Automated Characterization of Gene Expression Patterns with an Atlas of the Mouse Brain

J. P. Carson[1,2], T. Ju[3], C. Thaller[2], J. Warren[3], M. Bello[4], I. Kakadiaris[4], W. Chiu[1,2], G. Eichele[5]
[1]Graduate Program in Structural and Computational Biology and Molecular Biophysics, National Center for Macromolecular Imaging, [2]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX, USA
[3]Department of Computer Science, Rice University, Houston, TX, USA
[4]Visual Computing Lab, Department of Computer Science, University of Houston, TX, USA
[5]Max Planck Institute of Experimental Endocrinology, Hannover, Germany

Abstract—**A spatio-temporal map of gene activity in the brain would be an important contribution to the understanding of brain development, disease, and function. Such a resource is now possible using high-throughput *in situ* hybridization, a method for transcriptome-wide acquisition of cellular resolution gene expression patterns in serial tissue sections. However, querying an enormous quantity of image data requires computational methods for describing and organizing gene expression patterns in a consistent manner. In addressing this, we have developed procedures for automated annotation of gene expression patterns in the postnatal mouse brain.**

Keywords—**Brain atlas, gene expression, annotation, subdivision mesh**

I. Introduction

Understanding the functions of gene products is significantly aided by determining where and when genes express [1]. *In situ* hybridization (ISH) with non-radioactive probes is a histological method for marking the cells that express a particular gene. Expression distribution at the level of individual cells is important to develop an understanding of the role of genes in controlling cell identity and cell differentiation. Recently developed high-throughput (HT) ISH equipment reliably and rapidly performs ISH, thus allowing the determination of thousands of gene expression patterns in serially sectioned tissues [2]. By annotating the gene expression patterns in the resulting images, the ISH data can be much more readily queried and compared [3]. To accomplish this, we have combined a cell-based gene expression signal detection technique with a surface representation algorithm. This union of methods led to the development of a simple procedure for the rapid annotation of expression patterns.

II. Methodology

A. Data Production

Postnatal day 7 (P7) mouse brains were serially sliced to create sets of 20μm-thick sagittal tissue cryosections. Next, a Tecan Genesis robot platform carried out ISH on the brain sections. In this step, cellular mRNAs were hybridized with digoxygenin (DIG)-tagged riboprobes that in turn were detected by a sensitivity enhancing amplification reaction [4]. This resulted in distinctly localized blue/purple dye precipitates in the cells containing the mRNA transcript of interest (Fig. 1). Lastly, a Leica bright field microscope equipped with a motorized stage digitally imaged these sections at 50x magnification resulting 24-bit RGB images with a resolution of 3.3 microns per pixel (Fig. 1) [2].

B. Signal Detection

Estimating the quantity of dye precipitate in each cell throughout every image is a necessary step in determining the relative amount and distribution of mRNA transcripts. We developed *celldetekt* for this task. This program applied a sliding-window technique to determine the size and location of precipitate clusters, and produced color-coded maps representing the expression strengths of cells (Fig. 2) [5]. Cellular expression strength categories were set as strong (cell body filled with precipitate), moderate (about half filled with precipitate), weak (small amounts of precipitate), and no precipitate detected. *Celldetekt* isolated expression signal using an intensity threshold at 100 green [6]. Then, a 3x3 pixel window traversed the entire image and marked locations where signal completely occupied the window as strongly expressing cells. The average area of a cell body is 10μm in diameter, which is equivalent to the 3x3 pixel window size at the image resolution. The process repeated for moderately (2x2 window) and weakly (1x1 window) expressing cells. Non-expressing cells were detected by the same protocol using an intensity threshold of 240 gray.

This work was supported by the Burroughs Welcome Fund, the National Center for Research Resources (P41RR02250), the National Institute of Neurological Disorders and Stroke, and training fellowships from the Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (NLM Grant No. 5T15LM07093).

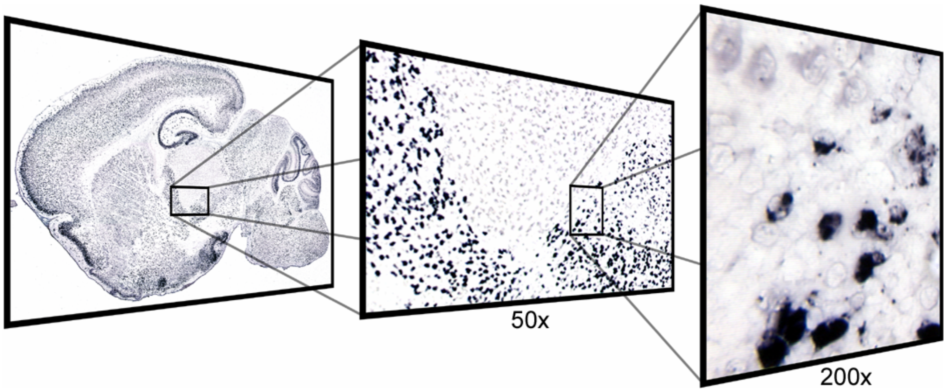

Fig. 1. Gene expression display using *in situ* hybridization. ISH marks the cells expressing a probed gene of interest. More dye precipitate (dark blue) indicates greater amounts of the specified mRNA in the cell.

