Spatio-temporal Registration of the Expression Patterns of Drosophila Segmentation Genes

Ekaterina M. Myasnikova† David Kosman‡

John Reinitz‡ and **Maria G. Samsonova†** *

† Institute of High Performance Computing and Data Bases
P.O. Box 71, St.Petersburg, 194291 Russia,
tel. 7-812-251-9092, fax 7-812-251-8314; samson@fn.csa.ru, miasnikov@pop3.rcom.ru
‡ Department of Molecular Biology and Biochemistry
Box 1020 Mt.Sinai School of Medicine, One Gustave L. Levy Place,
New York, NY 10029-6574, USA,
tel. 1-212-241-1952, fax 1-212-860-9279;
reinitz@kruppel.molbio.mssm.edu; dave@eve.molbio.mssm.edu

Abstract

The application of image registration techniques resulted in the construction of an integrated atlas of *Drosophila* segmentation gene expression in both space and time. The registration method was based on a quadratic spline approximation with flexible knots. A classifier for automatic attribution of an embryo to one of the temporal classes according to its gene expression pattern was developed.)

Keywords: gene expression data, image registration, genetic network, Drosophila

Introduction

The analysis of the integrated structure and behavior of the complex genetic regulatory systems underlying development is a central task for fundamental research in biology over the next few decades. Its accomplishment will require the acquisition of unprecedented amounts of gene expression data. Substantial progress in this direction has been already made. For example, a powerful approach to automated acquisition of gene expression data is based on monitoring the levels of expression of up to several thousand species of mRNA with DNA "chips"(Lockhart *et al.* 1996).

However this method, like other non-automated methods such as blotting, CAT assays, quantitative PCR *etc.* are based on the preparation of homogenates of cells as an initial step. Such methods are appropriate for studies of well differentiated tissue types, but have serious drawbacks for investigations of the early developmental stages in which the events in determination and pattern formation take place. At early embryogenesis the determination of cell fate and pattern formation proceeds in relatively small morphogenetic fields, in which the differences between future cell types are first traceable to relatively small spatial differences in the expression of a small number of genes. Thus the investigation of these processes requires a knowledge of the spatial and temporal distribution of gene expression *in situ*. Homogenate based methods are unable to capture spatial information about gene expression.

The investigation of early development requires methods for the acquisition of gene expression data which will allow us to simultaneously monitor the expression time course of all of the genes in a functional network at the resolution of a single cell. We are addressing this issue in the context of a particular biological system, namely the segment determination system of the fruit fly *Drosophila melanogaster* (Reinitz, Mjolsness, & Sharp 1995; Reinitz & Sharp 1995; Reinitz *et al.* 1998; Sharp & Reinitz 1998). Our ultimate goal is the deciphering of the molecular mechanisms which control this process, and a central part of this work is the construction of a quantitative atlas of segmentation gene expression at cellular resolution.

Like all other arthropods, the body of the fruit fly *Drosophila melanogaster* is made up of repeated units called segments. Before the 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. There are approximately forty genes specifically involved in segmentation, but the initial determination of the segments is a consequence of the expression of

* To whom correspondence should be addressed.

