Non-parametric Population Analysis of Cellular Phenotypes

Shantanu Singh[1], Firdaus Janoos[1], Thierry Pécot[1], Enrico Caserta[3], Kun Huang[2], Jens Rittscher[4], Gustavo Leone[3], and Raghu Machiraju[1]

[1] Dept. of Computer Science and Engg., The Ohio State University, U.S.A.
[2] Dept. of Biomedical Informatics, The Ohio State University, U.S.A.
[3] Dept. of Molecular Genetics, The Ohio State University, U.S.A.
[4] General Electric Global Research Center, U.S.A.

Abstract. Methods to quantify cellular–level phenotypic differences between genetic groups are a key tool in genomics research. In disease processes such as cancer, phenotypic changes at the cellular level frequently manifest in the modification of cell population *profiles*. These changes are hard to detect due the ambiguity in identifying distinct cell phenotypes within a population. We present a methodology which enables the detection of such changes by generating a phenotypic signature of cell populations in a data–derived feature–space. Further, this signature is used to estimate a model for the *redistribution* of phenotypes that was induced by the genetic change. Results are presented on an experiment involving deletion of a tumor–suppressor gene dominant in breast cancer, where the methodology is used to detect changes in nuclear morphology between control and knockout groups.

Keywords: Microscopy Image Analysis, Cell Nucleus, Shape Analysis.

1 Introduction

Gene targeting methods have elucidated the roles of several genes in disease processes such as cancer. In these experiments, a genetic perturbation such as a gene knockout is introduced in a model organism and the effects on the genetic as well as phenotypic makeup of the organism are examined. Recent advances in microscopy coupled with automated image analysis tools have enabled researchers to *quantify* a broad range of such phenotypic alterations at the cellular level [8]. In these studies, the phenotype of interest, such as cell morphology, is quantified using image features across a population of cells, and differences between normal and genetically perturbed populations are examined. In the event that a clear hypothesis can be posed based on a priori knowledge, it is possible to extract specific cellular features of interest [2] and investigate this hypothesis using standard statistical tests.

In the absence of such prior knowledge, investigations rely on exploratory data analysis techniques, making the detection of phenotypical changes more challenging. In addition, some of the changes affect only a certain subpopulation

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part II, LNCS 6892, pp. 343–351, 2011.

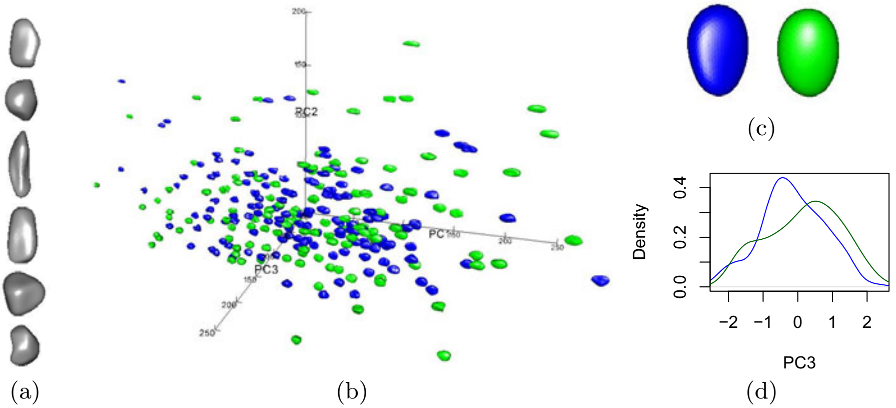


Fig. 1. Detecting changes in cell populations. (a) Examples of fibroblast nuclei from the study (b) Scatter plot of fibroblast nuclei obtained from a control group (Wild-type) and genetically modified group (*PTEN*-deleted) plotted in top three principal components (PC) of the shape space; no evidence of separation between the groups (c) Mean nuclear shapes of the two groups; difference between means is not statistically significant (Sec. 4.1) (d) Estimated (marginal) densities of the two groups in the third PC; *distributional* differences between the groups are statistically significant (Sec. 4.1).

of the cells, resulting in very subtle differences between groups. In this case it is no longer sufficient to analyze the global statistic of a specific cellular feature to detect the presence of an effect. In particular, during the onset of a disease such as the early stage cancer, small perturbations in the system are not easily detectable using existing methods. The quantification of such changes through the analysis of alterations in the population as a whole is the focus of this paper.

