

Unsupervised Analysis of Gene Expression in Neurological Animal Models

Kevin Hu

Massachusetts Academy of Math and Science

Abstract

Coming soon to a STEM fair near you.

Contents

| | | |
|-----------|---|-----------|
| 1 | Introduction | 3 |
| 2 | Literature review | 4 |
| 2.1 | Neurological disease modeling | 4 |
| 2.2 | Brain development | 5 |
| 2.3 | Gene expression regulation | 6 |
| 2.4 | Gene expression measurement | 7 |
| 2.5 | Gene expression analysis | 9 |
| 2.6 | The Allen Brain Atlas | 10 |
| 2.7 | Neural networks | 11 |
| 2.8 | Unsupervised learning | 17 |
| 3 | Research plan | 18 |
| 3.1 | Researchable question | 18 |
| 3.2 | Hypothesis | 18 |
| 3.3 | Procedure | 18 |
| 4 | Methodology | 20 |
| 5 | Results | 20 |
| 6 | Discussion | 20 |
| 7 | Conclusions | 20 |
| 8 | Limitations | 20 |
| 9 | Extensions | 20 |
| 10 | Acknowledgments | 20 |

1 Introduction

The use of animal models is an essential component of the drug development process. Animal models, in particular mouse models, allow diseases to be studied and the effects of candidate drugs to be observed in real time. This is crucial in neurological research where mice serve as valuable filters to faulty drugs prior to clinical trials.

However, genetic differences between humans and animal models often confound drug tests. Drug-gene interactions result when the products and effects of a gene interfere with the action of a drug. This interference may manifest as blockages of drug metabolism, mutated receptor proteins, and general physiological differences. Different patterns of spatial and temporal expression between humans and the model organism are factors in these interactions and are responsible for fundamental differences in morphology and cellular behavior.

Clustering, or the unsupervised categorization of data points into discrete sets, is widely used in gene expression analysis and enables the comparison of expression levels between two organisms through examination of the transcriptional landscapes. However, existing clustering algorithms rely heavily on measures of correlation that are prone to outlier leverage and high sample variance, both of which are common in biological samples. Recently, neural network algorithms have shown promise in the clustering of images such as those of the popular MNIST and USPS digit databases. The goal of this project was to examine the differences between human and mice gene expression in cerebral development using a recurrent convolutional neural network for the clustering of region-time expression matrices.

2 Literature review

2.1 Neurological disease modeling

Mice are extensively used in neurological disease modeling due to their relatively low time of growth and the fragility and lack of available human brain tissue. However, recent studies offer conflicting views regarding the similarity of mice and human gene expression patterns, which casts doubt on the accuracy of results of mouse trials. Some studies suggest that neural development expression is highly conserved between mice and humans (Lin et al., 2014). Others suggest that there exist fundamental differences between the two that undermine the use of mice as a viable model for diseases such as Alzheimer's and amyotrophic lateral sclerosis (ALS) (Burns et al., 2015).

With the advent of genetic engineering and regenerative biology, researchers have recently begun to switch to stem cells as a possible solution to the inaccuracy found in mouse models. Induced pluripotent stem cells (iPS cells) may be derived from nearly any sample of patient cells, allowing them to be generated from noninvasive skin grafts. These iPS cells mimic embryonic stem cells (ES cells) and are able to differentiate into any adult cells. iPS cells may specifically be exposed to a variety of molecular and physical factors that guide them towards differentiation into neurons. These neurons, which are genetically identical to the donor, may then be used in drug discovery, disease modeling, and cell-based therapy techniques. In addition, these cells offer insights into the genetic backgrounds and effects of disease-associated genes by allowing precise *in vitro* examination of their underlying mechanisms (Imaizumi and Okano, 2014).

In addition to culturing neurons, researchers have also managed to guide development of entire clusters of neurons that may mimic the patterns and timings of human brain development (Lancaster and Knoblich, 2014). These cell clusters, or cerebral organoids, may be generated by culturing patient-derived iPS cells within an environment that encourages agglomeration and growth (Nguyen et al., 2015). This project may also examine and cluster patterns of cerebral organoid gene expression in addition to those of mice and

humans.



Figure 1: A cerebral organoid viewed under fluorescence microscopy (Nguyen et al., 2015).

2.2 Brain development

The mammalian brain is among the most complex organized structures in biology. The development of the animal nervous system, termed neural induction, is initiated by the formation of several primitive cell clusters early in development. This begins with the formation of the neural plate on the ectoderm (outer layer) of the embryo, which folds inward as the neural fold/groove to create the neural tube and crest (Figure 2).

The neural tube is characterized by four regions which develop into distinct adult structures (Bakken et al., 2015). The first of these, the prosencephalon, develops into the telencephalon and diencephalon, which further develop into the forebrain, hypothalamus, and eyes. The second, the mesencephalon, develops into the midbrain. The third, the rhombencephalon (named for a rhombus-like shape), is responsible for the lower regions of the brain such as the cerebellum, pons, and medulla oblongata, which originate from the metencephalon and myelencephalon. The last of these regions is the spinal chord.

