Joint stage recognition and anatomical annotation of drosophila gene expression patterns
Xiao Cai, Hua Wang, Heng Huang, and Chris Ding
Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, 76019, USA

ABSTRACT

In this article, we propose a novel computational model for jointly stage classification and anatomical terms annotation of Drosophila gene expression patterns. We propose a novel Tri-Relational Graph (TG) model that comprises the data graph, anatomical term graph, developmental stage term graph, and connect them by two additional graphs induced from stage or annotation label assignments. Upon the TG model, we introduce a Preferential Random Walk (PRW) method to jointly recognize developmental stage and annotate anatomical terms by utilizing the interrelations between two tasks. The experimental results on two refined BDGP datasets demonstrate that our joint learning method can achieve superior prediction results on both tasks than the state-of-the-art methods.

Availability: http://ranger.uta.edu/%7eheng/Drosophila/
Contact: heng@uta.edu

1 INTRODUCTION

The mRNA in situ hybridization (ISH) provides an effective way to visualize gene expression patterns. The ISH technique can precisely document the localization of gene expression at cellular level via visualizing the probe by colorimetric or fluorescent microscopy. To produce the high-quality images recording the spatial location and intensity of the gene expression (Fowlkes et al., 2003; Hendriks et al., 2004; Lecuyer et al., 2007; Megason and Fraser, 2000), such spatial and temporal characteristics of expressions paved the way for inferring regulatory networks based on spatio-temporal dynamics. The raw data produced from such experiments includes digital images of the Drosophila embryo (examples are visualized in Fig. 1) using a particular gene expression pattern revealed by a gene-specific probe (Gumbrell et al., 2004; Lecuyer et al., 2003; Romancak et al., 2002; 2007). The fruit fly Drosophila melanogaster is one of the most used model organisms in developmental biology.

Traditionally, such ISH images are analyzed directly by the inspection of microscope images and available from well-known databases, such as the Berkeley Drosophila Genome Project (BDGP) gene expression pattern database (Romancak et al., 2002; 2007) and FlyISH (Lecuyer et al., 2007). To facilitate spatio-temporal Drosophila gene expression pattern studies, researchers needed to solve two challenging tasks first: Drosophila gene expression pattern stage recognition (temporal descriptions) and anatomical annotation (spatial descriptions). As shown in Figure 1, the Drosophila embryogenesis has been subdivided into 17 embryonic stages. These stages are defined by prominent features that are distinguishable in living Drosophila embryos (Weidmann et al., 2002). To recognize the stages of the Drosophila embryos, provide their time course patterns. On the other hand, the Drosophila gene expression patterns are often recorded by controlled vocabularies from the biologist’s perspective (Romancak et al., 2002). Such anatomical ontology terms describe the spatial biological patterns and often cross stages. What is more, because the ISH images are attached to each other collectively becoming bags of images, the corresponding stage label as well as anatomical controlled terms are the descriptions of the whole group of images instead of each individual image inside the bag. A Drosophila embryo ISH image bag belongs to only one stage, but has multiple related anatomical terms. Previously, those two tasks are tackled by domain experts. However, due to the rapid increase in the number of such images and the inevitable bias annotation by human curators, it is necessary to develop an automatic method to classify the developmental stage and annotate anatomical structure using controlled vocabulary.

Recently, a lot of research works have been proposed to solve the above two problems. They considered the stage recognition as a single-label multi-class classification problem while the anatomical annotation was treated as a multi-label multi-class classification problem. Kumar et al. (2003) first developed an embryo enclosing algorithm to find the embryo outline and extract the binary expression patterns via adaptive thresholding. Feng and Yeung (2004) and Feng et al. (2006) developed new ways to represent ISH images based on Gaussian mixture models, principal component analysis and wavelet functions. Besides that, they utilized min-redundancy max-relevance to do the feature selection and automatically classify gene expression patterns. Developmental stages. Recently, Annavarapu et al. (2011) constructed a system (called SPEx2) and concluded that the local regression (LR) method taking advantage of the controlled term interactions can get the best enhanced anatomical controlled term annotation results. The LR method was proposed by F et al. and developed based on their previous works.
**Drosophila gene expression patterns**

Fig. 1. Examples of *Drosophila* embryo ISH images and associated anatomical annotation terms in the stages 4–6, 7–8, 9–10, 11–12 and 13–16 in the BDGP database. The darker stained region highlights the place where the gene is expressed. The darker color the region has, the higher the gene expression level is.

