Automatic Mining of Fruit Fly Embryo Images *

Jia-Yu Pan, André Guilherme Ribeiro Balan†,
Eric P. Xing, Agma Juci Machado Traina†, Christos Faloutsos
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A.
{jypan,agrbalan,epxing,agma,christos}@cs.cmu.edu

ABSTRACT

We present FEMine, an automatic system for image-based gene expression analysis. We perform experiments on the largest publicly available collection of Drosophila ISH (*in situ* hybridization) images, showing that our FEMine system achieves excellent performance in classification, clustering, and content-based image retrieval. The major innovation of FEMine is the use of automatically discovered latent spatial "themes" of gene expressions, *LGEs*, in the whole-embryo context, as opposed to patterns in nearly disjoint portions of an embryo proposed in previous methods.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications — data mining, image databases; J.3 [Life and Medical Sciences]: Biology and Genetics

General Terms: Experimentation

Keywords: embryonic image analysis, Drosophila, gene expression, independent component analysis, Eigen-Embryo

1. INTRODUCTION

In multicellular organisms such as Drosophila and human, many important biological processes, including development and differentiation, are essentially ruled by gene expression activity[4, 5]. For multicellular organisms, gene expressions must be described in a spatio-temporal context, i.e., containing both spatial and temporal dynamics of gene activities.

In situ hybridization (ISH) analysis is an imaging method that reveal the spatial distribution of gene expression in tissues. Such spatial information is indispensable for in-depth analysis of the regulatory and developmental mechanisms in higher eukaryotic organisms [12]. The fast growing "Expression Pattern" database under the Berkeley Drosophila Genome Project (BDGP) now contains over 56,000 digital images of expression patterns of over 3,000 genes, and before long all genes in the Drosophila genome will be covered [1]. As of now, the only mining approach offered by the BDGP, for example, co-expressed genes or spatial (anatomical and histological) annotations of the gene expressions, is based on manual-labeling of the images by a domain expert using a controlled vocabulary [16]. Efforts of automating this process and grounding it on a more objective and robust feature description and distance measure have just begun [10, 14], and the tools available so far are clearly inadequate.

Despite its obvious importance and necessity, there has been little earlier work on automatic analysis and comparison of Drosophila embryo ISH images; recently developments mostly resort to simplistic image mining approaches with limited power [6, 8, 10]. The technique developed by Peng and Myers [14] is more robust and combines similarity measure at both local and global level. Nevertheless, these extant approaches offer limited flexibility for capturing complex gene ISH patterns that are present in a rich database such as the BDGP. For example, it is hard to capture the famous *stripe patterns* of pair rule genes in Drosophila embryo using a mixture of Gaussians. Furthermore, it is not clear how to define similarity functions among genes, to infer their dependencies (but see [13] for some recently developments).

In this paper, we present a novel image mining system, FlyEmbryo Miner (or FEMine), that can automatically extract, transform, compare, classify and cluster gene expression patterns in Drosophila embryos based on raw ISH images. The major innovation of FEMine is the use of automatically discovered latent spatial "themes" of gene expressions, referred to as *LGEs*, which capture basic patterns of gene expressions in the *whole*-embryo context, as opposed to localized patterns in nearly disjoint portions of the embryos that previous methods propose.

2. PROPOSED METHODOLOGY

The work-flow of FEMine contains three major steps: (a) Image processing, in which we perform segmentation and registration, (b) feature extraction with PCA [9] and ICA [7], and (c) data mining, with applications such as classification, clustering and content-based image retrieval. Table 1 summarizes the workflow of our FEMine system. In this

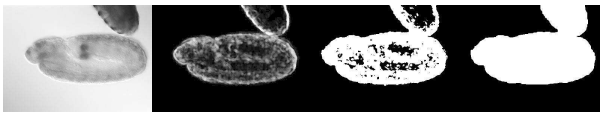
*We would like to thank Dr. Hanchuan Peng for inspiring discussions on aspects of ISH image analysis. This research is supported in part by NSF (grant nos. IIS-0209107, SENSOR-0329549, EF-0331657, IIS-0326322, and IIS-0534205), CAPES-Brazil (grant no. BEX- 1206/05-2), FAPESP-Brazil (grant nos. 03/01769-4 and 2005/04272-9), CNPq-Brazil (grant nos. 471950/2004-1 and 501214/2004-6), the Pennsylvania Infrastructure Technology Alliance, and an NSF CAREER Grant DBI-0546594 for E.P.X.

