Large-scale learning of cellular phenotypes from images

Vebjorn Ljosa[1], Piyush B. Gupta[1-2], Thouis R. Jones[1], Eric S. Lander[1–4], Anne E. Carpenter[1]

[1]Broad Institute of MIT and Harvard; [2]Dept. of Biology, MIT; [3]Whitehead Institute for Biomedical Research; [4]Dept. of Systems Biology, Harvard Medical School

Microscopy-based high-throughput screens can provide a broad view of biological responses and states at the resolution of single cells. Thousands of samples of cultured cells are perturbed by different chemicals or RNAi reagents. The samples are then stained and imaged, and samples that exhibit a phenotype of interest are chosen for further investigation. Some phenotypes are readily identifiable in captured image data; for instance, mitotic arrest can be detected by measuring the intensity of a fluorescent marker for mitosis. Other phenotypes, while apparent upon visual inspection, are much harder to identify computationally. As an example, when signaling pathways related to cell migration are stimulated, T47D breast cancer cells take on a motile appearance, but this phenotype is not easily captured in a sparse set of measurements. Classifiers trained on hand-curated training sets can identify such phenotypes [1, 2]. We present a method that can learn to recognize phenotypes without requiring hand-labeled cells for training. Instead, a classifier is learned from larger portions of the experiment known to be enriched (if only slightly) by the phenotype of interest. As an example, we use an RNAi screen of T47D breast cancer cells [2]. The screen was performed in duplicate, and the second replicate was treated with a protein stimulant of cellular migration. As a result, a migratory phenotype putatively related to metastasis was slightly more prevalent in the stimulated replicate (~55% vs. ~45%). Such noisy training sets are unsuitable for most machine learning methods, but large-scale machine learning [3] allows us to overcome the noise by using huge training sets (in our case, the millions of cells found in each replicate). As a result, a classifier specific for the response of cells to the stimulant can be created without manual classification of cells (Figure 1).Our goal is not to classify individual cells, but to decide whether each *sample* is enriched for a phenotype. Because the number of cells per sample varies greatly, computing the fraction of motile-looking cells is insufficient to estimate the underlying probability of motility. We therefore use the empirical distribution of classifier scores (Figure 1) to give each cell a probabilistic (i.e., soft) label. We can then compute probability density functions of the proportion of motile-looking cells per sample and derive enrichment scores.

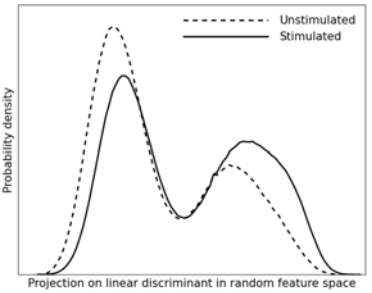

Figure 1: Histogram of per-cell classifier scores for the unstimulated and stimulated replicates, showing the slight shift of cells from a nonmigratory to a migratory phenotype upon stimulation. This histogram is the basis for our nonparametric scoring method.

[1] T.R. Jones et al. (2009) "Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning," *Proc. Acad. Sci. USA*, **106**:1826–1831.

[2] P.B. Gupta et al., "Identification of novel effectors of ErbB2/3-mediated cell migration with high-throughput image-based screening," submitted.

[3] A. Rahimi and B. Recht (2008) "Random features for large-scale kernel machines," *Advances in Information Processing Systems (NIPS)*, **20**:1177–1184.