Automatic Classification of Embryonic Fruit Fly Gene Expression Patterns

Andreas Heffel[1], Sonja J. Prohaska[1,2], Peter F. Stadler[1−5], Gerhard Kauer[6], Jens-Peer Kuska[1]

[1]Interdisciplinary Centre for Bioinformatics, University of Leipzig,
[2]Department of Computer Science, University of Leipzig,
[3]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, Leipzig;
[4]Institut for Theoretical Chemistry, University of Vienna; [5]Santa Fe Institute;
[6]Faculty Technical Sciences, University of Applied Sciences O/O/W, Emden
heffel@izbi.uni-leipzig.de

Abstract. Carefully studied in-situ hybridization Gene expression patterns (GEP) can provide a first glance at possible relationships among genes. Automatic comparative analysis tools are an indispensable requirement to manage the constantly growing amount of such GEP images. We present here an automated processing pipeline for Segmenting, Classification, and Clustering large-scale sets of Drosophila melanogaster GEP images that facilitates automatic GEP analysis.

1 Introduction

With the advent of large-scale automated whole-mount in-situ hybridization (ISH), the boom in high throughput techniques has also seized modern developmental biology [1, 2, 3, 4]. The spatio-temporal expression patterns of genes often provide rapid direct insight in the potential functional roles of novel or poorly studied genes and allows at least a first glance at possible relationships among genes. This can be exploited to develop new diagnostic methods. Where single genetic markers are not sufficient to distinguish healthy tissue from a diseased phenotype, combinations of genes can enhance diagnostic strength [5]. Microscopy and comparative analysis of images is therefore becoming an important tool to study the spatio-temporal distribution of multiple gene products. For Drosophila, gene expression patterns for all (about 13,500) genes across a set of developmental stages are being generated systematically, leading to the accumulation of a wealth of high dimensional data with complex encoding. Bioinformatics currently helps with data management, storage, access, and integration (e.g. BDGP [6, 4, 7, 8] and FlyBase [9]), but could also assist in data processing, data analysis and provide an observation tool for researchers in the field. Such an attempt has been made by Fly Express [10]. With 57,083 images mapping the expression patterns of 3,366 genes (April 30, 2008), it is currently the largest database that computes and holds standardized images. It offers the Basic Expression Search Tool, BEST, [11], to retrieve genes with expression patters similar to a given query pattern. An alternative and more robust set of

algorithms was proposed by Peng et al. [12, 13]. Here, we describe a processing pipeline for automatic segmentation, classification and clustering of ISH images, which is based on the representation of expression patterns by means of Bessel eigenfunctions. This spectral representation facilitates a faster and easier pattern classification once the coefficients are calculated. We apply our approach to a subset of 681 genes for which images are available in Fly Express for all 6 developmental stages.

2 Materials and methods

As previously explained in our recent publication [14] the elaborated processing pipeline include six basic steps. The preprocessing (i) consists of a shading correction and a contrast optimization method resulting in a "'clean image"'. The shape segmentation (ii) is executed on the clean image by computing the magnitude of the gradient followed by a fragmentation of the gradient feature space with the Gaussian Mixture Model [15, 16] instead of using a static threshold. The result is a probability map describing the probability for the membership to one of the classes (background, embryo). After filtering with a total variation filter, a smoothed mask is obtained with an unpredictable count of holes in it, because some areas in the embryo have feature values similar to the background. We close the holes and obtain the binary Segmentation Result. Gradient vector field snakes are used in case of several touching embryos, to isolate the embryo (iii) [17, 18]. The binary images are rigidly registered onto an ellipse prior to the snake segmentation, to allow an automatic placement of the initial contour.

The registration (iv) of the extracted embryo outline onto an ellipse is a necessary step to align the shapes and hence the patterns consequently. This is done in two steps: first, we apply a rigid registration [19] and second, a nonlinear registration onto the ellipse [20, 21]. The found transformations onto an ellipse, computed from the registration of the embryo outlines, are then applied to the masked embryo images to obtain the expression pattern mapped onto an ellipse. Due to the fact that the orthonormal system that we have chosen for the pattern representation is defined on a circle, the ellipse is finally stretched to circular shape.

The GEP extraction (v) step has been modified from the GMM method described in [14] to a HSV color space transformation. The pattern is extracted by taking the color intensity information from the V-channel and setting all values smaller than 20 percent to zero (denoising).

