical Shape Analysis of Brain Structures using SPHARM-PDM. *Insight Journal*. 2006:1–20.
18. Trimboli AJ, Cantemir-Stone CZ, Li F, Wallace JA, Merchant A, Creasap N, Thompson JC, Caserta E, Wang H, Chong JL, Naidu S, Wei G, Sharma SM, Stephens JA, Fernandez SA, Gurcan MN, Weinstein MB, Barsky SH, Yee L, Rosol TJ, Stromberg PC, Robinson ML, Pepin F, Hallett M, Park M, Ostrowski MC, Leone G. Pten in stromal fibroblasts suppresses mammary epithelial tumours. *Nature*. 2009; 461(7267):1084–1091. [PubMed: 19847259]
19. Van Dongen, S. A cluster algorithm for graphs. 2000.
20. Weinberger KQ, Saul LK. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*. 2009; 10:207–244.
21. Xing, EP.; Ng, AY.; Jordan, MI.; Russell, S. Distance metric learning, with application to clustering with side-information; *Advances in Neural Information Processing Systems*; 2003.
22. Zhu, X.; Ghahramani, Z.; Lafferty, J. Semi-supervised learning using gaussian fields and harmonic functions. Vol. 20. *ICML*; 2003. p. 912
23. Zink D, Fischer AH, Nickerson JA. Nuclear structure in cancer cells. *Nature reviews. Cancer*. 2004; 4(9):677–687.

Fig. 1. Different views of a cell nucleus

(a) Maximum intensity projection of cell nucleus with the automatic segmentation outlined. (b) Distance transform of nuclear mask to generate radial profile. (c) Spherical harmonics reconstruction of the nuclear surface. (d) Iso-surface of the tissue image with the selected nucleus in red.

Fig. 2.
Classification rates for the three cell-types

Fig. 3. Spatial neighborhood used to discriminate between cell-types in the absence of a significant shape differences. (Top) endothelial cell; (Bottom) fibroblast; (Right) visualization of shape difference. Surface of endothelial nucleus is in blue, arrows indicate the difference between nuclear shapes.

Fig. 4.

Nuclear subtypes within fibroblasts (FIB 1–3), macrophages (MAC 1–3) and endothelial cells (END 1–3). Rows correspond to different examples of the nuclear subtypes. Nuclei are colored in green, the cell-specific fluorescence in red.