6-15-2010

SPEX2: automated concise extraction of spatial gene expression patterns from Fly embryo ISH images.

Kriti Puniyani
Carnegie Mellon University

Christos Faloutsos
Carnegie Mellon University, christos@cs.cmu.edu

Eric P. Xing
Carnegie Mellon University, epxing@cs.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/machine_learning

Part of the Theory and Algorithms Commons

Published In
Bioinformatics, 26, 12, 47-56.
SPEX²: automated concise extraction of spatial gene expression patterns from Fly embryo ISH images

Kriti Puniyani, Christos Faloutsos and Eric P. Xing

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

Motivation: Microarray profiling of mRNA abundance is often ill suited for temporal–spatial analysis of gene expressions in multicellular organisms such as Drosophila. Recent progress in image-based genome-scale profiling of whole-body mRNA patterns via in situ hybridization (ISH) calls for development of accurate and automatic image analysis systems to facilitate efficient mining of complex temporal–spatial mRNA patterns, which will be essential for functional genomics and network inference in higher organisms.

Results: We present SPEX², an automatic system for embryonic ISH image processing, which can extract, transform, compare, classify and cluster spatial gene expression patterns in Drosophila embryos. Our pipeline for gene expression pattern extraction outputs the precise spatial locations and strengths of the gene expression. We performed experiments on the largest publicly available collection of Drosophila ISH images, and show that our method achieves excellent performance in automatic image annotation, and also finds clusters that are significantly enriched, both for gene ontology functional annotations, and for annotation terms from a controlled vocabulary used by human curators to describe these images.

Availability: Software will be available at http://www.sailing.cs.cmu.edu/

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

In multicellular organisms such as the metazoa, many important biological processes such as development and differentiation depend fundamentally on the spatial and temporal control of gene expression (Davidson, 2001; Gilbert, 2003). To date, the molecular basis and regulatory circuitry underlying metazoan gene regulation remains largely unknown. Numerous algorithmic approaches have been attempted to infer ‘networks’ of regulatory elements from high-throughput experimental data, such as microarray profiles (Dobra et al., 2004; Ong, 2002; Segal et al., 2003), ChIP-chip genome localization data (Bar-Joseph et al., 2003; Harbison et al., 2004) and protein-protein interaction data (Causier, 2004; Giot et al., 2003; Kelley et al., 2004), based on formalisms such as Bayesian networks (Cowell et al., 1999) or graph mining (Tanay et al., 2004). However, a key deficiency of these approaches is that they rely heavily on high-throughput biological data like microarrays that only capture average behaviour of the genes and proteins in a large cell population from, e.g. a cell culture, a dissected tissue or even a homogenized whole animal. For multicellular organisms such as Drosophila and human, gene expressions must be described in a spatiotemporal context, which reveals the histological specificities and temporal dynamics of the activities of the gene. Such information is not available from the standard whole-animal microarray data which record only the average expression of each gene over all cells in the body, nor is it easily obtainable from ‘tissue-specific’ microarray assays using advanced micro-dissection and cell-sorting techniques (Fig. 1).

In situ hybridization (ISH) assay is an imaging method to visualize mRNA expression in tissues and cells without homogenizing the specimens to be analyzed and therefore retains the original histological context of gene expression. Such information is indispensable for in-depth analysis of gene regulation networks, developmental mechanisms and oncogenic processes in higher eukaryotic organisms (Montalta-He and Reichert, 2003). Systematic profiling of ISH images capturing gene expressions over the entire span of Drosophila embryogenesis are now being undertaken at a whole-genome scale, offering an unprecedented opportunity for investigators to compare the spatio-temporal behavior of genes and begin assembling realistic pictures of gene regulatory networks underlying the developmental process (Tomancak et al., 2002). The fast growing ‘Expression Pattern’ database under the Berkeley Drosophila Genome Project (BDGP, 2005) now contains around 75 000 digital images of expression patterns of over 3400 genes.

As of now, the only mining approach offered by the BDGP to search for, for example, co-expressed genes, or anatomical and histological annotations of the gene expressions, is based on manual-labeling of the images by a domain expert using a controlled vocabulary. However, with the rapid growth of data volume, manual analysis is no longer feasible, and automatic analysis techniques...
are sorely needed, which require the development of new systems capable of noise removal, pattern extraction, feature description and similarity measures.

