











Automatic quantification of subtle cellular phenotypes in microscopybased high-throughput experiments

> Vebjorn Ljosa Broad Institute of MIT and Harvard Cambridge, MA, USA

With other members of the Imaging Platform and numerous collaborators

Fourth International Workshop on Machine Learning in Systems Biology (MLSB) Edinburgh, Scotland 2010-10-16







van Leeuwenhoek's microscope (late 1600s)



Image data are exciting.

You should join us in working on them.



Outline

- I. The Imaging Platform at the Broad Institute
- 2. CellProfiler image-analysis software
- 3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype
- 4. Large-scale training of a classifier for subtle phenotype changes
- 5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition
- 6. Discovering latent "phenotypes" by learning to scale image features using linear regression and a topic model





The Broad Institute's unusual organization



Faculty member + lab



Imaging platform



Chemical biology platform



Faculty member + lab



Sequencing platform

Faculty member + lab



Faculty member + lab

[phdcomics.com]

The Imaging Platform at the Broad Institute



Anne Carpenter Director

Methods development



Vebjorn Ljosa



Carolina Wählby Image assay development







David Logan

Mark

Bray



Software engineering



Lee Kamentsky

> Adam Fraser

> > 7

Outline

- I. The Imaging Platform at the Broad Institute
- 2. CellProfiler image-analysis software
- 3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype
- 4. Large-scale training of a classifier for subtle phenotype changes
- 5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition
- 6. Discovering latent "phenotypes" by learning to scale image features using linear regression and a topic model



CellProfiler—why was new software needed?

Assay-specific pre-packaged commercial software

Fast

Poor results on crowded cells or unusual cell types

Designed for standard assays

Cell adhesion Neurite outgrowth Micronucleus formation Protein translocation Cell cycle analysis Adipogenesis Reporter gene analysis Cell viability Apoptosis Cell migration Often inflexible

Proprietary methods

Expensive



IMAGING Platform







Published raw source code

Advanced algorithms

Heavily customized, not generalizable

Requires programming skills + a lot of time to adapt to new situations

Rarely applied outside the originating lab

function [rgOut, varargout] = ImDAPI2Rg(imDAPIin, LoGDim, LoGHW, MinArea)

wiendim=[5 5];

rgLoG=fspecial('log',LoGDim,LoGHW); imLoGout=imfilter(double(imDAPlin),rgLoG); imLoGoutW=wiener2(imLoGout,wiendim); rgNegCurve=imLoGoutW<-1;

%set outsides rgNegCurve([1 end],1:end)=1; rgNegCurve(1:end,[1 end])=1;

%Throw out noise, label regions rgArOpen=bwareaopen(rgNegCurve,MinArea,4);



Typical CellProfiler pipeline



Measure everything first, ask questions later



"Cytological profile": collection of measurements describing the appearance of a cell



Data plot: Noa Shefi 12

Successful image-based assays

Phenotypes:

Cell count Cell size DNA content Nuclear speckles Cytoplasm/nucleus localization Membrane localization Protein or phospho-protein levels Metastasis Wound healing Metaphase Anaphase/telophase Prophase Shape/texture Crescent-shaped nuclei Peas-in-a-pod Cells-on-the-move Long projections Crooked projections Hyphae-like fingers Actin at contractile ring/cell junctions Internal actin Actin circles Large spread cells Phospho-histone H3 nuclear dots **Bi/multinucleate**

Cell types:

Drosophila Kc167 cells Drosophila S2R+ cells Drosophila epithelial tissue Drosophila embryo Human HT29 cells Human A549 cells Human TOV21G cells Human H1299 lung carcinoma cells Human biopsied prostate gland tissue Human adult mesenchymal stem cells Mouse NIH/3T3 cells Mouse neural precursor cells derived from embryos Mouse lung tissue sections Mouse isolated germ cells Rat H9c2 cells C. elegans worms (preliminary) Neurons (preliminary) S. cerevisiae cells Yeast colonies Yeast growth patches Array grids



www.cellprofiler.org



[Download] (0.9 MB)

CellProfiler around the world



CellProfiler is downloaded 400x/month, 11,000x total (~50% USA, 90% non-profit institutions).



