Neuroinformatics for Genome-Wide 3D Gene Expression Mapping in the Mouse Brain

Lydia Ng, Sayan D. Pathak, Chihchau Kuan, Christopher Lau, Hongwei Dong, Andrew Sodt, Chinh Dang, Brian Avants, Paul Yushkevich, James C. Gee, David Haynor, Ed Lein, Allan Jones, and Mike Hawrylycz

Abstract—Large-scale gene expression studies in the mammalian brain offer the promise of understanding the topology, networks, and, ultimately, the function of its complex anatomy, opening previously unexplored avenues in neuroscience. High-throughput methods permit genome-wide searches to discover genes that are uniquely expressed in brain circuits and regions that control behavior. Previous gene expression mapping studies in model organisms have employed in situ hybridization (ISH), a technique that uses labeled nucleic acid probes to bind to specific mRNA transcripts in tissue sections. A key requirement for this effort is the development of fast and robust algorithms for anatomically mapping and quantifying gene expression for ISH. We describe a neuroinformatics pipeline for automatically mapping expression profiles of ISH data and its use to produce the first genomic scale 3D mapping of gene expression in a mammalian brain. The pipeline is fully automated and adaptable to other organisms and tissues. Our automated study of more than 20,000 genes indicates that at least 78.8 percent are expressed at some level in the adult C56BL/6J mouse brain. In addition to providing a platform for genomic scale search, high-resolution images and visualization tools for expression analysis are available at the Allen Brain Atlas web site (http://www.brain-map.org).

Index Terms—Bioinformatics (genome or protein) databases, data mining, registration, segmentation, information visualization.

1 INTRODUCTION

Several high-throughput efforts are underway to systematically analyze gene expression patterns in the mammalian central nervous system (CNS) [1], [2], [3]. These projects strive to gain insight into temporal and spatial expression of specific genes throughout CNS development and in the adult brain. Advances in genomic sequencing methods, high-throughput technology [4], [5], [6], [7], and bioinformatics through robust image processing [8], [9] now enable the neuroscience community to study nervous system function at the genomic scale. Central to these neuro-genomics efforts is understanding gene transcription in the context of the spatial anatomy and connectivity of the nervous system [10].

Of the several techniques presently available for exploring large-scale gene expression [1], [11], [12], in situ hybridization (ISH) is one of the most compelling because of its ability to study the anatomic localization of specific gene expression in the brain [13]. Perhaps more so than other currently available biotechnologies, ISH offers the promise of cellular level specificity and microstructural identification, two major interests of modern neurobiology. However, in order to localize expression patterns in the brain, it is critical to establish a universal coordinate system and a compatible anatomic ontology for the brain, ideally with multiple hierarchical levels, and methods of data presentation that facilitate comparison of the expression patterns of multiple genes [14], [15].

The Allen Brain Atlas (ABA) informatics pipeline addresses the challenge of automating the mapping of 3D gene expression patterns on a genomic scale. The primary goal of the pipeline is to acquire and process data on the expression of individual genes in a fashion that will enable online anatomic structural search, visualization, and data mining of ISH imagery. A secondary goal is to develop tools to facilitate visual and computational discovery. In this paper, we describe the ABA informatics platform and methodology and some of the publicly available tools and illustrate the power of the method via genomic level pattern searches.

The ABA informatics automated pipeline, illustrated in Fig. 1, consists of modules supporting the following functions:

- image preprocessing, including tile stitching and direct compression into JPEG2000 format,
- image storage and indexing,
- access to a novel online digital reference atlas for the adult C56Bl/6J mouse brain [16],
- and a compatible anatomic ontology for the brain, ideally with multiple hierarchical levels, and methods of data presentation that facilitate comparison of the expression patterns of multiple genes [14], [15].

The Allen Brain Atlas (ABA) informatics pipeline addresses the challenge of automating the mapping of 3D gene expression patterns on a genomic scale. The primary goal of the pipeline is to acquire and process data on the expression of individual genes in a fashion that will enable online anatomic structural search, visualization, and data mining of ISH imagery. A secondary goal is to develop tools to facilitate visual and computational discovery. In this paper, we describe the ABA informatics platform and methodology and some of the publicly available tools and illustrate the power of the method via genomic level pattern searches.

The ABA informatics automated pipeline, illustrated in Fig. 1, consists of modules supporting the following functions:

- image preprocessing, including tile stitching and direct compression into JPEG2000 format,
- image storage and indexing,
- access to a novel online digital reference atlas for the adult C56Bl/6J mouse brain [16],
3D image reconstruction and deformable registration to bring the ISH images into a common anatomic framework,

- signal detection and estimation for segmentation of expressing cells and tissues,

- compilation of gene expression results over 3D regions and presentation in an online searchable database, and

- visualization tools for examining 3D expression patterns of multiple genes in anatomic regions.

We describe each of the major components in this paper and supply additional details in the supplemental material, which can be found on the Computer Society Digital Library at http://computer.org/tcbb/archives.htm. While the present platform is optimized for the standard laboratory C57Bl/6J adult mouse, it is readily configurable for other mammalian organisms and tissues, including temporal developmental series and other image modalities.