†Affiliation of A. Balan and A. Traina: Instituto de Ciências Matemáticas e Computaão, Universidade de São Paulo, São Paulo, Brazil.

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

Table 1: Workflow of the FEMine system.

Input: The embryo image database: $\mathcal{D}=\{\mathcal{I}_1,\ldots,\mathcal{I}_N\}$.
Steps:
1. Image preprocessing (Section 2.1):
   For each image $\mathcal{I}_i \in \mathcal{D}$,
   (1.1) Identify and extract the major embryo.
   (1.2) Register the extracted embryo to a common form.
   Result: A registered image database: $D=\{I_1,\ldots,I_N\}$.
2. Extract the latent gene expression (LGE) patterns (Section 2.2):
   (2.1) Compute the Eigen-Embryos (Section 2.2.1).
   (2.2) Compute the LGEs (Section 2.2.2).
   Result: M' LGEs (biological meaningful image templates).
3. Data mining using FEMine (Section 3):
   (3.1) Represent each image $I_i \in D$ with LGEs (Eq. 1).
   (3.2) Tasks supported by FEMine:
   (a) Classification (Section 3.3).
   (b) Clustering (Section 3.4).
   (c) Content-based image retrieval (Section 3.5).



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

Figure 1: Embryo extraction steps. (a) Input image; (b) Local variance intensities; (c) Binary image after thresholding; (d) Final result after "holes filling".

section, we describe the first two of the three steps of the work-flow, and the third step will be described later.

2.1 Image Processing

The image processing stage is necessary in order to provide a standardized image database for gene expression pattern comparison. To deal with issues such as noise, occlusion, and inconsistent orientation in the embryo ISH images, we perform three image processing steps: *embryo extraction*, *main embryo isolation*, and *image registration*.

2.1.1 Embryo extraction

With very few exceptions, the embryos and the background have significantly different local texture properties. Embryos have a rougher texture with high local variance, while the background, a watery solution, has smooth tonal variations, which means pixels with low local variance. We calculate the variance of pixel intensity in a 3x3 window centered at each pixel of the image (Figure 1(b)), and set the pixel as foreground if the value is above a fixed threshold value. A result of this thresholding is shown at Figure 1(c). It is quite common to have embryo-pixels assigned as background, mainly at the center region of them. Thus, after obtaining the binary image, we apply a morphological binary operator to "fill the holes" inside the embryos' region. Figure 1 shows an example of each embryo extraction step.

2.1.2 Isolating the main embryo

An ISH image taken during a typical embryogenesis experiment usually contains a few dozens of embryos, with the most interesting embryo located at the center of the image. The embryos in an image may be touching the main

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

Figure 2: The shrink-expand method: (a-d) region erosion (e) two regions found.

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

Figure 3: Region growing: (a) Initialization by the shrink-expand method. (b-e) region growing.

embryo, or even occluding it (Figure 1(a)). To extract the main embryo, our first attempt was to use the well known "watershed transform" to partition the foreground region of the binary image. However, due to the noisy borders and concave shapes of the embryos, the watershed approach with a bad initial state tends to "over-segment" the embryos (i.e., too many regions at the final result).

To remedy this, we propose a novel approach to provide an initial state for the watershed algorithm. Basically, we perform a *shrink-expand* processing of the foreground region, i.e., first the region is continuously eroded until we find two separated regions, as shown in Figure 2. The two partitions of the foreground region are then the initial state for the watershed flooding algorithm. The algorithm "grows" back the regions, until they touch again, creating a watershed, as shown in the Figure 3.

For images with more than two embryos. We apply our "shrink-expand" algorithm recursively over the foreground region, keeping only the center-most region at each recursion step until the "shrink-expand" algorithm gives only one region. Figure 4 shows an example of recursive partitioning.

