Comparing *In situ m*RNA Expression Patterns of *Drosophila* Embryos

Hanchuan Peng
Lawrence Berkeley National Lab,
University of California, Berkeley
+1-919-225-3401
penghanchuan@yahoo.com

Eugene W. Myers
Computer Science Division,
University of California, Berkeley
+1-510-643-8220
gene@eecs.berkeley.edu

ABSTRACT

In situ staining of a target *m*RNA at several time points during the development of a *D. melanogaster* embryo gives one a detailed spatio-temporal view of the expression pattern of a given gene. We have developed algorithms and software for analyzing a database of such images with the goal of being able to identify coordinately expressed genes and further our understanding of *cis*-regulatory control during embryogenesis. Our approach combines measures of similarity at both the global and local levels, based on Gaussian Mixture Model (GMM) decompositions. At the global level, the observed distribution of pixel values is quantized using an adaptive GMM decomposition and then quantized images are compared using mutual information. At the local level, we decompose quantized images into 2-dimensional Gaussian kernels or "blobs" and then develop a blob-set matching method to search for the best matching traits in different pattern-images. A hybrid scoring method is proposed to combine both global and local matching results. We further develop a voting scheme to search for genes with similar spatial staining patterns over the time course of embryo development. To evaluate the effectiveness of our approach, we compare it with several global image matching schemes and a controlled vocabulary method. We then apply our method to 4400 images of 136 genes to detect potentially co-regulated genes that have similar spatio-temporal patterns, using expert-annotation to evaluate our results.

Categories and Subject Descriptors

H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval – *search process.* I.4.7 [**Image Processing And Computer Vision**] Feature Measurement – *feature representation, size and shape.* I.4.10 [**Image Processing And Computer Vision**]: Image Representation – s*tatistical, hierarchical, morphological.* I.5.4 [**Pattern Recognition**] Applications – signal processing. J.3 [**Life And Medical Sciences**] *Biology and genetics.*

RECOMB'04, March 27-31, 2004, San Diego, CA, USA.
Copyright 2004 ACM 1-58113-755-9/04/0003…$5.00.

General Terms

Algorithms, Design.

Keywords

In situ hybridization, Embryogenesis, Gene expression, Gaussian mixture model, Image matching, *Drosophila*.

1. INTRODUCTION

Understanding the roles of genes and their complicated relationships is one of the central themes of genome research. One popular approach is based on the analysis of microarray-gene-expression data (e.g. [5]), which currently can only be applied to a sizeable collection of cells such as an entire embryo or "tissue" sample. Such experiments thus reveal only the average expression levels over the sample, failing to observe any potentially pivotal spatial patterns of expression. It is possible that two completely unrelated genes could have similar global expression levels through a time series, but have completely different spatial patterns. To understand co-expression in any multi-cellular tissue or organism, it is clearly desirable to examine the spatial patterns of a gene's expression.

The *in situ* hybridization technique localizes specific *m*RNA sequences in morphologically preserved tissues/cells by hybridizing the complimentary strand of a nucleotide probe to the sequence of interest. *In situ* embryogenesis Staining Pattern (SP) images of *Drosophila melanogaster* are now available (e.g. [1][19][17]), for example, that of the Berkeley *Drosophila* Genome Project (BDGP) (www.fruitfly.org) [1]. At this time, this growing database contains 33699 images of 1711 genes. These SP images show where and when a target gene is expressed during embryogenesis.

It is believed that comparison of these SPs can be a very powerful way to understand the roles of genes and solve many related problems such as finding co-regulated genes [2][9][6][15]. The current method of classifying these images in the BDGP database is the manual assignment of terms from a controlled ontology vocabulary to each image [16]. This approach depends entirely on experts who are familiar with *Drosophila* embryogenesis. Alternatively, in this paper we propose image analysis algorithms to automatically compare *Drosophila melanogaster* embrogenesis SPs, with the goal of finding potentially co-regulated genes.

There are several possible ways of automatically matching SP images, including (1) global matching, (2) local feature (trait) matching, and (3) hybrid methods combining both global and local information. Global matching involves evaluating the similarity of the entire SP of one embryo to that of another, giving a ranking of similarity. Local matching involves finding portions (traits) of two SPs that are similar in shape and intensity variation. In this paper, we present a hybrid-Gaussian-Mixture-Model (GMM) paradigm to combine both global and local matching. This novel method is called hybrid-GMM-matching.

Our approach consists of four parts: image preprocessing (§2), global GMM decomposition and matching (§3), local GMM decomposition and matching (§4), and a hybrid method to detect genes with the similar spatio-temporal SPs (§5). A simple illustration can be seen in Fig.1: we use the image preprocessing module to extract the regions covered by an embryo (Fig.1 (b)) from input images (Fig.1 (a)), and register these embryos (Fig.1 (c)) so that they can be effectively compared. A global-GMM-decomposition method identifies and models the entire staining pattern (Fig.1 (d)). Our local-GMM-decomposition represents an SP as a set of blobs (Fig.1 (e)), upon which SP traits can be better matched.

Despite the fact that there exist many matching methods for general image retrieval, there is little earlier work in automatic analysis and comparison of embryogenesis SP images. For a dataset of about 900 embryo images of the early developmental stage, Kumar et al [9] binarized the embryo images and used a ratio of overlapping pixels as the score to measure the similarity between SPs. It can be classified as a global matching method. Other possible global matching methods include global correlation coefficient matching (or other similarity/distance scores), and histogram and shape matching (as used in color- and shape-based image retrieval [10][7]). In §6, we will compare our new methods and several global matching schemes, as well as the controlled-vocabulary annotation method of [16].

In §7, we apply the hybrid method to 4400 SP images to demonstrate the strength of the new method in finding potential co-regulated genes that have similar spatio-temporal SPs during embryo development. The automatic analysis results are also compared to the manual annotations.

