2013 IEEE Scientific Visualization Contest Winner

Observing Genomics and Phenotypical Patterns in the Developing Mouse Brain

Qihang Li
The Ohio State University

Gabriel Zachmann
University of Bremen

David Feng
Allen Institute for Brain Science

Kun Huang
The Ohio State University

Raghu Machiraju
The Ohio State University

Developmental neuroscience records and analyzes changes in the developing brain from embryogenesis to the end of life. Mammalian brain development provides useful and valuable information for neuroscience in the form of patterns of structural and functional changes. These patterns, in turn, serve as essential information to characterize brain development. So, it's pivotal and critical to observe, identify, visualize, and analyze the inherent patterns of change that can be used to infer genomic regulation. A systemic, comprehensive understanding of regulation during brain development will lead neuroscientists to biologically relevant insights and treatment strategies for neurological disorders and diseases.

Several studies have demonstrated that the underlying genomics have the most impact on mouse brain growth.[1,2] However, few researchers have studied spatiotemporal gene expression (messenger RNA) patterns instead of either spatial or temporal patterns in traditional developmental neuroscience. Understanding the spatiotemporal patterns will help clarify various genes' roles during brain development. Fortunately, data repositories from the Allen Institute for Brain Science (AIBS) have inherently captured structures and associated changes in gene expression in both the developing and adult brain at numerous stages. It now remains to discover these spatiotemporal patterns.

The 2013 IEEE Scientific Visualization Contest targeted the exciting domain of developmental neuroscience pertaining to the mouse brain. (For more on the contest, see the sidebar.) It's essential to record and characterize the significant patterns of mouse brain organization and development and to analyze their relationships to gene expression patterns. However, to explore the evolution and regulation of such complex processes and grasp the multitude of relationships between gene expression and brain structure, neuroscientists need not only a complete and comprehensive data-processing system but also an efficient and interactive visualization approach.

In response to the contest's challenges, we created a visual-analytics (VA) system that lets users observe, identify, visualize, and analyze the spatiotemporal patterns of gene expression during mouse brain development. The system employs specialized data-driven representations that capture emerging structure–function associations in such a comprehensive developmental process.

0272-1716/14/$31.00 © 2014 IEEE

The 2013 IEEE Scientific Visualization Contest

Gabriel Zachmann, *University of Bremen*
David Feng, *Allen Institute for Brain Science*

The annual IEEE Scientific Visualization Contest identifies challenging scientific problems with complex datasets and presents them to the visualization community to foster innovative solutions. Each contest addresses a different scientific domain. Contestants are provided with the dataset in an easily consumable format, along with documentation and starter scripts to jump-start their work. All the data from current and past contests are in a widely used repository available at http://sciviscontest.ieeevis.org.

The Contest Problem

The 2013 contest theme was developmental neurobiology. Using the publicly available Allen Developing Mouse Brain Atlas, the contestants visualized the expression levels for approximately 2,000 genes in a 3D mouse brain at six development stages (three embryonic and three postnatal). Each stage has its own reference volume, for which every voxel has a structure label. The expression volumes derive from in situ hybridization brain section images that have been registered into their appropriate reference volume and segmented for expressing cell signals. The structural annotations come from a hierarchical ontology that deepens over time.

The contestants had to address questions related to spatiotemporal gene expression pattern detection. Their visualization system should help identify genetic markers for particular structures during one or more stages, genes with interesting spatiotemporal gradients, and other features.

Evaluation

Contestants had to submit a two-page document describing their visualization and analysis, supplemented by a video of up to 10 minutes. Four scientists from the Allen Institute for Brain Science evaluated the submissions on how well they addressed the visualization tasks, the visualization quality, and the interaction quality.

The Winner

The winning submission was "Observing Genomics and Phenotypical Patterns in the Developing Mouse Brain," by Qihang Li, Kun Huang, and Raghu Machiraju from the Ohio State University. For details on this submission, see the main article.

Acknowledgments

The Allen Institute for Brain Science produced the Allen Developing Mouse Brain Atlas, which is available at http://developingmouse.brain-map.org. Thanks to Russell Taylor and Amit Chourasia for their continuing support of this contest.

Motivation and Tasks

The AIBS has organized several studies that captured and curated a large number of genes in the developing brain and has made this data publicly available. Researchers can now characterize various relationships between gene expression and structural patterns over space and time. Here, we describe the dataset used in the contest—the Allen Developing Mouse Brain Atlas (ADMBA; http://developingmouse.brain-map.org/static/atlas)—and the contest's tasks and challenges.

Data Collection

The ADMBA provides the gene expression levels of 2,105 genes in 2,691 anatomical structures across six stages:

- 13.5 embryonic days (stage E13.5),
- 15.5 embryonic days (E15.5),
- 18.5 embryonic days (E18.5),
- 4 postnatal days (P4),
- 14 postnatal days (P14), and
- 56 postnatal days (P56).

For each stage, a 3D reference mouse brain map and the corresponding annotated volume with structure labels are available. Figure 1 shows example slides in the sagittal plane from the ADMBA. Additionally, a hierarchical annotation of various evolving brain structures is available. Finally, the scrutinized genes are organized into 11 categories, allowing for further appropriate enrichment studies.

Tasks

On the basis of the AIBS-provided data, the contest posed four tasks:

- *Gradient identification*. Which genes exhibit directional expression patterns? Which categories do these genes belong to?
- *Structural patterns*. Which genes show strong expression in a small set of structures but little expression elsewhere? How do these patterns change throughout development?
- *Structural consistency*. Which structures have the most consistent expression patterns over time? Which structures are the least consistent?
- *Complementary patterns*. Which genes have expression patterns that complement each other in a structure? Are these patterns persistent during development?