Several advances were made during the development of the ABA in the areas of automated annotation and signal segmentation that should be applicable to other areas of biomedical imaging.

The Allen Brain Atlas is a database containing high-resolution image data of gene expression patterns of the mouse brain that enables search on 3D anatomy over high-level anatomic structures. The result of processing over 20,000 genes can be searched and visualized at www.brainmap.org, where the data is publicly available. In addition to 2D and 3D analysis and visualization tools, we have currently released the search results for 16 anatomic regions, over low, medium, and high levels of expression coverage and signal intensity.

2 MATERIALS AND METHODS

2.1 Tissue Processing

Tissue sections (25 μm thick) were generated from 8-week-old male C57BL/6J mouse brains (see [16] for phenotype classification). Each of these sections was labeled for a specific gene using a semiautomated ISH protocol originally developed at the Max Planck Institute (Hanover, Germany) and the Baylor College of Medicine (Houston, Texas) [17]. In general terms, Digoxigenin (DIG) labeled riboprobes, specifically designed for each particular gene, are hybridized to cellular mRNA transcripts on the tissue sections. Following tyramide signal amplification (TSA) to maximize sensitivity, colorimetric detection (NBT/BCIP labeling) of the bound probe produces a vivid blue/purple label in those cells expressing a particular gene. In this way the spatial expression pattern for a specific gene can be identified. An example ISH image is shown in Fig. 2a. A detailed description of the biochemical protocol can be found in [17], [18].

To enable a comprehensive analysis of gene expression and bound material usage, each physical brain is used to survey several independent genes. Specifically, the protocol divides the tissue sections into eight interleaved series such that the sections within one series are spaced 200 μm apart. Each series is either hybridized to a single gene or used for Nissl anatomical reference staining, a type of control stain that essentially stains all neurons [19] and shown on the left hemisphere of Fig. 2b. Spatial resolution in the z-plane for each gene is therefore 200 μm.

2.2 Automated Image Acquisition

The Image Capture System (ICS) consists of 10 independent Leica DM6000B microscopes, each with a Leica DC500 camera and mounted on an air table to isolate the microscope from external sources of vibration that would affect image quality. The image capture procedure is essentially fully automated and collects each image section at 10x magnification and resolution 1.07 μm² per pixel. Each brain section generates 80-100 TIFF tiles (3-8 MB) that are stitched and compressed directly into the JPEG2000 format. The wavelet-based JPEG2000 format is used for its superior compression quality and inherent multiresolution file format. A more detailed description of the configuration is given in the supplemental materials, which can be found on the Computer Society Digital Library at http://computer.org/tcbb/archives.htm.
2.3 Automatic Anatomic Mapping

The process of 3D mapping starts with atlas-based annotation\(^1\) using a novel high-resolution anatomic atlas, the Allen Reference Atlas (ARA), which was developed specifically for this project. The ARA consists of 528 coronal 25 \(\mu\)m thick, coronal Nissl stained sections of an unfixed, frozen mouse brain. In this atlas, several hundred anatomic brain structures were manually delineated on every fourth section in one hemisphere of the brain. To facilitate efficient in silico processing, each of the annotated regions is color coded with a unique RGB color value that enables programs to recognize the anatomic region that a given pixel of the image belongs to in constant time. The ARA can be independently viewed online at www.brain-map.org.

The first step toward registration of a 2D image series with the atlas is to reconstruct an approximate 3D volume from the individual images. Reconstruction is necessary for both the atlas image series and the ISH image series. Fig. 3 shows the result of a rigid reconstruction of the reference atlas, where each section is rigidly reoriented to match the adjacent images as closely as possible, forming an assembled 3D volume. Fig. 3a is one of the original atlas sections. Figs. 3b and 3c show reconstructed sagittal and horizontal planes through the assembled 3D volume; note the slightly jagged edge due to imperfect registration.\(^2\) A 1.5T low-resolution 3D averaged MRI volume was used as a template to ensure that the reconstruction resulted in a realistic volume. The reconstruction work was performed by collaborators in the Department of Radiology at the University of Pennsylvania [20]. A corresponding dense annotation volume is constructed from the 1:4 annotations by using shape-based interpolation [21] to interpolate across the nonannotated sections. The final 3D annotated template is then used to facilitate automatic annotation of the ISH imagery.

The overall gene mapping workflow is described as follows: Each ISH image series is treated independently. After cropping, white-balancing, and other preprocessing steps are applied to each gene series, the 2D ISH series is reconstructed as a 3D volume and approximately registered to the 3D assembled Nissl volume of the reference atlas. Multiresolution registration techniques, described in more detail below, are used to iteratively refine the alignment of the ISH volume to the Nissl volume so that the actual assembly of the 3D ISH volume occurs iteratively along with the mapping process.

Registration results in a deformation field that “best” maps the template reference atlas onto the ISH volume with respect to a specific registration optimization criterion. If the registration result is adequate, the deformation field can be used to facilitate automatic annotation of the ISH imagery.

