Extraction and comparison of gene expression patterns from 2D RNA in situ hybridization images

Daniel L. Mace1,2, Nicole Varnado2, Weiping Zhang2, Erwin Frise3, and Uwe Ohler2∗

1Computational Biology and Bioinformatics Graduate Program, Duke University, Durham, NC 27702
2Institute for Genome Sciences and Policy, Duke University, Durham, NC 27702
3Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

ABSTRACT

Motivation: Recent advancements in high-throughput imaging have created new large data sets with tens of thousands of gene expression images. Methods for capturing these spatial and/or temporal expression patterns include in situ hybridization or fluorescent reporter constructs or tags, and results are still frequently assessed by subjective qualitative comparisons. In order to deal with available large data sets, fully automated analysis methods must be developed to properly normalize and model spatial expression patterns.

Results: We have developed image segmentation and registration methods to identify and extract spatial gene expression patterns from RNA in situ hybridization experiments of Drosophila embryos. These methods allow us to normalize and extract expression information for 78,621 images from 3,724 genes across 6 time stages. The similarity between gene expression patterns is computed using four scoring metrics: mean squared error, haar wavelet distance, mutual information, and spatial mutual information. We additionally propose a strategy to calculate the significance of the similarity between two expression images, by generating surrogate data sets with similar spatial expression patterns using a Monte Carlo swap sampler. On data from an early development time stage, we show that spatial mutual information provides the most biologically relevant metric of comparison, and that our significance testing generalizes metrics to a strategy to calculate the significance of the similarity between two expression images, by generating surrogate data sets with similar spatial expression patterns using a Monte Carlo swap sampler. On data from an early development time stage, we show that spatial mutual information provides the most biologically relevant metric of comparison, and that our significance testing generalizes metrics to achieve similar performance. We exemplify the application of spatial metrics on the well-known Drosophila segmentation network.

Availability: An Java webstart application to register and compare patterns, as well as all source code, are available from: http://tools.genome.duke.edu/generegulation/image_analysis/insitu

Contact: uwe.ohler@duke.edu

1 INTRODUCTION

Advances in high-throughput microscopy have led to a rapid increase of digital image data in biology. New methods to image biological specimens at high resolution, and to visualize expression of genes of interest, have lead to a high interest in imaging in developmental and molecular biology, including the creation of virtual embryos to map expression profiles of important regulatory genes (Keller et al., 2008; Fowlkes et al., 2008). In the past, these images were analyzed in a manual fashion, e.g. comparing two expression patterns by qualitative visual inspection. In order to deal with these data sets more appropriately, it is necessary to develop automated methods for extracting and analyzing images. Methods to quantitatively describe spatial/temporal expression patterns is a relatively new area of research that has begun to be explored in model organisms (Megason and Fraser, 2007). In sea urchins, a comprehensive analysis of normalization methods was used for determining and quantifying spatial expression data (Damle et al., 2006). Recently, large databases of images for C. elegans have become available (Hunt-Newbury et al., 2007), accompanied by methods to analyze the data (Bao et al., 2006; Murray et al., 2008). For Arabidopsis root images, image registration techniques have been used to quantify tissue specific expression from green fluorescent protein (GFP) reporter lines (Mace et al., 2006; Lee et al., 2006).

A particularly interesting and fundamental problem that arises with the availability of image data, and which we address in this study, is to compare two samples on the level of their expression profiles: for instance, the same gene under different conditions or across different species, or different genes with the goal to cluster them akin to approaches developed for microarray data. Several general problems arise when comparing image expression data: (1) we need to develop methods to process the raw input images, to eliminate noise under a typical large range of imaging conditions (e.g. different viewpoints, different locations, multiple specimens per image) and to perform normalizations to decouple the variability in morphology from the variability in expression; (2) we need to represent the expression patterns and to specify appropriate similarity metrics capable of assessing spatial/temporal similarity; (3) and we need to assess the significance of observed similarities.

The majority of image analysis work in the context of development of model organisms has been carried out for Drosophila, and can be broadly grouped into two categories: quantitative high resolution analysis of a relatively small selected
subset of genes (Janssens et al., 2005a; Keranen et al., 2006; Fowlkes et al., 2008), and higher-throughput pattern analysis of thousands of genes (Peng et al., 2007; Kumar et al., 2002; Harmon et al., 2007). Fine-grained high resolution analysis often extracts quantitative expression values from image data, e.g. for the numerical simulations of specific regulatory pathways. This study follows previous high-throughput approaches which uses large datasets to quantitatively address expression in the context of spatial patterns: prediction of annotation terms, or clustering/classification of genes. We use a dataset of Drosophila embryonic expression patterns (Tomanack et al., 2007) to illustrate how we address the three questions outlined above.