In this paper a methodology is presented that is applied to detect the phenotypic effect of deletion of a tumor–suppressor gene (*PTEN*) in the context of breast cancer. Deletion of *PTEN* in fibroblasts has been shown to increase the predisposition to cancer in certain mouse models [11]. Of specific interest are changes at the cellular level in the early stages of cancer development which provide an important indication of what precursor activities occur prior to tumor metastasis. Motivated by the seminal role played by the cell nucleus in cancer [13], nuclear morphology has been used in this study as a proxy for the cell phenotype. The study focuses on detecting the effect of *PTEN* deletion on the nuclear shape of fibroblasts in an exploratory manner. The following data analysis approach, illustrated in Fig.1, explicates the guiding principles of the proposed method.

A spherical harmonics–based shape parameterization was used to represent nuclear morphology (Sec. 3.1) and PCA was used to reduce dimensionality. A plot of the data in the top three principal components is shown in Fig. 1b. No evidence of a discernible separation between the genotype groups was observed. The differences in the mean shapes (Fig. 1c) were not observed to be statistically significant based on a non–parametric test for differences between means (Sec. 4.1). However, the presence of an effect was observed when testing for differences in the *shape* of the distributions of the two populations (Fig. 1d). Note that the shape of the distribution essentially represents the phenotypic *population profile*, a characterization of the mix of the nuclear phenotypes present in a population. By using a non–parametric test for difference in distributions between the two groups, a difference between the marginal distributions in the third PC was observed at a significance of $p = 0.002$ (See Sec. 4.1 for further details).

The principles behind this analysis approach are summarized as follows. *First*, the feature set used is *agnostic* in that we did not seek specific structures (such as in [2]) to measure the cell phenotype. Rather, the biologically relevant modes of variation were learned from data using dimensionality reduction, which were used as the feature set. *Second*, the cell population exhibits a significant degree of phenotypic heterogeneity as seen in Fig. 1a. The change in the population profile corresponds to a change in the *composition* of this heterogeneity, which is detected through the analysis. In summary, the above approach enables the detection of phenotypic changes in *population profiles* in a *data–driven* manner.

In this paper, the proposed methodology builds on these principles to further estimate an *interpretable model* for the changes that have occurred. This is achieved by modeling the *redistribution* of cellular phenotypes that putatively occurred due to the induced perturbation. The formulation of the problem is equivalent to finding the Earth Mover's Distance (EMD) [10] in a data–derived feature space. The rest of the paper is organized as follows. A review of related work is presented in Section 2. The proposed methodology is discussed in Section 3 and results on breast cancer datasets are presented in Section 4. We conclude with a discussion in Section 5.

2 Related Work

Several methods for modeling cellular morphology have been proposed in image cytometry literature. A physically–based deformation model of a cell was proposed in [4] which served as an atlas to compare sub–cellular features. A multi–modal approach incorporating diverse cellular features was proposed in [6] that enabled simultaneous segmentation and classification of cells. A statistical methodology for the analysis of sub-cellular features was proposed in [1] that used descriptors from spatial statistics to characterize the internal structure of cells. The above approaches focused on the characterization of individual cells. A computational anatomy framework was used in [9] to quantify shape differences in cells by analyzing the deformation fields in mapping between the samples. This approach enables group analysis on cell populations in an agnostic framework, similar to the approach proposed in this paper. In contrast to the above approaches, this paper presents a methodology that enables the detection of subtle variations between population profiles as a whole, and additionally provides an interpretable model of differences between them.

