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

Jun Wang and Shih-Fu Chang
Department of Electrical Engineering
Columbia University
{jwang, sfchang}@ee.columbia.edu

Xiaobo Zhou and Stephen T. C. Wong
The Methodist Hospital Research Institute
Cornell University
{XZhou, STWong}@tmhs.org

Abstract

Systematic content screening of cell phenotypes in microscopic images has been shown promising in gene function understanding and drug design. However, manual annotation of cells and images in genome-wide studies is cost prohibitive. In this paper, we propose a highly efficient active annotation framework, in which a small amount of expert input is leveraged to rapidly and effectively infer the labels over the remaining unlabeled data. We formulate this as a graph based transductive learning problem and develop a novel method for label propagation. Specifically, a label regularizer method is proposed to handle the important label imbalance issue, typically seen in the cellular image screening applications. We also design a new scheme which breaks the graph into linear superposition of contributions from individual labeled samples. We take advantage of such a superposable representation to achieve fast annotation in an interactive setting. Extensive evaluations over toy data and realistic cellular images confirm the superiority of the proposed method over existing alternatives.

1. Introduction

Cellular Microscopic Screening: Gene function can be assessed by analyzing disruptive effects on a biological process caused by the absence or disruption of genes. With recent advances in fluorescence microscopy imaging and gene interference techniques like RNA interference (RNAi), genome-wide high-content screening (HCS) has emerged as a powerful approach to systematically study the functions of each individual gene. These microscopic screenings generate a large number of biological readouts, including cell size, cell viability, cell cycle, and cell morphology. A typical HCS cellular image usually contains a population of cells shown in multi-channel signals, such as DNA channel (indicating locations of nuclei) and F-actin channel (indicating information of cytoplasm) (Fig. 1).

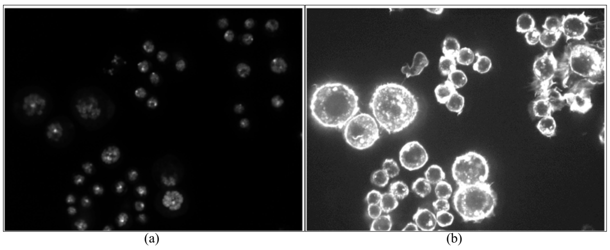


Figure 1. Typical microscopic images of Drosophila K_{c167} embryonic cells. (a) image of the DNA channel; (b) image of the F-actin channel after homomorphic enhancement.

Recently through manual analysis of fluorescence microscopy images, cellular phenotypes visible in RNAi cell images (e.g., cytoskeletal organization and cell shape) have been found important for HCS study [5]. Specifically, when an individual gene is ”turned off” by the RNAi technology, the resulting changes of the morphological structures of the cells in the images can be used to infer the function of the gene on the biological process under investigation (e.g., drug design, disease mechanism). However, a critical barrier preventing successful deployment of large-scale genome-wide HCS is the lack of efficient and robust methods for automating phenotype classification and quantitative evaluation of the rapidly increasing collection of HCS images

Interactive Microscopy Annotation: One important task in HCS is to rapidly retrieve the most relevant cellular images from the database given a certain cell phenotype of interest specified by biologists. Currently this is handled in a manual way - biologists first examine a few example images showing the phenotype of interest, and then manually browse through individual microscopic images, and assess the relevance of each image to the cellular phenotypes. Apparently, this manual procedure is very expensive and relies on well trained domain experts. Recently, a supervised learning manner based cellular phenotype identification system was developed [7]. However, it still replies much on the exhausted expert input.

In this paper, we propose an efficient interactive annotation framework for RNAi microscopic cellular images. Starting with the expert labeling of a few cells according to some predefined phenotypes, the system learns to infer the phenotype classes of unlabeled cells on the microscopic images. The learning is done in a semi-supervised manner that both the labeled and unlabeled data are utilized. Given the predicted phenotype label for the cells, image-level relevance scores are also computed. Then the system recommends the most relevant cell images to the biologist who will review the results and make further cell-level annotation. This interactive procedure is repeated until a sufficient number of relevant images are retrieved or no additional positive images can be found.

The objective of the proposed interactive system is to drastically improve the throughput of finding relevant images from a large RANi cellular image collection. The underlying technical goal is to develop a novel graph transductive learning approach that can execute accurate cell phenotype prediction, and also work in an incremental manner to handle new cell labels obtained from the interactive annotation procedure. Note the proposed system is different from the regular relevance feedback or active learning systems for image retrieval. Here the annotation is done at the cell level, while relevance scoring and recommendation are conducted at the image level.