Our work distinguishes itself from previous work on this Drosophila data by three main contributions. First, robust and fully automated image analysis techniques are used to process and register the raw images. Through the use of statistical shape models and partial mapping methods, these techniques are capable of handling sources of phototopic and imaging variability that limited previous approaches. Second, we comprehensively compare different similarity metrics, implement similarity measures that incorporate spatial dependencies to distinguish complex spatial patterns, and validate the different measures against visual annotation terms provided by experts. Third, we develop a new significance testing framework for spatial similarity scores through constrained realization Monte Carlo simulations and demonstrate how it can generalize similarity measures to achieve similar performance to spatial metrics. Finally, we illustrate this method of significance testing on known biological examples, emphasizing the importance for fully automated image registration/comparison models in the context of regulatory interactions. While we use a fly embryo dataset as example, the general registration and comparison approach is adaptable and thus of interest to the study of spatial expression patterns in a wide range of model organisms.

### 2 APPROACH

#### 2.1 Image Registration

Prior to any quantitative analysis of expression image data, it is necessary to normalize and register the images to a common frame of reference. By first mapping different specimens such as fly embryos to a common reference, we can subsequently apply a large variety of methods for univariate and multivariate data analysis for cross-subject comparisons (between replicates, genes, time stages, or different strains or species).

We have developed a fully automated registration method based on statistical shape models and improved numerical optimizers. This approach estimates both the average shape of an embryo, as well as the main components of variation of embryo shape, including orientation, from labeled data. It then uses this model to segment new images into foreground (a single complete embryo) and background (Figure 1). We applied this registration method to the complete set of 78,621 images in the latest release of the Berkeley Drosophila in situ database. Overall, the model-based registration addresses problems that limited the successful application of previous methods on the whole database, highlighting its ability of extracting embryos under a large variety of imaging conditions: multiple embryos or impeding boundaries, changes in lighting/microscope settings, and out of focus boundary regions.

To formally analyze registration accuracy, we uniformly sampled 200 images from 200 uniformly sampled genes from the dataset for quantitative assessment. Each image was manually segmented, registered and compared to the automated segmentation and registration. We found that the most common inconsistencies in registration resulted from incorrectly orienting the axis, and were observed in 20 cases (10%) for the anterior-posterior axis, and in 6 cases (3%) for the dorsal/ventral axis. While our anterior-posterior axis alignment of 90% is consistent with an 85% accuracy reported earlier (Gargesha et al., 2005), our approach differs from Gargesha’s approach as we encapsulate the alignment problem into the registration process by allowing the axis-corrected shape model to more properly fit to the embryos in the image. The total registration accuracy was assessed using a Test Point Error (TPE) measure (Zitova and Flusser, 2003) between the manually and automatically registered images. The average TPE was 0.94, with 190 images (95%) having a TPE within a reasonable accuracy rate of >0.90, as compared to the leading reported registration accuracy (78% (Pan et al., 2006)).

After normalization and registration of the images, expression patterns, or features useful for classification, can be extracted. Here, we are interested in comparing the global 2D expression pattern, and a choice of representation of the patterns has to be made: instead of working with the complete pixel-based 2D patterns, it is common to map them to a smaller set of features representing expression in small subregions of the specimen. As the registration maintains morphological variabilities (size and stretch), it prevents a one-to-one mapping on the pixel level. To be able to evaluate the effect of different metrics in a simple scenario, we therefore chose to project the staining intensities along the anterior-posterior x-axis and dorsal-ventral y-axis as described in 3.3. Projections have been used in a variety of recent applications in Drosophila (Janssens et al., 2005b; Segal et al., 2008) — as well as Mouse brain images (Liu et al., 2007) — and are particularly suited for the analysis of early embryonic expression as discussed below.

#### 2.2 Correspondence of Expression Similarities to Expert Annotations

Using the extracted expression patterns, we assess the importance of using spatial metrics (Haar wavelets and spatial mutual information) by comparing their performance to two previously used non-spatial metrics (mean squared error and mutual information). We determined how the similarity values computed by each metric corresponded to manually annotated expression terms in the in situ database. For this purpose, we focused on the subset of 27,157 images covering 3,127 genes acquired during the time window of developmental stages 4-6. Images in this stage window have been annotated with information on the view (lateral or dorsal/ventral) from which the images were taken (lateral or dorsal/ventral), and this crucial information is not yet provided for later stages. Furthermore, the images at this stage balance the frequency of spatially diverse expression patterns with the resolution and coverage in the database, and are not subject to the additional complexity that expression in later stages is conditional on earlier expression, which is also reflected in the annotation terms.

Genes in the selected window were annotated with 38 unique terms describing the spatial expression patterns. We removed all genes that had annotation terms which indicated the lack of spatial
variability (e.g., ubiquitous, maternal) and non-lateral views, leaving us with 209 genes, 1231 images, and 29 annotation terms. For each scoring metric, we calculate an enrichment significance for each annotation term describing how often genes annotated with a particular ontology term show the strongest similarity to genes annotated with the same term. Using a p-value cutoff of 0.05, spatial mutual information performed the best, with 22 of the 29 annotation terms being significantly enriched, while mutual information led to the second best result with 21. Both mean squared error and Haar wavelets led to 19 and 18 enriched terms respectively (Table 1).

