BIOINFORMATICS FOR NEUROANATOMICAL CONNECTIVITY

by

Leon French

B.Sc., The University of Windsor, 2003

M. Sc., The University of Windsor, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

January 2012

Abstract

Neuroscience research is increasingly dependent on bringing together large amounts of data collected at the molecular, anatomical, functional and behavioural levels. This data is disseminated in scientific articles and large online databases. I utilized these large resources to study the wiring diagram of the brain or 'connectome'. The aims of this thesis were to automatically collect large amounts of connectivity knowledge and to characterize relationships between connectivity and gene expression in the rodent brain. To extract the knowledge embedded in the neuroscience literature I created the first corpus of neuroscience abstracts annotated for brain regions and their connections. These connections describe long distance or macroconnectivity between brain regions. The collection of over 1,300 abstracts allowed accurate training of machine learning classifiers that mark brain region mentions (76% recall at 81% precision) and neuroanatomical connections between regions (50% sentence level recall at 70% precision). By automatically extracting connectivity statements from the Journal of Comparative Neurology I generated a literature based connectome of over 28,000 connections. Evaluations revealed that a large number of brain region descriptions are not found in existing lexicons. To address this challenge I developed novel methods that allow mapping of brain region terms to enclosing structures. To further study the connectome I moved from scientific articles to large online databases. By employing resources for gene expression and connectivity I showed that patterns of gene expression correlate with connectivity. First, two spatially anti-correlated patterns of mouse brain gene expression were identified. These signatures are associated with differences in expression of neuronal and oligodendrocyte markers, suggesting they reflect regional differences in

cellular populations. Expression level of these genes is correlated with connectivity degree, with regions expressing the neuron-enriched pattern having more incoming and outgoing connections with other regions. Finally, relationships between profiles of gene expression and connectivity were tested. Specifically, I showed that brain regions with similar expression profiles tend to have similar connectivity profiles. Further, optimized sets of connectivity linked genes are associated with neuronal development, axon guidance and autistic spectrum disorder. This demonstration of text mining and large scale analysis provides new foundations for neuroinformatics.

Preface

Together with my supervisor, Paul Pavlidis, I was responsible for the identification and design of the research program described in this thesis. I was the primary author for every chapter and corresponding publications. My supervisor, Paul Pavlidis contributed study design, supervision, concepts, text and editorial suggestions for all chapters.

An early version of Chapter 1 has been published as a review article (French L, Pavlidis P (2007). Informatics in neuroscience. Briefings in Bioinformatics 8:6. 446-456. doi:10.1093/bib/bbm047). The article has been edited and updated to reflect recent changes in the area of neuroinformatics.

A version of Chapter 2 has been published (French L, Lane S, Xu L and Pavlidis P (2009). Automated recognition of brain region mentions in neuroscience literature. Frontiers in Neuroinformatics. 3:29. doi:10.3389/neuro.11.029.2009). Suzanne Lane and Lydia Xu annotated the corpus for brain region mentions and neuroanatomical connectivity relations used in Chapter 2, 3 and 4.

A version of Chapter 3 has been accepted and published online (French L and Pavlidis P (2011). Using text mining to link journal articles to neuroanatomical databases. The Journal of Comparative Neurology. doi:10.1002/cne.23012, Copyright © 2011 Wiley-Liss, Inc.).

A version of Chapter 5 has been published (French L, Tan PPC and Pavlidis P (2011) Large-scale analysis of gene expression and connectivity in the rodent brain: insights through data

integration. Frontiers in Neuroinformatics. doi:10.3389/fninf.2011.00012). Patrick Tan contributed significantly to Chapter 5 and is a co-author of the corresponding publication. Specifically, Patrick contributed software, figures, concepts and editorial suggestions for Chapter 5.

A version of Chapter 6 has been published (French L, Pavlidis P (2011) Relationships between Gene Expression and Brain Wiring in the Adult Rodent Brain. PLoS Computational Biology 7(1): e1001049. doi:10.1371/journal.pcbi.1001049).

Table of Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

ABA	Allen Brain Atlas
ABAMS	Combination of ABA and BAMS abbreviations
AGEA	Anatomic Gene Expression Atlas
ASD	Autistic Spectrum Disorder
AUC	Area Under ROC Curve
BAMS	Brain Architecture Management System
BioCreAtIvE	Critical Assessment of Information Extraction Systems in Biology
BIRN	Biomedical Informatics Research Network
BIRNLex	Biomedical Informatics Research Network Lexicon
CoCoMac	Collations of Connectivity Data on the Macaque Brain
CRF	Condition Random Field
DTI	Diffusion Tensor Imaging
GATE	The General Architecture for Text Engineering
GO	Gene Ontology
INCF	International Neuroinformatics Coordinating Facility
ISH	*In Situ* Hybridization
LOOM	Lexical OWL Ontology Matcher
NE	Neuron Enriched
NeuroLex	Neuroscience Lexicon
NIF	Neuroscience Information Framework

NIFSTD	Neuroscience Information Framework Standard Ontology
OE	Oligodendrocyte Enriched
OWL	Web Ontology Language
RDF	Resource Description Framework
ROC	Receiver Operating Characteristic
SL	Shallow Linguistic Kernel
ρ	Spearman Rank Correlation Coefficient

Acknowledgements

I thank Paul Pavlidis, my graduate supervisor for his professional and kind guidance during my studies. He was always there to quickly address all my requests, questions and insecurities. Thanks go to my committee – Dan Goldowitz, Mark Wilkinson and Wyeth Wasserman. Their questions and advice successfully guided my transition from computer scientist to bioinformatician.

I am greatly indebted to the providers of the two main data sets relied upon in this thesis, the Allen Institute for Brain Research and The Brain Architecture Centre. In particular I thank Mike Hawrylycz, Lydia Ng and Susan Sunkin for providing the expression data and answering questions. I acknowledge John Barkley for creation of the semantic web version of the BAMS database. I thank Maryanne Martone, Stephen Larson and Anita Bandrowski for facilitating additions to NeuroLex. I acknowledge Amir Ghazvinian for kindly providing the source code of the LOOM simple mapping matcher. Dr. Cladia Krebbs provided guidance and instruction that significantly improved my understanding of neuroanatomy. I thank Suzanne Lane, Lydia Xu, and Tamryn Law for their hard work of annotating many abstracts and experiments. In particular I am grateful to Suzanne Lane for her months of tedious curation that created the large corpus I required for machine learning. I enjoyed and appreciated the research done with Patrick Tan; his hard work and curiosity helped me write Chapter 5 in record time.

I owe several people for their attention and review of my work. In particular, I appreciate the valuable comments from Kevin She on Chapter 5. In addition, I thank Michael

Hawrylycz, Susan Sunkin, Tim O'Conner and Warren Cheung for critical review of Chapter 6. I thank Jesse Gillis for insightful thoughts on all chapters. His ability to understand diverse topics provided comments that led to several clear improvements of my research.

Thank you to my fellow labmates and friends at CHiBi for helping me through the process. They all have helped this work move forward; in particular I thank Kelsey Hammer, Meeta Mistry, Warren Cheung, Raymond Lim, Chris Thachuk, Luke McCarthy, Thomas Sierocinski, Evan Morien and Ben Vandervalk. I thank Benjamin Good, the Griffith Brothers and Carri-Lyn Mead for their seasoned advice.

I am grateful to the National Science and Engineering Research Council of Canada for awarding me three years of funding.

Finally, heartfelt thanks are owed to my partner Nellie Chang and my parents for their understanding and support throughout the process.

Chapter 1: Introduction[1]

The brain is perhaps the most complex object known, and deciphering its workings is an enduring challenge. This complexity has limited discovery of causes and cures for many devastating brain disorders. However, over 2,500 years of analysis have led to a huge accumulation of information about the brain at levels ranging from molecular to the anatomical. New techniques applied over the past few decades are generating ever more detailed and comprehensive data sets. Recent examples include whole brain tractography and genome-wide expression atlases. These enormous datasets combined with the accumulated knowledge from past experiments require powerful methods for exploiting this amount and diversity of data. This thesis is about a small subset of this problem, focusing on two types of information that are both important and available in relatively global forms, namely macroconnectivity and gene expression patterns. I sought to address two questions: Can our current databases of connectivity be expanded using computational approaches? And how does gene expression relate to brain connectivity? The first question is an informatics one, addressing the gap between the vast literature on neuroanatomy and our ability to use it efficiently. The second question is motivated by biology, namely the relationship between the genome (and its genetic variation) and the structure and function of the brain. One specific motivating question is, why are so many genes expressed in complex patterns in the brain? In

[1] A previous version of this chapter has been published. French L, Pavlidis P (2007). Informatics in neuroscience. Briefings in Bioinformatics 8:6. 446-456. doi:10.1093/bib/bbm047). The article has been edited and updated to reflect recent changes in the area of neuroinformatics.

this introduction, I expand on these motivations, provide background on the state of the art, and put my work in the context of the nascent field of neuroinformatics. It is my hope that my research contributes to a better understanding of normal and abnormal brain function, and leads to new questions, resources, methodologies, experiments and analyses.

Neuroanatomical connectivity is a unifying theme in this thesis. At the cellular level these communication links are defined by trillions of synapses between billions of neurons and are fundamental to our understanding of the nervous system. At the macro level these links join together to form pathways that connect neuroanatomically defined brain regions. Recently, there has been a push to map all connectivity or the "connectome" of the human brain using neuroanatomical imaging techniques (2005). On the other hand, large amounts of connectivity information are already present in the literature, but this resource has not been fully exploited. Currently the majority of this information is fragmented across many reports and is not accessible for large scale analyses. We attempt to address this need, and furthermore explore a new area in the interpretation of gene expression patterns in the brain in light of connectivity data. The wiring diagram of the brain is only poorly understood. In part this is due to the complexity of the brain and the difficulty in collecting data. This complexity is also apparent in the mouse brain transcriptome with varied spatial gene expression patterns that have not been linked to specific function or structure. While many individual genes function in neurotransmission and forming connections, it is not clear how global molecular signatures relate to macro connectivity in the adult brain. For both the transcriptome and connectome, we suggest that informatics technologies can be applied to existing knowledge to make new discoveries and guide further experimentation. This

introduction reviews the challenges, methods and resources for investigating neuroscience from an informatics perspective.

1.1 Dissertation Overview

In this dissertation, I research new strategies to collect and analyze neuroanatomical connectivity data. In particular I focus on macro connectivity that describes connections between brain regions. To collect connectivity data I apply existing bioinformatics techniques to mine the neuroscience literature for connectivity statements. To analyze connectivity data I integrate several heterogeneous neuroscience data sources. These two objectives form the WhiteText project (Chapters 2-4) and ABAMS project (Chapters 5-6).

The WhiteText project addresses the following questions:

1. How accurately can neuroanatomical information, including connectivity be automatically or manually extracted from neuroscience literature?

2. What lexicographic, linguistic and semantic features are useful for extraction?

3. How much connectivity data is available in neuroscience abstracts?

The project aims to usefully apply text mining methods to brain connectivity. We will adapt and extend text-mining approaches previously used to analyze protein networks to extract data on connectivity of brain regions from free-text abstracts. We aim to manually generate a corpus for training and evaluation purposes. In addition to developing algorithms, we aim to

create a database of 15,000 connections among 1,000 mammalian brain regions. The database will become a tool for testing hypotheses about brain function, structure and development. The resulting resources will be made publicly available.

To understand connectivity we analyze a genome-wide atlas of the adult mouse brain that has characterized many heterogeneous expression patterns (Lein et al., 2007). Global relationships between these patterns and brain organization are suspected but few cases exist, especially beyond development. While expression patterns of single genes have been linked to specific structure, genome wide screens for associations between gene expression, macroconnectivity and cellular distributions have not been previously performed in the rodent brain. Such analyses may reveal genes that play potential roles in developing, maintaining and repairing brain architecture. Conversely, insight into neuroanatomy at the brain region level can be derived from functional annotations at the gene level.

In the context of the adult rodent brain, the ABAMS project addresses the following questions:

1. Can complex spatial expression patterns be explained by differences in cellular populations or number of neuroanatomical connections?

2. To what degree are expression and connectivity profiles correlated across many brain regions? Which genes carry the highest amount of information about connectivity in their expression levels? What are their functional associations?

3. To what degree are spatial and connectivity profiles correlated across many brain regions?

The objective of the project is to find and study relationships between gene expression and neuroanatomical connectivity. We hypothesize there is a statistical relationship between the connections and gene expression levels of individual brain regions. Based on this hypothesis we aimed to specify lists of genes that carry information about a given brain region's connectivity profile. This list of genes can aid interpretation of future experiments by suggesting connectivity-related functions, and candidate genes for understanding brain function in health and disease. Our methodology will focus on statistical data mining and heuristic search methods designed to find significant patterns in high-dimensional data.