Finally the pattern classification (vi) step can be computed on the extracted GEP sets. Thereby the patterns $\mathcal{P}(r, \phi)$ are described by a set of Fourier coefficients:

$$\mathcal{P}(r, \phi) = \sum_{j=1}^{\infty} \sum_{k=0}^{\infty} a_{j,k}\, \psi_{j,k}(r, \phi) \tag{1}$$

As basis, the eigenfunctions of the Laplace operator on a circle of radius ℓ,

$$\psi_{j,k}(r, \phi) = N_{j,k}\, e^{ik\phi} J_k \left(\frac{r j_{k,j}}{\ell} \right) , \tag{2}$$

are used, where ℓ is the radius of the circle, $J_k(z)$ are the k-th Bessel function, $j_{k,j}$ is the j zero of the k-th Bessel function and $N_{j,k}$ is a normalization factor.

3 Results

The gradient based GMM shape segmentation (ii) produces a reliable and accurate separation of the embryos on the border of the focal plane. About 30 percent of the examined images contain several touching embryos. In this case additional knowledge about the expected shape must be included to obtain the mask of a single embryo. This is accomplished by using the active contour approach (iii). The subsequently application of the rigid and nonlinear registration steps (iv) minimizes the distortions produced by the nonlinear registration. The pattern extraction step (v) using the HSV color space transformation is a simple but effective method. In cases of different staining colors or very dark (black) staining regions the HSV extraction method provides robust results.

The reason for choosing the Bessel base functions for the pattern representation is that they form a complete orthonormal system. Our empirical studies and reflections about the possible complexity of the GEP shows that every pattern can be adequately expressed by a set of 420 eigenfunctions ($k \in [0, \ldots, 20]$, $j \in [1, \ldots, 20]$). Fig. 1 summarizes the processing pipeline using the example of one image showing several coherent embryos.

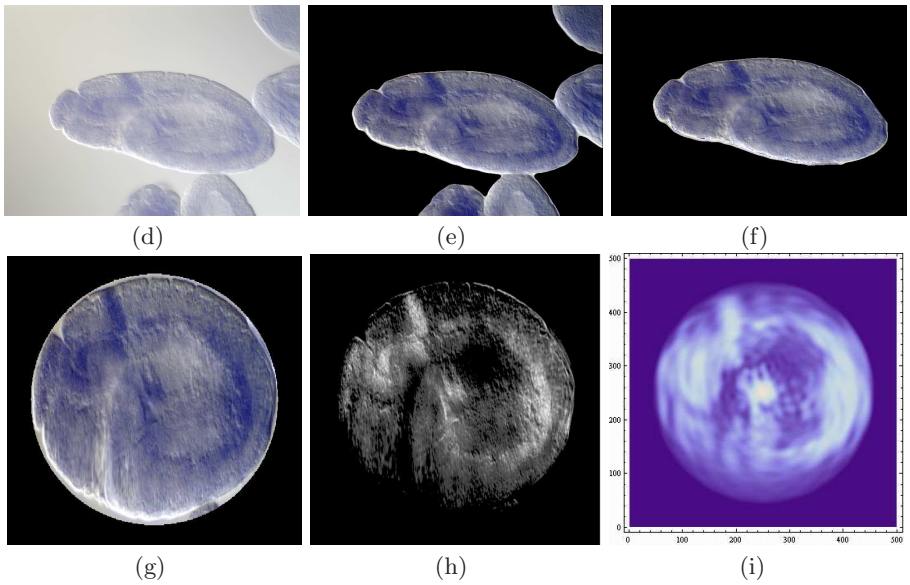
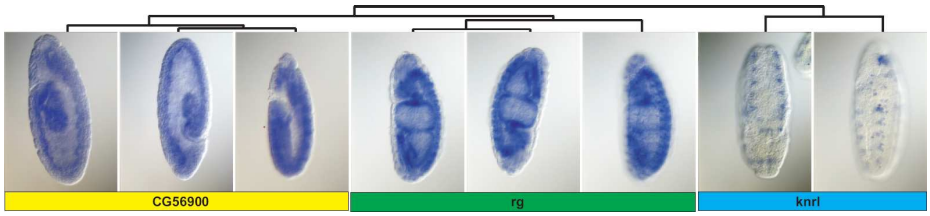


Fig. 1. Illustration of the process flow steps. (a) original image showing several coherent embryos. (b) preprocessing and shape segmentation result (c) isolation of the embryo by snake segmentation (d) registration to circular shape (e) GEP extraction (f) reconstructed GEP from the 420 Fourier coefficients.