While spatial mutual information and Haar wavelets are capable of incorporating the spatial structure of expression pattern, they do not fully incorporate the significance of the spatial pattern: genes with complex spatial patterning (e.g., even-skipped) will be scored similarly to genes with simpler spatial patterning (e.g., bicoid). To better account for this, and to provide actual significance values between expression patterns, we have developed simulation methods that account for the complexity of the spatial patterning. The significance values are computed by comparing the observed similarity value to a null distribution that preserves spatial dependencies between the two gene expression patterns, as described in Section 3.5.1 and shown in Figure 2. To accommodate for low signal/strong lighting effects, the significance values were converted into reweighted significance scores (RSS) as described in Section 3.6. The enrichment significance calculation was repeated for the RSS values and are shown in Table 1. By incorporating the spatial structure using RSS for each score, the scoring metrics performed similarly, with all the metrics resulting in 19-22 enriched significance scores.

While consistent differences between metrics are observed, and the significance estimates produce more stable results across metrics, the overall performance is not dramatically different. The advantage of using appropriate similarity metrics and significance estimates becomes more apparent in noisier scenarios, or when fewer features are used. Many, but not all, terms exhibit distinct patterns along the AP axis; however, when using projections onto the AP axis only, mean squared error results in only 15 terms compared to the 18 terms for both axis. In comparison, mutual information, spatial mutual information, and the RSS scores remain largely consistent regardless of whether the DV axis is incorporated into the score. The full results analogous to Table 1 are given in the supplement.

### 2.3 Expression Similarity of Known Co-regulated Genes

As application of the registration and expression comparison pipeline, we use spatial mutual information and significance tests to validate known biological interactions—suggesting their use for biological inferences. Gene regulation and spatial patterning are a tightly coupled process: transcription factors acting as activators for a gene are often co-expressed in similar spatial regions, while repressors are often expressed inversely to the targeted gene. We address how such regulatory relationships are reflected in spatial expression profiles in the context of the segmentation network, using a set of the gap, pair-rule, and segmentation genes adapted from (Schroeder et al., 2004). This network consists of a set of genes with many direct regulatory interactions and shared functional roles; in many cases, similarities in function/interaction are reflected in noticeable similarities of spatial expression patterns. Since the subset we are using does not contain the irregularities/noise of the full data set, we here calculated the unweighted significance scores of the anterior/posterior projections described in Section 3.5.1.

The significance values for the similarity scores from all pairwise comparisons are shown in Figure 2. Many of the genes which share functional roles are identified as being significant; for instance, pdm2 and nubbin (also known as pdm1) are paralogs with highly similar functional roles and interactions (Yeo et al., 1995). Significant similarity scores can also reflect regulatory relationships...
between genes with overlapping expression domains in the transcriptional network: *ocelliless* (also known as *orthodenticle*) is positively regulated by *bicoid* (Finkstein and Perrimon, 1990); *Ubx* indirectly regulates *dichaete* through the intermediate activation of *dpp* (Capovilla et al., 1994; Sanchez-Soriano and Russell, 2000). In addition to activators, we observe significant similarities for repressors, a property of mutual information which scores correlation and anti-correlation equally: *hunchback* represses the expression of *nubbin*, *pdm2* (Kambadur et al., 1998) and *Ubx* (Pirrotta et al., 1995); *giant* and *Krueppel* mutually repress each other (Kraut and Levine, 1991). Additional significant spatial localization can be a result of conditions in which genes act either in concert, or independently, to regulate other downstream genes. *Nubbin* shows significance with *dichaete*, and double knockout studies have shown that these two genes are essential for the proper formation of even-skipped stripes 1, 4, 5 and 6 (Ma et al., 1998).

Not all known interactions are detected as significant, nor is this to be expected given the data. Often, a spatial expression pattern of a gene is the result of complex interactions between many genes across several time stages. For example, the proper development of all even-skipped stripes requires the interaction of many genes; the stripes are encoded by distinct cis-regulatory regions; and some factors function to control only a subset of the stripes. In such cases, assessing similarity on the level of global gene expression pattern, as pursued here as illustrative example, may therefore be modified to deal with parts of patterns, such as disjoint expression domains or even the boundary of expression domains. We anticipate that the integration of quantitative spatial expression information with other “traditional” high-throughput data (e.g. binding, expression), will be useful to infer complex interactions in the regulatory networks of multicellular eukaryotic organisms (Fowlkes et al., 2008).

### 3 METHODS

#### 3.1 Berkeley Drosophila In Situ database

The Berkeley Drosophila in situ database consists of 78,621 images of 3,724 genes expressed in *Drosophila* embryos across 6 time windows (covering the developmental stages 1-3, 4-6, 7-8, 9-10, 11-12, 13-15). An established RNA in situ hybridization staining protocol was used to visualize spatial expression patterns as described in (Tomancak et al., 2002). Manual visual annotations are based on an ontology describing embryonic expression patterns, consisting of 314 terms, and were obtained from the latest release.