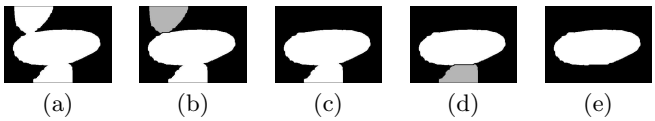
2.1.3 Image registration

Embryos extracted from the previous step could have different position, orientation, scale and shape in the images. For a better comparison between patterns in the extracted embryos, we perform an image registration step to transform the images, so that the comparison can be performed regardless their original position, orientation, scale and shape.

Given that embryos at the same developmental stages do not show considerable differences on shape (i.e., can be mostly rectified by an affine transformation), we select the method proposed by Thvenaz, Ruttimann and Unser [15]. The chosen method uses an efficient variant of the Marquardt-Levenberg algorithm for non-linear optimization and an elaborate multi-resolution structure to speed up the registration process. We configure the method so that after registration, each embryo will have an ellipsoidal shape, predefined size, and with its major axis aligned horizontally. The final product of our image processing are gray-level images with 352×160 pixels. Some examples are presented in Figure 5.

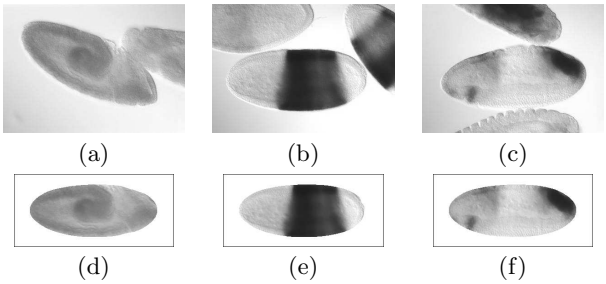
2.2 Latent Gene Expression Patterns (LGEs)

The main goal of our work is to extract biological meaningful patterns that captures the overall spatial gene expression patterns in an embryo ISH image. These patterns would also provide a representation of the embryo images, which is as compact as possible and suitable for clustering,



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

Figure 4: Recursive partitioning to separate the main embryo.



(a) (b) (c)

(d) (e) (f)

Figure 5: Processing the ISH images of Drosophila embryos: (a-c) original images (d-f) results.

classification and content-based retrieval. We refer to such patterns as "latent gene expression" (or, LGE) patterns.

We propose a two-step method to discover the latent spatial (gene expression) patterns from the embryo ISH images. The first step is to summarize the 1,763 images in our database into a few, more manageable, e.g., M=10, *typical* images. At the second step, from these typical images, we extract the *latent LGE patterns* (or *templates*) that compose the spatial gene expressions presented in the images.

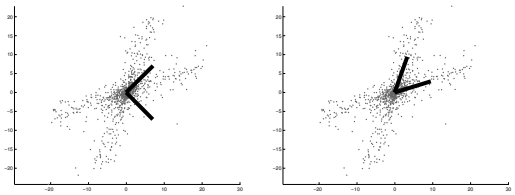
For the first step, we propose to use the Principal Component Analysis (PCA) [9] for discovering the typical images that summarize the database. For mining the LGE patterns at the second step, we propose to use the Independent Component Analysis (ICA) [7]. As we will show later, the independent "templates" of the embryo images found by ICA efficiently describe the global spatial gene expressions in the images. Next we present the details of these two steps.

2.2.1 Principal Component Analysis Pre-processing

We apply PCA to discover typical images in the image database. After our image processing and registration, every embryo is presented as a 352×160 pixel gray-scale image, and we propose to consider it as a point N-dimensional space, where N=352×160 is the number of pixels. Our database of 1,763 (registered) embryo images can be envisioned as a cloud of points in this N-dimensional space.

Computationally, let us consider a data set $D = [I_1, ..., I_M]$ of embryo images, where each image I_i $(i = 1, \ldots, M)$ is viewed as a 1-by-N vector of N=352×160 pixels, and is a point in the N-dimensional space. PCA computes the eigenvectors of the covariance matrix of the data (DD^T). Because these eigenvectors can also be treated as points in the N-dimensional pixel space, we can also visualize each of them as an 352×160 image. We refer to these images as *Eigen-Embryos*. Since these eigenvectors account for the major variations among the data points [9], Eigen-Embryos can be considered as the "typical" images that summarize the characteristics of images in the database.

