Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

# Drosophila Gene Expression Pattern Annotations via Multi-Instance Biological Relevance Learning

**Hua Wang[†], Cheng Deng[♮], Hao Zhang[†], Xinbo Gao[♮], Heng Huang[‡*]**
[†]Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401, USA
[‡]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA
[♮]School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, P. R. China
huawangcs@gmail.com, chdeng@mail.xidian.edu.cn, hzhang@mines.edu,
xbgao@mail.xidian.edu.cn, heng@uta.edu

## Abstract

Recent developments in biology have produced a large number of gene expression patterns, many of which have been annotated textually with anatomical and developmental terms. These terms spatially correspond to local regions of the images, which are attached collectively to groups of images. Because one does not know which term is assigned to which region of which image in the group, the developmental stage classification and anatomical term annotation turn out to be a multi-instance learning (MIL) problem, which considers input as bags of instances and labels are assigned to the bags. Most existing MIL methods routinely use the Bag-to-Bag (B2B) distances, which, however, are often computationally expensive and may not truly reflect the similarities between the anatomical and developmental terms. In this paper, we approach the MIL problem from a new perspective using the Class-to-Bag (C2B) distances, which directly assesses the relations between annotation terms and image panels. Taking into account the two challenging properties of multi-instance gene expression data, high heterogeneity and weak label association, we computes the C2B distance by introducing class specific distance metrics and locally adaptive significance coefficients. We apply our new approach to automatic gene expression pattern classification and annotation on the *Drosophila melanogaster* species. Extensive experiments have demonstrated the effectiveness of our new method.

The mRNA *in situ* hybridization (ISH) is crucial for gene expression pattern visualization. The ISH technique can precisely record the localization of gene expression at the cellular level via visualizing the probe by colorimetric or fluorescent microscopy to allow the production of high quality images recording the spatial location and intensity of the gene expression (L'ecuyer et al. 2007; Fowlkes et al. 2008). In literature, more than one hundred thousand images of gene expression patterns from early embryogenesis are available for *Drosophila melanogaster* (fruit fly) (Tomancak et al. 2002; Lyne et al. 2007; Grumbling, Strelets, and Consortium 2006). These images are a treasure trove for identifying co-expressed and co-regulated genes and to trace the changes in a gene's expression over time (Tomancak et al. 2002; Lyne et al. 2007; Grumbling, Strelets, and Consortium 2006). Such spatial and temporal characterizations of expressions paved the way for inferring regulatory networks based on spatio-temporal dynamics. Knowledge gained from analysis of the Drosophila expression patterns is widely important, because a large number of genes involved in fruit fly development are commonly found in humans and other species. Thus, research efforts into the spatial and temporal characteristics of Drosophila gene expression images have been at the leading-edge of scientific investigations into the fundamental principles of different species development (Tomancak et al. 2002; Walter et al. 2010; Osterfield et al. 2013).

The comparative analysis of gene expression patterns need analyze a large number of digital images of individual embryos. To facilitate the search and comparison of gene expression patterns during Drosophila embryogenesis, it is highly desirable to annotate the developmental stage and tissue-level anatomical ontology terms for ISH images. This annotation is of significant importance in studying developmental biology, because it provides a direct way to reveal the interactions and biological functions of genes based on gene expressions and enhance gene regulatory networks research. Due to the rapid increase in the number of ISH images and the inevitable biased annotation by human curators, it is necessary to develop an automatic system to classify the developmental stage and annotate anatomical structure using controlled vocabulary.

Recently some bioinformatics research works have been developed to solve the annotation and stage classification problems (Kumar et al. 2002; Peng et al. 2007; Puniyani, Faloutsos, and Xing 2010; Ji et al. 2010; Shuiwang et al. 2009; Li et al. 2009; Ji et al. 2008). Kuma *et al*. (Kumar et al. 2002) developed an embryo enclosing algorithm to find the embryo outline and extract the binary expression patterns via adaptive thresholding. Peng *et al*. (Peng et al. 2007) developed approaches to represent ISH images based on Gaussian mixture models, principal component analysis and wavelet functions. Besides, they also utilized min-redundancy max-relevance to do the feature selection and automatically clas-

*To whom all correspondence should be addressed. This work was partially supported by the following grants: NSF-IIS 1117965, NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NSF-IIS 1423591, NIH R01 AG049371.

sify gene expression pattern developmental stages. A system called SPEX$^2$ was recently constructed (Puniyani, Faloutsos, and Xing 2010), which concluded that the local regression (LR) method taking advantage of the controlled term-term interactions can get the best enhanced anatomical controlled term annotation results. All these methods have provided good computational solutions to classify or annotate Drosophila gene expression patterns captured by ISH. However, a major challenge of automatically annotating gene expression images lies in that the gene expression pattern of a specific anatomical and developmental ontology term is body-part related and presents in local regions of images, while in available gene expression image databases, the terms are attached collectively to groups of images with the identity and precise placement of the term remaining a mystery. Each image panel is assigned a group of annotation terms, but this does not mean that all the annotations apply to every image in a group, nor does it mean that the terms must appear together for a specific image.