1. We use the term *annotation* to refer to the determination of structural masks representing anatomic regions and reserve the term *segmentation* to refer to the specification of expressing cells or tissue in the images. These terms are often used interchangeably.

2. Deformable registration methods are intentionally not employed in the reconstruction of the 3D reference atlas as local tissue distortions in the reference template are not desirable.

---

**Fig. 2.** (a) ISH for gene Pvalb on a coronal section showing expression in the cortex, hippocampus, and thalamus. (b) Coronal section of the Allen Reference Atlas with Nissl stain in the left hemisphere and anatomic annotation in the right.

**Fig. 3.** Three views of the 3D rigidly reconstructed reference atlas Nissl volume overlaid with matching hierarchical brain structure annotation: (a) coronal, (b) sagittal, and (c) horizontal.
be used to warp the template annotation image, thus automatically creating annotations for the ISH volume.

2.3.1 3D Reconstruction and Registration

The template Nissl volume of the ARA is registered to each ISH image series using 3D intermodality registration techniques. In many cases, the boundaries between structures are ambiguous and automatic annotation is therefore a difficult problem. The mapping problem is further complicated by the “punctate” qualities of the NBT/BCIP labeling product itself, which is regionally inhomogeneous and of varying intensity.

In the present workflow, each gene series, with sections 200 μm apart, is processed independently and, although it is possible to use the entire brain series in the reconstruction process, we have determined that the gain in accuracy was not substantial in this context. For each gene series, there are essentially two registration problems: 1) 3D volume reconstruction of the ISH series and 2) subsequent registration of the template to the reconstructed volume.

In addition to inherent brain-to-brain variation, some of the major challenges in addressing these two problems include:

1. the comparatively large 200 μm spacing between images in a series, resulting in insufficient similarity between neighboring sections to align the images without the use of external constraints,
2. the relationship between the Nissl reference atlas image intensity and ISH image intensity is complex and different for each gene,
3. distortion and artifacts occur independently on each section, and
4. differential tissue bending before freezing results in large morphological variability between brains.

Further, to achieve a comprehensive analysis of the whole mouse genome (~20k genes) necessitated full automation of the registration process. To the authors’ knowledge, no other project using ISH imagery has reported a fully automated approach to annotation that has been applied in a large-scale manner. For example, the approach in a similar project described in [22] is to fit by hand a 2D subdivision mesh model of the boundaries onto sagittal ISH images in a semiautomated manner. Progress toward automation has been reported in [23], [24] by performing landmark detection using local neighborhood image intensity features in ~100 images. Our work, by contrast, is an automated 3D volume registration scheme that allows both coronal and sagittal data to be processed in a genome-scale context of over 400,000 images.

The robustness of the registration solution to artifacts, missing/damaged sections, and the spectrum of gene expression patterns is extremely important for achieving a genome-wide survey. Fig. 4 illustrates (from left to right) examples of a low expressing gene (Il1f9), thalamus enriched expression (Prkcd), and more ubiquitous expression (Tuba1). We found that an iterative refinement approach using both shape constraints and the mutual information [25] between the ISH image intensity and the Nissl images of the ARA was necessary to address these complexities of the data.

While mutual information (MI) has been successfully used in many medical imaging applications [26], using MI as the sole criteria is not suitable here. Due to the broad scope of this paper, there is insufficient space to provide comprehensive background on mutual information and deformable registration and the reader is referred to comprehensive texts such as [27], [28] for a detailed overview. As can be seen in Fig. 4, apart from dense cells in some regions (hippocampus, cerebellum, and olfactory bulb) and in the absence of differential gene expression, there is inadequate intensity contrast between regions to drive the registration, especially in the initial stages where there might be insufficient overlap. On the other hand, the overall brain shape when reconstructed in 3D is an extremely robust feature to get two brains in good global alignment. We propose a combination driving function of shape and MI to subsequently refine the registration. Our overall strategy is to interleave the ISH volume reconstruction with template-to-ISH volume registration, where the template is used to constrain the reconstruction. The method starts with a coarse reconstruction and registration and iteratively refines both in stages. The major steps of the algorithm are:

1. Foreground identification: 2D tissue mask images are produced using adaptive thresholding.
2. Coarse 3D ISH reconstruction: The 2D centroid [29] is computed for each ISH tissue mask in the series and a coarse 3D ISH volume is obtained by aligning the centroids.
3. Coarse template-to-ISH registration: The ARA template is linearly registered to the coarse reconstruction volume by minimizing the mean square difference between the 3D tissue binary mask volumes.
4. ISH reconstruction refinement: In this step, the template is used to help improve the ISH volume reconstruction. The template tissue mask is resampled into ISH volume coordinates. Each 2D ISH image is
then rigidly registered to the corresponding virtual 2D template section by minimizing the mean square difference between the 2D tissue masks.

5. **Deformable registration I (global)**: Using a three 3D B-spline [30] to parameterize the deformation field, the registration is refined by using a low-dimensional global deformable registration to warp the 3D template mask to the ISH volume mask.