In our experiments, we select only a reduced number M' ($M' < N$) of Eigen-Embryos, those associated with the highest eigenvalues. The value of M' is chosen so that the retained eigenvalues maintain 90% of the "total energy" (sum of squared eigenvalues).



(a) (b)

Figure 6: Artificial dataset presenting a X-like distribution. The dark lines are: (a) PCA basis vectors (Eigen-Embryos) (b) ICA basis vectors (LGEs).

2.2.2 Mining the Independent LGE Patterns

The Eigen-Embryos generated by PCA *together* provide a good description of the entire ISH image database [9]. However, individual Eigen-Embryo may not capture correct characteristic in the database. Figure 6 illustrates this situation using an artificial dataset – a cloud of 2-dimensional points in an 'X'-like shape. Figure 6(a) shows the two eigenvectors (or Eigen-Embryos): they are orthogonal to each other and together they can describe the major variations of the point set. However, each of them fails to capture the "X-like" distribution pattern of the data. A better set of vectors are shown in Figure 6(b), computed by ICA.

Therefore, we propose to use the Independent Component Analysis (ICA) to find a better set of image *templates*, which are more biologically significant and can show the latent gene expressions in the *whole*-embryo context. An intuitive way to describe ICA is through the *blind source separation* problem (*bss*): given a set of observed signals, ICA attempts to decompose them into a set of independent signals, without explicit knowledge about the signals (i.e. blind). The classical cocktail party problem is an example of *bss*. Consider being in a cocktail party which has several simultaneous conversations, music and noise. The properties of these sound sources are not known, but in most cases, they are not related and independent to each other. Given several different observations of the ambient sound, for example, the sound captured from different microphones located around the room, ICA takes advantage of the independent property of the sound sources and can recover the original conversations.

To relate this concept to our ISH images, Eigen-Embryos that summarize the ISH images in the database can be viewed as the microphone recordings. Our goal is to extract the hidden conversations – the latent gene expression patterns in the images.

Formally, let the M'-by-N matrix X be the collection of the M' Eigen-Embryos (each row is a Eigen-Embryo), and let the M'-by-N matrix S be the set of LGE patterns to be discovered. If B is the unknown M'-by-M' matrix that specifies the mixing of the latent patterns, then the "cocktail party" model is the following

$$X = BS.$$

Given a matrix X, ICA will compute the corresponding matrices B and S. Each row of S corresponds to a LGE and can be visualized as an image template (Figure 8). In our experiments, we use an implementation of ICA named *FastICA* [7].

Each image in the database can be represented using the LGEs. The *LGE-representation* I_{LGE} of an image I is com-

puted, by applying the following projection

$$I_{LGE} = S^{-1} I, \quad (1)$$

where S is the matrix formed by the LGEs as column vectors. We will also call I_{LGE} the "*LGE embedding*" of an image I.

3. EXPERIMENTAL RESULTS

In this section, we present the experimental results of FEMine on the following issues: (1) automatic processing of embryo images; (2) qualitative evaluation of the "Eigen-Embryo" and "LGE" patterns; and (3) the performance of classification, clustering, and content-based retrieval of ISH images using the LGE-representation.

3.1 Automatic Processing of Embryo Images

We download all the ISH images of Drosophila embryos in developmental stages 4-6 from the BDGP database. This dataset contains 8,566 ISH images of highly variable quality: images may contain multiple embryos or incomplete embryos (Figure 5, top panels). Using the image processing module described in Section 2.1 (which includes embryo extraction, isolation and registration), we recover 6,800 high quality, 8-bit gray-level and 352×160-pixel ISH images, each containing a single embryo (or occasionally a single piece of the embryo, due to incomplete coverage of the original image) and no background (Figure 5, bottom panel). Thus, our embryo extraction rate is nearly 78%, significantly higher than the roughly 30% yield (personal communication from Dr. H. Peng) by previous simpler methods.