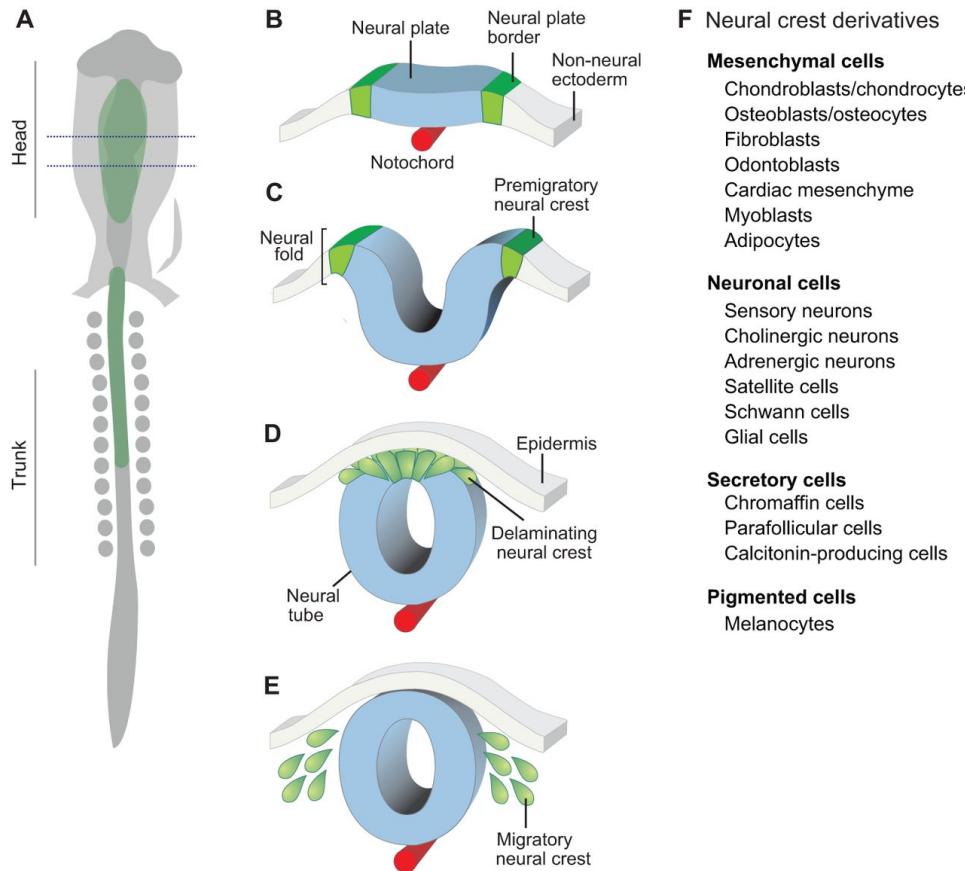


Figure 2: Development of the neural tube. (Simões-Costa and Bronner, 2015)

2.3 Gene expression regulation

Gene expression at the cellular level is regulated by a plethora of molecular means and is the foundation of an organism's development, growth, and survival. Although the vast majority of cells (except germ and certain immune cells) in an animal have identical DNA, their own structure and function is determined by these methods of regulation. This cellular regulation further determines the structure and function of the upper-level tissues, organs, and ultimately the organism itself.

Gene expression consists of the processes of DNA transcription to RNA and RNA translation to the amino acid chain that becomes the protein. Transcription is influenced by transcription factors, which affect how well an RNA polymerase (the enzyme primarily responsible for transcription) is able to bind to and form a transcription complex on a gene. Specific examples of molecules affecting gene expression include the addition of

acetyl or methyl groups to the DNA and proteins that by binding to the DNA, can either attract or block the binding of RNA polymerase. In eukaryotic cells such as those of mice and humans, the RNA transcript must be further processed prior to transcription. This post-transcriptional processing involves the addition of cap and tail sequences that protect the sequence from digestive enzymes and allow for translation to occur. The RNA transcript also contains exon and intron regions, with only the exon regions being incorporated into the final sequence to be translated. Once the RNA transcript has left the nucleus of the cell (through a nuclear pore), a ribosome (a large cellular machine that facilitates translation of RNA to protein) binds to the RNA and initiates construction of the amino acid chain. Once the amino acid chain is complete, the ribosome splits in two, and the protein may be further modified. For instance, the chain may be folded within a chaperonin, which provides an optimal environment for some proteins, or it may be tagged with small molecules that identify its destination. These methods, and many others, allow for extremely precise regulation of complex genetic pathways that are the drivers of embryonic development.

The embryonic phase of development is known for highly dynamic expression of genes across time and space, or spatiotemporal regulation. Differential gene expression over space and time is the feature that allows an organism to develop specialized cells. Conditional regulation of only a select few high-level "master genes" can determine the fate of a cell. (Bisceglia, 2010)

2.4 Gene expression measurement

Numerous methods exist for the quantification and visualization of gene expression data. In this study, the gene expression levels were computed from either *in situ* hybridization (ISH) or RNA sequencing (RNA-Seq).

