Machine Learning Based Methodology to Identify Cell Shape Phenotypes Associated with Microenvironmental Cues

Desu Chen
Biophysics Program
University of Maryland
College Park, MD, US
desuchen@umd.edu

Julian Candia, Meghan K. Driscoll, Wolfgang Losert
Department of Physics
University of Maryland
College Park, MD, US

Sumona Sarkar, Stephen J. Florczyk, Subhadip Bodhak, Carl G. Simon, Jr., Joy P. Dunkers
Biosystems & Biomaterials Division
National Institute of Standards & Technology
Gaithersburg, MD, US

***Abstract*— Cell shape has been demonstrated to be closely related to stem cell response and function in biomaterial environments. However, cell shape phenotyping in biomaterials with bioimage data is complicated by heterogeneous cell populations, microenvironment heterogeneity, and multi-parametric definitions of cell morphology. To associate cell morphology with cell-material interactions, we developed an analysis framework based on support vector machines (SVMs) with a multi-cell level "supercell" averaging method to build classifier boundaries that identify and predict microenvironment-driven morphology differences of cell populations. The "supercell" method reduces the influence of variability in single-cell morphology on the classification of cell populations with SVMs. We compared morphologies of human bone marrow stromal cells (hBMSCs) cultured on nanofiber scaffolds to those on flat films after one day of culture. Smaller cell size and more dendritic shape patterns were the major morphological responses of hBMSCs to nanofiber scaffolds.**

***Index Terms*— Cell morphology, biomaterial, support vector machine, supercell.**

I. Introduction

Cell morphology may be a valuable descriptor of cell behaviors, phenotypes and genotypes in different biological procedures such as immune response, cancer progress and differentiation [1-5]. High-throughput single-cell bioimaging has enabled the quantification of heterogeneous cell population with many cell shape features that are increasingly difficult to interpret. Innovative analytical tools must be developed to identify and combine key cell shape features correlated with biological outcome while accounting for both multi-parametric complexity and biological heterogeneity.

In this study, we investigated the morphology of human bone marrow stromal cells (hBMSCs) in nanofiber scaffolds compared to that of cells on flat films. Nanofiber scaffold structures have been demonstrated to uniquely induce osteogenic differentiation of human bone marrow stromal cells (hBMSCs) and alter cell shape, similarly to chemically induced differentiation [6]. However, only a few individual cell shape features have been investigated for their association with differentiation, and cell morphologies vary greatly across a nanofiber scaffold. To address this limitation, we have developed computational tools based on Support Vector Machines (SVMs) to identify cell morphological features associated with nanofiber [7] in a wide range of global or local shape metrics. Moreover, the resulting SVM classifiers provided a selection of reduced shape metrics to quantify hBMSC shape phenotypes in specific microenvironments. However, large variability in cell shape led to highly overlapping cell populations. In order to improve the training and prediction accuracies of the SVM classifiers, a method of averaging shape metrics over a small subset of randomly selected cells known as "supercell averaging" was implemented [8]. The random sampling used to generate supercells can introduce uncertainty in the SVM classifier. Therefore, by introducing a subsampling validation procedure, we studied the sample size as another important limiting factor in the construction of single-cell or supercell phenotypes and its effects on the tradeoff between prediction accuracy, supercell averaging and uncertainty in the classifier.

II. Material and Methods

A. Sample Preparation, Cell Culture and Imaging

Poly(ε-caprolactone) (PCL) films (SC) were generated with spin-coating and PCL nanofiber scaffolds (NF) were fabricated by electrospinning onto tissue culture polystyrene discs. hBMSCs were seeded and cultured on the PCL films and PCL nanofiber scaffolds for 24 hours (37° C, 5% CO2), with or without oseteogenic supplement (OS) of dexamethasone (10 nmol/L), β-glycerophosphate (20 mmol/L) and ascorbic acid (0.05 mmol/L). Cells were then fixed with 3.7% formaldehyde and permeabilized with 0.1% Triton-X, then stained with Alexa Fluor 546 phalloidin (0.33μM)) for actin and 4',6-diamidino-2-phenylindole (DAPI, 0.03mM) for nucleus. High-resolution 3-D z-stack images of hBMSCs were taken with a confocal microscope (Leica SP5) with 63x water immersion objective. A total of 121 hBMSCs in NF, 114 hBMSCs on SC, 125 hBMSCs in NF+OS and 116 hBMSCs in SC+OS were imaged.