C. Segmentation

The anatomy of the mammalian brain consists of major functional structures. An accurate segmentation of different structures is a prerequisite for the appropriate assignment of gene expression patterns to each structure. We defined a set of standard sagittal brain sections based upon the 11 sections illustrated in Valverde's atlas of the mouse brain [7] and marked the locations and boundaries of major brain structures (e.g., cerebellum, cortex, thalamus). Then we developed a method based on subdivision mesh technology to apply these structure boundaries to tissue sections from different brains. A subdivision mesh is a coarse mesh of quadrilaterals with a set of subdivision rules for generating smooth, increasingly fine quadrilateral meshes [8]. We created a set of meshes for the standard brain tissue sections and associated each quadrilateral in the coarse mesh with a specific anatomical structure. The fine mesh thus accurately represented the layout of anatomical structures in the standard tissue section [9]. Next, standard anatomical meshes were fit to the HT-ISH images collected. We used cross-correlation to detect key anatomical features in the set of images for each gene in order to match images with the correct mesh. The atlas meshes of the standard sections were then deformed to overlay the anatomical boundaries of the chosen ISH sections. (Fig 3.)

Automated fitting methods developed include an affine fit using principle component analysis to detect the primary axis of the brain, local fit of exterior boundaries to the tissue boundary using iterated least squares [9], and anatomical landmark detection for fitting interior boundaries [10]. After automated fitting, the mesh layout was visually verified to ensure proper boundary placement. Then each quadrilateral in the finely subdivided mesh was linked to the cellular expression strength quantities in the directly underlying tissue. Through these steps, the calculated cellular expression strengths were organized into a common subdivision-based atlas of the brain.

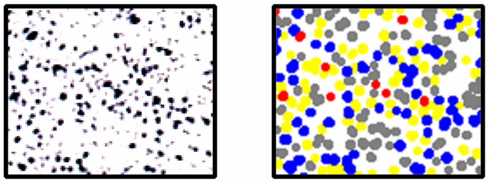

Fig. 2. Automated cell signal strength classification marks the different gene expression levels by color: red (strong), blue (moderate), weak (yellow), and gray (none)

D. Expression Pattern Annotation

ISH patterns in major anatomical structures are customarily annotated in a textual manner by classifying the distribution of gene expression as either ubiquitous (U), scattered (S), regional (R) or not detected (ND) (as in [11]). With the cellular expression strengths attached to the quadrilaterals in the subdivision mesh atlas, the total percentage of cells expressing (TPCE) in each structure was calculated for each gene. In addition, the scaled weighted deviation (SWD) in the percentage of cells expressing the gene across the quadrilaterals of the structure was also calculated. The weighting was by the number of cells in each quadrilateral and the scaling by the total percentage of expression in the structure. For our automated annotation method, ND was assigned to structures where the TPCE was less than 1%, and U assigned to structures with a TPCE greater than 50%. When TPCE was between 1%-50%, R was assigned for SWD greater than 0.7, and S for less than 0.7. These values were chosen based on patterns in the cortex for 30 genes. In cases where a particular structure exhibited multiple specific patterns for different expression signal strengths, the annotation for the strongest signal was assigned.

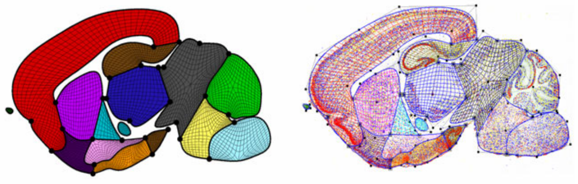

Fig. 3. A subdivision mesh rrespresentation of a brain tissue section is used to define boundaries of the anatomical structures.