3 Mathematical Approach

A non–parametric shape representation is adopted in order to support an unbiased exploratory study of nuclear phenotypes. Let $\psi : \mathcal{Z} \rightarrow \mathcal{X}$ denote the shape representation, where $\mathcal{Z}$ is the domain of nuclear shapes and $\mathcal{X}$ is the domain of shape features. A low–dimensional feature–space $\tilde{\mathcal{X}}$ corresponding to biologically plausible nuclear shapes is estimated by learning a transformation $\phi_D : \mathcal{X} \rightarrow \tilde{\mathcal{X}}$ where $D \subset \mathcal{X}$ is a large set of nuclear shapes. Given a normal cell population $P \subset \mathcal{Z}$ and a perturbed one $Q \subset \mathcal{Z}$, the set of most likely transitions that explain the apparent redistribution of phenotypes is estimated. A process similar to computing the Earth Mover's Distance metric [10] is used to measure the total cost of the transformation (Sec. 3.2).

3.1 Nuclear Morphology Representation

Nuclear morphology is modeled using a spherical harmonic (s.h.) representation of its surface [3]. The surface is mapped to a sphere by a distortion–minimizing transformation resulting in the representation of the original surface through the set of coordinate functions $[x(\phi, \theta), y(\phi, \theta), z(\phi, \theta)]$ on the unit sphere. These functions are represented in the s.h. basis as $x(\phi, \theta) \triangleq \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{l,m}^{x} Y_{l,m}(\phi, \theta)$ where $Y_{l,m}(\phi, \theta)$ are the s.h. basis functions of degree l and order m and $c_{l,m}^{x}$ is the corresponding coefficient [3]. The functions $y(\phi, \theta)$ and $z(\phi, \theta)$ are represented similarly. The resulting set of s.h. coefficients form the shape vector, with dimensionality $3(L+1)^2$, where L is the maximum degree of the harmonics considered.

Previous *in vivo* studies indicate that changes in morphology have low–dimensional modes of variation [5]. Linear dimensionality reduction methods are well suited to estimate these modes of variation [7]. Motivated by these studies, PCA is subsequently used to estimate a low–dimensional subspace of biologically relevant nuclear shapes. The resulting linear transform is represented as ϕ_D^{ϵ}, where ϵ is the fraction of variance that is discarded, resulting in a k–dimensional embedding.

3.2 Estimating Redistribution of Phenotypes

The differences between the distributions of P and Q represent the differences in the phenotypic profiles of the two populations of cells. This difference is modeled in terms of the putative *redistribution* of phenotypes that transformed P to Q. To estimate this, the optimal "moves" required to transform one population to another is computed under a cost model $d : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$. The choice of d is taken to be the Euclidean distance in ϕ_D^{ϵ} since this subspace captures the principal modes in which cell shapes vary [7]. It is noted here that the above is a formulation of the earth mover's distance [10] and can be solved using linear programming as follows.

Denote $P_\phi = \phi_D^{\epsilon} \circ \psi(P)$ where $\psi(P)$ corresponds to the shape representation and ϕ_D^{ϵ} to the low–dimensional embedding. P_ϕ and Q_ϕ are converted to empirical normalized histograms in the form $\hat{P}_\phi \triangleq \{(s_1, p_1), (s_1, p_1), \ldots, (s_1, p_n)\}$ and $\hat{Q}_\phi \triangleq \{(t_1, q_1), (t_1, q_1), \ldots, (t_1, q_m)\}$ where the pairs (x, c) indicate that $x \in \tilde{\mathcal{X}}$ is a bin center and c is the normalized frequency of points in the bin. The bin centers are obtained by vector quantization. The optimal solution is found in terms of the "flow" $R \triangleq (r_{ij})$ that minimizes the total cost $\sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} d_{ij}$ where $d_{ij} \triangleq d(s_i, t_j)$ and R obeys the constraints (i) $r_{ij} \geq 0$, $1 \leq i \leq m$, $1 \leq j \leq n$ (ii) $\sum_{i=1}^{m} r_{ij} \leq p_j$, $1 \leq j \leq n$ (iii) $\sum_{j=1}^{n} r_{ij} \leq q_i$, $1 \leq i \leq m$ (iii) $\sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} = min(\sum_{i=1}^{m} p_i, \sum_{j=1}^{n} q_j)$.