The CellProfiler project

free, at www.CellProfiler.org



CellProfiler's is the 5th most-accessed Genome Biology paper of all time



Experiments we have completed recently

Biological process/ phenotype	Laboratory	Samples tested	Number of fields of view (images) processed	
Meiosis	Terry Orr-Weaver (Whitehead)	RNAi	84,000	
Mitochondria	Vamsi Mootha (HMS/MGH)	chemicals	100,000	
Morphology	AstraZeneca	chemicals	109,200	
Cell cycle	AstraZeneca	chemicals	109,200	
Breast cancer/Heregulin	Eric Lander (Broad)	RNAi	144,798	
Tuberculosis	Deb Hung (Broad)	chemicals	164,000	
Glioma	David Sabatini & Bill Hahn (Whitehead, Harvard, Dana-Farber)	RNAi	286,000	
Polyploidy: AMKL	John Crispino (Northwestern University)	chemicals, some RNAi	530,000	
Hematopoetic stem cells	David Scadden and Stuart Schreiber (HMS, MGH, Broad)	chemicals, some RNAi	465,448	
Leukemic stem cells	Gary Gilliland (BWH/HMS)	chemicals, some RNAi	1,040,098	
Hepatotoxicity	Sangeeta Bhatia (MIT)	chemicals	1,135,093	

Outline

- I. The Imaging Platform at the Broad Institute
- 2. CellProfiler image-analysis software
- 3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype
- 4. Large-scale training of a classifier for subtle phenotype changes
- 5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition
- 6. Discovering latent "phenotypes" by learning to scale image features using linear regression and a topic model



Screens



Thousands of samples

Add thousands of chemicals or RNAi agents, each one in a different sample





Phosphorylated histone H3 – Mean





DAPI and mean pHH3





"Simple" phenotypes: one or two features is enough



X-axis: DNA content



Example of complex phenotype: motile T47D cells

Normal T47D cells

Features associated with cell motility: lamellipodia, filopodia, polarized cell shape, F-actin nucleation at filapodia, less clumping





Challenges

- How to get and crossvalidate with rare phenotypes?
- How to make classifier interpretable by biologist?
- Normalization of features?
- Dimensionality reduction?
- Prevent overfitting?
- Avoid having to tune parameters?



Rare phenotypes: HT29 colon cancer cells



[Jones et al., PNAS, 2009]

Iterative machine learning



Incorporated into CellProfiler Analyst

000	pyClassifier	
Fetch cells		
	Fetch 20 random cells from experiment Fetch!	
Train Classifier		
IF (Cells_NumberNeighbors_P IF (Cells_Intensity_CorrGreen) IF (Cells_NumberNeighbors_P	ntTouching > 57.870399, [-0.65022492, 0.65022492], [0.56773669, -0.56773669]) kle_StdIntensity > 0.032108199, [0.73954767, -0.73954767], [-0.34105974, 0.34105974]) ntTouching > 30.8176, [-0.2280075, 0.2280075], [0.99999994, -0.99999994]) Max number of rules: 20 (Find Rules) (Score All) (Score Image)	
Free Head (20)		ge
unclassified (20)		
	*	
-1 (92)	+1 (78)	
		18 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
ROAD IMAGING	CellProfiler Analyst data exploration software	

Rules for distinguishing HRG-stimulated T47D cells

 $(IF MeanSpeckles_AreaShape_Area > 12.00000, 0.827550, -0.350258) + \\(IF Cells_AreaShape_FormFactor > 0.449767, -0.331746, 0.706321) + \\(IF CellMembrane_Texture_3_CorrGreen_DifferenceVariance > 0.718124, 0.593955, -0.198424) + \\(IF Cells_Intensity_CorrGreenSpeckle_MaxIntensity > 0.370382, 0.787301, -0.204062) + \\(IF CellMembrane_Intensity_CorrGreenSpeckle_MinIntensityEdge > 0.001284, 0.275866, -0.500179) + \\(IF Cells_Texture_3_CorrGreen_SumEntropy > 1.710600, -0.199515, 0.700658) + \\(IF Nuclei_AreaShape_Perimeter > 96.669000, -0.177882, 0.788582) + \\(IF Cells_Texture_3_CorrGreen_GaborY > 0.147318, -0.215618, 0.613682) + \\(IF Cells_Intensity_CorrGreenSpeckle_StdIntensity > 0.033557, 0.742323, -0.169222) + \\(IF Cells_Intensity_CorrGreenSpeckle_StdIntensityEdge > 0.009328, -0.118830, 0.956272)$



