Active Microscopic Cellular Image Annotation by Superposable Graph Transduction with Imbalanced Labels

Jun Wang and Shih-Fu Chang *Department of Electrical Engineering, Columbia University, USA*

Xiaobo Zhou and Stephen T. C. Wong *The Methodist Hospital Research Institute, Cornell University, USA*

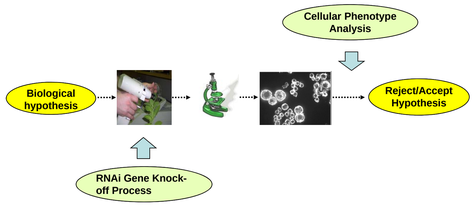
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

The Methodist Hospital System

Introduction

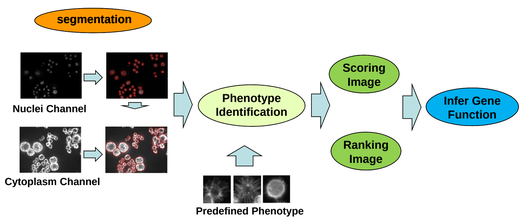
Scientific Motivation

- Cellular image analysis provides useful cues for gene function expression
- RANi is a powerful technique for manipulating effects of individual genes (gene knock-off), producing distinctive cell phenotypes
- Phenotype classification and filtering helps understanding roles of individual genes in biological processes
- Each genome-wide study produces a large data set (22,000 human genes, > 100,000 images)
- Require efficient analysis tools and systems
  → critical for RNAi Genome-wide High-Content Screening (HCS)



State of the Art of RANi Image Analysis

- CellProfiler allows for quality enhancement, cell segmentation, measurement, simple classification of predefined phenotypes (http://www.cellprofiler.org/)
- Our Prior Work:
  Wang, Chang, et al. Journal of Biomolecular Screening, 2008
  Wang, Chang, et al. Journal of Biomedical Informatics, 2008):
  - ➢ *Supervised phenotype classification*
  - ➢ *Biologists define phenotypes of interest, help annotate training samples*
  - ➢ *Computer extracts features and learns classifiers*



Our prior work of supervised phenotype classification and HCS analysis.

Objectives and Novel Contribution

- Provide flexibility for biologists to quickly define new phenotypes (Scalability).
- Achieve this via interactive annotation and relevance feedback (User in the Loop)
- Use transductive graph to propagate cell phenotype labels and rank image scores

Proposed Approach

Active Annotation via Graph Transduction

- Construct a graph with cells as nodes and cell similarity as edge weights (W), degree node (D)
- Cell phenotypes as multiple binary class Labels (Y)
  e.g., [1 0] as class 1, [1 0] as class 2, [0 0] as unlabeled
- Biologists Interact with the system to provide phenotype labels, but majority of cells are unlabeled → semi-supervised
- Machine predicts continuous-valued classification function (F)
- Goal: Propagate labels to the unlabeled cells via optimizing a objective functions (Q) including both "Smoothness" and "Fitness" components

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} \mathcal{Q}(\mathbf{F}) = \arg\min_{\mathbf{F}} \left[Q_{smooth}(\mathbf{F}) + Q_{fit}(\mathbf{F})\right]$$

- Different formulation of the loss function;
  - ➢Local and global consistency (Zhou et al NIPS 2004);

$$\mathcal{Q}(F) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left\|\frac{F_i}{\sqrt{D_{ii}}} - \frac{F_j}{\sqrt{D_{ii}}}\right\|^2 + \mu\sum_{i=1}^{n}\|F_i - Y_i\|^2$$

  - ➢Gaussian Field and Harmonic functions (Zhu et al ICML 2003);

$$\mathcal{Q}(F) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\|F_i - F_j\|^2$$

1) $\triangle F = 0$ on unlabeled data, where $\triangle = D - W$ is the graph Laplacian;
2) $F_i = Y_i$ on labeled data.