ISH is a technique that allows specific sequences of nucleic acids to be located in a section of cells or tissue. This localization is accomplished through the principle

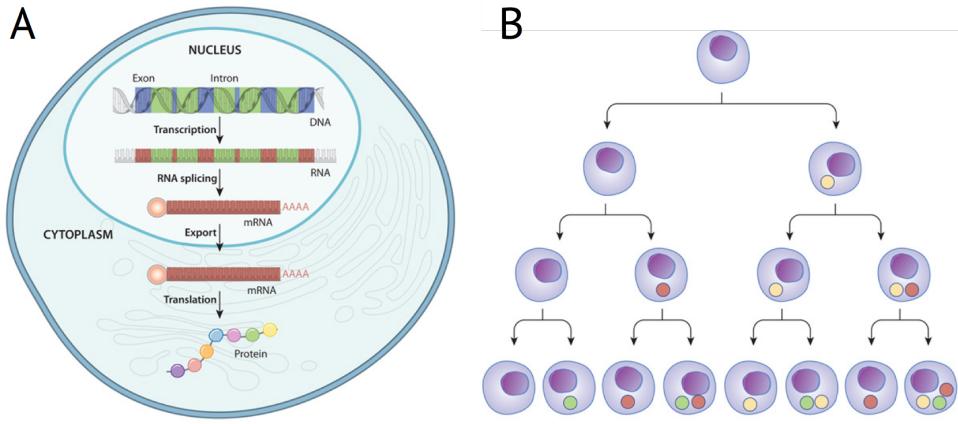


Figure 3: **A.** A general overview of the flow of information from DNA to RNA to protein. **B.** Demonstration of cell differentiation (Bisceglia, 2010).

that complementary nucleic acid sequences form hydrogen bonds (hybridize) with each other. The complementary sequence that detects the target sequence, or the *probe*, may be identified through fluorescent (fluorescent DNA components), radioactive (radioactive isotopes), or immunohistochemical (antibody-labeled) means. After the probe is allowed to hybridize to the target strand, RNases (RNA-digesting enzymes) hydrolyze any excess probes and the rest are washed off, leaving only bound probes. Following ISH, fluorescent microscopy may then be used to highlight regions of gene expression, which may then be scanned and analyzed using image analysis algorithms (Angerer and Angerer, 1991). Regions of high expression of the sequence of interest are shown as darker areas in an image (Figure 4).

In addition to ISH, RNA-Seq is also used in order to accurately measure gene expression. Because the primary function of RNA in the cell is as an intermediate in the transfer of information from DNA to protein, RNA expression levels can give an approximate indication of overall gene expression levels. In RNA-Seq, as the name suggests, next-generation nucleic acid sequencing is applied to a collection of RNA sampled from a tissue region. This RNA sample is first isolated from the source tissue through digestion using detergents and enzymes. Once the RNAs are isolated, a next-generation sequencing machine reads all the RNA to a digital storage device. These raw sequence data are

then formatted to identify exons and introns, which may be later used to identify gene expression levels. These gene expression values provide a value of the degree of a gene's expression within a cell (Wang et al., 2009).



(D'haeseleer, 2005). A significantly altered clustering would suggest that one variable affects the clustering to an abnormal degree, which may imply that the clustering method is chaotic.

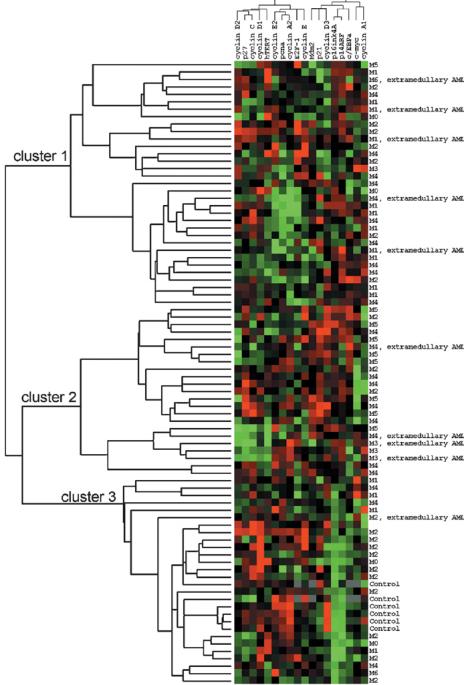


Figure 5: An example of a clustering algorithm applied to expression values of genes related to acute myeloid leukemia. Dendograms were created for regions (top) and sample (left) (Muller-Tidow et al., 2004).

2.6 The Allen Brain Atlas

Founded in 2003 by Microsoft co-founder Paul Allen, the Allen Institute for Brain Science is a nonprofit research organization dedicated to the public pursuit of brain research. Among the many free datasets provided by the Institute are the Allen Developing Mouse Brain Atlas (located at www.developingmouse.brain-map.org/) and the Allen Atlas of the Developing Human Brain (located at www.brainspan.org/).

