AUTOMATED ASSAY OF GENE EXPRESSION AT CELLULAR RESOLUTION

DAVID KOSMAN, JOHN REINITZ
Brookdale Center for Molecular Biology, Box 1126 Mt. Sinai Medical School, One Gustave L. Levy Place, New York, NY 10029 USA

DAVID H. SHARP
Theoretical Division MS B285, Los Alamos National Laboratory, Los Alamos, NM 87545 USA

We have recently developed an automated image processing method for obtaining quantitative values for average levels of gene expression at the resolution of a single cell. This method is described in the present paper. We place this method within a larger framework for the study of gene regulation in *Drosophila*, stressing that gene circuit models and improved data processing methods are mutually reinforcing approaches to this problem.

Introduction

The fundamental question of molecular biology as applied to multicellular organisms is: *How do cells containing the same genetic material come to express different sets of genes in a precise spatio-temporal pattern?* Implicit in this question is the central importance of studies of gene expression at the cellular level, with important consequences for both theory and experiment. With regard to theory, it emphasizes the importance of approaches which take explicit account of the fact that identical genetic material may have different expression states in different cells.[1] With regard to experimental work, it emphasizes the central importance of acquiring gene expression data at a fine scale of resolution.

Rapidly increasing amounts of gene expression data are becoming available. The complexity of the expression patterns and their underlying regulatory networks, the amount of data, and the fine scale resolution of the data pose severe challenges to the effective use of this data to answer important questions in developmental biology. There are two basic strategies for dealing with this problem. One is based on gene circuit models; the other on automated data processing.

Gene circuit models[2,3,4,5,6] address the problem by providing more powerful ways to organize and interpret the data, while automated data processing is a way to obtain and use more, and higher quality, data. Just as the field of ge-

nomics would be impossible without automated DNA sequencing, the study of genetic networks requires automated methods for the collection of expression data. Thus gene circuit models and improved data processing are mutually reinforcing ways to deal with the quantity and complexity of gene expression data.

A major focus of our recent work has been to develop an automated image processing method for obtaining quantitative values for average intensities of gene expression at the resolution of a single cell. The image processing component of this method is described in the present paper. The work reported here is part of a larger scale project to understand the process of segment determination in *Drosophila* by means of a specific dynamical model that is firmly based on experimental data. This application strongly influences our approach to image processing and for this reason we briefly outline, in the next section, the key features of the biological problem and of the modeling approach that we are using.

1 *Drosophila* Segmentation and Gene Circuits

Like all other insects, the body of the fruit fly *Drosophila melanogaster* is composed of repeated units called segments. Before these segments morphologically differentiate, their pattern is marked out by a chemical blueprint in a process called "determination". The chemical blueprint, or "prepattern" is constructed from patterns of proteins expressed from the segmentation genes, and so understanding segment determination is a matter of understanding how these patterns form. It is known from genetic experiments that the patterns are a result of mutual regulatory interactions among the segmentation genes, but the details of how this happens are too complex to be deduced solely by visual inspection of stained embryos. We are investigating the segment deter-

Figure 1: (Opposite Page) Late blastoderm stage embryo stained with antibodies to three different proteins. (A) FtzF1. (B) Hunchback. (C) Kruppel. (D) Kruppel (green) and Hunchback (red); note the region of overlapping expression is yellow. The inset in (D) shows the region selected for analysis with the mask superimposed in precise register. All four panels are taken from a scan of a single embryo. Nuclei are visible as small dots; anterior is to the left and dorsal is up.

Figure 3: (Opposite Page) Graph of the fluorescence intensity levels for Kruppel (green) and Hunchback (red) in each nucleus contained in the mask shown in Figure 1D as a function of position along the x axis of the image.