6. **Deformable registration II (regional)**: The registration is further refined by performing separate deformable registration over three anatomic areas (front, center, and back) using a weighted metric of mean squares difference in tissue mask and MI of the image intensity. Treating each region individually simplifies the underlying joint histogram of the MI.

7. **Deformable registration III (sectional)**: In this final step, the registration is refined on a per section basis, where registration only considers one ISH section at a time, using the same weighted metric as in the previous step. This step potentially allows independent section cutting/stretching distortions to be modeled.

Once registration is complete, the anatomic boundaries of the template are projected onto each 2D ISH image, using the transform computed in the registration. These boundaries are subsequently used later to compute anatomic region gene expression statistics. The multistage approach refining a rigid to deformable mapping was found to be essential because of the robustness required by high-throughput automation. A detailed analysis of annotation performance in the caudoputamen is given in the supplemental materials, which can be found on the Computer Society Digital Library at http://computer.org/tcbb/archives.htm.

### 2.4 Expression Segmentation

Automated image segmentation identifies the sites of higher visual concentrations of NBT (see Section 2.1) that correspond to sites of riboprobe binding. The TSA amplification step, inherent process variation, and lack of a common standard definition for ISH expression quantification makes defining “ground truth” a challenging endeavor. While approaches have been developed to directly quantify mRNA transcription via ISH [31], semiquantitative imaging-based methods are preferred due to the complexity of the process and amplification step [32].

Typical nonradioactive ISH images (Fig. 4) can have rather high levels of background intensity contaminated with nonspecific hybridization products that resemble low-level expressing cells. The automatic expression detection algorithm used in the ABA is designed to focus on achieving a high recognition probability for mid/high-level expressing cells while maintaining high specificity for rejecting nonspecific labeling and artifacts among low expressing genes. In contrast to a fixed threshold approach [32], we found that it was necessary to employ an adaptive algorithm that was sensitive to process variation and artifacts. In addition to a preprocessing step to color balance the images in a common gene series, there are three major stages in the expression detection algorithm:

1. **Tissue segmentation**: The goal is to obtain a reliable tissue mask for the major component of brain section. Using the 1.8 downsampled level of the image pyramid, an adaptive thresholding method [33] is applied to threshold the tissue area on image pixel intensity and tissue edge intensity distributions. Using this preliminary mask, a binary classifier examines the object area and compactness [34] of each of the segmented objects of the mask to reject nontissue objects such as small fragments, air bubbles, and other artifacts.

2. **Small and isolated object segmentation**: A $31 \times 31$ binomial kernel filtering [34] is performed on the high-resolution images to enhance the signal for cell-shaped objects of interest having the diameter of the neuron nucleus (e.g., $10-30 \mu m$) and to suppress the nonuniform background intensity. Next, thresholding by maximizing interclass variance of foreground (expression) pixels versus background (nonexpression) pixels is used to determine initial potential cells. An edge-enhanced image is also generated to capture objects of lower contrast which still have sufficient intensity strength. All detected objects are combined and passed through a multistage classification process to determine validity. Statistical and morphological characteristics of ISH expressing cells, such as size (area), shape (compactness, aspect ratio), and intensity (raw and filtered pixel values and integrated optical density) are combined from an offline precomputed database of expression patterns in several thousand images to enable efficient classification.

3. **Dense and clumped object segmentation**: Highly dense cell regions such as those found in the hippocampus and olfactory bulb have expressing tissue for which it is difficult to segment individual cell nuclei (Fig. 5a.) Image nonuniformity can produce more false positives due to the wider spatial span of these types of large objects. Object edge pixel strength and the orientation gradients [29] are used to segment the object boundary. The detection of these objects uses computed anatomic information such as the expected shape of the dentate gyrus of the hippocampus or cell densities in the olfactory bulb, as in Step 2. Each segmented object is a clump of densely packed cells for which the precise cell count is not available. Detection of these larger scale objects is performed at a lower-resolution (1:16) level of the image pyramid.

The segmentation results from the above three stages are finally combined into one mask. One of the key challenges with large-scale image data is to facilitate intuitive and effective visualization methods. We have found that an effective presentation is obtained by projecting the expression intensity level into a discrete level false color heat map whose categories are determined with the help of evaluation by trained neuroscientists. An analysis of the expression algorithm performance is given in the supplemental materials, which can be found on the Computer Society Digital Library at http://computer.org/tcbb/archives.htm.
The final output of this algorithm are expression maps presented on the ABA Website at www.brain-map.org.

2.5 Mapping Expression to the Atlas

The final stage of the pipeline is the classification of expressing cells and tissue in each of the mapped anatomic template regions. An effective mapping process should not lose access to the detailed properties of the image data and we use a 3D grid to facilitate this. The grid is a subdivision of the 3D reference atlas space into $100 \times 100 \times 100$ isotropic quadrat cells and serves as the highest resolution for quantifying expression information for the purposes of search and comparison. The reference atlas grid quadrats and corresponding structural annotation are mapped onto the ISH volume using the B-spline transform parameters computed earlier by the deformable registration module. The resulting deformed polyhedra from this mapping are subsequently projected onto each ISH section so that the 2D boundaries of a given structure on each original ISH image can be found. Fig. 6 shows an illustration of a deformed grid mapped onto an ISH cross-section. The use of grid techniques in atlas-based annotation is not new and has been proposed by [22].