The Developing Mouse Brain Atlas features *in situ* hybridization (ISH) data for over 2,100 genes across multiple stages of embryonic and postnatal mouse brain development. The raw images of the ISH scans total 434,946 images. These images were

assembled into 3-dimensional grids of expression values. In addition, an application programming interface (API) in the form of a bioinformatics pipeline allows researchers to access all data produced by the Developing Mouse Brain studies. The Developing Mouse API further allows researchers to obtain quantized expression values per region, which are calculated based off of analysis of the ISH scan images. These expression values may be found for several different developmental time points, at several discrete brain regions, for each of the 2,100+ genes (Thompson et al., 2014).

Similar to the Developing Mouse Brain Atlas, the Developing Human Brain Atlas also features a diverse array of ISH expression values for several thousand genes. These expression values are available for download as raw comma-separated-values (CSV) files from the Developing Human Brain Atlas Website, although an API is still available. In addition to ISH images, extremely detailed, cellular-level, magnetic resonance imaging (MRI) and microarray data are also included (Miller et al., 2014). The expression levels for these images may be summarized using provided image analysis techniques which yield region-time expression matrices (Figure 6).

2.7 Neural networks

Neural networks are a subfield of machine learning, which is often cited as the “field of study that gives computers the ability to learn without being explicitly programmed,” a definition given by Arthur Samuel in 1959. Although invented around the mid 19-th century, neural networks today are one of the most flexible and powerful machine learning algorithms. In several areas of research, neural networks offer state-of-the-art performance over traditional, hand-coded methods, such as in image pattern recognition, genomic analysis, and protein folding prediction. The basic neural network node, or *neuron*, is similar to its biological counterpart in that it receives a number of inputs and outputs an output (Figure 7). However, the two differ in both structure and function. For instance, the biological neuron asynchronously receives a combination of excitatory

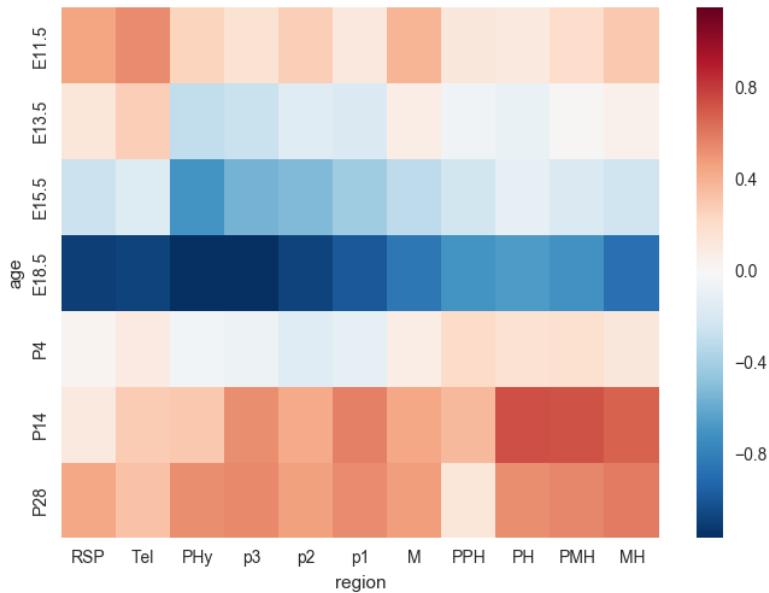


Figure 6: A heatmap of expression values for the *Abelson murine leukemia viral oncogene homolog 1* (*ABL1*) gene in the developing mouse brain. Expression values were obtained from ISH stains and computed for each region using specialized 3-D voxel counts provided by the Allen Brain Atlas API. The y-axis is the development stage and the x-axis is the brain region (abbreviated). The expression values were transformed to a logarithmic scale to account for skew.

and inhibitory impulses and sends a constant output signal. An artificial neuron instead is a mathematical construct that receives inputs synchronously and usually has a variable output value. In addition, whereas biological neural networks tend to be complicated in their topologies (many are not yet understood), artificial neural networks almost always have a tree-based structure of connectivity. The two further differ in their application: whereas brains are capable of extrapolation and learning of novel tasks, artificial neural networks are specialized and cannot readily adapt to new tasks given learned information.

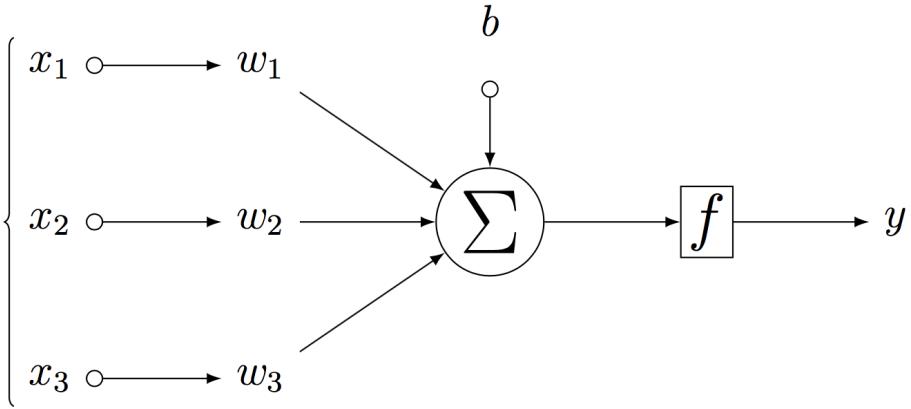


Figure 7: A diagram of a basic neuron with input values x_1, x_2 , and x_3 with respective weights w_1, w_2 , and w_3 and a bias b . The neuron computes the sum of these inputs and outputs a value y determined by activation function f .