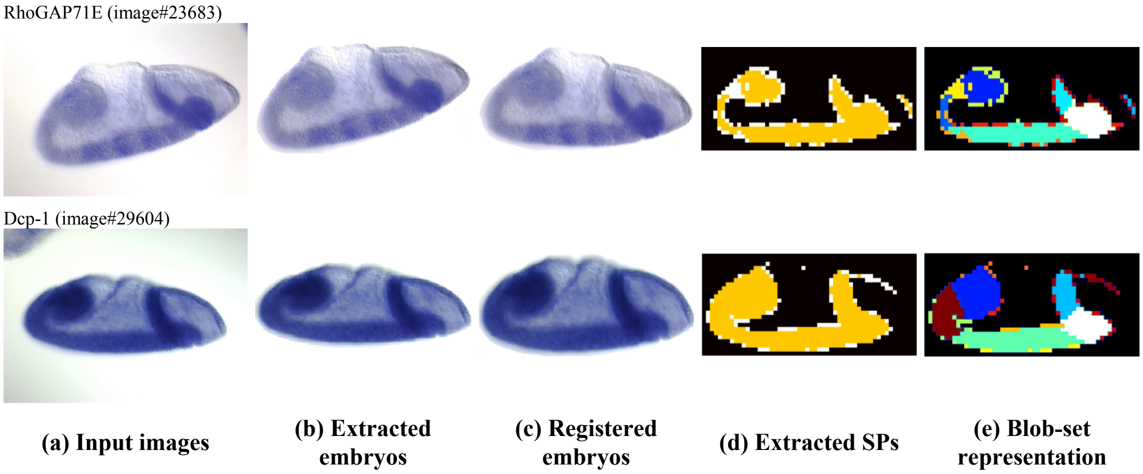


Figure 1. Steps of our method in building representations of embryogenesis Staining Patterns (SP). Each row corresponds to a SP image of a gene whose name and image number are given in (a). We first use basic image processing techniques to obtain standardized embryos (c). Then we use GMM decompositions to build both global representations (d) and local representations (e) of SPs. In (a)-(c), the darker the stain, the stronger the *in situ* expression of the respective gene. Different colors in (d) indicate different expression levels detected by clustering pixels based on intensity or color (see §3). Different colors in (e) indicate different SP traits detected by clustering pixels based on spatial locations (see §4).

2. IMAGE PREPROCESSING

Image preprocessing, including embryo extraction and registration, is a common module needed in any image comparison methods. For example, a simpler preprocessing method was also used in [9].

2.1 Embryo Extraction

The embryogenesis image-acquiring protocol requires that in a qualified SP image, there is a major centered embryo. The embryo extraction step is to segment an input image to separate the main embryo from the background, which presumably contains no information. Typically, the image background (water) often has shadows and the embryo boundary is often fuzzy. Hence, a single threshold on image-pixel intensity is insufficient to extract the embryo.

We note the image background and the embryo region have different local texture properties, i.e. the standard deviation of the embryo region is much larger than that of the background region. This is because the reflectance of the water-background is almost uniform everywhere and the illumination gradually changes; on the contrary, the central embryo has rich variation of the reflectance, and the illumination around the image-center is almost uniform.

Hence, for each image-pixel, we calculate the standard deviation of the local window (e.g. 3×3) around it. The pixel is binarized to "foreground" if the deviation is larger than a predefined threshold (e.g. 2), otherwise to "background". Binarization puts most embryo pixels as "foreground" and most background pixels "background". Next, a simple 8-neighbor-connectivity region-growing method is used to find the contour of the embryo. Every pixel within the contour is considered to be in the extract embryo. Two examples of extracted embryos are shown in Fig. 1 (b).

With the above method, the major embryo region that presumably covers the center of the image can be effectively extracted. For example, for the 4400 SP images used in §7, we visually examine the quality of the extracted embryos and find the above method can consistently produce satisfactory embryo regions, as long as an input image does not contain multiple touching embryos.

2.2 Embryo Registration

Embryo registration involves arranging the embryos so that they can be compared directly at the pixel level. This step is critical for global matching of embryos. For local matching, registration is also important because it provides a standard space where local features can be measured with the same "ruler" and identified more easily.

Our registration module performs an affine transform [7][8], and intensity rescaling. First the longest axis is detected using a standard principal component analysis method [7]. Then the embryo is rotated and scaled so that the longest axis is horizontal and its extent is a preset size (e.g. 300 pixels wide and 200 pixels high). Then pixel values are linearly transformed so the observed values span the interval [0,255]. Two examples of registered embryos are shown in Fig. 1 (c).

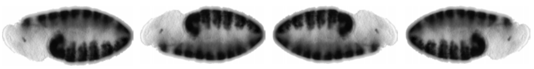


Figure 2. An example of the same embryo with four different orientations.

Note that this process does not determine the anterior - posterior or dorsal - ventral orientation of the embryo. Thus in what follows, whenever we compare a pair of embryos X and Y we consider flipping X vertically, horizontally, and both as in Fig. 2, and take the best correspondence over the four comparisons as the orientation-invariant similarity between X and Y.

Although it is possible to consider more sophisticated registration methods (e.g. automatically locate landmarks and warp to a preset template embryo), our visual inspection confirms that the above method is effective in producing meaningful registration results for the 4400 images used in §7.

3. A GLOBAL GMM MATCHING METHOD

The most direct way of global matching is pixel-by-pixel embryo comparison. Kumar et al [9] used the ratio of overlapping foreground pixels in binarized SP images. Analogous schemes that do not require binarization of pixel values include taking Euclidean distance, correlation coefficient, etc, across corresponding pixels.

A second type of global matching is to compare the pixel intensity distributions of embryos. The intuitive method is to compare the histograms of two embryos. However, histogram matching ignores the fact that expressions occurring at quantized levels vary from image to image and does not compare the *in situ* information in embryos.

Our global Gaussian-Mixture-Model (GMM) matching method combines both pixel comparison and the distribution comparison by first dynamically decomposing the intensity distribution into a number of staining levels using GMMs (§3.1), and then using a mutual information measure between quantized images (§3.2). This allows us to probe if two SPs have similar *in situ* distributions. By modeling only the *in situ* pixel-distributions, the unexpressed embryo regions are excluded from consideration. Like noise removal, this leads to better matching of the most interesting features of the SPs.

