Automated annotation of *Drosophila* gene expression patterns using a controlled vocabulary

Shuiwang Ji¹,², Liang Sun¹,², Rong Jin⁴, Sudhir Kumar²,³, and Jieping Ye¹,²

¹Department of Computer Science and Engineering, Arizona State University, Tempe, AZ, ²Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, AZ, ³School of Life Sciences, Arizona State University, Tempe, AZ, ⁴Department of Computer Science and Engineering, Michigan State University, East Lansing, MI

ABSTRACT

**Motivation**: Regulation of gene expression in space and time directs its localization to a specific subset of cells during development. Systematic determination of the spatiotemporal dynamics of gene expression plays an important role in understanding the regulatory networks driving development. An atlas for the gene expression patterns of fruit fly *Drosophila melanogaster* has been created by whole-mount *in situ* hybridization, and it documents the dynamic changes of gene expression pattern during *Drosophila* embryogenesis. The spatial and temporal patterns of gene expression are integrated by anatomical terms from a controlled vocabulary linking together intermediate tissues developed from one another. Currently, the terms are assigned to patterns manually. However, the number of patterns generated by high-throughput *in situ* hybridization is rapidly increasing. It is therefore tempting to approach this problem by employing computational methods.

**Results**: In this article, we present a novel computational framework for annotating gene expression patterns using a controlled vocabulary. In the currently available high-throughput data, annotation terms are assigned to groups of patterns rather than to individual images. We propose to extract invariant features from images, and construct pyramid match kernels to measure the similarity between sets of patterns. To exploit the complementary information conveyed by different features and incorporate the correlation among patterns sharing common structures, we propose efficient convex formulations to integrate the kernels derived from various features. The proposed framework is evaluated by comparing its annotation with that of human curators, and promising performance in terms of F1 score has been reported.

**Contact**: shuiwang.ji@asu.edu

1 INTRODUCTION

Detailed knowledge of the expression and interaction of genes is crucial to deciphering the mechanisms underlying cell-fate specification and tissue differentiation. DNA microarrays and RNA *in situ* hybridization are two primary methods for monitoring gene expression levels on a large scale. Microarrays provide a quantitative overview of the relative changes of expression levels of a large number of genes, but they do not often document the spatial information on individual genes. In contrast, RNA *in situ* hybridization uses gene-specific probes and can determine the spatial patterns of gene expression precisely. Recent high-throughput investigations have yielded spatiotemporal information for thousands of genes in organisms such as *Drosophila* (Tomancak et al., 2002; Lécuyer and et al., 2007) and mouse (Carson and et al., 2005; Lein and et al., 2006). These data have the potential to provide significant insights into the functions and interactions of genes (Kumar et al., 2002; Samsonova et al., 2007).

The fruit fly *Drosophila melanogaster* is one of the model organisms in developmental biology, and its patterns of gene expression have been studied extensively (Campos-Ortega and Hartenstein, 1997; Arbeitman et al., 2002; Tomancak and et al., 2002; Lécuyer et al., 2007). The comprehensive atlas of spatial patterns of gene expression during *Drosophila* embryogenesis has been created by *in situ* hybridization techniques, and the patterns are documented in the form of digital images (Tomancak and et al., 2002; Grumbling et al., 2006; Van Emden et al., 2006; Harmon et al., 2007). Comparative analysis of gene expression pattern images can potentially reveal new genetic interactions and yield insights into the complex regulatory networks governing embryonic development (Tomancak and et al., 2002; Kumar et al., 2002; Peng and Myers, 2004; Estrada et al., 2006).