<table>
<thead>
<tr>
<th>Annotation Term</th>
<th>Genes</th>
<th>MSE</th>
<th>Haar</th>
<th>MI</th>
<th>SMI</th>
<th>RSS</th>
<th>MSE</th>
<th>Haar</th>
<th>MI</th>
<th>SMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ammioserosa anlage in statu nascendi</td>
<td>19</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>anlage in statu nascendi</td>
<td>30</td>
<td>0.207</td>
<td>0.268</td>
<td>0.039</td>
<td>0.015</td>
<td>0.159</td>
<td>0.152</td>
<td>0.014</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>anterior endoderm anlage in statu nascendi</td>
<td>30</td>
<td>0.015</td>
<td>0.033</td>
<td>0.001</td>
<td>0.004</td>
<td>0.006</td>
<td>0.026</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>cellular blastoderm</td>
<td>27</td>
<td>0.126</td>
<td>0.142</td>
<td>0.312</td>
<td>0.368</td>
<td>0.265</td>
<td>0.193</td>
<td>0.463</td>
<td>0.355</td>
<td></td>
</tr>
<tr>
<td>clypealbrum anlage in statu nascendi</td>
<td>10</td>
<td>0.004</td>
<td>0.001</td>
<td>0.006</td>
<td>0.006</td>
<td>0.005</td>
<td>0.001</td>
<td>0.002</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>dorsal ectoderm anlage</td>
<td>15</td>
<td>0.008</td>
<td>0.009</td>
<td>0.002</td>
<td>0.010</td>
<td>0.010</td>
<td>0.006</td>
<td>0.003</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>dorsal ectoderm anlage in statu nascendi</td>
<td>72</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>ectoderm anlage in statu nascendi</td>
<td>13</td>
<td>0.017</td>
<td>0.030</td>
<td>0.017</td>
<td>0.029</td>
<td>0.017</td>
<td>0.035</td>
<td>0.015</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>endoderm anlage in statu nascendi</td>
<td>9</td>
<td>0.216</td>
<td>0.166</td>
<td>0.014</td>
<td>0.023</td>
<td>0.147</td>
<td>0.093</td>
<td>0.011</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td>foregut anlage in statu nascendi</td>
<td>25</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
<td>0.004</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>gap</td>
<td>27</td>
<td>0.101</td>
<td>0.053</td>
<td>0.020</td>
<td>0.027</td>
<td>0.046</td>
<td>0.049</td>
<td>0.036</td>
<td>0.030</td>
<td></td>
</tr>
<tr>
<td>head epidermis anlage in statu nascendi</td>
<td>4</td>
<td>0.005</td>
<td>0.009</td>
<td>0.005</td>
<td>0.006</td>
<td>0.004</td>
<td>0.005</td>
<td>0.004</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>head epidermis dorsal anlage in statu nascendi</td>
<td>10</td>
<td>0.007</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
<td>0.018</td>
<td>0.004</td>
<td>0.002</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>head mesoderm anlage</td>
<td>5</td>
<td>0.207</td>
<td>0.181</td>
<td>0.303</td>
<td>0.322</td>
<td>0.154</td>
<td>0.167</td>
<td>0.232</td>
<td>0.215</td>
<td></td>
</tr>
<tr>
<td>head mesoderm anlage in statu nascendi</td>
<td>15</td>
<td>0.075</td>
<td>0.169</td>
<td>0.095</td>
<td>0.029</td>
<td>0.075</td>
<td>0.135</td>
<td>0.112</td>
<td>0.118</td>
<td></td>
</tr>
<tr>
<td>hindgut anlage in statu nascendi</td>
<td>20</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>mesectoderm anlage in statu nascendi</td>
<td>12</td>
<td>0.007</td>
<td>0.021</td>
<td>0.032</td>
<td>0.008</td>
<td>0.007</td>
<td>0.012</td>
<td>0.025</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>mesoderm anlage in statu nascendi</td>
<td>18</td>
<td>0.180</td>
<td>0.166</td>
<td>0.161</td>
<td>0.097</td>
<td>0.219</td>
<td>0.101</td>
<td>0.196</td>
<td>0.189</td>
<td></td>
</tr>
<tr>
<td>pair rule</td>
<td>5</td>
<td>0.141</td>
<td>0.071</td>
<td>0.097</td>
<td>0.067</td>
<td>0.137</td>
<td>0.044</td>
<td>0.046</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td>pole cell</td>
<td>13</td>
<td>0.971</td>
<td>0.968</td>
<td>0.963</td>
<td>0.984</td>
<td>0.952</td>
<td>0.969</td>
<td>0.968</td>
<td>0.976</td>
<td></td>
</tr>
<tr>
<td>posterior endoderm anlage in statu nascendi</td>
<td>38</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>procephalic ectoderm anlage in statu nascendi</td>
<td>67</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>segmentally repeated</td>
<td>9</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>trunk mesoderm anlage</td>
<td>5</td>
<td>0.030</td>
<td>0.236</td>
<td>0.006</td>
<td>0.002</td>
<td>0.012</td>
<td>0.160</td>
<td>0.010</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>trunk mesoderm anlage in statu nascendi</td>
<td>17</td>
<td>0.017</td>
<td>0.045</td>
<td>0.162</td>
<td>0.086</td>
<td>0.008</td>
<td>0.066</td>
<td>0.269</td>
<td>0.208</td>
<td></td>
</tr>
<tr>
<td>ventral ectoderm anlage</td>
<td>11</td>
<td>0.008</td>
<td>0.047</td>
<td>0.004</td>
<td>0.004</td>
<td>0.023</td>
<td>0.043</td>
<td>0.003</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>ventral ectoderm anlage in statu nascendi</td>
<td>69</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>visual anlage in statu nascendi</td>
<td>17</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>yolk nuclei</td>
<td>35</td>
<td>0.419</td>
<td>0.485</td>
<td>0.823</td>
<td>0.818</td>
<td>0.639</td>
<td>0.592</td>
<td>0.841</td>
<td>0.805</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Table showing the enrichment terms for all genes within the second time window (stages 4-6) for all four similarity values: Mean Squared Error (MSE), Haar wavelets (Haar), Mutual Information (MI), and Spatial Mutual Information (SMI), given as Actual similarity scores, and reweighted significance scores (RSS). Each row represents an annotation term, light gray shading represents 0.05 significance, while dark gray represents 0.01 significance values.
of the database (Tomancak et al., 2007). The annotation set was curated by the BDGP group by manually inspecting the in situ images and providing ontology terms for each gene at every time stage. Additionally, information on the orientation of the embryos (i.e. the viewpoint) was manually curated for the stage window 4-6.

