33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Image Processing Challenges in the Creation of Spatiotemporal Gene Expression Atlases of Developing Embryos

Carlos Castro*, Miguel Ángel Luengo-Oroz*, Louise Douloquin†, Thierry Savy‡, Camilo Melani‡, Sophie Desnoulez†, Paul Bourgine‡, Nadine Peyriéras†, María Jesús Ledesma-Carbayo*, Andrés Santos*

Abstract— **To properly understand and model animal embryogenesis it is crucial to obtain detailed measurements, both in time and space, about their gene expression domains and cell dynamics. Such challenge has been confronted in recent years by a surge of atlases which integrate a statistically relevant number of different individuals to get robust, complete information about their spatiotemporal locations of gene patterns. This paper will discuss the fundamental image analysis strategies required to build such models and the most common problems found along the way. We also discuss the main challenges and future goals in the field.**

I. INTRODUCTION

Understanding the processes that pattern embryonic development is one of the mayor challenges in the post-genomic era. As a matter of fact, despite having nearly completed the genome sequence of most living systems [1], we are still very far from identifying how these genetic expressions relate to the spatiotemporal dynamics of cells and their differentiation into diverse tissues and organs.

The recent explosion of biological imaging techniques [2] [3] together with remarkable advances on biological labeling [4] [5] have changed matters dramatically by providing an enormous wealth of data with the required spatial and temporal resolution to tackle this problem. This flood of new data has challenged engineering-related disciplines [6] to develop novel image processing techniques with the ultimate goal of building a digital model of animal embryogenesis where we can unambiguously quantify, for every cell in the embryo and any time of development, the levels of expression of all the genetic products that define the system behavior [7]. This kind of study is highly relevant for the investigation of cancer diseases [8] and the pharmaceutical testing of new therapeutic drugs [9].

Common image acquisition procedures employed include confocal and multi-photon optic microscopy [10], capable of providing 3D data at the cellular resolution level, and Selective Plane Illumination Microscopy (SPIM) [11] which combines high-speed imaging and minimal phototoxicity to allow the fast recording of entire living embryos over long periods of time [12]. At the same time, Fluorescent In Situ Hybridization (FISH) techniques [13] permit the simultaneous visualization of several gene expression patterns within the same embryo. However, even though our ambitious goals above imply the quantification of hundreds of gene expressions patterns, FISH as well as optical filtering limitations do not permit to reveal more than a few colors at a time [14]. Thus, in order to allow direct comparisons between all the transcription factors involved in our system we need not only to perform a systematic acquisition of different embryos, each of them being stained for a different set of gene expressions, but also to develop the corresponding image processing methods to map all these different stacks into one common, canonical space where all the information coming from different acquisitions can be simultaneously studied. The output of these methods is usually referred in literature as an Atlas of genetic expression.

Long et al. [15] built a digital 3D atlas from confocal images of 15 C. elegans worms. They are small enough as to image their entire development at single-cell resolution and present a stereotyped cell lineage development which greatly facilitates quantitative comparisons between whole different individuals. Fowlkes et al. [16] integrated significantly variable data coming from hundreds of different biphoton acquisitions of Drosophila flies onto a common 3D framework spanning 95 different gene expressions at 6 time cohorts. Keller et al. [17] used SPIM to acquire 24-hour time-lapse recordings of 7 in-vivo zebrafish embryogenesis which allowed them to carry out a high-throughput qualitative study of 4D (3D+T) cell positions, divisions and migratory tracks. Finally, Lein et al. [18] achieved a comprehensive automated 3D expression atlas of the mouse brain. In spite of yielding relatively low temporal and spatial resolution and limited gene quantization, this project spans over $20,000$ genes which were composed together by using anatomical landmarks.