1.2 Literature Review

The application of informatics to neuroscience goes far beyond "traditional" bioinformatics modalities such as DNA sequences. In this section I describe how informatics is being used to study the nervous system at multiple levels, spanning scales from molecules to behaviour.

Neuroscience is a rich source of interesting computational and informatics problems and opportunities. These problems encompass a good deal of "traditional" bioinformatics (e.g., sequence analysis), applied to the neuroscience domain. In addition, perhaps more than any other field, neuroscience has been applying computation and informatics to domain-specific problems, giving rise to the term "neuroinformatics". Neuroinformatics includes the development of databases, standards, tools and models, and the development of simulations

and analytical techniques, spanning all levels of nervous system organization (from molecules to behaviour; Table 1). Much of the interest in neuroinformatics comes from the diverse types of neuroscience research and how they might be linked more effectively using informatics technologies.

I provide an overview of neuroinformatics, biased somewhat towards the viewpoint of practitioners of bioinformatics who are outside of neuroscience. Therefore we focus our attention on only a subset of areas within neuroinformatics. The large field of nervous system modeling and simulation is not reviewed, so I point readers to further resources covering models of single neurons (Crasto et al., 2007a), networks (Brette et al., 2007), and sensory/information processing (Destexhe and Contreras, 2006). In addition, neuroimaging informatics has been reviewed in detail recently (Brinkley and Rosse, 2002;Toga, 2002;Nielsen et al., 2006;Van Horn and Toga, 2009). Thus my focus is on other types of neuroscience data and knowledge databases, efforts towards integrating knowledge across domains, and especially on the analysis of the nervous system at the genetic, cell and molecular level. A summary of informatics resources covered in this review is given in Table 2.

1.2.1 Initiatives

Recent large scale initiatives motivate our neuroinformatics research. They provide ontologies for linking extracted data, portals for finding resources and communities for sharing ideas. The first is the Human Brain Project (HPB, http://www.nimh.nih.gov/neuroinformatics) started in 1993, based on recommendations

developed starting in 1989 (Shepherd et al., 1998). With leadership from the National Institute of Mental Health and other NIH institutes, HBP provided funding and guidance to many of the neuroinformatics projects mentioned in this chapter (Koslow, 2005). The HBP has been succeeded by the NIH Blueprint for Neuroscience research as neuroinformatics is increasingly folded into the mainstream of neuroscience and informatics (Huerta et al., 2006). The Blueprint for Neuroscience research is a large collaborative NIH effort for creating resources of general utility to neuroscience research (Baughman et al., 2006). The recently established International Neuroinformatics Coordinating Facility (INCF), funded by the EU and based in Stockholm, Sweden, with "nodes" in many European countries as well as the US and Japan, aims to "foster international activities in neuroinformatics", and is another signal of the seriousness with which the field is taking informatics (Amari et al., 2002). The INCF was founded in response to a report of the Organization for Economic Co-operation and Development (OECD) (Group, 2002).

The Society for Neuroscience (SfN) formed the Brain Information Group task force in 2003 and later the SfN Neuroinformatics committee. These were formed to examine the informatics needs of neuroscience and promote existing resources (Van Essen, 2007). The earliest result was the Neuroscience Database gateway (NDB). NDB organized 178 databases into five main categories, and 15 classes. Recently a consortium based effort funded by the National Institutes of Health constructed a successor to NDB, the Neuroscience Information Framework (http://www.neuinfo.org; NIF). In addition to cataloguing neuroscience resources NIF provides a dynamic portal to the resources and the data contained within them (Gardner et al., 2008). NIF is notable for its extensive efforts to

form accessible standards, services, ontologies and tools.

In addition to these organizational efforts, a large-scale project that stands out in neuroinformatics for its scale and scope is the NIH-backed Biomedical Informatics Research Network (BIRN) (Martone et al., 2004;Helmer et al., 2011). BIRN is focused on neuroinformatics, and emphasizes brain imaging in humans and mice, and acts as a "test bed for development of hardware, software, and protocols to effectively share and mine data in a site-independent manner for both basic and clinical research". BIRN is a network of research groups that at this writing involves work from at least two dozen laboratories across the United States and the United Kingdom. BIRN is discussed below in the context of several specific areas of neuroinformatics.

1.2.2 Towards integration

Because neuroscience data is highly heterogeneous, complex, and voluminous, it has been recognized that interoperation of tools and databases will be required to make the best use of available resources (Insel et al., 2003). This is evidenced throughout the thesis, especially in Chapters 5 and 6 which derive new insight from combined datasets. As in other areas of biology, efforts to standardize data representations and interfaces have been increasing, and neuroscience can clearly learn lessons from looking at how standards have developed in other fields of informatics. One side-effect of the interest in neuroinformatics and neuroscience data is the recognition that integration requires data sharing, and the subject has been widely discussed by neuroinformatics researchers (Koslow, 2000;Gardner et al.,

2001;Eckersley et al., 2003;Insel et al., 2003;Ascoli, 2006) .

1.2.3 Ontologies and vocabularies

Ontologies and controlled vocabularies are an important resource to enable informatics. Adherence to a specific terminology and/or data model can be constraining, but also greatly eases interoperability. One only has to look at the wide cross-referencing of the Gene Ontology (GO) (Ashburner et al., 2000) to see the power of standardized terminologies. Another example is BioPax, developed by the biological pathway database community to promote sharing of molecular pathway data (2006). BioPax has been widely adopted, and currently several pathway and interaction databases are available in BioPax format with more converted by third parties (Baitaluk et al., 2006;Kotecha N, 2006).

Currently many neuroscience databases use their own vocabularies for neuron, anatomical region and receptor types, but this situation is likely to change rapidly. For example, the BIRN includes an ontology "taskforce", and developed BIRNLex for use in BIRN projects (Martone et al., 2004), an ontology containing concepts from neuroanatomy, molecular species, experimental design and cognitive processes. BIRNLex terms are taken from existing resources whenever possible, with direct mappings given. Recently BIRNLex has been merged into the NIF standardized (NIFSTD) ontology (Bug et al., 2008). It is our hope that NIFSTD (or something like it) will have a wide impact and be adopted by other projects as a *de facto* standard. However, like many efforts to develop standards, it is difficult to please everybody, so it remains to be seen if a single standard can emerge soon enough.

Neuroanatomy is an example of an area where multiple standards have emerged (Bowden et al., 2007). There are two established nomenclatures for the rat brain (Swanson, 1999;Paxinos and Watson, 2007), and three for the mouse (Hof et al., 2000;Dong, 2007;Paxinos and Watson, 2007). Thankfully, mappings exist between the terminologies (Stephan et al., 2000;Bowden and Dubach, 2003;Bota and Swanson, 2010). These atlases provide hierarchical structured vocabularies. Usually a child term refers to a region that is volumetrically contained in the region described by the parent term (e.g., prefrontal cortex is part of the cortex). The most widely accepted nomenclature is NeuroNames which contains over 1,900 structures linked to over 7,500 terms describing the human, rodent and macaque brain (Bowden and Dubach, 2003). NeuroNames has been integrated into the Foundational Model of Anatomy, NIFSTD, and the Unified Medical Language System (Hole and Srinivasan, 2003). A web based interface to NeuroNames is provided by BrainInfo (Bowden and Dubach, 2002). For a given brain region external links are provided for connectivity, literature, cytoarchitecture, and gene expression.

The Brain Markup Language (BrainML) was developed as a set of XML schemas for exchange of neuroscience data (Gardner et al., 2001). BrainML encompasses representations of experimental protocols and designs, electrophysiology, measurement units, and other aspects important to representing neuroscience data, and forms a “base model” that is used to create additional specific components, such as describing animal experimental subjects. While BrainML does not yet appear to have undergone widespread adoption, as mentioned earlier, it was used to develop the Neuroscience Information Framework.

While purpose-built ontologies are clearly needed, many neuroscience concepts are contained in existing ontologies and terminologies that are not necessarily designed specifically for neuroscience. For example, a search for "hippocampus" at the National Center for Biomedical Ontology Bioportal (Noy et al., 2009) reveals 642 terms across 42 ontologies. The Gene Ontology also contains many neuroscience concepts such as ‘hippocampus development’ (biological process), ‘GABA receptor activity’ (molecular function) and ‘axon’ (cellular component) (Ashburner et al., 2000). Ideally new terminologies will meld seamlessly as possible with these existing terminologies and avoid reinventing the wheel.

1.2.4 Databases of molecules and cells

Several databases that are focused on collating data about specific neuron types and molecules motivate our work to database connectivity. They provide integration points and pioneering efforts into semi-automated curation. The most extensive purpose-built cell and molecular neuroscience knowledgebase is SenseLab, which includes seven databases covering pharmacology, ion channels, cell properties, olfactory pathways and neuronal models (Crasto et al., 2007a). Within the neuronal databases (CellPropDB, NeuronDB) entries are linked across scales of brain region, neuron, cell compartment, ion channel and receptor. Information about odorant molecules linked to receptors and maps of the olfactory bulb are provided in OdorDB, ORDB and OdorMapDB respectively. Links are provided to the Cell Centered Database (CCDB, consisting of cellular and subcellular imaging data), PubMed, GenBank and Ensembl. SenseLab increasingly spans a wide array of domains –

models, genetics, proteomics and imaging. SenseLab is largely curated manually, with some assistance from automated text-mining methods (Crasto et al., 2002) (Crasto et al., 2003).

1.2.5 Text mining in neuroscience

Text mining is the process of analyzing text to extract entities and relationships. It is usually performed on large collections of documents (a corpus). The field is closely related to natural language processing which seeks to computationally interpret human language. In this thesis we use several natural language tools for our tasks. One example is part of speech tagging where part-of-speech tags (noun or verb for example) are marked. Further examples are stemming (determine base form or stem of a word) or tokenization (determine end of a sentence). Although these tasks seem simple they are difficult to computationally perform, especially in biomedical text. The main problem is that there are not enough samples to statistically learn the complex rules (also known as the sparse data problem). Procedures like part-of-speech tagging and stemming ameliorate the problem by abstracting words into smaller categories. Ambiguity is another main challenge which especially limits text mining attempts to extract facts from text. This is clear in a seemingly simple task: extract and expand abbreviations in biomedical literature. Unfortunately, over 80% of biomedical abbreviations are ambiguous with over 16 expansions on average (Ao and Takagi, 2005). When both the short and long form of the abbreviation are provided in a document it is possible to correctly connect the two in roughly 95% of cases with 320 search patterns. The sentence boundary detection task of marking when a sentence begins and ends is also limited by ambiguity, problem examples include "congenic strains B10.D2/nSnJ" and "Hendricks et

al." (Xuan et al., 2007). Further ambiguity is observed when sentences are further segmented into words. Several methods can solve these tasks with reasonable accuracy and are applied in the first stages of most text mining systems.

Given the words and their annotations, rule-based or statistical aproaches are used to determine if they are names of important entities (named entity recognition). Example entities include genes, diseases, species and brain regions. Rule-based systems are derived from general knowledge about the specific domain and its entities. In contrast, statistical approaches employ machine learning tools to classify based on examples. After named entity recogntion, relationship extraction is required to text mine valuable information such as which genes are specifically expressed in a brain region. In this thesis we focus on extracting mentions of brain regions and statements that describe their connections.

Several projects have explored application of text mining techniques to neuroscience literature. Although the goals vary, the results are limited by standardized datasets for evaluating the methods. Textpresso for Neuroscience uses a text-mining approach to provide a neuroscience-focused search tool, indexing over 15,000 abstracts and full papers from the biomedical literature (Muller et al., 2008). The data in Textpresso is organized using a customized ontology based largely on selected terms from the Gene Ontology, combined with domain-specific concepts such as brain regions (Muller et al., 2004). The developers of NeuroExtract (Crasto et al., 2007b) rapidly built a neuroscience-focused database by searching for "brain" and "central nervous system" in three major bioinformatics resources (SwissProt, the Gene Expression Omnibus and the Protein Databank). These results and

associated abstracts were then filtered for 71 neuroscience related keywords (from cell types to brain regions). The authors show that their system returns more results than a keyword search performed on source websites (Crasto et al., 2007b). Similarly, the Synapse Database (SynDB), a database of genes involved in synaptic function, was populated by performing keyword searches on Interpro and UniProt databases followed by automatic then manual screening (Zhang et al., 2007). SynDB contains over 14,000 protein entries organized into a purpose-built 177-concept synapse ontology. While SynDB does not cross-reference to any neuroscience related databases, it provides links to eighteen general bioinformatics resources. SynDB has an extensive web browser interface, allowing a researcher to browse proteins using the ontology, functional categories, protein domains, species, chromosomal location and protein families. While these resources are still new, they represent efforts to make access to neuroscience knowledge easier and faster.