**2 DATA DESCRIPTORS**

As we known, the *Drosophila* embryos are 3D objects. However, the corresponding image data can only demonstrate 2D information from a certain view. Since recent study has shown that incorporating images from different views can improve the classification performance consistently (Ji et al., 2008, 2010; Li et al., 2009; Shuiwang et al., 2009). All of the above methods have provided new inspirations and insights for classifying or annotating *Drosophila* gene expression patterns captured by ISH. However, none of them considered doing those two tasks simultaneously. As we know, intuitively, anatomical controlled vocabulary terms provide evidence for the stage label and vice versa. For example, the early stage range is more likely annotated with the controlled terms such as ‘statu nascendi’ and ‘celluar’ than the terms ‘embryonic’ and ‘epidermis’. Therefore, besides the image–stage and image–annotation relationships which have been well studied and applied in the previous research, it is necessary to take advantage of the correlations between stage classes and annotation terms.

In this article, we propose a novel Tri-Relational Graph (TG) model that comprises the data graph, anatomical controlled terms graph, developmental stage label graph to jointly classify the stage of images and annotate anatomical terms simultaneously. Upon the TG model, we introduce a Preferential Random Walk (PRW) method to simultaneously produce image-to-stage, image-to-annotation, image-to-image, stage-to-image, stage-to-annotation, stage-to-stage, annotation-to-image, annotation-to-stage and annotation-to-annotation relevances to jointly learn the salient patterns among images that are predictive of their stage label and anatomical annotation terms. Our method achieves superior developmental stage classification performance and anatomical terms annotation results compared with the state-of-the-art methods.

- **(1)** This article is the first one to propose a novel solution to the questions ‘What is the developmental stage?’ and ‘What are the anatomical annotations’ simultaneously, given an unlabeled image bag.
- **(2)** Via the new TG model that we constructed, the relationships between stage label and anatomical controlled terms as well as the correlations among anatomical terms can be naturally and explicitly exploited by the graph-based semi-supervised learning methods.
- **(3)** We propose a new PRW method to seek the hidden annotation–annotation and annotation–stage relevances. Other than only using image-to-image relevance conducted by existing methods, we can directly predict the stage label and annotate anatomical controlled terms for unknown image bags.

### 2.1 Codebook construction

Usually the codebook is established by conducting the clustering algorithms on a subset of the local features, and the cluster centers are then chosen as the visual words of the codebook. In our study, we use K-means to do the clustering on the training image bags. Since the result of K-means depends on the initial centers, we repeat it with 10 random initializations from which the one resulting in the smallest objective function value is selected. The number of clusters is set to 1000, 500 and 250 for lateral, dorsal and ventral images, respectively, according to the total number of images for each view as shown in Table 1. (Other codebook sizes gave similar performance.)
expression images bags. At last, each image bag is of the number of images in each view and each bag, we normalized corresponding bag-of-words representation is a vector of zeroes if a gene expression pattern data, we have $G = \{x_1, \ldots, x_n\}$, where each image bag $x_i$ is classified into class $c_i$, and 0 otherwise. Meanwhile, each image bag $x_i$ is also annotated with a number of anatomical ontology terms $A = \{a_1, \ldots, a_k\}$, such that $y_{ij}(x_i) = 1$ if $x_i$ is annotated with term $a_j$, and 0 otherwise. Also, for convenience, we write $y_i = [y_{i1}, y_{i2}, \ldots, y_{ik}]^T \in \{0, 1\}^{k \times 1}$. Without loss of generality, we assume the first $i < n$ image bags are already labelled, which are denoted as $T = \{x_1, y_{i1}\}_{i=1}^n$. Our task is to learn a function $f: X \rightarrow \{0, 1\}^{k \times 1}$ from $T$ that is able to classify an unlabeled data point $x_i$ ($i \geq n$) into one stage class in $C$ and to annotate it with a number of anatomical terms in $A$ at the same time. For simplicity, we write $Y_i = [y_{i1}, y_{i2}, \ldots, y_{ik}]$ and $Y = \{x_1, \ldots, y_{ik}\}$. As introduced in Section 1, the stage class and anatomical terms have some relations. We utilize the following affinity matrix to model their interrelations, $F \in \mathbb{R}^{k \times k}$, where $R(i,j)$ indicates how closely class $c_i$ and term $a_j$ are related. In this work, we compute it as