To tackle this annotation ambiguity problem, Multi-Instance Learning (MIL) has been introduced (Li et al. 2009) where the image panel of a gene is considered as a *bag* and each image is considered as an *instance* inside the bag. Despite its success to capture the hierarchical structures of the gene expression data, it fails to identify the which image(s) in a panel truly corresponds to the annotated terms. With these recognitions, in this paper we explore the challenges, as well as the opportunities, in annotating gene expression image. Instead of studying the *Bag-to-Bag (B2B)* distance usually used in many existing MIL methods, we propose to directly assess the relevance between annotation terms and image panels by using the *Class-to-Bag (C2B)* distance for MIL (Wang et al. 2011b; 2011a; Wang, Nie, and Huang 2011; 2012). Specifically, we consider each annotation term as a "super-bag", which comprises all instances from the bags annotated to this term. The elementary distance from an instance in a super-bag to a data bag is first estimated, then the C2B distance from the term to the image bag is computed as the sum of the elementary distances from all the instances in the super-bag to the interested data bag. Moreover, we consider the relative importance of a training instance with respect to its annotated term by assigning it with a weight for each of its annotated terms, called as *Significance Coefficient (SC)*. Ideally, the learned SCs of an instance with respect to its truly associated terms should be as large as possible, whereas its SCs with respect other terms should be as small as possible. By further enhancing the C2B distance via term specific distance metrics to narrow down the gap between high-level annotation terms and low-level visual features, we call the resulted C2B distance as *Instance Specific Distance (ISD)* (Wang, Nie, and Huang 2011). Because the learned SCs explicitly give the ranks of the images with respect to their annotated terms, it solves the instance level labeling ambiguity problem.

## Learning ISD for Multi-Instance Classification

In this section, we will first introduce a ISD (Wang, Nie, and Huang 2011) to address the challenges of general MIL. ISD is a C2B distance parameterized by the proposed SCs and enhanced by class specific distance metrics. Then we will develop our optimization objective to learn the parameters of ISD, followed by an novel yet efficient algorithm to solve the proposed objective, whose convergence is rigorously proved. Finally, the classification rules using the learned ISD will be presented.

### Problem Formalization

We first formalize the MIL problem for Drosophila gene expression pattern annotations. Given a gene expression image annotation task, we have $N$ training image panels $\mathcal{X} = \{X_1, \ldots, X_N\}$ and $K$ annotation terms. Each image panel contains a number of images represented by a bag of instances $X_i = \left[\mathbf{x}_i^1, \ldots, \mathbf{x}_i^{n_i}\right] \in \mathbb{R}^{d \times n_i}$, where $n_i$ is the number of images (instances) in the image panel (bag). Each instance is abstracted as a vector $\mathbf{x}_i^j \in \mathbb{R}^d$ of $d$ dimensions. We are also given the class (annotation term) memberships of the input data, denoted as $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T \in \{0,1\}^{N \times K}$ whose row $\mathbf{y}_i^T$ is the label indication of $X_i$. In the setting of MIL, if there exists $j \in \{1, \ldots, n_i\}$ such that $\mathbf{x}_i^j$ belongs to the $k$-th class, $X_i$ is assigned to the $k$-th class and $Y_{ik} = 1$, otherwise $Y_{ik} = 0$. Yet the concrete value of the index $j$ remains unknown. To be more specific, the following assumptions are held in MIL settings:

- bag $X$ is assigned to the $k$-th class $\Longleftrightarrow$ at least one instance of $X$ belongs to the $k$-th class;
- bag $X$ is not assigned to the $k$-th class $\Longleftrightarrow$ no instance in $X$ belongs to the $k$-th class.

Our goal is to learn from the training data $\mathcal{D} = \{X_i, \mathbf{y}_i\}_{i=1}^N$ a classifier that is able to annotate terms for a new query image panel $X$.

### ISD for Multi-Instance Data

Because the major difficulty of general MIL problems are how to estimate the set-to-set distances and elucidate the instance level labeling ambiguity, we tackle these two difficulties by applying the ISD (Wang, Nie, and Huang 2011).

**C2B Distance for Multi-Instance Data** It is broadly accepted that (Boiman, Shechtman, and Irani 2008) traditional B2B distance is not the true similarity measurement of the class relationships between data objects (image panels). Thus in this paper we consider to directly assess the relevance between a class and a data object using the C2B distance (Wang et al. 2011b; 2011a; Wang, Nie, and Huang 2011; 2012).