A theme running through many cell and molecular databases is the use of information extraction from the biomedical literature. Literature mining is an active area in bioinformatics (for reviews see special issue of Briefings in Bioinformatics (Koehler, 2005)) and there are clearly additional interesting opportunities to apply natural language processing in domain-focused ways. Text mining shows up in our discussion of several other data modalities in the next sections.

1.2.6 Connectivity and connectomics

Several existing neuroinformatics resources focus on connectivity in a limited set of organisms. Although restricted, these resources provide valuable data for integration and evaluation throughout this thesis.

Brain connectivity can be thought of as a structural property of neurons (cell A connects to cell B) or of anatomical regions (inferior olive projects to the cerebellum). A “connectome” refers to a comprehensive map of connections at one of these scales. Measuring connectivity has a long history in neuroscience, and efforts to create exhaustive maps and databases are not new (Sporns et al., 2005). However, due to the difficulty of collecting connectivity data, the only complete nervous system connectivity map or connectome is for *C. elegans* (White et al., 1986). Clearly, having a good-quality map of human brain connectivity would serve as a cornerstone for understanding brain function and structure. The Human Connectome Project has recently begun to achieve this goal by magnetic resonance imaging 1,200 healthy adult brains (Marcus et al., 2011).

One current application of connectivity is in the development of models. For example, connectivity data has been used to create models of the relatively well-studied primate visual system (Itti and Koch, 2001;Serre et al., 2007). As an example of an ambitious modeling project that will need connectivity information, the Blue Brain project envisions computational modeling of the entire brain (Markram, 2006).

Currently connectivity data is sparse for humans so current databases focus on model

organisms. The Brain Architecture Management System (BAMS) focuses on connectivity in the rat brain, with over 40,000 records (Bota et al., 2003;, 2005). CoCoMac is a searchable database of connectivity data from over 400 literature reports in the Macaque monkey (Kotter, 2004). A related database, CoCoDat, contains detailed microcircuitry reports (Dyhrfjeld-Johnsen et al., 2005). Finally, the complete wiring diagram of the *C. elegans* nervous system can be downloaded from http://www.wormatlas.org/ (Chen et al., 2006).

An interesting experimental project to populate connectivity databases using natural language processing is part of the Neuroscholar project (Burns and Cheng, 2006;Burns et al., 2007). Neuroscholar is able to classify text with respect to several experimental parameters of interest in tract tracing studies with 80% precision (Burns et al., 2007). Another part of the Neuroscholar project, NeuARt II digitizes analog atlases to create a flexible brain mapping infrastructure (Burns et al., 2006). As currently implemented, Neuroscholar is designed to operate with human supervision to assist manual curation efforts that underlie projects like BAMS and CoCoMac.

There is interest in integrating connectivity data with other modalities. Recently the SenseLab team converted CoCoDat into OWL format, for integration with NeuronDB (Crasto et al., 2007a). BAMS is also involved in integration efforts, and provides links between neuron and cell associated molecules to brain regions.

1.2.7 Functional and morphometric imaging

Brain imaging refers to non- or minimally-invasive technologies for measuring brain anatomy or activity in live animals (often humans), perhaps the best known of which is functional magnetic resonance imaging (fMRI). These technologies provide results at the level of brain regions by providing functional associations and structural descriptions. These brain region linked results are published in the neuroscience literature at an increasing rate and present a potential target for our text mining work. There is extensive interest in making imaging data sharable and comparable for the purposes of archiving and meta-analysis, and in integration of imaging data with other modalities. As mentioned earlier, imaging informatics is a relatively well-developed field and the subject of recent review (Brinkley and Rosse, 2002;Toga, 2002;Nielsen et al., 2006), so we only give the briefest possible overview of this area.

Several repositories and databases of structural and functional MRI images exist, for example fMRIDC (Van Horn et al., 2004). Some systems provide extensive additional analysis tools. The Surface Management System Database (Van Essen, 2009), Brainmap (Laird et al., 2005), NeuroSynth (Yarkoni et al., 2011), and the Brede database (Nielsen et al., 2004) allow visualization of brain locations and searches based on a reference coordinate system (Nielsen and Hansen, 2004). The Brede database provides software and numerous cross references to a variety of bioinformatics resources. Brede entries link to genes, diseases, receptors (via SenseLab) and brain regions (BrainInfo, CoCoMac). The Brede database also provides correlated volumes for each experiment (Nielsen and Hansen, 2004), opening possibilities for

meta-analysis, and uses text mining to link articles to brain activation studies (Nielsen et al., 2004). NeuroSynth is unique in its extensive use of text mining to extract brain activation coordinates from fMRI studies. By parsing result tables of full text papers and associating informative keywords it automatically forms thousands of structure to function relationships (Yarkoni et al., 2011).

A specialized form of magnetic resonance imaging, diffusion tensor imaging (DTI), can be used to generate connectivity maps ("tractography") of living human brains (Le Bihan et al., 2001;Parker, 2004). For example, DTI has been used to describe connections between the thalamus and cortex (Behrens et al., 2003). Since DTI scans the whole brain non-invasively it has the potential to be used to collect connectivity data from large samples of humans and then related to other variables such as genetic variation and psychopathology; this is already an active area of study (Kubicki et al., 2007). Although a few DTI datasets are available online (Evans, 2006;Hermoye et al., 2006), to our knowledge there are no accessible databases of connectivity derived from DTI.

1.2.8 Genetics and gene expression

The wealth of bioinformatics methods and resources at the gene level facilitates our study of the brain. In Chapters 5 and 6 we employ this wealth to characterize relationships between anatomy and gene expression patterns in the rodent brain. It is generally much easier to analyze genes than behaviour or neuroanatomy, and the links between them have frequently been elusive, especially as applied to "higher" organisms. Recent advances in genome analysis (founded on detailed physical and genetic maps) and in expression analyses (e.g.,

using microarrays) have meant that bridging the gap between genotype and phenotype is getting easier, but is still limited by resolution at the organismal level. This is because behaviour and anatomy are highly complex and often thought to be heterogeneous.

To our knowledge the best-developed effort to bridge this gap is GeneNetwork (http://www.genenetwork.org/) (Wang et al., 2003). GeneNetwork uses RNA profiling data from recombinant inbred mice which have been extensively phenotyped (behaviourally and otherwise) and genotyped. Because of the inbred nature of these mice, but the relatively large genetic differences between lines, variability at the phenotypic level can be rapidly related to variability at the sequence level. Thus, using the GeneNetwork website, one can search for loci with variants that correlate with quantitative traits including expression levels (expression quantitative trait loci, eQTL) and behaviour. For example, Korostynski et al. used GeneNetwork to help identify candidate genes for variation in opioid preference between different mouse lines (Korostynski et al., 2006). Additional applications can be found referenced on the GeneNetwork website. We note that GeneNetwork is a part of BIRN and contributes to its goal of studying multi-modal data from mouse models of neurological disorders.

Understanding differences in the genes expressed in different brain regions and neurons has always been of value for generating hypotheses about how the brain works, even when uncoupled from genetic variation in individuals. For example, knowing what neurotransmitters are synthesized in a brain region gives a major clue as to what the neurons there are capable of doing. Spatially and temporally-organized gene expression during

development plays a crucial role in determining the ultimate structure of the nervous system. Besides GeneNetwork, there are two types of resources that have emerged in the analysis of expression in the nervous system: spatially resolved atlases, and expression profiling databases. The latter also include data from other high throughput techniques such as competitive genomic hybridization (CGH) and chromatin immunoprecipitation on microarrays (ChIP-chip).

I make extensive use of the Allen Mouse Brain Atlas (ABA), especially in Chapters 5 and 6. The ABA contains high resolution colorimetric *in situ* RNA hybridization data for most of the known mouse genes, in the adult brain (Lein et al., 2007). The ABA is primarily accessible via a sophisticated web-based graphical interface (Hochheiser and Yanowitz, 2007). The ABA allows searching for genes by similarity of expression patterns (NeuroBLAST), and is making summarized data on expression patterns available for download (Jones et al., 2009). ABA has also contributed a digital mouse brain atlas (Dong, 2007). There are a number of other atlases, which are lower coverage (hundreds to a few thousand genes) but complement ABA with additional features. The joint Brain Gene Expression Map (BGEM) and Gene Expression Nervous System Atlas (GENSAT) projects use radioactive *in situ* hybridization and fluorescent protein reporters, respectively. GENSAT is now a core database of NCBI's Entrez system. BGEM and GENSAT differ from ABA in that they include data from multiple embryonic stages as well as adults. Another distinction is that GENSAT's protein reporters often fill the neurons they are expressed in, revealing projection patterns as well as the cell bodies (Gong et al., 2003). Additional information and

comparison of these and other atlases are given in Sunkin (2006).

RNA expression profiling using microarrays or sequence-based approaches (SAGE and RNA-Seq) stands in contrast to atlases in that spatial resolution is (usually) ignored at the gain of simultaneous quantitative measurements of thousands of genes in one sample. This allows the creation of data sets surveying expression over many different conditions. As in many areas of biology there is much interest in using expression profiles to characterize the nervous system and its disorders (Mirnics and Pevsner, 2004). In some ways, expression profiling is poorly suited to analyzing the nervous system, as the tissues that are most easily available are highly heterogeneous. This heterogeneity results in dilution of biological signals: genes of interest may be expressed in only a few cells and lost in the background, or changes in expression might appear smaller than they really are. This makes the application of profiling to the nervous system a demanding activity that can push the technology to its limits. While this discourages some, it highlights the need to carefully design and analyze experiments, and take advantage of prior knowledge through integration (the approach of GeneNetwork) and meta-analysis.

Expression profiling data is readily found in public data repositories, the most important of which are GEO (Edgar et al., 2002) and ArrayExpress (Parkinson et al., 2007), which together contain hundreds of brain-related expression studies. A thorough review of expression studies in the brain (Aarnio et al., 2005) identified 448 papers as of June 2004, of which less than one in five had data available online. A more recent review identified about 400 brain-related studies with public data in GEO and ArrayExpress (Wan and Pavlidis,

2007).

To use this mass of data, more tools are needed. GEO and ArrayExpress offer a variety of useful analysis tools, but comparing data across studies is difficult. To that end, third-party data analysis tools are beginning to appear, and some of these are geared to neuroscience. Gemma (Lee et al., 2004) (http://chibi.ubc.ca/Gemma), which is developed in our laboratory, offers tools for the collective analysis of multiple brain expression data sets, and related tools without a neuroscience focus, are offered by a number of other systems (Rhodes et al., 2004;Assou et al., 2007;Pan et al., 2007). Integrating this type of analysis with spatially resolved atlases will be an important area of activity (Sunkin, 2006). Expression data from microarrays can be compared to in situ data such as the ABA, in order to aid interpretation (Lee et al., 2008).

1.2.9 Conclusion of literature review

An editorial by David Van Essen (2007), identifies a key area where effort is needed in neuroinformatics: well-populated databases that are able to efficiently interoperate. This requires standards and terminologies, and community acceptance of the idea of sharing data. Dr. Van Essen envisions a future in which it will be possible to use informatics resources to rapidly answer natural-language questions such as, "What parts of the brain are abnormal in individuals with autism?" (Van Essen, 2007). While this might still sound like science fiction, our review of the state of the field makes us optimistic that some of Van Essen's vision is reachable in the near future. There is a great deal to be done, but as I demonstrate in my thesis, a bioinformatician can already explore a wealth of neuroscience information

stored within general and domain-specific bioinformatics resources at multiple scales from molecules to behaviour.