$$R(i,j) = \cos(\theta_{i,j}) = \frac{y_i^T y_j}{\|y_i\| \|y_j\|}$$

where $y_i$ is the $i$-th row of $Y$, and $y_j$ is the $j$-th row of $Y$. Throughout this article, we denote a vector as a bold lowercase character and a matrix as an uppercase character. We denote the $i$-th entry of a vector $v$ as $v(i)$, and the entry at the $i$-th row and $j$-th column of a matrix $M$ as $M(i,j)$, $|v|$ denotes the Euclidean norm of vector $v$. And the inner product of two vector $v_1$ and $v_2$ is defined as $v_1^T v_2$. The construction of TG

Different from conventional single-label classification learning problem in which classes are mutual exclusive, the anatomical terms are interrelated with one another. Again, we resort to cosine similarity to calculate the controlled term affinity matrix $W_k$, where $W_k(i,j)$ indicates the correlation between $a_i$ and $a_j$. Thus, a graph $G = (V, E)$ is induced, where $V = \{x_1, \ldots, x_n\}$ and $E \subseteq V \times V$. And we use LNN graph. To be specific, we connect $x_i, x_j$ if one of them is among the other’s $k$ nearest neighbor and define the value of the edge connecting them by Equation (1). Because $G_k$ characterizes the dependencies among data points, it is usually called as data graph, such as the middle subgraph in Figure 1. Existing graph-based semi-supervised learning methods [22, 33] only make use of the data graph, on which the class label information is propagated.

$$W_k(i,j) = \exp(-\frac{d_{ij}}{2\sigma^2}), \quad i \neq j$$

where the vector $x$ is calculated using bag-of-word features for one image bag. Regarding the parameter $\sigma$, we resort to self-tuning method [34].
relevance between a class/annotation term vertex and a data point vertex. As each class/annotation term has a set of associated training data points, which convey the same biological record information as the class/annotation term, we consider both a class/annotation term vertex and its labeled training image bag vertices as a group set, 

\[ G_{\alpha} = \{G_{\alpha}, E_{\alpha}\} \]  

which is illustrated as the vertices with orange boundary in Fig. 3. As a result, instead of measuring vertex-to-vertex relevance between a class/annotation term vertex and an unlabeled data point vertex, we may measure the set-to-set relevance between the group set and the data point. Motivated by Brin and Page (1998), we consider both a class/annotation term vertex and its labeled training image bag vertices as a group set, and meanwhile it takes a preference to go to other vertices specified by \( \mathbf{h} \) with probability \( \alpha \). The equilibrium distribution of PRW in Equation (5) is determined by \( p^* = (1 - \alpha)M^T p^* + \alpha \mathbf{h} \) which leads to:

\[ p^* = \alpha [I - (1 - \alpha)M^T]^{-1} \mathbf{h}. \]  

Due to Perron-Frobenius theorem, the maximum eigenvalue of \( M \) is less than max \( \sum M(i,j) = 1 \). Thus, \( I - (1 - \alpha)M^T \) is positive definite and invertible. Equation (6) takes a similar form to two existing works: random walk with restart (RWR) methods (Tong et al. 2006) and PageRank algorithm (Brin and Page 1998). In the former, \( \mathbf{h} \) is a vector with all entries to be 0 except one entry to be 1 indicating the vertex where the random walk could be restarted; while in the latter, \( \mathbf{h} \) is a constant vector called damping factor (Brin and Page 1998). In contrast, the preferential distribution vector \( h \) in Equation (6) is a generic probability distribution, which is flexible thereby more powerful. Most importantly, through \( \mathbf{h} \) we can assess group-to-vertex relevance, while RWR and PageRank methods measure vertex-to-vertex relevance.