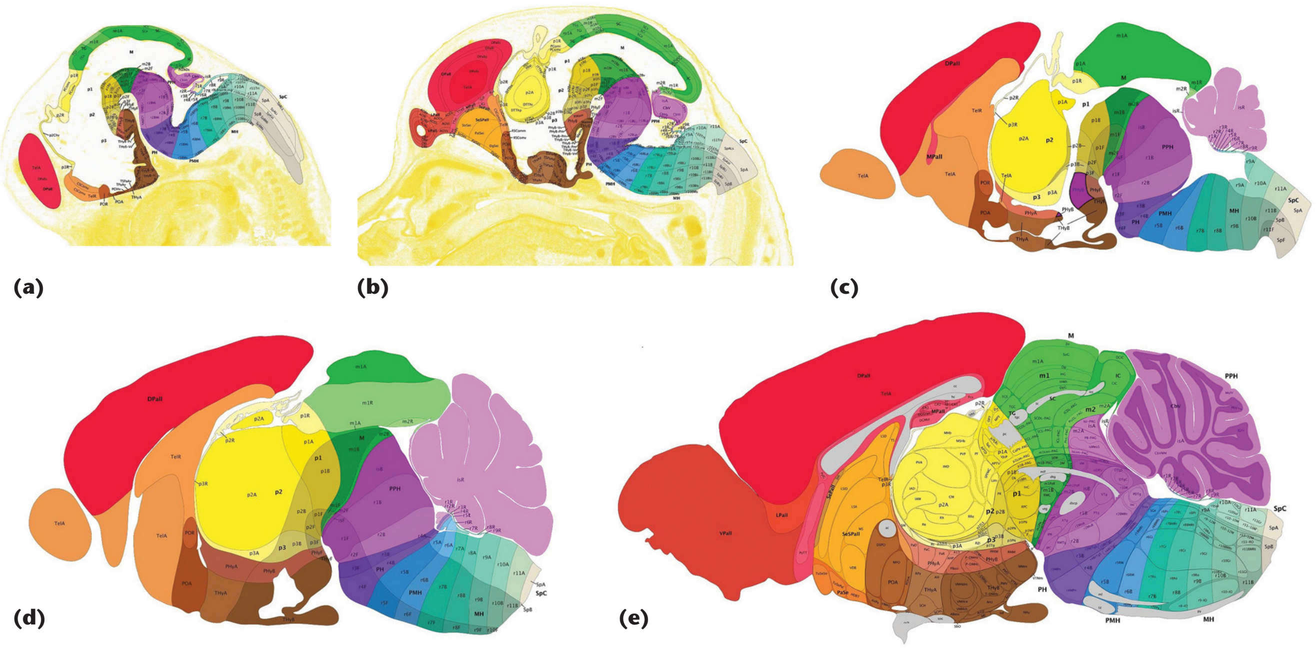

Figure 1. Slides in the sagittal plane from the Allen Developing Mouse Brain Atlas, for (a) 13.5 embryonic days (stage E13.5), (b) 15.5 embryonic days (E15.5), (c) 4 postnatal days (P4), (d) 14 postnatal days (P14), and (e) 56 postnatal days (P56).

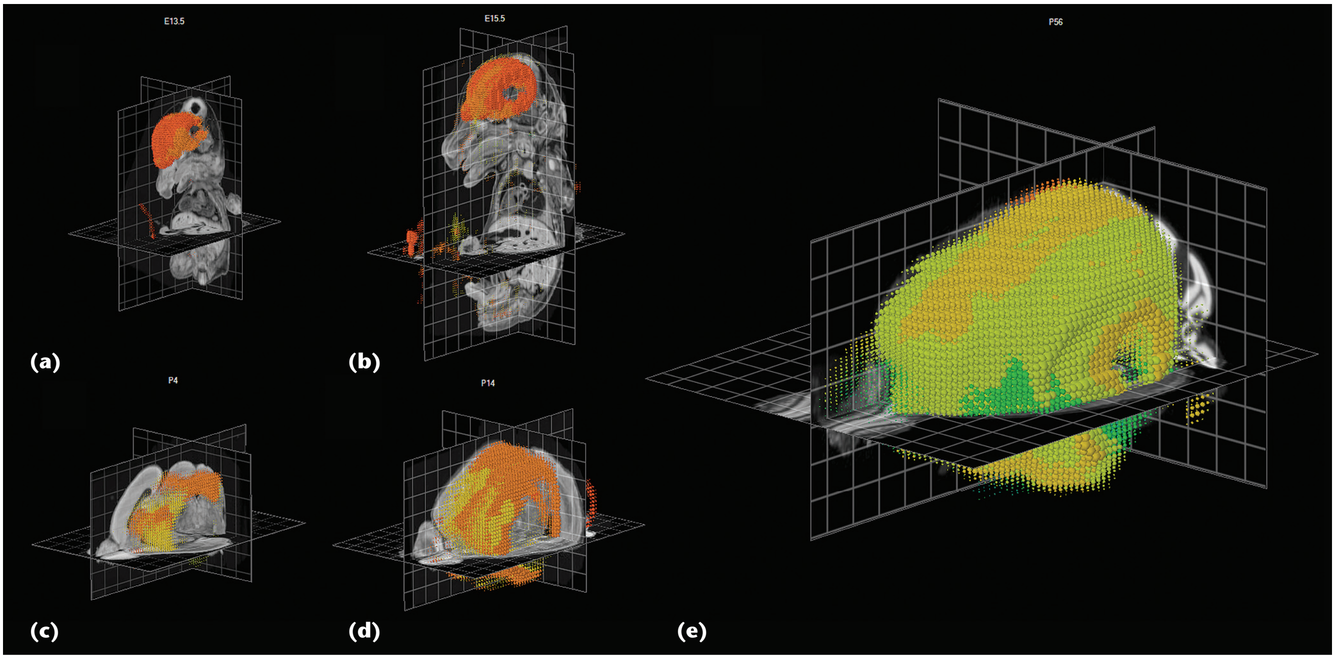

Figure 2. Brain Explorer 2 visualizations of the expression of gene TBR1, at (a) E13.5, (b) E15.5, (c) P4, (d) P14, and (e) P56 with more details. The visualizations employ a heat map in which the highest expression values are red and the lowest are green.

The characterization of such complex spatiotemporal patterns requires not only an efficient VA approach to represent the entire developmental process but also a flexible interrogative method to enable effective interactive queries. The Allen Institute's Brain Explorer 2 (http://developingmouse.brain-map.org/static/brainexplorer) is a visualization tool for exploring expression levels and marking their 3D locations in the mouse brain at a given development stage. Figure 2 shows Brain Explorer visualizations based on the gene expression data of TBR1.