Table 1 Data domains in neuroscience

Levels of Nervous system organization	**Examples of data modalities**
Organism	behaviour, physiology
Whole Brain	functional and anatomical imaging, brain region connectivity, connectomes
Brain region	microcircuitry, electrophysiology
Cells	neuronal morphology, electrophysiology
Cellular Compartments	protein localization
Molecules	genotypes, protein interactions, gene expression profiles

Table 2 Neuroinformatics resources

Project	Domain	URL
Allen Brain Atlas	Spatial gene expression	http://www.brainatlas.org/
BAMS	Brain architecture: molecular, cellular, and connectivity	http://brancusi.usc.edu/bkms/
BIRN	Research network	http://www.nbirn.net/
BrainInfo, Neuronames	Neuroanatomy	http://braininfo.rprc.washington.edu/
BrainMap	Functional neuroimaging	http://www.brainmap.org/
Brede database	Functional neuroimaging meta-analysis	http://hendrix.ei.dtu.dk/services/jerne/brede/
CCDB	Cellular and subcellular imaging	http://ccdb.ucsd.edu/
CoCoDat	Neuronal microcircuitry	http://www.cocomac.org/cocodat/
CoCoMac	Connectivity data of macaque	http://www.cocomac.org/
fMRIDC	Functional neuroimaging	http://www.fmridc.org/
Gemma	Gene expression meta-analysis	http://chibi.ubc.ca/Gemma/
Genenetwork	Systems genetics	http://www.genenetwork.org/
GENSAT	Spatial gene expression	http://www.gensat.org/
Neurodatabase	Neurophysiology	http://neurodatabase.org
Neuroinformatics Portal Pilot	Resource catalogue	http://www.neuroinf.de/
Neuroscience Information Framework	Resource catalogue and data portal	http://neuinfo.org/
NeuroSynth	Functional neuroimaging meta-analysis	http://neurosynth.org/
SenseLab	Neural systems, neurons, olfactory pathways, drugs	http://senselab.med.yale.edu/
SumsDB	Brain mapping	http://sumsdb.wustl.edu:8081/
SynDB	Synapse related proteins	http://syndb.cbi.pku.edu.cn/
WormAtlas	C. elegans neuronal connectivity	http://www.wormatlas.org/
Textpresso for Neuroscience	Genes, anatomy, drugs and other knowledge extracted from the literature	http://www.textpresso.org/neuroscience/

1.3 Chapter Summaries

The general aims of this thesis were to automatically collect large amounts of connectivity knowledge (WhiteText project, Chapters 2-4) and to characterize relationships between connectivity and gene expression (ABAMS project, Chapters 5-6).

The objective of the WhiteText project is to build a system capable of automatically extracting neuroanatomical connectivity statements from neuroscience abstracts. This objective is separated into three subtasks, each corresponding to a chapter:

Chapter 2 describes the first step of recognizing when an author refers to a brain region. The input is a biomedical abstract and the output is mentions of brain regions. In natural language processing research, this is known as named entity recognition. The chapter describes application of methods used by existing neuroscience databases and a state of the art statistical modeling method (conditional random field). Evaluation was performed against manually annotated brain region mentions.

Chapter 3 focuses on the task of converting brain region mentions to brain region concepts in a neuroanatomical lexicon. This conversion is also known as “normalization” or “resolution”, and is necessary because many names can refer to a single brain region concept. Manual evaluations were performed to gauge precision and coverage across a training dataset. For a given abstract, species of study explained a large amount of variance in the evaluation

measures. The tuned procedure was applied to an expanded corpus of over 12,000 abstracts, resulting in over one hundred thousand brain region mentions.

Chapter 4 builds on the work of Chapters 2 and 3 by evaluating methods for extracting connectivity relationships. Simple and advanced relationship extraction techniques are tested on a manually annotated set of 1,377 abstracts. This chapter tests if methods for extraction of protein-protein interaction statements generalize to extraction of connectivity relationships. The most accurate method was selected and applied to the large set of automatically extracted brain region mentions from Chapter 3. The result is over 28,000 predicted connectivity statements. A normalized set of these relationships is compared to an existing source of neuroanatomical connectivity.

The general aim of the ABAMS project was to apply and test computational approaches for elucidating the transcriptome and connectome. The objective was to analyze datasets describing cell-type-specific gene expression, connectivity, and regional expression. Specifically, we sought to test the hypothesis that there is a statistical relationship between the connections and gene expression levels of individual brain regions. This hypothesis is motivated by a large number of gene expression patterns that show unexplained spatial variation across the nervous system (Lein et al., 2007).

Chapter 5 describes a large-scale analysis of gene expression and connectivity in the rodent brain. Complex patterns of gene expression in the rodent brain are examined in the context of regional brain connectivity and differences in cellular populations. Two novel patterns of mouse brain gene expression showing a strong degree of anti-correlation are identified.

The patterns contain genes that mark neurons and oligodendrocytes, suggesting they reflect regional differences in cellular populations. In addition, the expression level of these patterns is correlated with connectivity degree, with regions expressing the neuron-enriched pattern having more connections with other regions.

Chapter 6 further examines relationships between gene expression and brain wiring in the adult rodent brain by analyzing shared connections. The analysis shows that adult gene expression signatures have a statistically significant relationship to connectivity and this effect is not entirely attributable to spatial correlations. Optimized lists of several hundred genes that carry significant information about connectivity are examined in detail. To overcome the effects of noise, replicate assays were used to create a smaller high confidence list of genes. Gene ontology analysis and literature review were performed on the lists to identify functional themes and associations to brain disorders.

Chapter 2: Automated recognition of brain region mentions[2]

2.1 Introduction

Bioinformatics has proven the value of databasing and formalizing knowledge. Traditionally much of the focus is on molecular biology but great opportunities exist in neuroscience (Akil et al., 2011). One means of building, or at least seeding, knowledge bases is text mining, or the automated extraction and formalization of information from free text sources such as the biomedical literature. There has been much interest in applying text mining to extracting information about genes and proteins. In the BioCreative 2 challenge, 44 teams competed to extract, resolve and link protein and gene mentions (Krallinger et al., 2008), and the methods work well enough to be of practical importance in creating databases (Leitner et al., 2008). There has been less work on how to apply such techniques to domain-specific knowledge in neuroscience.

One entity of interest in the neuroscience literature is mentions of neuroanatomical regions (which we call brain regions for short). By analogy to the task of extracting gene mentions, the ability to computationally extract mentions of brain regions would be of potential value in

[2] A version of this chapter has been published. French L, Lane S, Xu L and Pavlidis P (2009). Automated recognition of brain region mentions in neuroscience literature. Frontiers in Neuroinformatics. 3:29. doi:10.3389/neuro.11.029.2009

building neurobiological knowledge bases. This is because many neurobiological studies only make sense in the context of the specific brain regions studied. Furthermore anatomical or functional connections between regions are commonly described. Computationally extracting these locations would allow faster organization and mining of neuroscience data.

We hypothesize that many of the methods and approaches developed for extraction of information about genes can be applied to extraction of information about brain areas. This is an attractive approach because many of the challenges in analyzing text for information about genes are faced in trying to mine information about brain regions. These challenges include abbreviations, synonyms, lexical variation and ambiguity. For example, the gene "carbonic anhydrase 1" has synonyms including "carbonate dehydratase I", "Car1", and "CA-I". Its official symbol, CA1, is ambiguous in that it also matches a drug (the abbreviated form of coumermycin A1) and a brain region (the CA1 field of the hippocampus). Similarly brain regions have a variety of names and abbreviations, and can be confused with other types of entities. Approaches have been developed for addressing these problems for genes, so it seems reasonable to expect that the lessons learned will apply at least partly to other domains. However, before these approaches can be applied to brain regions, a "gold standard" corpus is needed. Such a corpus is needed both as training data for algorithms and for evaluation of methods. To our knowledge, no such resource exists for neuroscience text mining.

Past efforts in neuroscience text mining provided limited ability to retrieve brain region mentions, by looking for exact matches of brain region names from small lists (Crasto et al.,

2003;Crasto et al., 2007b;Muller et al., 2008). This limits the recall to a small number of (usually broad or large) brain regions. The most extensive effort is “Textpresso for Neuroscience”, with a list of 4,800 brain region terms (Muller et al., 2008). Unfortunately evaluations of these tools are lacking, as the methods were not checked against a gold standard set of annotated abstracts, leaving accuracy in question. The Neuroscholar project was the first to explore advanced natural language processing methods to extraction of neuroscience data (Burns et al., 2007). Focusing on neuroanatomical connectivity, Burns et al. sought to extract and annotate detailed statements from full-text articles. Their goal was extraction of relatively detailed experimental parameters and descriptions of results. They manually annotated 1047 sentences from 21 articles. Text spans were tagged with five different labels including two that represented brain regions. These annotations provided the test and training examples for a CRF that was able to produce the same tags at an overall 79% F-Measure (performance for brain-region recognition alone was not reported). Although it was a small dataset they found the CRF could be joined with manual curation to increase annotation rate by 255%.

The goals of the current work are two-fold. First, we provide a reasonably large corpus of article abstracts manually annotated for brain region mentions. Second, we develop and evaluate methods for extraction of brain region mentions from text, using the corpus. This sets the stage for further efforts at improving and applying text-mining methods to neuroanatomical questions.

2.2 Methods

2.2.1 Corpus creation

Articles for the corpus were initially selected manually but later an automated procedure was employed. The first 119 articles in the corpus were selected with the help of PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) searches using keywords such as "afferent" and "efferent". The process was then automated to increase speed of curation and reduce bias in selection. The automated procedure picks random articles from the Journal of Comparative Neurology. There was no limitation placed on the topic organism (rat and cat were most common but insects were the topics of some abstracts). We experimented with other search strategies, for example the MeSH keyword of "Neural Pathways". The Journal of Comparative Neurology was chosen to maximize the number of abstracts that included brain region mentions. It has also been used in previous work (Burns et al., 2007). A total of 1,377 abstracts were used.

The selected abstracts were retrieved in MEDLINE XML format for preprocessing. For each abstract the PubMed identifier, title and abstract were stored. The abstract text was then processed by the Schwartz and Hearst abbreviation expansion algorithm (Schwartz and Hearst, 2003). This identifies the short and long forms of abbreviations in the abstract with high accuracy. All short forms of the abbreviation are replaced with the long form followed by its short form in parentheses. Thirty-two abstracts (2.3%) were reloaded without expansion due to encoding errors. The abbreviation expansion changes are expressed in the XML markup and can be reversed. Finally, annotators are provided the abstract and title

for annotation. The General Architecture for Text Engineering (GATE, http://gate.ac.uk/) was used to create, compare and visualize the document annotations. Additionally, GATE provided a helpful interface and API for managing the document collections.

2.2.2 Manual annotation guidelines

The annotators were presented with the title and abstract text in the GATE interactive document display. Using the computer mouse, regions of text were selected and then "tagged" as representing a brain region mention. One annotator (the "primary" annotator) annotated all abstracts. A secondary annotator re-annotated a random subset of abstracts annotated by the primary annotator (to allow estimation of the human component in annotation accuracy). The annotators used their own knowledge of neuroanatomy, supplemented by online resources such as medical dictionaries, neuroanatomical atlases and BrainInfo (http://braininfo.rprc.washington.edu). An initial set of guidelines were developed prior to the annotation starting; these guidelines were amended in response to the outcome of periodic discussion of problems and manual review of the corpus.

Brain (and spinal cord) regions were the primary targets of our manual annotation efforts. We annotated all mentions of brain regions in both the abstracts and titles according to our guidelines. Although we annotated all brain region mentions, our guidelines are influenced by our interest in mentions that describe higher-level features such as neuroanatomical connections.

A key set of guidelines involves the level of detail. In particular, we did not attempt to

annotate details such as specific cortical layers, in part because they cover the whole cortex but also because these were judged to present an additional challenge that would be a topic of future work. Conversely, broad mentions of “systems” were not annotated (e.g. “orexin/hypocretin system” or “vestibular system”). However, mentions such as “cortex” were captured. Further, mentions of white matter tracts or fasciculi were not annotated. Annotations also included text that modified the mention. An example is "motor related areas of the hippocampus". We annotated the adjective forms of brain regions, for example "thamalic" or "cortical". We annotated parts that were identified by a number (primarily this applied to Brodmann areas or cortical regions such as V1). Brain region mentions were not extended to include organism name, so “rat hippocampus” would always be annotated only as “hippocampus”. We annotated text segments that referred to a specific region but might not be resolvable without more context. For example, in an abstract about the cerebellum we might find mentions of “medial zone”. As a fragment, “medial zone” cannot be assigned to a specific region.

One particular problem area is conjunctions or coordination ellipses that connect two entities together. Examples are "dorsal and ventral cortex" or "lower thoracic and lumbosacral segments" which could be expanded to “dorsal cortex and ventral cortex” and “lower thoracic segments and lower lumbosacral segments”. The difficulty is determining whether these should be broken up into two brain region mentions or treated together. Past annotation efforts have recognized this difficulty (Tanabe et al., 2005). Unlike abbreviations there is no reliable method to automatically expand such expressions (Buyko et al., 2007). In the corpus, annotation of conjunctions varies except in the abstracts annotated by both annotators where

consistency was enforced. To achieve this, the whole conjunction was annotated if the contained brain region names have been shortened.