The 8,566 images record the expression of roughly 2000 genes. For the following experiments, we prepare a data set by manually select at most 2 images per gene, keeping only the images with the most relevant gene expression patterns. After this screening, our experimental database contains images of 1,763 different genes.

3.2 LGEs versus Eigen-Embryos

From our database of 1763 processed embryo images, we build a training set with 127 handpicked images containing a variety of salient gene expression patterns. Then we apply FEMine to extract the Eigen-Embryos and the LGEs of these images.

Figure 7 shows the top 10 Eigen-Embryos that amounts to about 90% of the total energy. A visual inspection suggest that Eigen-Embryos correspond to periodic spatial patterns over an embryo image, each with a different spatial frequency. Together, the Eigen-Embryos are able to reconstruct the numerical signals in the embryo images. However, each Eigen-Embryo does not correspond to biologically interpretable spatial patterns of gene expressions, such as that of the gap genes and the well-known stripe structures of the pair-rule gene expressions.

On the other hand, the LGEs perform much better in capturing biologically significant patterns. Figure 8 shows the 10 LGEs extracted from the top 10 Eigen-Embryos. A number of typical early developmental expression pattern of the Drosophila genes are captured in the LGEs automatically, such as the segmentally repeated pattern, the anterior and posterior patterns, and so on. In the following experiments, we focus on LGEs only. With the 10 extracted LGEs, each of the 1,763 images can be represented as a 10-dim feature vector, using the LGE-embedding (Eq. 1). Each 10-dim image vector is also to unit length, to reduce the variations in pixel intensity among the images.

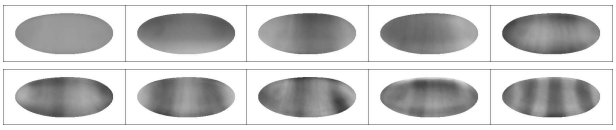


Figure 7: Eigen-Embryos

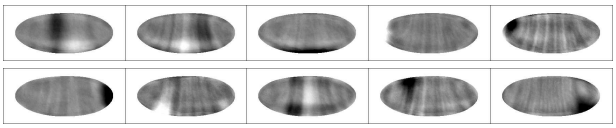


Figure 8: LGEs

3.3 ISH Pattern Classification

To provide an objective and quantitative validation of the usefulness of LGE-based representations of ISH images, we first conduct a classification experiment. The goal is to show that the LGE-based representation preserves essential features of ISH images which can be picked up by a classifier to distinguish images of different type. We experiment with two kinds of classifiers: the support vector machine (SVM) [2, 3] and the k-nearest neighbor (k-NN) classifier [11]. In our experiments, we observe that SVM classifiers always perform better than a k-NN classifier, and therefore only the results of SVM classifiers are reported.

We manually label ISH images for the classification experiment and construct 7 major classes of patterns (minor patterns with fewer than 8 images are not used in our experiment). Images in the same class either have the same body part annotation given by the biologist, or they have similar visual presentation. Among the 7 classes, the number of images in a class ranges from 16 to 25. Totally, there are 131 images in 7 classes. Figure 9 shows the representative images for the 7 classes in our experiments. For example, the images in class 1 have the pattern of "anterior endoderm", and those in class 3 show the pattern of "mesoderm". Class 4 contains images that show the body and tail of a fruit fly, while class 5 contains the interesting patterns which are called "pair rule".

We apply a standard SVM (i.e., the libsvm package [3]) for multiway classification. We experiment with various kernels, including the radial basis kernel, polynomial kernel, and linear kernel, and different parameter settings.

The goal of the classification experiment is to evaluate the representation ability of the LGEs. For a fair evaluation, we reduce the influence from the classifier by searching a good parameter setting for the SVM classifier. The search of a good parameters for the SVM classifier is done by uniformly sampled from the parameter space. For each candidate kernel/parameter configuration, we conduct 5-fold cross-validation on a subset of the database (in total, 96 images randomly sampled from each class). The configuration with the best cross-validation performance is chosen.

The next step is to applied the SVM with selected configuration to the entire database (the entire set of 131 images), to obtain a realistic evaluation of the expression power of LGEs. Again, 5-fold cross-validation is performed to obtain the classification accuracy, and we report the mean and standard deviation of the classification accuracy.

