Joint Analysis of In-situ Hybridization and Gene Expression Data

Lennart Opitz[1], Alexander Schliep[2] and Stefan Posch[1]

[1] Institut für Informatik, Martin-Luther-Universiät Halle-Wittenberg, D-06120 Halle, Germany; {opitz, posch}@informatik.uni-halle.de

[2] Department Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63–73, D-14195 Berlin, Germany; schliep@molgen.mpg.de

Abstract. To understand transcriptional regulation during development a detailed analysis of gene expression is needed. In-situ hybridization experiments measure the spatial distribution of mRNA-molecules and thus complement DNA-microarray experiments. This is of very high biological relevance, as co-location is a necessary condition for possible molecular interactions.

We use publicly available in-situ data from embryonal development of Drosophila and derive a co-location index for pairs of genes. Our image processing pipeline for in-situ images provides a simpler alternative for the image processing part at comparable performance compared to published prior work. We formulate a mixture model which can use the pair-wise co-location indices as constraints in a mixture estimation on gene expression time-courses.

1 Introduction

The cellular processes constituting life as we know it are controlled by highly complex interaction mechanisms, where the most important form of control is transcriptional regulation. That is the control of the amount of proteins which are produced for a gene in the genome. As quantifying protein levels is experimentally difficult, the intermediate product, messenger RNA (mRNA) which is transcribed from a gene and gets translated to a protein, has received a lot of attention. DNA-microarrays are an experimental technique based on hybridization reactions to quantify levels of mRNA-levels for thousands of genes simultaneously. However, these experiments give a view of transcriptional regulation averaged over many cells or tissues. To understand development of organisms and the necessary differentiation of cells with the same genome, it is necessary to obtain a finer grained picture of gene expression.

In-situ hybridization experiments measure the abundance and spatial distribution of specific mRNA-molecules in organisms through staining cells proportionally to mRNA-concentration. Although the experimental technique is quite expensive as experiments have to be repeated for each gene reasonably large amounts of data exist. For example, the Berkeley Drosophila Genome Project (BDGP, http://www.fruitfly.org) provides a database of images for expression patterns during embryonal development. There are problems with data quality due to the experimental errors and the imaging process, however the data provides a unique opportunity to augment gene expression time-courses over embryonal development with co-location information. This is of very high biological relevance, as co-location is a necessary condition for possible interaction.

We introduce an image processing pipeline for processing in-situ hybridization data and a simple co-location index, which performs as well as published results (Peng et al. (2004)) even though it is substantially simpler (for more details see Opitz (2005)). We also formulate a statistical mixture model which allows the use of the co-location data in the form of pair-wise constraints in a mixture estimation on gene expression time-courses. This will provide a self-contained framework for joint analysis of in-situ hybridization and gene expression data.

2 Method

2.1 Image processing pipeline

The majority of hybridization images in the BDGP database contain the projection of exactly one centered embryo. However, there is a substantial portion of images with multiple touching or partially projected embryos. To exploit as much data as possible, the goal of image preprocessing is to locate and extract exactly one complete embryo from each image, even for touching embryos. Subsequently this embryo is registered to standardized orientation and size to allow for comparison of different expression patterns. Figure 1 shows the steps of the image processing pipeline for one example image.

To distinguish between embryo and non-embryo pixels we employ a similar approach as Peng et al. (2004) estimating the local variance of grey level intensities for each pixel in a 3 × 3 neighborhood. As noted in Peng et al. (2004) the background is much more homogeneous and as a consequence it suffices to apply a fixed predefined threshold for segmentation using variance

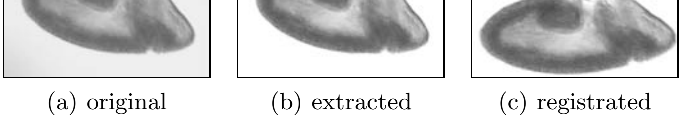
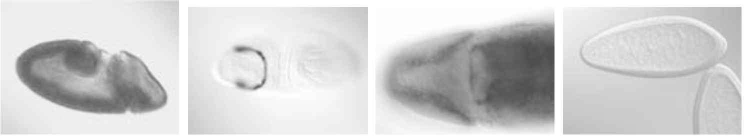

(a) original (b) extracted (c) registrated

Fig. 1. Example for steps of the image processing pipeline.



(a) type I (b) type II (c) type III (d) type IV

Fig. 2. The four categories of embryo images.

estimates. To eliminate small holes and erroneous embryo regions a sequence of morphological closing and opening using a circular mask of radius 4 is applied (see e.g. Gonzalez (1991)). Finally, the largest connected component is extracted and remaining holes are filled.