2.2.3 Dictionary matching

To test dictionary matching approaches we created term lists from neuroanatomical nomenclature sources. Although several lexicons exist we focused on Neuronames, the largest source of brain region names (Bowden and Dubach, 2003). We extracted terms from both Nomenclatures of Canonical Mouse and Rat Brain Atlases and the Ontology of Human and Macaque Neuroanatomy. From the later, a total of 6,462 terms were extracted from the primary names, synonyms, ancillary structures and Latin terms. We additionally extracted 1,900 terms from the Nomenclatures of Canonical Mouse and Rat Brain Atlases that organizes terms from mouse (Hof et al., 2000;Paxinos and Franklin, 2001;Dong, 2007) and rat brain atlases (Swanson, 1999). Since we expand abbreviations within the abstracts we excluded abbreviations contained in Neuronames.

To match the Neuronames terms to the document text we used a GATE Gazetteer. The gazetteer identifies occurances of names based on provided lists. Bracketed text in Neuronames terms were removed before matching. We set the Gazetteer to use case insensitive exact string matching. Resulting annotations were joined to remove overlapping matches.

To compare our method to that used by “Textpresso for neuroscience” we used the lexicon files from http://www.textpresso.org/neuroscience, with case sensitive exact matching. To

further replicate conditions used by Textpresso we reverted the expanded abbreviations in the abstracts and did not filter abbreviation terms from the lexicon.

2.2.4 Conditional random field

For automated annotation of brain region mentions, we applied a linear chain conditional random field (CRF) using the Mallet software toolkit (Lafferty et al., 2001;McCallum, 2002). A linear chain CRF is similar to a hidden Markov model (HMM). Like an HMM, a CRF is a method for sequence processing that takes a series of symbols (in our case, words) as input and provides as output the predicted state (in our case, whether the symbol is part of a brain region mention or not). Unlike HMM's, in which state probabilities are conditioned only on the state of the previous token, CRF state probabilities are computed by conditioning on the entire input sequence. Therefore, it cannot compare the probabilities of labellings across sentences. In return CRF models allow token descriptions (features) with complex dependencies. For example, HMM's use the current token type but a CRF feature design can examine the previous and next two tokens.

To start, the CRF model must be trained, by computing features for tokens with known label sequences (training set). In our case each feature has a Boolean value (details on the features are given in the next section). For example a feature named "text=red" is true if the current token is "red". These features combined with the state transitions form feature functions. The feature functions are then given weights, so that a specific feature can influence the likelihood of specific state transition. The weights are learned from the known state sequences using an optimization procedure. For example, in Table 5 we can see that the

probability of the label sequence changing from outside of a brain region to inside is increased when the preceding token is “the”. For test sequences or sentences, probabilities of state sequences are computed. The most probable state sequence then forms the predicted brain region mention spans. For further detail we point our readers to a more complete introduction of CRFs (Wallach, 2004).

The GATE software was used to segment the abstracts into sentences and tokens. For Mallet we used default CRF settings from the SimpleTagger class except Gaussian variance was set to 1.

2.2.5 Features

As mentioned, all of the features we used were binary. Thus the representation of each token was a long binary vector representing, for each feature, whether it was present for the given token. The simplest feature is the token itself, generated for every word/token in the corpus. We tested orthographic features, for example an uppercase first letter or presence of a numerical digit. The part of speech tag and lemma of the word were computed and tested. Like the text features, the lemmas of every word become a feature that is set to true if a word’s canonical form matches that lemma. To determine lemmas and tags we employed a model for the TreeTagger software (Schmid, 1994) that was extensively trained on the GENIA biomedical corpus (Kim et al., 2003) for STRING-IE (Saric et al., 2006).

The token is compared to several term lists and lexical resources. For complete matching a word and neighbouring words must exactly match a brain region name in one of many

neuroanatomical lexicons. Further we segmented the brain region names into word n-grams. For example “ventral anterior nucleus” is fragmented into the 2-grams of “ventral anterior” and “anterior nucleus“. The tokens are then matched against these n-grams allowing relaxed matches to the lexicons. We further employed word lists for neuroanatomical terms describing boundaries or regions (e.g. bank, sulci, surface, area), neuroanatomical directions (e.g. dorsal, superior), root neuroscience terms (e.g. chiasm, raphe, striated) and stop words (e.g. on, this, is). Root neuroscience terms were extracted from Dr. Eric Chudler’s resource for neuroanatomical, neurophysiological and neuropsychological terminology (http://faculty.washington.edu/chudler/neuroroot.html). We used the stop word list from the Snowball small string processing language software (http://snowball.tartarus.org). We added regular expression features that match common templates, for example Brodmann's areas and spinal vertebrae. Finally, we employed window features that add context information to the current words feature set. This is done by encoding features from previous and following words into the current word's set.

To rank the context features we averaged feature weights from eight cross-validation folds. The weights are from CRFs using only the text feature with a context window of two tokens on each side. We show the top weights for the state transition of outside a brain region mention into inside one, which occurs at the first word of a brain region mention. We filtered out the direct features from the current word to leave only the weights and rankings of features derived from the neighbouring words. Next we calculated a normalized score by multiplying the weight by the natural logarithm of its frequency.

2.2.6 Experiment setup

Manual feature design and initial tests were performed using eight fold cross-validation on the 1,146 abstracts annotated only by the primary annotator. Annotations from both curators were merged by a logical OR operation at the character level (if an annotator marked that character as a brain region then it was kept). Sentences of an abstract were not split between training and testing sets. Each sentence became an input instance for the CRF. Final results were generated on the same eight fold cross-validation across all abstracts.

2.2.7 Evaluation

We used standard evaluation measures that ignore true negatives and operate at the annotation level instead of the token. Precision is defined as the proportion of predictions matching the annotated spans with recall being the proportion of annotated spans that match a prediction. F-Measure is the harmonic mean of precision and recall. In the strict case annotation spans must match exactly. Lenient measures are computed by counting partially overlapping spans as matches.

2.3 Results

In total 1,377 abstracts were annotated by the primary curator. A second curator annotated 231 of those abstracts for agreement evaluation. The average number of brain region annotations per abstract from the primary curator was 13.2 and 14.6 for the second. Interannotator agreement was 90.7% (F-measure), increasing to 96.7% for the lenient

measure. Table 3 displays the top forty occurring mentions and their frequencies in the corpus.

The GATE tokenizer split the corpus into 17,247 sentences then 461,552 tokens with 46,340 labelled as brain regions. On average each brain region is 2.3 tokens in length. We observed a large vocabulary of 17,901 token types.

Lexicon-based methods directed from neuroanatomical atlases performed poorly on the dataset, reaching 43.8% F-measure (precision=57.2%, recall=35.5%). We expected a higher level of precision; we believe variances in applying the annotation guidelines account for some of the false positives. Neuronames contains terms for layers, systems and tracts all of which we did not annotate. In addition, TextPresso contains abbreviations which possibly cause additional false positives.

The next best performance of 66.4% F-Measure was attained by a CRF using 625 features we derived primarily from neuroanatomical lexicons. The lemma and text based CRF's demonstrate the effect of the context window. These classifiers only look at the token type, or word. Without the window features the text based CRF achieves 66.7% F-measure. Adding information about the previous and next two words increases F-Measure to 76.1%. By combining any two of the designed, lemma and text feature sets the CRF reaches F-measures in the range 76-78%. Combining the text and lemma features only slightly improves on text alone suggesting the features are very similar. By combing all three feature sets, the F-Measure peaks at 78.6%, with most of the gain from recall. This CRF that

combined all features perfectly predicted all brain region mentions for 174 abstracts that had on average 6.8 brain region mentions per abstract.

We were unable to clearly determine which of our designed features contributed most to the final performance. This is due to the high dependency between the designed features and the simple text features. Furthermore, F-Measure varies by about 1% across different cross-validation splits, so improvements of less than 1% are not significant. Throughout Table 4 the recall rate is below precision. This suggests many novel brain regions are left unrecognized, also known as out-of-vocabulary error. Indeed, we find that on average 19.3% of text features are observed in the test folds but not in the training folds. To test the impact of this effect, we repeated the experiment but allowing the sentences of an abstract to be spread across training and testing sets. This decreases unseen words to 10.4% because new terms are often mentioned many times throughout an abstract. At this sentence level performance improves; F-measure reaches 0.813 with the gain in recall twice that of precision. This suggests that, not surprisingly, performance can be improved simply by having more diverse training data.

We found some of the poorly classified examples were studies of brain regions from insects or other organisms underrepresented in the corpus. These abstracts tended to lack relevant training samples, and the regions they mention are not contained in the brain region lexicons we collected, resulting in poor recall. To examine this effect in more detail, we used a subset of abstracts for which we annotated the organism of study. This subset was further reduced to those studying monkey, cat, rat and mouse brains. A full featured CRF trained on this set of

214 common organism abstracts demonstrates much higher performance than a CRF trained on a random subset of the same size. This is demonstrated primarily by recall which increases to 75.7% from 67.6%, combined with a small increase in precision we find F-Measure increases to 77.8% from 72.5%. In terms of unseen features, the random set has 20.2% compared to 17.6% for the common organism set. This suggests that both sets have a similar out-of-vocabulary error.

We began by assuming that expanding abbreviations to the full forms would increase performance. As a test of this assumption, we reverted the expanded abbreviations back to the original, resulting in an F-Measure decrease of only 2.1 (to 76.5%). If we include the Neuronames abbreviation terms as an added feature this difference is reduced to 1.4.

We observed that coordinating conjunctions (see Methods) cause a significant amount of error. Examples are “middle and caudal amygdala” or “hippocampus and amygdala”. Five percent of annotations have a similar form with 893 annotations in 403 of the abstracts containing “and”, “or”, comma, semicolon, or a slash. By removing these abstracts we remove annotations that span conjunctions, the remaining abstracts still have conjunctions but each part is annotated separately. By training and testing the CRF on the reduced set of 974 the F-Measure increases to 79.9. This is significant compared to 76.5% reached by a CRF trained on a random set of the same size. With these consistently annotated conjunctions the strict precision gains the most, while lenient precision is almost unchanged. This suggests both datasets produce similar predictions but consistent annotations produce more precise spans.

Table 5 presents the context feature weights derived from a text only CRF. The window size ranged from the two preceding and following tokens. Although we only display the top 20, this CRF has over 300,000 weights for 17,901 token types times 5 token locations across four state changes. As expected common prepositions or adpositions are the most informative. Interestingly, “rat” and “monkey” have top scores. It seems the CRF learned that an organism name often precedes a brain region mention. Another entry is “projections” that is informative when seen two words before the current token. The importance of this connectivity-related term makes sense given the high number of tract tracing experiments in the Journal of Comparative Neurology.

We found several techniques frequently used in general and biomedical named entity recognition research did not improve performance. Guided by work on gene name extraction we experimented with bidirectional parsing and beginning-inside-outside labels (Hsu et al., 2008). We processed the text using MMTx and extracted rich semantic features (Aronson, 2006). We tested feature induction (McCallum, 2003), an extension of the CRF framework. To treat the abstract as a whole we tested treating each abstract as a sequence instead of its sentences and carried the features from the first mention of a word to all the following. The large vocabulary suggested semi-supervised learning may help; we tested a self training approach using an additional set of 3,881 unlabelled abstracts. Unfortunately, all of these methods failed to produce a significant increase in performance when compared to our best results.