However, Brain Explorer is too limited to observe spatiotemporal patterns hidden in the data or to complete any of the four tasks. These tasks require exhuming significant patterns in various phenotypes from the data. To achieve this, we developed our system.

A Visual-Analytics Solution

Our system integrates two VA representations: the *gene expression flow matrix* (GEFM) and *hierarchical orientation structural tree* (HOS-tree). The GEFM uses a data-mining algorithm to find the associations between gene expression patterns and structures. The HOS-tree organizes the brain anatomy as an oriented tree and allows for intuitive visualization of the development process. The associations found in the GEFM drive the visualization of the patterns during development and hence the HOS-tree. Furthermore, users can trace a collection of gene expression patterns back to various anatomical structures. Figure 3 gives a concise overview of our approach.

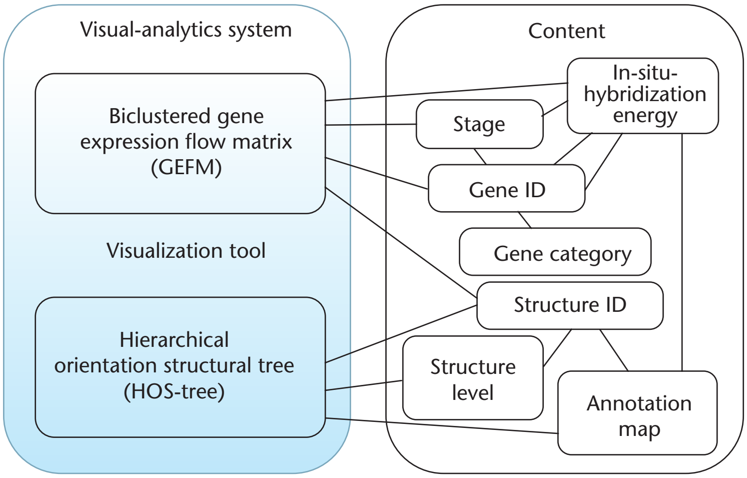
Visual-analytics system
Content
Biclustered gene expression flow matrix (GEFM)
Visualization tool
Hierarchical orientation structural tree (HOS-tree)
In-situ-hybridization energy
Stage
Gene ID
Gene category
Structure ID
Structure level
Annotation map

Figure 3. An overview of our visual-analytics system, which helps researchers find associations between gene expression and mouse brain structures.

The GEFM

To encode the structural patterns' variability over the six stages, the GEFM uses a 2D matrix to represent the associations between gene expression and structures. Five flow models represent the likely trends of gene expression—from a rapidly increasing profile, to a constant profile, to a rapidly decreasing profile (see Figure 4a). In the GEFM, each model has been assigned a different color. Figure 4b shows the GEFM for all 2,105 genes (rows) and 2,691 structures (columns).

Subsequently, we used *k*-means biclustering to search for correlated subsets in the original GEFM.[3] Figure 4c shows the biclustered GEFM using the Java Treeview tool.[4] Compared with the original GEFM, the biclustered GEFM lets users detect correlated sections more clearly by inspecting the predominant color in each section.

The HOS-Tree

The HOS-tree imposes a hierarchical structure over the developmental stages and is well suited for observing structural patterns' evolution. Nodes represent structures; an edge between two nodes implies that they share a common lineage during development.

Figure 5a shows annotated structures for the first four expression levels; Figure 5b is the basic tree derived from this annotation.

To reveal the emerging directionality during brain development, we use the RGB color system to delineate the developmental orientation from preceding to succeeding structures. We define the developmental orientation as the normalized vector between neighboring structures' spatial centers. Each primary color's intensity denotes the measurement on each color axis as determined by the projection of the developmental-orientation vector (see Figure 5c). Red is along the *x*-axis (medial–lateral), green is along the *y*-axis (ventral–dorsal), and blue is along the *z*-axis (caudal–rostral). For unannotated structures, we estimate colors on the basis of the successor nodes. We borrowed this coloring scheme from the literature on diffusion tensor imaging.[5]

In the HOS-tree (see Figure 5d), solid lines link structural nodes attributed with expression levels, whereas dashed lines link nodes with missing measurements of expression levels. Furthermore, we distribute nodes along concentric rings where the radius indicates the hierarchy level. So, the HOS-tree shows not only the patterns of structural development that occurs over expression levels but also the significant directional trends of spatial development.

Interactive Visualization

Our system has two major functions: *global section exploration* (GSE) and *gene profile exploration* (GPE). GSE lets users explore the potential global spatial gene expression gradients hidden in the data. It unearths the gene–structure correlated sections in the biclustered GEFM. Once the user selects a section of interest, the system highlights the corresponding structure nodes (using dark circles) in the HOS-tree. Because each clustered section in the biclustered GEFM represents a highly correlated gene–structure subset, users can observe a spatial gene expression gradient pertaining to those genes and structures across all stages.

GPE is more useful for learning developmental patterns. It focuses on a single gene profile and displays the corresponding pattern at any stage. After the user selects a gene of interest, the system indicates the measured expression level by the width of the corresponding edges between structural nodes in the HOS-tree. Because the edges' colors indicate the developmental orientation and