The resulting region may be the projection of a single complete or partial embryo or the projection of a set of multiple touching embryos. For further processing we define four types of regions: (I) one complete high quality embryo, (II) one complete blurred embryo, (III) one partially projected embryo, and (IV) a set of touching embryos (see Figure 2). Embryos of type II and III images are to be eliminated from further processing as they do not allow reliable co-location comparison. For type IV regions the aim of the processing pipeline is to separate the individual embryos and to extract one complete high quality embryo. This classification is realized by a series of simple filters. First, as a measure of ellipticity we compute the deviation of the region extracted from an elliptical. Second, the compactness is computed as the ratio of the squared circumference and the area of the region. Type I regions are defined by a linear separating line in the resulting two-dimensional feature space which is trained by a SVM using a set of 100 training images randomly selected from the BDGP database. To select type II images a threshold on the ellipticity is applied to the remaining images. Type III regions are identified using the area of the region and the number of pixels coincident with its bounding box. The threshold for both filters are determined from the same set of 100 images. To separate multiple touching embryos a new approach has been developed. First, a hypothesis of the coarse location of the centroid of the central embryo is derived using simple heuristic. Using a set of concentric rays emerging from this centroid the contour points of the complete embryo region are computed (see Figure 3). If the distance of two neighboring contour points exceeds the mean of distances by more the 20% it is considered a cut point between different embryos. The set of all detected cut points is used to finally separate the central embryo from the remaining ones. To eliminate invalid embryos separated by this method, an additional set of 50 type IV images is used to train a SVM using again ellipticity and compactness as features.

The final step of preprocessing is to register the embryos extracted to a standardized orientation and size. The embryo is rotated to horizontally align the principal axis. Subsequently the bounding box is scaled to a standard size (658 × 279 pixels for our experiments). After this registration there is

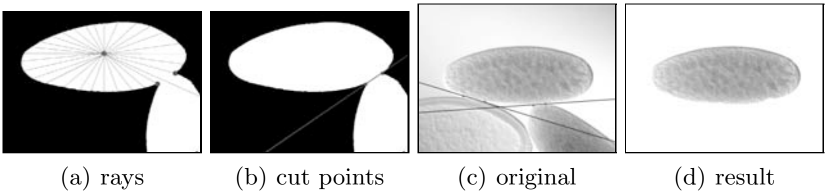


Fig. 3. Principle and example for separation of multiple embryos. (a) detection of contour points using a set of concentric rays; (b) detected cut points and separating line; (c) example image; (d) extracted embryo.

still an ambiguity in orientation, which may correspond to dorsal vs ventral and anterior vs posterior position. A part of the BDGP images is taken in a lateral position, giving again rise to four orientations (an example is given in Figure 4). We do not make an attempt for normalization at this step. Rather when comparing a pair of embryos for co-location we take all four orientations into account and use the best result as similarity score.

2.2 Co-location index

To compare in-situ hybridization patterns between genes and/or developmental stages, we developed a simple co-location index for a pair of registered embryos which is directly based on the intensity levels of staining. We prefer this approach to binarization of intensities (cf. Kumar et al. (2002)) or quantization into a set of discrete staining levels (cf. Peng (2004)). Binarization seems too coarse an approximation and disregards completely valuable information on the abundance of mRNA. On the other hand, there is little evidence for a fixed number of homogeneously distributed staining levels across a wide range of genes and developmental stages. As an alternative we propose to use the correlation coefficient of two registered embryo images where intensities of corresponding pixels for two registered embryos are considered as paired data. This score takes both the spatial distribution and the strength of hybridization into account. Using this correlation coefficient, we achieve invariance of the co-location with respect to linear scaling of intensities. This is of importance, as images are acquired under different illumination conditions and this invariance eliminates the need of image normalization. As a consequence, also expression patterns which differ by a uniform scaling of intensities not due to differing illumination are scored as very similar (see Section 3 for an example). Depending on the application this may or may not be desired. In the latter

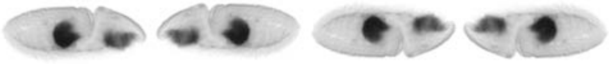


Fig. 4. Four possible orientations of an embryo.

case one may use (unnormalized) cross-correlation that is the scalar product of the expression patterns as a substitute for the correlation coefficient with prior normalization of illumination differences.

2.3 Joint clustering

Mixture models (McLachlan et al. (2000)) are the method of choice for clustering gene expression time-courses; see Bar-Joseph (2004) for a recent review. We extend a framework (Schliep et al. (2005)) using linear Hidden Markov Models as components to allow the joint analysis of gene expression time-courses and co-location information obtained from in-situ experiments. Instead of unsupervised learning we use a partially supervised approach, where constraints between genes are taken from the in-situ data. The pairwise constraints are used in the EM-algorithm with extensions proposed by Lu et al. (2005) and Lange et al. (2005).