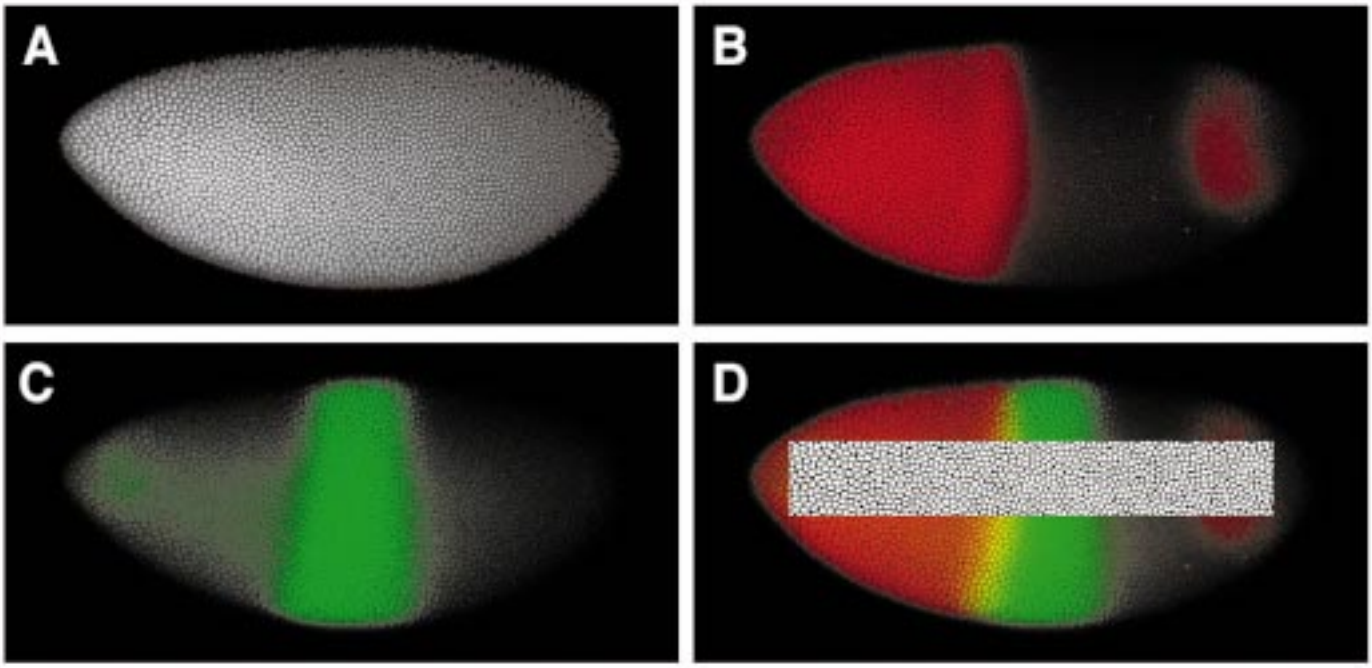

Figure 1

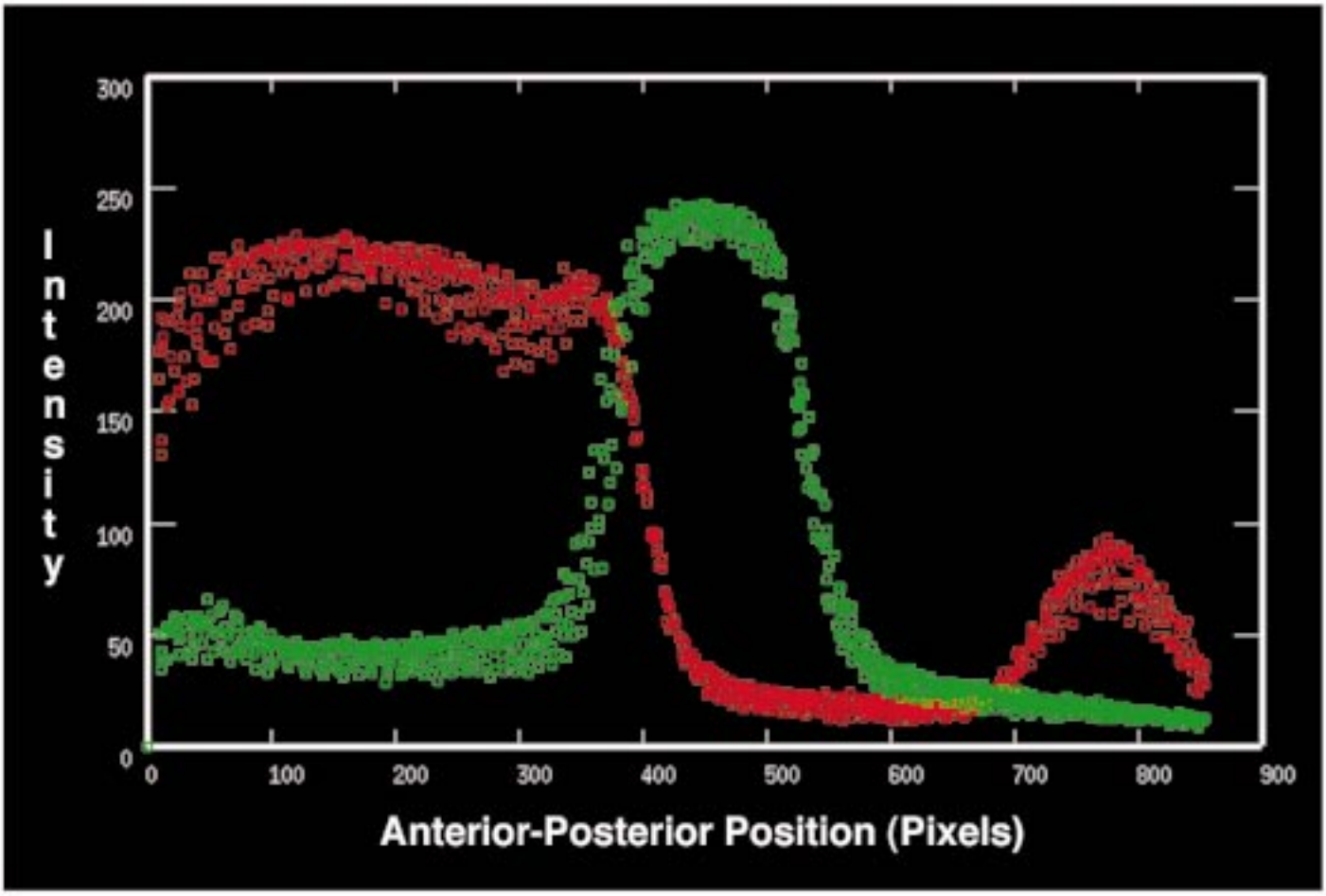

Figure 3

mination process using a model that is based on a dynamical equation for the time rate of change of protein concentrations. The parameters in this equation are determined by fits to gene expression data, as described below.

1.1 The Biological System

Segment determination takes place during the latter part of what is known as the "blastoderm" stage of development. During the time when segments are determined, the embryo consists of a roughly prolate spheroid of about 5000 nuclei. At the beginning of the segment determination process the blastoderm is syncytial (not divided into cells). During the determination process membranes invaginate and separate the nuclei into discrete cells, a process that comes to completion at the same time as segment determination. When cellularization is complete, gastrulation begins, terminating the blastoderm stage. A larva hatches about 22 hours later.

The segmentation genes are divided into 4 classes, the expression of each of which is to good approximation a function only of position along the Anterior-Posterior (A-P) axis of the embryo. The maternal coordinate genes are expressed from the maternal genome in the form of concentration gradients. Only two affect the generation of segments: *bicoid* (*bcd*) and maternally expressed *hunchback* (hb^{mat}). The gap and pair-rule genes are the main players in segmentation. Gap genes are expressed in broad domains 10-20 nuclei in width that are localized from the time that they are first detected, although they undergo some refinement as the blastoderm stage proceeds. The expression patterns of two gap genes, *Kruppel* (*Kr*) and *hunchback* (*hb*) are shown in Figure 1. We remark that *hb* is unusual in that it is both a zygotic gap gene and a maternal coordinate gene. Pair-rule genes are typically expressed in patterns of 7 stripes, each about 3 nuclei wide. The pair-rule genes acting together activate the segment polarity genes at the onset of gastrulation in expression domains only one cell wide that stably specify segments. [7,8,9]

1.2 Gene Circuits

Our approach to modeling gene circuits incorporates, as a basic design principle, the fundamental features of the gene regulation problem for multicellular organisms as discussed in the Introduction. The model describes a collection of cells or cell nuclei, each containing a common regulatory circuit. We use concentrations of protein products of genes as state variables. This is a very important choice, because protein concentrations are observable using currently available techniques. It is this fact that ties the modeling component of our effort to the image processing work reported here.

