The Processing and Analysis of *in situ* Gene Expression Images of the Mouse Brain

Manjunatha N. Jagalur[1], **Chris Pal**[1], **Erik Learned-Miller**[1], **R. T. Zoeller**[2] **and David Kulp**[1]
[1]Department of Computer Science,
[2]Department of Biology & The Laboratory of Molecular and Cellular Neurobiology
University of Massachusetts, Amherst, MA 01003
{manju,pal,elm,dkulp}@cs.umass.edu, tzoeller@bio.umass.edu

Abstract

Many important high throughput projects use *in situ* gene expression detection technology and require the analysis of images of spatial cross sections of organisms taken at *cellular level resolution*. Projects creating gene expression atlases at unprecedented scales for the embryonic fruit fly as well as the embryonic and adult mouse already involve the analysis of hundreds of thousands of high resolution experimental images. We present an end-to-end approach for processing raw *in situ* expression imagery and performing subsequent analysis. We use a non-linear image registration technique specifically adapted for mapping expression images to anatomical annotations and a method for extracting expression information within an anatomical region. We also present a new approach for jointly clustering the rows and columns of a matrix and we relate clustered patterns to Gene Ontology (GO) annotations. Our approach should be applicable to a variety of *in situ* experiments but we focus here on imagery and experiments of the mouse brain – an application with tremendous potential for increasing our fundamental understanding of neural information processing systems.

1 Introduction

Many large scale molecular biology experiments now use cDNA microarray technology for measuring expression levels of a large number of genes for a small tissue sample or cell. However, there are a number of projects underway to map spatial patterns of gene expression using *in situ* hybridization (ISH) technology for tens of thousands of genes in different organisms. In contrast to microarray based methods, these projects can produce huge archives of high-resolution 2D and 3D images and involve the analysis of complex spatial patterns of expression in the context of anatomical structures, tissues and cells. Examples of these projects include: the Berkeley ISH embryonic fruit fly (*Drosophila*) experiments [16], the ISH mouse embryo experiments at the Max-Planck Institute [3], and the particularly massive scale ISH experiments involving over $21,000$ genes, and roughly 300, 5000×5000 pixel images per gene for adult mouse brains in the Allen Brain Atlas [1].

The processing and analysis of ISH experiments, linking of atlas based experimental archives with relevant scientific literature and comparing results with existing knowledge has the potential for tremendous impact on the scientific community. In our experiments here we focus on the processing and analysis of ISH experiments of the adult mouse brain. In this paper we outline some emerging problems and challenges and describe our end-to-end system. Our system consists of a non-linear, information theoretic, adaptive landmark based procedure for registering high resolution ISH imagery to a reference, a method to obtain mappings to spatial and anatomical regions and a novel method for jointly clustering the rows and columns of a matrix using a variational learning method and a sequential optimization approach. Finally, we show that the results of our analysis allows

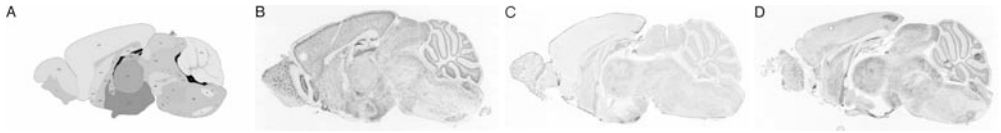

Figure 1: A. Reference is a centrally located sagittal image. Expression images for B. (Abr) is one of the best quality images. Most are of quality C (Adcy5) and D (Astn1) is amongst the worst with substantial tissue damage and distortion.

one to discover clusters of biological relevance, focusing our experiments on relating clusters to the Gene Ontology (GO) database [2].