To facilitate pattern comparison and searching, the images of *Drosophila* gene expression patterns are annotated with anatomical and developmental ontology terms using a controlled vocabulary (Tomancak and et al., 2002; Grumbling et al., 2006). The basic requirement for annotation is to assign a unique term, not only for each terminally differentiated embryonic structure, but also for the developmental intermediates that correspond to it. Four general classes of terms, called anlage *in statu nascendi*, anlage, primordium, and organ (ordered in terms of developmental time), are used in the annotation. Such an elaborate naming scheme describes a developing “path”, starting from the cellular blastoderm stage until organs are formed, that documents the dynamic process of *Drosophila* embryogenesis. Due to the overwhelming complexity of this task, the images are currently annotated manually by human experts. However, the number of available images produced by high-throughput *in situ* hybridization is now rapidly increasing (Kumar et al., 2002; Gurunathan et al., 2004; Peng and Myers, 2004; Ye et al., 2006; Tomancak and et al., 2007). It is therefore tempting to design computational methods for the automated annotation of gene expression patterns.

The automated annotation of *Drosophila* gene expression patterns was originally considered difficult due to the lack of a large reference data set from which to learn. Moreover, the “variation in morphology and incomplete knowledge of the shape and position of various embryonic structures” have made this task more elusive (Tomancak and et al., 2002). We attempt to address this...
problem by resorting to advanced tools developed recently in the computer vision and machine learning research communities and on the large set of annotated data available from the Berkeley Drosophila Genome Project (BDGP) (Tomancak and et al., 2002). There are several challenging questions that need to be addressed when approaching this problem by computational methods. As has been stated in Tomancak and et al. (2002), the first challenge is to deal with the issue that the same embryonic structure can appear in different shapes and positions due to the distortions caused by the image acquisition process. Fortunately, recent advances in object recognition research have led to robust methods that can detect interest regions and extract features that are invariant to a class of transformations from these regions. These two correlated lines of research have reached some maturity now (see Mikolajczyk et al. (2005) and Mikolajczyk and Schmid (2005) for an overview).

The second challenge of this task lies in the data representation. The embryogenesis of Drosophila has been divided into six discrete stage ranges (1-3, 4-6, 7-8, 9-10, 11-12, and 13-16) in the BDGP high-throughput study (Tomancak and et al., 2002). Gene expression patterns are documented collectively by a group of biological applications (Lanckriet et al., 2004b; De Bie et al., 2007). For the problem of gene expression pattern annotation, a variable number of terms from the controlled vocabulary can be assigned to a group of patterns. Hence, this problem belongs to the more general framework of multi-label learning. We propose methods based on hypergraph (Agarwal et al., 2006; Zhou et al., 2007) to project and combine the multiple kernel matrices for multi-label data. The proposed formulation can capture the correlation among patterns sharing a common embryonic structure by including them in a common edge in hypergraph. We also show that kernel canonical correlation analysis (Hardoon et al., 2004) is a special case of the proposed formulation. The overall flowchart of the proposed framework is depicted in Fig. 2.

![Diagram](image.jpg)

**Fig. 1.** Sample image sets and the associated terms in the Berkeley Drosophila Genome Project (BDGP) database in two stage ranges. Only images taken from lateral view with the anterior to the left are shown.

For the problem of gene expression pattern annotation, a variable number of terms from the controlled vocabulary can be assigned to a group of patterns. Hence, this problem belongs to the more general framework of multi-label learning. We propose methods based on hypergraph (Agarwal et al., 2006; Zhou et al., 2007) to project and combine the multiple kernel matrices for multi-label data. The proposed formulation can capture the correlation among patterns sharing a common embryonic structure by including them in a common edge in hypergraph. We also show that kernel canonical correlation analysis (Hardoon et al., 2004) is a special case of the proposed formulation. The overall flowchart of the proposed framework is depicted in Fig. 2.

**Fig. 2.** Illustration of the proposed framework for multi-label multiple kernel learning. The proposed formulation can capture the correlation among patterns sharing a common embryonic structure by including them in a common edge in hypergraph. We also show that kernel canonical correlation analysis (Hardoon et al., 2004) is a special case of the proposed formulation. The overall flowchart of the proposed framework is depicted in Fig. 2.

**2 FEATURE GENERATION AND KERNEL CONSTRUCTION**

In this section, we present our methods for extracting features from gene expression pattern images and constructing kernels between sets of patterns.