Following the mapping of the grid onto the ISH image, local image statistics, such as the number of expressing cells, average cell size, average intensity, etc., are calculated for each quadrat. Since a one-to-one mapping exists between deformed quadrats in ISH space and quadrats in reference atlas space, gene expression statistics can be collected and mapped back into atlas space. Quadrats in the atlas space are labeled with the anatomic structures they intersect and, so, after calculating image features at the quadrat level, the features are combined into summary measurements for each brain structure. This facilitates intergene comparisons, as we show below, and provides the foundation for developing an effective anatomic-based search application.

3 RESULTS

3.1 The ABA Search Facility

More than 20,000 genes of the adult C56Bl/6J adult mouse have been processed in the pipeline and released to the public. The online search interface allows for the query of expression coverage (a measure of expression density) and expression intensity (a measure of punctate color) in each of 11 anatomic regions. The current search regions consist of cerebellum, cerebral cortex, hippocampal formation, hypothalamus, medulla, midbrain, olfactory bulb, pallidum, pons, striatum, and thalamus. Users can perform conjunctive or disjunctive (AND/OR) searches for up to three simultaneous structures. The search engine returns results that are ranked by specificity of expression, where the measure of specificity used is the number of structures showing expression for a particular gene. This assumes that the user will be most interested in genes that are specific to the structure of interest or at least a small number of structures. Genes with highly specific expression appear at the top of the list of query results, while genes having more widespread expression appear toward the bottom.

As an example, we performed a search using the public Website for genes demonstrating high expression coverage and high intensity in the striatum. A total of 209 genes met

![Fig. 6. Local deformed 2D projection of quadrats from reference atlas space.](image)
the search conditions. Sample ISH sections from the top 12 genes in the list are shown in Fig. 7, where section orientation variability occurs because the data is displayed in the original unregistered form. These genes all exhibit marked strong expression in the striatum and show good concordance with the literature for striatum expression in rat, mouse, or human tissue, [2], [35], [36], [37], [38]. The expression pattern for one of the genes, 2010300C02Rik, is novel and undocumented. Genes returned beyond the 12 also exhibit interesting striatal expression, yet with increasingly higher expression in other global structures. The choice of ranking prioritizes strong expression in the striatum, yet this is not necessarily an indication of the lower significance of the genes lower in the return list.

Fig. 8a shows a detailed comparison of genes 6 and 7 from the list, Rgs9 (Regulator of G-protein signaling) and Pdyn (prodynorphinin), both expressing in the caudoputamen (CP). Rgs9 and Pdyn expression is dramatically increased following neuroadaptation in the striatum due to cocaine exposure [35], [39]. Closer examination of the CP for these two genes, shown in Fig. 8a, reveals a subtle expression variation in the structural boundary of the CP. A detailed look at gene 8, brain-enriched Itpka (inositol 1,4,5-trisphosphate 3-kinase(A)), shows regional expression in CA1/CA2 and the dentate gyrus of the hippocampus (Fig. 8b). Itpka has been shown to be expressed in the dorsal striatum, cerebral cortex, hippocampus, and cerebellum (GNF Data set www.symatlas.org, GNF 1M, gcRMA) via microarray analysis and its deletion results in impairment of long-term potentiation in the CA1 region of the hippocampus [38]. Coexpression studies such as these, facilitated by appropriate online tools, illustrate the power of spatially mapped databases and concurrent image viewing.

3.2 3D Visualization Tools

The registration, segmentation, and expression measurements over grid quadrats enable 3D models of gene expression to be presented in reference atlas coordinate space. By using 3D viewing applications, users can quickly get a global sense of the distribution of expression across the brain for a particular gene. Fig. 9a shows Brain Explorer, an application we developed to visualize 3D gene expression in the reference atlas coordinate space. The spheres (gene expression for gene Pde10a) correspond to expression content in quadrats, where the size of the spheres is directly proportional to the number of expressing cells detected in a quadrat. The model is fully interactive, with controls for rotation, zooming, panning, and cutting planes as well as domain-specific controls to show atlas structures and filter expression to selected structures. Fig. 9b illustrates
expression for Pde10a again with color rendering by intensity (a visualization option in the Brain Explorer) together with the coronal and sagittal atlas planes and projected structures. The Brain Explorer application was used extensively in the validation and analysis of gene expression patterns for pipeline. The Brain Explorer executable is available from the ABA Website together with 3D expression summaries of all processed genes. It is also integrated with the search facility of the main Web application.