Motivation: A major challenge in developing effective solutions for the aforementioned applications is a robust cell phenotype prediction method that we may use to recommend relevant images throughout the process. To meet this objective, we propose an efficient learning method that leverages the power of graph transduction. There have been some promising graph based transductive learning approaches proposed recently, such as local and global consistency (LGC) [10], and the method based on Gaussian fields and harmonic functions (GFHF) [13]. However, there are two major problems in applying such techniques to the cellular image annotation task. First, the manual cell labeling on microscopic images easily generates imbalanced labels since the browsed HCS is usually bias to a certain phenotype. In such situations, existing methods like LGC and GFHF tend to fail, as illustrated in a toy example in Fig. 2 (a) (b). Second, the interactive annotation system needs respond fast to the incremental labels to fulfil the realistic annotation application. To solve these problems, we propose a novel graph propagation technique with label regularizer to handle the imbalanced label issue, and a new superposable graph propagation approach to achieve the incremental learning in terms of new labels. Through extensive experiments over synthetic data and realistic RANi cellular images, we demonstrate the proposed techniques can improve annotation accuracy while improving the speed at the same time.

The remainder of this paper is organized as follows. In Section 2, we briefly review two existing graph transductive learning approaches, LGC and GFHF, then propose the new approach of superposable graph transduction with label regularizer. Section 3 shows the experimental evaluation of label regularizer method. Section 4 reports the experimental results of interactive annotation results on real microscopic images. Concluding remarks and discussion are given in Section 5.

2. Methodology

First we describe the notation used in the paper. Given the dataset as $\mathcal{X} = (\mathcal{X}_l, \mathcal{X}_u) = \{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_n\}$ and the labels of a small portion of the data $\{y_1, \cdots, y_l\}$, where $y_i \in \mathcal{L} = \{1, \cdots, c\}$. The objective is to infer the labels $\{\mathbf{y}_{l+1}, \cdots, \mathbf{y}_n\}$ of the unlabeled data $\{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_n\}$. The graph is represented as $\mathcal{G} = \{\mathcal{X}, E\}$, where $\mathcal{X} = \{\mathbf{x}_i\}$ and $E = \{e_{ij}\}$. The sample $\mathbf{x}_i$ is treated as the node on the graph and the weight of edge e_{ij} is w_{ij}. So the weight matrix is denoted as $W = \{w_{ij}\}$ and the node degree matrix $D = diag(d_{ii})$ is defined as $d_{ii} = \sum_{j=1}^{n} w_{ij}$, where d_{ii} is degree of node $\mathbf{x}_i$. The label matrix Y is described as $Y \in \mathcal{R}^{n \times c}$ with $Y_{ij} = 1$ if $\mathbf{x}_i$ is with label $y_i = j$ and $Y_{ij} = 0$ otherwise. Moreover, the ith row and jth column vectors are denoted as $Y_{i\cdot}$ and $Y_{\cdot j}$, respectively.

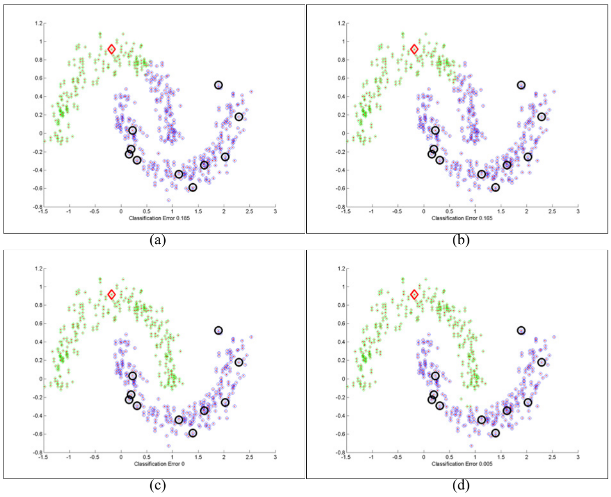
2.1. Brief survey on graph transductive learning

Graph based semi-supervised methods commonly treat the samples (labeled and unlabeled) as the nodes on a graph and the edge as the affinity evaluation between nodes. A classification function $F \in \mathcal{R}^{n \times c}$ is estimated on the graph to minimize a predefined loss function $\mathcal{Q}(F)$, which usually reflects the global smoothness and the local fitting on labeled nodes. Since the mincuts method proposed by Blum and Chawla [3], there are a lot of related work has been done in the past a few years. Here we briefly summarized two emerging graph transductive learning approaches, Gaussian fields and harmonic functions (*GFHF*) and local and global consistency (*LGC*) . A more detailed survey paper can be found in [12].

Gaussian Fields and Harmonic Functions (*GFHF*) [13]: In this approach, the Gaussian random fields is viewed as the quadratic loss function with infinity weight to lock the labeled nodes by the given labels. The graph regularizer based loss function is defined as:

$$\mathcal{Q}(F) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \|F_{i\cdot} - F_{j\cdot}\|^2 + M \sum_{i=1}^{l} \|F_{i\cdot} - Y_{i\cdot}\|^2 \tag{1}$$

The row vector of $F_{i\cdot} \in \mathcal{R}^c$ is the function value at the node $\mathbf{x}_i$, which reflects the likelihood of this node belongs



(a) (b) (c) (d)

Figure 2. The demonstration of imbalance labels issue on the widely used two-moon toy data. The large markers denote the labeled samples and the small color markers show the classification results. (a) *LGC* with error rate 0.185; (b) *GFHF* with error rate 0.165; (c) *LR-LGC* with error rate 0.00; (d) *LR-GFHF* with error rate 0.005.

to different class. The coefficient $M \longrightarrow \infty$ is used to clamp the given labels. Therefore, in order to minimize the above graph cost function, we force $F_{i\cdot} = Y_{i\cdot}$ for $\mathbf{x}_i \in \mathcal{X}_l$.

