31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

An Automatic Quantification and Registration Strategy to Create a Gene Expression Atlas of Zebrafish Embryogenesis

C. Castro, M. A. Luengo-Oroz, S. Desnoulez, L. Duloquin, L. Fernández-de-Manuel, S. Montagna, M. J. Ledesma-Carbayo, P. Bourgine, N. Peyrieras and A. Santos

Abstract—**In order to properly understand and model the gene regulatory networks in animals development, it is crucial to obtain detailed measurements, both in time and space, about their gene expression domains. In this paper, we propose a complete computational framework to fulfill this task and create a 3D Atlas of the early *zebrafish* embryogenesis annotated with both the cellular localizations and the level of expression of different genes at different developmental stages. The strategy to construct such an Atlas is described here with the expression pattern of 5 different genes at 6 hours of development post fertilization.**

I. Introduction

UNDESTANDING the role that genes play during the embryogenesis of living animals is a major question in developmental biology. Morphogenetic processes [1] and underlying genetic activity [2-3] have been extensively studied. However, it is still an open challenge to comprehend all these processes and many approaches are being proposed in order to adequately model and predict the evolution of the gene regulatory networks [4]. To achieve this goal, it is crucial to provide accurate, quantitative data about gene expression in time and space at the single cell resolution. Recent approaches have addressed this issue in invertebrate animal models such as *Drosophila* [5].

We focused on the *zebrafish* embryo that provides a model system for the vertebrate development and offers advantages such as fast development and optical clarity during embryonic stages. The embryos were treated for double fluorescent in situ hybridization (FISH) procedures [6] to reveal the expression patterns of pairs of genes. In addition, the embryo nuclei were counter-stained to allow proper detection of every single cell. Acquisition was then achieved thanks to confocal, bi-photon laser scanning microscopy techniques [7] resulting in high-throughput 3D digital images throughout different time stages. However, FISH as well as optical limitations do not permit to reveal more than three or four fluorescent gene expression patterns at a time [8]. Nevertheless, the reconstruction of transcriptional regulatory networks involves hundreds of genes and any gene receives at least 5 or more inputs [9]. The final goal of this paper is to overcome the limitations in the detection of multiple gene expression patterns in order to reconstruct the most reliable 3D Atlas suitable for providing quantitative data at the single cell resolution level for gene network architecture construction and dynamics modeling.

The proposal first provides a method to automatically make quantitative, unbiased measurements of the levels of several gene expressions and their 3D cell-wise location based on the intensity of the FISH signal. This is achieved by acquiring high-resolution images of *zebrafish* embryos stained by FISH for two genes expressions. Being constrained by the cellular resolution, these images do not comprise the whole *zebrafish* embryo and only cover limited views restricted to the stained regions. After measuring their expression in every cell, these various fluorescent signals are all registered into the same, complete *zebrafish* template leading to the completion of a 3D Atlas where simultaneous comparison of a large number of gene products can be achieved. To do so, it was necessary to acquire lower resolution images that comprised the whole *zebrafish* embryo. In all the samples, the gene *goosecoid (gsc)* served as a reference, allowing precise matching of other gene patterns. *Gsc* is a transcriptional factor, early expressed in the region named the embryonic shield, which has been described to have the proper properties to be considered as a dorsal organizer [2]. The *gsc* pattern, together with the embryo geometrical shape given by the nuclei, were employed to perform the spatial registration that maps other gene expression data within the common template.

The paper organization starts with an overview of the complete strategy. Next, the quantification and registration techniques are described. Then, the results of the scheme proposed are illustrated for a specific time cohort -the *shield* stage at 6 hours post fertilization (hpf)- and 5 different genes -*gsc*, *chordin (chd), antivin (atv), spadetail (spt) and snail*- to finally end with the conclusion and future work.

Manuscript received April 7, 2009.

This work was supported by the Spanish Ministry of Education, through the FPU grant program, by the European projects Embryomics (NEST 012916) and BioEmergences (NEST 028892) and by the joint research project Spain-France MORPHONET (HF2007-0074).

C. Castro, M. A. Luengo-Oroz, L. Fernández-de-Manuel, M. J. Ledesma-Carbayo and A. Santos are with the Group of Biomedical Image Technologies, ETSIT, Universidad Politécnica de Madrid, Madrid 28040 Spain. {ccastro, maluengo, lfernandez, mledesma, andres}@die.upm.es.