Neural networks learn by adjusting weights and biases to minimize a loss or objective function which determines the degree of error present in the predictions of a network. When a network is first initialized, the weights present are usually set to zero or randomly determined using a probability distribution function. As the network is fed data, the outputs are compared to target values and a loss value is computed. This loss value is then used to adjust the weights of the network to better fit the inputs to the output using a method called backpropagation in which the effect of each neuron on the loss function is computed using the chain rule. The weights are then updated using a method of gradient descent in order to minimize the loss value. This iterative updating allows the network to eventually learn a complex function given enough training data and time.

There exist a multitude of activation functions which determine the output of an artificial neuron (Figure 8). Many of these functions are designed such that the loss gradients have smooth derivatives that can be handled by backpropagation. For instance, a step function would not be able to be backpropagated. The activation functions will also partly determine the computational workload; a ReLU activation gradient is much faster to compute than a sigmoid activation gradient.

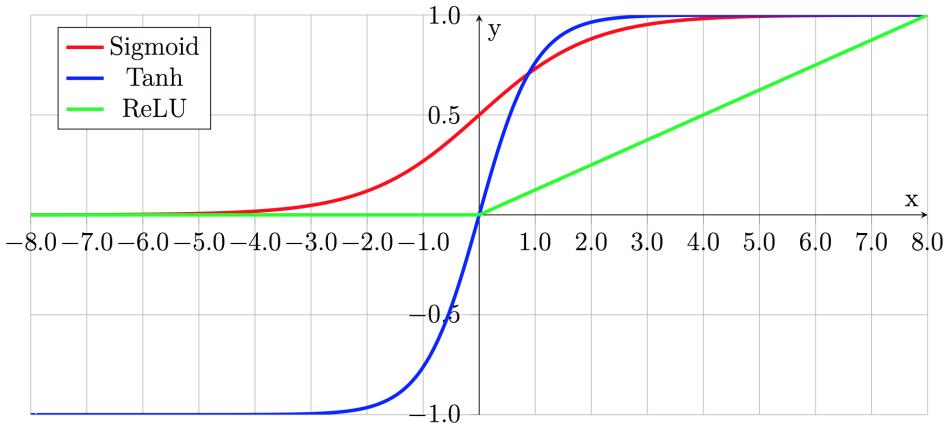


Figure 8: Graphs of common neuron activation functions.

Table 1: Equations of the activation functions shown in Figure 8.

| Linear | Sigmoid | Tanh | ReLU |
|------------|-----------------------------|--|--|
| $a(x) = x$ | $a(x) = \frac{1}{1+e^{-x}}$ | $a(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ | $a(x) = 0 \text{ if } x \leq 0, x \text{ if } x > 0$ |

Neural networks are able to approximate complex functions such as classifiers through abstraction by multiple layers. In general, larger neural networks tend to yield more accurate predictions and lower losses. However, large networks are very computationally expensive to train and may also overfit the data when there are not enough training examples.

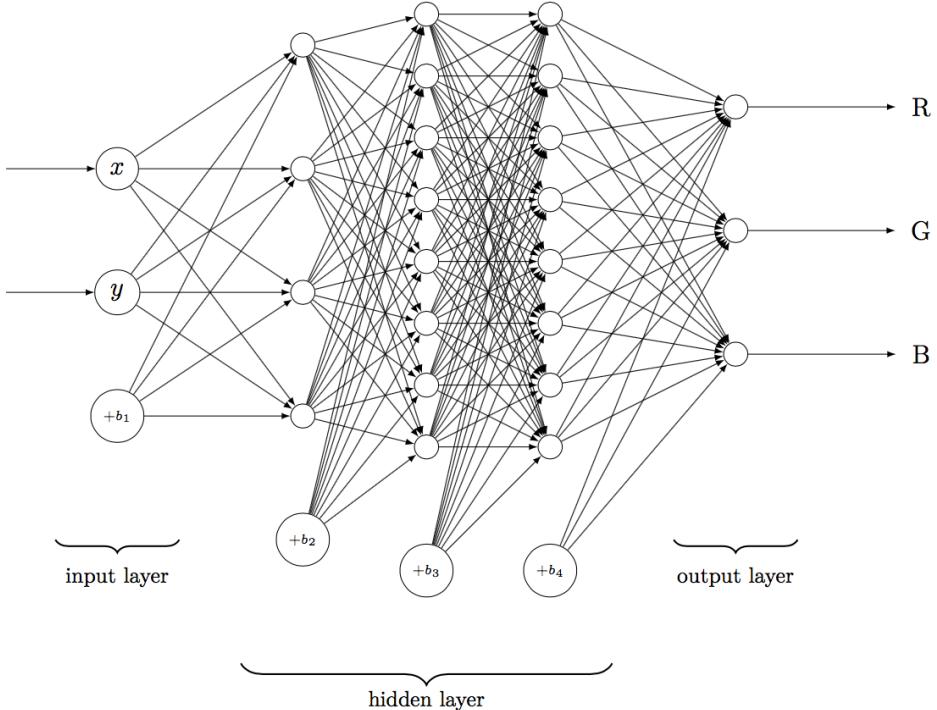


Figure 9: An example neural network that learns to generate an image given x and y pixel position inputs. The inputs are the x and y coordinates of the pixel, and the outputs are the three red-green-blue (RGB) values that define the color of a pixel.