First we represent a class as a *super-bag* that comprises all the instances contained in the training bags labeled with the class of interest:

$$C_k = \left\{ \mathbf{x}_i^j \mid i \in \pi_k \right\} \ , \tag{1}$$

where $\pi_k = \{i \mid Y_{ik} = 1\}$ is the index set of all the training bags that belong to the $k$-th class. We denote the number of instances in $C_k$ as $m_k$, *i.e.*, $|C_k| = m_k$.

Note that, in single-label classification tasks (such as embryonic developmental stage classification) where each image panel belongs to exactly one class, *i.e.*, $\sum_{i=1}^K Y_{ik} = 1$,

we have

$$\begin{cases} C_k \cap C_l = \varnothing \ \ (\forall\, k \neq l) \ \ , \\ \sum_{k=1}^{K} m_k = \sum_{i=1}^{N} n_i \ \ . \end{cases} \tag{2}$$

In multi-label classification tasks (such as anatomical term annotations) (Wang, Huang, and Ding 2009; Wang, Ding, and Huang 2010; Wang, Huang, and Ding 2011) where each image panel (thereby each instance) may belong to more than one class, *i.e.*, $\sum_{i=1}^{K} Y_{ik} \geq 1$, we have

$$\begin{cases} C_k \cap C_l \neq \varnothing \ \ (\forall\, k \neq l) \ \ , \\ \sum_{k=1}^{K} m_k \geq \sum_{i=1}^{N} n_i \ \ . \end{cases} \tag{3}$$

That is, different super-bags may overlap and one instance $\mathbf{x}_i^j$ may appear in multiple super-bags.

Then we define the elementary distance from an instance $\mathbf{x}_i^j$ of a super-bag $C_k$ to a data bag $X_{i'}$ using the distance between $\mathbf{x}_i^j$ and its nearest neighbor instance in $X_{i'}$ as:

$$d_k\left(\mathbf{x}_i^j, X_{i'}\right) = \left\| \mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right) \right\|^2 \quad , \quad \forall i \in \pi_k \ , \tag{4}$$

where $\mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)$ denotes the nearest neighbor of $\mathbf{x}_i^j$ in $X_{i'}$.

Finally, the C2B distance from $C_k$ to $X_{i'}$ is computed as:

$$\begin{aligned} D\left(C_k, X_{i'}\right) &= \sum_{i \in \pi_k} \sum_{j=1}^{n_i} d_k\left(\mathbf{x}_i^j, X_{i'}\right) \\ &= \sum_{i \in \pi_k} \sum_{j=1}^{n_i} \left\| \mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right) \right\|^2 \ . \end{aligned} \tag{5}$$

**ISD— Parameterized C2B Distance** Because the C2B distance defined in Eq. (5) does not take into account the the instance level labeling ambiguity in MIL, we further develop it by weighting the instances in a super-bag upon their relevances to the corresponding classes.

Due to the ambiguous associations between instances and labels, not all the instances in a super-bag really characterize the corresponding class. To address this, we define $s_{ik}^j$ to be the weight for $\mathbf{x}_i^j$ with respect to the $k$-th class, we compute the C2B distance from $C_k$ to $X_{i'}$ as following:

$$D\left(C_k, X_{i'}\right) = \sum_{i \in \pi_k} \sum_{j=1}^{n_i} s_{ik}^j \left\| \mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right) \right\|^2 \ . \tag{6}$$

Because $s_{ik}^j$ reflects the relative importance of instance $\mathbf{x}_i^j$ when determining the label for the $k$-th class, we call it as the Significance Coefficient (SC) of $\mathbf{x}_i^j$ with respect to the $k$-th class, and the resulted C2B distance computed by Eq. (6) as the ISD as per (Wang, Nie, and Huang 2011).

SC is the most important contribution of this work from learning perspective of view, because it *explicitly* ranks the relevances of the training instances of a class. If the learned SCs make sense, the instance level labeling ambiguity in MIL is solved. Moreover, through the learned SCs, a clear picture of the insight of the input image panels of gene expressions can be seen.

**Refined ISD by Class Specific Distance Metrics** The ISD defined in Eq. (6) by definition is a weighted Euclidean distance, which is independent of input data. Similar to many other learning models, using Mahalanobis distance with an appropriate distance metric to capture the second-order statistics of input data is desirable for gene expression image annotation. Taking into account the high heterogeneity of gene expression data, instead of learning a global distance metric for all classes as in existing many works (Jin, Wang, and Zhou 2009; Guillaumin, Verbeek, and Schmid 2010), we learn $K$ different class specific distance metrics $\{M_k \succ 0\}_{k=1}^{K} \subset \mathbb{R}^{d \times d}$, one for each class. Note that, using class specific distance metrics is only feasible with the C2B distance, because we are only concerned with intra-class distance. However, traditional B2B distance needs to compute distances between bags belonging to different classes that involve inter-class distance metrics, which inevitably complicates the problem.