The biochemical mechanisms governing the regulatory behavior of eucaryotic genes are far from well understood. As a result, there is no reliable *in vitro* assay by which one can determine the regulatory interactions of the model from a biochemical starting point. Moreover, even if the necessary understanding did exist, it would not provide a practical approach to modeling gene regulation, owing to the complexity of the biochemical description. These facts have an important consequence: At present, effective modeling of networks of eucaryotic genes must be based on a phenomenological approach that takes gene expression data as input and produces a regulatory circuit as output. Because our modeling approach is data driven in this way, the results reported here are central to our overall effort.

We represent a circuit by the elements of a matrix $\mathbf{T}$. Each element T^{ab} of this matrix characterizes the regulatory effect of one gene on another by a single real number for each possible pair a and b. Thus if T^{ab} is positive gene b activates gene a; if T^{ab} is negative gene b represses gene a, and if T^{ab} is zero gene b has no effect on gene a. In a very basic sense, this is a minimal model of gene regulation: We do not know, *a priori*, which gene regulates which other gene. Consequently, we must allow for the possibility that each gene regulates every other gene. For N genes, this leads to an $N \times N$ matrix. This matrix is the fundamental theoretical object in our model. Note that each cell nucleus contains a copy of the regulatory circuit defined by the $\mathbf{T}$-matrix, and that the same regulatory circuit occurs in each cell nucleus. This is a reflection of the fundamental biological fact that the cell nuclei in a multicellular organism contain identical genetic material.

The change with respect to time of concentrations of proteins is governed by three basic processes: Direct regulation of protein synthesis from a given gene by the protein products of other genes (including auto-regulation as a special case); transport of molecules between cell nuclei; and decay of protein concentrations.

We combine these considerations into a coarse-grained chemical kinetic equation as follows. Let the position of a cell nucleus along the A-P axis be indexed by i, such that nucleus $i+1$ is immediately posterior to nucleus i. Each cell nucleus contains a copy of a regulatory circuit composed of N genes, determined by the matrix $\mathbf{T}$. The concentration of the ath gene product in nucleus i is a function of time, denoted by $v_i^a(t)$. Then

$$\frac{dv_i^a}{dt} = R_a g_a\Big(\sum_{b=1}^{N} T^{ab} v_i^b + m^a v_i^{bcd} + h^a\Big) + D^a(n)\Big[(v_{i-1}^a - v_i^a) + (v_{i+1}^a - v_i^a)\Big] - \lambda_a v_i^a, \tag{1}$$

where N is the number of zygotic genes included in the circuit. The first term

on the right hand side of the equation describes gene regulation and protein synthesis, the second describes exchange of gene products between neighboring cell nuclei, and the third represents the decay of gene products.

In (1), T^{ab} is the previously discussed matrix of genetic regulatory coefficients. The *bcd* input is given by $m^a v_i^{bcd}$, where v_i^{bcd} is the concentration of *bcd* protein in nucleus i and m^a is the regulatory coefficient of *bcd* acting on zygotic gene a. g_a is a "regulation-expression function", which we assume takes the form $g_a(u^a) = (1/2)\left[\left(u/\sqrt{u^2+1}\right)+1\right]$ for all a, where $u^a = \sum_{b=1}^{N} T^{ab} v_i^b + m^a v_i^{bcd} + h^a$. R_a is the maximum rate of synthesis from gene a, and h^a summarizes the effect of general transcription factors on gene a. The diffusion parameter $D^a(n)$ depends on the number n of cell divisions that have taken place, and varies inversely with the square of the distance between nuclei. λ_a is the decay rate of the product of gene a. Nuclear divisions are incorporated by shutting down synthesis for a time equivalent to one mitosis, and doubling the number of nuclei.