In image analysis, a specific type of neural network termed the convolutional neural network is often used. Convolutional neural networks are loosely based on the structure of the animal visual cortex, which contains overlapping receptive fields. Similarly, in the convolutional neural network architecture, the inputs to a neuron are shared, convolving the input image. In a convolutional layer, the inputs are processed using filters, which are the two-dimensional representations of the weights used by each unit, or kernel, of the layer. These filters are gradually adjusted during training to recognize specific features in the input. The results of the convolution are then processed using a pooling method (Figure 10) that produces a more abstract representation. In deep learning, convolutional and pooling layers are often coupled in a sequence such that higher level features may be learned from the data. For example, while the first layer of convolution may be trained to recognize edges, later layers may recognize shapes and eventually discrete objects. The primary advantage of convolutional neural networks lie in the fact that they are locally

scale-invariant: convolution allows a feature of the image to be recognized regardless of position.

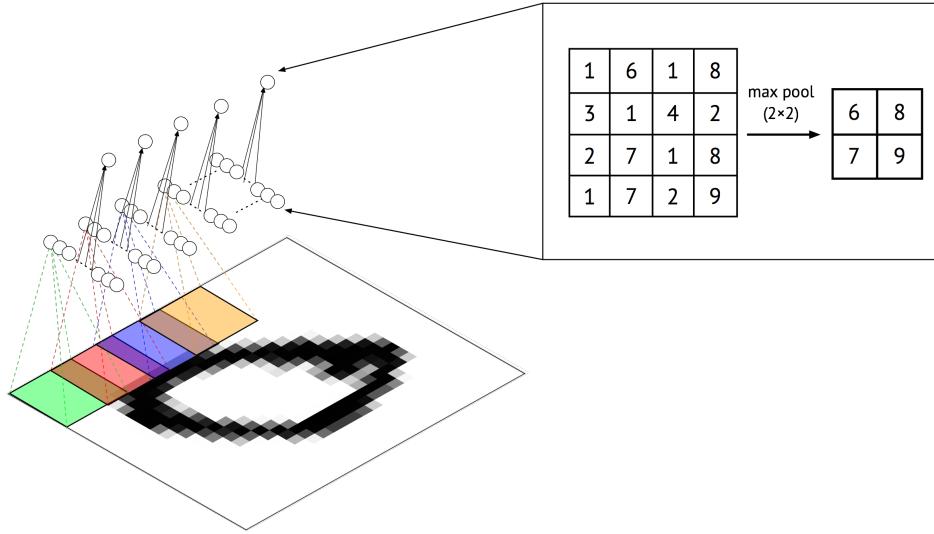


Figure 10: An example of a simple convolutional network designed for digit classification of the MNIST dataset.

Following convolution and pooling layers, a traditional dense layer of neurons is implemented to classify the features obtained (Figure 11). Such networks are currently the state-of-the-art in nearly all image classification tasks.

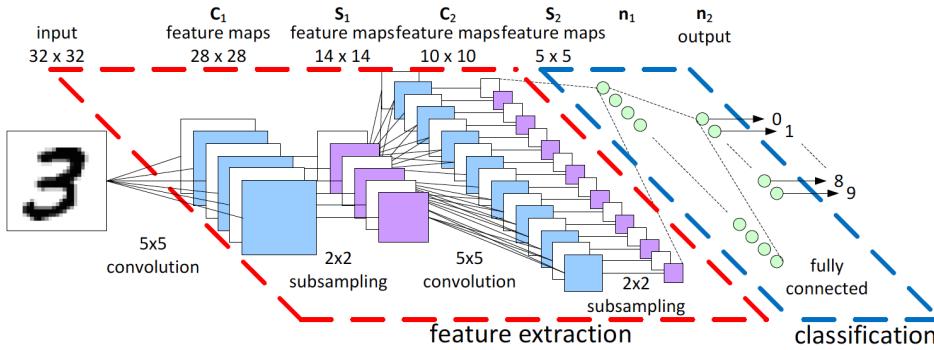


Figure 11: An example of a multi-layer CNN for classifying handwritten digits from the MNIST dataset. The feature extraction section repeatedly convolves the input image to higher level features, which are fed as inputs to the classification stage. In this case, the classification stage is composed of two fully connected layers, with the final layer outputting the predicted digit classification (Peemen et al., 2011).