S. Desnoulez, L. Duloquien and N. Peyrieras are with DEPSN, CNRS, Institut de Neurobiologie Alfred Fessard, Gif-sur-Yvette 91190, France. {desnoulez, louise.duloquin, nadine.peyrieras}@inaf.cnrs-gif.fr

P. Bourgine is with CREA-Ecole Polytechnique, Paris 75015, France. {paul.bourgine}@polytechnique.edu

S. Montagana is with the Department of Electronics, Informatics and Systems, Universita di Bologna, Cesena 47023, Italy. {sara.montagna}@unibo.it

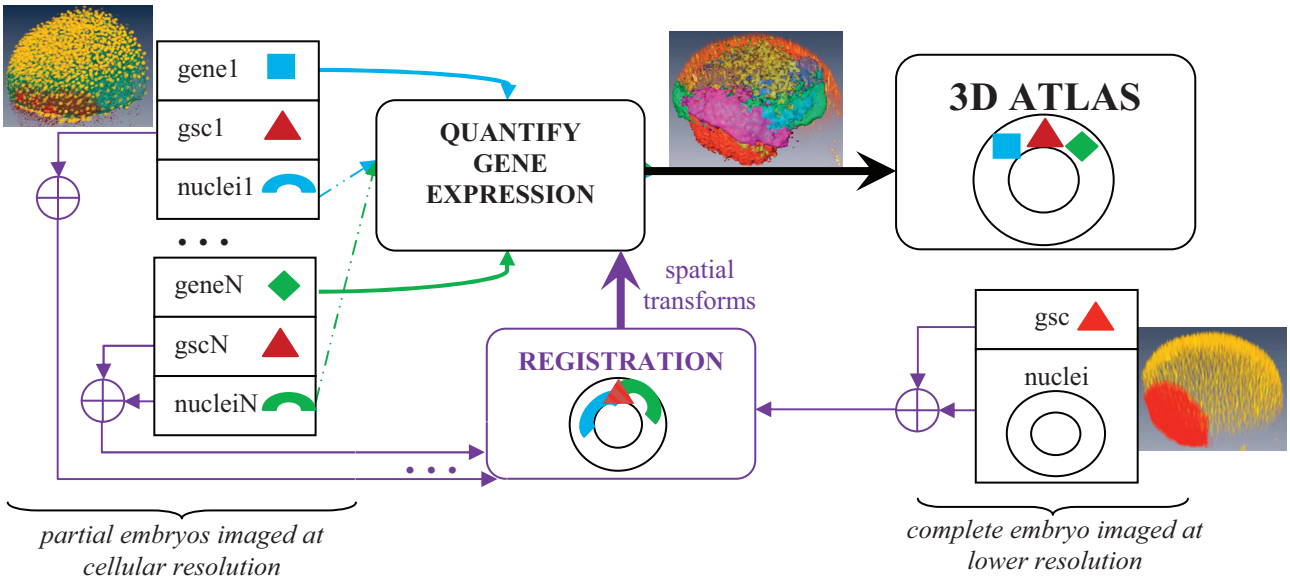

Fig. 1. Block diagram depicting our proposal to get a *zebrafish* Atlas.

II. Methodology

Our strategy to create a gene expression Atlas of the *zebrafish* embryo is depicted in Fig. 1 and involves two main processes: 1) A gene expression quantification scheme that works at the cellular resolution with partial images of the embryo centered around the gene of interest. The nuclei color channel of these images is used to identify the cells while the gene expression color channel is used to locate and quantify the relative amount of RNA provided by its transcriptional activity within those cells. All the datasets include a *gsc* gene color channel used as common reference. 2) A spatial registration method that works at lower resolution levels with images comprising the whole embryo and labeled to include a *gsc* channel too. The low-resolution, blurred nuclei serve to give an anatomical reference of the embryo where to map the partial views. The *gsc* pattern is also used as a reference to sort out the appropriate spatial transformations that match partial, high-resolution views into the complete, low-resolution template. These spatial transformations allow gathering into the final 3D Atlas information about the cell-wise location and relative amount of RNA of as many genes as required at a given developmental stage. The following Sections describe the details about these 2 main algorithms.

A. Gene expression quantification

The goal is to obtain quantitative measures about gene expression in each cell of individual embryos. Each embryo was fluorescently labeled to stain its nuclei and reveal two different gene expression patterns. One of them always being the referential *gsc* whose transcripts were marked with the red fluorescent dye Cy5, while the transcripts of the other genes under study were always marked with the green fluorescent dye FITC [6].

The strategy to segment and measure gene expression is depicted in Fig. 2. The first step involves the segmentation of the cellular domains within the embryo. To do so, we developed an automatic nuclei detection algorithm that works over the nuclei marker. Keeping residues of morphological area openings [10] with the appropriate sizes we isolated the elements having the typical volume of a cell nucleus. After that, the centroids of remaining connected components correctly identified the cells nuclei.