3.1 Staining Level Extraction Based on GMM Decomposition

In an embryo, there are typically several distinct expression levels and each is assumed to generate a roughly Gaussian distribution of pixel values centered around the respective mean expression level. Thus we expect the distribution of pixel values, $h^{I\text{-}Embryo}(c)$, for c in [0,255] to be a mixture of Gaussians, i.e.

$$h^{I-Embryo}(c) = \sum_{k=1}^{K} u_k G_k(c), \qquad (1)$$

where c is the pixel intensity (or color), K is the unknown number of stain levels, $G_k(c)$ is the kth Gaussian kernel to represent the color, and u_k is the weight of the respective kernel. Note that if we have 3-component color information, then G_k is 3-dimensional Gaussian kernel, and if we have grayscale information, then G_k is a 1-dimensional kernel. Given this assumption of distribution, we seek to find the number of levels K and their Gaussians.

With the model in Eq. (1), we want to find the best parameters to solve the following posterior maximization problem,

$$\max P(h^{I-Embryo} \mid D^{I-Embryo}), \qquad (2)$$

where $D^{I\text{-}Embryo}$ is the part of embryo data being considered. For an assumed K, we can use the Expectation-Maximization (EM) method to solve Eq. (2) and find the optimal pixel-clustering results. A description of the standard EM algorithm is omitted here but can be found in many textbooks [11][18].

Because the stronger *in situ* expression levels correspond to the darker stains in embryo, among the K Gaussian kernels, the one with the smallest mean value most likely represents an expression level. This idea can be explicitly described as the following algorithm, which progressively identifies all expression staining levels.

(1) Initially, the data $D^{I\text{-}Embryo}$ contains all pixels in an embryo. The mean intensity of all embryo pixels is m.

(2) The EM method is used to find a GMM with K kernels. Denote the kernel with the smallest mean value as G_1^*, and the mean intensity of all pixels represented by this kernel as m_1. If $m_1 < m$, then we consider G_1^* to represent a stain level.

(3) We assign all pixels represented by G_1^* as having the stain level 1, eliminate them from the pixel set $D^{I\text{-}Embryo}$. We repeat step (2) to find a new GMM-decomposition whose ker-

nel with smallest mean is G_2^* with mean intensity m_2. If $m_2 < m$, then G_2^* is considered to represent a new stain level. Otherwise the algorithm stops, the reason being that if the most probable (darkest) kernel does not represent a stain level, then other less likely pixels will not be stain.

(4) Eventually, this algorithm returns N^{Global} Gaussian kernels, each of which is obtained in a run of the GMM decomposition.

The result of this procedure is that we identify N^{Global} positive staining levels each modeled by a Gaussian G_i^* and weight u_i that in aggregate give us a model of the distribution of *in situ* pixel values, $h^{I-Stain}$:

$$h^{I-Stain} = \sum_{i=1}^{N^{Global}} u_i G_i^* . \qquad (3)$$

Two SP examples are shown in Fig. 1 (d), where different colors are used to mark the most probable staining level at each pixel. Note that by modeling only the *in situ* distributions of the SP, but not the whole embryo, we exclude irrelevant information. This improves both accuracy and robustness of the analyses that follow.

We call the above procedure the global-GMM-decomposition algorithm. There are three remaining issues. First, we used a preset number of kernels, K, in each run of the EM algorithm in Step 2. Because we aim to accurately estimate only the kernel with the smallest mean out of K kernels, a convenient approach is to simply set K as intermediate value like 4 or 5. Setting K too small introduces many false-positive pixels in G_i^*, while setting K too large will exclude many false-negative pixels. A mid-ranged K leads to kernels with reasonably compact and homogeneous distribution. An alternative is to adaptively choose K to maximize the posterior probability of a K-part GMM, e.g. $\text{argmax}_K P(h^{I-Embryo} \mid D^{I-Embryo})$ [14]. We have tried both methods and found they give comparable results.

Second, our method considers only the distribution of pixel intensity. We refine this method by integrating local spatial information. A general way to do this is to find small patches of pixels within which pixel-intensity is relatively homogeneous, and then use the mean intensity of the patch to represent all pixels within this patch. Instead of local image scanning to seek arbitrarily shaped small patches, we use a simple method of generating small square patches (e.g. 9×9 pixels), which is effectively the image downsampling. Compared to other possible methods such as Gaussian smoothing, this method does not consider overlapping patches.

Finally, by considering the embryo as a collection of homogeneously colored patches, the number of data points is greatly reduced. For example, a 9×9 patch reduces the number of data points to less than 2% of the original amount. Accordingly, the speed of EM method in Step 2 is much faster than the original one and the slow speed of EM is no longer an issue for our application.

3.2 Matching with Mutual Information

Eq. (3) is a distribution function of SP pixels at specific spatial locations. During the decomposition algorithm each non-background pixel is assigned to a quantized expression level when it is subtracted from the remnant distribution in Step 3. We can then ask questions like "do two embryos have similar *in situ* stain distributions of expression levels?"

Given two quantized SPs X and Y, we use mutual information to examine whether their *in situ* stain distributions have strong dependency on each other. Mutual information [4] has the form $s^{MI}(X,Y) = \Sigma_{X,Y} P(X,Y)\log\{P(X,Y)/[P(X)P(Y)]\}$, where $P(X,Y)$ is the joint distribution, $P(X)$ and $P(Y)$ are the marginal distributions. Mutual information describes the statistical dependency of SPs. The larger $s^{MI}(X,Y)$, the stronger dependency between X and Y. Because $s^{MI}(X,Y)$ is upper-bounded by the minimum of the entropies of X and Y (i.e. $\min(H(X),H(Y))$), and lower-bounded by 0, we use the normalized mutual information (range is [0,1]) as the global similarity score of any SP-pairs,

$$s^{Global}(X,Y) = \frac{2 \cdot s^{MI}(X,Y)}{H(X)+H(Y)} . \qquad (4)$$

Effective calculation of mutual information often requires X and Y are categorical variables with a small number of states. Note that the extracted SP using our global GMM decomposition has only a limited number, N^{Global}, of expression levels. Hence, this representation is suitable for mutual information calculation. From a perspective of signal/image processing, our global GMM decomposition is an adaptive quantization scheme.