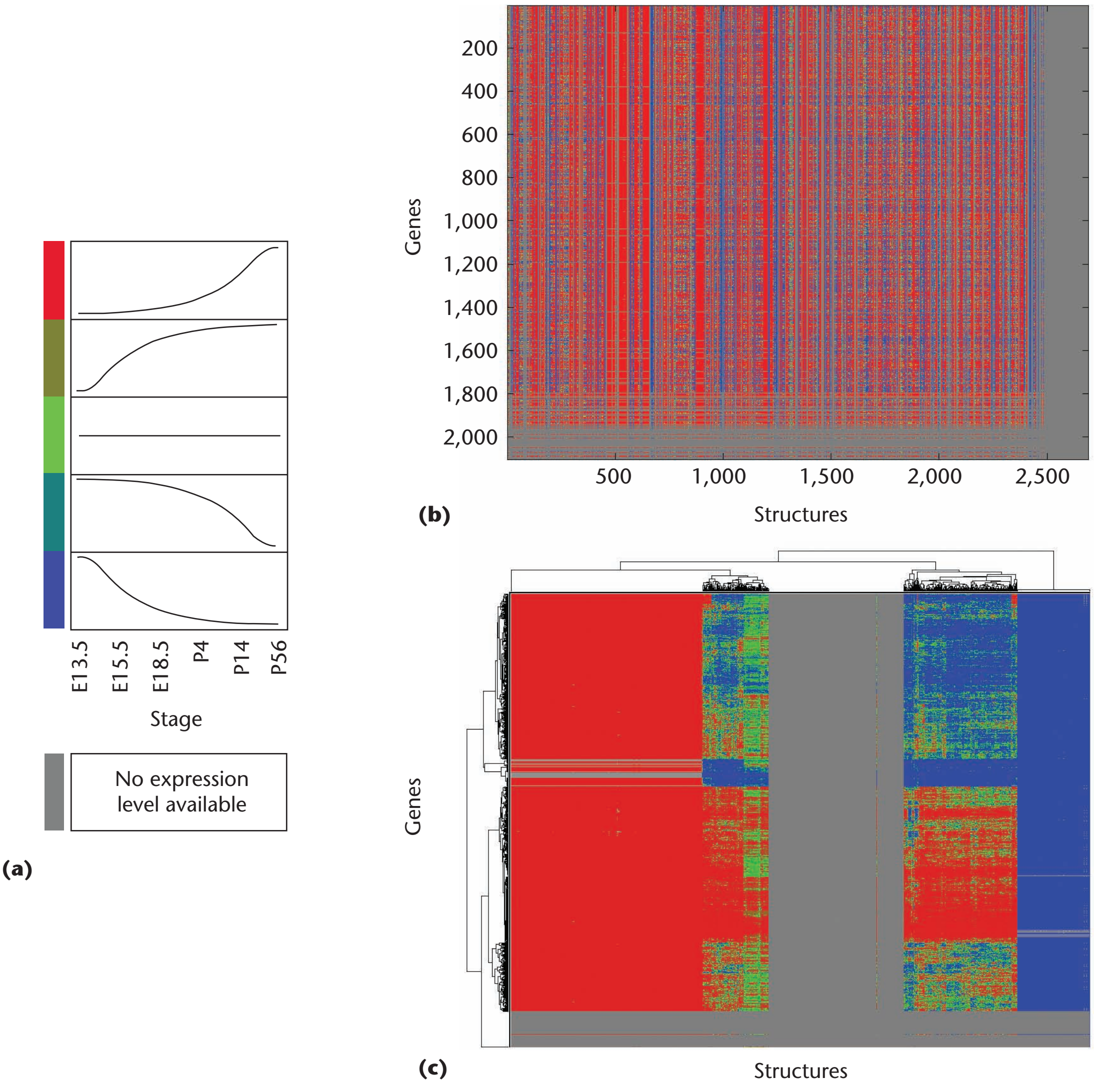

Figure 4. The GEFM. (a) We used five flow models with corresponding colors for the expression levels. (b) We employed the Pearson correlation coefficient to mark the temporal expression flow of each gene (rows) in each structure (columns). (c) We applied *k*-means biclustering on the derived GEFM to delineate tightly correlated subsets.

their width measures gene expression levels, users can also inspect the directional expression patterns from the HOS-tree. Moreover, they can compare six directional expression patterns (one for each stage) to find the expression gradient changes throughout the brain's development.

The User Interface

Our system's interface facilitates easy inspection and analysis of the spatial and temporal patterns of global sections and single-gene profiles. It has three areas:

- the preprocessed biclustered GEFM, for GSE (see Figure 6a);
- a list of genes and a slider for stage selection, for GPE (see Figure 6b); and
- the HOS-tree (see Figure 6c).

Our system implements these areas according to the rules we described in the section "Interactive Visualization." Thus, our system offers a straightforward, interactive way to observe, explore, and analyze the expression levels, the expression levels' gradients, and the structural orientation.

Case Studies

To demonstrate our system's robustness and effectiveness, we analyzed several gene profiles and correlated sections identified in the biclustered GEFM. Here we describe our solutions for the contest's four tasks.

Gradient Identification

This case study compared spatial expression patterns for gene TBR1 at P14 and P56. TBR1 is an important transcription factor in vertebrate embryonic development and is critical for neuron differentiation and migration in brain development. Furthermore, it's considered extremely important for regulating human cortex development.

Figure 7a shows TBR1's spatiotemporal profile at P14. Many of the wider edges are green or red. This pattern provides strong evidence that the expression energy is much stronger along the *x*- and *y*-axes