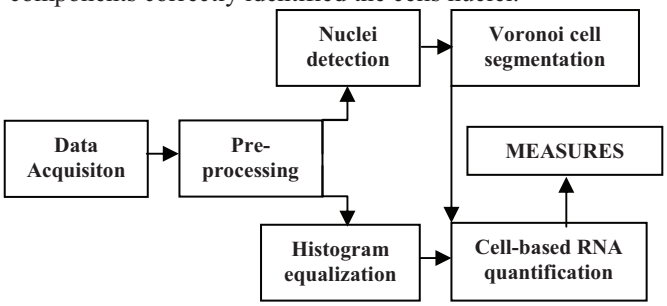

Fig. 2. Block diagram depicting the gene expression quantification scheme.

Once the nuclei were detected, they were used as seeds to generate their corresponding Voronoi diagrams, i.e. the regions of influence surrounding each nucleus. Previous work [11] showed that Voronoi diagrams provide a fairly accurate model of the true cell geometries. Consequently, all measurements and quantifications can be assigned to individual cells.

After that, we assumed that the gene expression level is proportional to the fluorescence brightness within the embryo [12]. Thus, we focused on the mean intensity values present in each cell in order to assign a transcriptional activity score ranging from 0 to 1 (Fig. 3). However, the different acquisition conditions, especially those concerning gain and saturation, proved to significantly bias intensity measurements from one embryo to another. Moreover, the staining provided by the two fluorescent dyes employed - FITC and Cy5- presents intrinsic variations in brightness and half-life [6]. All these factors yielded non-consistent scores, which made some kind of normalization necessary.

To address this difficulty a set of embryos were imaged such that the *gsc* reference was labeled twice, both with FITC and Cy5 staining. This allowed finding out the proper transformation to obtain equivalent measures under different imaging conditions. Given that the gray-level histogram carries the information about the gene intensity pattern, we implemented a histogram matching algorithm [13] that equalized the FITC histogram of the *gsc* into the Cy5 one so that the score differences between them were minimal. Then, this FITC to Cy5 histogram transformation, obtained from the *gsc-gsc* set, was taken as the model to equalize the rest of FITC-marked gene patterns into the referential Cy5-marked *gsc*.

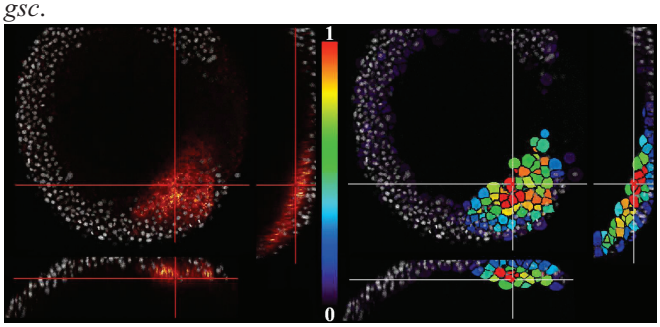

Fig. 3. On the left, cross sections of the *zebrafish* nuclei with the *gsc* expression labeled by Cy5. On the right, their corresponding Voronoi cells after being scored according to their gene expression level.

Finally, once the scores were corrected to impartially reflect the transcriptional activity, we quantified the gene expression patterns. We found the number and location of cells affected by the gene expression domains, calculated their volumes and surfaces and computed their cross volume with respect to the referential *gsc*. Since the gene expression domains in our datasets do not have sharp boundaries, most of the previous measures were somewhat subjective. In order to deal with this issue, we implemented *fuzzy* versions of the previous measurements based on the 0 to 1 real scores assigned to each cell. Finally, we constructed bi-dimensional histograms that easily show the joint distribution of scores between any gene and *gsc*. We believe these histograms are a valuable tool to evaluate the degree of spatial correlation between different gene expression domains (Fig. 4).

B. Registration

Our spatial registration algorithm always works with two sets of images: On one side, the partial, high-resolution views of the *zebrafish* embryos acting as the moving images that have to be mapped into the common template. On the other side, the complete, low-resolution image of a *zebrafish* instance, fixed during the registration process and acting as the common template mentioned above.

In both set of images, fixed and moving, two different channels were employed during the registration process: the nuclei information, which gives details about the anatomical constraints, and the *gsc* pattern, which provides universal reference about where the gene samples are placed.