To be more precise, instead of using Eq. (6), we compute the ISD using the Mahalanobis distance as following:

$$\begin{aligned} &D\left(C_k, X_{i'}\right) \\ &= \sum_{i \in \pi_k} \sum_{j=1}^{n_i} s_{ik}^j \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T M_k \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right] \ . \end{aligned} \tag{7}$$

We refer to $D\left(C_k, X_{i'}\right)$ computed in Eq. (7) as the proposed ISD in the sequel of this paper.

## Optimization Objective

Equipped with the ISD defined in Eq. (7), following the standard learning strategy, we learn its two set of parameters, $s_{ik}^j$ and $M_k$, by maximizing the data separability, *i.e.*, we minimize the overall ISD from a class to all its belonging bags, whilst maximizing the overall ISD from the same class to all the bags not belonging to it. Formally, for a given class, say $C_k$, we solve the following optimization problem:

$$\min_{\substack{M_k \succ 0,\ \mathbf{s}_{ik} \geq 0, \\ \mathbf{s}_{ik}^T \mathbf{e} = 1}} \frac{\sum_{i' \in \pi_k} D\left(C_k, X_{i'}\right) + \gamma \sum_{i \in \pi_k} \mathbf{s}_{ik}^T \mathbf{s}_{ik}}{\sum_{i' \notin \pi_k} D\left(C_k, X_{i'}\right)} \ , \tag{8}$$

where $\mathbf{s}_{ik} = \left[s_{ik}^1, \ldots, s_{ik}^{n_i}\right]^T$ is the SC vector of $X_i$ with respect to the $k$-th class. In Eq. (8), $\mathbf{e} = [1, \ldots, 1]^T$ is a constant vector with all entries to be 1. The second term in the numerator of Eq. (8) is to avoid over-fitting and increase the numerical stability. Here we constrain the overall weight of a single bag with respect to a class to be unit, *i.e.*, $\mathbf{s}_{ik} \geq 0$, $\mathbf{s}_{ik}^T \mathbf{e} = 1$, such that all the training bags are fairly used. This constraint is equivalent to require the $\ell_1$-norm of $\mathbf{s}_{ik}$ to be 1 and implicitly enforce sparsity on $\mathbf{s}_{ik}$ (Tibshirani 1996), which is in accordance with the fact that one annotation term of an image bag usually arises from only one or a few of its images, but not all.

Because the class specific distance metric $M_k$ is positive definite, we can reasonably write it as $M_k = U_k U_k^T$ where $U_k$ is an orthonormal matrix such that $U_k^T U_k = I$. Thus, the

optimization problem in Eq. (8) is transformed as:

$$\min_{\substack{U_k^T U_k = I, \\ \mathbf{s}_{ik} \geq 0,\, \mathbf{s}_{ik}^T \mathbf{e} = 1}} \frac{\sum_{i' \in \pi_k} D\left(C_k, X_{i'}\right) + \gamma \sum_{i \in \pi_k} \mathbf{s}_{ik}^T \mathbf{s}_{ik}}{\sum_{i' \notin \pi_k} D\left(C_k, X_{i'}\right)} \ , \tag{9}$$

where the distance $D\left(C_k, X_{i'}\right)$ is defined as

$$\begin{aligned} & D\left(C_k, X_{i'}\right) \\ & = \sum_{i \in \pi_k} \sum_{j=1}^{n_i} s_{ik}^j \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T U_k U_k^T \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right] \ . \end{aligned} \tag{10}$$

## Optimization Algorithm

In order to solve the optimization problem in Eq. (9), we first present the following useful theorems.

**Theorem 1** *The global solution of the following general optimization problem:*

$$\min_{\mathbf{v} \in \mathcal{C}} \frac{f(\mathbf{v})}{g(\mathbf{v})} \ , \quad \textit{where } g(\mathbf{v}) \geq 0 \ \ (\forall\ \mathbf{v} \in \mathcal{C}) \ , \tag{11}$$

*is given by the root of the following function:*

$$h(\lambda) = \min_{\mathbf{v} \in \mathcal{C}} \ f(\mathbf{v}) - \lambda g(\mathbf{v}) \ , \tag{12}$$