At the outset, we don't know what the values of T^{ab} and the other parameters in (1) are. We do know what the experimentally observed *solutions* of (1) are: they are simply the observed gene expression patterns. For fixed initial conditions, solutions of (1) depend on what parameters are chosen: We seek the set of parameters that minimize the summed squared deviations between the observed data and the solutions of (1). This is a least squares optimization problem, which we solve by methods described elsewhere.[6]

In previous work expression levels were visually estimated from photomicrographs. Each embryo can be stained for at most 3 gene products, and so a complete map of expression patterns could be constructed only by a laborious manual registration of expression domains. Visual estimation of expression levels and manual image registration are both severe problems in terms of the time needed to construct a dataset and the accuracy of the result. In the following Sections, we solve the first of these problems.

2 Materials and Methods

Antibodies to Kruppel and Hunchback proteins were raised as follows. Expression plasmids (pAR 3040 vector) for Hunchback and Kruppel were provided by Steve Small. Full-length proteins were produced in bacteria and purified by SDS-PAGE followed by electroelution. We repeatedly immunized a rat (Hb) and a guinea pig (Kr) with 200 microgram doses of the purified protein over a period of several months, and obtained serum containing polyclonal antibodies against these gene products.

Drosophila embryos were collected on apple juice plates and fixed according to the standard protocol,[10] with the exception that 4% paraformaldehyde in PBS was substituted for formaldehyde in buffer B. Embryos were incubated with diluted anti-Hb (1:500), anti-Kr (1:300), and rabbit anti-FtzF1 (1:400; provided by Leslie Pick) in PBS with 0.1% Tween, washed, blocked in 5% non-fat dry milk, then incubated in a cocktail of fluorescent conjugated secondary antibodies diluted 1:200 (Jackson Labs). We used FITC anti-guinea pig, Texas Red anti-rabbit, and Cy5 anti-rat. After washing, embryos were mounted on slides and assayed using the 16X oil immersion Plan objective of a Leica TCS4D confocal microscope.

Three 8 bit channels were used to detect the proteins separately. We excited the dyes with a single wavelength at a time to ensure no leakage between channels, using the BP-FITC filter with the 488 nm excitation line (FITC), the BP-60030 filter with the 568 nm excitation line (Texas Red), and the RG665 filter with the 647 nm excitation line (Cy5). For each stain, three 1024x1024 pixel images of the blastoderm at two-micron depth intervals were obtained. These three images were averaged prior to further processing.

The processing and visualization of data was performed using the Khoros system,[11] available at `http://www.khoral.com`.

3 Results

We will illustrate our results and methods for the case of a particular embryo, shown in Figure 1. This embryo was fluorescently stained for the protein products of the gap genes *hb* and *Kr* and the maternally expressed transcription factor FtzF1. The embryo was scanned in a tangential plane which intersects the lateral portion of the blastoderm, revealing its surface. Because the three proteins visualized in Figure 1 are localized to cell nuclei, the fluorescent signal appears in many small spots each corresponding to an individual nucleus. Thus we wish to monitor fluorescent signals over a two-dimensional surface of the embryo.

We seek to determine the average level of gene expression in each nucleus, which is the site where the proteins we study exert their biological function. This is a two-step process. The first step is to define which pixels lie on each nucleus by creating a labeled mask. Using this mask, we average the fluorescence intensity level over each nucleus. In this section we explain in detail how these steps can be carried out, and we show an illustrative example of the results that can be obtained with this technique. For this example, we take the expression domains for *Kr* and *hb* shown in Figure 1 and quantitate them.

We have accomplished this by creating a binary image, or mask, based on a fluorescent counterstain of a maternally deposited transcriptional regulator, FtzF1.[12,13] FtzF1 protein is uniformly distributed throughout the blastoderm in all nuclei.