Prior to the registration process it was necessary to preprocess the original input images to adequate them to the chosen registration method. For the sake of simplicity, we made a single image including the information of both the nuclei and the *gsc* channels. This was performed for both the fixed and the moving image stacks. The original nuclei channels were first smoothed and downsampled. Then, their domain was segmented and turned into a binary mask. This way, we were able to speed up the registration process while keeping generic information of the *zebrafish* shape that does not outline specific geometric variations across specimens. Meanwhile, the original *gsc* expression patterns were smoothed and downsampled likewise. Then, their domain of expression was segmented before adding them to their corresponding nuclei binary masks. This addition was made by enhancing the gene patterns so that their weight guiding the spatial transformations was stronger.

We chose a 3D rigid registration scheme that performs translations and rotations on the moving image to maximize its similarity to the fixed image, measured by a specific metric. To speed up convergence, this registration process was first initialized by aligning the center of mass in both datasets. Then, the optimization process until reaching maxima was broken up in two steps: First, a Differential Evolution optimizer, implemented according to [14], was used to thoroughly search for a global minimum rather than focusing on a local version of it. Then, the area of this global minimum was further explored by means of an iterative Gradient Descent optimizer that refined the final solution. The metric evaluated during the optimizers iterations was the cross-correlation between the fixed and the moving image. Finally, the moving image had to be resampled by linear interpolation into the fixed image in order to adjust their pixel resolution differences.

As a result, we obtained a complete, adequate transform between both sets. This transformation was later applied to the expression patterns of those genes co-stained with *gsc* in the high-resolution images. This process led to the final formation of the 3D *zebrafish* Atlas annotated with the quantifications obtained from the previous phase (Fig. 1).

III. EXPERIMENTS

Image acquisition was performed with a Leica SP2 bi-photon and confocal laser scanning microscope equipped with a Leica objective 10x HCX APO 0,5W U-V-I for the low-resolution, complete sets and with a 20x objective, for the high-resolution, partial volumes. The corresponding acquisition times, image sizes and resolutions were 90 minutes, 1024x1024x333 voxels and 0.569x0.569x1.17 µm and 45 minutes, 1024x1024x135 voxels and 0.521x0.521x1.04 µm respectively.

The nuclei detection approach presented in Section II.A was manually tested over more than 300 nuclei yielding a total detection rate around 94%, 3,1% of the total nuclei were false positives and 2,5% were false negatives.

For the quantification method, we validated our histogram normalization method by comparing all the Cy5 vs. FITC

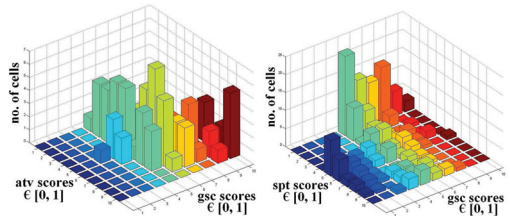

Fig. 4. On the left, the primacy of the diagonal shows a high correlation between *atv* and *gsc*. On the right, the primacy of the XY axes proves *spt* and *gsc* are mainly uncorrelated.

cell scores detected in the *gsc-gsc* stacks. Thus, a total of 12 embryos and 15.973 cells were evaluated in order to robustly assess the accuracy of the final normalized scores. 67% of all the involved cell scores differed by less than 15% before histogram normalization. This rate raised up to 87% after the histogram matching. Quantification was then performed on 5 different gene sets: *gsc, atv, chd, snail* and *spt*. Then, the volume and surface of these gene patterns were measured and correlated against their *gsc* referential. In addition, all their cells were scored according to their genetic activity. The joint distribution of these scores was correlated against those of the *gsc* referential in 2D histograms (Fig. 4).

For the registration, the final correlation metrics between the *gsc* patterns in the high-dimensional images and the one in the low-dimensional *zebrafish* turned out to be: 0,54, 0,68, 0.69 and 0,70 out of a maximum of 1 for the *atv, chd, snail* and *spt* datasets respectively. Processing time was 10 minutes per dataset in an Intel CoreDuo at 3GHz. The correlation metric found by the human expert when trying to align the gene patterns manually was around 0,47. In addition, visual inspection of the final result, shown in Fig. 5, and its comparison to the existing qualitative description [3] of *zebrafish* gene expressions positions at the *shield* stage (6 hpf) appeared to be satisfactory.

IV. Conclusion

We presented a complete methodology and framework to construct a 3D+t gene expression Atlas. The extension of the presented results to the representation of new genes and time stages is straight-forward.

Future work involves contributing with additional genes and time points to this Atlas so that it can be employed in biological simulations. At the same time, the quantification and registration schemes will be further optimized and validated against the available qualitative data. Inter-embryo variability will be taken into account by proposing the construction of a statistical *zebrafish* embryo template.