1.1 Highlights of this article

In this article, we present SPEX2 (SPatial gene EXPression pattern EXtractor), a highly effective and reliable image processing pipeline for automated and concise extraction of bona fide gene expression patterns (rather than generic shaded areas as usually recognized by naive pattern extracting procedures), from Drosophila embryonic ISH results imaged from the lateral view. Such patterns offer a high-fidelity surrogate of the spatial patterns of gene expression in a developing embryo or if necessary other subjects in question (Fig. 2c), nearly free of misleading non-expression patterns due to poor quality staining/washing, body texture, color condensation caused by body anatomy, embryo shape and contour, etc., which often fool standard pattern extracting procedures, as endogenous gene expression patterns (Fig. 2b). These patterns allow highly informative and specific feature representations of each gene to be generated, which can be used in a variety of downstream analysis like functional clustering, gene annotation and network inference.

Specifically, we address the following questions in this article:

1. Given an ISH image of a Drosophila embryo, how to find the pixels that correspond specifically to the spatial expression pattern, rather than other non-expressional entities such as body anatomies and textures, in the embryo?
2. How should a good representation of the gene expression pattern be constructed?
3. How should this representation be used for further clustering and classification tasks?

Comparisons of gene expression patterns from different ISH images must be performed with respect to the embryo, and not the image. The position, orientation, size, shape contour, lighting condition and texture of the embryo within the image do not matter, as long as the comparison is dependent on the location and strength of the gene expression within the embryo. This requires automated detection of the embryo in an image. Additionally, the orientation of the embryo needs to be identified and standardized, and the embryo must be registered to a standard shape. Furthermore, the ISH image contains noise in addition to the gene expression itself, due to staining artifacts. The correct expression pattern must be extracted from the registered image before conducting further analysis.

SPEX2 converts every raw ISH image of Drosophila embryo into a feature representation of the spatial gene expression pattern thereof suitable for downstream quantitative analysis, based on the following three steps : (i) embryo standardization, via embryo extraction, orientation correction and registration, (ii) gene expression extraction via stain extraction and pattern segmentation and (iii) feature extraction. Each step in the pipeline uses image processing and machine-learning algorithms to extract the correct output. Automated error control methods detect and reject images if they are not being correctly analyzed, or if they are unsuitable for analysis due to imaging artifacts.

The resultant feature representation can be directly used for tasks like classification, clustering, standard correlation analysis and network inference of Drosophila genes in a metric space. Our techniques are automatic, and are not specific to any data set. Our pipeline also outputs spatial patterns of gene expression, that are amenable to easy interpretation by biologists.

As proof of concept, we demonstrate our technique on lateral view images from the Berkeley Drosophila Genome Project (BDGP) gene expression pattern database, from the time stage 13–16. To evaluate our pipeline, we cluster the genes based on the features extracted by SPEX2, and report enrichment analysis, conducted using gene ontology (GO) functional annotations, as well as enrichment of manual annotations describing the spatial expression localization using a controlled vocabulary. We also learn a classifier to annotate gene expression patterns during embryogenesis using a controlled vocabulary, and report classification accuracy. We find that we significantly outperform other standard feature extraction techniques from the computer vision community, as well as the techniques reported in previous work.

1.2 Related work

We build upon the first steps taken by earlier work to construct our analysis pipeline for Drosophila ISH images. The system BEST, developed by Kumar et al. (2002), performs a direct pixel-level comparison of binarized images, using the intersection of the foreground regions as a similarity measure for gene expression patterns. They develop an embryo enclosing algorithm to find the embryo outline, and extract the binary expression pattern via adaptive thresholding.

Li et al. (2009) propose multi-instance multi-label learning via appropriate kernels to improve performance specifically for annotating images using a controlled vocabulary. An extension was proposed by Li et al. (2009) to model term-term interactions in a regression framework that has improved performance for this task. They extract position invariant features using a sparse codebook on aligned images, and apply a local regularization framework on these features for automatic image annotation.

Peng and Myers (2004), and Zhou and Peng (2007) developed techniques to represent ISH images, based on Gaussian mixture models, principal component analysis and wavelet functions. They use the wavelet features, with min-redundancy max-relevance feature selection, to automatically annotate images. Heffel et al. (2008) have also proposed a pipeline for this task, using embryo outline extraction, transformation of the embryo into a circular outline and conversion to fourier-coefficients-based feature representation. They report a visual clustering of seven images using their pipeline.