2 Image Registration and Expression Level Extraction

Since our goal is to gather statistics about common expression patterns in anatomical structures across experiments, it is important that we achieve an accurate and robust registration. Due to the intrusive nature of processing steps in ISH experiments such as organ extraction and cutting into slices, substantial deformations, artifacts and tissue damage can arise. Fig. 1 illustrates some of the challenges involved with the processing and analysis of these types of experiments. A good review of existing brain warping techniques is given in [15]. For many ISH experiments, image resolutions and raw image sizes are much greater than typical medical imagery. Our registration approach is particularly adapted to these properties.

For our experiments here, we use 100 centrally located, sagittal 2D expression experiment slices and find the closest corresponding reference slices in the Allen Atlas [1]. In [18] a coarse to fine 3D approach was used to register a histology reference volume of the Allen Atlas imagery. We will restrict our discussion and experiments here to robust non-linear 2D registrations between experimental and reference slices. While full 3D registration is desirable, high spatial resolution in the image plane and relatively widely spaced slices across the volume for expression imagery make such approaches particularly challenging.

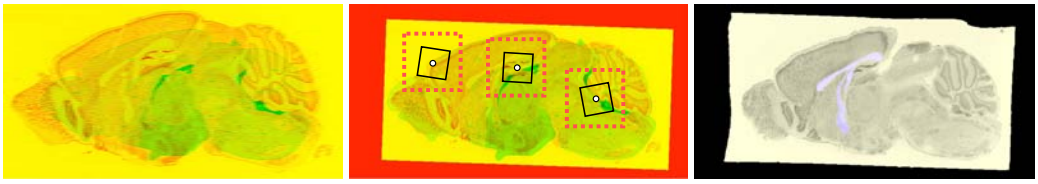

Figure 2: Green channel is the expression image, red channel is the reference image, each are at 400×800 pixel resolution. (Left) Before registration. (Middle) After approximate registration where dotted squares indicate search spaces for smaller anchor patches within. (Right) After our adaptive non-linear registration step anatomical patterns are extracted using a mask (shown in blue).

We start with a Nissl stain reference image with hand annotated anatomical regions (fig. 1A). For each experimental ISH expression image, we perform a coarse registration with a global affine transformation (fig. 2, middle) and use a discrete search with a mutual information based matching criterion [17]. We use a joint histogram based approach similar to [10] but with 16 equally spaced intensity bins. 75 high entropy 100×100 pixel patches are identified in the reference image which are then used to define anchor locations within each expression image for our refined registration. Patches may overlap, but we limit the maximal overlap to be 50%. Entropy is a measure of information about local shape and structure in a region and choosing high entropy patches reduces our chances of spurious registrations. Other authors have also used entropy measures to identify *salient regions* for more general correspondence problems in computer vision [9].

The anchor patches are more precisely mapped to the experimental image by performing a local search over a small subset of affine transformations in a 150×150 pixel window defined by the initial anchor position in the reference. This approach is similar to the montages of transformed latent

images introduced in [13]. However, here again we use a local mutual information based matching criterion. After local registration, the centers of the patches in the reference and experimental images serve as the key pixel correspondences between the images. From these points, two corresponding Delaunay triangulations are constructed in the two images and we perform an outlier rejection step to eliminate inconsistent correspondences. The final, refined registration is obtained by performing a bi-cubic interpolation of each point with respect to the encompassed triangle. Once the expression images are registered, the annotated anatomical regions in the reference image are simply mapped to the experimental image, allowing for the expression levels within each region to be easily extracted in a semi-automated way.

We have experimented with simple summary metrics for each anatomical region such as the mean and median expression level, however, we have found that a robust, quantile measure results in superior performance. Fig. 3 shows the 70th percentile for each of 38 anatomical regions for 100 genes. Once this form of matrix summarization has been constructed we can apply cluster analysis techniques in a manner similar to the analysis of cDNA microarray experiments. One approach is to independently cluster the rows and columns of the data matrix [5], ignoring any dependencies between the two clustering problems. Other approaches based on *direct clustering* [7] or *biclustering* [11] simultaneously cluster both the rows and columns of a data matrix and have been applied to microarray data [4] as well as other heterogeneous data [14]. Other approaches have cast bi-clustering in various cost minimization frameworks. We now present and apply a novel approach to bi-clustering in which we formulate the problem as one of inference and optimization in a formal probability model, a joint row-column mixture model.