**Proof.** Suppose $\mathbf{v}^*$ is the global solution of the problem (11), and $\lambda^*$ is the corresponding global minimal objective value, the following holds:

$$\frac{f(\mathbf{v}^*)}{g(\mathbf{v}^*)} = \lambda^* \ . \tag{13}$$

Thus $\forall\ \mathbf{v} \in \mathcal{C}$, we have

$$\frac{f(\mathbf{v})}{g(\mathbf{v})} \geq \lambda^* \quad \Longrightarrow \quad f(\mathbf{v}) - \lambda^* g(\mathbf{v}) \geq 0 \ , \tag{14}$$

which means:

$$\min_{\mathbf{v} \in \mathcal{C}} \ f(\mathbf{v}) - \lambda^* g(\mathbf{v}) = 0 \quad \Longleftrightarrow \quad h(\lambda^*) = 0 \ . \tag{15}$$

That is, the global minimal objective value $\lambda^*$ of the problem (11) is the root of the function $h(\lambda)$, which complete the proof of Theorem 1. □

**Theorem 2** *Algorithm 1 decreases the objective value of the problem (11) in each iteration.*

**Proof.** In the Algorithm 1, according to step 2 we know that

$$f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1}) \leq f(\mathbf{v}_t) - \lambda_t g(\mathbf{v}_t) \ , \tag{16}$$

According to step 1, we know that

$$f(\mathbf{v}_t) - \lambda_t g(\mathbf{v}_t) = 0 \ . \tag{17}$$

Thus we have

$$f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1}) \leq 0 \ , \tag{18}$$

which indicates that

$$\frac{f(\mathbf{v}_{t+1})}{g(\mathbf{v}_{t+1})} \leq \lambda_t = \frac{f(\mathbf{v}_t)}{g(\mathbf{v}_t)} \ , \tag{19}$$

and completes that proof. □

**Theorem 3** *Algorithm 1 is a Newton's method to find the root of the function $h(\lambda)$ in Eq.* (12).

**Proof.** According to step 2 in the Algorithm 1, we have

$$h(\lambda_t) = f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1}) \ . \tag{20}$$

Thus we have

$$h'(\lambda_t) = -g(\mathbf{v}_{t+1}) \ . \tag{21}$$

In Newton's method, the updated solution should be

$$\begin{aligned} \lambda_{t+1} &= \lambda_t - \frac{h(\lambda_t)}{h'(\lambda_t)} \\ &= \lambda_t - \frac{f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1})}{-g(\mathbf{v}_{t+1})} \\ &= \frac{f(\mathbf{v}_{t+1})}{g(\mathbf{v}_{t+1})} \ , \end{aligned} \tag{22}$$

which is exactly the step 1 in Algorithm 1. Namely, Algorithm 1 is a Newton's method to find the root of the function $h(\lambda)$. □

**Algorithm 1:** The algorithm to solve the problem (11).

$t = 1$. Initialize $\mathbf{v}_t \in \mathcal{C}$ ;
**while** *not converge* **do**
1. Calculate $\lambda_t = \frac{f(\mathbf{v}_t)}{g(\mathbf{v}_t)}$ ;
2. Calculate $\mathbf{v}_{t+1} = \arg\min_{\mathbf{v} \in \mathcal{C}} f(\mathbf{v}) - \lambda_t g(\mathbf{v})$ ;
3. $t = t + 1$ ;

Theorem 3 indicates that Algorithm 1 converges very fast and the convergence rate is quadratic convergence, *i.e.*, the difference between the current objective value and the optimal objective value is smaller than $\frac{1}{c^{c^t}}$($c > 1$ is a certain constant) at the $t$-th iteration. Therefore, Algorithm 1 scales well to large data sets in gene expression patterns classification tasks, which adds to its practical value.

Based upon Theorem 1 and Theorem 3, we employ the alternatively iterative method to solve the optimization problem in Eq. (9) using Algorithm 1 as following.

First, when $\mathbf{s}_{ik}$ is fixed, the problem in Eq. (9) can be written as following:

$$\min_{U_k^T U_k = I} \frac{Tr(U_k^T A_k U_k)}{Tr(U_k^T B_k U_k)} \ , \tag{23}$$

where the matrices $A_k$ and $B_k$ are defined as:

$$\begin{aligned} A_k = & \sum_{i' \in \pi_k} \sum_{i \in \pi_k} \sum_{j=1}^{n_i} s_{ik}^j \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right] \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T \\ & + \gamma \sum_{i \in \pi_k} \mathbf{s}_{ik}^T \mathbf{s}_{ik} I \ , \end{aligned} \tag{24}$$

$$B_k = \sum_{i' \notin \pi_k} \sum_{i \in \pi_k} \sum_{j=1}^{n_i} s_{ik}^j \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right] \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T \ . \tag{25}$$

According to the step 2 in the Algorithm 1, we need to solve the following problem:

$$\min_{U_k^T U_k = I} Tr(U_k^T A_k U_k) - \lambda_t Tr(U_k^T B_k U_k) \ , \tag{26}$$

which is known to have optimal solution with eigenvalue decomposition of $A_k - \lambda_t B_k$.