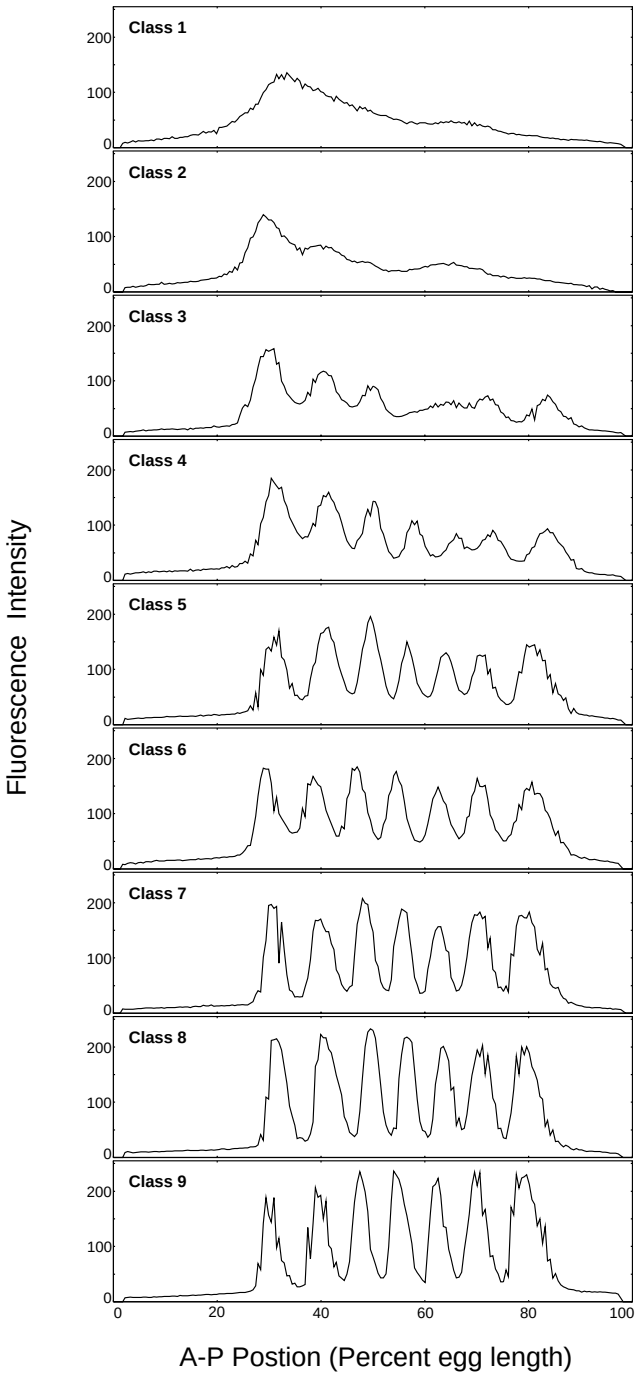

Figure 1: Representative expression data of the pair-rule gene *eve* for each of the temporal classes. Data for each graph was obtained from a portion of a single embryo processed text file.

only about fifteen genes. A few of these are "maternal coordinate genes" expressed from the mother to provide asymmetric initial conditions. Most are zygotic genes that are expressed in patterns that become more spatially refined over time. This refinement is a consequence of members of the network regulating one another's synthesis in a precise manner. Of particular importance are members of the "gap" and "pair-rule" classes of segmentation genes. Gap genes are expressed with unimodal or bimodal concentration profiles which become gradually steeper, while pair-rule genes initially express protein in a a single very broad domain that restricts to seven narrow domains over a relatively short time interval (Akam 1987; Ingham 1988). In general, each gap and pair-rule gene expresses protein in a different set of locations, but these have a characteristic overlap with one another. Thus one can view the expression pattern as a collection of "domains", each of which is a region of expression containing one concentration maximum.

Expression of segmentation genes is largely a function of position on the anterior-posterior (A-P), and so can be well represented in one dimension. Gene expression is monitored by confocal scanning of fixed embryos stained with fluorescence tagged antibodies. Each embryo is observed at the moment of development when it is fixed; each embryo can be scanned for the expression of up to three genes at once. The application of image segmentation techniques allows us to determine the average fluorescence level on each nucleus (Kosman, Reinitz, & Sharp 1997). Images are processed so that each nucleus in the processed dataset is labeled numerically and specified by the following features: x and y coordinates of its centroid together with quantitative values for the average intensities of gene expression for up to three genes over each nucleus.

Here we describe how data of this type can be combined into an integrated atlas which will contain the information about simultaneous expression of all of the network genes over time to the resolution of a single cell. The creation of such an atlas is accomplished through the application of image registration techniques in both space and time. These methods allow us to construct a map of all relevant expression domains from a series of embryos of the same age and to develop a classifier for automatic attribution of an embryo to one of the temporal classes according to its gene expression pattern.