The purpose of this paper is to briefly review the main techniques and problems, available tools and future challenges of biological image methods related to the construction of genetic atlas in developing embryos. Section II depicts the fundamental image processing techniques to build a gene expression atlas. Then, Section III describes the future challenges to be faced to finally discuss the conclusions in Section IV. While in this short article it is difficult to include all the important work and to explain the details of the introduced applications and computing methods, we hope

This work was supported by the FPU grant program (Spanish Ministry of Education), TEC2008-06715-C02-02 (Spain), ANR and ARC (France), FP6 New Emerging Science and Technology EC program and European Regional Development Funds (FEDER).

*C. Castro, M.A. Luengo-Oroz, M.J. Ledesma and A. Santos are with Biomedical Image Technologies, Universidad Politécnica de Madrid, 28040, Spain. {ccastro, maluengo, mledesma, andres}@die.upm.es

†L. Douloquin, S. Desnoulez and N. Peyriéras are with N&D, CNRS, Institut de Neurobiologie Alfred Fessard, Gif-sur-Yvette 91190, France. {louise.duloquin, nadine.peyrieras}@inaf.cnrs-gif.fr

‡T. Savy, C. Melani and P. Bourgine are with CREA-École Polytechnique, Paris 75015, France. {paul.bourgine}@polytechnique.edu

that the presented facts and pointers can be helpful for both researches in the field and general audiences who may have interest in learning the basic ideas of gene atlas creation.

II. IMAGE PROCESSING METHODS

As engineers, our contribution to the ambitious biological objectives depicted above consists in developing the required computational tools to detect and track every cell and quantify the interaction of every gene product in images of any developing tissue in a standardized, cell-based manner. In other words, we aim at achieving a complete definition of the (x, y, z, t, G) data universe where we could describe the levels of expression of every protein in the genome (G) at all developmental times (t) for every cellular position (x, y, z) in a developing organism. The fundamental techniques to accomplish this task are depicted below. However, there is no universal solution to the problems tackled by these techniques. Algorithms that work well on one set of images will often not work on another set due to significant differences in terms of embryo shape, cell density, signal-to-noise ratio, image resolution, bio-markers employed, etc.

A. Nuclei Detection

Methods to determine cellular positions (x, y, z) are often based on segmenting cell nuclei rather than whole cells because nuclei tend to have a more compact shape and less overlap with their neighbors. In [19] we used an approach based on morphological operators: First, a granulometry of nuclei images provides the archetypal upper and lower sizes of the objects within. Then, using these sizes to keep residues of morphological area openings [20] directly isolates all elements with the typical volume of a cell nucleus. Other approximations to the problem include the numerical solution of a 3D nonlinear advection-diffusion equation as proposed in [21]. Adaptive thresholding, gradient methods, 3D watershed algorithms or training-based schemes based on support vector machine classifiers are also popular solutions as described in [15].

B. Gene Quantification

Image-based quantification of the levels of expression of one specific gene product within one specific cell, (x, y, z, g_1), requires the prior application of cell shape segmentation algorithms. In this respect, models based on Voronoi diagrams built from the nuclei centers [22] showed an interesting correlation with the actual true cell geometries. Other strategies based on differential-equation segmentations of the true shapes have also been successfully applied [23]. Once cell volume has been extracted, gene expression grades can be extracted based on the general assumption of a linear relationship between the fluorescence intensity enclosed in a cell and the protein concentration within it [24]. However, while this approach has proven valid when referred to one individual embryo, differences on the acquisition conditions and non-linear effects such as photobleaching, the fluorescent dye's extinction coefficient or the depth-dependent signal decay make it arguable whether these figures can be directly compared when coming from different probes. With this problem in mind, Damle et al. [25] achieved a proper fluorescence-to-protein conversion by applying depth-extinction corrections (based on measurements of the signal decay of a second freely diffusible internal fluorescent standard) and by deriving a realistic conversion constant obtained after imaging a set of probes with a known quantity of their target fluorescent protein.