4. A LOCAL GMM MATCHING METHOD

In §3, the intensity distribution of an extracted SP is modeled using GMMs (Eq. (3)). However, spatially the SP is irregularly distributed in the 2-dimensional image plane (i.e. it is irregularly shaped, as shown in Fig. 1 (d)). It is often hard to directly compare the shapes of entire SPs. In addition, under many circumstances it is not necessary to compare the whole SP against others, because co-regulated genes do not necessarily have exactly the same stains at every spot. Instead, they might share only some local traits. This raises the issue of how to represent a stained image in term of traits that will serve as the basic components in matching.

We develop a local GMM decomposition method to model an SP as a collection of spatially distributed traits or "blobs" (§4.1). This provides a way to identify the traits of any SP. We also propose a score for SPs based on comparing the blob decomposition (§4.2).

4.1 Blob Extraction Based on GMM Decomposition

We assume that the spatial distribution of a SP can be modeled as the union of a series of component blocks. Within each block, the spatial coordinates of SP are compact enough to comprise a small spatially distributed cluster. A simplest way is the rectangular blocks. While this has been shown to be very useful in other applications (e.g. [13][12]), it is less optimal for SPs because most SP components are closer to oval than rectangle. Hence, we propose to model the SP spatial distribution $h^{S-Stain}(x,y)$ (x and y are spatial coordinates) using GMM:

$$h^{S-Stain}(x,y) = \sum_{k=1}^{K} v_k G_k(x,y), \qquad (5)$$

where v_k is the coefficient of the kth 2-dimensional Gaussian kernel $G_k(x,y)$.

The EM algorithm is used again to find the parameters of Eq. (5), so that the following posterior is maximized:

$$\max P(h^{S-Stain} \mid D^{S-Stain}), \qquad (6)$$

where the data $D^{S-Stain}$ contains the pixel coordinates (x,y) for the current expression level. We use the adaptive method [14] to determine the optimal number of kernels, K.

We do the local-GMM-decomposition for every stain level in the SP. Note that for each stain level, the intensity-distribution of the respective pixels has been assumed homogeneous. Thus, there is no need to incorporate the pixel-intensity information in Eq. (6).

Local-GMM-decomposition leads to a new representation of the SP, consisting of a set of blobs (ellipses). We call it the blob-set representation B, similar to the blobworld representation for general image retrieval [3]. One natural way to represent each blob is a Gaussian kernel, as suggested in Eq. (5). However, a more effective way is to represent the blob using the pixels covered by this blob. Hence, each blob b is a function of the stain level of the associated pixels, g, and the spatial locations of all these pixels, l. Suppose that there are N blobs in a SP, the blob-set representation B has the following form,

$$B = \bigcup_{i=1}^{N} b(g,l). \qquad (7)$$

Since there are N^{Global} stain levels in a SP, and for the ith stain-level we can solve Eq. (6) and obtain K_i blobs, the total number of blobs in an SP is $N = \Sigma_{i=1}^{N^{Global}} K_i$.

Compared to the blobworld representation developed in [3] which models the joint color-texture-location distribution of pixels using Gaussian kernels, our blob-set representation is based only on the pixel location information, of each different stain level. While simpler our model has several advantages for this application domain: (1) it allows generation of both global and local representations and more flexible SP matching schemes, (2) it is computationally more efficient, and (3) it is empirically more accurate because GMM decomposition degrades with increasing dimensionality.

4.2 Matching with Blob Sets

The blob-set representation provides rich information for a variety of SP-analysis. Here we only consider using this representation to examine the similarity of SPs. Suppose that we have two blob-sets B^X and B^Y of two SPs X and Y (with N^X and N^Y blobs, respectively). For every blob b_i^X in B^X, we search in B^Y the blob with the largest similarity, denoted as b_{i*}^Y. The respective similarity is denoted as $s^{Blob}(b_i^X, b_{i*}^Y)$. Similarly, for every blob b_j^Y in B^Y, we search in B^X the blob b_{j*}^X with the largest similarity $s^{Blob}(b_j^Y, b_{j*}^X)$. We define the following similarity to score how two blob-sets resemble each other,

$$s^{Local}(B^X, B^Y) = \sum_{i=1}^{N^X} s^{Blob}(b_i^X, b_{i*}^Y) + \sum_{j=1}^{N^Y} s^{Blob}(b_j^Y, b_{j*}^X) \qquad (8)$$

where the similarity of two blobs, $s^{Blob}(.)$, is defined below (Eq. (9)). Note that Eq. (8) allows a blob in one set be matched to multiple blobs in the other set. The symmetric form makes the score more robust to local variations.

If two blobs are spatially overlapping, they are similar to each other to some degree. The larger the overlapping, the more similar these two blobs. Hence, the ratio of overlapping pixels out of the total area of the two blobs is an index of their similarity. Additionally, two overlapping blobs are more similar if their stain levels g^X and g^Y are closer. These constraints can be written as the similarity between two blobs in Eq. (9),

$$s^{Blob}(b^X, b^Y) = [1 - \frac{|g^X - g^Y|}{256}] \cdot \frac{\Omega(l^X \cap l^Y)}{\Omega(l^X \cup l^Y)} \qquad (9)$$

where $\Omega(.)$ is the operator to calculate area, l^X and l^Y are respective spatial locations of blob pixels.

5. HYBRID MATCHING METHOD

5.1 Combining Global and Local Matching

We propose the multiplication in Eq. (10) to combine the global-matching score s^{Global} and the local matching score s^{Local}. This avoids the problem of the different scales of s^{Global} and s^{Local}.

$$s^{Hybrid} = s^{Global} \cdot s^{Local}. \qquad (10)$$

In image matching, an efficient way is to first use the global matching to filter out the unlikely matching SPs for a query SP, then the local matching is done in a significantly smaller pool of candidate SPs. Eq. (10) is still used to score the SPs that pass the global matching. We sort the scores from large to small and produce a ranking list; the top ranking SPs are most similar to the query SP.