The optimal $F^\star = \arg\min_F \mathcal{Q}(F)$ is a harmonic function, which satisfies two conditions:

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

2) $F_{i\cdot} = Y_{i\cdot}$ on labeled data.

The above optimization can be obtained by solving the harmonic function with closed form of matrix manipulations. The weight matrix W is re-permutated as labeled and unlabeled sets:

$$W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix} \tag{2}$$

Correspondingly, let $F = [F_l \ F_u]'$ and $D = diag(D_{ll}, D_{uu})$. From $\triangle F = 0$ on the unlabeled data, the values of classification function on unlabeled nodes are derived as:

$$F_u = (D_{uu} - W_{uu})^{-1} W_{ul} F_l \tag{3}$$

Local and Global Consistency (*LGC*) : Considering local and global consistency, a new elastic regularizer framework is proposed in [10].

$$\mathcal{Q}(F) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \left\| \frac{F_{i\cdot}}{\sqrt{D_{ii}}} - \frac{F_{j\cdot}}{\sqrt{D_{ii}}} \right\|^2 + \mu \sum_{i=1}^{n} \|F_{i\cdot} - Y_{i\cdot}\|^2 \tag{4}$$

Let $S = D^{-1/2} W D^{1/2}$, the above cost function can be approximated in the matrix form as:

$$\mathcal{Q}(F) = \frac{1}{2}\mathrm{tr}\left\{F'F - F'SF + \mu(F-Y)'(F-Y)\right\} \tag{5}$$

The optimization of the above graph regularization can be achieved by calculating the partial derivative.

$$\frac{\partial \mathcal{Q}}{\partial F} = 0 \Longrightarrow F = \beta(I - \alpha S)^{-1} Y \tag{6}$$

where $\alpha = 1/(1+\mu), \beta = \mu/(1+\mu)$. Comparing to *GFHF* approach, *LGC* is more flexible since there is no force term to clamp the given labels. However, this advantage could bring more drawbacks in case of imbalanced labels since the given minority labels can be changed to the majority class after propagation.

2.2. Superposition Law

The label matrix Y can be decomposed to the sum of a series individual sample label mask. For each individual labeled sample $\mathbf{x}_i$, the label mask is defined as $\hat{Y}_i = \{\hat{y}_{ij}\} \in \mathcal{R}^{n \times c}$, where only one nonzero element $\hat{y}_{ij} = 1$ if $y_i = j$. So we can write $Y = \sum_{i=1}^{l} \hat{Y}_i$. Replace Y in Eq. 6 by the the sum of individual label mask, we can get:

$$F = \beta(I - \alpha S)^{-1} \sum_{i=1}^{l} \hat{Y}_i = \sum_{i=1}^{l} \beta(I - \alpha S)^{-1} \hat{Y}_i = \sum_{i=1}^{l} \hat{F}_i \tag{7}$$

where $\hat{F}_i = \beta(I - \alpha S)^{-1} \hat{Y}_i$ is the classification function propagated only by labeled sample $\mathbf{x}_i$. From this equation, we can conclude that the classification function F obtained by graph propagating using the labeled sample set $\mathcal{X}_l = \{\mathbf{x}_1, \cdots, \mathbf{x}_l\}$ equals to the sum of a functional set $\mathcal{F} = \{\hat{F}_1, \cdots, \hat{F}_l\}$, where each element of $\mathcal{F}$ is the classification function propagated from a individual sample in $\mathcal{X}_l$. We call this as superposition law in graph propagation procedure. This principle motivated us that the classification function F can be incrementally updated as to new labeled samples instead of recalculating the propagation from the entire label set. Besides the supposition law on individual labels, it also can be described on each class as:

$$F = \sum_{j=1}^{c}\sum_{y_i=j} \beta(I - \alpha S)^{-1} \hat{Y}_i = \sum_{j=1}^{c}\sum_{y_i=j} \hat{F}_i \tag{8}$$

$\sum_{y_i=j} \hat{F}_i$ denote the propagated component by the labels from class j. Apparently, it only has the jth column vector nonzero, which numerically equals $F_{\cdot j}$.