3.3 Large Scale Expression Correlation

A preliminary study based on scanning a subset of the ABA database of more than 20,000 genes indicates that about 78.8 percent are expressed at some level in the murine brain. An independent check for gross level expression was manually performed by a team of trained neuroscientists and achieves a comparable number. This number is considerably higher than that predicted by previous microarray experiments that put the number closer to 55 percent [3]. While, from the biological perspective, gene expression is deterministic, variation in biological specimens, image capture methods, artifacts, and inherent protocol variability indicate a more probabilistic informatics treatment. Determination of expression levels and the absolute categorization of “expressed” versus “nonexpressed” in a given structure is better treated from a statistical viewpoint. Fig. 10 illustrates the concept. By mapping 20,000+ ISH gene specimens (including both coronal and sagittal data) to the reference atlas and binning detected expression into a 300 μm resolution reference atlas coordinate system we can obtain a 3D map of the averaged fractional expressing area (ratio of expressing pixels over all pixels) for each 3D spatial bin in the mouse brain. In Fig. 10, we show one sagittal and one horizontal plane of the heat map in which brighter (white/yellow) color values indicate a higher fractional expressing area. The spatial variation in the heat map is due to both 1) differences in cell density between spatial regions and 2) the difference in gene expression frequency (i.e., number of genes expressing in a
local region). For example, the brightness of the neo-cortex potentially indicates that it is a region of high cell density and high frequency of gene expression. On the other hand, the darkness of the fiber tracts potentially indicates both low cell density and low frequency of gene expression. Further analysis is being undertaken to interpret this data and results from gene-to-gene correlation studies. As a side note, Fig. 10 also represents the co-registration of 20,000+ data sets and the ability to discern the corpus callosum and fimbria additionally confirms the accuracy of the automated registration to at least a 300 μm resolution.

4 DISCUSSION

4.1 Related Projects

In addition to our effort at the Allen Brain Atlas, related work includes Genepaint, (http://www.genepaint.org), Emage (http://genex.hgu.mrc.ac.uk/Emage), and BGEM (Brain Gene Expression Map; http://www.stjudebgem.org), a component of GENSAT (Gene Expression Nervous System Atlas; http://www.ncbi.nlm.nih.gov/projects/gensat). Other significant efforts include the UT Mouse Brain Library http://www.nervenet.org/ and the Harvard Brain Atlas http://mahoney.chip.org/mahoney/. These projects encompass a wide variety in the approach to data generation, level of anatomic mapping, and downstream processing, although a common shared characteristic is that they generate a large number of images of expression patterns. Overall, they represent an impressive array of initiatives using high-throughput methods, including colorimetric and radioactive ISH. Space does not permit a thorough review of the details and differences between these projects, but the reader is referred to the review article [40] for an overview.

With the exception of Genepaint (www.genepaint.org), most of the aforementioned projects use hand curated annotation as a means of associating expression patterns with anatomic regions and, as noted in [41], this approach is unsuitable to very large-scale image databases. Genepaint involves a more systematic geometric modeling effort based on techniques of Catmull-Clark subdivision meshes [22], [42] and enables a fine recursively editable mesh to be placed on each section of the ISH image. Queries to the database are enabled by the inherent multiresolution property of the data structure. As this project represents the only semiautomated approach to mapping the mouse brain transcriptome, a more detailed comparison with the present project is merited.

The technique used by the technique of Carson et al. in Genepaint is for a postnatal 7-day old (P7) atlas of the same C57BL/6j strain. Methods of anatomic mapping are accomplished by positioning the mesh with control points on each section by a semiautomated method that first uses a principal components analysis to find a general tissue alignment, subsequently requiring human intervention to best position the mesh. If the mesh is well-aligned, this method is powerful in that it enables considerable localization of anatomic features. The subdivision mesh is also attractive as a mechanism for capturing expression as the data structure is recursively refined at query time. The automated Allen Brain Atlas pipeline is, by contrast, a coarser anatomic mapping as registration is performed entirely without intervention. In the ABA, gene expression is mapped by deforming the 3D atlas volume to reconstructed ISH volumes and so is a 3D mapping from the starting point of 2D sections. The extension of the Catmull-Clark technique to 3D is an exciting possibility, but is problematic in that serious challenges exist in the positioning of the mesh control points in an automated manner. The throughput required by the ABA project demanded a more automated approach and, whereas it is possible to map 10-20 genes/day via the Genepaint approach, the ABA pipeline maps more than 500 genes/day. While this entails a lower specificity of mapping, it is precisely the rate necessary to accomplish a genomic scan. The expression detection algorithm used by Carson et al. is a threshold-based method [41]. Given the throughput constraints, full automation, as well as consequent variability in data quality, we found that a considerably more robust algorithm was necessary. The details of the algorithm discussed in Section 4.4 were necessary to account for artifacts and process variance.

4.2 Pipeline Issues

There are particular difficulties in developing a pipeline to perform anatomic mapping on a genomic scale and several innovative approaches have been developed [32], [43], [44]. The complexity and sheer volume of information generated necessitates specialized algorithms and a strong integrated processing environment. To achieve this, there was a need to incorporate capabilities from several areas of bioinformatics, medical image processing, and automation. Processing image data on a genomic scale has its own special challenges, including massive storage requirements, multimodality image registration problems, measurement and characterization of signal, and appropriate database construction to enable effective search. See the supplemental material, which can be found on the Computer Society Digital Library at http://computer.org/tcbb/archives.htm, for further details.