C. 3D Atlas Construction

The composition of an adequate 3D $(x, y, z, g_1, ..., g_N)$ map containing a sufficiently representative number of different genes, N, is achieved by registration procedures. These algorithms are designed to process sets of image stacks coming from different individuals -which differ in size and orientation- so that they fit into the same referential space making the positions of their nuclei directly comparable. Such "virtual multiplexing" makes it practical to examine the relations in the expression of all these genes without having to directly co-stain embryos for all possible pairs. Most registration procedures in literature include an initialization algorithm which aims at removing some of the geometrical variation among embryos by coarsely aligning anatomical landmarks [18] or the animal mayor axes [15] [16] [2]. Posterior fine registration procedures include pixel-based alignment methods adapted from medical image analysis such as mutual information registration [19] or spline elastic registration [26]. However, object-based registration is also possible: work in [15] and [16] includes modelization of the statistical cell positions allowing the identification of correspondences between differently extracted nuclei. Both approaches have pros and cons: statistical modeling helps ruling out inter-embryo variability due to individuals differences in shape, size or deformations but implies working in an object-based domain where we lose track of the originally acquired data. On the contrary, in a pixel-based approach we are able to keep the original renders, surfaces and volumes of the gene expressions but the non-trivial variations in the extent of gene patterns and number and density of involved nuclei makes it challenging to discern whether the final variations in the atlas correspond to biological or methodological sources (e.g. registration errors, imperfect synchronization between samples, etc.)

D. Cell Tracking and Lineage Reconstruction

Tracking algorithms applied to cellular movements are capable of following nuclei positions through time, that is to say, specifying their (x, y, z, t) coordinates. Solutions implemented include 3D methods which must be first given the identified (x, y, z) positions for every instant of a time-lapse series of images. That is the case in [27] where the vector field is estimated from registration between two consecutive time steps. This vector field is used to predict which position should each detected nuclei at t have at $t + 1$ to then assign the closest detected nuclei in $t + 1$ as the continuation of the trajectory. There are also tracking algorithms which work entirely in 4D [28] by detecting

connected paths in the 3D+T space thanks to morphological reconstructions from a manual or automatic marker. Once cell trajectories have been identified, it is useful to perform mitosis detection algorithms which allow to connect tracks across cell divisions and reconstruct the cell lineage across the embryo development. This kind of analysis was recently applied successfully to unveil the phenomenological behavior of cell proliferation and division waves in early zebrafish development [29].

E. Integration of Cell Tracking and Gene Quantification

Since the emergence of transgenic embryo lines [30], animal models can be engineered to express a fluorescent labeling as they continue to live and carry a normal development. Contrary to FISH techniques, where embryos need to be fixated, these methods, combined with fast, time-lapse microscopy, open up the possibility to perform developmental studies in vivo. In [31], the complete (x, y, z, t, g_1) coordinates of a transgenic fish line, labeled to show one specific gene expression g_1, were determined by successively applying nuclei detection, cell shape extraction, gene quantification and cell tracking techniques. Such approach allows a straightforward study about how cell dynamics and lineage are influenced by genetic activity: Maps of cell speeds show parallelisms to those indicating levels of gene expression and studies about how gene intensities evolve through a cell progeny indicated that gene propagation is not necessarily attached to family links.

F. Validation

All the previously introduced algorithms are usually designed to operate in an automatic way so that they can systematically handle the enormous amount of data involved in the creation of a gene atlas (N big 3D images where $N =$ no. of genes $\times$ no. of individuals per gene $\times$ no. of time steps). However, due to the lack of gold standards in the field, output results generally require different levels of supervision with manual validation and correction not being uncommon [15]. Given the intrinsic spatial nature of the atlas creation problem, visual assessment is the common validation norm for virtually all previously described methods. Consequently, the development of sophisticated visualization tools adapted to each particular dataset becomes ineludible, see Section II-G. Registration performance is perhaps the exception to this norm since visual assessment becomes challenging when lacking specific anatomical landmarks [16]. Indirect validation criteria are then used: when variability measurements between gene patterns in the atlas are in the same range as the variability that one embryo directly co-stained for those gene patterns would show when compared to another individual, then we can take it as a proof that our registration process succeeded in its task of factoring out geometric variability to keep just biological disparity.