2.3. Label Regularizer

In the traditional graph regularization formulation such as Eq. 1 and Eq. 4, the weights of the labeled nodes have not been considered. Here, we propose the label regularization term to solve the imbalance labels issue. First, let's see how the imbalance labels problem generated. Without losing any generality, we here analyze two-class case here. Assume the number of labels are $y_{1,\cdots,l_1} = 1, y_{l_1+1,\cdots,l_1+l_2} = -1$, where $l_1 + l_2 = l$. From the class superposition equation 8,

$$F = \sum_{i=1}^{l} \hat{F}_i = \sum_{i=1}^{l_1} \hat{F}_i + \sum_{i=l_1+1}^{l} \hat{F}_i \tag{9}$$

If $l_1 << l_2$ and assume the graph is connected, the derived classification function will have bias to samples from the majority class. In other word, the nodes will mostly be labeled as the majority class. We illustrated this imbalanced labels issue on graph propagation using the widely used two-moon toy data, as shown in Fig. 2. For the two class problem, the numbers of the positive and negative labels are 1 (red large diamond marker) and 10 (black large circle marker), respectively. The propagation results by *LGC* and *GFHF* are shown in (a) and (b). The graph construction follows the approach in [10] with Gaussian kernel size $\delta = 0.1$. Fig. 2 (c) and (d) shows the results by label regularizer approaches, which will be discussed in the below.

Here we propose the label-regularized LGC (*LR-LGC*) approach to handle the imbalanced label problem:

$$\begin{aligned} F &= \sum_{i=1}^{l} v_{ii}\hat{F}_i = \sum_{i=1}^{l} \beta(I-\alpha S)^{-1} v_{ii}\hat{Y}_i \\ &= \beta(I-\alpha S)^{-1} V Y \end{aligned} \tag{10}$$

where the node weight matrix $V = \{v_{ii}\} \in \mathcal{R}^{n\times n}$ is a diagonal matrix and the value v_{ii} is normalized node degree whin each individual class, which is computed as:

$$v_{ii} = d_{ii}/D^j = d_{ii}/\sum_{i=1}^{l} d_{ii}Y_{ij} \tag{11}$$

where assume $\mathbf{x}_i$ is with label $y_i = j$, then $D^j = \sum_{i=1}^{l} d_{ii}Y_{ij}$ is the total degree of the labeled nodes in class j. If we trace back to the graph regularization framework in Eq. 5, the revised loss function with label regularizer is:

$$\mathcal{Q}(F) = \frac{1}{2}\mathrm{tr}\left\{F'F - F'SF + \mu(F - VY)'(F - VY)\right\} \tag{12}$$

Conducting the partial differential on $\mathcal{Q}(F)$ as to F will result the same optimal F as Eq 10.

Similarly, we can apply the label regularizer term to the harmonic function formulation to derive label-regularized *GFHF* (*LR-GFHF*) approach. We rewrite the label regularizer matrix as:

$$V = \begin{bmatrix} V_{ll} & 0 \\ 0 & 0 \end{bmatrix} \tag{13}$$

Note that $F = [F_l \;\; F_u]'$ $\;Y = [Y_l \;\; Y_u]'$ and the harmonic conditions requires $F_l = Y_l$, we can rewrite the classification function as $F = [Y_l \;\; F_u]'$. From $\triangle F = 0$ on unlabeled data, we can get the function value of F on unlabeled data as:

$$F_u = (D_{uu} - W_{uu})^{-1} W_{ul} V_{ll} Y_l \tag{14}$$

2.4. Active Graph Transduction for Interactive Annotation

In the application of cellular microscopic image annotation, the expert interaction incrementally provide more labeled cell samples. Therefore, the graph propagation will be updated for each round of annotation. From the superposition law, we know that the graph propagation can be incremental in terms of new labels since the propagated functional components from new labeled data can be superposed to the previous optimized classification functions. Considering label regularizer, the new labeled data will change the weights v_{ii} on individual nodes. We hereby proposed the following active graph transduction approach.

The classification function F and label matrix Y can be written as the concatenation of column vectors as $F = [F_{\cdot 1} \cdots F_{\cdot j} \cdots F_{\cdot c}]$, and $Y = [Y_{\cdot 1} \cdots Y_{\cdot j} \cdots Y_{\cdot c}]$, where $F_{\cdot j}, Y_{\cdot j} (j = 1, \cdots, c)$ is corresponding to labeled samples from class j. Considering the superposition principle, the column vector $F_{\cdot j}$ is computed as:

$$F_{\cdot j} = \beta(I - \alpha S)^{-1} V Y_{\cdot j} \tag{15}$$

The above equation can be seen as the vector version of superposition law of Eq. 8. Assume the new labeled sample $\mathbf{x}_s$ with degree d_{ss} is with class $y_s = j$. From the discussion above, the label matrix is only updated in the jth column, which is vector $Y_{\cdot j}$. Thereby, from Eq. 15, only the vector $F_{\cdot j}$ need to be renewed. Let D^j denotes the total degree of the labeles in class j without counting new labeled sample $\mathbf{x}_s$, we can calculate two coefficients λ, γ as:

$$\lambda = \frac{D^j}{D^j + d_{ss}} \qquad \gamma = \frac{d_{ss}}{D^j + d_{ss}} \tag{16}$$