3 Row Column Mixtures for Cluster Analysis

Consider a data matrix $\mathbf{X}$ where elements of the matrix are written as $x_{i,j}$. In our model each row i of the matrix is associated with a row class random variable $r_i \in \{1, ..., n_r\}$, where n_r is the number of possible row classes. Each column j of the matrix is associated with a column class random variable $c_j \in \{1, ..., n_c\}$, where n_c is the number of column classes. The conditional distribution for element $x_{i,j}$ is then a function of the random variable associated with row i and column j. As such, the joint distribution of the data $\mathbf{X}$, row classes r_i and column classes c_j can be written:

$$P(\mathbf{X}, R, C) = \prod_{i}^{N_r} \prod_{j}^{N_c} p(x_{i,j}|r_i, c_j) P(r_i) P(c_j). \tag{1}$$

Here we will use Gaussian models where $P(x_{i,j}|r_i, c_j) = \mathcal{N}(x_{i,j}; \Theta_{r_i,c_j})$, where $\Theta_{r_i,c_j} = \{\mu_{r_i,c_j}, \sigma^2_{r_i,c_j}\}$ although other choices of distribution are possible. The unconditional distribution for each row class r_i and column class c_j is given by $P(r_i) = \pi_{r_i}$ and $P(c_j) = \pi_{c_j}$ respectively. Let all the row and column classes be written as $R = \{r_1, ..., r_{N_r}\}$, where N_r is the number of rows in the matrix and let $C = \{c_1, ..., c_{N_c}\}$, where N_c is the number of columns. It is insightful to contrast row column mixtures with a traditional mixture of Gaussians for the rows of a matrix where the joint distribution for the data matrix and the row classes is given by: $P(\mathbf{X}, R) = \prod_i \prod_j N(x_{i,j}; \mu_{r_i,j}, \sigma^2)\pi_{r_i}$,where $\mu_{r_i,j}$ now represents elements of vectors $\boldsymbol{\mu}_j$. If, in the joint row column model we assign each column to its own class then the models are equivalent.

One way to optimize parameters Θ of a row column mixture is to use a variational [8] Expectation Maximization (EM) approach. We use an approximation to the posterior distribution $P(R, C|\mathbf{X})$ consisting of $Q(R, C) = \prod_i Q(r_i) \prod_j Q(c_j)$. To optimize a variational bound on the log probability of the data we start with initial guesses (e.g. uniform distributions) and iteratively update $Q(r_i)$s and $Q(c_j)$s. Starting with an initial guess for $\tilde{\Theta}$, we repeat the following two steps until convergence: (1) Variational E-steps for one or more rounds updating Qs, then (2) An M-step, updating $\tilde{\Theta}$. To express our variational E-steps succinctly, define hidden row and column class membership or indicator variables as $H = \{R, C\}$. It can be shown that the variational updates for fully factorized Qs can be written $Q^*_i(\{H\}_k) = \exp\left[E_{Q_{l\neq k}}\langle \ln P(\tilde{\mathbf{X}}, H)\rangle\right]\left[\sum_{\{H\}_k} \exp\left[E_{Q_{l\neq k}}\langle \ln P(\tilde{\mathbf{X}}, H)\rangle\right]\right]^{-1}$ where $E_{Q_{l\neq k}}\langle\cdot\rangle$ denotes the expectation under all $Q_{l\neq k}$ and $\tilde{\mathbf{X}}$ represents an observed data matrix $\mathbf{X}$. Since one holds $Q_{l\neq k}$ constant, these computations are performed locally in the graphical model of (1). Variables are updated in turn under random permutations of their ordering over itera-

tions. The updates of parameters of the model are computed via a Maximization or M-step, setting $\frac{\partial}{\partial\theta} E_Q \langle \log P(\tilde{\mathbf{X}}, R, C) \rangle = 0$, giving us closed form updates for Θ.