Problems with Previous Approaches

- Label imbalance: # of labeled data highly uneven among classes
  - ➢ *Because cells in the same image shown to users tend to belong to the same class*
  - ➢ *We propose a new node regulizer to overcome the imbalance issue*
- Need Active and incremental Learning. How to incorporate new labels obtained from user interaction without redoing the entire graph propagation?
  - ➢ *Apply superposition rule on graph transduction to efficient label propagation*

Idea 1: Superposition on Graph

- The classification function obtained by graph propagation using the labeled sample set equals the sum of a functional set, where each element is contributed by an individual labeled sample (x_i).

$$F = \beta(I-\alpha S)^{-1}\sum_{i=1}^{l}\hat{Y}_i = \sum_{i=1}^{l}\beta(I-\alpha S)^{-1}\hat{Y}_i = \sum_{i=1}^{l}\hat{F}_i$$

- Superposition among classes;

$$F = \sum_{j=1}^{c}\sum_{y_i=j}\beta(I-\alpha S)^{-1}\hat{Y}_i = \sum_{j=1}^{c}\sum_{y_i=j}\hat{F}_i \qquad \mathbf{Y} = \begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix} = \begin{bmatrix}1&0\\0&0\\0&0\end{bmatrix} + \begin{bmatrix}0&0\\0&1\\0&0\end{bmatrix}$$

($\hat{Y}_1$, $\hat{Y}_2$)

Idea 2: Node Regularizer

- Node regularizer is the class-normalized weight matrix by given labels
  - ➢ *Sum of node regularizer for each class equals 1*
  - ➢ *Higher degree nodes have higher value of node regularizer*

Hadamard product

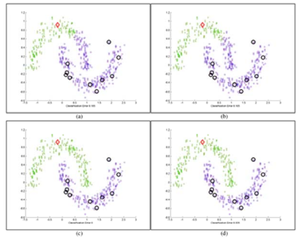
$$F = \sum_{i=1}^{l} v_{ii}\hat{F}_i = \sum_{i=1}^{l}\beta(I-\alpha S)^{-1}v_{ii}\hat{Y}_i = \beta(I-\alpha S)^{-1}VY \qquad \mathbf{v} = \sum_{j=1}^{c}\frac{\mathbf{Y}_{\cdot j}\odot \mathbf{D}\vec{1}}{\mathbf{Y}_{\cdot j}^T\mathbf{D}\vec{1}}$$

$$\mathbf{D} = \begin{bmatrix}1&0&0&0\\0&2&0&0\\0&0&3&0\\0&0&0&4\end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix}1&0\\0&1\\1&0\\0&0\end{bmatrix} \longrightarrow \mathbf{v} = \begin{bmatrix}\frac{1}{1+3}&0&0&0\\0&\frac{2}{2}&0&0\\0&0&\frac{3}{1+3}&0\\0&0&0&\frac{0}{4}\end{bmatrix}$$

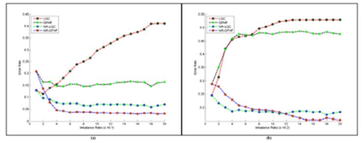
- Incorporate with prior class knowledge;

Node Regularizer Experimental Validation

- Toy data of Two-Moon set



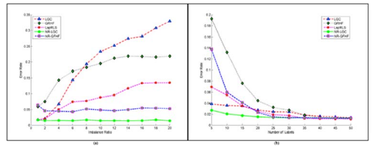
The proposed node regularizer method (NR-LGC and NR-GFHF) successful overcome the label imbalance problem.



The proposed node regularizer methods (NR-LGC and NR-GFHF) are effective with different graph construction (kernel size) and ratio of imbalanced labels.

- Handwritten Digits

Figure The examples of USPS handwritten digitals.

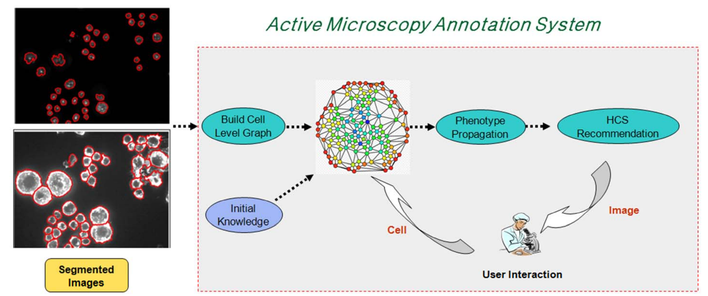


Vary the ratio of # of labeled samples among classes from 1 to 20.