There were also notable challenges in architecting the registration pipeline as a limited number of conventional imaging controls were available. For example, block face imaging, whereby a static image of the cryotome mounted
brain is first imaged prior to cutting, interleaving Nissl sections, might have improved the registration if available. The neuroscientist’s requirement for a high resolution coronal reference atlas while having the gene survey performed in the sagittal plane mandated cross plane registration methods, thereby limiting the number of reliably identified structures to a comparatively small number. There were no fiducials or other landmark methods used and the registration and quantification was done in a fully automated manner. A certain amount of replicate data was available by virtue of the coronal and sagittal series run for certain genes, although no other replicates were used in the informatics processing. Limited section availability of 20 sagittal sections per gene and 54 for coronal series limits the accuracy of registration as well. Reconstruction must effectively interpolate information of missing sections. Nevertheless, our internal studies show that the registration is accurate to about 300 μm, which represents comparatively small structures such as represented by the three substructures of the striatum discussed in Section 3.2.

The ABA pipeline was designed foremost for automation and high-throughput to meet the challenge of mapping the entire brain transcriptome within a comparatively short time frame. Recent advances in software and the Web services model are well suited to the modern bioinformatics workflow that includes rapid configurability, remote user services and data sources, and distributed computing. Two excellent platforms for this model in neuroscience are the LONI pipeline [44] and SCIRun [45]. It is possible to configure these environments for high-throughput processing; however, as our goal was less toward building an environment to serve the broadest user group and more toward a highly specialized application, we opted for a less generalized implementation. The pipeline is adaptable within the domain of image-based signal data that is reconstructable from 2D sections and is certainly reconfigurable to run other mammalian tissues in ISH, such as the rat.

4.3 Data Mining and Future Directions

Global expression statistics and basic correlation by gross structural anatomy shed little light on microanatomy and cell type function and process, two of the hallmarks of modern neuroscience. A much closer look at the data based on the present work is possible by examining expression correlation at the 300 μm grid level and this is facilitated by the tools we have developed. While this is still not at a level of sufficient detail to define cell morphology except in an averaged way per grid cell, the potential for understanding coexpression and network pathway prototypes seems promising. We are presently developing tools to search and explore the genome by spatial expression correlation at the grid level using Brain Explorer for visualization. This would, in particular, yield solutions to the query of finding genes like a candidate set and is consistent with the probabilistic approach we mentioned earlier.

5 Conclusions

We have addressed the problem of developing an automated neuroinformatics pipeline capable of processing ISH data in a 3D framework on a genomic scale. More than 20,000 genes have been processed to date through the ABA informatics pipeline. While the workflow is optimized for high-throughput performance and 3D mapping on a genomic scale, rather than rapid reconfiguration or interoperability, it would be straightforward to adapt the platform to survey and map other mammalian organisms. Several advances were made in the areas of automated annotation and signal segmentation that should be applicable to other modalities of imaging data. The database and its organization lead to the possibility of spatial expression clustering studies in the mouse brain at an unprecedented level of resolution.

Acknowledgments

This work was sponsored by the Allen Institute for Brain Science. The authors wish to thank the Allen Institute founders, Paul G. Allen and Jody Patton, for their vision, encouragement, and support. The authors enjoyed important and fruitful discussions regarding ideas and methodology with David Anderson and Art Toga.

References

P. Fagergren et al., “Temporal Upregulation of Prodynorphin”


Lydia Ng received the BSc degree in computer science and the BE degree in electrical engineering in 1994 from the University of New South Wales, Sydney, Australia, and the PhD degree in electronics in 2000 from Macquarie University, Sydney, Australia. She is a member of the Informatics Group at the Allen Institute for Brain Science, Seattle, Washington. Her research interests include image registration, PDE-based image processing methods, data mining, and bioinformatics. She is a member of the IEEE.

Sayan D. Pathak received the BTech degree in instrumentation engineering from the Indian Institute of Technology, Kharagpur, India, in 1994 and the MS and PhD degrees in bioengineering from the University of Washington, Seattle, in 1996 and 2000, respectively. He is a senior analyst in bioinformatics at the Allen Institute for Brain Science, Seattle, and is an affiliate assistant professor in the Departments of Electrical Engineering and Bioengineering at the University of Washington, Seattle. His interests are in image/signal processing, multivariate data analysis, classification, databases, and search algorithms applied to solving biomedical problems. He has been the principal investigator on several US National Institutes of Health (NIH) sponsored projects and was one of the initial contributors to the advanced segmentation and registration toolkit (http://www.ilt.org), sponsored by the National Library of Medicine. He is a member of the IEEE.

Chihchau Kuan received the BS degree in electrical engineering from National Tsinghua University in 1982 and the MS degree in computer science from the State University of New York in 1988. Currently, he is a senior analyst of informatics at the Allen Institute for Brain Science, Seattle, Washington. In the past, he was among the lead researchers on several major projects applying image processing and pattern recognition techniques in quantitative cytology and histology, semiconductor inspection, security, and industrial automation, among which was the first ever FDA-approved fully automated cervical cancer primary screener. His research interests include image processing, visual information extraction and management, pattern recognition, and neural information processing. He is the author or coauthor of more than 25 technical papers, 19 US patents, and several patents in other countries or pending.