Fig. 2. Hierarchical clustering on a set of 8 images that demonstrates that images representing the same genes (CG5690, rg, knrl) in the same developmental stage (Stage 5) are correctly grouped together.



4 Discussion

Naturally, the pictures of Drosophila embryos are much simpler in structure than many other ISH data, which show expression patterns superimposed on complex morphologies (such as mouse or zebrafish embryos). However, the examined images show a lot of difficulties such as blurred contours, background shading, coherent partial embryos, and different staining colors. Our approach is capable of dealing with such complications and achieve a considerable reduction of the data amount needed to represent the patterns. The Euclidean distance in the truncated 420 dimensional space of Fourier coefficients can be used for hierarchical clustering. This allows a classification of expression patterns. As an example, Fig. 2 shows that similar expression patterns, in this case different image representing the same gene, are indeed clustered together. Larger data sets of images show a similar agreement with the visual expectation (data not shown). In particular, images showing the same gene in the same developmental stage are most often grouped together.

In practical applications, the orientation of the embryo with respect to the two axis of symmetry of the ellipse is important. At present, the correct orientation is not generated automatically. Unless annotated in the initial image, all four orientations are included in the clustering and in the end an expert control [6] is inevitable.

References

1. Gawantka V, Pollet N, Delius H, et al. Gene expression screening in Xenopus identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. Mech Dev. 1998;77:95–141.
2. Neidhardt L, Gasca S, Wertz K, et al. Large-scale screen for genes controlling mammalian embryogenesis, using high-throughput gene expression analysis in mouse embryos. Mech Dev. 2000;98:77–93.
3. Kudoh T, Tsang M, Hukriede NA, et al. A gene expression screen in zebrafish embryogenesis. Genome Res. 2001;11:1979–1987.
4. Tomancak P, Beaton A, Weiszmann R, et al. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biology. 2002;3:0088.

5. Schubert W, Bonnekoh B, Pommer AJ, et al. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. Nat Biotechnol. 2006;24:1270–1278.
6. BDGP: Berkeley Drosophila Genome Project. http://www.fruitfly.org/.
7. Tomancak P, Berman BP, Beaton A, et al. Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biology. 2007;8:R145.
8. Montalta-He H, Reichert H. Impressive expressions: Developing a systematic database of gene-expression patterns in Drosophila embryogenesis. Genome Biology. 2003; p. 205–205.
9. FlyBase: A Database of Drosophila Genes and Genomes. http://flybase.bio.indiana.edu/.
10. Van Emden B, Ramos H, Panchanathan S, et al. FlyExpress: An image-matching web-tool for finding genes with overlapping patterns of expression in Drosophila embryos; 2006. www.flyexpress.net.
11. Kumar S, Jayaraman K, Panchanathan S, et al. BEST: A novel computational approach for comparing gene expression patterns from early stages of Drosophila melanogaster development. Genetics. 2002;162:2037–2047.
12. Peng H, Myers EW. Comparing in situ mRNA expression patterns of Drosophila embryos. In: Proc Int Conf Res Comput Mol Biol; 2004. p. 157–166.
13. Peng H, Long F, Zhou J, et al. Automatic image analysis for gene expression patterns of fly embryos. BMC Cell Biology. 2007;8 (Suppl 1):S7.
14. Heffel A, Stadler PF, Prohaska SJ, et al. Process flow for classification and clustering of fruit fly gene expression patterns. Proc IEEE ICIP. 2008; p. 721. http://www-video.eecs.berkeley.edu/Proceedings/ICIP2008/ICIP2008.html.
15. Pernkopf F, Bouchaffra D. Genetic-based EM algorithm for learning Gaussian mixture models. IEEE Trans Pattern Anal Mach Intell. 2005;27:1344–1348.
16. Moon TK. The expectation-maximization algorithm. IEEE Signal Processing Magazine. 1996;(6):47–60.
17. Xu C, Prince JL. Snakes, shapes, and gradient vector flow. IEEE Trans Image Process. 1998;7:359–369.
18. Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. Int J Computer Vis. 2004;1:321–331.
19. Gottesfeld Brown L. A survey of image registration techniques. ACM Computing Surveys. 1992;24:325–376.
20. Braumann UD, Kuska JP. A new equation for nonlinear image registration with control over the vortex structure in the displacement field. Proc IEEE Int Conf Image Process. 2006; p. 329–332.
21. Braumann UD, Kuska JP. Influence of the boundary conditions on the result of non-linear image registration. Proc IEEE Int Conf Image Process. 2005; p. I-1129–I-1132.