Materials and Methods

Materials

We raised a panel of antibodies to *Drosophila* segmentation genes and used it to fluorescently stain embryos as described (Kosman & Reinitz 1998). Confocal images of 429 embryos were obtained. Images were automatically rotated and cropped by making a thresholded mask of the entire embryo, finding the principal moments, rotating the embryo to bring the principal moments in line with the x and y axes, and cropping the image to the edge of the mask. Correct dorsoventral (D-V) and anterior-posterior (A-P) orientation was then obtained by visual inspection and manual flipping if needed. Each image was segmented and reduced

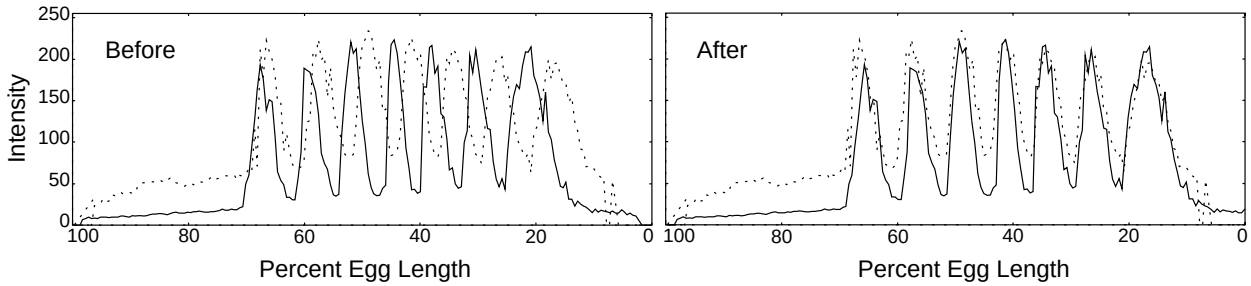

Figure 2: Example of image registration. We show the graphs of common Eve domains in two embryos, one solid and one dotted before (left) and after (right) registration.

to an text file containing a list of identified nuclei together with their x and y coordinates and the average fluorescence levels of three gene products as described (Kosman, Reinitz, & Sharp 1997), with the following modifications. First, masks were constructed from an image containing the maximum pixel from the three channels used, without use of a nuclear counterstain. Second, the watershed method (instead of skeletonization) was used for the correction step following edge detection (Bowler, Reinitz, and Kosman, manuscript in preparation). A single image gives about 2000 processed nuclei.

The 429 embryo images used in this study yielded text files of quantitative data on the expression of 12 segmentation genes. These comprised the maternal coordinate genes *bicoid* (*bcd*) and *caudal* (*cad*), the gap genes *Kruppel* (*Kr*), *knirps* (*kni*), *giant* (*gt*), and *hunchback* (*hb*), and the pair-rule genes *even-skipped* (*eve*), *runt* (*run*), *hairy* (*h*), *odd-skipped* (*odd*), *fushi-tarazu* (*ftz*), and *paired* (*prd*) distributed as follows:

- 54 embryos stained for the *ftz*, *runt*, and *eve* genes;
- 5 embryos stained for the *eve*, *cad* and *bcd* genes;
- 26 embryos stained for the *eve*, *kni* and *gt* genes;
- 144 embryos stained for the *eve*, *Kr* and *gt* genes;
- 55 embryos stained for the *eve*, *odd* and *h* genes;
- 112 embryos stained for the *eve*, *kni* and *hb* genes;
- 28 embryos stained for the *eve*, *Kr* and *hb* genes;
- 5 embryos stained for the *eve*, *prd* and *odd* genes.

All the embryos belong to cleavage cycle 14 (Foe & Alberts 1983). This cycle is about an hour long and is characterized by a rapid transition of the pair-rule gene expression patterns, which culminates in the formation of 7 stripes.

We divided the 429 embryos into temporal classes by an extensive and thorough visual analysis of images and graphs of individual embryos that will be published elsewhere (Kosman et al., in preparation). Each image is allocated to one of 9 temporal classes on the basis of visual inspection of pair-rule gene expression patterns, particularly that of *eve*. Each embryo was stained for *eve*, the pattern of which is highly dynamic.