Tomancak et al. (2007) analyzed the global gene expression patterns in the BDGP data set, using only the manual annotations available for each gene from a controlled vocabulary. They reported clustering results on joint clustering of microarray data and annotation terms, and found interesting clusters that could not be found using microarray data alone.
The SPEX 2 system consists of three major components: (i) embryo standardization, (ii) gene expression pattern extraction and (iii) feature representation. An illustration of the pipeline is given in Figure 3. Below, we describe each component in detail.

2 METHODS

The SPEX2 system consists of three major components: (i) embryo standardization, (ii) gene expression pattern extraction and (iii) feature representation. An illustration of the pipeline is given in Figure 3. Below, we describe each component in detail.

2.1 Embryo standardization

Given a raw ISH image, SPEX2 uses an embryo standardization process to convert it into a standardized form suitable for subsequent expression extraction and pattern comparison. The embryo is extracted from the ISH image, and aligned along its anterio/posterior (A/P) and dorsal/ventral (D/V) axis correcting for the orientation, thereby ensuring the anterior (of the embryo) is to the left and the dorsal surface is to the top of the image. Finally, the embryo is registered to a standard shape and size.

2.1.1 Embryo outline extraction. Our embryo extraction procedure works in two steps. First, a foreground object extractor is used to extract potential embryos in the image. Second, a series of increasingly complex tests filter out foreground objects that are not embryos, or are embryos not suitable for analysis.

The object extractor uses the Canny edge operator to identify regions with fast-changing color and high variance. A series of morphological operations (dilations and erosions) are used to smooth out the edges and close holes to find the foreground objects.

A sequence of tests are then applied to each foreground object to test whether it’s an embryo suitable for further analysis; rejected items include erroneous outlines, partial embryos, multiple embryos physically touching or overlapping with each other, and excessively dried or otherwise mishandled embryos.

(1) Objects touching the image boundary are rejected, since these may be partially imaged embryos.

(2) Objects that are too small or too large are rejected. Small objects imply that a part of the actual embryo was potentially missed by the object extractor. Large objects are either partial embryos imaged using a large magnification, or incorrect outlines that include a portion of the background in the foreground object.

(3) If the maximum distance between the object outline and the convex outline of the object is large, the image is rejected; ensuring that the embryo outline is almost convex.

(4) Scale-independent shape features of the object outline are extracted and compared with expected shape features of a standard embryo. Scale independence is required since the size of the embryo varies across images. Examples of shape features include: (i) the ratio between the major and minor axes of the object must match the expected ratio for a *Drosophila* embryo. This ensures that the object is not too thin and narrow, nor is it too circular. (ii) the centroid of the foreground object must be close to the centroid of its outlining rectangle (ensures symmetry). (iii) the maximum (and mean) curvature of the object outline must be similar to the values expected for an embryo (filters out deformed embryos). If the value of any of the above features is >20% away from the feature value computed from a single correctly identified embryo, then the image is rejected.

Some examples of embryo outlines extracted by our algorithm are shown in Figure 4. Embryo extraction works well in presence of varying illumination (Fig. 4a), when the background is lighter than the foreground (Fig. 4b), in the absence of stain in the embryo (Fig. 4c), and when there are multiple embryos touching each other (Fig. 4d).

2.1.2 Alignment, orientation detection and registration. To align all embryos for later comparisons, we assume the camera angle is perpendicular to the surface of the embryo, which is the case with most imaging technologies with zoom-in. An ellipse is fitted to the detected embryo outline, with the major axis of the ellipse assumed to be the A/P axis, and minor axis the D/V axis of the embryo; and the embryo is rotated so that the A/P axis is horizontal.

Next, the correct orientation of the aligned embryo is identified and standardized so that the head is to the left, tail to the right, dorsal part of the embryo at the top, and ventral part at the base. This is akin to a binary classification task, for which we need to determine whether to flip the embryo horizontally to correctly position the anterior part of the embryo to the left, and vertically to position the dorsal side to top. Gargesha et al. (2005) proposed a technique to automatically annotate the A/P sides of the embryo. However, their technique is supervised, requiring a large amount of pre-labeled data, which is tedious and expensive to generate. Additionally, their technique is based on a heuristic that does not utilize the knowledge of the expected gene expression patterns. As for finding the D/V sides of the embryo, to our knowledge, no reported result is available so far.