Second, when fixing $U_k$, we define a vector $\mathbf{d}_{ii'k} \in \mathbb{R}^{n_i}$, where the $j$-th element is

$$\left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T U_k U_k^T \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right] \ . \tag{27}$$

Then we can rewrite the problem in Eq. (9) as:

$$\min_{\mathbf{s}_{ik}\geq 0,\ \mathbf{s}_{ik}^T\mathbf{e}=1} \frac{\sum_{i'\in\pi_k}\sum_{i\in\pi_k}\mathbf{s}_{ik}^T\mathbf{d}_{ii'k} + \gamma\sum_{i\in\pi_k}\mathbf{s}_{ik}^T\mathbf{s}_{ik}}{\sum_{i'\notin\pi_k}\sum_{i\in\pi_k}\mathbf{s}_{ik}^T\mathbf{d}_{ii'k}} \ . \tag{28}$$

We define

$$\mathbf{d}_{ik}^w = \sum\nolimits_{i'\in\pi_k} \mathbf{d}_{ii'k} \in \mathbb{R}^{n_i} \ , \tag{29}$$

and

$$\mathbf{d}_{ik}^b = \sum\nolimits_{i'\notin\pi_k} \mathbf{d}_{ii'k} \in \mathbb{R}^{n_i} \ , \tag{30}$$

by which we further rewrite Eq. (28) as:

$$\min_{\mathbf{s}_{ik}\geq 0,\ \mathbf{s}_{ik}^T\mathbf{e}=1} \frac{\sum_{i\in\pi_k}(\mathbf{s}_{ik}^T\mathbf{d}_{ik}^w + \gamma\mathbf{s}_{ik}^T\mathbf{s}_{ik})}{\sum_{i\in\pi_k}\mathbf{s}_{ik}^T\mathbf{d}_{ik}^b} \ . \tag{31}$$

According to the step 2 of Algorithm 1, we solve:

$$\min_{\mathbf{s}_{ik}\geq 0,\ \mathbf{s}_{ik}^T\mathbf{e}=1} \sum_{i\in\pi_k}(\mathbf{s}_{ik}^T\mathbf{d}_{ik}^w + \gamma\mathbf{s}_{ik}^T\mathbf{s}_{ik}) - \lambda\sum_{i\in\pi_k}\mathbf{s}_{ik}^T\mathbf{d}_{ik}^b \ . \tag{32}$$

We define

$$\mathbf{d}_{ik} = \mathbf{d}_{ik}^w - \lambda\mathbf{d}_{ik}^b \ , \tag{33}$$

by which we rewrite the problem in Eq. (32) as:

$$\min_{\mathbf{s}_{ik}\geq 0,\ \mathbf{s}_{ik}^T\mathbf{e}=1} \sum_{i\in\pi_k}(\mathbf{s}_{ik}^T\mathbf{d}_{ik} + \gamma\mathbf{s}_{ik}^T\mathbf{s}_{ik}) \ . \tag{34}$$

We can see that the problem in Eq. (34) can be decoupled to solve the following subproblems separately for each $i \in \pi_k$:

$$\min_{\mathbf{s}_{ik}\geq 0,\ \mathbf{s}_{ik}^T\mathbf{e}=1} \mathbf{s}_{ik}^T\mathbf{d}_{ik} + \gamma\mathbf{s}_{ik}^T\mathbf{s}_{ik} \ , \tag{35}$$

which are convex quadratic programming (QP) problem, and can be efficiently solved because $\mathbf{s}_{ik} \in \mathbb{R}^{n_i}$ and the value of $n_i$ is usually not large in MIL problems.

## Classification Using ISD

Given a query videop clip $X$, using the learned class specific distance metrics and SCs,

$$\begin{cases} M_k \ \ (1 \leq k \leq K) \ , \\ s_{ik}^j \ \ (1 \leq k \leq K, 1 \leq i \leq N, 1 \leq j \leq n_i) \ , \end{cases} \tag{36}$$

we can compute $D(C_k, X) \ \ (1 \leq k \leq K)$ from all the classes to the query image using Eq. (7). Sorting $D(C_k, X)$, we can easily assign labels to the query image.

For single-label classification tasks, in which each image panel belongs to one and only one class, we assign $X$ to the class with minimum ISD, *i.e.*,

$$l(X) = \arg\min_k D(C_k, X) \ . \tag{37}$$

For multi-label classification tasks, in which one image panel may be assigned with more than one class label, we need a threshold to make prediction. For every class, we learn the adaptive decision boundary $b_k$ (Wang, Huang, and Ding 2009; 2013), which is then used to determine the class membership for $X$ using the following rule: assign $X$ to the $k$-th class if $D(C_k, X) < b_k$, and not otherwise.

Learning ISD by solving Eq. (8) and classifying query image panels using the rules above, our Explicit Instance Ranking (EIR) method for multi-instance classification is proposed.