Obviously, the coefficients λ, γ satisfy $\lambda + \gamma = 1$. Then the new vector F_j^{new} can be updated as:

$$F_{\cdot j}^{new} = \lambda F_{\cdot j} + \gamma \hat{F}_s = \lambda F_{\cdot j} + \gamma P_{\cdot s} \tag{17}$$

where $\hat{F}_s$ is the propagated component with only labels $\mathbf{x}_s$. Let $P = \beta(I - \alpha S)^{-1}$, $\hat{F}_{\cdot s}$ is exactly the sth column vector of P, i.e. $\hat{F}_s = P_{\cdot s}$. Based on the superposition law discussed in the previous section, the updating of F by replacing the kth column with $F_{\cdot j}^{new}$ is equivalent to the the optimization result directly obtained from Eq. 12. However, the superposition approach shows more efficiency in

```
Input: cell samples $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_n\}$, labeled sample $Y_l = \{\mathbf{y}_1, \cdots, \mathbf{y}_l\}$, class $\mathcal{L} = \{1, \cdots, c\}$, microscopic image set $\mathcal{Z}$, each of which contains a cell sample subset $\hat{\mathcal{X}}$.
  1. Graph Construction:
     Calculate the affinity matrix $W = \{w_{ij}\}$, node degree matrix $D = diag(d_{ii})$, total degree of labeled samples for each class $D^j$, and propagation matrix $P$;
  2. Initialization Propagation:
     Calculate the propagated function: $F = PY$, and $F_{\cdot j} = PY_{\cdot j}, j = 1, \cdots, c$;
  3. Given a new labeled sample $\mathbf{x}_s$ and $y_s = j$:
     Compute the coefficients: $\lambda = \frac{D^j}{D^j + d_{ss}}, \gamma = \frac{d_{ss}}{D^j + d_{ss}}$;
  4. Update classification function $F$:
     With the calculated $\lambda, \gamma$, update the function $F$ in the $j$th column as $F_{\cdot j}^{new} = \lambda F_{\cdot j} + \gamma P_{\cdot s}$ and ${D^j}^{new} = D^j + d_{ss}$;
  5. If there are more new labeled samples, go to [3], else go to [6];
  6. Update image relevance to cellular phenotypes:
     For each microscopic image, update the image relevance score using Eq. 18.
Output: The image relevance score to cellular phenotypes.
```

Figure 3. Active annotation by superposable graph transductive algorithm with label regularizer.

terms of time cost since we reduce the computation from matrix multiplication to scale multiplication and vector addition. Note that the superposition framework can not be simply extended to *LR-GFHF* since the updating as to new labels requires to calculating the inverse of a dimensional-decreasing matrix, as shown in Eq. 14.

During active annotation for cellular microscopic images, the cells in each screening are propagated and finally assigned with soft labels denoted by the classification function F. Assume that the microscopic image $\mathbf{z}_t$ contains the a subset of cell samples $\hat{\mathcal{X}}_t$. The image relevance vector $\mathbf{r} = \{r_j\}, (j = 1, \cdots, c)$ representing the relevance score of this microscopy to each cellular phenotype is computed using the normalized soft labels.

$$r_j = \sum_{\mathbf{x}_t \in \hat{\mathcal{X}}_t} F_{tj} / n_t \tag{18}$$

where r_j is the relevance score as to the cellular phenotype j and n_t is the number of cells in this image. The recommended microscopic screening corresponding to a certain cellular phenotype query is based on the ranking of these relevance scores. we summarize the superposable transductive learning algorithm for interactive annotation in Fig. 3.

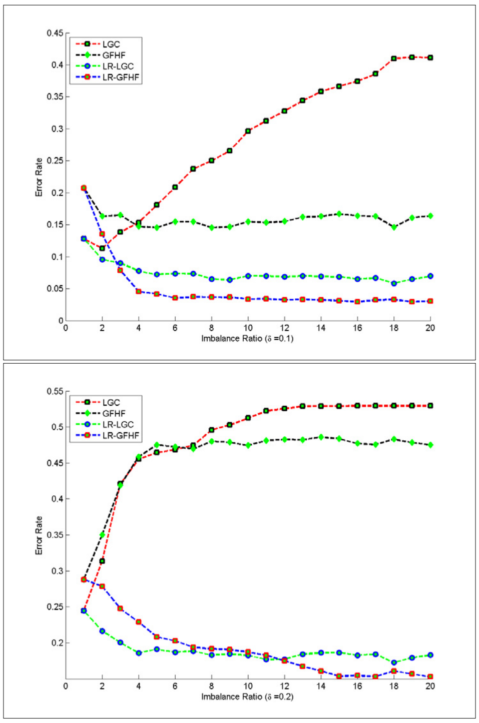


Figure 4. The performance comparison on the two-moon toy data. The kernel size is $\delta = 0.1$ (top row) and $\delta = 0.2$ (bottom row).

3. Experiments for Validating Label Regularizer Method

3.1. Toy Data

One of the illustration of the experiments on two-moon toy data has been show in Fig. 2, including 318 positive samples and 282 negative samples. Although in previous literatures , this two moon data has the perfect propagation results with reasonable setting [10][6]. However, the classification results have been empirically shown sensitive to the location of the given labels and ratio between two classes. Here we conduct more systematic experiments on this two-moon data. We fix one class with only one given label and the other class has number of the labeled samples from 1 to 20. The accuracy is based on the average of the 100 rounds random selections of the labeled samples. Fig. 4 shows the performance curves of the proposed label regularizer approaches, *LR-LGC* and *LR-GFHF*, compared with the standard *LGC* and *GFHF* methods. From the figure, we can see the label regularized approaches are much more robust to the imbalance labels and graph construction (different Gaussian kernel size δ).