3.2 Image Registration

The approach for segmenting and registering images is based on statistical shape models using signed distance maps to describe object contours (Leventon et al., 2000). Signed distance maps are a representation of contours which contain the distance from the contour of the object for every point in the image: negative distance values depict regions that are inside the object, positive distances are outside, and the magnitude represents the actual distance. Signed distance maps are an attractive choice for shape modeling as they provide a continuous representation of a discrete space which is easily interchangeable (the signed distance map can be directly converted from the contour, and the contour can be determined from the signed distance map by calculating the zero crossing of the distance map).

A *Drosophila* shape model was automatically created from a manually curated set of 120 embryo images. First, the contours of the embryo were manually segmented and transformed into signed distance maps. The objects were then automatically normalized in size by minimizing the distance of each individual signed distance map to the mean signed distance map. The resulting normalized maps were analyzed using a hierarchical Principal Component Analysis (PCA) decomposition (Johan A. Westerhuis, 1998). Let $X$ be the set of all training images, and $X^B \subseteq X$ be the subset of training images belong to time stage $b$, where $b \in B, B = [1, 6]$, and $X^i \cap X^j = \emptyset \mid i, j \in B, i \neq j$. In standard PCA, a new set of bases $w_i \in W$ is selected such that $w_i = \arg \max \text{var} \{x_i x'_{-1}\}$ where $x'_{-1}$ is the deflated matrix from the previous iteration. Each vector $w_i$ is then converted into a two dimensional signed distance principal component image. Hierarchical PCA extends upon PCA by normalizing the contribution of each basis to each individual block.

In addition to providing characteristic priors on the shape through hierarchical PCA, we also model the filtered intensity values around the contour of the embryos. We create a set of histograms of the intensity values by binning observed intensity values by their respective signed distance. We concentrate on the intensity values close to the contour, i.e. we bin the intensities observed in distances from -25 to 25 in 1 unit increments, while remaining bins are not included.

The task of image registration is to find the optimal set of parameters, $\theta$, such that the images are accurately aligned to a common frame of reference. The parameters we optimize over can be separated into two categories: rigid transformation parameters of the image $\theta_r$, and the principle shape

![Fig. 2. Generation of background data sets and pairwise significance tests for 2D expression patterns. Upper Right Panel: Computation of significance values, as exemplified on in situ images for the genes nubbin (A) and dichaete (C). The extracted expression vectors (bottom of (A,C)) of nubbin and dichaete are used to calculate the background distribution specific for the comparison of these two expression patterns. For each gene, a set of random realizations with constraints on the correlation between spatially adjacent expression values is created as described in 3.5 (B,D). The constrained realizations are used to compute a background distribution of similarity values (histogram in (E)). The observed similarity on the expression patterns is indicated with an orange line, and is used to calculate an empirical $p$ value for each comparison. Lower Left Panel: A set of previously described gap, pair-rule, and segmentation genes (Schroeder et al., 2004) was used to evaluate the significance testing. Images for each gene were registered, and their respective column expression vectors were calculated. Using the constrained realizations (histograms) and observed similarity values (orange lines), significance values were calculated for each similarity score. The color of the histogram represents the significance of the pairwise score (blue: $> 0.1$, green: $(0.1, 0.05]$, yellow: $(0.05, 0.01]$, red: $< 0.01$). The results in this example explore anterior-posterior patterning and are based on the x-axis-projections of the expression patterns; for the comparisons in Table 1, projections to both x- and y-axes were used.](http://bioinformatics.oxfordjournals.org/)
components of our shape prior $\theta_s$. The principle component parameters provide a non-rigid component to the registration, allowing the underlying shape model to take on a variety of possible shapes which are defined by the signed distance maps. The rigid transformation parameters are used to rotate, translate, scale, and flip (horizontally and vertically) each image so that it maps onto the evolving shape model. These two sets of parameters are simultaneously optimized using an in house implementation of a Particle Swarm Optimizer (Kennedy, 1995).