G. Visualization

As discussed previously in Section II-F, the development of an advanced visualization platform is an essential requirement of any system dealing with gene expression atlas. Ideally, this visualization platform should be used not only to picture the multidimensional input data and output results but also to run the previously described algorithms on request while providing the necessary tools to correct, validate, annotate and quantify their outcomes. Several software packages described in literature cover some of these aspects, each of them being specifically designed to the needs and data of each project. Some relevant examples are FlyEx [32], MovIt [33], GoFigure [34] or PointCloudExplore [35] [36].

III. FUTURE CHALLENGES

In Section II we discussed some of the latest methods in the lead of achieving a quantitative depiction of joint genetic and cellular dynamics. None of these projects, however, has yet accomplished the ultimate goal of a complete spatio-temporal, (x, y, z, t, G), representation of an embryo's genome throughout its embryogenesis. Here we present two of the mayor breakthroughs that will help completing this challenge in the next few years.

A. Multiplex in-situ

The next generation of in situ hybridization techniques is expected to overcome the limitations on the number of possible gene expressions that can be simultaneously labeled. Pierce et al. have just developed a new multiplexing technique that allows the fluorescent labeling of up to 5 different mRNA targets at a time [37]. Compared to the current double in-situ hybridization techniques, this scheme will drastically ease the addition of new genetic probes. Registration methods will also benefit as some of the 5 available tags could be used for referential purposes providing the necessary landmarks to increase the final accuracy.

B. 4D Atlas and Cell Lineage

The introduction of transgenic animal lines -that can be labeled to show the expression of a certain gene in vivo- introduce the possibility of developing registration methods integrally working in 4D which would directly yield $(x, y, z, t, g_1 ... g_N)$ maps. This scheme would bring a big leap in the field since gene expression could be continuously quantified in time rather than having discrete samples at the different developmental stages where in situ hybridizations are defined. The big challenge underlying here will be to make a comprehensive study correlating the genome dynamics with cell migration and differentiation.

IV. DISCUSSIONS AND CONCLUSION

We have showed the interest of creating digital atlases that provide cellular locations and intensity levels of different gene expressions throughout embryo development. We have later introduced the key image processing tools used in literature to segment, register, track, quantify, validate and visualize the results concerning the construction of such atlases. Main applications include unveiling the mechanism of the gene regulatory networks that control the embryogenesis [16] or examining hypothesis about an animal model anatomy [15]. Such studies are made possible thanks to the

quantitative information that a gene expression atlas can provide. Counting the number of cells within each gene pattern, measuring co-expression between different domains, quantifying their volumes and contact surfaces and seeing how these figures evolve with time are key aspects to model a realistic gene regulatory network [38]. On the other hand, animal atlases, which comprise average gene patterns and cell evolutions of dozens of different individuals, are the perfect framework to identify cell populations consistently diverging from the norm and to investigate its nature. Future challenges notably include a further exploration linking gene expression behavior with cell lineage evolution.

REFERENCES

[1] A.F. Schier and W.S. Talbot. Molecular genetics of axis formation in zebrafish. *Annu. Rev. Genet.*, 39:561–613, 2005.

[2] S.G. Megason and S.E. Fraser. Imaging in Systems Biology. *Cell*, 130(5):784–795, 2007.

[3] A. Abbott. Microscopic marvels: Seeing the system. *Nature*, 459(7247):630, 2009.

[4] D.M. Chudakov, S. Lukyanov, and K.A. Lukyanov. Fluorescent proteins as a toolkit for in vivo imaging. *Trends in biotechnology*, 23(12):605–613, 2005.