Our experiments show that the LGE representation of the ISH images achieves good classification accuracy, indicating good representation power of the LGEs. Table 2 shows

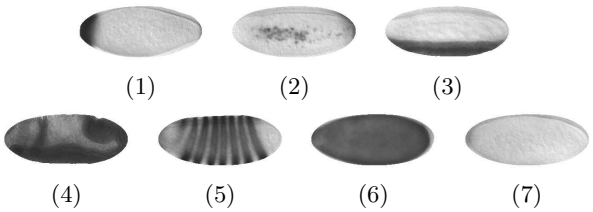


Figure 9: Sample ISH images of the 7 classes

Table 2: Classification with the SVM classifier

Kernel	Radial Basis	Polynomial	Linear
Accuracy	82% (± 9%)	**85%** (± 7%)	77% (± 9%)

the best cross-validation accuracy of a SVM classifier using different choices of kernel functions (with best parameters found). We find that using a SVM with polynomial kernel achieves the best mean classification accuracy 85.42% with a standard deviation 7.22%. In fact, the LGE-representation consistently gives good classification accuracy regardless of the kernel choice.

Table 3 shows the confusion matrix of the classification of the 7 ISH image classes. In the table, each row corresponds to a class, and each column corresponds to a predicted class. The number at the i-th row and the j-th column shows the percentage of images in class i are classified as class j. Therefore, the values in the cells on the diagonal correspond to correct prediction. In this case, images in classes 3 and 6 are always classified correctly (100%). Also, all classes (except class 4) are classified with more than 75% accuracy.

3.4 ISH Clustering

We perform a clustering analysis of the ISH images, to see whether we can discover genes with correlated expressions, and major expression patterns in our database. Among the 1763 images, there are several ones that do not show significant body part ISH staining – we refer to these images as *"uninteresting"* images. We propose a two-stage clustering procedure that first removes the uninteresting images via an initial clustering, and then re-clusters the rest of the images that exhibit significant diversity and variability ISH staining.

In the first stage of our clustering procedure, we apply either K-means or the Gaussian Mixture Model (GMM) [11] to the whole dataset (i.e., 1763 images), and we empirically set the total number of clusters K to be K=20 and K=30 in two different trials. By visual inspection, we exclude several large clusters of "uninteresting" images (not shown, but similar to the ones representing class 6 and class 7 in Figure 9). Thus we obtain a "filtered" dataset of 268 ISH images which potentially contain the interesting patterns of ISH staining in the embryonic body parts.

In the second stage of our clustering procedure, both K-means and the GMM are applied on the filtered dataset of 268 images to find K=20 clusters. Figure 10 shows sample images from the clusters found by the Gaussian mixture model. Again for brevity, we only show images from 10 clusters (out of total 20 clusters). The clustering obtained via K-mean is similar and not reported here. Remarkably, each of the clusters we discover appeared to correspond to a unique gene expression pattern revealed by ISH. In particular, we notice that earlier blob-based approaches without the LGE embedding can not pick up delicate patterns such as the segmental repeats of the pair-rule genes (e.g., cluster 9 in Figure 10).

Table 3: Confusion matrix of the best SVM result

	Predicted						
Truth	1	2	3	4	5	6	7
1	**75%**	-	-	-	-	8%	17%
2	12%	**76%**	-	-	12%	-	-
3	-	-	**100%**	-	-	-	-
4	-	-	-	**59%**	8%	8%	25%
5	-	8%	-	-	**75%**	-	17%
6	-	-	-	-	-	**100%**	-
7	-	6%	-	-	-	-	**94%**

In our experiments, the clustering algorithm (either K-means or GMM) is repeated for 1,000 times with random initialization. Among results from repeated K-means runs, we pick the one with *minimum distortion*, that is, the sum of distances from data points to its cluster centroid is minimum. Among the results of the GMM runs, we keep the one with the highest likelihood given the data set.