2.8 Unsupervised learning

Machine learning can be divided into the two subfields of unsupervised and supervised learning. The primary difference between the two is in the training of the model: whereas supervised learning trains the model using a predefined set of classifications, or *labels*, unsupervised learning trains the model without such guidance. Figure 12 gives a basic illustration of the difference between the two: in the left diagram, the datapoints are labeled as being red hexagons or blue triangles. With these labels, one could train supervised programs such as classifiers that predict the label of a data point given its x-y position. On the figure on the right, however, none of the datapoints are labeled (all of them are red hexagons). For these data, one could apply an unsupervised method such as a clustering analysis program to group datapoints by how similar they are to each other. This allows researchers to make inferences regarding patterns and structures in the data.

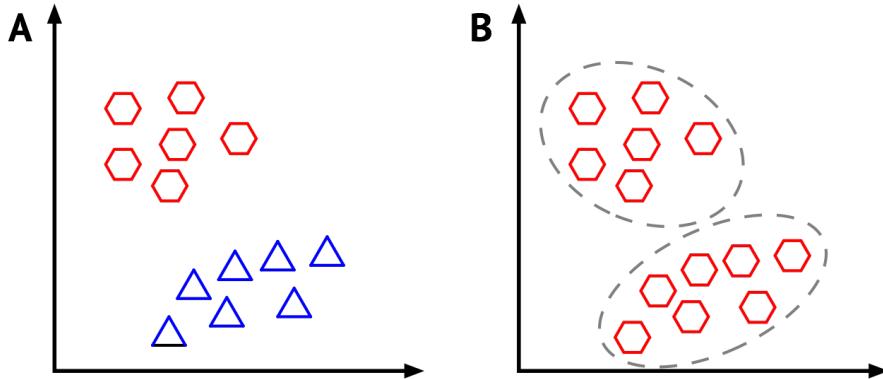


Figure 12: An illustration of the primary difference between supervised (**A**) and unsupervised (**B**) learning.

Clustering is an essential method of unsupervised learning. Clustering analysis is often performed following dimensionality reduction methods that allow the data to be reduced to a simple two or three-dimensional representation while maintaining the original variance and structure in the data. Algorithms such as K-means, DBSCAN, and complete linkage clustering may then be employed to calculate clusters. In image classification tasks, classification and dimensionality reduction allow the learned results

of a neural network to be visualized (Figure 13). Furthermore, a convolutional neural network may be directly coupled with a clustering algorithm as a loss function in order to gradually produce high quality clusters as the network is trained (Yang et al., 2016).

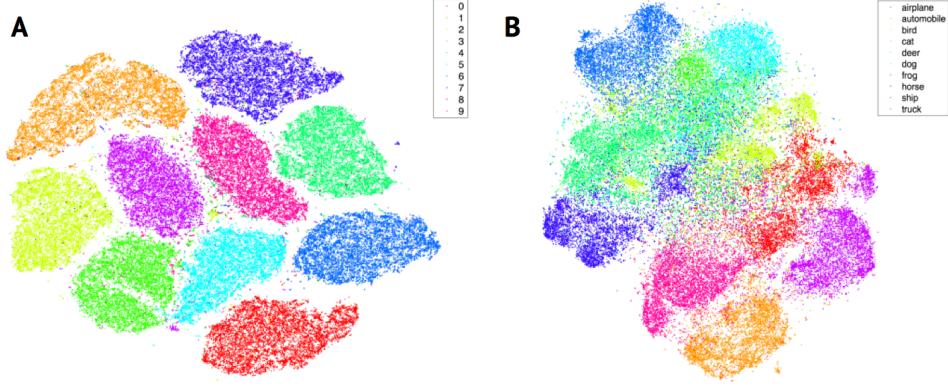


Figure 13: The results of a *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) dimensionality reduction on a clustering of the MNIST (**A**) and CIFAR-10 (**B**) datasets (van der Maaten, 2014).

3 Research plan

3.1 Researchable question

Are there significant differences between patterns of gene expression in developing mouse and human central nervous systems?

3.2 Hypothesis

It is hypothesized that there exist significant differences between the gene expression profiles of developing mouse and human central nervous systems.

3.3 Procedure

This project will be entirely performed on a computer. Gene expression profiles for developing mouse and human brains will be first downloaded from the Allen Brain Atlas. Next, the data will be formatted and transformed into matrices of region versus time

expression expression matrices. At this point, there will be two datasets, a developing human dataset and a developing mouse dataset.

For each dataset, a convolutional neural network algorithm will be applied in order to automatically classify genes into discrete clusters based on patterns of expression. The relative positions of genes in the human and mouse clusterings will then be examined. A different clustering would suggest that the respective transcriptomic landscape is different. The gene expression clusterings may be applied to highlight discrete drug-gene interactions in human versus mouse brains. Potential differing expressions may account for anatomical differences in mouse and human brains. Time permitting, the similarity of expression patterns in cerebral organoids will also be considered. The insights from this project may also be applied to single-cell cancer sequencing, where clusterings may be used to cluster and identify stages of cancer cell development.

4 Methodology

5 Results

6 Discussion

7 Conclusions

8 Limitations

9 Extensions

10 Acknowledgments

References

Angerer, L. M. and Angerer, R. C. (1991). In Situ Hybridization—A Guided Tour.