[5] C. Vonesch, F. Aguet, J.L. Vonesch, and M. Unser. The colored revolution of bioimaging. *Signal Processing Magazine, IEEE*, 23(3):20–31, 2006.

[6] H. Peng. Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24(17):1827, 2008.

[7] A.C. Oates, N. Gorfinkiel, M. González-Gaitán, and C.P. Heisenberg. Quantitative approaches in developmental biology. *Nature Reviews Genetics*, 10(8):517–530, 2009.

[8] J.F. Amatruda, J.L. Shepard, H.M. Stern, and L.I. Zon. Zebrafish as a cancer model system. *Cancer Cell*, 1(3):229–231, 2002.

[9] L. Yang, N.Y. Ho, R. Alshut, J. Legradi, C. Weiss, M. Reischl, R. Mikut, U. Liebel, F. Müller, and U. Strähle. Zebrafish embryos as models for embryotoxic and teratological effects of chemicals. *Reproductive Toxicology*, 28(2):245–253, 2009.

[10] 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. *Mechanisms of development*, 120(11):1407–1420, 2003.

[11] J. Huisken, J. Swoger, F. Del Bene, J. Wittbrodt, and E.H.K. Stelzer. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science*, 305(5686):1007, 2004.

[12] P.J. Keller and E.H.K. Stelzer. Quantitative in vivo imaging of entire embryos with digital scanned laser light sheet fluorescence microscopy. *Current opinion in neurobiology*, 18(6):624–632, 2008.

[13] H. Clay and L. Ramakrishnan. Multiplex fluorescent in situ hybridization in zebrafish embryos using tyramide signal amplification. *Zebrafish*, 2(2):105–111, 2005.

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

[15] F. Long, H. Peng, X. Liu, S.K. Kim, and E. Myers. A 3D digital atlas of C. elegans and its application to single-cell analyses. *Nature Methods*, 6:667–672, 2009.

[16] C.C. Fowlkes, C.L.L. Hendriks, S.V.E. Keranen, G.H. Weber, O. Rubel, M.Y. Huang, S. Chatoor, A.H. DePace, L. Simirenko, C. Henriquez, et al. A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell*, 133(2):364–374, 2008.

[17] P.J. Keller, A.D. Schmidt, J. Wittbrodt, and E.H.K. Stelzer. Reconstruction of Zebrafish Early Embryonic Development by Scanned Light Sheet Microscopy. *Science*, 322(5904):1065–1069, 2008.

[18] ES Lein, MJ Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, AF Boe, MS Boguski, KS Brockway, EJ Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2007.

[19] C. Castro, MA Luengo-Oroz, S. Desnoulez, L. Duloquin, S. Montagna, MJ Ledesma-Carbayo, P. Bourgine, N. Peyrieras, and A. Santos. An automatic quantification and registration strategy to create a gene expression atlas of zebrafish embryogenesis. In *Proc. IEEE EMBS Conference*, pages 1469–1472, 2009.

[20] L. Vincent. Morphological area openings and closings for grey-scale images. *NATO ASI Series of Computer and Systems Sciences*, 126:197–197, 1994.

[21] O. Drblikova, M. Komornikova, M. Remesikova, P. Bourgine, K. Mikula, N. Peyriéras, and A. Sarte. Estimate of the cell number growth rate using PDE methods of image processing and time series analysis. *J. Electr. Eng.*, 58(7):86–92, 2007.

[22] MA Luengo-Oroz, L. Duloquin, C. Castro, T. Savy, E. Faure, B. Lombardot, R. Bourgine, N. Peyriéras, and A. Santos. Can voronoi diagram model cell geometries in early sea-urchin embryogenesis? In *Proc. IEEE ISBI Conference*, pages 504–507, 2008.

[23] C. Zanella, M. Campana, B. Rizzi, C. Melani, G. Sanguinetti, P. Bourgine, K. Mikula, N. Peyrieras, and A. Sarti. Cells segmentation from 3-d confocal images of early zebrafish embryogenesis. *Image Processing, IEEE Transactions on*, 19:770–781, 2010.