3.5 Content-Based Image Retrieval (CBIR)

We would like to emphasize that FEMine is designed to be a user friendly expert system for image-based gene expression analysis. An important function offered by our FEMine system is the content-based ISH-image retrieval for an arbitrary query image. In Figure 11, we present a screen-shot of FEMine, to highlight its interface. Notice that it can display tables containing the thumbnails of the retrieved images, as well important annotations of each image, such as the name of the image file, the name of the gene expressed, and the body part description contained in BDGP, provided by specialists. It also provides an automatic and instantaneous evaluation of the query results, by plotting the correspondent Precision vs. Recall curve, when the ground truth of a retrieval query is available.

4. CONCLUSION

We have developed FEMine, for automatic pattern mining in a Drosophila embryonic ISH image database. FEMine offers a wide range of tools for analysis, starting from basic image processing, feature extraction, and high-level pattern recognition functionality, such as pattern classification, image clustering, and content-based image retrieval. One of the main novelties of FEMine is the introduction and automatic extraction of "LGEs" — the latent *whole-embryo* gene expression themes in Drosophila embryos. It is our belief that FEMine can offer an integrated toolbox to comprehend and analyze existing and rapidly growing repositories of Drosophila embryo ISH images for a variety of scientific research problems.

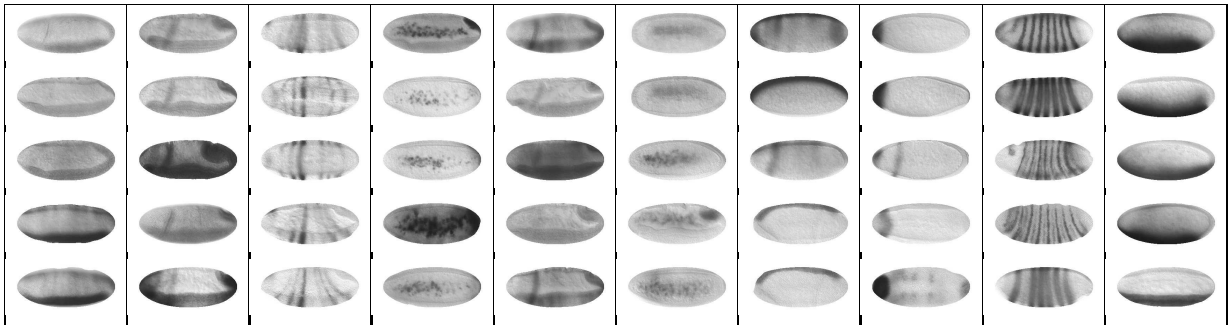
5. REFERENCES

[1] BDGP. Patterns of gene expression in Drosophila embryogenesis, 2005.

[2] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001.

[4] E. H. Davidson. *Genomic Regulatory Systems*. Academic Press, 2001.



(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
18 images	25 images	5 images	25 images	13 images	23 images	27 images	22 images	25 images	18 images

Figure 10: Clusters of ISH images using GMM: 10 (out of total 20) clusters are displayed. Also shown are the number of images in each cluster.