We propose an algorithm for unsupervised embryo orientation detection, based on the insight that images of the same gene at the same time stage must have similar expression patterns. We start with a heuristic assignment to each embryo, and change the assignment of a particular embryo if it
different images by comparing the pixel-level expression values.

structures, enabling comparison of the spatial patterns of gene expression in

of the standardization process, for all the processed images, there is a

an exact map from pixel space to body part of the embryo. At the end

to register the embryo to a standard ellipse shape. This enables us to obtain

examples of orientation detection and correction of embryos is shown in

the heuristic that the dorsal side is less curved than the ventral side of the

stained for a single gene is outlined in Figure 5, and is run for all genes

increases its similarity with other embryos stained with the same gene, in a

greedy manner. The algorithm for A/P orientation detection for all embryos

identified by pre-processing the image is a noisy measurement of the true

effect, distorted due to poor quality staining/washing, body texture, color condensation caused by body anatomy, embryo shape and contour, etc. Since the expression patterns are noisy with no sharp edges, standard edge-based segmentation algorithms are unable to find the correct stain pattern; adaptive thresholding methods also fail due to the presence of a large variance in the amount of staining in different images. Hence, we correct these issues by using a MRF-based segmentation algorithm to remove noise from the expression pattern. Furthermore, given wide differences of expression patterns in different images, using a standard MRF with fixed parameters across images is hardly adaptive; therefore we fit image-specific MRFS in an unsupervised manner.

2.2 Concise Gene Expression Pattern Extraction

Given a standardized embryonic image, SPEX2 extracts concise spatial gene expression patterns therein via a two-step procedure. First, standardized embryonic images are pre-processed to extract ISH stains. Then, noise in the stains are removed using a Markov Random Field (MRF) model-based image segmentation. Our algorithm constructs the MRF graph structure and finds image-specific parameters for the image segmentation in a completely unsupervised way.
We used $\beta$ where $\Phi_1$ is the node potential, which captures the effect that pixel $y_i$ has on the label of $x_i$. $\Psi$ is the edge potential, which captures how the label of $x_i$ is influenced by the labels of its neighbors, and $E$ is the set of edges we found over the super-pixels.

The node potential $\Phi(y_i, y_j)$ is assumed to be Gaussian with parameters $(\mu_f, \sigma_f)$ if $x_i$ is foreground, and $(\mu_b, \sigma_b)$ if $x_i$ is background. The edge potential is defined as

$$
\Psi(x_i, y_j) = \exp(-\beta I(x_i = y_i, x_j = y_j)),
$$

where $I$ is an indicator function. $\beta$ defines the penalty given for neighboring pixels to disagree, i.e. one of the pixels is foreground and the other is background, and there is an edge connecting them. $\beta$ captures the strength of the penalty. As $\beta$ increases, we encourage smoother foreground assignments. We used $\beta = 2$, and found that it gave reasonably good performance.

2.2.4 Learning MRF parameters For the MRF defined above, the parameters $(\mu_f, \sigma_f, \mu_b, \sigma_b)$ must be defined for each image. Learning the MRF parameters for every image, by using classical unsupervised MRF learning techniques, is usually slow and inconvenient to process thousands of images.

We propose a simple heuristic to determine the graph parameters. If the penalty parameter $\beta$ is zero, then the edge potentials are constant. The MRF then reduces to a mixture of Gaussians, where every super-pixel value is generated from one of two Gaussians, corresponding to the foreground and background, respectively. The Gaussian parameters can then be learnt efficiently by computing the histogram of the image, and fitting a mixture of two Gaussians to the histogram using EM. If we want to impose the smoothness of the estimates, we add a small uniform prior (1% of the mass of the histogram) to the image histogram before running EM. The parameters of the two Gaussians are then treated as approximations to the MRF parameters, i.e. $\mu_1, \mu_2, \sigma_1, \sigma_2$.

2.2.5 Loopy belief propagation for inference A standard approximate inference technique, loopy belief propagation (LBP), is used to find the maximum a posteriori (MAP) assignment to each $x_i$, as foreground or background. Although LBP is not always guaranteed to converge, in our experiments, a small number (3–10) of iterations were sufficient for convergence, for all input images. At the end of this inference procedure, all background nodes are set to zero, and the foreground expression value is used as the final gene expression pattern obtained at the end of our image analysis pipeline. A small flowchart of our gene expression pattern extraction process is shown in Figure 8. Some examples of the gene expression patterns found by our MRF image segmentation algorithm are shown in Figure 9.