**2.1 Feature generation**

There are two primary methods for extracting features. When the images are not well-aligned, the covariant region detector is first applied on the

\[ K_1 \oplus K_2 \oplus \ldots K_p \]

\[ \sum \theta_i k_i \]
images to detect interest regions. Then, local descriptor is used to extract features from the detected regions. An alternative approach is to apply local descriptor on a dense regular grid, instead of interest regions (Grauman and Darrell, 2007; Lazebnik et al., 2006). Such an approach is motivated from the bag-of-words model from the text-modeling literature, and competitive performance has been achieved on image applications (Fei-Fei and Perona, 2005). Since the images in our FlyExpress (Van Emden et al., 2006) database are already well-aligned, we take the second approach in this article (Fig. 2). Instead of tuning the local descriptor and grid size manually, we apply several popular local descriptors on regular grids of different sizes, and rely on the MKL framework to select the appropriate local descriptors and grid size. More details on feature generation are described in Section 4.

2.2 Pyramid match kernels

In kernel methods, a symmetric function is called a ‘kernel function’ if it satisfies the Mercer’s condition (Schölkopf and Smola, 2002). When used for a finite number of samples in practice, this condition amounts to requiring that the kernel matrix is positive semidefinite. The wide applicability of kernel methods stems from the fact that they only require the characterization of similarities between objects by the use of the kernel trick. The pyramid match algorithm (Grauman and Darrell, 2005, 2006, 2007) computes kernels for variable-sized sets of feature vectors. The main idea of this approach is to convert sets of features to multi-dimensional, multi-resolution histograms, and then compute the similarity between the corresponding histograms based on histogram intersections. The final similarity between two sets of vectors is computed as a weighted sum of the similarities at the histogram levels. This similarity is an approximation to the similarity of the best partial matching between the feature sets. The resulting similarity matrix based on this measure is probably positive definite, and it can be used in existing kernel-based learning algorithms. Details on the pyramid match algorithm can be found in the Supplement.

The pyramid match algorithms proposed in Grauman and Darrell (2005, 2006, 2007) treat the sets of features to be matched as orderless. In some applications, the spatial layout of features within a set may convey critical discriminative information. Lazebnik et al. (2006) proposed the spatial pyramid matching algorithm to perform pyramid matching in the two-dimensional image space, thus taking the spatial information into account directly. The main idea of this approach is to quantize the local features in images into a number of discrete types by applying clustering algorithms, and then place multi-resolution histogram pyramid on the two-dimensional images. It is also possible to integrate geometric information directly into the original pyramid match algorithm by adding the image coordinates as two additional dimensions into each feature vector (Lazebnik et al., 2006; Grauman, 2007), and we adopt this approach in this article. Note that the original pyramid match algorithms are proposed to match two images, and that we extend them to match two sets of images.

3 LEARNING WITH MULTIPLE KERNELS

In this section, we present a multi-label, multiple kernel learning formulation based on hypergraph for integrating the kernel matrices derived from various local descriptors. Results in Section 4 show that the integrated kernels yield better performance than that of the best individual kernel.

3.1 Hypergraph spectral learning

Hypergraph generalizes traditional graph by allowing edges, known as “hyperedges”, to connect more than two vertices, thus capturing the joint relationship among multiple vertices. We propose to construct a hypergraph (for the collection of gene expression patterns in question) in which each pattern is represented as a vertex. To document the joint similarity among patterns annotated with a common term, we propose to construct a hyperedge for each term in the vocabulary, and include all patterns annotated with a common term into one hyperedge. Hence, the number of hyperedges in this hypergraph equals the number of terms in the vocabulary. Laplacian is commonly used to learn from a graph (Chung, 1997). To learn from a hypergraph, one can either define hypergraph Laplacian directly, or expand it into a traditional graph for which Laplacian is constructed. Since it has been shown that the Laplacians defined in both ways are similar (Agarwal et al., 2006), we use the expansion-based approaches in this article. The star and clique expansions are two commonly-used schemes for expanding hypergraphs. Following the spectral graph embedding theory (Chung, 1997), we propose to project the patterns into a low-dimensional space in which patterns sharing a common term are close to each other. When formulated in the kernel-induced feature space, this can be achieved by solving the following optimization problem:

$$\max_{B} \quad \text{trace} \left( B^T (KCK) B \right)$$

subject to

$$B^T (K^2 + \lambda K) B = I,$$

where $K \in \mathbb{R}^{n \times n}$ is the kernel matrix, $n$ is the number of image sets, $C = I - L$ in which $L$ is the normalized Laplacian matrix derived from the hypergraph, $B$ is the coefficient matrix for reconstructing the projection in feature space, and $\lambda > 0$ is the regularization parameter.