Similarly to RWR (Tong et al. 2006), when we set the \( \mathbf{h} \) to be a probability distribution in which all the entries are 0 except for those corresponding to \( \{G_{\alpha}, E_{\alpha}\} \) measures how relevant the \( \alpha \)-th group is to the \( j \)-th vertex on \( G \).

### 3.3 Preferential random walk on TG

In order to classify and annotate unlabeled data points using the equilibrium probabilities in Equation (6) of the PRW on TG, we need to construct the transition matrix \( M \) and the preferential distribution \( \mathbf{h} \) from \( G \).

**Construction of the transition matrix \( M \):**

Let

\[ M = \begin{bmatrix} M_G & M_{MC} & M_{MCX} \\ M_{CX} & M_{CX} & M_{CX} \\ M_{AX} & M_{AX} & M_{AX} \end{bmatrix}. \]  

where \( M_G, M_{MC}, \) and \( M_{CX} \) are the intrasubgraph transition matrices of \( G \), \( G_C \), and \( G_{CX} \) respectively, and the rest six sub-matrices are the intersubgraph transition matrices among \( G \), \( G_C \), and \( G_{CX} \). Let \( \beta_l \in [0, 1] \) be the jumping probability, i.e., the probability that a random walker hops from \( G \) to \( G_C \) and vice versa. And let \( \beta_r \in [0, 1] \) be the jumping probability from \( G_C \) to
Therefore, $p_1$ and $p_2$ regulate the reinforcement between $G_X$ and one of the other two subgraphs. When both $p_1 = 0$ and $p_2 = 0$, the random walk are performed independently on $G_X$, which is equivalent to existing graph-based semi-supervised learning methods only using the data graph $G_X$. Similarly, we define $\lambda$ as the jumping probability from $G_Y$ to $G_X$ or vice versa.

During a random walk process, if the random walker is on a vertex of the data subgraph which has at least one connection to the label subgraph, such as vertex $x_i$ in Figure 4, she can hops to the class label or annotation subgraph with probability $p_1$, or annotation subgraph with probability $p_2$, or stay on the data subgraph with probability $1 - p_1 - p_2$ and hop to other vertices of the data subgraph. If the random walker is on a vertex of the data subgraph without a connection to the class label or annotation subgraph, she stays on the data subgraph and hops to other vertices on it as in standard random walk process. To be more precise, let $d^0_i = \sum_j T(i,j)$, the transition probability from $x_i$ to $x_j$ is defined as following:

$$p(x_i|x_j) = \frac{\beta_1 T(i,j)}{d_Y(i,j)}, \quad d_Y(i,j) > 0,$$

Similarly, let $d^0_i = \sum_i Y(i,j)$, the transition probability from $x_i$ to $x_j$ is:

$$p(x_i|x_j) = \frac{\beta_2 Y(i,j)}{d_Y(i,j)} , \quad d_Y(i,j) > 0.$$}

Following the same definition, the rest four inter-subgraph transition probability matrices are defined as:

$$p(a_i|x_j) = \frac{\beta_3 Y(i,j)}{d_Y(i,j)} , \quad d_Y(i,j) > 0,$$

$$p(a_i|x_j) = \frac{\beta_4 Y(i,j)}{d_Y(i,j)} , \quad d_Y(i,j) > 0.$$}

where $d_Y(i,j) = \sum_i Y(i,j)$ and $d_Z(i,j) = \sum_i Z(i,j)$, and

$$p(x_i|x_j) = \frac{\lambda I(i,j)}{d_Y(i,j)} , \quad d_Y(i,j) > 0.$$}

where $d_Y(i,j) = \sum_i Y(i,j)$ and $d_Z(i,j) = \sum_i Z(i,j)$. Let $d_Y = \sum_i Y(i,j), d_Z = \sum_i Z(i,j)$, $d_{Y^2} = \sum_i Y(i,j)^2, d_{Z^2} = \sum_i Z(i,j)^2$ where $Q_Y = Q_Y + Q_Y$ and $Q_Y = Q_Y - Q_Y$. The data subgraph intra transition probability from $x_i$ to $x_j$ is computed as:

$$p(x_j|x_i) = \frac{\lambda I(i,j)}{d_Y(i,j)} , \quad d_Y(i,j) > 0.$$}

Similarly, let $d^0_i = \sum_i W_Y(i,j)$, the annotation label subgraph intra transition probability from $a_i$ to $a_j$ is:

$$p(a_j|a_i) = \frac{(1 - \beta_1 - \beta_2)W_Y(i,j)}{d_Y(i,j)}, \quad d_Y(i,j) > 0.$$}

and

$$p(a_j|a_i) = \frac{\beta_3 W_Y(i,j)}{d_Y(i,j)} , \quad d_Y(i,j) > 0.$$}

where $d_Y(i,j) = \sum_i W_Y(i,j)$, the classification label subgraph intra transition probability from $a_i$ to $a_j$ is:

$$p(a_j|a_i) = \frac{\beta_4 W_Y(i,j)}{d_Y(i,j)} , \quad d_Y(i,j) > 0.$$}

It can be easily verified that, $\sum M(i,j) = 1$, i.e. $M$ is a stochastic matrix.

Construction of the preferential distribution $H$: the preferential distribution vector specifies a group of vertices to which the random walker prefers to moving in every iteration step. The relevance between this group and an vertex is measured by the equilibrium distribution of the random walk process. Therefore, we construct $K = K_1 \cup K_2$ preferential distribution vectors, one for each semantic group $G_i$:

$$h^{(i)}_i = \left[ \begin{array}{c} \frac{\gamma h^{(i)}_X}{(1 - \gamma)h^{(i)}_K} \\ \frac{(1 - \gamma) h^{(i)}_X}{(1 - \gamma)h^{(i)}_K} \end{array} \right] \in \mathbb{R}^{n \times K},$$

where $h^{(i)}_X(i) = 1 / \sum_j h_X(j)$ if $y(i) = 1$ and $h^{(i)}_X(i) = 0$, otherwise; $h^{(i)}_X(i) = 1$, if $i = k$; $\gamma \in [0, 1]$ controls how much the random walker prefers to go to the data subgraph $G_X$ and other two subgraphs $G_Y, G_A$. It can be verified that $\sum h^{(i)}_X(i) = 1$, i.e. $h^{(i)}_X$ is a probability distribution. Let $I_K$ be the identity matrix of size $K \times K$, we write

$$H = [h^{(1)}_X, \cdots, h^{(K)}_X] = \left[ \begin{array}{c} \gamma H_X \\ (1 - \gamma)H_X \end{array} \right] \in \mathbb{R}^{K \times K},$$

PRW on TG: given the TG of a dataset, using the transition matrix $M$ defined in Equation (4) and the preferential probability matrix $H$ defined in Equation (13), we can perform PRW on the TG. According to Equation (6), its equilibrium distribution matrix $P^n$ is computed as:

$$P^n = \beta(l - (1 - \omega) N) - 1, H,$$

$$P^n = [p^n_1, \cdots, p^n_K] \in \mathbb{R}^{n \times K},$$

and the annotated controlled terms for $x_i$ using the adaptive decision boundary method (Rame et al. 2020) on the submatrix $P_{na}^n$ by Equation (8) gives the stage prediction is a single-label classification problem, we set the stage label $R(i)$ with the maximum probability as the stage label for image bag $x_i$.

Since the stage prediction is a single-label classification problem, we set the stage label $R(i)$ with the maximum probability as the stage label for image bag $x_i$.

$$y(i) = \beta(p(i)), \forall n = 0, 1, 2, \ldots, n.$$

where $y(i)$ is the $i$-th row vector of matrix $P_{na}^n$. Therefore, we can do the stage classification and anatomical controlled term annotation simultaneously.

4 DATA REFINEMENT

In this section, we will introduce the details of the data used in our experiment. *Drosophila* embryogenesis has 17 stages, which are divided into 6 major ranges, i.e. stages 1–3, 4–6, 7–8, 9–10, 11–12 and 13–16 (the stage 17 is usually studied individually), in the BDPG database (Fonmanak et al. 2004). Each image bag is labeled with one stage term and many controlled vocabulary terms. The total number of anatomical controlled vocabulary terms is 303.