Table 3 Top 40 frequently occurring mentions

Mention	Frequency
retina	313
retinal	280
spinal cord	256
cortical	239
superior colliculus	142
cortex	140
olfactory bulb	134
brainstem	127
thalamic	122
thalamus	115
hippocampus	108
hypothalamus	100
lateral geniculate nucleus	92
olfactory	92
cerebellum	86
thalamocortical	85
suprachiasmatic nucleus	83
amygdala	78
hippocampal	76
optic nerve	74
forebrain	73
striatum	73
inferior colliculus	72
visual cortex	71
cerebral cortex	69
basal forebrain	68
nucleus of the solitary tract	64
spinal	64
cerebellar	63
globus pallidus	61
midbrain	60
periaqueductal gray	60
locus coeruleus	59
basal ganglia	57
nucleus accumbens	55
substantia nigra	55
v2	55
area 17	54
prefrontal cortex	52

Table 4 Results from evaluated techniques

	Strict			Lenient		
Name	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
TextPresso Lexicon	0.529	0.185	0.274	0.824	0.288	0.427
Neuronames Lexicon	0.572	0.355	0.438	0.839	0.521	0.643
Features CRF	0.751	0.595	0.664	0.889	0.704	0.786
Lemma CRF	0.773	0.681	0.724	0.890	0.784	0.834
Text CRF	0.811	0.717	0.761	0.924	0.818	0.868
Features + Lemma + Text CRF	0.813	0.761	**0.786**	0.916	0.857	0.886

Table 5 Top 20 context features from text only CRF

Token Type	Position	Count	CRF Weight	Normalized Score
the	previous token	28376	11.4	117.2
and	previous token	13109	10.8	102.8
period	previous token	16811	10.4	101.3
from	previous token	2295	10.4	80.6
in	previous token	12203	8.5	80.1
to	previous token	6630	9.1	79.9
with	previous token	2957	9.8	78.1
that	previous token	3581	9.2	75.5
rat	previous token	777	10.4	69.2
into	previous token	758	9.6	63.9
monkey	previous token	216	11.8	63.6
left bracket	previous token	10944	6.7	61.9
labeled	previous token	785	9.0	60.2
projections	second preceding token	904	8.6	58.3
The	previous token	3274	7.0	56.4
or	previous token	1198	7.9	56.2
mouse	previous token	171	10.9	56.0
and	next token	13108	5.8	54.7
of	previous token	19205	5.5	54.6

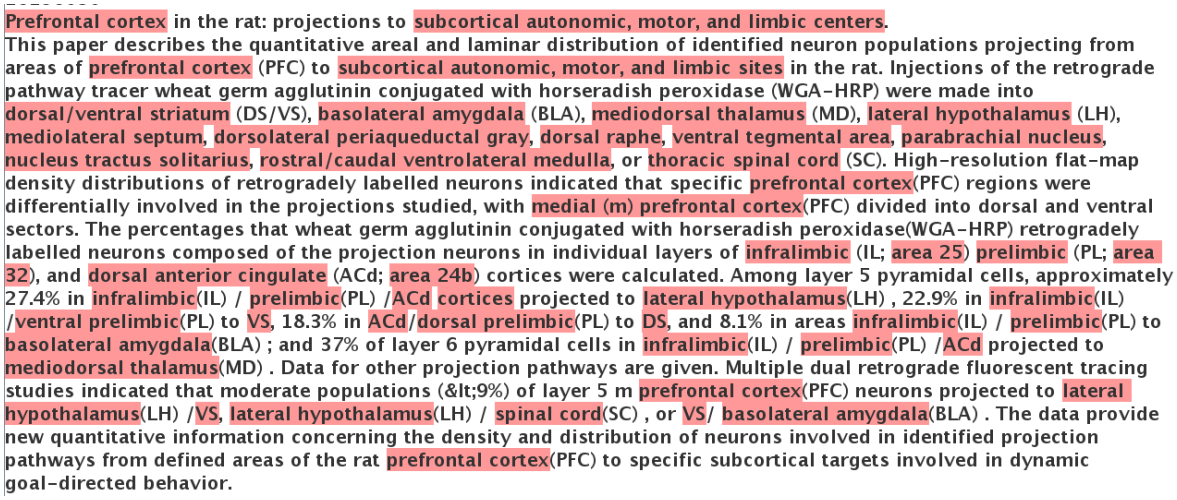
Prefrontal cortex in the rat: projections to subcortical autonomic, motor, and limbic centers.
This paper describes the quantitative areal and laminar distribution of identified neuron populations projecting from areas of prefrontal cortex (PFC) to subcortical autonomic, motor, and limbic sites in the rat. Injections of the retrograde pathway tracer wheat germ agglutinin conjugated with horseradish peroxidase (WGA-HRP) were made into dorsal/ventral striatum (DS/VS), basolateral amygdala (BLA), mediodorsal thalamus (MD), lateral hypothalamus (LH), mediolateral septum, dorsolateral periaqueductal gray, dorsal raphe, ventral tegmental area, parabrachial nucleus, nucleus tractus solitarius, rostral/caudal ventrolateral medulla, or thoracic spinal cord (SC). High-resolution flat-map density distributions of retrogradely labelled neurons indicated that specific prefrontal cortex(PFC) regions were differentially involved in the projections studied, with medial (m) prefrontal cortex(PFC) divided into dorsal and ventral sectors. The percentages that wheat germ agglutinin conjugated with horseradish peroxidase(WGA-HRP) retrogradely labelled neurons composed of the projection neurons in individual layers of infralimbic (IL; area 25) prelimbic (PL; area 32), and dorsal anterior cingulate (ACd; area 24b) cortices were calculated. Among layer 5 pyramidal cells, approximately 27.4% in infralimbic(IL) / prelimbic(PL) /ACd cortices projected to lateral hypothalamus(LH) , 22.9% in infralimbic(IL) /ventral prelimbic(PL) to VS, 18.3% in ACd/dorsal prelimbic(PL) to DS, and 8.1% in areas infralimbic(IL) / prelimbic(PL) to basolateral amygdala(BLA) ; and 37% of layer 6 pyramidal cells in infralimbic(IL) / prelimbic(PL) /ACd projected to mediodorsal thalamus(MD) . Data for other projection pathways are given. Multiple dual retrograde fluorescent tracing studies indicated that moderate populations (<9%) of layer 5 m prefrontal cortex(PFC) neurons projected to lateral hypothalamus(LH) /VS, lateral hypothalamus(LH) / spinal cord(SC) , or VS/ basolateral amygdala(BLA) . The data provide new quantitative information concerning the density and distribution of neurons involved in identified projection pathways from defined areas of the rat prefrontal cortex(PFC) to specific subcortical targets involved in dynamic goal-directed behavior.

Figure 1 A representative annotated abstract with several expanded abbreviations

The original source abstract is from Gabbott and colleagues (2005).

2.4 Discussion

We have provided the first corpus of manually annotated brain region mentions in biomedical abstracts. The corpus is large enough to allow statistical models to learn the nomenclature. This is demonstrated by the text-based CRF which reached a 76.1% F-Measure without outside resources. We found context windows, lemmatization and abbreviation expansion to be the most informative features for CRF labelling. A CRF using all the features provided the best performance of 78.6% F-Measure.

Compared to more advanced techniques, the dictionary approach based on neuroanatomical lexicons performed poorly. However, it has the advantage of speed and easier resolution to standardized names. Furthermore, features derived from these lexicons provide valuable

information to the CRF models.

We demonstrated that significant amounts of error are due to coordinating conjunctions, previously unseen words and brain regions of less commonly studied organisms. The poor performance of the lexicon combined with recall values consistently below precision suggest that lexical resources for neuroscience need to be improved. Current resources are based primary on neuroanatomical atlases of a few organisms. With open initiatives like NeuroLex we hope richer resources will be generated by a broader audience (http://neurolex.org/wiki/).

We performed a preliminary examination of normalization of mentions to standardized identifiers. This task is more difficult than mention extraction alone, as demonstrated by our baseline methods covering just over one third of mentions. One reason for the difficulty of the normalization task is that researchers do not use standardized nomenclatures for brain regions in their papers. This is a recognized problem for resolving gene mentions (where aliases are common) which has been ameliorated to some extent by efforts by nomenclature standardization committees (Wain et al., 2004). Such efforts would be of obvious value in neuroscience (Bug et al., 2008). When combined with organism identification it grows in difficulty.

Chapter 3: Using text mining to link journal articles to neuroanatomical databases[3]

3.1 Introduction

The last 15 years has seen increasing interest in formally encoding and bringing together existing neuroscience databases (Shepherd et al., 1998;Koslow, 2005;Akil et al., 2011). These databases are often focused on a specific domain, and thus must be linked together or otherwise integrated to fulfill their potential for enabling discovery. A major challenge is that the majority of neuroscience data, results and conclusions are stored in scientific articles. The sheer mass of this material and its relative inaccessibility to integration with other databases is a bottleneck. A key step in enabling efficient mining and integration of the neuroscience literature is the formal encoding of concepts they contain. While this can be done manually, high-throughput methods based on natural language processing techniques ("text mining") are attractive. An obvious target for formal encoding of neuroscience data is the anatomical brain region where an experiment was performed. This chapter describes tools for analyzing brain anatomy information from free text and provides novel applications.

---

[3] A version of this chapter has been accepted and published online. French L, Pavlidis P (2011). Using text mining to link journal articles to neuroanatomical databases. Journal of Comparative Neurology, Copyright © 2011 Wiley-Liss, Inc..

While free-text searches (such as those supported by major web search engines) are effective at finding documents that contain user-specified text, there are many advantages to formally linking text describing a concept to a fixed, formalized identifier. For example, the text strings “ventral tegmentum”, “ventral tegmental area” and “VTA” can all be mapped to concepts in standardized terminologies. These concepts are identified by unique and stable concept identifiers such as “birnlex_1415”. This mapping (also referred to as the process of “normalization”) separates the concept of the VTA from the way it is presented in text. One advantage of this formal mapping is that it enables query expansion to include sub-parts of brain regions (Gardner et al., 2008). Using the structure of the terminology, software can infer that a query for “midbrain” should include information that refers to the VTA, because the terminology encodes the fact that the VTA is part of the midbrain. In contrast, a purely text-based search for “midbrain” would not be guaranteed to retrieve information on the VTA. A second advantage is integration across data modalities. The fact that the VTA contains dopaminergic neurons is formally encoded in the terminology and can be discovered automatically (Gardner et al., 2008). Software can be used to automatically learn that tyrosine hydroxylase is one of the genes most specifically expressed in the VTA, using the formal brain region encodings in the Allen Brain Atlas (Lein et al., 2007). Genome-scale expression experiments that studied the VTA can be identified by links to the Gemma system (Lee et al., 2004). Similar integration approaches will reveal patterns of anatomical connectivity (Bota et al., 2005) or functional imaging results (Nielsen, 2003). The third advantage of formal encoding of brain regions is to discover patterns of information hidden in text. For example, the co-occurrence of mentions of a brain region might be used to infer a functional or structural connection between them. Similarly associations between brain

region mentions could be linked to other concepts found in text such as “addiction”. The critical step in enabling all of these scenarios is identifying that a piece of text such as a journal article abstract refers to a specific brain region concept. The same principles naturally apply to other conceptual domains such as drugs or diseases, which have their own formalized terminologies.

Our focus in this chapter is the development of methods for automatically linking free text to formal brain region identifiers. In Chapter 2, we presented high-performance methods for the first step needed to perform this task, which is identifying which parts of a document refer to brain regions. This recognition step only highlights textual spans or mentions that represent brain region mentions. In this chapter we address the step of automatically normalizing the mentions by resolving them to identifiers in neuroanatomical atlases and ontologies. By normalization (referred to as “standardization” or “resolution” in the text mining literature), we mean the mapping of a piece of text (a mention) to the concepts referred to by the text, in a formal way that can be used by computers. This addresses the difference between the concept of, for example, the substantia nigra pars compacta and the text “substantia nigra pars compacta”. In a computer system, we want all mentions of the concept “substantia nigra pars compacta” to be accessible in a consistent way. For example, the text “SNPC”, in the appropriate context, might refer to the same concept. If the computer system stores the information for occurrences of the text “SNPC” separately from that for the text “substantia nigra pars compacta”, queries accessing the latter will not successfully retrieve information linked to the former.

To our knowledge, little formal work has explored automatically normalizing brain region mentions to database identifiers. The most relevant studies are by Srinivas at al., who extracted compound terms from thalamic atlases and manually filtered them for neuroanatomical concepts (Srinivas et al., 2003;Srinivas et al., 2005). They then attempted to map the acronyms and terms across the atlases of cat, primate, human and monkey. Their results were focused on recall over precision to list possible mappings that can be evaluated manually for thesauri creation. Most other efforts in this area are focused on similar information retrieval tasks - a query brain region string is given and used to search a database (Bowden and Dubach, 2002;Nielsen, 2003). The most advanced example is the Neuroscience information framework (NIF) which matches user queries to existing brain region terminologies to expand the input query with synonyms (Gardner et al., 2008). Other literature retrieval search engines attempt to match mentions to lists of regions but they do not have explicit identifiers in an atlas or ontology (Crasto et al., 2003;Muller et al., 2008). Several general purpose tools extract biomedical concepts from the literature (Jonquet et al., 2009;Aronson and Lang, 2010). These tools discover terms from large biomedical terminologies which include some brain region concepts. For this work we choose to explore several simple methods and customize them for neuroanatomy.

Several challenges prevent full resolution of brain region mentions. Ambiguity often prevents confident resolution, for example the hypothalamus, medulla and thalamus each have an “arcuate nucleus”. Detailed studies present further challenges as authors often modify region names beyond existing nomenclature. This is done by adding directional or descriptive prefixes like “dorsal” and “agranular”. The wide range of organisms used in neuroscience

study presents another problem as taxa are described with different neuroanatomical terminologies. Here we present several novel solutions to these problems and evaluate their effectiveness, and yield a method that provides a high level of accuracy in mapping text to brain region concepts. We apply our approach to the analysis of a large set of abstracts from the Journal of Comparative Neurology, providing information on the distributions of brain region mentions. Our results are a starting point for linking diverse neuroinformatics data sources to literature-based information on brain regions. Finally, we highlight several remaining challenges.