Kernel canonical correlation analysis (kCCA) (Hardoon et al., 2004) is a widely-used method for dimensionality reduction. It can be shown that kCCA involves the following optimization problem:

$$\max_{B} \quad \text{trace} \left( B^T K (Y^T Y)^{-1} Y^T K B \right)$$

subject to

$$B^T (K^2 + \lambda K) B = I,$$

where $Y$ is the label matrix. Thus, kCCA is a special case of our proposed formulation based on hypergraph.

3.2 A convex formulation

It follows from the theory of reproducing kernels (Schölkopf and Smola, 2002) that the kernel $K$ in Eq. (1) uniquely determines a mapping of the patterns to some feature space. Thus, kernel selection (learning) is one of the central issues in kernel methods. Following the multiple kernel learning framework (Lanckriet et al., 2004a), we propose to obtain an optimal kernel matrix by integrating multiple kernel matrices constructed from various features, that is, $K = \sum_{j=1}^{p} \theta_j K_j$ where $\{K_j\}_{j=1}^{p}$ are the $p$ kernels constructed from various local descriptors and $\{\theta_j\}_{j=1}^{p}$ are the weights satisfying $\sum_{j=1}^{p} \theta_j = 1$. We show that the optimal weights that maximize the objective function in Eq. (1) can be obtained by solving a semi-infinite linear program (SILP) (Hettich and Kortanek, 1993) in which a linear objective is optimized subject to an infinite number of linear constraints. This is summarized in the following theorem (Proof given in the Supplement):
4 RESULTS

In this section, we apply the proposed framework for annotating gene expression patterns. We use a collection of images obtained from the FlyExpress database (Van Emden et al., 2006), which contains standardized and aligned images. All the images used are taken from lateral view with the anterior to the left. The size of each raw image is 128 × 320.

4.1 Experimental setup

We apply nine local descriptors on regular grids of two different sizes on each image. The nine local descriptors are SIFT, shape context, PCA-SIFT, spin image, steerable filters, differential invariants, complex filters, moment invariants, and cross correlation. These local descriptors are commonly used for objection recognition (more details can be found in Mikolajczyk and Schmid (2005)). The sizes of the grids we used are 16 and 32 pixels in radius and spacing (Fig. 2), and 133 and 27 local features are produced for each image, respectively.

It is known that local textures are important discriminative features of gene expression pattern images, and features constructed from filter banks and raw pixel intensities are effective in capturing such information (Varma and Zisserman, 2003). We therefore apply Gabor filters with different wavelet scales and filter orientations on each image to obtain global features of 384 and 2592 dimensions. We also sample the pixel values of each image using a bilinear technique, and obtain features of 10240, 2560, and 640 dimensions.

The resulting features are called “global features”.

After generating the features, we apply the vocabulary-guided pyramid match algorithm (Grauman and Darrell, 2006) to construct kernels between image sets. A total of 23 kernel matrices (2 grid sizes × 9 local descriptors + 2 Gabor + 3 pixel) are obtained. Then, the proposed MKL formulation is employed to obtain the optimal integrated kernel matrix. The performance of kernel matrices (either single or integrated) is evaluated by applying the support vector machine (SVM) for each term and treating image sets annotated with this term as positive, and all other image sets as negative. We extract different numbers of terms from the FlyExpress database and use various numbers of image sets annotated with the selected terms for the experiments.