Random sampling of data from different classes.

Overall Framework

System Diagram for Cellular Image Annotation



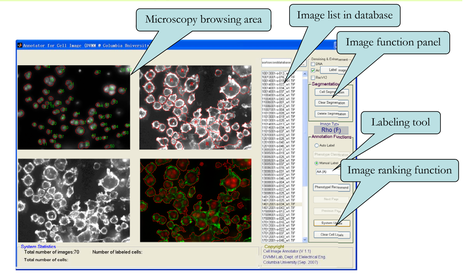
Active Annotation with Continuous User Input

- Initial cell annotation, obtain prediction function F.
- Fuse cell scores to obtain image ranks
- System displays top images for phenotype of interest (e.g., class k)
- User select and label one new sample (x_s) of class k
- Calculate new node regularizer
- Incrementally update classification function
- Update is very efficient since only scores of class k need to be computed

$$F_{\cdot k}^{new} = \lambda F_{\cdot k} + \gamma \hat{F}_{\cdot s} \quad \lambda = \frac{D^k}{D^k + d_{ss}} \quad \gamma = \frac{d_{ss}}{D^k + d_{ss}}$$

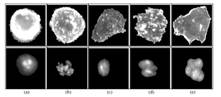
Evaluation

GUI of Microscopy Active Annotation System



Cell Phenotypes

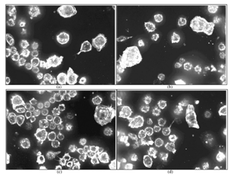
Sample cellular phenotypes identified by biologists (with distinct appearance and geometry features)



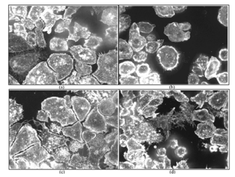
The top row is cytoplasm channel and the bottom row is nuclei channel.

Cell Phenotype	*Appearance Description*
Actin Accumulation (*AA*)	actin accumulation in the cell body, bright intensity, may have non-round nuclei;
Cell Cycle Arrest (*Cyc4-ori*)	large size, round cells with multi-nuclei;
Longthin LPA (*LL*)	resulted long punctuate actin, with cell shape as prolonged water drop or long thin poles shape;
LS-Fla (*LF*)	cells with large spiky and filamentous structure;
Rho	large and flat shape, with multi-nuclei, non-round.

Rank Microscopic Images

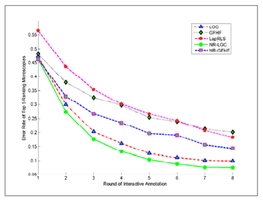


Active annotation results on the top four ranked images under the query of *AA* cellular phenotype.



Active annotation results on the top four ranked images under the query of *Rho* cellular phenotype.

Statistics of Annotation Performance



The performance evaluation. X coordinate coordinate denotes the interaction rounds and Y coordinate denotes the precision of top 5 ranked microscopy images.

➢210 images (multiple channels), 3162cells, 5 phenotypes, roughly even phenotype size;
➢Simulate active annotation scenario
➢Add 10 randomly chosen ground truth cells in each iteration;
➢System update the prediction function
➢Evaluate the precision of the top ranking images after each iteration

Method	LGC	GFHF	LapRLS	NR-LGC	NR-GFHF
Computation Cost (sec.)	0.81	70.05	218.9	0.14	70.28

Table 2. Computation cost of active annotation (8 rounds) on the microscopic cellular images.

Incremental graph update reduce the LGC process time by 16 times. But it did not improve the GFHF method.

Acknowledgments

Norbert Perrimon, Chris Bakal, Wei Liu, and Zheng Yin for technical discussion and assistance in data set acquisition.