A second optimization method is to start with a hard assignment for row and column classes and search for new Maximum a Posteriori (MAP) parameter $\{\tilde{\Theta}, \tilde{R}\}$ and variable $\tilde{C}$ assignments by considering individual row $\tilde{r}_i$ and column $\tilde{c}_j$ assignment changes sequentially. To perform this optimization, we cycle through the data under a random permutation, removing the contribution of each row or column and compute the optimal class re-assignment *after* the parameters have also been updated. We refer to this type of algorithm as a sequential optimization.

4 Results, Discussion and Conclusions

To illustrate the robustness and quality of our registrations figures 3, 4 and 5 show a progression of good to poorly registered images. Approximately 55, 40 and 5 images of our 100 image set could be characterized as belonging to these good, moderate and poor registration categories respectively.

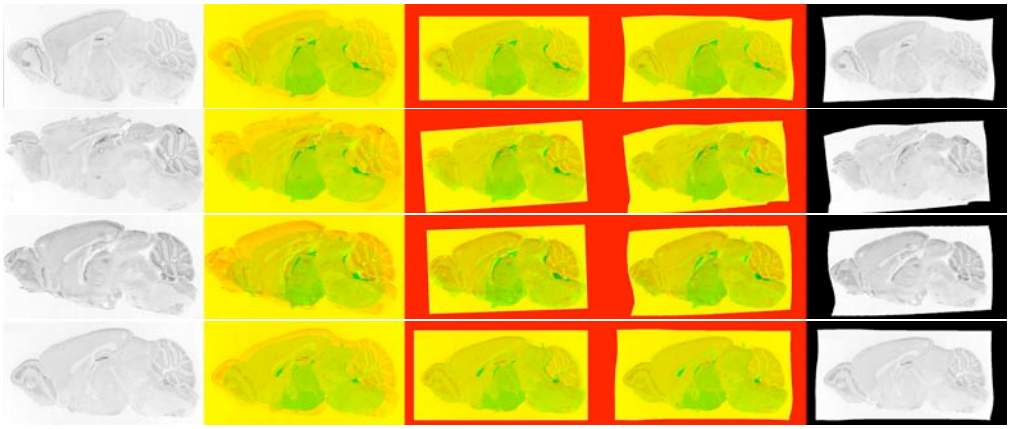

Figure 3: A selection of relatively well registered images. Leftmost image is the original image and rightmost is the final registered image. Other images show the difference between the reference image (red channel) and expression image (green channel) first as the original images, then after coarse registration and then the final registration respectively. Figures 4 and 5 use the same arrangement.

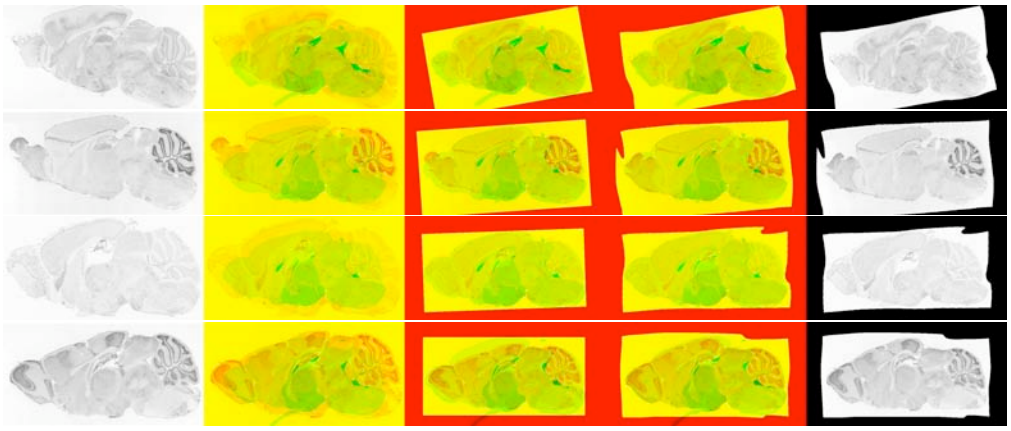

Figure 4: A selection of the moderately well registered images.