Precision and recall are two commonly-used criteria for evaluating the performance of multi-label classification systems (Datta et al., 2007). For each term, let Π and Λ denote the indices of patterns that are annotated with this term by the proposed framework and by human curators in BDGP, respectively. Then, precision and recall for this term are defined to be $P = |\Pi \cap \Lambda|/|\Pi|$ and $R = |\Pi \cap \Lambda|/|\Lambda|$, respectively, where $|\cdot|$ denotes the set cardinality. The F1 score is the harmonic mean of precision and recall as $F1 = (2 \times P \times R) / (P + R)$. To measure performance across multiple terms, we use both the macro F1 (average of F1 across all terms) and the micro F1 (F1 computed from the sum of per-term contingency tables) scores, which are commonly used in text and image applications (Datta et al., 2007). In each case, the entire data set is randomly partitioned into training and test sets with ratio 1:1. This process is repeated ten times, and the averaged performance is reported. We report the performance of each individual kernel and compare it with methods based on multi-instance learning on a data set of 10 terms and 1000 image sets in the Supplement.

Results indicate that kernels constructed from the SIFT and PCA-SIFT descriptors yield the highest performance.

4.2 Annotation results

We apply the proposed formulations (star, clique, and kCCA) to combine the various kernel matrices derived from different local descriptors. The performance of multiple kernel learning based on the soft margin 1-norm SVM (SVM1) criterion proposed in Lanckriet et al. (2004a) is also reported. Since the SVM1 formulation is only applicable to binary-class problems, we apply the formulation for each term by treating image sets annotated with this term as positive, and all other image sets as negative. To demonstrate the effectiveness of the proposed formulation for integrating kernels, we also report results obtained by combining the candidate kernels with uniform weight, along with the performance of the best individual kernel (among the 23 kernels) for each data set.

To compare with the existing method proposed in Zhou and Peng (2007), we extract wavelet features from images and apply the min-redundancy max-relevance feature selection algorithm to select a subset of features. As was done in Zhou and Peng (2007), we assign terms to individual images and apply linear discriminant analysis to annotate each image. Note that this setup does not consider the image group information and is the same as the one proposed in Zhou and Peng (2007). The annotation results measured by F1 score and precision and recall are summarized in Tables 1–4.

It can be observed from the results that in terms of both macro and micro F1 scores, the kernels integrated by either star or clique expansions achieve the highest performance on all but one of the data sets. This shows that the proposed formulation is effective in combining multiple kernels and potentially exploiting the complementary information contained in different kernels. For all data sets, the integrated kernels outperform the best individual kernel. In terms of precision and recall, our results indicate that SVM1 and Uniform achieve higher precision than the proposed formulations, while they both yield significantly lower recall. On the other hand, the best individual kernel produces slightly higher recall than the proposed formulations, while it yields significantly lower precision. Note that precision and recall are two competing criteria, and one can always achieve a perfect score on one of them at the price of the other. Hence, the proposed formulation achieves a harmonic balance between precision and recall, as indicated by the F1 scores. Note that BIK can have both higher precision and higher recall than the proposed formulation, since we report the highest precision and the highest recall among all of the candidate kernels separately. Hence, the BIK for precision and recall may not correspond to the same kernel. For all the four measures, the proposed formulations outperform the method proposed in Zhou and Peng (2007) significantly. This shows that the annotation performance can be improved by considering the image group information.

Fig. 3 shows some annotation results obtained by clique expansion for sample patterns in each stage range. Note that the pyramid match algorithm can compute kernels between variable-sized sets of images. Thus, terms can be predicted for image sets of any size. Overall, the proposed computational framework achieves promising performance on annotating gene expression patterns. Meanwhile, we realize that the current framework suffers from some potential limitations. By comparing the BDGP terms and the predicted terms for patterns in stage ranges 7-8 and 9-10, we can
We first select a number of terms and then extract certain number of image sets annotated with at least one of the selected terms. The number of terms used are 10, 20, 30, 40, 50, and 60, and the number of image sets used are 1000, 1500, and 2000 in each case. The first three rows report the F1 scores obtained by kernels combined with star expansion, clique expansion, and CCA, respectively. The fourth row presents the F1 scores achieved by kernels combined with the soft margin 1-norm SVM (SMV) formulation in which an optimal kernel is learned for each term separately. The fifth row shows the F1 scores achieved by kernels combined from the candidate kernels with uniform weights. The performance of the best individual kernel (BIK) over all local descriptors and grid sizes on the same data set is reported in the sixth row. The results obtained by the method proposed in Zhou and Peng (2007) are reported in the last row. The performance shown in this table is the averaged scores over ten random partitions of the entire data set into training and test sets with ratio 1:1.