We used the following way to refine the dataset. First, we only keep the image bag data with lateral, dorsal and ventral view information. And then, we eliminate six common annotation terms, that is, ‘no staining’, ‘ubiquitous’, ‘strong ubiquitous’, ‘faint ubiquitous',
We used ‘inverse’ 5-fold cross-validation to determine the values of the following five parameters, that is, using 1-fold for training and using the remaining 4-folds for testing to mimic the scenario in the real application where the number of training data is much less than the testing data. In our experiment, we found that the following five parameters are not sensitive in certain ranges with less than 50 image bags as the data points for each stage. At last, the summary of the refined dataset is shown in Table 2.

5 EXPERIMENT

In this section, we will conduct experiments to evaluate PRW empirically on the refined dataset and compare it with other state-of-art classification methods. Since our method can do joint classification, in order to evaluate the benefit of joint learning, we compare its performance with that of the state-of-art multiclass single label or multiclass multilabel algorithms which can only handle either stage classification or anatomical term annotation problem. Our procedure is to train our model with stage labeled and anatomical term annotated image bags. All testing image bags are unlabeled with developmental stage and unannotated with anatomical controlled terms.

5.1 Experimental setup

When constructing PRW on TG, we used $k$-NN graph setting $k=9$. We used ‘inverse’ 5-fold cross-validation to determine the values of the following five parameters, that is, using 1-fold for training and using the remaining 4-folds for testing to mimic the scenario in the real application where the number of training data is much less than the testing data. In our experiment, we found that the following five parameters are not sensitive in certain ranges with good performances. $\beta_1$, $\beta_2$, and $\lambda$ controls the jumping between different subgraphs and cannot affect the result much if they are assigned in the range of $(0.1, 0.45)$. $\alpha$ controls initial preference of the random walker and will get stable result if it is assigned in the range of $(0.01, 0.1)$. $\gamma$ controls how much the random walker prefers to go to the data subgraph or to go to two other subgraphs and it is usually in the range of $(0.1, 0.3)$.

Besides those parameters, we also need to initialize the stage as well as anatomical controlled terms for the testing image bag $x_i$, where $i=1 \ldots l, l$ is the number of training image bag. In our experiment, we used $k$-nearest neighbor (KNN) method to do the initializations for both stage classification and anatomical term annotations tasks because of its simplicity and clear intuition. To be specific, we use $k=1$ and we abbreviate it as 1NN. Our joint classification framework will self-consistently amend the incorrect labels for stage and controlled terms. We perform 10 random splits of the data and report the average performance over the 10 trials. Please note that, in each trial, we still do ‘inverse’ 5-fold cross validation and record the average performance result as the result of that trial.

5.2 Image bag stage classification

*Drosophila* gene expression pattern stage categorization is a single-label multi-class problem. We compare the result of our method with that of support vector machine (SVM) with radial basis function (RBF) kernel $\mathcal{L}$ (Zhang and Lin, 2001). We use the optimal parameter values for $C$ and $\gamma$ got from cross-validation as well. We also compare the classification result of 1NN that we use to do the initialization. We assess the classification in terms of the average classification accuracy and the average confusion matrices. Since the data that we used is class balanced, the mean value of the entries on the diagonal of the confusion matrix is also the average classification accuracy. From the resulting average confusion matrix shown in Figure 6, we can see that the average prediction accuracy of our method is better than that of the other two state-of-art methods, especially in the last stage 13–16, where the number of anatomical terms is greatly larger than that of the other stages.