[24] J.Q. Wu and T.D. Pollard. Counting cytokinesis proteins globally and locally in fission yeast. *Science*, 310(5746):310, 2005.

[25] S. Damle, B. Hanser, E.H. Davidson, and S.E. Fraser. Confocal quantification of cis-regulatory reporter gene expression in living sea urchin. *Developmental biology*, 299(2):543–550, 2006.

[26] K. Rohr, M. Fornefett, and HS Stiehl. Spline-based elastic image registration: integration of landmark errors and orientation attributes. *Computer Vision and Image Understanding*, 90(2):153–168, 2003.

[27] C. Melani, M. Campana, B. Lombardot, B. Rizzi, F. Veronesi, C. Zanella, P. Bourgine, K. Mikula, N. Peyriéras, and A. Sarti. Cells tracking in a live zebrafish embryo. In *Proc. IEEE EMBS Conference*, pages 1631–1634, 2007.

[28] D. Pastor, MA Luengo-Oroz, B. Lombardot, I. Gonzalvez, L. Duloquin, T. Savy, P. Bourgine, N. Peyrieras, and A. Santos. Cell tracking in fluorescence images of embryogenesis processes with morphological reconstruction by 4D-tubular structuring elements. In *Proc. IEEE EMBS Conference*, pages 970–973, 2009.

[29] N. Olivier, M.A. Luengo-Oroz, L. Duloquin, E. Faure, T. Savy, I. Veilleux, X. Solinas, D. Débarre, P. Bourgine, A. Santos, et al. Cell lineage reconstruction of early zebrafish embryos using label-free nonlinear microscopy. *Science*, 329(5994):967, 2010.

[30] S. Pauls, B. Geldmacher-Voss, and J.A. Campos-Ortega. A zebrafish histone variant H2A. F/Z and a transgenic H2A. F/Z: GFP fusion protein for in vivo studies of embryonic development. *Development genes and evolution*, 211(12):603–610, 2001.

[31] C. Castro, M.A. Luengo-Oroz, L. Douloquin, T. Savy, C. Melani, S. Desnoulez, M.J. Ledesma-Carbayo, P. Bourginey, N. Peyriéras, and A. Santos. Towards a digital model of zebrafish embryogenesis. Integration of cell tracking and gene expression quantification. In *Proc. IEEE EMBS Conference*, pages 5520–5523, 2010.

[32] Andrei Pisarev, Ekaterina Poustelnikova, Maria Samsonova, and John Reinitz. Flyex, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Res*, 37(Database issue):D560–D566, Jan 2009.

[33] T. Savy and EMBRYOMICS. MOVIT: Morphogenesis Visualization Tool. unpublished, 2011.

[34] A. Gouaillard, T. Brown, M. Bronner-Fraser, S.E. Fraser, and S.G. Megason. GoFigure and The Digital Fish Project: Open tools and open data for an imaging based approach to system biology. *Insight Journal*, 2007.

[35] O. Rübel, G.H. Weber, M.Y. Huang, E.W. Bethel, M.D. Biggin, C.C. Fowlkes, C.L.L. Hendriks, S.V.E. Keränen, M.B. Eisen, D.W. Knowles, et al. Integrating data clustering and visualization for the analysis of 3d gene expression data. *IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 64–79, 2010.

[36] G. H. Weber, O. Rubel, Min-Yu Huang, A. H. DePace, C. C. Fowlkes, S. V. E. Keranen, 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. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6(2):296–309, 2009.

[37] H.M.T. Choi, J.Y. Chang, L.A. Trinh, J.E. Padilla, S.E. Fraser, and N.A. Pierce. Programmable in situ amplification for multiplexed imaging of mrna expression,. *Nat Biotechnol*, 28(11):1208–1212, Nov 2010.

[38] E.H. Davidson and D.H. Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796, 2006.