By using the shape parameters as well as the empirical histogram values, we assess how well the image is registered with the following metric:

$$ f(\theta_x, \theta_y) = \alpha_1 g(s,g) + \alpha_2 h(s,i) \quad (1) $$

$s = \theta_s W$ is the signed distance map created from the linear combination of signed distance principal component images. $g(s,g) = \sum_{j=1}^{n} \left( \frac{1}{1 + s_j^2} (s_j - g_j) \right)$ is an extension to Leventon’s original scoring function, and $h(s,i) = \sum_{j} p(s_j, t_{ij})$ is the empirical probability of observing a pixel value of $i$ at the signed distance location $s_j$ over all $s_j \in (-25, 25)$. To summarize, our method differs from Leventon’s method by three main distinctions. First, the effect of the intensity distribution is limited to the area around the zero crossing of the evolving shape. This is necessary to limit the contribution of internal staining and multiple/impedding embryos to the score. It also allows for a narrow band approach to be used with minimal effect on the overall score which increases the runtime performance. Second, the variability in size of the objects is normalized prior to creating the shape model. This is required to provide accurate representations of the actual differences in the shape of the Drosophila embryo which are not simply related to size. The size of the embryo then becomes an additional parameter in our optimization step. Lastly, we provide an additional term that is based on the direct values of the filtered image. This term is not decomposed into individual components, but rather describes the distribution of the filtered image in relation to the signed distance location. This allows the registration to accurately detect and align the shape model to the correct position of the image.

3.3 Representation of Expression Patterns

After registration, the transformed image $T$ and shape model $S$ are used to calculate column and row vectors of expression data. To allow a dyadic decomposition for the Haar wavelets, 64 columns and 32 rows element vectors are used. The row and column vectors are created by dividing the bounded shape image into equally spaced rows and columns, and computing the mean pixel intensity value of the second channel of the masked image for each column and row entry $c_i$ and $r_j$. Where $c_i = \frac{1}{n_{\text{row}}} \sum_{j} p(s_j, t_{ij})$, and $o_{c1}(j)$ is a mapping for column region $c_1$ of all pixels within the shape model ($o_{c1} = 1$) and $n_{\text{row}} = \sum n_{c1}$. The same representation is used for the rows, resulting in two vectors of expression denoted as $c$ and $r$.

3.4 Metrics and Evaluation Measures

Four metrics were evaluated to determine the similarity between expression patterns: mean squared error (MSE), mutual information (MI), Haar wavelets (HW), and spatial mutual information (SMI). MSE and MI are two frequently used measures for comparing vectors of data. However, both assume that the samples within each vector are independent, which is not the case for spatial and time series data. Interaction terms between individual elements of a sample are capable of describing higher order structures, such as the formation of gradients, or the alternating striped pattern of odd and even-skipped. These dependencies are relevant not only from the perspective of kinetics and molecular diffusion, but also in the interactions between genes. It is for this reason that similarity metrics that account for spatial dependency can be expected to provide more biologically relevant measures for comparing images. We implement two such measures, Haar wavelets and spatial mutual information, which are the spatial counterparts of mean squared error and mutual information, respectively. For all scoring metrics, we compute similarities independently on row and column vectors, and then combine them in a sum weighted by the vector size (i.e. in our case, with 1/3 and 2/3 respectively).

3.4.1 Mean Square Error

Mean squared error scoring metrics have been previously used to compare gene expression patterns a and b (e.g. in Liu et al., 2007) on RNA in situ brain images). For each row and column we sum up the difference between the elements as defined:

$$ d_{ab}^{MSE} = \frac{1}{n} \sum_{j} \left( C_j^a - C_j^b \right)^2 \quad (2) $$

3.4.2 Mutual Information

We use the standard mutual information, e.g. as defined in (Steuer et al., 2002):

$$ d_{ab}^{MI} = H(A) + H(B) - H(A,B), \quad (3) $$

where $H(A)$ and $H(B)$ define the entropy of each variable (or in this case, the column and row vectors for each gene) defined as:

$$ H(A) = -\sum_{i=1}^{M_A} p(a_i) \log_p(a_i) \quad (4) $$

And $H(A,B)$ is the joint entropy:

$$ H(A,B) = -\sum_{i=1}^{M_A} \sum_{j=1}^{M_B} p(a_i,b_j) \log_p(a_i,b_j) \quad (5) $$

3.4.3 Haar Wavelets

Wavelet analysis allows one to simultaneously examine the frequency and resolution components of a signal. This is accomplished by iteratively decomposing the signal with high and low bandpass filters. The low frequency filter serves the purpose of downsampling the image into progressively smoother dyadic scales. Let $s \in (1, 6)$ denote the dyadic scaling factor of the low frequency filter. For the Haar wavelets, this low pass filter can then be written as: $L^s_A = 0.5 * \left( L_{2s+1}^A + L_{2s+1}^B \right)$, where $i$ represents a specific spatial location. The high pass filter is responsible for creating the coefficients of the wavelet which ultimately will be used to represent the similarity between patterns. The high pass filter is written as: $H^s_A = 0.5 * \left( L_{2s+1}^A - L_{2s+1}^B \right)$. The Haar distance can then be calculated using the Mean Squared Error: $d_{ab}^{HW} = \sum_{s} \left( H_{a,s} - H_{b,s} \right)$.