Given the optimal solution $R^* \triangleq (r_{ij}^*)$, the putative redistribution is obtained by *removing* r_{ij} instances of phenotype s_i from $\hat{P}_\phi$ and *adding* an equal number of instances of phenotype t_j to $\hat{Q}_\phi$. Table 2 provides an illustrative example for this interpretation. Thus by the above formulation, we are able to *interpret* the difference between the two populations in terms of the most likely phenotypic transitions that occurred. Further, the difference between the two distributions is measured by the total cost of the transformation given by $D(P_\phi, Q_\phi) \triangleq \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij}^* d_{ij} / \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij}^*$.

In experiments with small sample sizes, histogram estimation in the dimensionality of the embedded space may result in unstable estimates. In this event, the following strategy is used to retain the dimensions in which the marginal distributional differences are (statistically) most significant. Denote $\phi_i(\cdot) \triangleq [\phi(\cdot)]_i$. The statistic $D(P_{\phi_i}, Q_{\phi_i})$ is computed for $i \in \{1 \ldots k\}$ and its significance is obtained through permutation testing. The bases are ranked by increasing order of p-value, and a cutoff p-value is used to select the top l features.

4 Results

The method was applied to the knockout study discussed in Section 1 to determine the phenotypic changes resulting from the selective deletion of the *PTEN* gene. The study focused on the morphological changes in fibroblasts, the cell type in which the gene was selectively deleted.

Data Collection and Feature Extraction. Tissue sections were collected from two–month old female mice belonging to control (P) and *PTEN*-deleted (Q) groups, with three mice per genotype group. The sections were stained with DAPI, a fluorescent marker for DNA, to identify cell nuclei. Endogenously expressed cell-specific fluorescence was used to detect fibroblasts. 3D images of the tissue were acquired with a confocal microscope at an in-plane resolution $0.31 \mu m$ and axial resolution of $0.5 \mu m$. Nuclei were segmented using Otsu thresholding followed by morphological closing to fill holes in the volume. Segmentation errors were manually corrected using a semi-automatic segmentation tool [12]. For each nucleus, s.h. coefficients were computed to the fifth degree resulting in a 108-dimensional feature–vector. In all, a random subset of 125 fibroblast nuclei from each group were selected for the study.

4.1 Statistical Tests

A set of statistical tests were performed on the cell populations to test for differences between the *PTEN*-deleted and control groups as follows. Tests were performed in *each dimension* of (a) the s.h. feature–space $\mathcal{S}_{sh}$ and (b) the principal components feature–space $\mathcal{S}_{pc}$, obtained by reducing dimensionality of $\mathcal{S}_{sh}$ keeping 95% of the total variation. Test for difference between the group means was performed using the Student's t test (ST). The Kolmogorov-Smirnov test (KS) was used to test for difference between the distributions of the groups. Permutation testing was performed over 20,000 iterations. In each iteration, the labels of the data were randomly permuted and the relevant test statistic was computed to generate the null distribution. Results from these tests are reported in Table 1. The columns $p^{(1)}$ through $p^{(5)}$ list the five most significant *uncorrected* (w.r.t. multiple comparisons) p-values in ascending order. The index of the component is shown in parentheses. The p-values that are significant *after* multiple test correction using false discovery rate method are underlined. For both, $\mathcal{S}_{sh}$ and $\mathcal{S}_{pc}$, ST did not report an effect in any of the features, an indicator that the heterogeneity of the cellular population precludes the use of a simple test of difference between means to identify differences. The KS test on the other hand, reported a statistically significant difference in the third PC. None of the components in $\mathcal{S}_{sh}$ reported significance in the KS test due to the effect of multiple comparison correction over 108 comparisons.