The evolution of the expression patterns of *eve* during cleavage cycle 14 is illustrated in Figure 1. Time classes 1, 2, 3, and 4 do not have seven well-defined eve stripes, and could be grouped according to the degree of formation of individual stripes. The remaining groups (classes 5, 6, 7, 8 and 9) do have seven well-defined eve stripes and were classified by features of the overall eve pattern along with the patterns of other pair-rule genes. After seven *eve* stripes are clearly visible their intensities increase in the posterior portion of the embryo. This process is accompanied by drastic intensity changes between adjacent nuclei. By the end of cycle 14, all *eve* stripes have reached maximum and equal intensity and maximum sharpness (i.e. maximum intensity difference between adjacent nuclei). Note that late *eve* stripes (class 9) are distinctly sharper on the anterior edge.

Registration in space

The simplest approach is a 1-D registration technique, which is performed for the raw data extracted from the central 10% of y-values on the midline of an embryo in the A-P direction (x-coordinate). The y-values of these data are then ignored, and the patterns, demonstrating the variation of gene expression along the x-axis, are presented as diagrams. The stripes and interstripe areas on the curve have the form of peaks and valleys respectively. To identify the most essential features of each pattern and to get rid of noise on the curves, some approximation method should be applied. One widely used class of approximation methods is that of splines, in which a function is approximated piecewise by polynomials of a given order which are constrained to follow certain continuity constraints at the borders ("knots") of the approximated segments. This method provides a smooth approximation of the curves, and classifies each pattern by a set of parameters. The simplest smooth approximation is provided by a quadratic spline with M flexible knots, where first derivatives are constrained to be continuous at each knot. The system of knots, $\widetilde{x}_1, \widetilde{x}_2, \ldots \widetilde{x}_M$, is introduced as a set of distinct points on the x-axis, which border the area of each peak. Each knot is located at the point of inflection between a peak and the adjacent valley. The number of knots is equal to the total number of minima and maxima on the given curve. The quadratic spline approximation

class	**4** [37]	**5** [55]	**6** [55]	**7** [19]	**8** [71]	**9** [26]
stripe 1	31.92(0.36)	31.56(0.44)	31.54(0.44)	31.63(0.39)	31.99(1.52)	31.74 (0.41)
1/2	37.28(0.36)	36.63(0.30)	36.42(0.30)	36.34(0.22)	36.54(1.1)	36.19(0.20)
stripe 2	41.96(0.24)	41.56(0.30)	41.31(0.22)	41.10(0.22)	41.12(0.67)	40.72(0.20)
2/3	46.79(0.24)	46.06(0.30)	45.60(0.30)	45.30(0.22)	45.15(0.42)	44.90(1.43)
stripe 3	50.32(0.30)	49.83(0.22)	49.45(0.30)	49.13(0.22)	48.96(0.25)	48.53(0.41)
3/4	54.07(0.43)	53.72(0.30)	53.22(0.30)	52.89(0.22)	52.55(0.38)	52.06(0.36)
stripe 4	57.52(0.36)	57.14(0.37)	56.67(0.30)	56.45(0.26)	56.06(0.59)	55.51(0.36)
4/5	61.41(0.49)	61.08(0.44)	60.50(0.30)	60.10(0.30)	59.60(0.93)	58.91(0.31)
stripe 5	65.34(0.36)	64.86(0.37)	64.22(0.22)	63.76(0.22)	63.05(1.18)	62.30(0.25)
5/6	68.55(0.30)	68.39(0.37)	67.82(0.30)	67.32(0.26)	66.53(1.52)	66.07(0.82)
stripe 6	71.98(0.36)	72.09(0.30)	71.59(0.22)	70.98(0.22)	70.23(1.77)	69.58(0.46)
6/7	77.03(0.30)	76.81(0.30)	75.94(0.22)	75.19(0.22)	74.46(2.19)	73.94(0.31)
stripe 7	82.64(0.36)	82.27(0.52)	81.05(0.37)	80.08(0.48)	79.23(2.61)	78.60(0.46)