3 RESULTS

We apply SPEX$^2$ to the ISH images from the BIDGP (2005). Since our system performs automatic analysis for images in the lateral position, we picked 2689 images from the 13–16 time stage of the data set, which represent the expression patterns of 1432 genes. After automatic filtering of unqualified images in the standardization phase, 1904 images of 1011 genes entered the pattern extraction phase. We analyzed these expression patterns and report results on
two exemplary tasks: automatic annotation of images, and image clustering.

3.1 Image annotation

The expression patterns in BDGP *Drosophila* ISH images were annotated with anatomical and development ontology terms from a controlled vocabulary by human curators. Automatic annotation of images with terms from a controlled vocabulary represents a unique challenge itself. Since the main goal of SPEX2 is to extract concise spatial expression patterns from ISH images for generic downstream applications of any user, rather than offering a perfect annotator, we will demonstrate the quality of the SPEX2 output (e.g. expression features) using standard off-the-shelf annotation classifiers.

We focus on the 10 most frequent annotation terms in BDGP, and treat every term as an independent binary classification task. Each binary classifier is a standard SVM with a Gaussian kernel (we used libsvm (Chang and Lin, 2001) for our experiments). We use 10-fold cross-validation over a small set of values to pick the tuning parameter of SVMs—the cost of misclassification $C$.

We compare our results with two benchmark systems representing the state-of-the-art. In System I, we implement the feature extraction and classification procedure proposed by Zhou and Peng (2007). Their system extracts the embryo outline by using an adaptive thresholding method (Peng and Myers, 2004), and registers the embryo using affine transformation and intensity scaling. The A/P orientation is determined by maximizing total gene similarity across all images. Subsequently, 2D wavelet embryo features are used, with min-redundancy max-relevance feature selection to pick the best features. Finally, binary classification on each annotation term is obtained via LDA (linear discriminant analysis). In System II, Ji et al. (2009) used dense SIFT feature descriptors that are converted into sparse codes to form a codebook to represent their aligned images, and proposed an elegant local regularization (LR) procedure for multi-label learning. Details on how to obtain well-aligned images were not given, but the work by the same group in Ye et al. (2006) used a image standardization procedure outlined in Kumar et al. (2002), followed by histogram equalization for improved contrast in images. Hence, we use the above procedure when implementing this system, using the LR code from that group.

We evaluate the performance using accuracy and $F_1$ score (Goutte and Gaussier, 2005). $F_1$ score is the harmonic mean between the precision and recall of the results, and lies between 0 and 1, with higher $F_1$ representing better performance. Figure 10 shows the classification accuracy based on the SPEX2 features, in comparison with the benchmarks. In terms of mean accuracy, SPEX2 outperforms both the systems, while maintaining the same $F_1$ score. It is noteworthy that our result is obtained with a standard SVM, because our goal here is to demonstrate the quality of the SPEX2 features, not that of the annotation algorithm. Indeed, we observe that using the sophisticated LR annotation algorithm of System II with our SPEX2 features, increases our $F_1$ score, at the cost of a very small reduction in accuracy. Using the paired $t$ test, the difference in accuracy between SPEX2 with LR and System II was found significant with $P = 9.51e−5$.

In addition, we visualize the information captured in the extracted expression patterns from SPEX2 and the two systems we compare with, by computing the SVD of the expression patterns (Pan et al., 2006). The set of eigen vectors can then be represented as images. We call these images eigen-expression patterns, like eigenfaces used in facial recognition (Pentland and Turk, 1991). The top 25 eigen-expression patterns produced by SPEX2 seem to find localized regions of expression that correspond well to known gene expression patterns.
We do most of our analysis on 15 clusters, the mean image of each

Next, we evaluate the SPEX features on clustering, using a popular

3.2 Gene expression clustering

We conduct enrichment analysis using both sets of information.

3.2.1 Hypothesis test for enrichment Given a single cluster, and a

3.2.2 Annotation terms enrichment If the data is well clustered,

3.2.3 GO functional enrichment It is believed that similar

source is the GO functional annotations, associated with the gene.

We conduct enrichment analysis using both sets of information.

Given a single cluster, and a single annotation term (from controlled vocabulary or GO ontology), a P-value can be obtained by using an exact hypergeometric test. However, since we test each cluster for multiple annotations, a correction for multiple hypothesis is needed. Standard corrections for multiple hypothesis testing are usually found to be either very conservative, or having low power. We instead convert the P-values into q-values, that control the positive false discovery rate (pFDR), by using the procedure described by Storey (2002). The pFDR is the expected proportion of erroneous rejections among all rejections, thus a pFDR value of 5% means that 5% of predicted significant features will be truly null. The q-value measures the strength of the observed statistic, with respect to pFDR, and automatically corrects for multiple hypothesis testing, it is therefore a much more powerful test scheme.