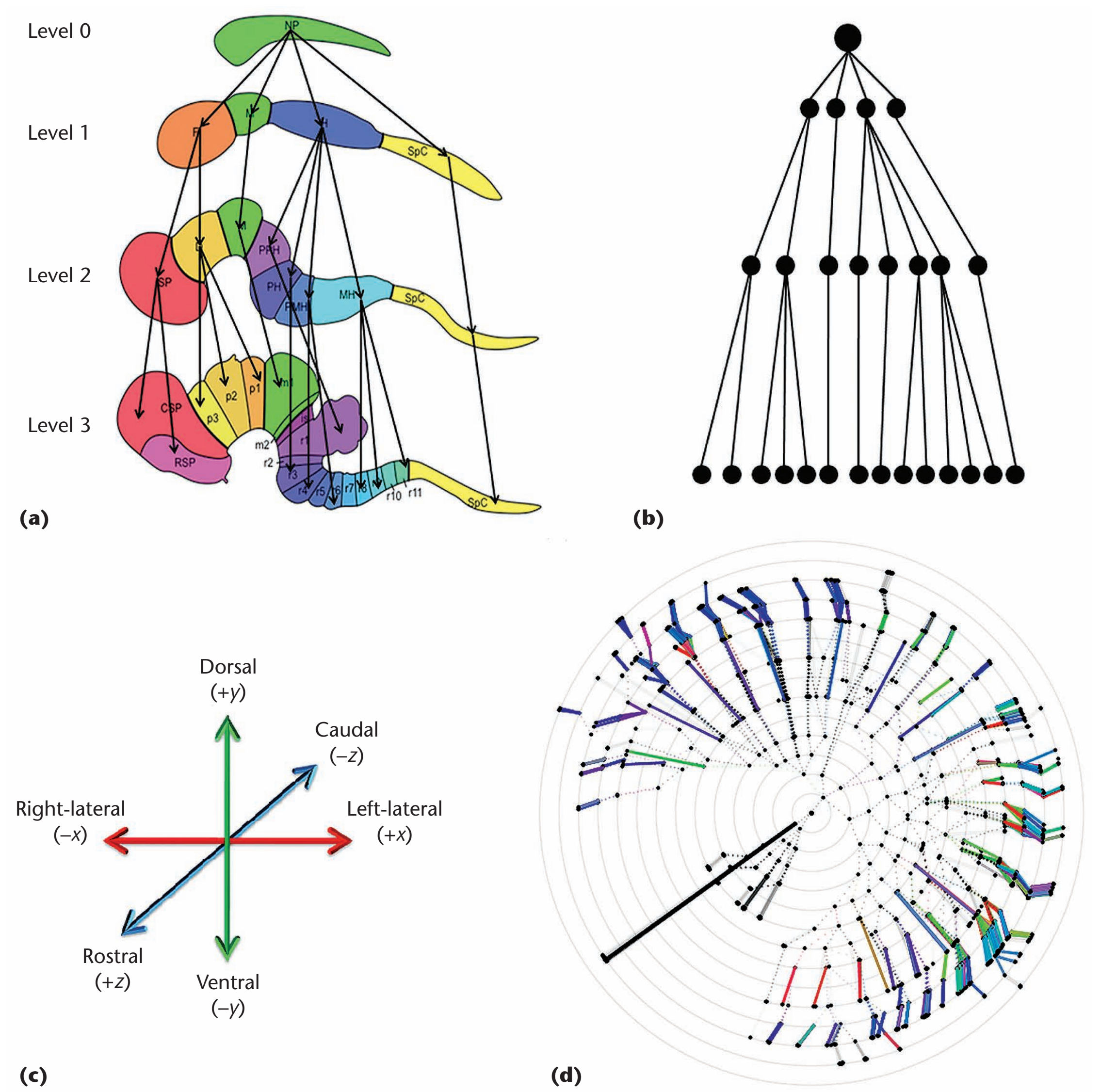

Figure 5. The HOS-tree. (a) Annotated structures for the first four expression levels. (b) The hierarchical tree derived from Figure 5a. (c) The RGB coordinate space we use to indicate direction in the HOS-tree. (d) The HOS-tree for all 2,691 structures.

than the *z*-axis. Because the *x*- and *y*-axes estimate horizontal and vertical spatial gradients, the spatial expression pattern closely matches often-observed major cortical development. However, at P56 (see Figure 7b), the expression energy is much weaker. So, we believe TBR1 is functionally expressed before P56.

Therefore, we believe our system can robustly and efficiently identify gene expression gradients. Here's our solution to this task:

> By using GPE to analyze the gene profile, we can detect both spatial and temporal gene expression patterns by examining the HOS-tree edges' width and color.

Structural Patterns

This case study used GSE to analyze a section of the biclustered GEFM and then used GPE to examine a specific gene profile.

We first chose a section in the biclustered GEFM that was mostly blue (see the correlated gene–structure section X in Figure 8a, bordered in yellow). We asserted that the genes in this section were expressed at a high energy level only at the end of the early postnatal stage. The corresponding structures highlighted in the HOS-tree (see Figure 8b) include the anterior olfactory region, terminal hypothalamus, hippocampus, corpus callosum, commissural pretectal domain, isthmus, and pallial septum. We also observed that these structures occurred mostly at a relatively late stage, between hierarchy levels 9 and 10.

So, we claim that the genes in section X are expressed strongly in specifically delineated structures at later development stages. More important, as part of the essential development in the early postnatal stage, synapses begin to form pathways and eventually connect the brain's different parts.[6] Therefore, we believe these correlated structures are primary regions where synapses develop at this stage.

Next, we considered gene SIM2, well known as the homolog of one of the *Drosophila* single-minded

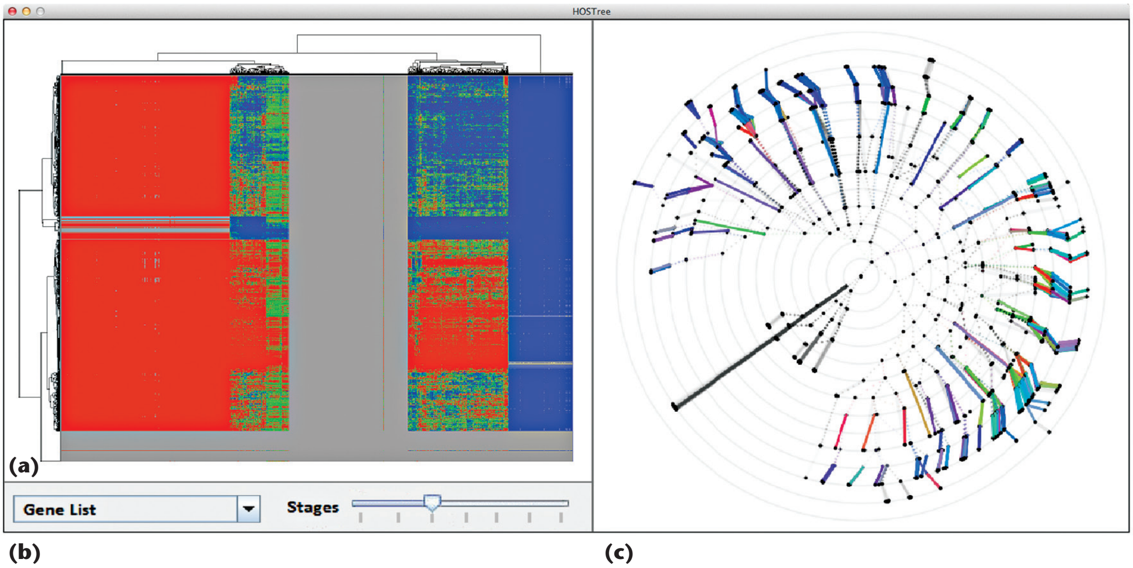

Figure 6. Our system's user interface comprises (a) the global section exploration area, showing the preprocessed biclustered GEFM; (b) the gene profile exploration area, containing the gene list and a slider for stage selection; and (c) the HOS-tree.