3.2 Methods

3.2.1 Annotated corpus

We used our previously described annotated corpus of brain region mentions in journal abstracts (Chapter 2). The corpus of 1377 abstracts consists of 1258 abstracts randomly chosen from the Journal of Comparative Neurology and 119 abstracts selected from other neuroscience journals. Although the text spans are manually curated, the corpus provides no normalization of the brain region mentions into database identifiers. Previously, an unsupervised abbreviation expansion algorithm was applied to all abstracts in the corpus (Schwartz and Hearst, 2003). All extracted mentions of an abbreviation short form were expanded to their long forms in a given abstract.

3.2.2 Extraction of lexicons

The complete lexicon was compiled from NeuroNames (Bowden et al., 2007), NIFSTD (Bug et al., 2008), Brede Database (Nielsen, 2003), the Brain Architecture Management System (BAMS) (Bota and Swanson, 2008) and the Allen Mouse Brain Reference Atlas (ABA) (Dong, 2007). All terms were converted to lowercase and are linked to the provided identifiers. Of the five lexicons only BAMS and ABA are true neuroanatomical atlases that provide direct links between brain region names and 3D volumes in a digital or print format. The Brede database provides similar spatial data with 3D coordinates for named regions of interest. We did not add abbreviation terms to the lexicon because we expand abbreviations as described above.

NeuroNames terms were extracted from all worksheets in the NeuroNames Ontology of Mammalian Neuroanatomy (NN2010) and Nomenclatures of Canonical Mouse and Rat Brain Atlases (NN2007) excel files (Bowden and Dubach, 2002). Classical, ancillary, Latin and synonym terms were added to the lexicon. Further, terms from all four mouse and rat atlases were added to the lexicon. Overall 9,188 unique terms were extracted to represent 3,238 Neuroname concepts.

The 2,391 NIFSTD terms were extracted from the 1,272 classes in the Anatomy subontology. Synonyms and the main labels were extracted for ontology classes that were regional parts of the eye, ear, brain, spine and ganglion of the peripheral nervous system.

Terms from the Brede Database were extracted from the worois.xml file. Terms were

obtained from all name and variation XML tags. Hemispheric “left” and “right” prefixes were removed to be consistent with the rest of the lexicon. In total 1,006 terms were extracted from Brede to represent 763 concepts.

For BAMS, we extracted terms from the primary lexicon - Swanson-1998 (Swanson, 1999). This lexicon allows linking to the rich connectivity information curated into BAMS. The version of the BAMS database we use contains 962 rat brain region terms and is accessible via bulk download (http://brancusi.usc.edu/bkms/xml/swanson-98.xml). Instead of parsing the original XML, we used a converted semantic web version created by John Barkley (http://sw.neurocommons.org/2007/kb-sources/bams-from-swanson-98-4-23-07.owl).

Allen Brain Atlas terms were obtained from the OWL formatted version downloaded from the Allen Brain Atlas API documentation. Like the above sources, abbreviations were excluded from the extraction. In total 910 terms and concepts were extracted (no synonym information).

The total number of concepts in these five lexicons is 7,145, but it is clear that there is extensive redundancy among them (even after accounting for species-specificity of concepts). Unfortunately, because there are limited direct mappings of concepts across the terminologies, it is difficult to estimate how many different brain regions are represented in total. We arrive at a rough estimate of 1,000 different mammalian brain region concepts based on the sizes of four of the lexicons, and the fact that the much larger NeuroNames (at over 3,000 concepts) has an expanded concept of “brain region” that includes “ancillary”

terms that tend not to be recognized as distinct concepts by the other lexicons.

3.2.3 Resolvers

We employed five methods of matching textual mentions to region names in the ontologies and atlases. The most basic is the Exact String Matching Resolver. This resolver simply converts the mention to lower case and attempts to match all characters to a region name in the lexicon. The next step is implemented in the Bag of Words Resolver which splits the mention strings into words (tokenization) and then looks for exact string matches for each of these words. This is a common information retrieval technique that matches the same text but ignores word order.

To remove lexical variation we again tokenized the phrases into words. We converted the words into a base form by using a stemmer. A stemmer normalizes words to their base form by removing common endings. For example “ventral striatopallidal parts of the basal ganglia” is stemmed to “ventr striatopallis part of th bas gangl”. We employed the Lovin’s stemmer as implemented by Eibe Frank (Lovins, 1968). We created two resolvers analogous to the ones above. After tokenizing and stemming, the first resolver will match the stemmed tokens to the stemmed terms in the lexicons (Stem Resolver). The second will match them in any order (Bag of Stems Resolver). The Bag of Stems resolver is similar to the orderless gap-edit global string-matching algorithm used by Srinivas et al. (2005). In their implementation they allowed half of the stems to match for terms greater than two words in length (uninformative common words excluded). Our Bag of Stems method is slightly different with

use of a different stemmer and a strict requirement of all words to match (our mentions might be modified to remove specific terms).

To compare these to an externally designed method we employed the Lexical OWL Ontology Matcher (LOOM). LOOM is a simple method for mapping across biomedical ontologies. While LOOM is not designed for matching free text mentions we found its approximate string matching technique to be of value. LOOM uses a string comparison function that requires an exact match for words longer than four characters and allows one character mismatch for longer strings (after removing spaces and parentheses). The LOOM authors show it provides comparable performance to more complicated tools for ontology mapping (Ghazvinian et al., 2009).

3.2.4 Mention editors

To improve the resolution of mentions we employed several techniques that edit the mentions. In total nine editors are employed. The final three are considered to be lossy because they remove important words from the mention (Table 6). Each mention editor is applied in the order presented in Table 6 and is only applied to unmatched mentions. The result of the editor does not replace the original mention, but instead expands it by adding modified versions. Each editor is executed once except the Direction Remover which is run a second time at the end to extract more general regions from very specific mentions.

3.2.5 Species extraction

We employed LINNAEUS, a species name identification system for biomedical literature for extracting species mentions from the corpus (Gerner et al., 2010). LINNAEUS provided an open and accurate tool for quantifying species mentions with accuracies above 90%. We used the default configuration properties to tag the abstracts for NCBI species identifiers. Of the 209 species found we manually deemed 44 to be not relevant. These primarily included mentions of reagents for tract tracing (“horseradish”, “phaseous vulgaris”, “pseudorabies virus”). We noted some false positives, including brain regions that were tagged as species (“n. superficialis”, “n. ambiguus”).

3.2.6 Data model

To capture the relations between abstracts, mentions, terms and ontology concepts we employed a resource description framework (RDF) model (W3C, 2004). The general RDF structure is guided by the relationships and entities in Figure 2. For NIFSTD, BAMS, ABA and Brede concepts we link to the original identifiers for future integration. The full RDF dataset is available on our supplement website at http://www.chibi.ubc.ca/WhiteText/ .

3.2.7 Manually created term to concept links

Several evaluations and manual modifications were applied to test and improve the normalization procedures. During our first test we noticed many commonly used synonyms were not mapping to the lexicon. Examples include unexpanded abbreviations and region

names that have been used as adjectives (“cortical”, “thalamic”). We were able to manually map 42 of 122 unresolved mentions that had more than 9 mentions in the corpus. We provide evaluation statistics with and without these annotated synonyms, because these hand-tunings were done post-hoc.

3.2.8 Evaluation

By automatically testing for exact string matches we were able to review the complete set of mention to region pairings. The exact string matches were automatically accepted while the remaining pairings were manually evaluated. Each mention-to-concept pairing was marked as accept, reject or specific-to-general (partitive relationship). A specific-to-general marking applies to mentions where the region was mapped to an enclosing region (e.g. “nucleus deiters dorsalis” mapped to “nucleus of deiters”). This applies to many cases, as several of our mention editors discard information. Resolutions of ambiguous terms were accepted only if they matched a majority of the contexts. For example, all mappings of “arcuate nucleus” were rejected because the abstracts in which they occur are not consistently referencing the arcuate nucleus of the thalamus, medulla or hypothalamus. To reduce redundant evaluations, pairings were grouped when the main text label for the matched region is the same across ontologies. The abstracts in which the mention occurred were used to judge the context and correctness of the resolution. Resolutions were accepted across species unless it was a specific parcellation scheme for a species, for example - “area 10a of Vogts”.

Normalization coverage represents the proportion of mentions that have been mapped to at least one brain region concept. This proportion of mapped mentions is dominated by

frequently occurring terms like "cortex". To control for mention popularity we provide two additional measures of coverage. The first ignores the number of times a mention occurs and treats each unique mention equally (rare mentions are given equal weight as common terms). The second ignores repeat mentions of a mention within an abstract and weights each mention by the number of abstracts it appears in.

Normalization accuracy was measured by dividing the number of accepted concept to mention links by all total mention-to-concept resolutions made. We take into account frequency of the mention by multiplying the concept to mention links by number of abstracts the mention appears in. We considered specific-to-general mappings to be an accepted resolution while also measuring their frequency individually.

Although the species name recognizer we choose has been previously evaluated we compared it to a subset of our abstracts that we previously annotated with species. Because the annotated tags were entered in free text we employed LINNAEUS to convert them to NCBI taxonomy identifiers. These converted identifiers were then compared to those extracted from the abstracts automatically.

3.3 Results

Figure 2 shows an overview of the system we developed, starting from journal abstract to mapped concept. In developing the approach, we examined the properties of the input terminologies, and carefully evaluated the quality of the mappings we obtained, as described in the next sections. In the final section we describe the application of the pipeline to a large

set of JCN abstracts and present findings on the patterns of brain region concept usage.

3.3.1 Summary of the terminologies

We first established the basic properties of the target terminologies (or "lexicons") we used for mapping. It is important that these terminologies encompass the range of concepts used in the literature. In total we extracted 11,909 terms from five terminologies. These terms represent a total of an estimated 1,000 different mammalian brain regions (see Methods). On average a concept in the aggregated terminologies had 1.6 terms or labels (for example representing synonyms; note that we must distinguish between "concepts" and their textual representation as "terms"). While we estimate that concept overlaps among the terminologies are high, term overlap across terminologies was remarkably low, with terms being linked to just 1.3 of the five terminologies on average, with 79.8% of the terms appearing in only one terminology. Across the ontologies the highest amount of overlap was between ABA and NeuroNames with 62.7% of the ABA terms appearing in the much larger NeuroNames set. In addition 53.5% of the NIFSTD terms appear in NeuroNames. This is expected because NIFSTD was originally based on NeuroNames (Bug et al., 2008). Although the NeuroNames curators have imported some of the ABA and BAMS terminology, it is not complete. While some "singleton" terms are minor variants of terms found in other terminologies (e.g., raphé vs. raphe), the lexicons contain many apparently obscure or rarely-used terms such as "area 22 of mauss 1908".

3.3.2 Evaluation of concept resolution

We ran our resolvers on the corpus of 17,585 brain region mentions (See Methods). To evaluate the results, we first examined coverage, providing a simple measure to compare across the different methods. We compute it as the proportion of mentions that are resolved to the lexicons. We provide three ways of measuring coverage that account for how the mention occurs in the corpus. To measure the coverage rate of unique mentions we weight each mention equally by ignoring the number of times it occurs. The two remaining coverage measures weight each mention by the total number of occurances in the corpus and the number of abstracts the mention occurs in (disregards multiple mentions in a single abstract). These measures show that popular brain regions are more likely to be resolved to existing lexicons than rare terms that appear only once in the corpus. In Table 7, these measures show that popular brain regions are more likely to be resolved to existing lexicons than rare terms that appear only once in the corpus with the coverage rate of unique mentions (18.8%) at half that of all mentions (47.7%).

The accuracy of the mappings was evaluated for all non-exact string mappings. Table 7 shows accuracy and coverage across the resolvers. We found the Simple Mapping Matcher performed the worst with 3.6% of unique mention mappings rejected. Overall the combined set of resolvers result in 4.3% of unique mentions being rejected and 52.9% of mentions failing to map.

3.3.3 Tuning and final evaluation

After reviewing the unmatched mentions, we decided to modify our pipeline and input lexicons. The first change was the addition of manually-created mention-to-term links in the lexicons. This is akin to adding synonyms to the ontologies. We were able to create these links for 42 of the 122 top unmatched mentions that occurred more than 9 times in the corpus. Examples include “cortical”, “thalamic” and “si”. Although we sought to remove acronyms and abbreviations, several occur in the list. These are primarily terms that the automatic abbreviation expander failed to resolve or a long form was not provided by the author. The rate of these errors is roughly 5%, and is similar to the tested accuracy of the abbreviation expander (Schwartz and Hearst, 2003). The addition of these links produces a 7.7 percentage point increase in mention coverage (Table 8). We list these values separately because they reflect post-hoc additions to the pipeline, but they provide true increases in coverage expected for a production system.