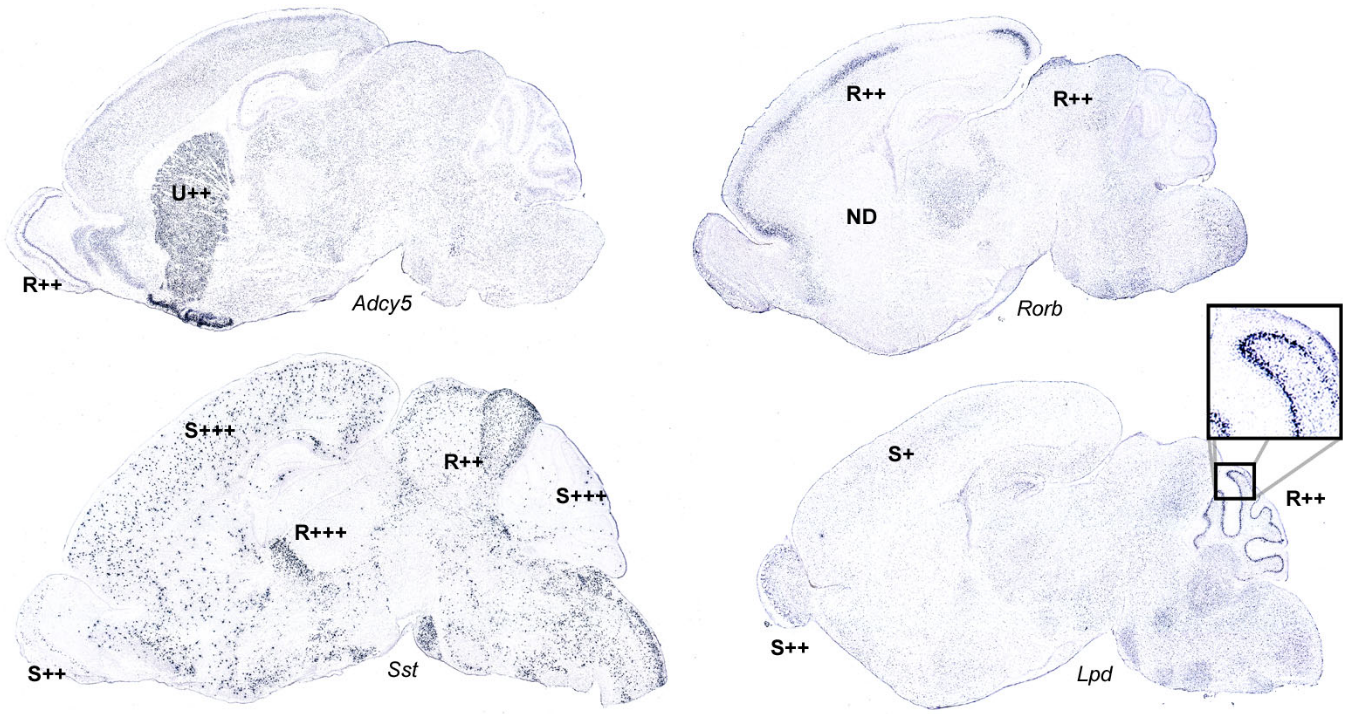

Fig. 4. Illustrations of automated annotation results for four different genes. Examples are shown for *adenylate cyclase 5* (*Adcy5*), *RAR-related orphan receptor beta (Rorb)*, *somatostatin (Sst),* and *lipidosin (Lpd).* Expression strength of patterns are either strong (+++), moderate (++), or weak (+).

III. Results

The accuracy of *celldetekt* was evaluated by comparing its signal strength classification for 255 cells at 50x magnification with classifications performed visually by an expert at 200x resolution (see Table I). 85% correctly matched, with the remainder placed into adjacent categories. Importantly, there was a low incidence of false positives of gene expression. These results suggest that automated cellular gene expression signal strength classification can be performed at 3.3µm/pixel resolution.

Examples of annotation results are shown in Fig. 4. The validity of automated pattern annotation was assessed through a comparison with visual pattern assessment of 178 patterns across different structures and genes (see Table II). Automated annotation matched visual annotation 90% of the time, with the majority of the discrepancies occurring when the automated method assigned a scattered pattern. All instances of structures visually identified as expressing were detected by the automated method as a pattern.

IV. Discussion

Each of the two computational steps has the potential for error. The sources of these errors are inherent to the data itself. The tissue sections can contain dust or air bubbles that produce a dark signal, which is then detected as gene expression. It may be possible to develop a method to digitally identify and remove such artifacts prior to precipitate detection. The other source of error is due to inherent differences between anatomical structures. The thresholds used for categorizing patterns based on SWD and

Table I
Cellular Gene Expression Signal Classfication

Visually Observed	Automated Classification			
	None	Weak	Moderate	Strong
None	61	1	0	0
Weak	15	51	7	0
Moderate	0	18	48	2
Strong	0	0	15	57

Table II
Gene Expression PATTERN Classfication

Visually Observed	Automated Classification			
	None	Scattered	Regional	Ubiquitous
None	35	3	1	0
Scattered	0	41	2	0
Regional	0	4	56	2
Ubiquitous	0	6	0	28

TPCE were set using only one structure, the cortex. However, other structures of the brain have different shapes and layouts of substructures. It may be possible to address this by using different threshold values for each structure.