3.2. USPS digital data

In order to comparing the experiments in [10], we use the same data for our handwritten digital experiments. A total of 3874 USPS digital samples, containing 1269, 929, 824, and 852 samples for the four digital $1, 2, 3, 4$ is used to eval-

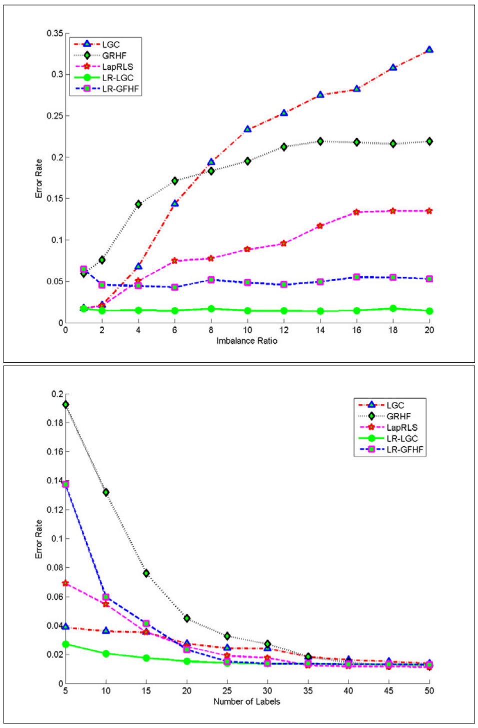


Figure 5. Performance comparing on USPS handwritten digital database: imbalance test (top row) and random test (bottom row).

uate the proposed approaches. We design two experimental strategies for comparison studies. First we deliberately create the imbalance label cases by using bias labels for a certain digital. For instance, we use $\tilde{l}$ labels for each digital of 1,2, and 3 and $r \cdot \tilde{l}$ labels of digital 4. We called r imbalance ratio, which is from 1 to 20 in the experiments. The second strategy is that we random choose some labels from the data, guaranteeing at least one label for each digital. Besides standard *LGC* and *GFHF*, another manifold regularization approach, Laplacian Regularized Least Squares (*LapRLS*) [2], is also tested for comparison study. Fig. 5 shows the experimental results. The error rate is based on the average on 100 trials.

Moreover, the procedure of building an efficient and robust graph is the key part of the graph based methods. The graph construction issue mostly means the calculation of the affinity matrix W. Usually, people prefer to use RBF kernel matrix [10] [4]. The value of the kernel size δ is not learnable in case of small labeled data. Previous research has shown that the propagation results highly depend on the kernel size δ selection [6]. However, this fixed size of kernel is not feasible to real data since the samples may not be sampled evenly and uniformly. There are some methods proposed to improve the graph construction, such as local scalling [9], local linear approximation [6], and adaptive kernel size selection [4]. In our experiments on real data (the above USPS handwritten digital and later cell images), we use an adaptive kernel size based on the mean distance

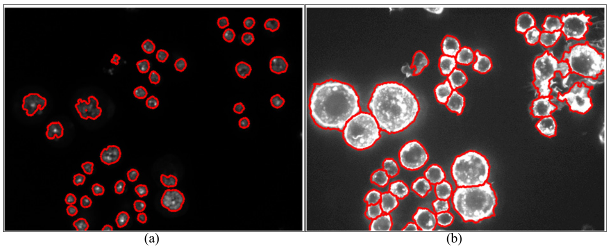


Figure 6. The automatic segmentation result of the microscopy image of Fig. 1. (a) nuclei segmentation; (b) extracted cell bodies.

of K-nearest neighborhoods. The number of nearest neighbors is empirically set as $K = 6$ for the experimental study.

From the comparison in Fig. 5, we can conclude that the label regularizer can improve the performance of *LGC* and *GFHF* and in both imbalance case (highly improved) and random case (slightly improved as the number of labels is increased). Especially, *LR-LGC* achieved the best performance in most cases.

4. Experiments on Active Annotation of Cellular Microscopy

4.1. Material and Preprocessing

In our experiments, we use the microscopic images of *Drosophila* K_{c167} embryonic cells to validate the active annotation approach in both accuracy and time cost. The images are acquired by automated microscopy with a Universal Imaging AutoScope Nikon TE300 [7]. The previous biological study on this dataset shows that the image appearance in the cell level reflected the underlying gene function expression [1]. However, it requires a huge burden of manual searching the positive cell samples and annotating the cellular phenotypes. Here we use 70 HCS microscopy screening sets, containing 210 cell images of three channels (only DNA and F-actin images are used for analysis). First we conduct homomorphic filtering on the raw data to get enhancement and denoising. Since the DNA signal is fairly strong, protruding out from a relatively uniform dark background; thus, nuclei are easily segmented by a histogram thresholding technique. However, cytoplasmic segmentation remains a challenging task due to intensity variation and cellular phenotype diversity. Starting from the well-segmented nuclei region, we applied a seeded watershed algorithm combining deformable model refining to separate both isolated and attached cell bodies as presented in [11] and [8]. Fig. 6 shows the cell segments of a cellular microscopic image. After segmentation, we obtained a total of 3162 valid cell segments, among of which 191 (6%) cells were manually labeled.