B. Cell Shape Analysis

Max projections of the z stacks were processed with snake algorithm [9] to define cell outlines with sub-pixel resolution in MATLAB (Fig. 1. a). 22 Shape metrics were calculated with the outline of each hBMSC and normalized with z-score (Fig. 1.b). For the cell population of each culture condition, 120 supercells were generated by averaging the shape metrics over a certain number (supercell size) of randomly picked original single cells with replacement. SVMs classifiers with linear kernel (implemented with kernlab package in R) were trained on different random supercell data sets for 100 times in pairwise comparisons. The final classifier hyperplane orientation was defined by the average normal vector $\bar{\mathbf{n}}$ over all normal vectors $\mathbf{n}$ of each machine learning repeat, i.e.

$$\bar{\mathbf{n}} = \sum_{all} \mathbf{n} \Big/ \left\| \sum_{all} \mathbf{n} \right\| \tag{1}$$

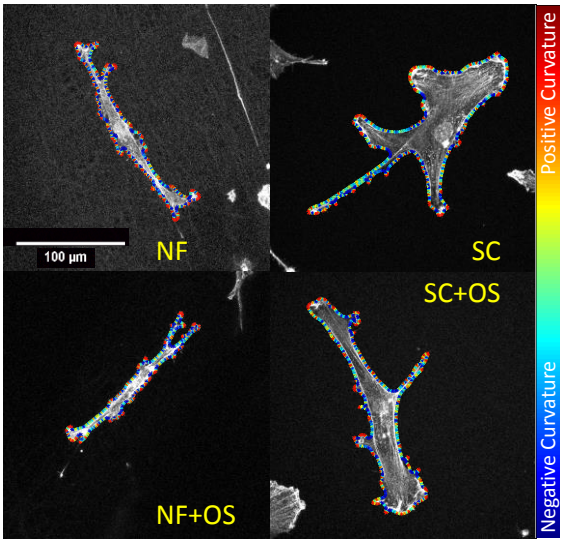
The classifier hyperplane stability was then measured as the average cosine function <cosθ> (inner product) of the angle θ between the instant normal vector of each machine learning procedure and the average classifier normal vector.

$$\langle \cos\theta \rangle = \langle \bar{\mathbf{n}} \cdot \mathbf{n} \rangle \tag{2}$$

Shape metrics showing statistically significant differences ($p < 0.01$) between micro-environments, were preselected based on 1-way ANOVA and Tukey multi-comparison test. Then, all combinations of 3 shape metrics were used to build different metric spaces for the subsequent SVM analysis. The combination of 3 shape metrics with the highest training classification accuracy among those that satisfy a certain classifier hyperplane stability criterion was finally selected to represent the population morphology difference.

With the selected shape metrics, a subsampling validation procedure was employed to decide which training data size and supercell size are appropriate to build the classifier hyperplane. In this procedure, a training subsample of a certain size was randomly picked from the original cell population and then randomly generated 120 supercells of a certain supercell size. The SVM/supercell paradigm was applied to these data sets to train a classifier hyperplane. 120 supercells of the same supercell size were also randomly made with the remaining sample to form a test subsample. The hyperplane achieved with the training subsample was utilized to predict the test subsample. This subsampling validation procedure was repeated for 200 times for a certain training sample size and supercell size. The classifier hyperplane stability calculated with Eq. 2. Both prediction accuracy and the classifier hyperplane stability were taken into account to decide the appropriate training sample size and supercell size.

a



b

Metric Category	Metrics
Cell Size	Area (μm²) Perimeter (μm) Mean Boundary Distance (μm) Max Boundary Distance (μm) Major Axis Length (μm) Minor Axis Length (μm) Max Major Branch Length* (μm) Mean Major Branch Length* (μm) Max Major Branch Width* (μm) Mean Major Branch Width* (μm)
Global Shape Pattern	Circularity Solidity Aspect Ratio Num. of Major Branches* Max Branch Length/ Major Axis* Mean Branch Length/ Equivalent Radius*
Local Feature	Mean Negative Curvature (μm⁻²) Invagination (μm⁻¹) Tortuosity (μm⁻⁴) Num. of Tiny Protrusions* Num. of Protrusions per Branch* Average Boundary Interval Between Protrusions* (μm)

Fig. 1. Quantification of cell shapes in different microenvironment (NF and SC with or without OS). (a) Outlines of hBMSCs were obtained with snake algorithm which allowed calculation of local curvature. Boundary regions were colored differently according to local curvatures. (b) 22 metrics were quantified to describe hBMSC shapes and sorted into 3 categories about different aspects of cell shape. 12 metrics were obtained with the snake outlines (without asterisks) and 10 metrics were obtained from branch analysis (with asterisks).

III. Results

To distinguish morphologies of hBMSCs in NF and SC, the optimal combination of 3 shape metrics were identified as minor axis length, solidity and mean negative curvature (supercell size = 5, hyperplane stability threshold <cosθ> > 0.99). The accuracy of the classifier training is 99.3%±0.6% (Fig. 2.a). The average normal vector of the classifier hyperplane is (-0.86±0.04, -0.43±0.06, 0.24±0.08).