Table 2. Performance of integrated kernels on gene expression pattern annotation in terms of macro F1 score.

<table>
<thead>
<tr>
<th># of terms</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sets</td>
<td>1000</td>
<td>1500</td>
<td>2000</td>
<td>1000</td>
<td>1500</td>
<td>2000</td>
</tr>
<tr>
<td>Star</td>
<td>0.5661</td>
<td>0.5741</td>
<td>0.5434</td>
<td>0.5436</td>
<td>0.4903</td>
<td>0.4575</td>
</tr>
<tr>
<td>Clique</td>
<td>0.5251</td>
<td>0.5220</td>
<td>0.4876</td>
<td>0.4536</td>
<td>0.5125</td>
<td>0.4926</td>
</tr>
<tr>
<td>kCCA</td>
<td>0.5487</td>
<td>0.5608</td>
<td>0.5323</td>
<td>0.3987</td>
<td>0.4635</td>
<td>0.4477</td>
</tr>
<tr>
<td>SVM1</td>
<td>0.4924</td>
<td>0.5413</td>
<td>0.5333</td>
<td>0.3780</td>
<td>0.4640</td>
<td>0.4356</td>
</tr>
<tr>
<td>Uniform</td>
<td>0.4947</td>
<td>0.5498</td>
<td>0.5418</td>
<td>0.3727</td>
<td>0.4703</td>
<td>0.4480</td>
</tr>
<tr>
<td>BIK</td>
<td>0.5418</td>
<td>0.5430</td>
<td>0.5185</td>
<td>0.4241</td>
<td>0.4515</td>
<td>0.4344</td>
</tr>
<tr>
<td>Z&amp;P</td>
<td>0.3756</td>
<td>0.3810</td>
<td>0.3775</td>
<td>0.2695</td>
<td>0.2759</td>
<td>0.2804</td>
</tr>
</tbody>
</table>

This table shows the performance of each method in terms of micro F1 score. See the footnotes of Table 1 for detailed explanations.

5 CONCLUSIONS AND DISCUSSIONS

In this article, we have presented a computational framework for annotating gene expression patterns of *Drosophila*. We propose to extract invariant features from gene expression pattern images and construct kernels between these sets of features. To integrate multiple kernels effectively, we propose multi-label, multiple kernel learning formulations based on hypegraph. Experimental evaluation shows that the integrated kernels consistently outperform the best individual kernel. Currently, the annotation of patterns by human curators requires multiple passes, and the proposed framework can be used as a preprocessing step whose annotation is further refined by human curators.

In future work, we plan to perform a detailed analysis of the weights obtained by the MKL formulation, and investigate how they are related to the relevance of each kernel. Our experimental results show that features extracted on smaller grids tend to yield better results. However, computational resource limitations prevent the use of a grid size smaller than 16 pixels. We plan to explore ways to overcome this problem. Retrieving gene expression patterns by combining information from images and annotations is an interesting and challenging research issue. The proposed framework can assign a probability of associating each term to each image, producing a probability vector for unannotated images from various high-throughput experiments. Such information can potentially be exploited to facilitate pattern retrieval. Detailed analysis of the annotation results produced by the proposed framework indicates that integration of gene expression pattern images taken from multiple views can potentially improve the annotation performance.
In this case, the current pyramid match algorithms need to be adapted so that only images taken from the same view are matched. It can be seen from the third and fifth images in Fig. 4 that the annotation terms can also be associated with partial patterns in each image. These partial patterns have been removed in our FlyExpress database (Fig. 3), so these terms cannot be predicted correctly by the proposed framework. We plan to explore ways to incorporate these partial patterns in the future.