[1] abbreviated as CycA-sti since this cellular phenotype is frequently found in case of knocking down gene *CycA* and *sti* by RNAi.

Cell Phenotype	*Appearance Description*
Actin Accumulation (*AA*)	actin accumulation in the cell body, bright intensity, may have non-round nuclei;
Cell Cycle Arrest (*CycA-sti*) [1]	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.

Table 1. Biologically pre-defined cellular phenotypes and the appearance description.

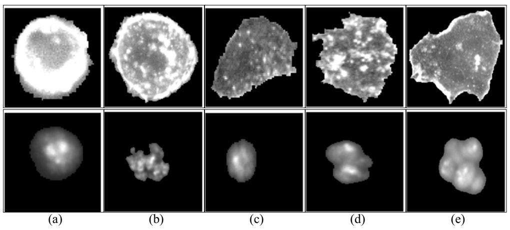


Figure 7. The cell segments examples of predefined cellular phenotype prototypes. The top row is the cytoplasm and the bottom row is the corresponding nuclei. (a) Actin Accumulation (*AA*); (b) Cell Cycle Arrest (*CycA-sti*); (c) Longthin-LPA (*LL*); (d) LS-Fla (*LF*); and (e) Rho.

For these cell segments, biologists pre-defined five distinct cellular phenotypes (Table 1). All these cellular phenotypes exhibit unique texture and geometric characteristics, as the cell prototypes shown in Fig. 7. In order to capture the morphological and appearance properties of different cellular phenotypes, a total of 214 dimensional attributes, including wavelet features, Zernike moments features, Haralick features, region properties, were computed from the cell segments [7].

4.2. Active Transduction for Interactive Annotation

Since each microscopic image contains a population of cells, some of which belongs to different phenotypes. However, the most dominant cellular phenotype in a certain microscopy reflects the underlying gene 'turn down' function expression. Hence, the microscopic images are categorized in five types, corresponding to the five phenotype in cell level. The task of annotating the image class is to ranking the image based on the relevance to a certain cell phenotype query. It can help the scientists rapidly target the most relevant genes related to a biological hypothesis. Moreover, it also can assist to collect the positive samples for further mining task.

In these experiments, we show how the active annotation framework improves the procedure of discovering the relevant microscopies given a small portion of labelled cells. In each annotation iteration, the values $F = \{F_{ij}\}$ for individual cells are obtained to compute the image relevance scores. Staring from 10 initial cellular labels, at least one for each phenotypes, we simulated the interactive annotation procedure by subsequentially adding 10 more cell labels in the next round. Fig. 8 gives the performance comparison of the five approaches. We can see that the annotation accuracy on the microscopies increases as to getting more cell labels. The label regularizer adjustment improved both *LGC* and *GFHF*. Eventually, after 8 rounds annotation, only 80 cell labels (around 2.5% of the total cell segments) can achieve 92.6% annotation accuracy (by LR-LGC). Fig. 9 and 10 shows two examples of the top four recommended microscopies by the system under the cellular phenotype query of *AA* and *Rho*. Meanwhile, the merit on computational cost of the active annotation is presented in Table 2. Since the graph construction can be executed off line, the table only provides the computation cost during active annotation procedure. The superposable frame work with LR-LGC highly reduced the computation burden, which can satisfy the requirement of online realtime annotation.

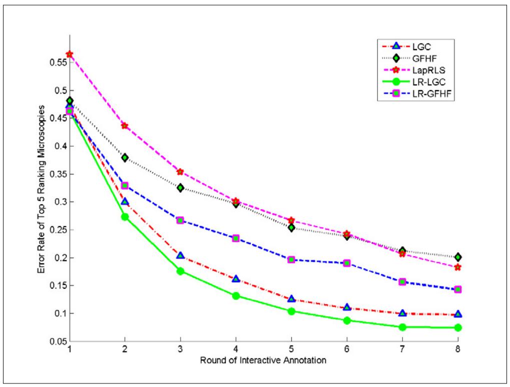


Figure 8. The performance of active annotation using graph transductive learning approach. X coordinate denotes the interaction rounds and Y coordinate denotes the accuracy of top 5 ranked microscopy images.

Method	LGC	GFHF	LapRLS	LR-LGC	LR-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.