Let the real-valued N-by-T matrix $X = \{x_i\}_{i=1}^{N}$ denote the N gene expression time-courses of length T. A mixture model is a convex combination of K component models; note that here the choice of component model is not important. We write $[x_i|\Theta] = \sum_{k=1}^{K} \alpha_k [x_i|\theta_k]$, where the non-negative α_j sum to one and the θ_j denote the parameters of the components. Then $\Theta = (\alpha_1, ..., \alpha_K, \theta_1, ..., \theta_K)$ is the set of parameters. Following (Lu et al. (2005)) we assume that—recall the complete data likelihood $[X, Y|\Theta] = [X|Y, \Theta][Y|\Theta]$—there is further dependence on pair-wise constraints $W^+, W^- \in R^{N \times N}$, yielding $[Y|\Theta, W^+, W^-] \propto [Y|\Theta][W^+, W^-|Y, \Theta]$. A positive W_{ij}^+ indicates that genes i and j should be accounted for by the same component, and a positive W_{ij}^- that they should not. Furthermore, we assume that

$$[W|\Theta, Y] = \frac{1}{Z} \exp\left(\sum_i \sum_{j \neq i} -\lambda^+ W_{ij}^+ 1\{y_j \neq y_i\} - \lambda^- W_{ij}^- 1\{y_j = y_i\} \right),$$

where λ^+ and λ^- are global weights of the constraints. The estimation problem can be solved using Gibbs-sampling or mean field approximations (Lu et. al (2005), Lange et al. (2005)). In consequence, when we set entries of W^+, W^- to zero except for strong correlations or anti-correlations, we obtain clusterings in which clusters contain co-located genes with similar expression time-courses. Results will be reported elsewhere.

3 Results

We tested the image processing pipeline using an independent set of 300 randomly chosen images from the BDGP which were manually labeled according to the types introduced in Subsection 2.1. Table 1 shows the proportion of types for manual as well as automatic labeling using the image processing pipeline proposed. About 87% of the images are suited for comparison where

Table 1. Distribution of image types for manual and automatic classification (left); Classification accuracy where the positive class are images of type I and IV (right).

Type	I	II	III	IV
manual	70.7	5.7	7.7	16.0
automatic	71.7	4.3	10.3	13.7

	True	False	Σ
Positive	80,7%	1%	81,7%
Negative	3%	15,3%	18,3%
Σ	83,7%	16,3%	100%

only 71% would be used by approaches like Peng et al. (2004), Kumar et al. (2002). Of the sets of touching embryos 73% could be separated correctly and as a consequence a total of 82% images were rendered as usable with only 1% false positive. The second important issue is the quality of the embryos extracted. Five persons were asked to assess the accuracy of the embryo contours into one of the categories (see Figure 5 for examples). With 69.4% of all images judged as good and 24% as average, the method proves well suited to register and extract embryos.

To evaluate the comparison of expression pattern with the co-location index we first used the same data set of 11 images annotated as "posterior endoderm anlage" as in Peng et al. (2004). Figure 6 shows the ranking comparing *Acf1* as query image to the remaining ten images. These results are as expected, both with regard to detailed annotation from ImaGO (compare Peng et al. (2004)) and with regard to visual impression. We note that our ranks using the co-location index deviates from the results for a correlation coefficient given in Peng (2004) which may be due to different extraction of embryos or quantization of intensities. The rankings obtained between the three methods are very similar, although the co-location index is computationally a much simpler method. Note, that the ranking of *CG5525* ahead of *Slbp* with the co-location index is due to the invariance with respect to illumination changes, see Subsection 2.2.

For a more comprehensive test we extracted ten further groups of images for developmental stage 7-8 with identical annotation. For each of the annotation terms used, we randomly selected ten images from BDGP for the set of genes returned by an ImaGO query. For the ten queries ImaGO return an average of 38.6 genes for each annotation term combination (intersection of 2 different terms) which in turn resulted in an average of 179.5 images in BDGP. We now compare each of the 100 images in turn against the complete

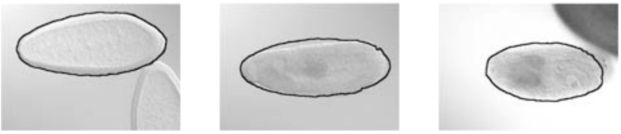

(a) good (b) average (c) bad

Fig. 5. Examples of the three categories to asses the accuracy of embryo contours.

