Registering *Drosophila* Embryos at Cellular Resolution to Build a Quantitative 3D Atlas of Gene Expression Patterns and Morphology

Charless C. Fowlkes[1], Cris L. Luengo Hendriks[2], Soile V. E. Keränen[2], Mark D. Biggin[2], David W. Knowles[2], Damir Sudar[2], Jitendra Malik[1]

Berkeley Drosophila Transcription Network Project
1. Department of Electrical Engineering and Computer Science, University of California, Berkeley
2. Life Sciences and Genomics Divisions, Lawrence Berkeley National Laboratory

Abstract

The Berkeley Drosophila Transcription Network Project is developing a suite of methods to convert volumetric data generated by confocal fluorescence microscopy into numerical, three dimensional representations of gene expression at cellular resolution. One key difficulty is that fluorescence microscopy can only capture expression levels for a few gene products in a given animal. We report on a method for registering 3D expression data from different Drosophila embryos stained for overlapping subsets of gene products in order to build a composite atlas, ultimately containing co-expression information for thousands of genes. Our techniques have also allowed the discovery of a complex pattern of cell density across the blastula that changes over time and may play a role in gastrulation.

1. Introduction

The output of animal gene transcription networks are complex 3D patterns of expression. These dynamically changing patterns can vary radically from one cell to the next and it is these spatial variations that principally define and determine the morphology and physiology of the animal. It is a grand challenge of biology to learn how to decipher the transcriptional information in the genome to the point where we can predict and model such intricate patterns. Researchers typically analyze spatial patterns of gene expression in multicellular organisms by visual inspection of photographic images of in situ hybridization experiments. A profound limitation of this approach is that it captures little 3D structure and provides only rough quantitative information. If we are to fully describe and understand such systems, we first need a means to quantitatively record, with cellular resolution, these phenomenally complex expression patterns.

Our focus is on the network controlling early *Drosophila melanogaster* embryogenesis prior to gastrulation. This transcription network is initialized by the maternal deposition of mRNAs and proteins from a handful of regulators into the egg. Zygotic transcription begins two hours after fertilization at which point the embryo is a syncytium with around 2,000 nuclei located along the perimeter. Thirty or so zygotic regulatory genes are transcribed initially, controlled by the maternally deposited factors. The Berkeley Drosophila Transcription Network Project [1] is focused on understanding interactions among these transcriptional regulators and their $\sim$1,000 target genes during the first 100 minutes of zygotic transcription. The availability of a complete genome sequence, powerful molecular and genetic tools, and the fact that there are only tens of specific factors regulating the early spatial patterning of transcription make *Drosophila* a desirable system for studying transcription networks.

Laser scanning confocal microscopy is a key enabling technology which allows us to measure the relative concentrations of gene products over a 3D volume. Using state of the art 2-photon equipment it is possible to image the entire *Drosophila* blastula with a resolution at which individual nuclei are still distinguishable. While fluorescence imagery is an incredibly rich source of data, it falls short of our goals since it can only provide expression data for 2 or 3 gene products in a given animal.[1] *How can we capture the expression levels of a thousand different genes over the entire blastula at cellular resolution?*

Our approach is to composite expression measurements made from multiple images into a common model where the spatial distribution of more than three gene products can be studied simultaneously. The key step is identifying corresponding nuclei in two different embryos and then using this correspondence to transfer the expression levels from

[1] When expression patterns are spatially disjoint, careful multiplexing techniques have provided up to 10 channels [9, 8]

Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops (CSBW'05)
0-7695-2442-7/05 $20.00 © 2005 IEEE

one to the other. If the embryos were identical and imaged in the exact same pose, finding correspondences would be easy. Unfortunately this is not the case, as morphology, including the number of cells, varies even between embryos at the exact same developmental stage. In addition, the imaging process introduces further complications in the form of variability in the pose, staining, and deformation of the embryo during fixation. Our contribution is a technique for finding correspondence between embryos which is robust to these modes of variation.

To our knowledge this three-dimensional registration problem at cellular resolution has not been studied in the past. The closest work to ours is that of [12] who use spline and wavelet based registration techniques in 1D along the anterior-posterior axis of the *Drosophila* embryo . More recently they have considered a 2D version of the problem in order to register the lateral surface of an embryo which has been "flattened" before imaging [13].

2. Finding Correspondences

Figure 1 gives an overview of our technique. We stain each embryo with a marker for DNA, a gene of interest, and a common "reference gene" which guides the registration process. Using the DNA marker, we segment out the location of each nuclei and record its 3D location along with expression levels in the surrounding cytoplasm [10]. Given two such sets of extracted nuclei locations, we formulate correspondence as an optimization problem whose solution simultaneously yields a mapping between nuclei as well as a smooth, non-linear spatial deformation that warps one embryo onto the other.