Multiple classes

Classifier 2.0 - C:\Trunk\CPAnalyst\properties\2009_02_1	9_MijungKwon_Centrosomes.pr	operties		
File CY3 CY5 Blue FITC Display Help				
-Fetch cells Fetch	20 random 💌 cells from experim	ient 💽 Fe	stch!	
Train Classifier				
<pre>IF (Spindle_Texture_GaborY_RescCY3_10 > 18.502899, [-1.0000011, -0.883 IF (MitoNud_Texture_DifferenceVariance_RescCY3_2 > 0.60490501, [-0.940 0.12799327])</pre>	60709, -0.76721311, -0.80046856, -0.9 38278, 0.20617817, -0.062676564, 0.25	7453994, 0.42582536], [-0.58094478, - 692186, -0.2703512, -0.75822949], [0.	0.61087728, -0.64080977, -0.63225764 31957713, -0.82919544, -0.2336835, -0	, -0.58749253, -0.94761813])).94829345, -0.066388384, -
IF (Spindle_Intensity_StdIntensityEdge_RescCY5 > 0.00881173, [-0.0407383	14, 0.10163302, -0.52159816, -0.27694	175, 0.12328041, 0.018147739], [-1.0,	, -1.0000001, 0.78679597, 0.25150594,	-1.0000001, -1.0000002])
		Maxin	number of rules: 15 Find Rules	Score All Score Image +
unclassified (20)				
			0 3 6	
BipolarMonastral (35) Multipolar (14)	Other (70)	Prophase (49)	Bipolar (64)	Monopolar (16)
fetching 20 random cells from whole experiment				



QC and Tracking Down Hits

	(000														
		Source:				Plate:	Plate: 2002-01-W01-02-01-CN00002412-Brerun 💠									
		Data sour Screen1f Measurem	ce: eb2509run2 ients:	_Per_Image	¢ A	01 02 03	04 05 (06 07 (08 09	10 11	12 13	14 15 :		3 19 20	21 22	23 24
		Image_C	oun 🛊		С											ÖÖ
	(Data Aggreg	gation:		D											
		Aggregati cv%	on method:		F											
	ſ	View Option Color Map	s:):		H											
	En	jet	+ +	ene	J											
dene	Counts	Counts	p(Enriched)	p(Enriched)	Enriche	d Score										
gene	positive	negative	positive	negative	pos	itive										
NME1	179	2248	0.96984534	0.03015463	1.507348	0914										
PRPS2	140	2002	0.91686196	0.08313715	1.042504	19841										
TK1	146	2358	0.84006921	0.15993923	0.720383	014599										
PMS1	86	1356	0.83575721	0.16424677	0.706593	839064										
GALK1	119	1938	0.82593326	0.17408002	0.676229	179724				2	1			1		
MAPK13	148	2908	0.64246398	0.35755690	0.254528	98308					.215		31.3	36		38.456
Gabra2	112	2208	0.63344543	0.36661946	0.237570	577506					ð	_	_	_		
PHKG2	72	1431	0.61455659	0.38561178	0.202601	261265				//	ð					
MAP2K3	181	3784	0.56842729	0.43162143	0.119620	956956				//	0					
STK19	105	2377	0.46921976	0.53066779	-0.05353	84459292					0					
Gabra1	92	2086	0.46803014	0.53249026	-0.05561	31897801					0					
MAP2K6	91	2169	0.41092623	0.58938132	-0.15640	582072					0					
Gpr12	79	1978	0.35990666	0.64030450	-0.25005	34158										
PDXK	112	2829	0.34267214	0.65696491	-0.28290	3238922										
MAPK11	48	1360	0.26241467	0.73790298	-0.44882	4138167										
Gabra3	84	2504	0.19606696	0.80500784	-0.61281	54381										

Outline

- I. The Imaging Platform at the Broad Institute
- 2. CellProfiler image-analysis software
- 3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype
- 4. Large-scale training of a classifier for subtle phenotype changes
- 5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition
- 6. Discovering latent "phenotypes" by learning to scale image features using linear regression and a topic model



The phenotype of motile T47D cells

Normal T47D cells

Features associated with cell motility: lamellipodia, filopodia, polarized cell shape, F-actin nucleation at filopodia, less clumping