Toxicology Methods, 1(1):2–29.

Bakken, T. E., Miller, J. A., Luo, R., Bernard, A., Bennett, J. L., Lee, C.-K., Bertagnolli, D., Parikshak, N. N., Smith, K. A., Sunkin, S. M., Amaral, D. G., Geschwind, D. H., and Lein, E. S. (2015). Spatiotemporal dynamics of the postnatal developing primate brain transcriptome. *Hum. Mol. Genet.*, 24(15):4327–4339.

Bisceglia, N. (2010). *Regulation of gene expression*. Nature Education.

Burns, T. C., Li, M. D., Mehta, S., Awad, A. J., and Morgan, A. A. (2015). Mouse models rarely mimic the transcriptome of human neurodegenerative diseases: A systematic bioinformatics-based critique of preclinical models. *European Journal of Pharmacology*, 759:101–117.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nat Biotech*, 23(12):1499–1501.

Imaizumi, Y. and Okano, H. (2014). Modeling human neurological disorders with induced pluripotent stem cells. *Journal of neurochemistry*, 129 3:388–99.

Lancaster, M. A. and Knoblich, J. A. (2014). Generation of cerebral organoids from human pluripotent stem cells. *Nat Protoc*, 9(10):2329–2340.

Lin, S., Lin, Y., Nery, J. R., Urich, M. A., Breschi, A., Davis, C. A., Dobin, A., Zaleski, C., Beer, M. A., Chapman, W. C., Gingeras, T. R., Ecker, J. R., and Snyder, M. P. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, 111(48):17224–17229.

Miller, J., Ding, S., Sunkin, S., Smith, K., Ng, L., Szafer, A., Ebbert, A., Riley, Z., Royall, J., Aiona, K., Arnold, J., Bennet, C., Bertagnolli, D., Brouner, K., Butler, S., Caldejon, S., Carey, A., Cuhaciyan, C., Dalley, R., Dee, N., Dolbeare, T., Facer, B., Feng, D., Fliss, T., Gee, G., Goldy, J., Gourley, L., Gregor, B., Gu, G., Howard, R., Jochim, J., Kuan, C., Lau, C., Lee, C., Lee, F., Lemon, T., Lesnar, P., McMurray, B., Mastan, N., Mosqueda, N., Naluai-Cecchini, T., Ngo, N., Nyhus, J., Oldre, A., Olson, E., Parente, J., Parker, P., Parry, S., Stevens, A., Pletikos, M., Reding, M., Roll, K., Sandman, D., Sarreal, M., Shapouri, S., Shapovalova, N., Shen, E., Sjoquist, N., Slaughterbeck, C., Smith, M., Sodt, A., Williams, D., Züller, L., Fischl, B., Gerstein, M., Geschwind, D., Glass, I., Hawrylycz, M., Hevner, R., Huang, H., Jones, A., Knowles, J., Levitt, P., Phillips, J., Sestan, N., Wohnoutka, P., Dang, C., Bernard, A., Hohmann, J., and Lein, E. (2014). Transcriptional landscape of the prenatal human brain. *Nature*, 508:199–206.

Muller-Tidow, C., Metzelder, S. K., Buerger, H., Packeisen, J., Ganser, A., Heil, G., Kugler, K., Adiguzel, G., Schwable, J., Steffen, B., Ludwig, W.-D., Heinecke, A.,

- Buchner, T., Berdel, W. E., and Serve, H. (2004). Expression of the p14arf tumor suppressor predicts survival in acute myeloid leukemia. *Leukemia*, 18(4):720–726.
- Nguyen, L., Wang, Y., and Nikolakopoulou, A. (2015). *Cerebral organoid derived from ALS patient stem cells*. University of Southern California.
- Peemen, M., Mesman, B., and Corporaal, H. (2011). Speed Sign Detection and Recognition by Convolutional Neural Networks. *International Automotive Congress*.
- Simões-Costa, M. and Bronner, M. E. (2015). Establishing neural crest identity: a gene regulatory recipe. *Development*, 142(2):242–257.
- Thompson, C. L., Ng, L., Menon, V., Martinez, S., Lee, C.-K., Glattfelder, K., Sunkin, S. M., Henry, A., Lau, C., Dang, C., Garcia-Lopez, R., Martinez-Ferre, A., Pombero, A., Rubenstein, J. L., Wakeman, W. B., Hohmann, J., Dee, N., Sodt, A. J., Young, R., Smith, K., Nguyen, T.-N., Kidney, J., Kuan, L., Jeromin, A., Kaykas, A., Miller, J., Page, D., Orta, G., Bernard, A., Riley, Z., Smith, S., Wohnoutka, P., Hawrylycz, M. J., Puelles, L., and Jones, A. R. (2014). A High-Resolution Spatiotemporal Atlas of Gene Expression of the Developing Mouse Brain. *Neuron*, 83(2):309–323.
- van der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, (15):1–21.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.
- Yang, J., Parikh, D., and Batra, D. (2016). Joint Unsupervised Learning of Deep Representations and Image Clusters. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.