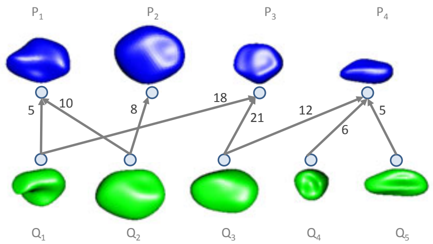
Table 1. Permutation testing results

Feat	Test	$p^{(1)}$	$p^{(2)}$	$p^{(3)}$	$p^{(4)}$	$p^{(5)}$
$\mathcal{S}_{sh}$	ST	0.177 (17)	0.209 (27)	0.387 (12)	0.513 (35)	0.551 (31)
$\mathcal{S}_{sh}$	KS	0.011 (12)	0.021 (23)	0.210 (17)	0.319 (31)	0.332 (27)
$\mathcal{S}_{pc}$	ST	0.191 (2)	0.262 (1)	0.291 (3)	0.655 (7)	0.706 (10)
$\mathcal{S}_{pc}$	KS	$\underline{0.002}$ (3)	0.050 (4)	0.181 (10)	0.237 (5)	0.286 (7)

4.2 Estimating Redistribution of Phenotypes

While the KS test established the presence of a difference between the distributions, there was no interpretation provided by the test about *how* the phenotype transformation occurred. Such an interpretation is provided by using the method described in Section 4, where the most likely redistribution that explains the differences is modeled as follows.

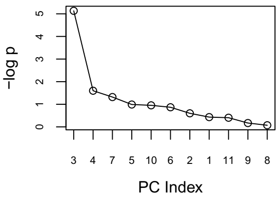
Estimating Embedding from Data. A collection of 1600 fibroblast nuclei collected across several images in the same animal model were used to obtain the biologically relevant modes of variation using PCA. By using a larger data set, a larger gamut of the biological variability is captured and results in a better characterization of the natural modes of variation. Keeping 95% of the total variation resulted in an 11-dimensional feature-space. The representation of the two populations in this feature space were computed.



	P_1	P_2	P_3	P_4
Q_1	**4.8(1.2)**	3.1(1.2)	**18.2(1.8)**	0.3(0.9)
Q_2	**10.1(2.1)**	**7.8(1.9)**	0.4(0.1)	0.3(0.1)
Q_3	1.1(0.6)	1.2(0.3)	**21.2(2.4)**	**12.1(1.8)**
Q_4	0.8(0.3)	2.1(0.8)	1.3(0.4)	**5.8(2.9)**
Q_5	0.4(0.1)	1.8(0.4)	0.3(0.1)	**6.9(1.2)**

Table 2. Estimating the redistribution of phenotypes. The nuclei $P_1 ... P_4$ and $Q_1 ... Q_5$ represent the cluster centers within the groups Wild-type and *PTEN*-deleted respectively. Edge weights correspond to total percentage of the optimal flow from one phenotype subgroup to another. Edges with weight less than five are not shown for clarity. The table shows the bootstrap estimates of edges weights demonstrating the stability of the solution (1 s.d. shown in brackets). Dominant flows are highlighted.

Estimating Redistribution. Since the sizes of the datasets in the experiment were small ($n = 125$/group), the dimensionality was further reduced as described in Sec. 3.2 in order to get stable estimates of histograms. The sorted plot of the negative log p-values are shown in the inset. A sharp knee was observed at PC4 and was used as the criterion for selecting PC3 and PC4 for the rest of the analysis. Next, the empirical histograms $\hat{P}_\phi$ and $\hat{Q}_\phi$ in this space — representing the population profiles — were computed by adaptive binning, for which centroids were obtained using using k-means clustering. A total of $n_p = 4$ and $n_q = 5$ centroids were selected for the two groups based on the AIC criterion. The centroids for P and Q are visualized in Table 2 (left). This solution is visualized through the directed graph shown in the figure. In order to establish stability of the result, bootstrap estimates of the edges weights were estimated over 1000 iterations. It is observed that phenotypes P_4, small elongated nuclei were transformed to Q_3, significantly larger nuclei and to Q_4, smaller and rounded ones, while a certain amount of them remained relatively unaltered Q_5. Further, a significant fraction of rounded nuclei P_4 acquired a curved morphology Q_1. This model of redistribution thus provides an interpretation of the phenotypic shifts that occurred, giving additional insights into the nature of the genetic perturbation.