Our approach to modeling shape variations falls into the category of "deformable templates" (e.g. [5]). To register a pair of embryos, we repeatedly alternate between two steps:

- **Correspondence**: For each nuclei in the first embryo, locate the "best" corresponding nuclei in the second embryo. This is formulated as a linear assignment problem with outliers where the matching weight includes a distance term and an appearance term dependent on the expression pattern. An efficient algorithm [4] finds solutions for thousands of cells in a few seconds

- **Warping**: Use the correspondences to estimate a coordinate transformation that maps the first embryo to the second. We perform warping using the regularized thin-plate spline (TPS) which is a non-parametric model for representing flexible, global coordinate transformations [6, 11]. The thin plate spline is the higher dimensional generalization of the cubic spline. It has been shown effective for modeling variation in biological forms [3] and can be easily fit by solving a compact system of linear equations [14].

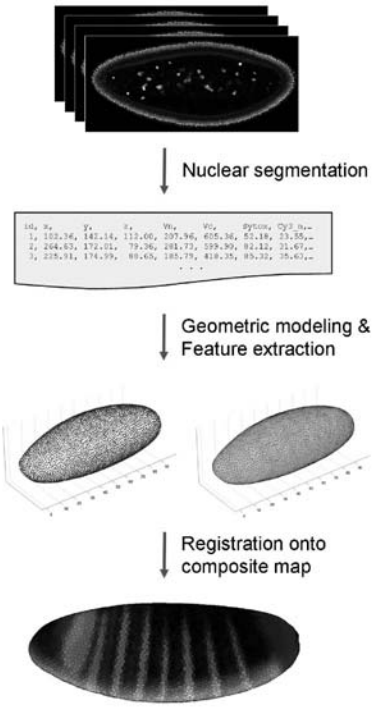

Figure 1. Data-flow: Imaging each embryo produces a high-resolution volumetric image. Individual nuclei are segmented out using a DNA stain. The result is a "point cloud" containing the location of each nucleus and the apical, basal and nuclear fluorescence levels of the two other channels imaged, along with morphological measurements such as the estimated nuclear and cellular volumes [10]. From this data, we produce a geometric model and extract features which capture the local pattern of expression for a reference gene in the neighborhood of each cell. These features along with the nuclei locations are then used in order to identify corresponding cells and register the embryo onto a composite model.

After warping, the embryos are better aligned making it easier to solve the correspondence problem, which improves the estimate of the transformation, bettering the alignment... and so on.

The typical problem encountered when applying deformable template methods is that of establishing the initial correspondence between two embryos. The syncytial blastoderm has very little in the way of unique morphological structures for identifying particular nuclei. Even using the expression levels of one or two gene products at a point is not sufficient. We overcome this problem by using a richer spatial descriptor termed *shape context* (see Figure 2) which has been successfully applied to a range of 2D and 3D shape matching problems in computer vision [2, 7].

Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops (CSBW'05)
0-7695-2442-7/05 $20.00 © 2005 IEEE

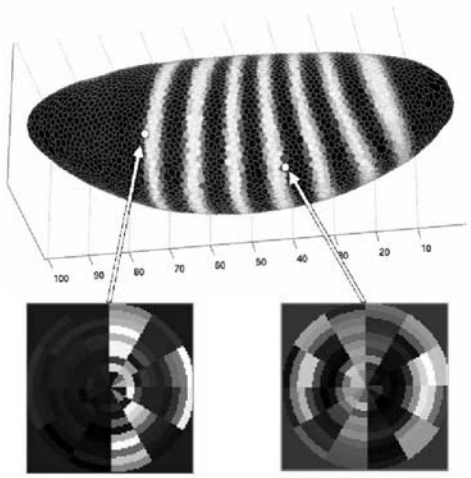

Figure 2. Shape Context Descriptor: For each nucleus we associate a shape context descriptor which is a spatial histogram of the expression levels of cells in the neighborhood of that point on the blastula surface. The polar spatial structure of the histogram bins provides robustness by making the descriptor more sensitive to the expression levels of nearby cells than to those of points farther away. Linear assignment finds corresponding cells in two embryos which are nearby and have similar shape context descriptors.