Table 1: Results of the spatial registration of *eve* expression patterns for temporal classes 4-9. The size of each class is given in brackets. Rows contain the x-coordinates of maxima (stripes) and minima (interstripes; $N/(N+1)$ denotes the interstripe between stripe N and $N+1$) of Eve concentration with the standard deviations shown in parentheses.

with M knots can be represented as

$$\mathrm{sp}_2(x) = \sum_{k=0}^{2} \zeta_{k0} x^k + \sum_{n=1}^{M} \zeta_{n2} (x - \widetilde{x}_n)_+^2, \qquad (1)$$

where $(x)_+ = \max(x, 0)$.

The parameters of $\mathrm{sp}_2(x)$ given in (1) are estimated by the least squares method. This is done as a two step procedure, in which the knot locations are found by the Nelder-Mead "downhill" simplex optimization procedure (Press *et al.* 1988) in conjunction with the linear method for least squares estimation of other parameters of the spline. Given the sample of $\{x_i\}_{i=1}^N$, the x-coordinates of N nuclei under consideration and $\{\lambda_i\}_{i=1}^N$, the values of protein intensity in each nucleus, the $(3+2M)$ parameters of this approximation are to be estimated by minimizing the cost function

$$S_1 = \sum_{i=1}^{N} \{\mathrm{sp}_2(x_i) - \lambda_i\}^2 \to \min .$$

The partial derivatives of S_1 with respect to the spline parameters are set to zero and the parameters $\{\zeta_{0k}\}_{k=0}^2$ and $\{\zeta_{n2}\}_{n=1}^M$ are estimated by solving the system of $(3+M)$ linear equations for the vector of knots $\{\widetilde{x}_j\}_{j=1}^M$ obtained at the current step of the nonlinear optimization. The initial location of knots is chosen automatically. As a result of such approximation each curve is specified by the set of knots $\{\widetilde{x}_j\}_{j=1}^M$ and by the spline parameters $\{\zeta_{k0}\}_{k=0}^2$ and $\{\zeta_{n2}\}_{n=1}^M$. From these parameters the x-coordinate of the kth peak's extremum is given by

$$X_k = \frac{2\sum_{n=1}^{k} \zeta_{n2}\widetilde{x}_n - \zeta_{10}}{2(\zeta_{20} + \sum_{n=1}^{k} \zeta_{n2})}, \qquad k = 1, ..., M.$$

The registration of the images is performed by resizing the patterns along the x-axis by the affine transformation $x' = x\rho + \Delta$, so that the total distance between the x-coordinates of all the extrema $\{X_{kj}\}_{k=1}^M$ is minimized. An example of such registration is presented in Figure 2.

Temporal Classification of Embryos

The spline approximation is a good tool for spatial registration, since it gives a good estimation of the extrema location. Nevertheless this method cannot provide a good fit to the real concentration profiles, in which not only the location but also the shape of peaks is essential. The sharp peaks, typical for later stages of embryo development, are smoothed strongly when the spline-based method is applied. A better fit can be achieved by excluding from the spline approximation (1) the condition of coinciding of the first derivatives at knots, keeping only the continuity condition. Such modification of the formula (1) will be given by

$$\mathrm{sp}_1(x) = \sum_{k=0}^{2} \zeta_{k0} x^k + \sum_{n=1}^{M} \zeta_{n1} (x - \widetilde{x}_n)_+ + \zeta_{n2} (x - \widetilde{x}_n)_+^2.$$

As a result of the modified approximation the coordinates of the extrema $(X_k, Y_k)_{k=1}^M$ can be estimated from the least square estimation of the parameters of $\mathrm{sp}_1(x)$. The peak located between two knots $\widetilde{x}_k$ and $\widetilde{x}_{k+1}$ is approximated by the parabola $y = Y_k + c_k(x - X_k)^2$, where $c_k = \zeta_{20} + \sum_{n=1}^{k} \zeta_{n2}$. Each curve is specified now by the set of triples $\{X_k, Y_k, c_k\}_{k=1}^M$, which are dealt with instead of raw data.