We conduct enrichment analysis using the procedure outlined by Arava et al. (2003), which allows us to estimate q-values for multiple hypothesis tests, even when the statistics being measured are correlated (as is the case for GO and pattern annotations).

If the data is well clustered, then a single cluster of images must be enriched for specific annotation terms that the images have been annotated with. Table 1 shows a partial enrichment analysis for 15 clusters. All clusters were significantly enriched for at least one term, with a total of 90 enriched terms. Since the number of terms is higher than the number of clusters, each cluster is enriched for a combination of multiple terms. For example, cluster one with 149 images is enriched for images that have been annotated with embryonic brain and central nervous system, while cluster three with 100 images is enriched for a combination of embryonic brain with embryonic midgut and ventral nerve cord. Images annotated with only ventral nerve cord have been clustered into a separate cluster (having 139 images).

To assess the advantage of concise expression information extracted by SPEX2 over benchmark systems, we performed the same clustering analysis based on features generated by the two systems discussed above. We counted the number of clusters from there that have at least one significant annotation at $q=0.05$. Figure 13 shows the number of significant clusters found by the three methods, as we vary the number of clusters from 5 to 100. We observe that SPEX2 works better than the other two methods, with an average of 18.39% more significant clusters obtained than its closest competitor System I.

Since we are analyzing spatial patterns of genes that are differentially expressed in the embryonic stage, without any analysis across time, we expect to find enrichment of smaller, more precise

Fig. 12. Each image is the mean of a single cluster found by using processed images from different systems. The intensity of any pixel in the mean image is the average intensity of that pixel in all images assigned to this cluster. As can be seen, clustering using unprocessed images only finds clusters based on embryo position and illumination. The clusters produced by SPEX2 have very low noise, and visually look pure in terms of patterns clustered.

The literature on clustering specifies a variety of evaluation

Given a single cluster, and a single annotation term (from controlled vocabulary or GO ontology), a P-value can be obtained by using an exact hypergeometric test. However, since we test each cluster for multiple annotations, a correction for multiple hypothesis is needed. Standard corrections for multiple hypothesis testing are usually found to be either very conservative, or having low power. We instead convert the P-values into q-values, that control the positive false discovery rate (pFDR), by using the procedure described by Storey (2002). The pFDR is the expected proportion of erroneous rejections among all rejections, thus a pFDR value of 5% means that 5% of predicted significant features will be truly null. The q-value measures the strength of the observed statistic, with respect to pFDR, and automatically corrects for multiple hypothesis testing, it is therefore a much more powerful test scheme.

We conduct enrichment analysis using the procedure outlined by Arava et al. (2003), which allows us to estimate q-values for multiple hypothesis tests, even when the statistics being measured are correlated (as is the case for GO and pattern annotations).

If the data is well clustered, then a single cluster of images must be enriched for specific annotation terms that the images have been annotated with. Table 1 shows a partial enrichment analysis for 15 clusters. All clusters were significantly enriched for at least one term, with a total of 90 enriched terms. Since the number of terms is higher than the number of clusters, each cluster is enriched for a combination of multiple terms. For example, cluster one with 149 images is enriched for images that have been annotated with embryonic brain and central nervous system, while cluster three with 100 images is enriched for a combination of embryonic brain with embryonic midgut and ventral nerve cord. Images annotated with only ventral nerve cord have been clustered into a separate cluster (having 139 images).

To assess the advantage of concise expression information extracted by SPEX2 over benchmark systems, we performed the same clustering analysis based on features generated by the two systems discussed above. We counted the number of clusters from there that have at least one significant annotation at $q=0.05$. Figure 13 shows the number of significant clusters found by the three methods, as we vary the number of clusters from 5 to 100. We observe that SPEX2 works better than the other two methods, with an average of 18.39% more significant clusters obtained than its closest competitor System I.