Expression characterization is most significant when it is reproducible. However, visual annotation is prone to human error and subjectivity. Automated annotation is an objective process. More importantly, automated methods take advantage of modern computational power to accomplish the task rapidly. This is especially beneficial when dealing with large data quantities as in the case of genome-wide analysis of cellular gene expression.

Automated annotation would also allow one to develop better characterizations of patterns beyond the traditional categories. Whereas many cases of expression patterns are clearly ubiquitous, scattered, or regional, often the appropriate pattern category is not clear. Reporting instead the numerical values of SWD and TPCE would provide additional information, allowing one to know exactly how much expression there is and how evenly it is distributed across a structure.

In a common spatial context, comparison of expression patterns should provide a mechanism for identifying candidate genes involved in specific biological or pathological processes. We have begun the process of creating a platform for organizing and mining gene expression patterns. This prototype resource is currently accessible at *http://www.geneatlas.org*. By providing web access to the data and powerful search tools, users can query data with a minimum investment in time and effort. Expanding the dataset to include thousands of annotated gene expression patterns should aid in the identification of genes that express in similar fashions, e.g. synexpression groups [12], and in uncovering general rules of spatial expression regulation.

V. Conclusion

Placing gene expression patterns into a common spatial framework and generating unbiased characterizations of patterns will greatly facilitate the mining of gene activity data for biologically important information. The combination of HT-ISH data collection, automated cellular expression signal detection, and subdivision mesh-based segmentation enables the automated annotation of gene expression patterns. ISH relies on RNA probes that can be readily designed for all known transcripts of the mouse genome, and thus it is possible to generate in the near future a transcriptome-wide digital gene expression pattern atlas of the mouse brain using the techniques described here.

Acknowledgment

J. P. Carson thanks G. Alvarez-Bolado and A. Liang for their assistance.

References

[1] S. Gong, C. Zheng, M. L. Doughty, K. Losos, N. Didkovsky, U. B. Schambra, N. J. Nowak, A. Joyner, G. Leblanc, M. E. Hatten, and N. Heintz, "A gene expression atlas of the central nervous system based on bacterial artificial chromosomes," *Nature*, vol. 425, pp. 917-25, 2003.

[2] J. P. Carson, C. Thaller, and G. Eichele, "A transcriptome atlas of the mouse brain at cellular resolution," *Curr Opin Neurobiol*, vol. 12, pp. 562-5, 2002.

[3] D. C. Van Essen, "Windows on the brain: the emerging role of atlases and databases in neuroscience," *Curr Opin Neurobiol*, vol. 12, pp. 574-9, 2002.

[4] H. M. Kerstens, P. J. Poddighe, and A. G. Hanselaar, "A novel in situ hybridization signal amplification method based on the deposition of biotinylated tyramine," *J Histochem Cytochem*, vol. 43, pp. 347-52., 1995.

[5] J. P. Carson, "Quantitative annotation and analysis of gene expression patterns with an atlas of the mouse brain," Ph.D. dissertation, Structural and Computational Biology and Molecular Biophysics program. Baylor College of Medicine, Houston, 2004.

[6] J. C. Russ, *The image processing handbook*, 4th ed. Boca Raton, Fla.: CRC Press, 2002.

[7] F. Valverde, *Golgi atlas of the postnatal mouse brain*. Wien ; New York: Springer-Verlag, 1998.

[8] J. Warren and H. Weimer, *Subdivision methods for geometric design : a constructive approach*. San Francisco: Morgan Kaufmann, 2002.

[9] T. Ju, J. Warren, G. Eichele, C. Thaller, W. Chiu, and J. Carson, "A geometric database for gene expression data," presented at Eurographics Symposium on Geometry Processing, Aachen, Germany, 2003.

[10] I. Kakadiaris, M. Bello, S. Arunachalam, W. Kang, T. Ju, J. Warren, J. Carson, W. Chiu, C. Thaller, and G. Eichele, "Landmark-driven, Atlas-based Segmentation of Mouse Brain Tissue Slices Containing Gene Expression Data," unpublished.

[11] O. Braissant, T. Gotoh, M. Loup, M. Mori, and C. Bachmann, "L-arginine uptake, the citrulline-NO cycle and arginase II in the rat brain: an in situ hybridization study," *Brain Res Mol Brain Res*, vol. 70, pp. 231-41, 1999.

[12] C. Niehrs and N. Pollet, "Synexpression groups in eukaryotes," *Nature*, vol. 402, pp. 483-7, 1999.