In a subsampling validation procedure to test the classifier built with the selected shape metric combination of minor axis length, solidity, and mean negative curvature, both the training subsample size and the supercell size to build the classifier varied. The classifier hyperplane stability was improved with increasing number of cells in the training set to build the classifier. The classifier stability threshold of $\langle\cos\theta\rangle > 0.99$ was still assumed to define stable classifier hyperplanes. In Fig. 2.b, classifier hyperplane stability and prediction accuracy were combined to quantify effect of data size and supercell size on the classifier for selected shape metrics. Figure 2.b showed that morphology difference should be quantified with appropriate selections of supercell size and training data size and supports the efficacy to train a stable classifier hyperplane with the selected shape metrics at supercell size of 5 and current data size (121 hBMSCs of NF and 114 hBMSCs of SC).

IV. Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

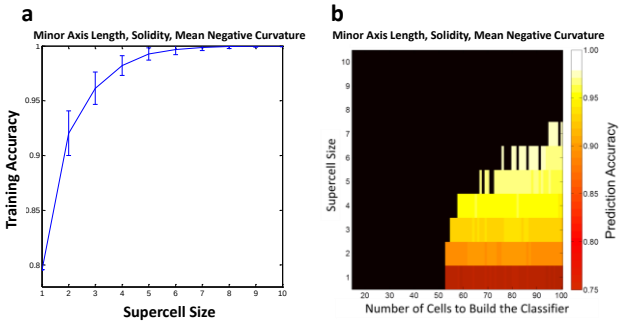


Fig. 2. Results of SVM analysis and associated subsampling test of selecting 3 shape metrics to compare morphological difference of hBMSC populations of NF and SC. (a) Training classification accuracy of the selected shape metric combination in SVM training with supercell implementation. All error bars represent standard deviation.
(b) Prediction accuracy of the classifier hyperplanes in the subsampling validation when the built classifiers were tested with the rest of the total sample at different supercell sizes. The dark region represented combinations of training data size and supercell size causing unstable classifier hyperplane. In the stable region, combinations of the training data size and supercell size were colored according to the prediction accuracy.

V. Acknowledgment

WL and DC acknowledge NIST grant 70NANB14H282 and WL and JC acknowledge NSF grant PHY120596.

VI. References

[1] R. McBeath, D. M. Pirone, C. M. Nelson, K. Bhadriraju, and C. S. Chen, "Cell Shape, Cytoskeletal Tension, and RhoA Regulate Stem Cell Lineage Commitment," Developmental Cell, 2004. **6**(4): p. 483-495.

[2] K. A. Kilian, B. Bugarija, B. T. Lahn, and M. Mrksich, "Geometric cues for directing the differentiation of mesenchymal stem cells," Proc Natl Acad Sci U S A, 2010. **107**(11): p. 4872-7.

[3] M. D. Treiser, E. H. Yang, S. Gordonov, D. M. Cohen, I. P. Androulakis, J. Kohn, et al., "Cytoskeleton-based forecasting of stem cell lineage fates," Proc Natl Acad Sci U S A, 2010. **107**(2): p. 610-5.

[4] H. V. Unadkat, N. Groen, J. Doorn, B. Fischer, A. M. Barradas, M. Hulsman, et al., "High content imaging in the screening of biomaterial-induced MSC behavior," Biomaterials, 2013. **34**(5): p. 1498-505.

[5] T. L. Downing, J. Soto, C. Morez, T. Houssin, A. Fritz, F. Yuan, et al., "Biophysical regulation of epigenetic state and cell reprogramming," Nat Mater, 2013. **12**(12): p. 1154-62.

[6] G. Kumar, C. K. Tison, K. Chatterjee, P. S. Pine, J. H. McDaniel, M. L. Salit, et al., "The determination of stem cell fate by 3D scaffold structures through the control of cell shape," Biomaterials, 2011. **32**(35): p. 9188-96.

[7] N. Cristianini and J. Shawe-Taylor, "An introduction to support vector machines : and other kernel-based learning methods." 2000, Cambridge, U.K. ; New York: Cambridge University Press. xiii, 189 p.

[8] J. Candia, R. Maunu, M. Driscoll, A. Biancotto, P. Dagur, J. P. McCoy, Jr., et al., "From cellular characteristics to disease diagnosis: uncovering phenotypes with supercells," PLoS Comput Biol, 2013. **9**(9): p. e1003215.

[9] X. Chenyang and J. L. Prince, "Snakes, shapes, and gradient vector flow," Image Processing, IEEE Transactions on, 1998. **7**(3): p. 359-369.