3.4.4 Spatial Mutual Information

Spatial mutual information is an extension of mutual information to include neighboring dependencies and has been used in different image analysis applications (e.g. (Rodriguez-Carranza and Loew, 1998)). Instead of defining the entropy as the one dimensional probability of observing an event (or in this case, an expression value), we instead define it as the joint probability of observing an value $i$ while its neighboring expression value is $j$. The entropy values for each variable (gene) then become a two-dimensional entropy.

$$ H(A, \hat{A}) = -\sum_{i,j \in N_i} p(a_i, \hat{a}_j) \log_p(a_i, \hat{a}_j) \quad (6) $$

where $N_i$ are the neighbors of $i$, and $\hat{a}_j = |a_i - a_j|$ is the difference between neighboring elements.

We chose to use neighboring elements of distance 2. This value was selected as it preserves biological significant spatial patterns (local gradients; patterns such as pair-rule or gap), while still being computationally tractable. The cross-image comparison then becomes a 4-dimensional joint entropy:
### 3.5 Biological significance testing

The significance of a scoring metric can be computed based on a series of surrogate data sets used for hypothesis testing. We create appropriate surrogate data by drawing constrained realizations (Theiler and Prichard, 1996) which account for both the marginal distributions of the intensity values as well as the joint spatial dependencies between neighboring variables. This strategy is based on surrogate data whose spatial complexity is representative of the underlying spatial processes; genes with high spatial dependencies (e.g. smooth gradients) will result in a surrogate data set with similar gradient patterns, while those having low spatial dependencies (e.g. hard gradients) will have dissimilar patterns. This approach requires the use of a sampler capable of drawing values from both the marginal as well as the spatial dependencies.

To meet these requirements, we sampled as follows: For each column vector, we initialize every element by drawing a sample from the marginal distribution. This initialization provides us with column vectors that have no spatial dependency. To account for the spatial dependencies, an iterative swap sampler is used on this random initialization. Let \(p^1(a_{i,j})\), \(p^p(a_{i,j})\), \(p^p(a_{i,j})\) be the probability of observing an intensity \(a_{i,j}\) while its first, second, and third neighbor elements are \(a_{i,j}, a_{i,l}, a_{i,k}\), where \(j \in N_3^3, k \in N_3^2, l \in N_3^2\) are neighborhood association sets. To account for edge effects, neighbor relationships are considered on a torus. Let \(p(a_{i,j})\) be the cumulative probability of observing \(a_{i,j}\). For each iteration, four random locations are chosen: \(K = k_i, k_j, k_e, k_f\) with observed values \(L = \{o(k_i), o(k_j), o(k_e), o(k_f)\}\) and \(h : K \rightarrow K\) be a permutation of the locations with \(h = h_1, \ldots, h_4\) being the set of all permutations of \(K\) with \(h_i(k_i, k_j, k_e, k_f) = h_i\) being the permuted locations with observed values \(L_i = \{o(k_{i,j}), o(k_{i,l}), o(k_{i,k}), o(k_f)\}\). The probability of a swap is calculated for all permutations (including the identity permutation) and the most probable swap is chosen. By iteratively resampling from this distribution, surrogate data sets are created that account for both the marginal as well as the spatial dependency structure.

#### 3.5.1 Significance Testing

40 constrained realizations are created for each gene. The background similarity value is constructed by calculating all 1,600 pairwise similarity values between the constrained realizations for each gene denoted \(b^s_{i,j}\) for \(s \in \{1, 1600\}\). The empirical \(p\)-value of the expression between gene \(i\) and \(j, p_{i,j}\), is calculated by counting the number

\[
H(A, \hat{A}, \hat{B}, \hat{B}) = -\sum_{i=1}^{N} \sum_{j \in N_k} \sum_{l \in N_{-k}} d_{s,t}^{a_{i,l}} \log p(a_{i,l}) + \log(1 - p(a_{i,l}))
\]

\[
p(x) = \frac{1}{N h^{2}} \sum_{i=1}^{N} \exp \left( \frac{-(x - x_i)^2}{2h^2} \right),
\]

where \(h\) is the bandwidth parameter and controls the smoothness of the estimate. This window allows us to provide a smoother and more robust mutual information calculation when dealing with sparse data.