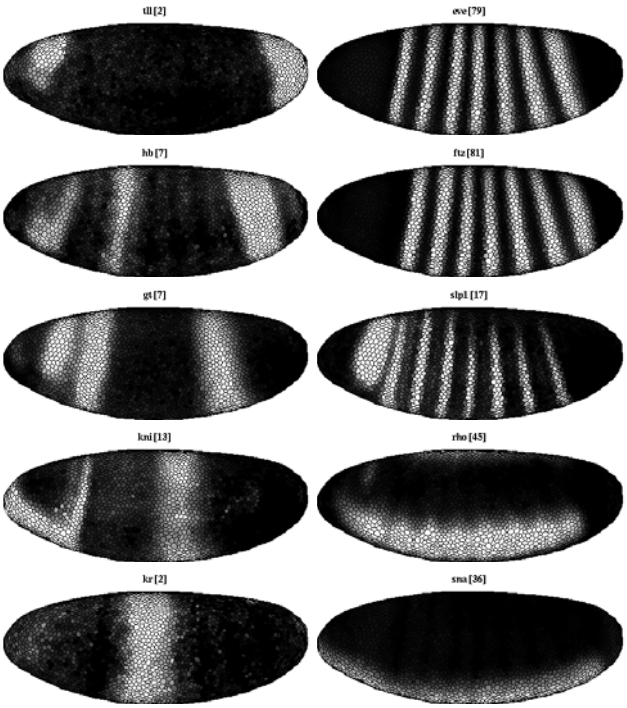

Figure 3. Composite Atlas: Views of a preliminary composite atlas of gene expression demonstrating the output of our algorithm. For each nucleus in the composite, we have constructed an estimate of the expression levels for 14 different genes, a subset of which are shown here. When multiple embryos are stained for the same gene, we can utilize the correspondence to average these measurements on the composite model. As the staining process is stochastic, averaging multiple fluorescence measurements can decrease the variance in our estimate of the "true" in vivo expression levels. Numbers in brackets indicate how many embryos stained for that gene were averaged together.

3. Exploiting Correspondences to Build Composite 3D Maps

Iterating the two steps of correspondence and warping will, given two similar but not identical embryos, yield a set of correspondences between nuclei in the two animals. Once a collection of embryos have been registered onto a common target embryo, we can use the correspondence to transfer expression levels onto that target embryo. Figure 3 shows a preliminary composite atlas constructed in this fashion from 150 confocal images using the pair rule genes *ftz* and *eve* as the reference for registration.

Once constructed, we can easily visualize and study any combination of genes in the composite atlas. This capability should be quite useful in a high throughput setting with 1000s of genes since even producing the $O(N^2)$ raw images required to see every pairwise combination involves an infeasible number of in situ hybridizations. In contrast, our method only requires $O(N)$ experiments, one for each of N genes paired with a reference gene.

In addition to expression patterns, the correspondence can also be used to average other quantities such as morphological measurements. For instance, we have begun to study the density of cellular packing in the blastoderm. We measure the spacing between segmented nuclei in each source embryo point-cloud and then use the correspondence to transfer these noisy raw measurements onto the target embryo where they are averaged together. This method has revealed a complex spatial pattern of nuclear densities which changes over time and may play an interesting role in the mechanics of gastrulation. Such subtle morphological features are not readily apparent in visual examination of a single embryo but are easily discovered in our quantitative composite atlas.

References

[1] `http://bdtnp.lbl.gov`

[2] S. Belongie, J. Malik, J. Puzicha (2002). "Shape Matching and Object Recognition Using Shape Con-

Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops (CSBW'05)
0-7695-2442-7/05 $20.00 © 2005 IEEE

texts" IEEE Trans. Pattern Analysis and Machine Intelligence, vol 24, pp. 509-522.

[3] F.L. Bookstein, (1989). "Principal Warps: Thin-Plate Splines and Decomposition of Deformations," IEEE Trans. Pattern Analysis and Machine Intelligence, vol 11, no. 6, pp. 567-585.

[4] B. V. Cherkassky and A. V. Goldberg (1995) "On Implementing Push-Relabel Method for the Maximum Flow Problem," th Integer Programming and Combinatorial Optimization Conference, p 157-171

[5] H. Chui, and A. Rangarajan. (2000). "A New Algorithm for Non-Rigid Point Matching," Proc IEEE Conf. Computer Vision and Pattern Recognition, pp 44-51.

[6] J. Duchon, (1977). "Splines Minimizing Rotation-Invariant Semi-Norms in Sobolev Spaces," Constructive Theory of Functions of Several Variables, W. Schempp and K. Zeller, eds., pp. 85-100, Berlin: Springer-Verlag.

[7] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik. "Recognizing Objects in Range Data Using Regional Point Descriptors," European Conference on Computer Vision, Prague, Czech Republic, 2004

[8] D. Kosman, C. M. Mizutani, D. Lemons, W. G. Cox, W. McGinnis, E. Bier. (2004). "Multiplex Detection of RNA Expression in *Drosophila* Embryos." Science 305: 846-846

[9] J. M. Levsky, S. M. Shenoy, R. C. Pezo, R. H. Singer (2002) "Single-Cell Gene Expression Profiling" Science 297: 836-840

[10] C. L. Luengo Hendriks, S. V. E. Keränen, G. H. Weber, B. Hamman, M. D. Biggin, D. W. Knowles, D. Sudar *In preparation*

[11] J. Meinguet. (1979). "Multivariate Interpolation at Arbitrary Points made Simple," J. Applied Math. Physics (ZAMP), vol 5, pp. 439-468.

[12] E. Myasnikova, A. Samsanova, K. Kozlov, M. Samsonova, J. Reinitz (2001). "Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods," Bioinformatics, Vol. 17(1), pp. 3-12.

[13] A. Spirov, A. Kazansky, D. Timakin, J. Reinitz. (2002) "Reconstruction of the Dynamics of *Drosophila* Genes Expression from Sets of Images Sharing a Common Pattern" Real-Time Imaging 8: p 507-518

[14] G. Wahba (1990) "Spline Models for Observational Data" SIAM.

Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops (CSBW'05)
0-7695-2442-7/05 $20.00 © 2005 IEEE