Our experiments have shown that *when block constant patterns are indeed present within the data*, both the variational and sequential methods described in section 3 produce comparable and improved quality clustering results in comparison to independently applied row and column clustering methods and a variety of other optimization algorithms [12]. We have used the sequential method for

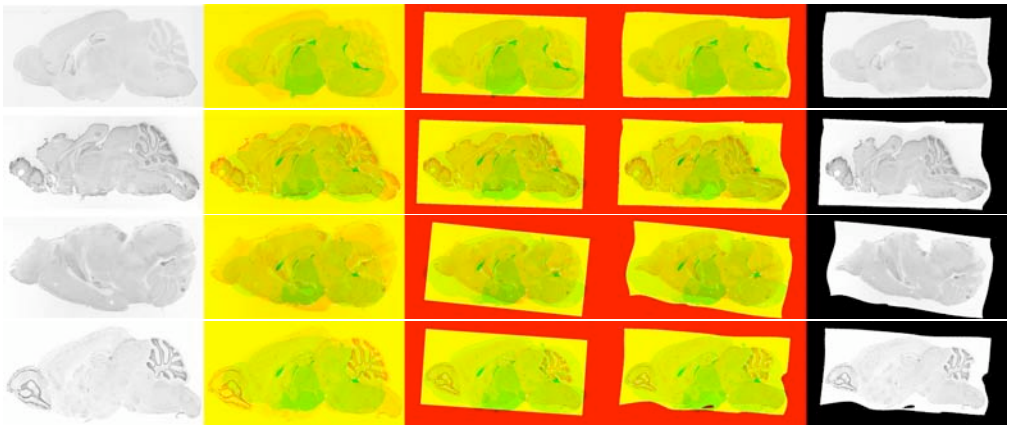

Figure 5: A selection of more challenging and poorly registered images.

the results shown in fig. 4 as our implementation runs faster for matrices of the size examined here. We used the best result over 10 runs for a 5 row and 5 column class model, but we see annealing approaches, variational MCMC hybrids and automated model selection methods as having great potential for this model. Based on other experiments [12] we anticipate that variational methods should have superior running time performance for data matrices on the order of $20,000$ experiments and hundreds to thousands of anatomical structures and sub-structures. Given these initial results, there appears to be reasonable block constant structure present in the data matrix. However, recent related methods allowing overlapping groups such as Matrix Tile Analysis [6] seem promising. Further, methods which also account for spatial proximity of brain regions seem promising.