**Ontology Annotation Term Assessment**

The fly embryo database frequently contains more than one image per gene. A pairwise blocked matrix was created to represent the similarities between genes where \(d_{s,t}^{a_{i,l}}\) represents the distance from the \(s\)th image of the \(i\)th gene to the \(t\)th image of the \(j\)th gene. Let \(D_{i,j}\) represent the distance between gene \(i\) and \(j\) where \(D_{i,j} = \min(d_{s,t}^{a_{i,l}})\) for the MSE and Haar metrics and \(D_{i,j} = \max(d_{s,t}^{a_{i,l}})\) for MI and SMI. For each gene \(i\) and each annotation term \(k\) present for that gene \(a_{i,k} = 1\), we perform a Mann-Whitney-Wilcoxon test. To accommodate for multiple disjoint annotations, we performed the rank test on all genes \(j\) with the same annotation \(a_{j,k} = 1\), or genes with different annotations, but did not share any additional annotations of the original gene \(a_{j,k} = 0\) if \(a_{j,k} = 1\). Let \(U_{i,k}\) represent the Wilcoxon signed rank value for gene \(i\) and annotation term \(k\). The significance for each annotation term, \(k\), was calculated by taking the expectation of the U statistic \(E(U_{i,k}) = \frac{1}{2} \sum U_{i,k}^2\), and calculating its resulting z-score \(z^k = (E(U_{i,k}) - \mu_{U_{i,k}}) / \sigma_{U_{i,k}}\). The \(p\)-values for each annotation is calculated directly from the z-scores.
of background scores greater than the observed score \( |b_{i,j}^s| > d_{i,j} \), and dividing it by the number of elements \( p_{i,j} = |b_{i,j}^s| > \frac{d_{i,j}^M}{1600} \).

### 3.6 Reweighted Significance Scores

To account for low staining and high lighting variability commonly observed in the data, we compute a reweighted significance score (RSS) between images \( s, j: RSS_{i,j} = p_{i,j} + (1 - p_{i,j})/X \), where \( X = \exp(\sqrt{\text{var}(s)} \times \text{var}(j)) \). The RSS score is a smooth alternative to thresholding: as the variability in the images increases (stronger signal), the RSS score approaches its actual p-value. Images with low variability (weak signal/stain), have their scores reweighted closer to 1.

### 4 DISCUSSION

To address questions that arise with the analysis of spatial gene expression patterns, we have implemented a complete pipeline for 2D *Drosophila* RNA in situ imaging datasets pertaining to developmental biology. Our work is certainly not the first computational study based on *Drosophila* digital microscopy images. Some other fly datasets (Janssens et al., 2005b; Keranen et al., 2006; Fowlkes et al., 2008) are obtained with the purpose to study the expression of few genes but at much higher quality in terms of resolution and quantification of expression. Each registration method on these complex images is in general tailored to a specific scenario, often with well controlled imaging protocols/environments (e.g., single embryos, consistent staining/lighting) and small datasets (<500 images).

Some recent studies have also addressed *Drosophila* 2D RNA in situ images, on the same or similar data that we used here (Peng et al., 2007; Kumar et al., 2002). Our goal was to compare and identify pairwise spatial expression patterns, with an example in gene regulatory interactions, and is thus most similar to the work of (Kumar et al., 2002), a system for database retrieval of RNA in situ images based on the global expression pattern. Other studies focused on classifying or clustering patterns, e.g. with the purpose of classifying images into gene ontology annotation terms (Ye et al., 2008; Zhou and Peng, 2007). These different goals show the breadth of possible applications on this new type of high-throughput data; however, this also means that direct comparative assessments between them are difficult, even for shared subtasks such as image registration.

Our shape-based registration method is fully automated and applied on all 78,621 images in the current release of the expression pattern database. In contrast, other approaches typically selected small application-oriented subsets, or used a registration process that was manually curated, and these sets do not show enough overlap to create a common gold standard set at this point. We did however compare the performance of several distance metrics used within other systems (mean squared error (Liu et al., 2007), mutual information (Peng and Myers, 2004), and wavelets (Peng et al., 2007)), with the result that the proposed spatial mutual information is more adequate, and that many of the previous metrics can achieve similar performance by incorporating the spatial structure in a significance score.

While our method for the automated registration of images is more robust and complete than many previous approaches, it cannot be an off-the-shelf solution for all image registration problems. Differences in sample preparation and imaging technology, as well as differences in the morphology of specimens which are imaged, pose restrictions on the possible choices of analysis methods. For instance, our previous work on Arabidopsis root images (Mace et al., 2006) deals with confocal images of GFP reporter constructs. In addition to the different morphology, the fly dataset poses additional challenges of dealing with multiple objects, and a high variability in imaging conditions; conversely, we needed to apply non-rigid registration for the plant roots. Further, while our method for significance testing and surrogate data set generation are well suited to the syncytium formation of the *Drosophila* embryo, this is generally not applicable across all organisms. For example, in *C. elegans*, gene expression is heavily influenced by cell lineage, and a more appropriate representation for neighboring dependencies would be between neighboring lineage elements, and not direct spatial location. Additionally, different sampling methods may need to be applied to deal with differences in the structure of the organism. While general platforms for cell-based image analysis have been published (Carpenter et al., 2006), a universal framework for complete multicellular organs or organisms has to be placed at a higher level and poses a challenge for the automated analysis of image expression datasets pertaining to developmental biology.
5 CONCLUSION
As images are increasingly becoming a source of high-throughput genomics data, theoretically sound approaches to evaluate them become necessary. Whatever the specific context, appropriate quantitative methods along the lines proposed in this study will allow us to move image expression data from qualitative descriptions to the quantification of gene expression, and to its use as phenotype for which we can assess the significance of changes under perturbations of the genotype or the environment.

REFERENCES