Unstimulated

Stimulated by heregulin



Built training set of ~300 cells

000	pyClassifier
Fetch cells	
	Fetch 20 random 🛟 cells from experiment 🛟 (Fetch!)
Train Classifier	
IF (Cells_NumberNeighbor IF (Cells_Intensity_CorrGre IF (Cells_NumberNeighbor	ercentTouching > 57.870399, [-0.65022492, 0.65022492], [0.56773669, -0.56773669]) Speckle_StdIntensity > 0.032108199, [0.73954767, -0.73954767], [-0.34105974, 0.34105974]) ercentTouching > 30.8176, [-0.2280075, 0.2280075], [0.99999994, -0.99999994]) Max number of rules: 20 Find Rules Score All Score Image +
unclassified (20)	
-1 (92)	+1 (78)
ROAD IMAGIN	CellProfiler Analyst data exploration software

Why cut out the human?



Labeling for automatic training set



Unstimulated

Stimulated by heregulin

45 % motile cells

55 % motile cells



35

Two ways to improve the classifier



Training set size



Random Fourier features

$$\cos(\omega'\mathbf{x}+b)$$



Kernel Name	$k(\Delta)$	$p(\omega)$
Gaussian Laplacian	$e^{-rac{\ \Delta\ _2^2}{2}} e^{-\ \Delta\ _1}$	$(2\pi)^{-rac{D}{2}}e^{-rac{\ \omega\ _2^2}{2}} \ \Pi_d rac{1}{\pi(1+\omega^2)}$
Cauchy	$\prod_d \frac{2}{1+\Delta_d^2}$	$e^{-\ \Delta\ _1}$



[Rahimi and Recht, NIPS, 2007] $_{\rm _{37}}$

Random features



Linear discriminant on random features

7.6 million training cells,130 measurements

Mapped into 250dimensional random feature space

Trained Fisher's linear discriminant





Automatic vs. hand training





Outline

- I. The Imaging Platform at the Broad Institute
- 2. CellProfiler image-analysis software
- 3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype
- 4. Large-scale training of a classifier for subtle phenotype changes
- 5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition
- 6. Discovering latent "phenotypes" by learning to scale image features using linear regression and a topic model



Expected likelihood kernel

$$k(A,B) = \int \Pr(x|A) \,\Pr(x|B) \,\mathrm{d}x$$



[Jebara et al., 2004]

z(x) projection of x into random-feature space

$$\langle z(a), z(b) \rangle \approx k(a, b)$$

$$v_A = \frac{1}{|A|} \sum_{x \in A} z(x)$$

$$\approx \int \Pr(x|A) \Pr(x|B) dx$$





Correlation coefficient of dot products



Outline

- I. The Imaging Platform at the Broad Institute
- 2. CellProfiler image-analysis software
- 3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype
- 4. Large-scale training of a classifier for subtle phenotype changes
- 5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition
- 6. Discovering latent "phenotypes" by learning to scale image features using linear regression and a topic model



Histone deacetylases

- 11 enzymes
- Component of chromatin-regulating complexes
- Also target many non-histone proteins
- Broad relevance to cell signaling and cell state
- Induce differentiation and inhibit proliferation in cancer models
- Inhibitors used clinically for cutaneous T-cell lymphoma, others in trials for other cancers



Screening question: find specific HDAC inhibitors

Inhibition potency data from biochemical assays



Training



Samples

 $T = K X^{-1}$





Testing



Cells





Convert scores to soft labels by logistic transform





Proportions of cells for each topic





5 I

Correlations between classes match HDAC phylogeny





Summary

- I. The Imaging Platform at the Broad Institute
- 2. CellProfiler image-analysis software
- 3. Iterative training of a boosting classifier for a particular (possibly rare) cellular phenotype
- 4. Large-scale training of a classifier for subtle phenotype changes
- 5. Comparing heterogeneous populations of cells perturbed by small molecules or RNA inhibition
- 6. Discovering latent "phenotypes" by learning to scale image features using linear regression and a topic model



Thank you!

Imaging Platform and alumni:

Mark Bray Anne Carpenter Adam Fraser Thouis (Ray) Jones Lee Kamentsky David Logan Kate Madden Tejas Shah Carolina Wählby

Collaborators:

Peter Caie Neil Carragher Paul Clemons Emma Cooke Christopher Denz Piyush Gupta Sigrun Gustafsdottir Tom Houslay Melissa Kemp Angela Koehler Mijung Kwon Melissa Passino Eric Lander Aly Shamji