# Experiment

In this section, we will conduct experiments to evaluate the proposed method empirically on Drosophila gene expression data and compare it with other state-of-art classification methods for both stage classification and anatomical term annotation. Note that, the former task is a single-label classification term, because each image panel can belongs to one and only one class (development stage), while the latter task is a multi-label classification task, because one image panel is usually annotated with more than one anatomical terms. Besides, we also study the effectiveness of the proposed SCs when elucidating the usefulness of each image in an image panel with respect a certain annotation term.

## Data Descriptions

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), we will use the images taken from multiple views instead of one perspective as the data descriptor. We only consider the lateral, dorsal, and ventral images in our experiment due to the fact that the number of images taken from other views is much less than that of the above three views. Following our prior work (Cai et al. 2012), all the images from the Berkeley Drosophila Genome Project (BDGP) database[1] have been pre-processed, including alignment and resizing to $128 \times 320$ gray images. For the sake of simplicity, we extract the popular SIFT (Lowe 2004) features from the regular patches with the radius as well as the spacing as 16 pixels (Shuiwang et al. 2009). Specifically, we extract one SIFT descriptor with 128 dimensions on each patch and each image is represented by 133 $(7 \times 19)$ SIFT descriptors. As a result, each image is represented by a fixed-length vector, while each gene expression pattern contains a number of images that forms an image panel.

[1] http://www.fruitfly.org/

Table 1: Stage classification results in terms of average classification accuracy over 5 developmental stages.

| Method | Accuracy |
|---|---|
| SVM | 0.845 |
| 1NN | 0.774 |
| PRW | 0.852 |
| Our method | 0.869 |

### Developmental Stage Classification of Image Bags

*Drosophila* gene expression pattern stage categorization is a single-label multi-class problem. We compare our method with support vector machine (SVM) with radial basis function (RBF) kernel (Chang and Lin 2001) and 1-Nearest Neighbor (1NN) classifier. We use the optimal parameter values for $C$ and $\gamma$ obtained from cross-validation for SVM. We also compare the classification result of the preferential random walk (PRW) method (Cai et al. 2012), which is one of the most recent simultaneous developmental stage classification and anatomical term annotation method and has reported state-of-the-art results on *Drosophila* gene expression patterns. We assess the classification in terms of the average classification accuracy as shown in Table 1. The results in Table 1 have shown that the average prediction accuracy of our method is better than that of all the three competing methods, which demonstrate the effectiveness of the proposed method in developmental stage classification for *Drosophila* gene expression patterns.

### Controlled Vocabulary Terms Annotation on Image Bags

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 multi-instance 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 2007), macro and micro average of precision and F1 score are used to assess the overall performance across multiple labels. We compared five state-of-art-multi-label classification methods: local shared subspace (LS) (Ji et al. 2008), local regression(LR) (Ji et al. 2009), harmonic function (HF) (Zhu, Ghahramani, and Lafferty 2003), random walk (RW) (Zhou and Schölkopf 2004) and PRW (Cai et al. 2012). All of them are proposed recently to solve the multi-label 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 web sites. 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 multi-label 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 shows the average anatomical annotation performance of 79-term dataset. Compared to the above five stat-of-the-art methods, our method has the best results by all metrics.

## Conclusions

In this paper, we proposed to explicitly learn the relative importance of each gene expression image in a collectively annotated image bag with respect to each annotation terms. As a result, we can clearly see the insight of the image bags and utilize the gene expression pattern data for better analysis. Our new method computes a novel Class-to-Bag (C2B) distance, which address the two major challenges in multi-instance learning — high heterogeneity and weak label association. Both developmental stage classification and anatomical controlled term annotation tasks are tested by our new method on the *Drosophila melanogaster* species. We evaluated the proposed method using one refined BDGP dataset. The experimental results demonstrated in the real application, our new learning method can achieve superior prediction results on both tasks than the state-of-the-art methods. Moreover, the learned Significance Coefficients have clear biological meanings, which additionally confirm the correctness of our new method.

## References


Boiman, O.; Shechtman, E.; and Irani, M. 2008. In defense of nearest-neighbor based image classification. In *IEEE CVPR*.

Cai, X.; Wang, H.; Huang, H.; and Ding, C. 2012. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics* 28(12):i16–i24.

Chang, C., and Lin, C. 2001. LIBSVM: a library for support vector machines.

Fowlkes, C.; Hendriks, C. L.; Keranen, S.; and et al. 2008. A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell* 133:364–374.

Grumbling, G.; Strelets, V.; and Consortium, T. F. 2006. FlyBase: anatomical data, images and queries. *Nucl. Acids Res.* 34:D484–488.

Guillaumin, M.; Verbeek, J.; and Schmid, C. 2010. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*.