5.3 Image bag controlled vocabulary terms annotation

Besides the stage classification task, we also validate our method by predicting the anatomical controlled terms for the *Drosophila* gene expression patterns, which can be considered as a multi-class multi-label classification problem. The conventional classification performance metrics in statistical learning, precision and F1 score, are utilized to evaluate the proposed methods. For every anatomical term, the precision and F1 score are computed following the standard definition for the binary classification problem. To address the multi-label scenario, following Tsoumakas and Vlahavas (2009), macro and micro average of precision and F1 score are used to assess the overall performance across multiple labels. We compared four state-of-art multi-label classification methods: local shared subspace (LSS) (Li et al., 2008), local regression (LR) (Li et al., 2009), harmonic function (HF) (Zhu et al., 2003), and random walk (RW) (Zhou and Schölkopf, 2004). All of them are proposed recently to solve the multilabel annotation problem. In addition, we compare the results of 1NN as well. For the first three methods we use the published codes posted on the corresponding author’s websites. And we implement the RW method following the original work (Zhou and Schölkopf, 2004). For HF and RW methods, we follow the original work to solve the multilabel annotation only. Therefore, we only evaluate those two methods on data subgraph and annotation label subgraph without using any information derived from the classification label subgraph such as the stage-term correlation. Table 4 shows the average anatomical annotation performance of the 79-term dataset. Compared to the above five state-of-art methods, our method has the best results by all metrics. Figure 6 illustrates the average Micro F1 score of our method, 1NN, LSS, LR, RW and HF approaches for all the anatomical terms on 79-term dataset. And again, our method consistently achieves best performance for most of the anatomical controlled terms.

5.4 The advantage of joint learning

Unlike the traditional work, our proposed method can take advantage of all the information to do the stage classification and anatomical term annotation simultaneously. Therefore, when the number of training data is scare, we can resort to both intrarelations and
Fig. 4. The middle part demonstrates the terms–stages correlation and the right part shows the terms–terms correlation of 79 terms. The stage unknown test data shown in the left part is classified correctly as Stage 13–16, because of the strong correlation between the predicted stage and its predicted anatomical terms and vice versa, NOT the similarity of its first and second nearest neighboring data induced from the data graph only.

Table 2. Annotation prediction performance comparison on the 79-term dataset

| Method | Ma Pre | Ma F1 | Mi Pre | Mi F1 |
|--------|--------|--------|--------|--------|
| 1NN    | 0.3455 | 0.3595 | 0.2318 | 0.2230 |
| LS     | 0.5640 | 0.3778 | 0.3516 | 0.1903 |
| LR     | 0.6049 | 0.4425 | 0.3953 | 0.2243 |
| RW     | 0.4019 | 0.3385 | 0.2808 | 0.1835 |
| HF     | 0.3727 | 0.3296 | 0.2756 | 0.1733 |
| Our method | 0.6125 | 0.4434 | 0.4057 | 0.2336 |

Ma Pre, Avg. Macro Precision; Ma F1, Avg. Macro F1; Mi Pre, Avg. Micro Precision; Mi F1, Avg. Micro F1.

(iii) How to make better stage classification and anatomical controlled term annotation?

Interrelations to make the decision for stage classification and anatomical controlled term annotation simultaneously. When there are strong correlation between those two tasks, we expect that the performance of both tasks will be enhanced by joint learning work than treating them individually and independently. Figure 4 shows the pairwise label correlations of the 79 terms and stage–term correlations between 5 stages and 79 terms. As highlighted by purple arrows, we can observe that there are high pairwise correlations between the terms ‘embryonic brain’, ‘ventral nerve cord’ as well as ‘embryonic/larval muscle system’. Moreover, all the above three terms have high correlations with the stage 13–16, which can provide strong evidence that the given testing image bag could belong to the last developmental stage besides the induction from the data graph only. If our joint classification framework annotates it with all those three terms, although from the data similarity we cannot get strong evidence for the stage prediction, we can take advantage of the term–term as well as term–stage high correlations to adjust its stage to stage 13–16. In other words, relevant anatomical terms could help us to predict the stage label since they provide the spatial and temporal information of local structure corresponding to a specific embryo development stage. Nevertheless, not all anatomical terms will definitely benefit stage classification, which is consistent with our stage classification result. From Figure 4 we can see that our method may have competitive result compared with SVM with respect to some certain stage. However, given more anatomical term information, the performance of our method will gradually outperform the other methods, especially for the prediction result of stage 13–16.