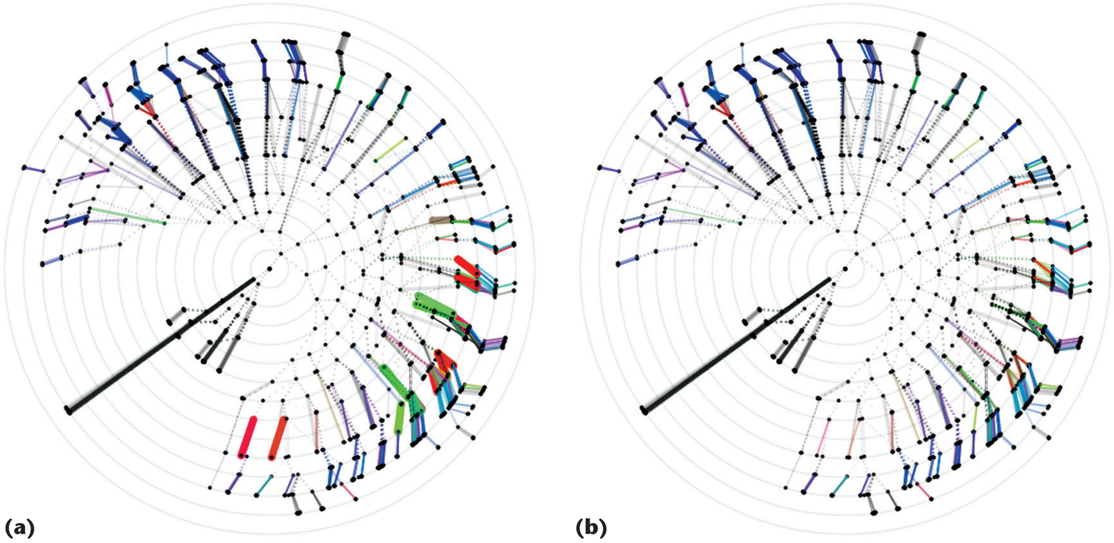

Figure 7. A spatiotemporal profile of TBR1. (a) At P14, the wider edges are mostly green and red, indicating gradients along the *x*- and *y*-axes. (b) At P56, the expression energy is much weaker, suggesting that TBR1 is functionally expressed before P56.

(SIM) genes. It expresses preferentially in the diencephalon during early embryogenesis to mediate neuroendocrine hormone gene expression.[7,8]

In Figure 9, the nodes with the wider edges are the structures that SIM2 strongly expressed. At E13.5 (see Figure 9a), the highly expressed areas are clearly mid-range (between levels 9 and 11). At P56 (see Figure 9b), the highly expressed areas diffuse into the late range (between levels 11 and 13).

Here's our solution to this task:

> By using GSE to focus on the color representation in the biclustered GEFM, we can detect significant expression–structure patterns by examining the highlighted nodes. By further using GPE to analyze variations in patterns across stages, we can detect hierarchical and temporal expression patterns by examining the edge widths.

Structural Consistency

This case study investigated gene expression consistency over the six stages. We define consistency as an expression's invariability over time or across stages. We can observe consistent and inconsistent gene–structure pairs by detecting their colors in

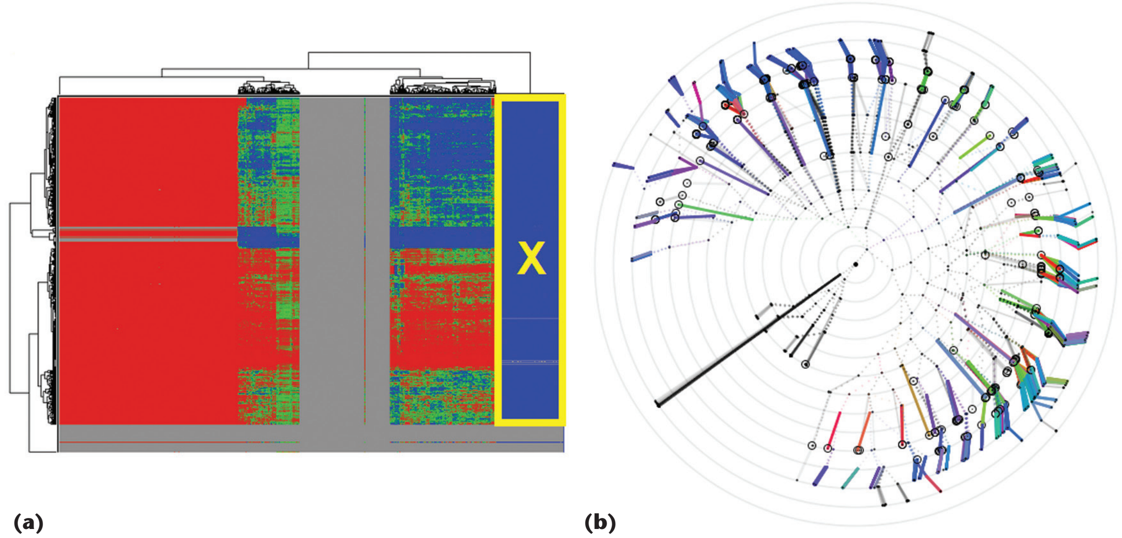

Figure 8. Analysis of the correlated gene–structure section X of the biclustered GEFM. The selected gene was expressed much more strongly at the end of the early postnatal stage. (a) Section X, bordered in yellow. (b) The corresponding structures highlighted in the HOS-tree. These structures were distributed mostly between hierarchy levels 9 and 10.