This approach to the pattern description along with the spatial registration provides an opportunity to create a "standard pattern" for each protein, which is defined by the average location of knots and by the average estimations of spline parameters over an age group. The standard patterns possess all the typical features of a given protein at different developmental stages. Even

class 4 [37]	**class 5** [55]	**class 6** [55]	**class 7** [19]	**class 8** [71]	**class 9** [26]
94.6%	61.7%	72.7%	68.0%	66.1%	73.1%

Table 2: Recognition results for temporal class compared to visual analysis. For each class the percent of correct attributions of *eve* expression patterns is presented. The class sizes are given in brackets.

more interesting from a biological perspective, the variance of these patterns can be estimated. This enables to assess the variability of the protein concentrations.

The essential step for pattern classification in time is to characterize the patterns in terms of some observable parameters, and then use these parameters to read off the time. Human observers classify the developmental stage of an embryo by careful study of its pattern, since each stripe possesses its own features at any stage of an embryo development. These features can be expressed in terms of the parameter triples estimated above and used in determining an embryo age and hence, the temporal classification of pattern images.

For each class a standard pattern is obtained and average characteristic features, i.e. centroids, are calculated. To attribute an embryo to a certain temporal class we compute a distance between the given pattern and centroids of all the classes. The embryo is assigned to the least distant class.

We introduce characteristic features of a standard pattern as the set of parameters $\{d_j, c_j\}_{j=1}^{M}$, where $d_j = Y_j - Y_{j-1}$; $d_1 = Y_1$. These parameters completely describe shape and height of the peaks, while after the spatial registration x-coordinates of the peaks $\{X_j\}_{j=1}^{M}$ can be ignored. The distance between two patterns at different time stages is defined as the total Euclidean distance between the pairs $\sqrt{\sum_j \{(d_{kj} - d_{lj})^2 + w(c_{kj} - c_{lj})^2\}}$, where k and l are the labels of the patterns compared, w is a weight coefficient.

We note that a distance metric of this type is appropriate for intermediate and late developmental stages when a full set of domains has formed. For the early developmental stages the number, location and shape of peaks change rapidly and therefore more qualitative and specific characteristics of the standard patterns are required.

Both methods for spatial registration and temporal classification are implemented in C.

Results

Spatial registration

Our goal is to construct a map of all relevant expression domains at a given time from a series of embryos of the same temporal class. In our experiments each embryo has been stained for three gene products, one of them always being the product of the pair-rule gene *eve*. Embryos of approximately the same age stained for different combinations of proteins are subjected to spatial registration against the common *eve* expression domain. As a result for each temporal class the standard expression patterns of all the segmentation proteins are constructed (see fig. 3). Table 1 presents the accuracy of the spatial registration of *Eve* protein for all the temporal classes with the exception of the very early ones. The standard deviations of the stripe locations are quite small, in most cases less than 0.5% egg length. A single nucleus is about 1% egg length in diameter, so this represents a high level of accuracy. The highest standard deviations are associated with the posterior area of the pattern for the largest group, class 8.

The standard expression patterns of 8 segmentation genes attributed by visual inspection to temporal class 8 are presented in figure 3. The temporal class 8 contains the largest number of embryos and has seven well-defined *eve* stripes. The standard patterns were constructed only for those genes which have been stained in more than 9 embryos.

Temporal classification of images

In order to determine the segmentation gene circuits, quantitative data on segmentation gene expression domains must be arranged in time. While the attribution of gene expression patterns to predefined temporal classes can be done by visual inspection, it is a very time consuming process. Automated temporal classification will constitute a fundamental advance in pattern analysis.