Fig. 5. Stage classification results in terms of confusion matrices on 79-term dataset: (a) the confusion matrix calculated by SVM (b) the confusion matrix calculated by 1NN. (c) the confusion matrix calculated by our method. (a) SVM: acc. 84.50% (b) 1NN: acc. 77.40%; (c) our: acc. 85.20%
5.5 The more meaningful asymmetric correlation matrix

When we build the TG, at first, we assume the term–term correlation and stage–term correlation are both symmetric, since we used cosine similarity to represent their correlations. However, the above assumption does not always hold in the real data. In *Drosophila* embryo gene expression images, we found that the conditional probability of the occurrence of term ‘ventral nerve cord’ given term ‘embryonic brain’ is higher than that of the ‘embryonic brain’ given ‘ventral nerve cord’, which satisfies the biology meaning that ventral nerve cord occurs earlier than embryonic brain. After learning, our method can automatically discover the above hidden asymmetric correlation information, that is,

$$P^a = \begin{bmatrix}
 P^a((n+K_c+1, K_c+1) & \cdots & P^a((n+K_c+1, K) \\
 \vdots & \ddots & \vdots \\
 P^a(n+K, K_c+1) & \cdots & P^a(n+K, K)
\end{bmatrix}
$$  

(23)

In order to see the learned asymmetric term–term correlation more clearly, in Figure 7, we show the difference matrix got by $P^a - P^a_T$. Taking those more accurate asymmetric correlation into consideration, our method can potentially improve both stage classification and anatomical annotation results even more.

6 CONCLUSION

In this article, we proposed a novel TG model to learn the task interrelations between stage recognition and anatomical terms annotation of *Drosophila* gene expression patterns. The standard bag-of-word features and three major views (lateral, dorsal and ventral) were used to describe the 3D *Drosophila* images. A new PRW method was introduced to simultaneously propagate the stage labels and anatomical controlled terms via TG model. Both stage classification and anatomical controlled term annotation tasks are jointly completed. We evaluated the proposed method using one refined BDGP dataset. The experimental results demonstrated in the real application, when the number of training data is scarce, our joint learning method can achieve superior prediction results on both tasks than the state-of-the-art methods. What is more, we can discover more accurate asymmetric term–term correlation, which can potentially improve the results of both tasks even more.

ACKNOWLEDGEMENT

The author would like to thank Dr Sudhir Kumar for his help in data collection.

Funding: [This research was supported by National Science Foundation Grants CCF-0830780, CCF-0917274, DMS-0915228 and IIS-1117965].

REFERENCES

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. In *International Conference on World Wide Web (WWW)*, Elsevier Science Publishers, pp. 107–117.

Chang, C. and Lin, C. (2011) LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 1–27.

Fowlkes, C. et al. (2008) A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm. *Cell*, 133, 364–374.

Grumbling, G. et al. (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res.*, 34, D484–D488.

Hendriks, C. L. et al. (2006) Three dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution I: data acquisition pipeline. *Genome Biol.*, 7, R123.
X.Cai et al.

S.X. et al. (2008) Extracting shared subspace for multi-label classification. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2009, pp. 381–389.

S.X. et al. (2009) Drosophila gene expression pattern annotation using sparse features and ternary term interactions. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 407–416.

S.X. et al. (2010) A shared-subspace learning framework for multi-label classification. ACM Transactions on Knowledge Discovery from Data (TKDD), 4, 1–29.

Kang,F. et al. (2009) Correlated label propagation with application to multi-label learning. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1719–1726.

Kumar,S. et al. (2006) BEST: A novel computational approach for comparing gene expression patterns from early stages of Drosophila melanogaster development. Genetics, 162, 2037–2047.

Lecuyer,E. et al. (2007) Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell, 131, 174–187.

L.Y. et al. (2009) Drosophila gene expression pattern annotation through multi-instance multi-label learning. In Proceedings of the 21st International Joint Conference on Artificial Intelligence, AAAI press, pp. 1445–1450.

Lowe,D. (2004) Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis., 60, 91–110.

Lyne,R. et al. (2007) FlyMine: an integrated database for Drosophila and anopheles genomics. Genome Biol., 8, R129.

Megason,S. and Fraser,S. (2007) Imaging in systems biology. Cell, 130, 786–795.

Peng,H. and Myers,E.W. (2004) Comparing in situ mRNA expression patterns of drosophila embryos. In International Conference on Research in Computational Molecular Biology (RECOMB), ACM, pp. 157–166.

Peng,H. et al. (2007) Automatic image analysis for gene expression patterns of fly embryos. BMC Cell Biol., 8(Suppl. 1), S7.