3.1 Computer Identification of Nuclei

The first step is to rotate an image so that the A-P axis of the embryo lies along the horizontal direction. A rectangular field of nuclei is chosen for analysis. The rectangular area chosen for this problem is shown by an inset of the completed mask in Figure 1D. A portion of this region is shown at high magnification in Figure 2. The 12 panels in this figure illustrate the major steps in the processing. This provides a roadmap for the discussion below.

Preliminary Processing and Edge Detection

Figure 2A shows a highly magnified view of the FtzF1 channel of the embryo in Figure 1. In Figure 2B, the contrast of the nuclei relative to the interstices is enhanced by local histogram equalization of image pixels.[14] This operation flattens the brightness histogram over a region about the size of a nucleus as much as possible while still maintaining the relative brightness ranking of pixels. The image is expanded by a factor of two (not shown), and then filtered in order to smooth the edges of nuclei. The filtering operation replaces each pixel value with the median value obtained from itself and its eight neighbors, as shown in Figure 2C. Figure 2D shows the result of applying the Shen-Castan edge extraction algorithm[15] to the image in Figure 2C in order to identify the boundaries of each nucleus. Because we are actually interested in the nuclei rather than their boundaries, all "off" (zero-valued, black) pixels enclosed by the extracted edges are turned on (Figure 2E).

Mask Correction

Inspection of Figure 2E shows that many nuclei are represented by isolated blobs of non-zero pixels, as desired. Others, however, are fused into dumbbell shaped objects, whereas they should form disjoint objects. We correct the mask using the following procedure. The overall strategy is to use the imperfect mask (Figure 2E) to surround each blob with a buffer zone that provides

Figure 2: (Opposite Page) High magnification view of the steps involved in creating a mask. See text for details.