Ji, S.; Tang, L.; Yu, S.; and Ye, J. 2008. Extracting shared subspace for multi-label classification. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 381–389. ACM.

Ji, S.; Yuan, L.; Li, Y.; Zhou, Z.; Kumar, S.; and Ye, J. 2009. Drosophila gene expression pattern annotation using sparse features and term-term interactions. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 407–416.

Ji, S.; Tang, L.; Yu, S.; and Ye, J. 2010. A shared-subspace learning framework for multi-label classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4(2):1–29.

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

| Method | Macro average | | Micro average | |
|---|---|---|---|---|
| | Precision | F1 | Precision | 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 |
| PRW | 0.6125 | 0.4434 | 0.4057 | 0.2336 |
| Our method | **0.6431** | **0.4815** | **0.4319** | **0.2711** |


Jin, R.; Wang, S.; and Zhou, Z. 2009. Learning a distance metric from multi-instance multi-label data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE.

Kumar, S.; Jayaraman, K.; Panchanathan, S.; Gurunathan, R.; Marti-Subirana, A.; and Newfeld, S. 2002. BEST: A novel computational approach for comparing gene expression patterns from early stages of Drosophila melanogaster development. *Genetics* 162:2037–2047.

L'ecuyer, E.; Yoshida, H.; Parthasarathy, N.; and et al. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131:174–187.

Li, Y.; Ji, S.; Kumar, S.; Ye, J.; and Zhou, Z. 2009. Drosophila gene expression pattern annotation through multi-instance multi-label learning. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, CA*.

Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2):91–110.

Lyne, R.; Smith, R.; Rutherford, K.; and et al. 2007. FlyMine: an integrated database for Drosophila and anopheles genomics. *Genome Biol.* 8:R129.

Osterfield, M.; Du, X.; Schupbach, T.; Wieschaus, E.; and Shvartsman, S. 2013. Three-dimensional epithelial morphogenesis in the developing drosophila egg. *Dev Cell* 24(4):400–10.

Peng, H.; Long, F.; Zhou, J.; Leung, G.; Eisen, M.; and Myers, E. 2007. Automatic image analysis for gene expression patterns of fly embryos. *BMC Cell Biology* 8(Suppl 1):S7.

Puniyani, K.; Faloutsos, C.; and Xing, E. P. 2010. SPEX$^2$: Automated Concise Extraction of Spatial Gene Expression Patterns from Fly Embryo ISH Images. *Intelligent Systems for Molecular Biology(ISMB)* 26(12):i47–i56.

Shuiwang, J.; Ying-Xin, L.; Zhi-Hua, Z.; Sudhir, K.; and Jieping, Y. 2009. A bag-of-words approach for Drosophila gene expression pattern annotation. *BMC Bioinformatics* 10.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. *J. Royal. Statist. Soc B.* 58:267–288.

Tomancak, P.; Beaton, A.; Weiszmann, R.; and et al. 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.* 3:88.

Tsoumakas, G., and Vlahavas, I. P. 2007. Random - labelsets: An ensemble method for multilabel classification. In *ECML*, 406–417.

Walter, T.; Shattuck, D. W.; Baldock, R.; and et al. 2010. Visualization of image data from cells to organisms. *Nat Methods* 7(3):S26–41.

Wang, H.; Huang, H.; Kamangar, F.; Nie, F.; and Ding, C. H. 2011a. Maximum margin multi-instance learning. In *Advances in neural information processing systems (NIPS)*, 1–9.

Wang, H.; Nie, F.; Huang, H.; and Yang, Y. 2011b. Learning frame relevance for video classification. In *Proceedings of the 19th ACM international conference on Multimedia*, 1345–1348. ACM.

Wang, H.; Ding, C.; and Huang, H. 2010. Multi-label linear discriminant analysis. In *Proc. of ECCV*, 126–139.

Wang, H.; Huang, H.; and Ding, C. 2009. Image annotation using multi-label correlated green's function. In *IEEE ICCV 2009*, 2029–2034. IEEE.

Wang, H.; Huang, H.; and Ding, C. 2011. Image annotation using bi-relational graph of images and semantic labels. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 793–800. IEEE.

Wang, H.; Huang, H.; and Ding, C. 2013. Function–function correlated multi-label protein function prediction over interaction networks. *Journal of Computational Biology* 20(4):322–343.

Wang, H.; Nie, F.; and Huang, H. 2011. Learning instance specific distance for multi-instance classification. In *AAAI*.

Wang, H.; Nie, F.; and Huang, H. 2012. Robust and discriminative distance for multi-instance learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2919–2924. IEEE.

Zhou, D., and Schölkopf, B. 2004. Learning from labeled and unlabeled data using random walks. In *DAGM-Symposium*, 237–244.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. D. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 912–919.