5.2 A Voting Method to Detect Genes with Similar Spatio-Temporal Patterns

Genes co-regulated by the same transcriptional factor are likely to have the similar spatio-temporal SPs over a range of embryo developmental stages. Based on the score in Eq. (10), we design the following voting method to detect genes that have similar spatio-temporal SPs.

Voting means summarization of the image matching results of all developmental stages. The 16 developmental stages of *Drosophila* embryogenesis are often categorized as 6 main phases (stages 1-3, stages 4-6, stages 7-8, stages 9-10, stages 11-12, and stages 13-16) [1]. Thus, we do a phase-by-phase spatial matching. Suppose the input data consist of SP images of a query gene Q and a pool of candidate genes L in which the spatio-temporally similar patterns are searched. For each phase, we search the matching genes in L whose SPs are similar to those of the query Q. Since that for any phase usually there are multiple images for a gene, we say that two genes have matching patterns if they have at least one similar SP-pair for this phase. The similarity of this SP-pair is used to indicate the similarity of these two genes at this phase (if there are multiple matching pairs, then use the largest similarity value). As a result, for each phase we obtain a ranking list of matching genes. These ranking lists are summarized to produce the spatio-temporal matching list. Obviously, if a candidate gene appears at top of multiple ranking lists of multiple phases, this gene is highly probable to have similar spatio-temporal patterns with the query gene *Q*.

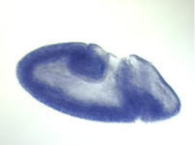
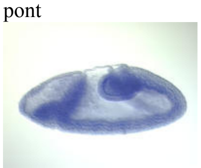
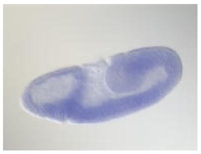
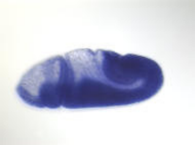
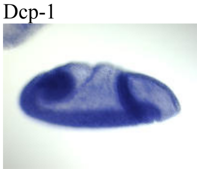
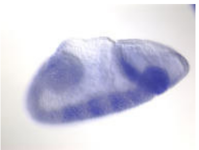
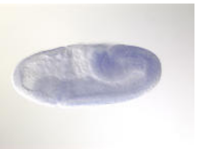
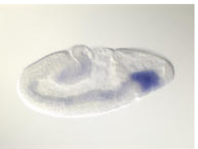
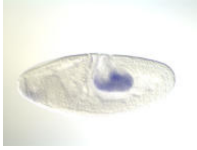
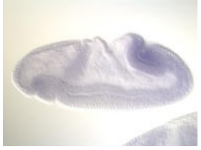
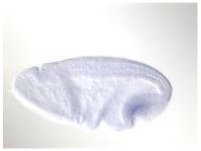
Acf1	pont	mam	Slbp	Dcp-1	RhoGAP71E
					
ventral ectoderm anlage posterior endoderm anlage hindgut anlage anterior endoderm anlage trunk mesoderm anlage head mesoderm anlage	ventral ectoderm anlage posterior endoderm anlage hindgut anlage anterior endoderm anlage trunk mesoderm anlage head mesoderm anlage	ventral ectoderm anlage posterior endoderm anlage hindgut anlage trunk mesoderm anlage head mesoderm anlage	ventral ectoderm anlage posterior endoderm anlage hindgut anlage anterior endoderm anlage trunk mesoderm anlage procephalic ectoderm anlage head mesoderm anlage	ventral ectoderm anlage posterior endoderm anlage anterior endoderm anlage	ventral ectoderm anlage posterior endoderm anlage anterior endoderm anlage
cl	CG33099	GATAe	CG5525	CG6051	
					
posterior endoderm anlage trunk mesoderm anlage head mesoderm anlage	posterior endoderm anlage anterior endoderm anlage trunk mesoderm anlage head mesoderm anlage	posterior endoderm anlage anterior endoderm anlage	posterior endoderm anlage	posterior endoderm anlage	

Figure 3. 11 SP images with the expert-annotation "posterior endoderm anlage". Each image is arbitrarily selected for a gene, just for the experimental illustration. The input images are 24-bit RGB images (1520×1080 pixels), which are converted to grayscale image in matching. The extracted embryos are registered as 400×200 images. We underline the annotation "ventral ectoderm anlage" to indicate the most reasonable matching images when the query-image belongs to gene Acf1.

6. EXPERIMENTAL COMPARISON OF DIFFERENT MATCHING METHODS

Before addressing a real application in §7, we use some small datasets to investigate whether or not our methods improve matching compared to several other scoring methods. Due to space limitations, we only include one example in this section.

We use SP images of the top genes returned by the BDGP image anatomy ontology server, ImaGO [1], for the term "posterior endoderm anlage", which is present in embryonic developmental stages 7-8, or phase 3. These genes and their SP images are shown in Fig. 3. For image matching we convert the blue color images to grayscale.

The expert-annotation results extracted from the BDGP database [1] (also listed in Fig.3) are used as the "ground truth" to evaluate the results. Since these 11 images share the same annotation "posterior endoderm anlage", they must have some extent of similarity. However, as indicated by the additional annotations for every image, other traits of these images are not necessary the same. Hence, the most similar SP images should have the largest number of common annotations. For example, if we use the image of gene Acf1 as the query, according to the annotation results in Fig.3, the top ranking images should belong to the following genes: pont, mam, Slbp, Dcp-1, and RhoGAP71E, because they share the most annotations such as "ventral ectoderm anlage". The image of pont should be in the first place in the matching list, since it shares the same set of 6 annotations with Acf1.

Fig. 4 shows the image matching results for the query Acf1. The 2nd to 4th columns are results using three whole-embryo matching methods. There are some clear problems in these results. First, some images' ranks are incorrect. For example, the image of gene mam ranks 10th using the OverlapRatio method (ratio of overlapping staining pixels, as used in [9]), 8th using the CorrCoef method (correlation coefficient), and 11th using whole-embryo mutual information scoring (the 4th column). These results are clearly inconsistent with what are expected from the expert-annotations (Fig. 3) or a simple visual inspection, because the mam image is similar to Acf1, and should have a relatively high ranking. Second, similarity scores are very close to each other in value, so that for a given query, it is unclear where the border is between similar and dissimilar images. For example, for the OverlapRatio method, scores for the 3rd to 11th best match range from 0.62 to 0.51. It is hard to use such undifferentiated scores for retrieval or clustering.