We applied the classification method described above to attribute each embryo from the dataset under study to one of the nine temporal classes defined by visual analysis. The visual and numerical analyses were conducted independently. The parameters of the standard *eve* expression pattern were considered as the features of a temporal class. To attribute an embryo to a given age class we tested it against the training set, which consisted of all embryos of this age with exception of the tested embryo. The test was performed for all temporal classes with the exception of the three earliest ones, for which seven well shaped stripes have not yet formed. The recognition results are shown in the Table 2. The class 4 embryos are best recognized (the attribution is 94.6%). Other embryos are allocated to their classes with the accuracy of 61.7 - 73.1 %. Thus the classification method is not very reliable (maximal attribution error is 38%).The reason for the high accuracy of recognition of embryos belonging to temporal class 4 may be the strong difference of the *eve* expression pattern at this age in comparison with other temporal classes It is evident from figure 1 that class 4 embryos

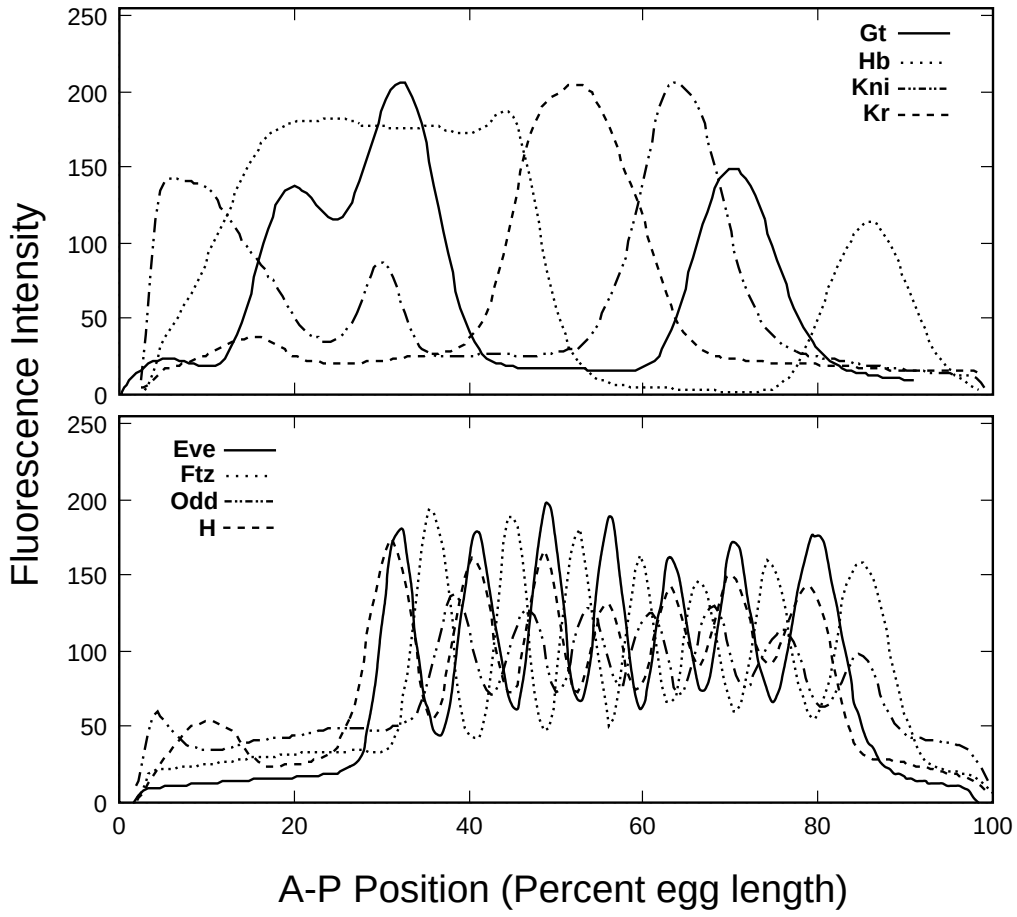

Figure 3: Results of spatial registration of the expression patterns of 8 segmentation genes. Original data was from embryos belonging to time class 8. The top graph shows the patterns of four of the gap genes; the bottom graph, the patterns of four of the pair-rule genes.

have significantly less well developed stripes than the other later classes.