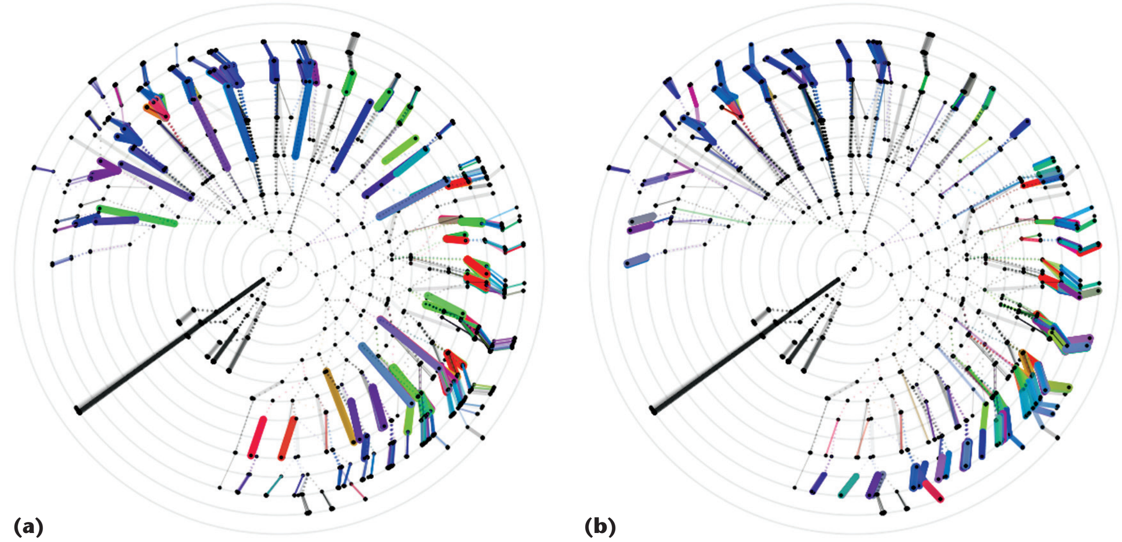

Figure 9. Gene profile exploration of gene SIM2. (a) At E13.5, the strongly expressed areas are mostly mid-range, between levels 9 and 11. (b) At P56, the strongly expressed areas diffuse into the late range, between levels 11 and 13.

the GEFM. Green indicates the least variation. So, a gene that's expressed consistently would be green in both the GEFM and biclustered GEFM.

Figure 10a shows the selected correlated gene–structure section Y (bordered in yellow), in which most genes are green. As Figure 10b shows, the highlighted structures are mostly between levels 5 and 10, and the colored edges among them are in various shades.

Because the neural system requires the longest time to develop and genes must be functionally expressed during the entire development process,[9] we hypothesize that the genes expressed in a time-invariant fashion (the genes in section Y in Figure 10a) are functionally related to neural development. To confirm our hypothesis, on the basis of the expressed genes' functional categories, we created a histogram of the functional roles of the genes indicated in green. As Figure 10c shows, the top three categories of the genes in this cluster are associated with the transcription factor, nervous-system development, and neurological-system development. Because the result verifies our hypothesis, we believe those time-invariant expressed genes are highly correlated to neural system development.

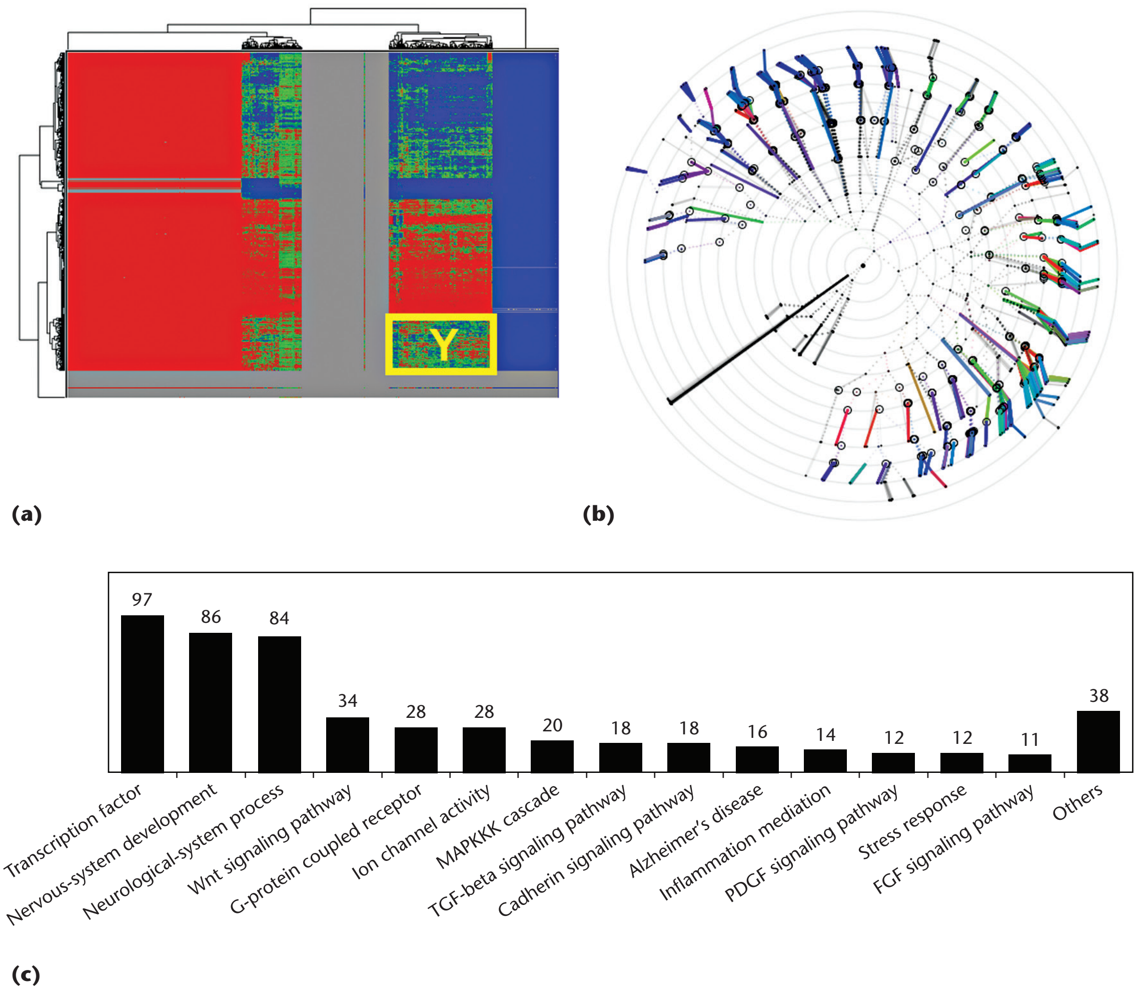

Figure 10. Analysis of the correlated gene–structure section Y of the biclustered GEFM, focusing on genes expressed consistently over all six stages. (a) Section Y, bordered in yellow. (b) The corresponding structures highlighted in the HOS-tree. These structures are distributed mostly between levels 5 and 10, in various colors. (c) A histogram of the frequency of categories for selected genes. We claim that these genes play a functional role in neural-tissue development.