The reason for these unsatisfactory results is that scoring the correspondence in the background and minor variations in expression level clusters creates unwanted noise. Our new methods, as shown in the 3 rightmost columns, overcome these problems. For example, in the 5th column, the global GMM matching method extracts SPs from embryos and compares them using mutual information (as in §3.2). Unlike the whole-embryo mutual information matching results in the 4th column, the global-GMM-matching results do not have apparent inconsistency with the expert annotations. In addition, the small scores at the bottom rows of Fig.4 (e.g. 0.02 between GATAe and Acf1) clearly indicate that these images are dissimilar from the query. These show that the explicit extraction of an SP model and subsequent comparison of the models is superior to whole-embryo pixel comparisons.

The local GMM matching (blob-set matching, in the 6th column) performs comparably to global GMM matching. However, because local traits are identified and compared in this method, it is able to tell that the mam image has more similar traits to Acf1 than Dcp-1 does. This is consistent with the expert-annotations.

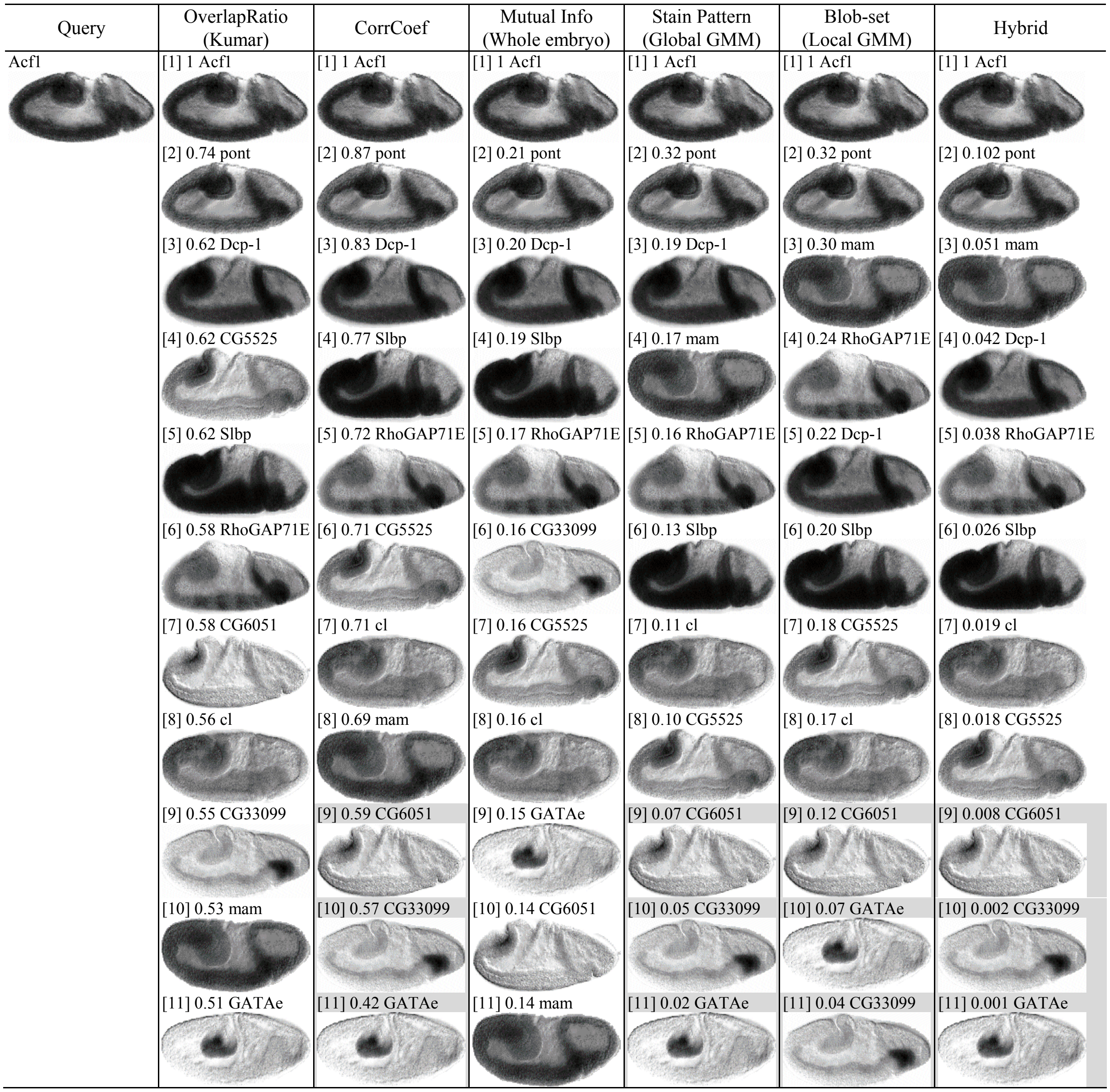


Figure 4. Matching results for the query Acfl. Each column gives the ranking list for a method. On top of the each image, the string shows information about the match of this image to the query: the number in brackets is the ranking; the next number is the similarity score between this image and the query; and the gene name is shown last. The background of some of the lower cells is marked gray to indicate that their similarity with the query is not as significant as that of the images above them. Note that we also put the query image of Acfl itself in the ranking and it is the first-matched image for every method. The respective scores are provided as the baseline to normalize the similarity between other images and the query.