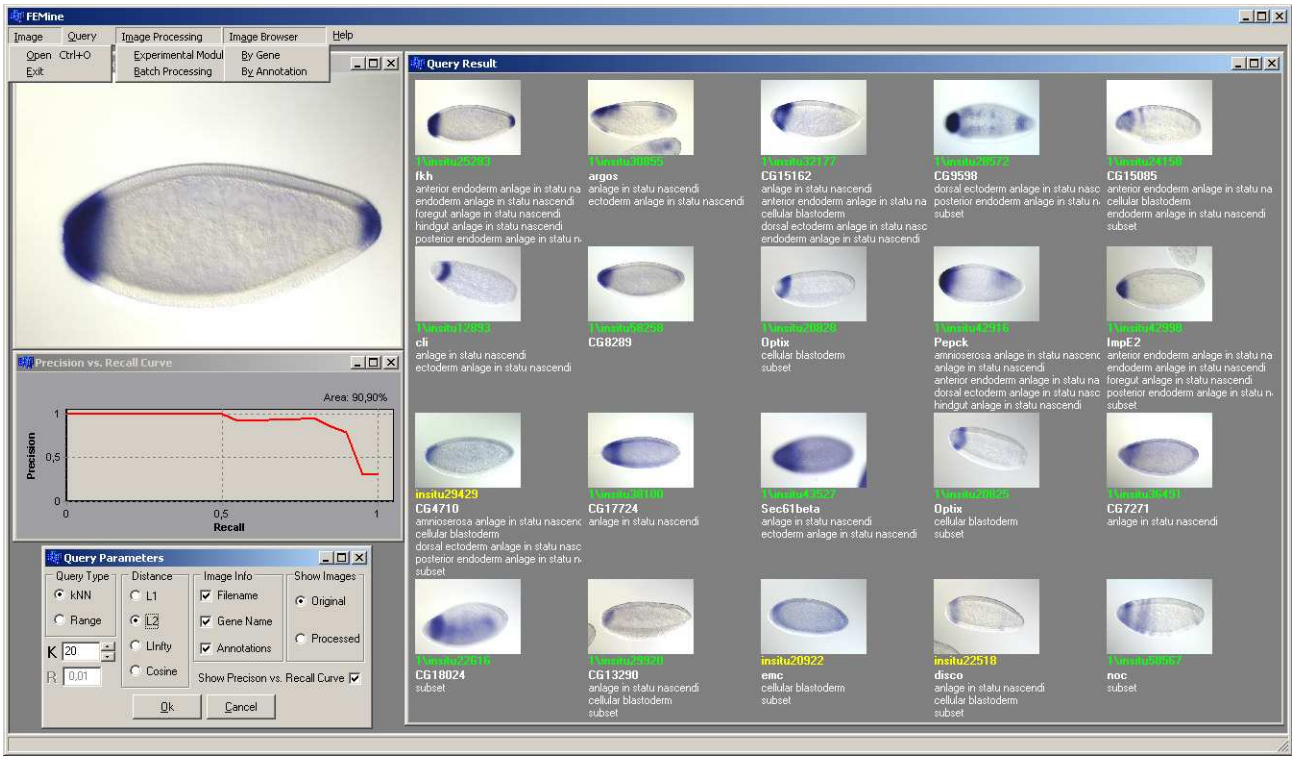


Figure 11: The FEMine GUI for content-base image retrieval. The query image has the pattern of an "anterior endoderm". Hits are shown with the original (pre-processed) images.

[5] S. F. Gilbert. *Developmental Biology, Seventh Edition*. Sinauer Associates, 2003.

[6] R. Gurunathan, B. V. Emden, S. Panchanathan, and S. Kumar. Identifying spatially similar gene expression patterns in early stage fruit fly embryo images: binary feature versus invariant moment digital representations. *BMC Bioinformatics*, 5(202):13, December 2004.

[7] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.

[8] K. Jayaraman, S. Panchanathan, and S. Kumar. Classification and indexing of gene expression images. In A. G. Tescher, editor, *Proceedings of SPIE - Applications of Digital Image Processing XXIV*, volume 4472, pages 471–481, 2001.

[9] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[10] S. Kumar, K. Jayaraman, S. Panchanathan, R. Gurunathan, A. Marti-Subirana, and S. Newfeld. BEST: a novel computational approach for comparing gene expression patterns from early stages of Drosophila melanogaster development. *Genetics*, 162(4):2037–47, 2002.

[11] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[12] H. Montalta-He and H. Reichert. Impressive expressions: developing a systematic database of gene-expression patterns in Drosophila embryogenesis. *Genome Biol*, 4(2):205, 2003.

[13] H. Peng, F. Long, M. Eisen, and E. Myers. Clustering gene expression patterns of fly embryos. In *IEEE 2006 International Symposium on Biomedical Imaging (ISBI 2006)*, 2006.

[14] H. Peng and E. Myers. Comparing *in situ* mRNA expressions of Drosophila embryos. In *Proc. 8th Annual Int. Conf. on Research in Computational Molecular Biology (RECOMB 2004)*, pages 157–166, 2004.

[15] P. Thvenaz, U. E. Ruttimann, and M. Unser. A pyramid approach to subpixel registration based on intensity. *IEEE Transactions On Image Processing*, 7(1):27–41, January 1998.

[16] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. Celniker, and G. Rubin. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol*, 3(2):14, 2002.