ACKNOWLEDGEMENTS

We thank Kristen Grauman, John Lee, and Zhi-Hua Zhou for fruitful discussions, Hanchuan Peng for the feature selection code, Kristi Garboushian for editorial support, and Bernard Van Emden for help with access to the gene expression data. This work is supported in part by research grants from National Institutes of Health and National Science Foundation.

REFERENCES


<table>
<thead>
<tr>
<th>Stage range</th>
<th>BDGP terms</th>
<th>Predicted terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>maternal</td>
<td>maternal</td>
</tr>
<tr>
<td></td>
<td>cellular blastoderm</td>
<td>cellular blastoderm</td>
</tr>
<tr>
<td>4-6</td>
<td>trunk mesoderm anlage</td>
<td>anterior endoderm anlage</td>
</tr>
<tr>
<td></td>
<td>posterior endoderm anlage</td>
<td>posterior endoderm anlage</td>
</tr>
<tr>
<td>7-8</td>
<td>trunk mesoderm primordium</td>
<td>trunk mesoderm primordium</td>
</tr>
<tr>
<td></td>
<td>anterior endoderm primordium</td>
<td>anterior endoderm primordium</td>
</tr>
<tr>
<td></td>
<td>posterior endoderm primordium</td>
<td>posterior endoderm primordium</td>
</tr>
<tr>
<td></td>
<td>inclusive hindgut primordium</td>
<td>inclusive hindgut primordium</td>
</tr>
<tr>
<td>9-10</td>
<td>trunk mesoderm anlage</td>
<td>anterior endoderm anlage</td>
</tr>
<tr>
<td></td>
<td>posterior endoderm anlage</td>
<td>posterior endoderm anlage</td>
</tr>
<tr>
<td>11-12</td>
<td>embryonic central brain glia</td>
<td>embryonic central brain glia</td>
</tr>
<tr>
<td></td>
<td>lateral cord glia</td>
<td>lateral cord glia</td>
</tr>
<tr>
<td></td>
<td>neuroblasts of ventral nervous system</td>
<td>procephalic neuroblasts</td>
</tr>
<tr>
<td></td>
<td>procephalic neuroblasts</td>
<td>embryonic central brain neuron</td>
</tr>
<tr>
<td></td>
<td>lateral cord neuron</td>
<td>lateral cord neuron</td>
</tr>
<tr>
<td></td>
<td>lateral midline</td>
<td>lateral midline</td>
</tr>
<tr>
<td></td>
<td>lateral cord glia</td>
<td>lateral cord glia</td>
</tr>
<tr>
<td></td>
<td>embryonic central brain glia</td>
<td>embryonic central brain glia</td>
</tr>
<tr>
<td>13-16</td>
<td>embryonic central nervous system</td>
<td>embryonic central nervous system</td>
</tr>
<tr>
<td></td>
<td>ventral nerve cord</td>
<td>ventral nerve cord</td>
</tr>
<tr>
<td></td>
<td>embryonic central brain neuron</td>
<td>embryonic central brain neuron</td>
</tr>
<tr>
<td></td>
<td>lateral cord neuron</td>
<td>lateral cord neuron</td>
</tr>
<tr>
<td></td>
<td>lateral midline</td>
<td>lateral midline</td>
</tr>
<tr>
<td></td>
<td>lateral cord glia</td>
<td>lateral cord glia</td>
</tr>
<tr>
<td></td>
<td>embryonic central brain glia</td>
<td>embryonic central brain glia</td>
</tr>
<tr>
<td></td>
<td>embryonic central brain</td>
<td>embryonic central brain</td>
</tr>
</tbody>
</table>

Fig. 3. Annotation results for sample patterns in the six stage ranges. BDGP terms denote terms that are assigned by human curators in the Berkeley Drosophila Genome Project (Tomancak and et al., 2002), and predicted terms denote terms predicted by the proposed computational framework.

Fig. 4. The original five images in stage range 13-16 from BDGP. The first and the third images are taken from lateral view; the second and the fourth images are taken from ventral view; the fifth image is taken from dorsal view. Only the first and the third images are used in our experiments shown in the bottom of Fig. 3.