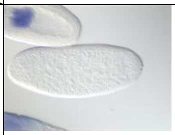
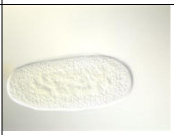
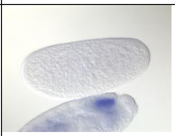
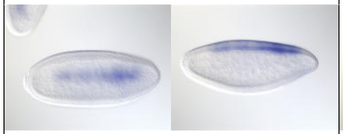
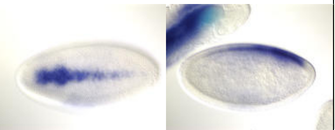
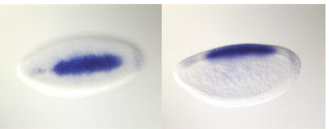
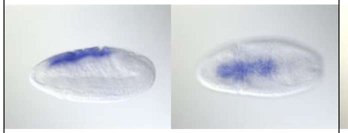
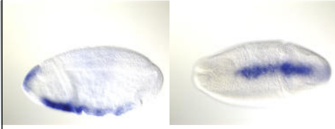
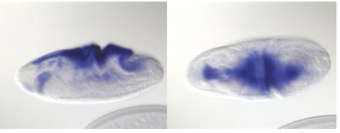
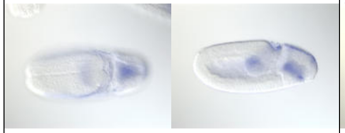
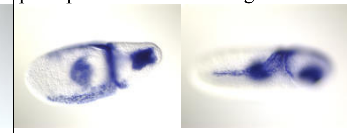
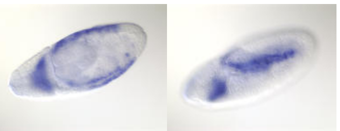
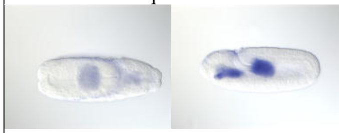
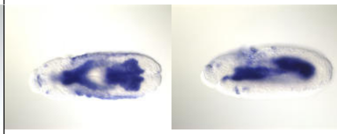
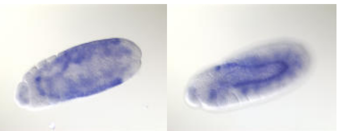
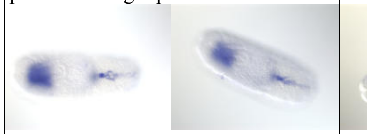
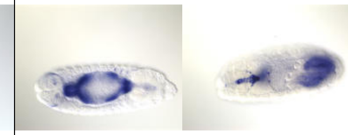
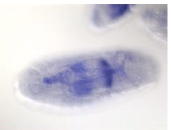
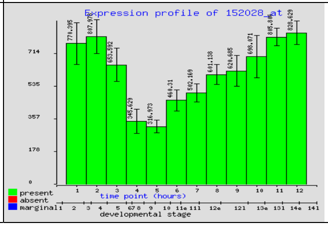
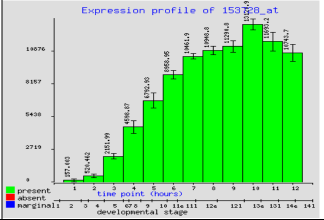
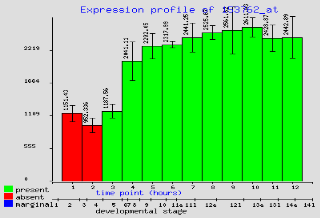
Gene Phase		CG3136	Ance	ush
1	stage1-3	 maternal		
2	stage4-6	 anlage in statu nascendi cellular blastoderm ♥ subset ♦♥	 subset ♦	 cellular blastoderm ♥ subset ♥ yolk nuclei
3	stage7-8	 procephalic ectoderm primordium dorsal ectoderm anlage ♦♥ procephalic ectoderm anlage ♦♥ dorsal ectoderm	 hindgut anlage anterior endoderm anlage dorsal ectoderm anlage ♦ procephalic ectoderm anlage ♦	 dorsal ectoderm anlage ♥ procephalic ectoderm anlage ♥ head mesoderm anlage
4	stage9-10	 anterior endoderm primordium ♦ posterior endoderm primordium ♦ external foregut primordium proventriculus primordium dorsal ectoderm primordium ♥	 anterior endoderm primordium ♦ inclusive hindgut primordium posterior endoderm primordium ♦	 dorsal ectoderm primordium ♥ head mesoderm P2 primordium
5	stage11-12	 anterior midgut primordium ♦ dorsal epidermis primordium ♥ proventriculus primordium posterior midgut primordium ♦	 anterior midgut primordium ♦ posterior midgut primordium ♦ Malpighian tubule primordium	 antennal primordium1 dorsal epidermis primordium ♥
6	stage13-16	 amnioserosa ♦ embryonic/larval endocrine system ♦ embryonic proventriculus ♦	 amnioserosa ♦ embryonic/larval endocrine system ♦ lymph gland embryonic hindgut embryonic proventriculus ♦	 sensory nervous system primordium ventral midline embryonic head epidermis need new term(sensory system head) embryonic dorsal epidermis
Microarray expression level		 	 	

Figure 5. An example of the detected genes CG3136 (query), Ance, and ush, which have similar spatio-temporal patterns. For each gene at each phase, one or two representative embryo images are shown, followed by the expert-annotations (extracted from [1]). A "♦" is used to mark the common annotations between genes CG3136 and Ance, and a "♥" for those between CG3136 and ush. For comparison, the microarray expression levels during the embryogenesis are also shown at the bottom.

Our hybrid GMM matching method (the right-most column) gives the best results. The rankings of pont, mam, Dcp-1, etc are consistent with the expert-annotations, as well as both the global- and local-GMM-matching results. The similarity score shown in brackets on top of each image is a product of the respective scores of global- and local-GMM-matching. For example, for mam the score is 0.051, which is the product of 0.17 (the global-GMM-matching score) and 0.30 (the local-GMM-matching score). We see that the hybrid method presents a nice balance of the global and local matching schemes.