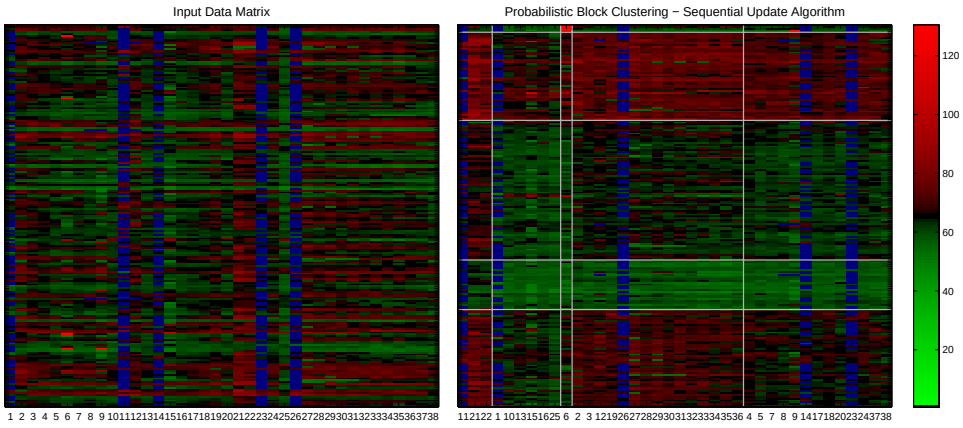

Figure 6: (Left) The original data matrix where each row corresponds to a gene expression experiment and each column corresponds to an anatomical region. Matrix elements represent expression levels. (Right) A permutation of the original matrix $\mathbf{X}$ after row column clustering.

The following observations give support that our registration and mask based feature extraction methods are of good quality. The set of highly expressed genes for mask 1, the olfactory bulb, is enriched for feeding behavior genes with a p-value of .008. Learning and memory genes are highly expressed in mask 11, the medial habenula which is just below the hippocampal formation, (p-value .001). Analysis of the clustering in Fig. 4 also suggests that this aspect of our approach can yield biologically meaningful information. We found that many genes within our clusters had high expression values in organs consistent with their GO annotations. For example, gene/row cluster 2 contains genes Aff2, Prkar1b, Shc3, Tmod2, Abi2 and is therefore enriched for category GO:0007611 "learning and/or memory" with a p-value of 10^{-3}. All p-values were computed using hypergeometric based enrichment tests. In conclusion, we believe the methods we have developed

here should be applicable to other contexts and organisms and should scale to higher resolution and genome scale data.

Acknowledgements

We thank Michael Hawrylycz and the Allen Brain Atlas for feedback and for providing access to data. This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. This work is also supported in part by CRI: Computational Biology Facility for Western Massachusetts, award number CNS 0551500, NIH (1R01HG003880), NSF (CNS 0551500), Microsoft Research under the eScience and Memex funding programs and by Kodak. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

References

[1] The Allen Brain Atlas, http://www.brain-map.org.

[2] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[3] J. Carson, C. Thaller, and G. Eichele. A transcriptome atlas of the mouse brain at cellular resolution. *Current Opinion in Neurobiology*, 12:562–565, 2002.

[4] Y. Cheng and G. Church. Biclustering of expression data. In *Proc Int Conf Intell Syst Mol Biol*, volume 8, pages 93–103. 2000.

[5] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[6] I. Givoni, V. Cheung, and B. Frey. Matrix tile analysis. In *The Proceedings of UAI*. July 2006.

[7] J. Hartigan. Direct clustering of a data matrix. *JASA*, 67:123–129, 1972.

[8] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. Kluwer, 1998.

[9] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[10] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *Medical Imaging, IEEE Transactions on*, 16(2):187–198, 1997.

[11] B. Mirkin. *Mathematical Classification and Clustering*. Dordrecht: Kluwer, 1996.

[12] C. Pal. *Probability Models for Information Processing and Machine Perception*. PhD thesis, University of Waterloo, 2004.

[13] C. Pal, B. Frey, and N. Jojic. Learning montages of transformed latent images as representations of objects that change in appearance. In *ECCV '02: Proceedings of the European Conference on Computer Vision, Springer-Verlag lecture notes in Computer Science*, volume 4, pages 715–731. Springer-Verlag, 2002.

[14] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA*, 101(9):2981–2986, March 2 2004.

[15] A. W. Toga. *Brain Warping*. Academic Press, 1999.

[16] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. Celniker, and G. Rubin. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biology*, 3(12), 2002.

[17] P. Viola and William M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

[18] P. A. Yushkevich, B. B. Avants, L. Ng, M. Hawrylycz, P. D. Burstein, H. Zhang, and J. C. Gee. 3d mouse brain reconstruction from histology using a coarse-to-fine approach. In *The proceedings of the Third International Workshop on Biomedical Image Registration (WBIR)*, pages 230–237. 2006.