In addition we note that in constructing the classifier we used the features of the *eve* expression pattern only, while the visual analysis classified embryos with consideration of all the stained proteins simultaneously. We believe that characterization of the temporal classes by the features of other proteins along with *eve* may substantially improve the classification.

Conclusions

In this work the spline based registration techniques were applied to construct of a map of all relevant expression domains from a series of embryos of the same age, as well as to development of a classifier for automatic attribution of an embryo to one of the temporal classes according to its gene expression pattern.

The registration method is characterized by unprecedented high accuracy. We have mapped the location of expression domains to within less than 0.5% egg length, and often to as little as 0.2% egg length, using tens of embryos. Similar analyses by hand (Frasch & Levine 1987) used 14 or fewer embryos and obtained standard deviations of 1.6% for *eve* stripe locations. The well defined set of nine temporal classes of expression is of some biological interest, since the degree of reproducibility of these patterns is a significant biological question.

At present the method of temporal classification of embryos on the basis of their expression patterns is not very reliable (maximal attribution error is 38%), so these results must be regarded as preliminary. One important question that must be answered is the accuracy of both the visual and numerical temporal analyses against physical clock time. Certain cytological events (Merrill, Sweeton, & Wieschaus 1988) can be used as physical developmental clock in fixed tissue, and an analysis of this data is underway.

We are confident that the analysis of this data, to be reported elsewhere, will have interesting implications for developmental biology.

Acknowledgements This work was supported by grant RO1-RR-07801 from the US NIH, grant 00014-97-1-0422 from the US ONR, and by the Ministry of Science and Technologies of the Russian Federation.

References

Akam, M. 1987. The molecular basis for metameric pattern in the *Drosophila* embryo. *Development* 101:1–22.

Foe, V. A., and Alberts, B. M. 1983. Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in *Drosophila* embryogenesis. *Journal of Cell Science* 61:31–70.

Frasch, M., and Levine, M. 1987. Complementary patterns of *even-skipped* and *fushi tarazu* expression involve their differential regulation by a common set of segmentation genes in *Drosophila*. *Genes and Development* 1:981–995.

Ingham, P. W. 1988. The molecular genetics of embryonic pattern formation in *Drosophila*. *Nature* 335:25–34.

Kosman, D., and Reinitz, J. 1998. Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. *Development, Genes, and Evolution* 208:290–294.

Kosman, D.; Reinitz, J.; and Sharp, D. H. 1997. Automated assay of gene expression at cellular resolution. In Altman, R.; Dunker, K.; Hunter, L.; and Klein, T., eds., *Proceedings of the 1998 Pacific Symposium on Biocomputing*, 6–17. Singapore: World Scientific Press. http://www.smi.stanford.edu/projects/helix/psb98/-kosman.pdf.

Lockhart, D. J.; Dong, H.; Byrne, M. C.; Follettie, M. T.; Gallo, M. V.; Chee, M. S.; Mittmann, M.; Wang, C.; Kobayashi, M.; Horton, H.; and Brown, E. L. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* 14:1675–1680.

Merrill, P. T.; Sweeton, D.; and Wieschaus, E. 1988. Requirements for autosomal gene activity during precellular stages of *Drosophila melanogaster*. *Development* 104:495–509.

Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; and Vetterling, W. T. 1988. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Reinitz, J., and Sharp, D. H. 1995. Mechanism of formation of eve stripes. *Mechanisms of Development* 49:133–158.

Reinitz, J.; Kosman, D.; Vanario-Alonso, C. E.; and Sharp, D. 1998. Stripe forming architecture of the gap gene system. *Developmental Genetics* 23:11–27.

Reinitz, J.; Mjolsness, E.; and Sharp, D. H. 1995. Cooperative control of positional information in *Drosophila* by *bicoid* and maternal *hunchback*. *Journal of Experimental Zoology* 271:47–56.

Sharp, D. H., and Reinitz, J. 1998. Prediction of mutant expression patterns using gene circuits. *BioSystems* 47:79–90.