References

[1] C. B. Kimmel, W. W. Ballard, S. R. Kimmel, B. Ullmann and T. F. Schilling. *"Stages of embryonic development of the zebrafish"*, *Developmental Dynamics*, vol. 203, issue 3, pp. 253-310, 1995.

[2] A. F. Schier and W. S. Talbot. *"Molecular genetics of axis formation in zebrafish"*, *Ann. Rev. Genet*, vol. 39, pp. 561-613, Dec. 2005.

[3] www.zebrafishgrns.org.

[4] R. Doursat. *"The growing canvas of biological development: Multiscale pattern generation on an expanding lattice of gene regulatory networks", InterJournal: Complex Systems*, 1809, 2006.

[5] C. C. Fowlkes, C. L. Luengo Hendriks, Soile V. E. Keränen, G. H. Weber, O. Rübel, M. Y. Huang, S. Chatoor, A. H. DePace, L. Simirenko, C. Henriquez, A. Beaton, R. Weiszmann, S. Celniker, B. Hamann, D. W. Knowles, M. D. Biggin, M. B. Eisen and J. Malik. *"A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm", Cell*, 133, pp. 364-374, April 2008.

[6] T. Brend and S. A. Holley. *"Zebrafish whole mount high-resolution double fluorescent in situ hybridization", Journal of Visualized Experiments,* 2009.

[7] S. G. Megason and S. E. Fraser. *"Digitizing life at the level of the cell: high-performance laser-scanning microscopy and image analysis for in toto imaging of development"*, *Mech.Dev.*, vol. 120, pp. 1407–1420, 2003.

[8] B. N. Giepmans, S. R. Adams, M. H. Ellisman and R. Y. Tsien. *"The fluorescent toolbox for assessing protein location and function", Science,* 312(5771):217-24, April 2006.

[9] T. M. Chan, C. H. Chao, H. D. Wang, Y. J. Yu and C. H. Yuh. *"Functional Analysis of the Evolutionarily Conserved Cis-Regulatory Elements on the Sox17 Gene in Zebrafish", Developmental Biology,* 326, pp. 456-470, November 2008.

[10] L. Vincent, *"Morphological area opening and closing for grayscale images"*, *Shape in Picture Workshop*. NATO, Springer-Verlag, pp. 197–208, 1992.

[11] M. A. Luengo-Oroz., L. Duloquin, C. Castro, T. Savy, E. Faure, B. Lombardot, P. Bourgine, N. Peyriéras and A. Santos. *"Can Voronoi Diagram Model Cell Geometries in Early Sea-Urchin Embryogenesis?". 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2008),* pp. 504-507, May 2008.

[12] G. H. Weber, O. Rübel, M. Y. Huang, A. H. DePace, C .C. Fowlkes, S. V. E. Keränen, C .L. Luengo Hendriks, H. Hagen, D .W. Knowles, J. Malik, M. D. Biggin and B. Hamann. *"Visual exploration of three-dimensional gene expression using physical views and linked abstract views", IEEE Transactions on Computational Biology and Bioinformatics*, 2008.

[13] K. R. Castleman. *"Digital Image Processing", Prentice Hall,* pp. 91-94, 1996.

[14] R. Storn and K. Price. *"Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continous Spaces", Journal of Global Optimization,* vol. 11, pp. 341-359, 1997.

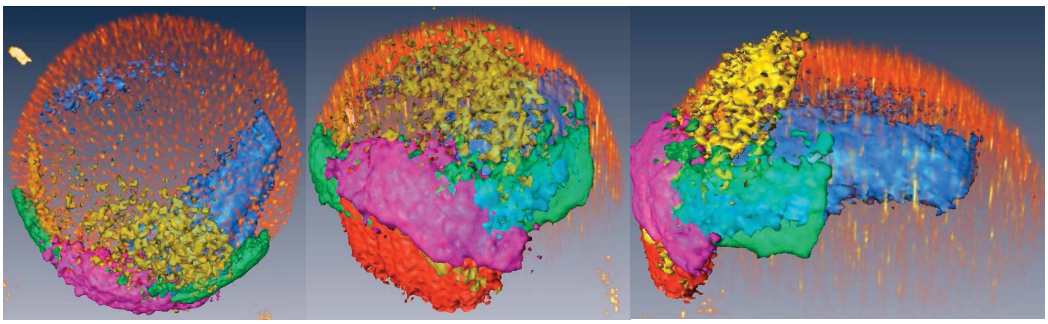

Fig. 5. From left to right: animal pole, dorsal and lateral views of the final *zebrafish* 3D Atlas where the 5 genes were registered: *gsc* (red), *chd* (purple), *spt* (green), *snail* (blue) and *atv* (yellow).