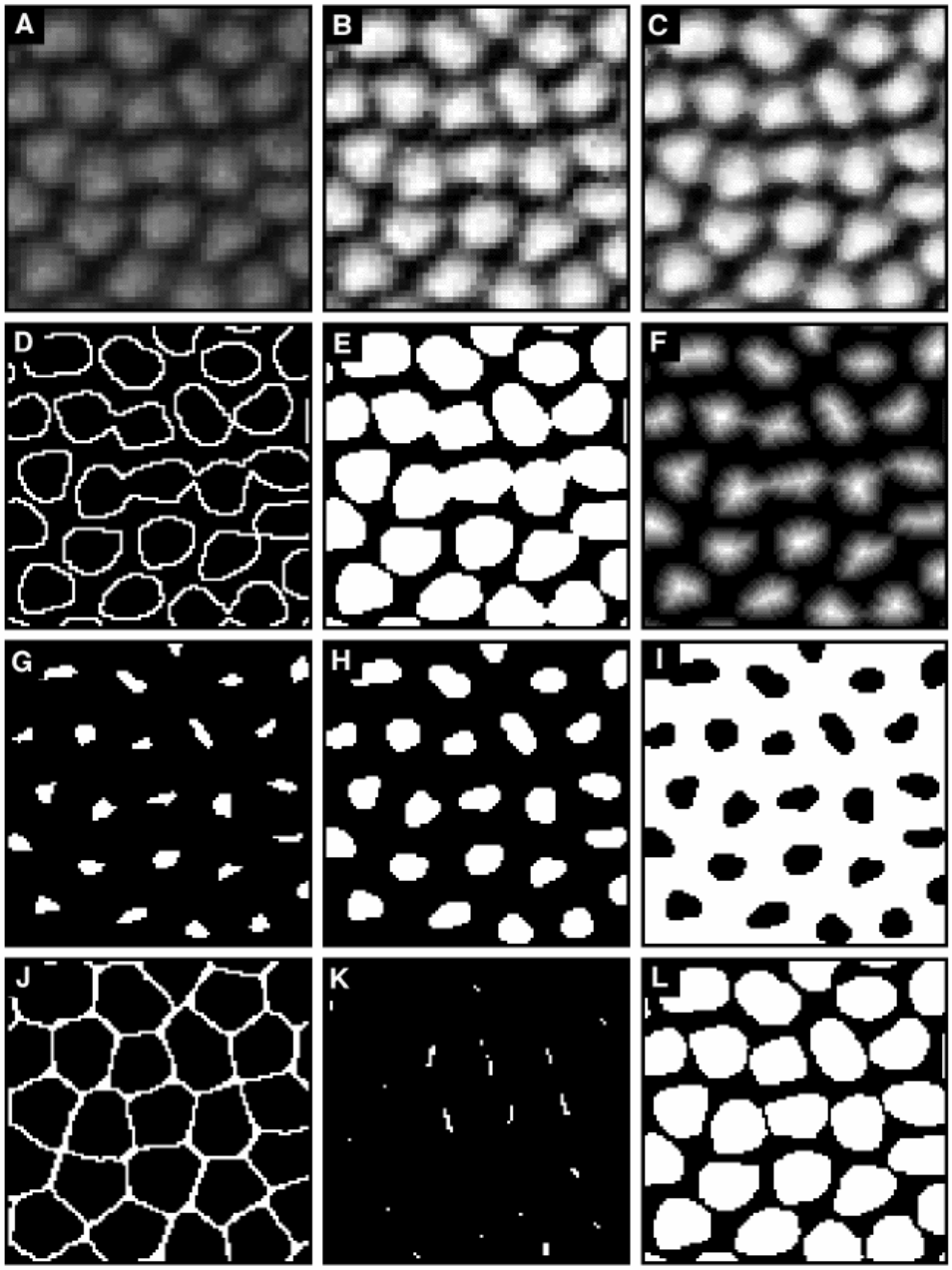

Figure 2

a boundary between two joined features. To form the buffer zone, we first create a Euclidean distance map of the imperfect mask,[14] shown in Figure 2F. The values of pixels in Figure 2F have been reassigned according to their distance from the closest edge of a feature. Thus, pixels at the center of features receive a higher value than pixels near their edges. Joins between features are characteristically close to edges, and so we set all pixels from Figure 2F whose intensity is above a certain threshold to white (Figure 2G), while preserving the core of all features. Figure 2H is obtained from 2G by one cycle of dilation,[14] a process in which "off" pixels with at least one "on" pixel in a neighborhood defined by a "structuring element" are activated. In this case the element used is a disk of 5 pixels diameter. The purpose of this step is to prevent artifactual right angles present in 2G from being amplified into erroneous features in the reticular structure in Figure 2J. A binary reversal of this image, its complement, is shown in Figure 2I. This set of "on" pixels provides the base from which the buffer zone is constructed. The white area in Figure 2I is topologically equivalent to the boundaries between nuclei, but it is much too thick. This area is converted into a boundary by a process known as "erosion", in which "on" pixels with "off" neighbors are themselves turned off.[14] Figure 2I is transformed into Figure 2J by multiple cycles of erosion with the constraint that no chains of "on" pixels may break. This compound operation is known as "skeletonization",[14] and the skeletonized image shown in Figure 2J is the desired buffer zone between distinct nuclei. Figure 2K shows the results of a Boolean AND operation using Figs. 2E and 2J as operands: these are precisely the points to be removed from Figure 2E. This is done by subtracting Figure 2K from Figure 2E to yield the final mask shown in Figure 2L and Figure 1D.

3.2 Obtaining Fluorescence Averages and Nuclear Positions

Now we have an image that tells where nuclei are. We need to be able to refer to individual nuclei, so we give each nucleus a unique numerical identifier. Four averages are now taken over each nucleus. We average the x and y coordinates of each pixel lying within that nucleus so as to obtain its centroid. We also average the fluorescence level of each segmentation gene product being assayed. For each nucleus, this results in a data structure containing five components: The nuclear identifier; The x and y position of the nucleus, and the average intensities of each gene product being assayed.

The results of the entire procedure are shown in Figure 3. This figure shows a plot of average *Kr* and *hb* expression in each nucleus as a function of position along the x axis which we identify, as an approximation, with the

A-P axis. Altogether, intensities from 811 nuclei are shown. Information about the y dorso-ventral (D-V) positions of the nuclei is not shown in this figure. We stress, however, that full information about the position of nuclei on the 2D surface of the embryo is provided by this image processing method. This information is essential for constructing complete maps of all expression domains via image registration. Precisely the same procedure has been used, without change, on many embryos, so that it is clear that the method is suitable for large scale studies.