5 Discussion

In this paper, a methodology was presented to quantify differences between two cell populations in terms of their phenotypic profiles and an explanatory model by which the differences can be interpreted was provided. Using an example of perturbations induced by the deletion of a tumor–suppressor gene in the mouse model, it was first shown that the phenotypic differences between control and perturbed fibroblast cell populations can be characterized in terms of their differences in distribution in a data–derived phenotype space. The proposed methodology was further used to estimate the putative *redistribution* of nuclear phenotypes. This model of redistribution provides a hypothesis about the nature of changes induced by genetic perturbations in the animal system.

The experiment described was conducted in the context of understanding the role of the tumor microenvironment in breast cancer [11]. A characterization of the changes that take place in fibroblasts — cells that play a major role in this microenvironment — can lead to the generation of new hypothesis about cancer progression. While nuclear morphology was used as a proxy for cell phenotype in this study, the proposed methodology can be applied to any phenotypic characterization of a cell, so long as an appropriate cost model d can be estimated for the same. In future studies, we plan to integrate more morphological features including cell image texture and cellular context for further characterization of the effects of genetic perturbation in the tumor microenvironment.

References

1. Andrey, P., et al.: Statistical Analysis of 3D Images Detects Regular Spatial Distributions of Centromeres and Chromocenters in Animal and Plant Nuclei. PLoS Computational Biology 6(7) (July 2010)
2. Boland, M.V., Murphy, R.F.: A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics 17(12), 1213–1223 (2001)
3. Brechbühler, C.: Parametrization of Closed Surfaces for 3-D Shape Description. Computer Vision and Image Understanding 61(2), 154–170 (1995)
4. 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 231(Pt. 1), 105–114 (2008)
5. Keren, K., Pincus, Z., Allen, G.M., Barnhart, E.L., Marriott, G., Mogilner, A., Theriot, J.A.: Mechanism of shape determination in motile cells. Nature 453(7194), 475–480 (2008)
6. Lin, G., Chawla, M.K., Olson, K., Barnes, C.A., Guzowski, J.F., Bjornsson, C., Shain, W., Roysam, B.: A multi-model approach to simultaneous segmentation and classification of heterogeneous populations of cell nuclei in 3D confocal microscope images. Cytometry Part A 71(9), 724–736 (2007)
7. Pincus, Z., Theriot, J.A.: Comparison of quantitative methods for cell-shape analysis. Journal of Microscopy 227(Pt. 2), 140–156 (2007)
8. Rittscher, J.: Characterization of Biological Processes through Automated Image Analysis. Annual Review of Biomedical Engineering 12, 315–344 (2010)
9. Rohde, G.K., Ribeiro, A.J.S., Dahl, K.N., Murphy, R.F.: Deformation-based nuclear morphometry: capturing nuclear shape variation in HeLa cells. Cytometry Part A 73(4), 341–350 (2008)

10. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Movers Distance as a Metric for Image Retrieval. International Journal of Computer Vision 40(2), 99–121 (2000)
11. Trimboli, A.J., Fukino, K., de Bruin, A., Wei, G., Shen, L., Tanner, S.M., Creasap, N., Rosol, T.J., Robinson, M.L., Eng, C., Ostrowski, M.C., Leone, G.: Direct evidence for epithelial-mesenchymal transitions in breast cancer. Cancer Research 68(3), 937–945 (2008)
12. Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage 31(3), 1116–1128 (2006)
13. Zink, D., Fischer, A.H., Nickerson, J.A.: Nuclear structure in cancer cells. Nature reviews. Cancer 4(9), 677–687 (2004)