Query	Rank	co-location index	Local GMM	Hybrid
Acf1 	1	pont 0.69 	pont 0.32 	pont 0.102
	2	mam 0.51 	mam 0.30 	mam 0.051
	3	RhoGAP71E 0.44 	RhoGAP71E 0.24 	Dcp-1 0.042
	4	Dcp-1 0.42 	Dcp-1 0.22 	RhoGAP71E 0.038
	5	CG5525 0.37 	Slbp 0.20 	Slbp 0.026
	6	Slbp 0.31 	CG5525 0.18 	cl 0.019
	7	cl 0.31 	cl 0.17 	CG5525 0.018
	8	CG6051 0.29 	CG6051 0.12 	CG6051 0.008
	9	CG33099 0.18 	GATAe 0.07 	CG33099 0.002
	10	GATAe -0.21 	CG33099 0.04 	GATAe 0.001

Fig. 6. Comparing the expression pattern of *Acf1* as query against ten other genes annotated as "posterior endoderm anlage" from ImaGO. Columns 3 to 5 give the ranking using our co-location index compared to the two favored methods of Peng et al. (2004). For each image the name of the gene and the score is given. Ranks and scores for the two matching methods of Peng et al. (2004) are take from the reference, for display we use our imageprocessing method.

BDGP dataset for developmental stage 7-8 comprising a total of 2893 images for 1768 different genes and derive the ranking using the co-location index. For the remaining nine images from the same group as the query we determine the distribution of the resulting ranks shown in Figure 7. As before the expectation is, that these nine remaining images should show up at the top of the ranking list as they share the same annotation with the query. However we can not expect to rank them exactly at the first top positions. First

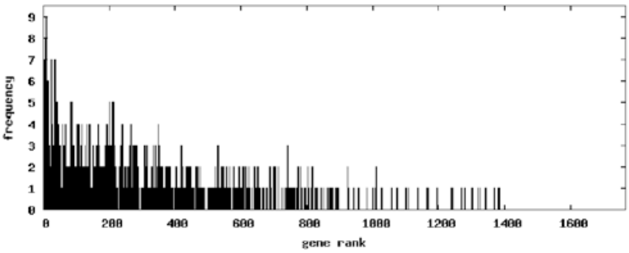


Fig. 7. Distribution of ranks from the 10 genes to each other for 10 different groups.

there are in general more images for genes identically annotated, as we choose only ten images for each group for reasons of computational efforts. Second there are ambiguities with respect to orientation (see Subsection 2.1) and also experimental and annotation inaccuracies. The histogram conforms with our expectation as the nine images are in most cases ranked ($> 67\%$) in the first third of a query result.

4 Conclusions

We presented a reliable yet simple image processing pipeline which allows to compute pair-wise co-location indices for genes from in-situ hybridization images. Furthermore, we formulate a mixture approach to use this co-location data as constraints for clustering gene expression time-courses, potentially leading to more relevant clusters of functionally related, interacting genes. This will be evaluated for the Drosophila development in further studies.

In our image processing pipeline we perform rigid transformations of the embryos with anisotropic scaling. Due to the variations in embryo shape, elastic, non-rigid transformations (Neumann et al. (1999)) might increase robustness of the co-location index. Furthermore, detection of embryo orientation, dorsal/ventral versus lateral position, is a relevant problem which needs to be addressed. Extensions to three-dimensional data as well as more intricate clustering formulations are further exciting questions to pursue.

References

BAR-JOSEPH, Z. (2004): Analyzing Time Series Gene Expression Data. *Bioinformatics, 20, 16, 2493–2503*

GONZALES, R. and WINTZ, P. (1991): *Digital Image Processing.* Addison-Wesley.

KUMAR, S., JAYARAMAN, K., PANCHANATHAN, S., GURUNATHAN, R., MARTI-SUBIRANA, A. and NEWFIELD, S. (2002): BEST - A Novel Computational Approach for Comparing Gene Expression Patterns from Early Stages of Drosophila Melanogaster Development. *Genetics, 169, 2037–2047.*

LANGE, T., LAW, M.H., JAIN, A.K. and BUHMANN, J.M. (2005): Learning with Constrained and Unlabeled Data. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 731–738.*

LU, Z. and LEEN, T. (2005): Semi-supervised Learning with Penalized Probabilistic Clustering. *NIPS 17, 849–856.*

MCLACHLAN, G. and PEEL, D. (2000): *Finite Mixture Models.* Wiley, New-York.

NEUMANN, S., POSCH, S. and SAGERER, G. (1999): Towards Evaluation of Docking Hypothesis Using Elastic Matching. *Proceedings of the GCB, 220.*

OPITZ, L. (2005): *Analyse von Bildern der mRNA-in Situ-Hybridisierung.* Master thesis, Institut für Informatik, Universität Halle-Wittenberg.

PENG, H. and MYERS, E.W. (2004): Comparing in situ mRNA Expression Patterns of Drosophila Embryos. *RECOMB'04, 157–166.*

SCHLIEP, A., COSTA, I.G., STEINHOFF, C. and SCHÖNHUTH (2005): Analyzing Gene Expression Time-courses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2, 3, 179–193.*