5. Discussion and Conclusion

In this work, we proposed a novel graph transduction learning framework for the application of interactive RNAi cellular image annotation. To handle the fundamental problem in predicting cell phenotypes from a small set of training samples and highly imbalanced cell labels, we incorporate the label regularizer term to develop a new graph propagation approach. The merits of the proposed technique have been validated by significant performance gains over the toy data, USPS digital data, and real RANi cellular

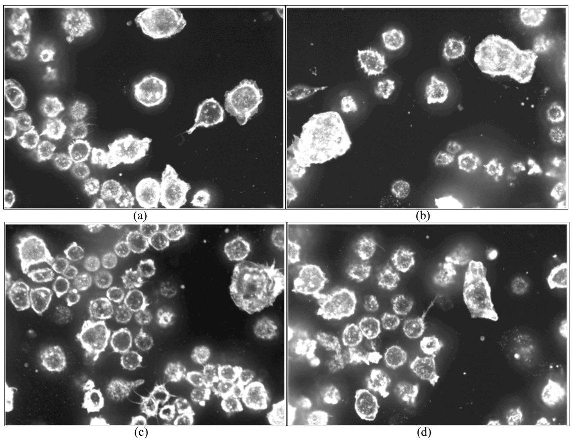


Figure 9. Active annotation result on the top four ranked microscopies under the query of *AA* cellular phenotype. The ranking scores are 0.8871, 0.8269, 0.7732, and 0.6245, respectively.

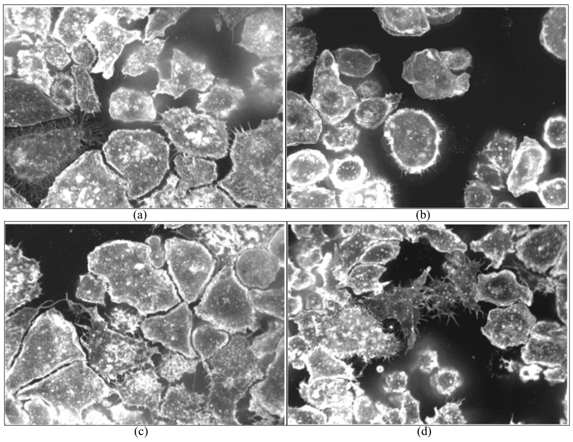


Figure 10. Active annotation result on the top four ranked microscopies under the query of *Rho* cellular phenotype. The ranking scores are 0.6667, 0.4286, 0.4242, and 0.4186, respectively.

images. Furthermore, in order to facilitate realtime interaction, we developed a superposable transductive learning algorithm to achieve the fast updating of cellular label propagation which adapts to incremental new cell labels generated from the interactive annotation.

The contributions in the application aspect include a novel framework for real-time analysis of bimolecular screening. We model the cellular image annotation task as a joint procedure of cell label propagation and image relevance ranking. Biologists may use the system to annotate and retrieve cellular images showing a large variety of cell phenotypes which are critical for various applications such as large scale gene function study and drug designs. To the best of our knowledge, this is the first multi-level graph transduction learning system successfully validated over real microscopic cellular images. The superposable graph transductive learning, real-time interaction designs make the system a truly scalable option for handling the explosively growing amount of cellular images in biological applications.

6. Acknowledgments

The authors would like to thank Dr. Norbert Perrimon and Dr. Chris Bakal for their valuable assistance and feedback. We also thank Mr. Wei Liu for the useful discussion on toy experiments and Mr. Zheng Yin for the help in collecting the cell images.

References

[1] C. Bakal, J. Aach, G. Church, and N. Perrimon. Quantitative Morphological Signatures Define Local Signaling Networks Regulating Cell Morphology. *Science*, 316(5832):1753, 2007. 6

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006. 6

[3] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proc. 18th ICML*, pages 19–26, 2001. 2

[4] M. Hein and M. Maier. Manifold denoising. *Proc. NIPS*, 19, 2006. 6

[5] A. A. Kiger, B. Baum, J. S, M. R. Jones, A. Coulson, C. Echeverri, and N. Perrimon. A functional genomic analysis of cell morphology using rna interference. *Journal of Biology*, 2:27, 2003. 1

[6] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *Proc. 23th ICML*, pages 985–992, 2006. 5, 6

[7] J. Wang, X. Zhou, P. L. Bradley, S.-F. Chang, N. Perrimon, and S. T.C. Wong. Cellular Phenotype Recognition for High-Content RNAi Genome-Wide Screening. *Journal of Biomolecular Screening*, 13(1):29–39, Feb. 2008. 1, 6, 7

[8] G. Xiong, X. Zhou, and L. Ji. Automated Segmentation of Drosophila RNAi Fluorescence Cellular Images Using Deformable Models. *IEEE Transactions on Circuits and Systems I*, 53(11):2415–2424, 2006. 6

[9] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. *Proc. NIPS*, 17:1601–1608, 2004. 6

[10] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *Proc. NIPS*, volume 16, pages 321–328, 2004. 2, 3, 4, 5, 6

[11] X. Zhou, K. Liu, P. Bradley, N. Perrimon, and S. Wong. Towards automated cellular image segmentation for RNAi genome-wide screening. In *Lecture Notes in Computer Science, MICCAI*. Spring-Verlag, 2005. 6

[12] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005. 2

[13] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. 20th ICML*, 2003. 2