On the other hand, after inspecting the 327 correlated genes in section Y, we found that

- 27 genes contain homeobox, a protein domain essential for development-related functions;
- 30 genes are involved in the Wnt signaling pathway, which is critical to neuronal development;
- 58 genes are known to cause an abnormal somatic nervous system morphology in mice; and
- seven genes are known to be involved in human bipolar disorder.

Here, we claim that the selected genes are certainly salient for the development of the neural system, with a strong influence on organization.

Here's our solution to this task:

> The biclustered GEFM identifies the correlated gene–structure subsets and uses the five-color representation to identify each expression flow type over the six stages. In particular, the genes expressed consistently over time are green.

Complementary Patterns

Complementary genes are those expressed cooperatively in a structure. Unfortunately, exploring complementary patterns isn't feasible with our system. However, users can still observe genes that follow complementary flow patterns in various structures.

As in the structural-patterns case study, genes expressed with the same flow pattern have the same color in the GEFM. For example, section X in Figure 8a displays all the genes that were expressed strongly as the fifth pattern (in blue) in the corresponding structures.

Because we cluster the genes in every column (and hence structure), users can observe the gene subsets expressed by the same flow. So, once users select a column in the biclustered GEFM, they can observe significant gene subsets solely on the basis of the predominant color.

On the other hand, because opponent colors (red–blue and olive–sapphire) represent opposite flow shapes, the related genes are highly likely to be complementary. So, we believe that the observation of such patterns aids the exploration of complementary patterns.

Although we haven't observed salient complementary patterns, our system is well suited for exploring gene subsets described by the same flow patterns in any structure. Here's our solution to this task:

> For each structure, the corresponding column in the biclustered GEFM includes all gene expression flows. Users can note the genes (and gene subsets) described by the same flow model and hence having the same color. So, genes (and gene subsets) represented by opponent colors will likely be complementary.

From the case studies, we conclude that our system not only addresses most of the contest challenges but also provides a useful tool for exploring gene expression in a developing brain. More important, the exploration is interactive, and the observed spatiotemporal patterns will likely lead to further analysis and hypotheses.

Although our system enables observations of the spatiotemporal patterns of gene expression over the six stages, it has noticeable limitations. First, the temporal patterns are derived from predefined models. In this case some of the natural expression patterns wouldn't be observed. Another limitation is with the HOS-tree's 3D vector system. Our approach doesn't account for the different coordinate frame systems underlying the various annotation volumes over stages. So, we plan to develop a precise approach that learns the inherent and natural temporal patterns, as well as a robust method that detects the precise developmental orientations.

Acknowledgments

We thank the 2013 IEEE Scientific Visualization Contest organizers and the Allen Institute for Brain Science for providing the dataset and posing the analytic challenges. We thank the anonymous reviewers for their comments.

References

1. E.S. Lein et al., "Genome-Wide Atlas of Gene Expression in the Adult Mouse Brain," *Nature*, vol. 445, no. 7124, 2007, pp. 168–176.
2. P.A. Gray et al., "Mouse Brain Organization Revealed through Direct Genome-Scale TF Expression Analysis," *Science*, vol. 306, no. 5705, 2004, pp. 2255–2257.
3. M.J.L. de Hoon et al., "Open Source Clustering Software," *Bioinformatics*, vol. 20, no. 9, 2004, pp. 1453–1454.
4. A.J. Saldanha, "Java Treeview—Extensible Visualization of Microarray Data," *Bioinformatics*, vol. 20, no. 17, 2004, pp. 3246–3248.
5. S. Mori et al., *MRI Atlas of Human White Matter*, Elsevier, 2005.
6. N. Tian, "Visual Experience and Maturation of Retinal Synaptic Pathways," *Vision Research*, vol. 44, no. 28, 2004, pp. 3307–3316.
7. A. Yamaki et al., "Molecular Mechanisms of Human Single-Minded 2 (SIM2) Gene Expression: Identification of a Promoter Site in the SIM2 Genomic Sequence," *Gene*, vol. 270, nos. 1–2, 2001, pp. 265–275.
8. E. Goshu et al., "SIM2 Contributes to Neuroendocrine Hormone Gene Expression in the Anterior Hypothalamus," *Molecular Endocrinology*, vol. 18, no. 5, 2004, pp. 1251–1262.
9. D. Rice and J.S. Barone, "Critical Periods of Vulnerability for the Developing Nervous System: Evidence from Humans and Animal Models," *Environmental Health Perspectives*, vol. 108, supplement 3, 2000, pp. 511–533.

Qihang Li *is a doctoral candidate in the Ohio State University's Department of Computer Science and Engineering. Contact him at li.1376@osu.edu.*

Gabriel Zachmann *is a professor of computer graphics, visual computing, and virtual reality at the University of Bremen. Contact him at zach@cs.uni-bremen.de.*

David Feng *is a software engineer at the Allen Institute for Brain Science. Contact him at davidf@alleninstitute.org.*

Kun Huang *is an associate professor in the Ohio State University's Department of Biomedical Informatics. Contact him at kun.huang@osumc.edu.*

Raghu Machiraju *is a professor in the Ohio State University's Departments of Computer Science and Engineering and Biomedical Informatics. Contact him at machiraju.1@osu.edu.*

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com.