All of our three methods have the strength to differentiate dissimilar SPs, as indicated by the three bottom rows in Fig. 4. For example, for hybrid-GMM-matching, there is a clear scale-gap of scores between the 8th match (CG5525, score=0.018) and the 9th match (CG6051, score=0.008). Compared to the relative minor differences of score-scale of the OverlapRatio method (or the other two whole-embryo matching methods), our new methods indicate genes of the 2nd ~ 8th matches share significant stain features with the query gene, and the genes ranked worse than the 8th match have only partial common stain features with the query gene. The large difference of score-scale is very useful in clustering similar images. These partially shared stain-traits can also be detected using the local-GMM-matching, as we plan to show in a future work.

Experiments of this section indicate that our new methods, especially the hybrid method, are good for detecting similar embryogenesis SPs.

7. FINDING CO-REGULATED GENES

Based on the hybrid GMM matching method, we use the voting scheme in §5.2 to find genes for which the spatio-temporal expression patterns are similar. We use 4400 images of 136 genes from the BDGP website [1]. These genes have the largest number (greater than 30 each) of images in the database. The images are down sized to 400×200 pixels for the matching. Different sizes are also tested with comparable results.

An example of our preliminary analysis is given in Fig. 5, where the query is gene CG3136. Two genes, Ance and ush, are found to have similar spatio-temporal patterns for 5 or 6 embryo developmental phases. Their similarity has not been previously reported in the literature to our knowledge.

We visually compare the images of each developmental phase for the returned gene-pairs, and find that they do have similar expression staining patterns in almost every phase, as seen in Fig. 5. This visual evidence is further confirmed by the expert-annotations from the BDGP database: CG3136 shares 1~3 annotations with Ance (marked using a "♦") for any of phases 2~6, and 1~2 annotations with ush (marked using a "♥") for each of phases 2~5. For example, for phase 3, the three genes have two common annotations "dorsal ectoderm anlage" and "procephalic ectoderm anlage". Accordingly, our method finds several well-matched images as shown in Fig. 5. For phase 1, the images are also very similar, but the annotations for Ance and ush are missing in the BDGP database.

The large agreement between our results and the expert-annotations indicates that our hybrid-GMM-matching and voting method is a meaningful way to detect genes with the similar spatio-temporal SPs. This further suggests this method may be of utility in finding potentially co-regulated genes.

In Fig. 5, we also show the microarray expression levels of the three genes during embryo development (the horizontal axis) [1]. It is clear that there is little correlation between the microarray expression levels of CG3136 and Ance, though they have similar *in situ* expression patterns. This indicates a simple comparison of microarray expression levels is not sufficient or suitable for detecting potentially co-regulated genes.

8. ACKNOWLEDGMENTS

We thank the BDGP group in Lawrence Berkeley National Laboratory for providing the invaluable data online, and Dr. Fuhui Long in Duke Univ. for discussions about image analysis and probabilistic matching. We also thank the anonymous reviewers for helpful comments. This work is supported by National Institutes of Health (NIH), under contract No. R01 GM70444-01.

9. REFERENCES

[1] BDGP: *Berkeley Drosophila Genome Project*, http://www.fruitfly.org.

[2] Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B., Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome, *Proc Natl Acad Sci USA*, 99 (2002), 757-762.

[3] Carson, C., Belongie, S., Greenspan, H., and, Malik, J., Blobworld: image segmentation using expectation-maximization and its application to image querying, *IEEE Trans Pattern Analysis and Machine Intelligence*, 24, 8 (2002), 1026-1038.

[4] Cover, T., and Thomas, J., *Elements of Information Theory*, New York: Wiley, 1991.

[5] De Gregorio, E., Spellman, P.T., Rubin, G.M., and Lemaitre, B., Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays, *Proc Natl Acad Sci USA*, 98, 22 (2001), 12590-12595.

[6] Gieseler, K., Wilder, E., Mariol, M.C., Buratovitch, M., Berenger, H., et al., DWnt4 and wingless elicit similar cellular responses during imaginal development, *Dev. Biol.*, 232 (2001), 339-350.

[7] Gonzalez, R.C., and Woods, R.E., *Digital Image Processing*, (2nd Ed), Addison-Wesley, 2002.

[8] Jain, A., *Fundamentals of Digital Image Processing*, Prentice-Hall, 1986.

[9] Kumar, S., Jayaraman, K., Panchanathan, S., Gurunathan, R., Marti-Subirana, A., and Newfeld, S., BEST: a novel computational approach for comparing gene expression patterns from early stages of Drosophlia melanogaster develeopment, *Genetics*, 169 (2002), 2037-2047.

[10] Long, F.H., Zhang, H.J., and Feng, D.G., Fundamental of content-based image retrieval, In *Multimedia Information Retrieval and Management*, Feng, D., Siu, W.C., Zhang, H. J. (Eds.), Springer-Verlag, (2003), 1-12.

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

[12] Peng, H.C., Long F.H., Chi Z., and Siu W., Template document image matching based on component block list, *Pattern Recognition Letters*, 22, 9 (2001), 1033-1042.

[13] Peng, H.C., Long, F.H., and Chi, Z., Document image recognition based on template matching of component block projections, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25, 9 (2003), 1188-1192.

[14] Roberts, S.J., Husmeier, D., Rezek, I., and Penny, W., Bayesian approaches to Gaussian mixture modeling, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20, 11 (1998), 1133-1142.

[15] Takaesu, N.T., Johnson, A.N., Sultani, O.H., and Newfeld, S.J., Combinatorial signaling by an unconventional Wg pathway and the Dpp pathway requires Nejire (CBP/p300) to regulate dpp expression in posterior tracheal branches, *Dev. Biol.*, 247 (2002), 225-236.

[16] Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., and Rubin, G.M., Systematic determination of patterns of gene expression during Drosophila embryogenesis, *Genome Biology*, 3, 12 (2002).

[17] Tsai, C.C., Kramer, S.G., and Gergen, J.P., Pair-rule gene runt restricts orthodenticle expression to the presumptive head of the Drosophila embryo, *Dev. Genet.*, 23 (1998), 35–44.

[18] Webb, A., *Statistical Pattern Recognition*, Arnold, 1999.

[19] Zhang, H., and Levine, M., Groucho and dCtBP mediate separate pathways of transcriptional repression in the Drosophila embryo, *Proc. Natl. Acad. Sci. USA*, 96 (1999), 535–540.