Since we are analyzing spatial patterns of genes that are differentially expressed in the embryonic stage, without any analysis across time, we expect to find enrichment of smaller, more precise

The literature on clustering specifies a variety of evaluation

Given a single cluster, and a single annotation term (from controlled vocabulary or GO ontology), a P-value can be obtained by using an exact hypergeometric test. However, since we test each cluster for multiple annotations, a correction for multiple hypothesis is needed. Standard corrections for multiple hypothesis testing are usually found to be either very conservative, or having low power. We instead convert the P-values into q-values, that control the positive false discovery rate (pFDR), by using the procedure described by Storey (2002). The pFDR is the expected proportion of erroneous rejections among all rejections, thus a pFDR value of 5% means that 5% of predicted significant features will be truly null. The q-value measures the strength of the observed statistic, with respect to pFDR, and automatically corrects for multiple hypothesis testing, it is therefore a much more powerful test scheme.

We conduct enrichment analysis using the procedure outlined by Arava et al. (2003), which allows us to estimate q-values for multiple hypothesis tests, even when the statistics being measured are correlated (as is the case for GO and pattern annotations).

If the data is well clustered, then a single cluster of images must be enriched for specific annotation terms that the images have been annotated with. Table 1 shows a partial enrichment analysis for 15 clusters. All clusters were significantly enriched for at least one term, with a total of 90 enriched terms. Since the number of terms is higher than the number of clusters, each cluster is enriched for a combination of multiple terms. For example, cluster one with 149 images is enriched for images that have been annotated with embryonic brain and central nervous system, while cluster three with 100 images is enriched for a combination of embryonic brain with embryonic midgut and ventral nerve cord. Images annotated with only ventral nerve cord have been clustered into a separate cluster (having 139 images).

To assess the advantage of concise expression information extracted by SPEX2 over benchmark systems, we performed the same clustering analysis based on features generated by the two systems discussed above. We counted the number of clusters from there that have at least one significant annotation at $q=0.05$. Figure 13 shows the number of significant clusters found by the three methods, as we vary the number of clusters from 5 to 100. We observe that SPEX2 works better than the other two methods, with an average of 18.39% more significant clusters obtained than its closest competitor System I.

Since we are analyzing spatial patterns of genes that are differentially expressed in the embryonic stage, without any analysis across time, we expect to find enrichment of smaller, more precise
Table 1. Enrichment analysis for 15 clusters, using terms from the controlled vocabulary

<table>
<thead>
<tr>
<th>Cluster size</th>
<th>Term annotation</th>
<th>Annotation probability</th>
<th>Overlap q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>140</td>
<td>Embryonic brain</td>
<td>0.298</td>
<td>113</td>
</tr>
<tr>
<td>194</td>
<td>Embryonic midgut</td>
<td>0.282</td>
<td>109</td>
</tr>
<tr>
<td>194</td>
<td>Embryonic larval muscle system</td>
<td>0.150</td>
<td>67</td>
</tr>
<tr>
<td>149</td>
<td>Embryonic Malpighian tubules</td>
<td>0.074</td>
<td>41</td>
</tr>
<tr>
<td>104</td>
<td>Embryonic midgut</td>
<td>0.282</td>
<td>56</td>
</tr>
<tr>
<td>139</td>
<td>Ventral nerve cord</td>
<td>0.327</td>
<td>75</td>
</tr>
<tr>
<td>39</td>
<td>Embryonic central brain pars intercerebralis</td>
<td>0.0094</td>
<td>3</td>
</tr>
<tr>
<td>110</td>
<td>Aminotransferase</td>
<td>0.01577</td>
<td>13</td>
</tr>
<tr>
<td>140</td>
<td>Embryonic epiphatons</td>
<td>0.0678</td>
<td>27</td>
</tr>
<tr>
<td>165</td>
<td>Embryonic hypopharynx</td>
<td>0.168</td>
<td>51</td>
</tr>
<tr>
<td>168</td>
<td>Yolk nucleus</td>
<td>0.037</td>
<td>6</td>
</tr>
<tr>
<td>78</td>
<td>Embryonic brain</td>
<td>0.298</td>
<td>47</td>
</tr>
<tr>
<td>70</td>
<td>Embryonic epiphatons</td>
<td>0.168</td>
<td>28</td>
</tr>
<tr>
<td>96</td>
<td>Embryonic large intestine</td>
<td>0.035</td>
<td>13</td>
</tr>
<tr>
<td>103</td>
<td>Embryonic larval somatic muscle</td>
<td>0.070</td>
<td>31</td>
</tr>
<tr>
<td>93</td>
<td>Ventral nerve cord</td>
<td>0.327</td>
<td>51</td>
</tr>
</tbody>
</table>

The first column shows the size of the cluster, the next two columns show the term annotation, and the probability that a given gene will be annotated with this term. The fourth column gives the number of images in this cluster annotated with this term, with the last column giving the q-value of the overlap.