4 Discussion

The importance of improved methods for the acquisition of gene expression data is widely recognized. For example, a powerful approach to automated data processing is based on monitoring of mRNA levels with DNA "chips".[16] These will allow investigators to apply the large scale automated methods now used for sequencing studies to studies of gene expression. These "chips", like non-automated methods such as blotting, CAT assays, quantitative PCR *etc.*, are based on the preparation of homogenates of cells as an initial step. Thus they do not capture spatial information about gene expression.

Homogenate based methods are most useful for studies of well differentiated tissue types. They have serious drawbacks, however, for investigations of early events in determination and pattern formation. These processes take place in relatively small morphogenetic fields where the differences between future cell types are first traceable to relatively small spatial differences in the expression of a small number of genes. Tracing such processes requires a knowledge of the spatial distribution of gene expression *in situ* at cellular resolution, which has been accomplished by the method presented here.

A major focus of our ongoing work is to characterize the accuracy and reproducibility of the quantitative output of this method. We are also using other image processing techniques, such as watershed segmentation,[14] to improve the mask creation procedure. Methods for automated image registration and for relating observed intensities to standard concentration curves are under development.

In summary, conclusions about regulatory mechanisms which govern fundamental events in *Drosophila* development are based on inferences from observed gene expression patterns either directly, or to establish the relevance of *in vitro* studies. The methods we have developed in this paper will greatly facilitate this procedure.

Acknowledgements

We thank Steve Small for gifts of expression plasmids, and Leslie Pick for the gift of FtzF1 antibody. This work was supported by grant RR 07801 from the National Institutes of Health.

References

1. Eric Mjolsness, David H. Sharp, and John Reinitz. *Journal of Theoretical Biology*, 152:429–453, 1991.
2. Denis Theiffry, M. Colet, and Rene Thomas. *Mathematical Modelling and Scientific Computing*, 2:144–151, 1993.
3. Harley H. McAdams and Lucy Shapiro. *Science*, 269:650–656, 1995.
4. Roland Somogyi and Carol Ann Sniegoski. *Complexity*, 1:45–63, 1996.
5. John Reinitz, Eric Mjolsness, and David H. Sharp. *Journal of Experimental Zoology*, 271:47–56, 1995.
6. John Reinitz and David H. Sharp. *Mechanisms of Development*, 49:133–158, 1995.
7. Michael Akam. *Development*, 101:1–22, 1987.
8. P. W. Ingham, N. E. Baker, and A. Martinez-Arias. *Nature*, 331:73–75, 1988.
9. Peter A. Lawrence. *The Making of a Fly*. Blackwell Scientific Publications, Oxford, UK, 1992.
10. Manfred Frasch, Tim Hoey, Christine Rushlow, Helen J. Doyle, and Michael Levine. *The EMBO Journal*, 6:749–759, 1987.
11. John Rasure and Mark Young. In *1992 SPIE/IS&T Symposium on Electronic Imaging*, volume 1659 of *SPIE Proceedings*. SPIE, 1992.
12. Giovanni Lavorgna, Hitoshi Ueda, Joachim Clos, and Carl Wu. *Science*, 252:848–851, 1991.
13. Yan Yu, Willis Li, Kai Su, Miyuki Yussa, Wei Han, Norbert Perrimon, and Leslie Pick. *Nature*, 385:552–555, 1997.
14. John C. Russ. *The Image Processing Handbook Second Edition*. CRC Press, 1995.
15. J. Shen and S. Castan. In *Proceedings CVPR '86*, Miami, 1986.
16. David J. Lockhart, Helin Dong, Michael C. Byrne, Maximillian T. Follettie, Michael V. Gallo, Mark S. Chee, Michael Mittmann, Chunwei Wang, Michiko Kobayashi, Heidi Horton, and Eugene L. Brown. *Nature Biotechnology*, 14:1675–1680, 1996.