Christopher Lau received the BS degree in biological science from Carnegie Mellon University in 1997 and the PhD degree in bioengineering from the University of Washington in 2003. He is currently working on bioinformatics and visualization projects at the Allen Institute for Brain Science in Seattle, Washington.
Hongwei Dong, MD PhD, is a neuroanatomist who was formerly a postdoctoral student of Professor Larry Swanson at the University of Southern California. Dr. Dong led the team in the identification and designation of each of the neuroanatomic landmarks indicated in the Allen Reference Atlas and was responsible for all graphic drawings. The Allen Reference Atlas plays a central role in determining the coordinate system for atlas data mapping.

Andrew Sodt is a senior software engineer with more than 20 years of experience in designing and implementing large-scale grid and distributed processing environments. He is a member of the Allen Institute of Brain Science technology team and was responsible for the overall pipeline computing architecture for the Allen Brain Atlas.

Chinh Dang received the BA degree in biology from the University of Denver and the MS degree in information systems from the University of Colorado. She worked in the field of drug discovery and bioinformatics software to develop and implement data management systems. She leads the Allen Institute of Brain Science technology team in the design, development, and maintenance of internal information systems that provide the data management capabilities to support laboratory processes, data processing and the Allen Brain Atlas data pipeline.

Brian Avants received the PhD degree in medical imaging from the Penn Image Computing and Science Laboratory of the University of Pennsylvania. His interests include deformable mapping theory and biomedical applications, as well as the applications of computational and statistical techniques in image processing.

Paul Yushkevich received the BS degree in computer science and mathematics from the University of North Carolina at Charlotte and the MS and PhD degrees in computer science from the University of North Carolina at Chapel Hill, where he worked under the supervision of Dr. Stephen M. Pizer. He is a research assistant professor with the Penn Image Computing and Science Laboratory (PICSL) in the Department of Radiology at the University of Pennsylvania. His primary research interests involve the application of high-dimensional geometrical and statistical techniques to medical image analysis, with a particular focus on shape modeling via medial representations.

James C. Gee, PhD, is an associate professor of radiologic science and computer and information science, director of the Penn Image Computing and Science Laboratory (PICSL), and program codirector of the HHMI-NIBIB Integrated Graduate Training Program in Clinical Imaging and Informational Sciences, all at the University of Pennsylvania, Philadelphia. Well known for its contributions to biomedical image analysis and, in particular, the emerging field of computational anatomy, PICSL’s work focuses on the development of methods for quantifying the ways in which anatomy can vary in nature, over time, or as a consequence of disease or intervention. Some recent developments include approaches for geodesic image registration and temporal parameterization in the space of diffeomorphisms, construction of population templates using manifold theory, and statistical shape characterization with the medial representation. He is a member of the IEEE and the IEEE Computer Society.

David Haynor is a professor of radiology at the University of Washington. His research involves applications of image processing, such as morphometrics, functional MRI, brain perfusion, and white matter tract mapping. He is particularly interested in using these techniques to track structural brain differences in fetal alcohol syndrome, disorders of development and neurodegenerative disease.

Ed Lein received the BS degree in biochemistry from Purdue University in 1993 and the PhD degree in neurobiology from the University of California at Berkeley in 1999. His postdoctoral work at the Salk Institute for Biological Studies involved using laser capture microdissection, DNA microarrays, and high throughput in situ hybridization to examine highly specific gene expression in subregions and layers of the hippocampus and neocortex. A focus of his current work at the Allen Institute for Brain Science is utilizing gene expression as a tool to define specific neuronal subtypes in the brain and genetic manipulation of these cells as a means of understanding their roles in higher-level behaviors.

Allan Jones received the BS degree in biology from Duke University and the PhD degree in genetics and developmental biology from the Washington University School of Medicine and did postdoctoral training at the University of Pennsylvania. He joined the Allen Institute for Brain Science team in early 2003. He was instrumental in setting up scientific operations and hiring core teams responsible for the Atlas project. He was promoted to senior director of the Allen Brain Atlas project in 2004 and has led the project toward execution and delivery of its aggressive milestones, with completion occurring in mid-2006. Working closely with the founders, scientific, and business advisors, he has helped to establish the future scientific direction and programs for the Allen Institute and was promoted to the position of Chief Scientific Officer in 2006. He has broad scientific experience in genetics, molecular biology, and development. Prior to joining the Allen Institute, he held various scientific and management positions with Avitech Diagnostics, Rosetta Inpharmatics, and Merok and Co.

Mike Hawrylycz received the PhD degree in applied mathematics from the Massachusetts Institute of Technology and subsequently was a postdoctoral researcher in the Computer Research and Applications Group at the Los Alamos National Laboratory. He has worked in a variety of applied mathematics and computer science areas, addressing problems in consumer and investment finance, electrical engineering and image processing, and computational biology and genomics. At the Allen Institute for Brain Science, he is responsible for the direction of the informatics effort, including image processing, bioinformatics, and data analysis.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.