Table 2. Enrichment analysis for 15 clusters, using GO functional annotations

<table>
<thead>
<tr>
<th>Cluster size</th>
<th>GO category</th>
<th>GO function</th>
<th>GO category</th>
<th>GO function</th>
<th>Overlap q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>149</td>
<td>GO:0007520</td>
<td>Myoblast fusion</td>
<td>103</td>
<td>GO:0007150</td>
<td>DNA double-strand RNA binding</td>
</tr>
<tr>
<td>187</td>
<td>GO:0007342</td>
<td>Open tracheal system development</td>
<td>0.001015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>126</td>
<td>GO:0008407</td>
<td>Bristle morphogenesis</td>
<td>0.001015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>GO:0035193</td>
<td>Larval central nervous system remodeling</td>
<td>0.001015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>174</td>
<td>GO:0007352</td>
<td>Metamorphosis</td>
<td>0.06963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>116</td>
<td>GO:0007662</td>
<td>Ecdysis</td>
<td>0.001015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>GO:0008508</td>
<td>Polyhedral morphogenesis</td>
<td>0.001015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>GO:0008466</td>
<td>Enolpyruvate inhibitor activity</td>
<td>0.06963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>GO:0004497</td>
<td>Mesoorygogenase activity</td>
<td>0.001015</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first column shows the size of each cluster, the next three columns show the GO category, function, and number of genes in the dataset having that GO function. The fifth column gives the number of genes with the particular GO function present in this cluster, and the last column gives the q-value of the overlap.

All 10 genes related to 'progression of morphogenetic furrow during compound eye morphogenesis' are found in the same cluster, and five of the nine genes related to segment specification, are also clustered together. Additionally, all genes related to 'larval central nervous system remodeling' are found in a single cluster, and five of the six genes related to 'bristle morphogenesis' are also co-clustered.
This seems to imply that genes involved in larval stage development are already showing spatial coherence in the embryonic stage. Thus, the SPEX² clusters are able to capture fine-grained GO functional annotations. In contrast, clustering using features extracted by System I found only six significant clusters out of 15. Our method thus improves the number of significantly enriched clusters by 50%. System II returned only one significantly enriched cluster out of 15, at $q = 0.05$.

4 DISCUSSION

SPEX² represents the first step towards automatic functional analysis of ISH images of Drosophila embryos, namely concise extraction of spatial gene expression patterns. Our extraction system employs a pipeline of analytical techniques to first standardize the embryo via embryo outline extraction, orientation detection and correction, and registration; and then extracts spatial expression signal via filters and probabilistic segmenters. Finally, it converts the spatial signals into a low-dimensional feature representation, suitable for advanced analysis. We evaluated our system by using the resultant features for automatic pattern annotation and clustering. Using simple classification techniques and our sophisticated feature extraction pipeline, we achieved a significant improvement in annotation accuracy over existing systems. We also clustered the Drosophila ISH images, and conducted enrichment analysis on both pattern term annotations, and GO functional annotations. We found significant enrichment in both scenarios.

The next step is a more detailed analysis of ISH images using this feature representation. The current work has focused on clustering images from a single time stage—in the future, we plan to study image analysis across time. Another important question to be addressed is how to combine microarray data with ISH image data to be able to be able to leverage two independent sources for joint analysis.

The concise spatial pattern of genes extracted from ISH images by SPEX² can also be used as a token of gene expression and applied to infer a gene regulation network, as with microarray data. A detailed study along this direction involves some additional technicalities, and is therefore beyond the scope of this paper.

Finally, another direction of future research would be to find time-varying gene regulatory networks using this data. Such analysis would allow us to capture spatial variations at a single time stage, as well as varying relationships between genes across time. A first step in this direction has been taken for microarray data by Ahmed and Xing (2009). We intend to extend developments of this model for Drosophila ISH images, thus enabling us to discover spatial-temporal gene regulation networks.

ACKNOWLEDGEMENTS

We thank Fan Guo and Lei Li for assistance and helpful discussions during initial processing of the BDGP images.

Funding: National Science Foundation DBI (0640543, 0546594); Alfred P. Sloan Fellowship (to E.P.X.)

Conflict of Interest: none declared.

REFERENCES


Pan,J.-Y. et al. (2006) Automatic mining of fruit fly embryo images. In ACM SIGKDD conference on Knowledge Discovery and Data Mining.