Further modifications derived from observations of unmatched mention patterns were implemented as “mention editors” (Table 6). The editors all perform slight modifications of the original mention and, as a last resort, remove qualifying terms such as “medial” from the string. When combined, the mention editors raise the coverage of abstract-mention pairs to 58.4% and 35.9% of unique mentions with 39.4% specific-to-general mappings. As Table 8 shows, they each provide a modest contribution to the increase while providing accurate mappings. Finally, the lossy editors (designed to discard qualifiers) created primarily general-to-specific mappings for the mentions that failed to match after applying the

preceding mention editors.

To gain insight into mentions that failed to match, we manually examined a random subset of 100 unmatched mentions. A quarter of the sampled mentions were references to brain regions that are not contained in the lexicons we used, and instead refer to regions in other species. Fourteen of the 100 unique unmatched mentions can be explained by annotation errors in our corpus, including text spans that missed the first character of a term and annotations of tracts. This result was expected given previously measured rates of annotator agreement (Chapter 2). The remaining unmatched mentions can be categorized as unique variants, very specific mentions and ambiguous mentions (a complete list is available as Appendix A). Beyond the annotation errors, it is not clear how to map the unmapped mentions without extending the lexicons.

3.3.4 Species-specific evaluation

We hypothesized that the quality of resolution would depend on the organism used in the study, as the lexicons are species-specific and many taxa lack lexicons. To filter for species of study we ran LINNAEUS to identify species mentions in the abstracts (Gerner et al., 2010). We compared the automatically tagged species information to a subset of 396 abstracts with manually annotated species information. LINNAEUS was able to recall 97.4% of the annotated species mentions that could be mapped to a specific species. Precision could not be fully evaluated because many mentions of species are too general and refer to a genus or other taxonomic level. LINNAEUS does not extract these terms and as a result terms like “Macaque monkey”, “pigeon” and “squirrel monkey” could not be extracted (but were still

annotated manually for the subset). Overall, LINNAEUS identified species terms in 88% of abstracts. Co-occurrence of species within abstracts is relatively low; the most common pair of species is rat and human which occur together in 30 abstracts.

As predicted, the coverage of mentions and specific-to-general matches varied greatly across species. Table 9 presents the results for a selected set of top occurring species. Species that lacked lexicons resolved less well and specific-to-general mappings occurred much more often. The top occurring species benefited from lexicons of their species. To determine the accuracy of the commonly studied species targeted by our lexicons we combined the mentions that co-occur with rat, mouse, human, rhesus monkey and *macaca fasicularis* mentions. Coverage of unique mentions for this grouping increases by 7 percentage points, and specific-to-general mappings are reduced to 33.7% from 39.4% on all mentions. Accepted mappings slightly increased from 95.1% to 96.6%.

3.3.5 Analysis of all Journal of Comparative Neurology abstracts

We ran our final method on 12,557 JCN abstracts that are not already in our corpus (covering 1975 to January 2011). This required first running the abbreviation expander, then the brain region mention detector as described previously (Chapter 2), followed by the tuned normalization pipeline described above. In total we found 142,178 brain region mentions. Of these 95,895 were resolved to a concept in a lexicon, representing 7,923 unique region mentions and 57,185 unique abstract-region pairs (on average 4.6 per abstract; 86% of abstracts having at least one). The resolution results resemble those from the manually annotated abstracts with 67.5% of mentions resolved and 27.4% of unique mentions

matched to a lexicon entry. For the subset of commonly studied species that cover the lexicons coverage reaches 71.6% of mentions and 32.3% of unique terms. The slight increase in mention coverage and decrease in unique coverage is expected from a larger corpus size generating a larger set of rare terms. Table 10 presents the top 25 most frequently occurring NIFSTD concepts. The types of unmatched mentions are similar to those found previously with many broad terms that are not explicitly in the lexicons and several insect brain regions such as “mushroom bodies”.

We examined the extent to which terms in the lexicons are used. We found that 44.1% of the 7,145 available concepts are used at least once. Viewed another way, over 55% of the concepts (and 77% of terms) in the lexicons do not appear to be used in any JCN abstract. These results suggest that many of the concepts (and terms) in the lexicons are rarely used by working scientists.

Because our analysis includes information on species and publication date as well as brain region use, the final data set allows interesting temporal analyses of the JCN. We first asked whether there is a tendency for more recent articles to use more narrowly defined brain regions. By comparing the publication year with the proportion of specific-to-general mappings in the training set we observe a slight but non-significant positive trend (Spearman correlation 0.18; p-value = 0.31). Our analysis is also able to reveal trends in the “popularity” of brain regions over the years. For example, we found that there was an abrupt dip in the mentions of “superior colliculus” in the early 1990s, while the hippocampus and amygdala enjoyed rising mentions until recently (Figure 3). A similar analysis of species of study

shows that mentions of mouse and humans are increasing, while rat and Rhesus monkey mentions are fading (Figure 4).

Table 6 Mention editor descriptions and examples

Example input and output mention strings are separated by ">>". The Direction splitter mention editor expands the single input string into two mentions. Methods that discard important words from the mention are classified as 'Lossy'.

Name	Description	Lossy	Example		
Direction splitter	Splits conjunctions that use neuroanatomical directions	No	dorsal and posterior hypothalamic areas	>>	[dorsal hypothalamic areas, posterior hypothalamic areas]
Hemisphere stripper	Removes prefixes that specify hemisphere	No	contralateral inferior olivary	>>	inferior olivary
Bracketed text remover	Removes text that is enclosed by brackets	No	secondary somatosensory (sii) cortex	>>	secondary somatosensory cortex
"n." expander	Expands "n." to nucleus	No	n. ambiguus	>>	nucleus ambiguus
"of the" remover	Removes subdivision descriptors	No	medial portion of the entorhinal cortex	>>	medial entorhinal cortex
Region suffix remover	Removes "region" suffixes	No	posterior cingulate region	>>	posterior cingulate
Cyto prefix remover	Removes prefixes that mention cytoarchitectural descriptions	Yes	parvocellular red nucleus	>>	red nucleus
Direction remover	Removes neuroanatomical direction specifiers	Yes	caudal cuneate nucleus	>>	cuneate nucleus
"nucleus of the" remover	Removes nucleus specifiers	Yes	nucleus of the pontobulbar body	>>	pontobulbar body

Table 7 Mention coverage and rejection rates across resolvers

Coverage is provided at three different levels to quantify repeated mentions. For the "Unique Mentions" and "Reject Unique" columns the number of times a mention occurs is ignored (rare terms are given equal weight as common terms). The "Abstract-Mentions" and "Reject Abs-Mention pairs" statistics ignores the number of times a mention occurs in an abstract. The "Mentions" and "Reject Frequency" columns weight each unique mention by the number of times it occurs in the corpus.

		Coverage		Mapping Accuracy		
Resolver	**Mentions**	**Abstract-Mentions**	**Unique Mentions**	**Reject Frequency**	**Reject Abs-Mention pairs**	**Reject Unique**
Exact String Match	41.1%	36.0%	14.3%	0.0%	0.0%	0.0%
Bag of Words	42.1%	37.1%	15.8%	0.2%	0.2%	1.0%
Stem	45.1%	39.4%	16.2%	0.5%	0.5%	1.0%
Bag of Stems	46.4%	40.8%	18.0%	0.7%	0.7%	1.9%
LOOM Matcher	41.1%	35.8%	14.3%	2.5%	2.5%	3.6%
All	47.1%	41.6%	18.8%	3.1%	3.2%	4.3%

Table 8 Incremental improvements from several additional methods

Added editor	New Mentions Matched	Added Mappings	Percent accepted	Specific to General Mappings	Matched Mentions
Baseline	8280	2963	95.7%	0.8%	47.1%
Manual to Mention to Concept links	1346	91	97.8%	0.0%	54.7%
Direction splitting editor	35	109	86.2%	7.3%	54.9%
Hemisphere Strip Mention Editor	131	265	100.0%	0.0%	55.7%
Bracketed text remover	29	73	90.4%	2.7%	55.8%
Converts n. to nucleus	5	14	100.0%	0.0%	55.9%
Remover of "of the" type phrases	27	67	85.1%	10.4%	56.0%
Region[s] suffix remover	36	56	100.0%	0.0%	56.2%
Cytoarchitecture prefix remover	37	76	97.4%	96.1%	56.4%
Direction prefix and suffix remover	1092	2204	95.8%	94.4%	62.7%
Remover of "nucleus of the" phrase	21	72	100.0%	100.0%	62.8%
Direction prefix and suffix remover	125	205	84.9%	84.9%	63.5%

Table 9 Resolution of species linked mentions

The “Species terms” column list all recognized terms for a given species. Coverage is provided at two levels by counting mention frequency (“Mention Coverage”) and ignoring it (“Unique Coverage”).

Species	Species terms	Mentions	Unique Coverage	Mention Coverage	Rejected Mappings	Specific-to-general Mappings
Cat	cats, kitten, cat, Cat, kittens	1001	42.5%	60.6%	3.0%	24.3%
Rabbit	rabbit, rabbits	200	60.0%	73.3%	4.0%	16.5%
Pigeon	Columba livia	157	40.1%	45.0%	1.1%	19.6%
Clawed frog	clawed frog, Xenopus laevis, African clawed frog, X. Laevis	107	57.0%	66.8%	8.1%	37.1%
Rat	rat, rats, Norway rat, Sprague-Dawley rats, Wistar rats, Sprague-Dawley rat	2434	44.6%	67.7%	3.2%	31.9%
Mouse	mice, mouse, murine, transgenic mice	396	55.8%	75.7%	2.6%	13.6%
Human	patient, patients, human, infant, children, humans, infants, people, participants, man	409	57.7%	73.5%	4.2%	11.6%
Rhesus Monkey	rhesus monkey, rhesus monkeys, Macaca mulatta	406	49.5%	63.6%	2.5%	29.6%
Macaca f.	macaca fascicularis, cynomolgus monkey, cynomolgus monkeys	143	64.3%	67.9%	5.9%	17.5%
Macaca f., Rhesus, Human, Mouse and Rat		3061	42.9%	68.6%	3.4%	33.7%
All		5941	35.9%	63.5%	4.9%	39.4%

Table 10 Top 25 most frequent brain region concepts in the Journal of Comparative Neurology

Regions are limited to the NIFSTD terminology with frequency determined from the full JCN corpus. Both the manually curated and automatically predicted brain region spans were used as input to the resolution approach.

Region	Frequency
Retina	7341
Cerebral cortex	5578
Spinal cord	3915
Thalamus	2290
Hippocampus	2098
Cerebellum	1953
Hypothalamus	1800
Olfactory bulb	1551
Brainstem	1512
Superior colliculus	1457
Neostriatum	1343
Amygdala	1312
Midbrain tectum	1109
Midbrain	1093
Forebrain	962
Solitary nucleus	819
Locus ceruleus	769
Substantia nigra	764
Cochlea	762
Entorhinal cortex	712
Lateral geniculate body	705
Dentate gyrus	684
Central gray substance of midbrain	662
Telencephalon	660
Cochlear nuclear complex	651

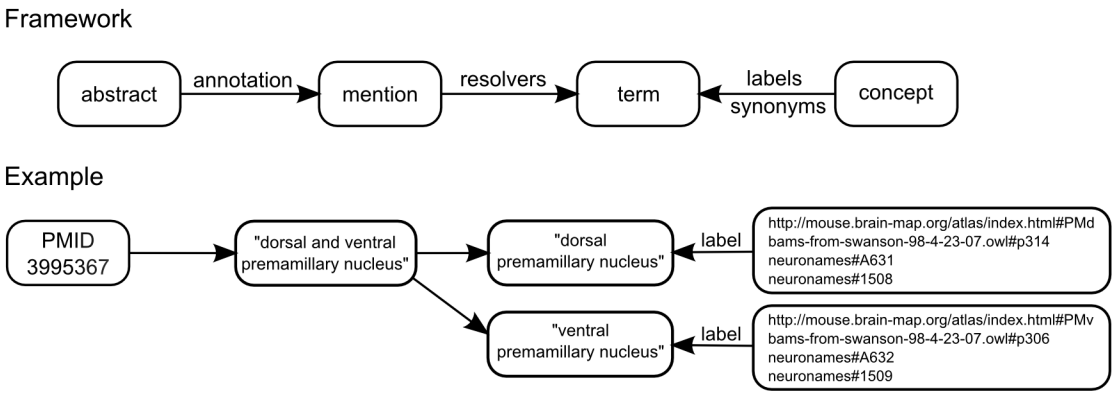

Figure 2 Representation of the system and an example

The procedure starts with an abstract that is manually or automatically scanned to find brain region mentions. The extracted mentions are then processed by mention editors and resolvers. For this example all resolvers including the exact string matcher resolve the direction split strings to the correct concepts.

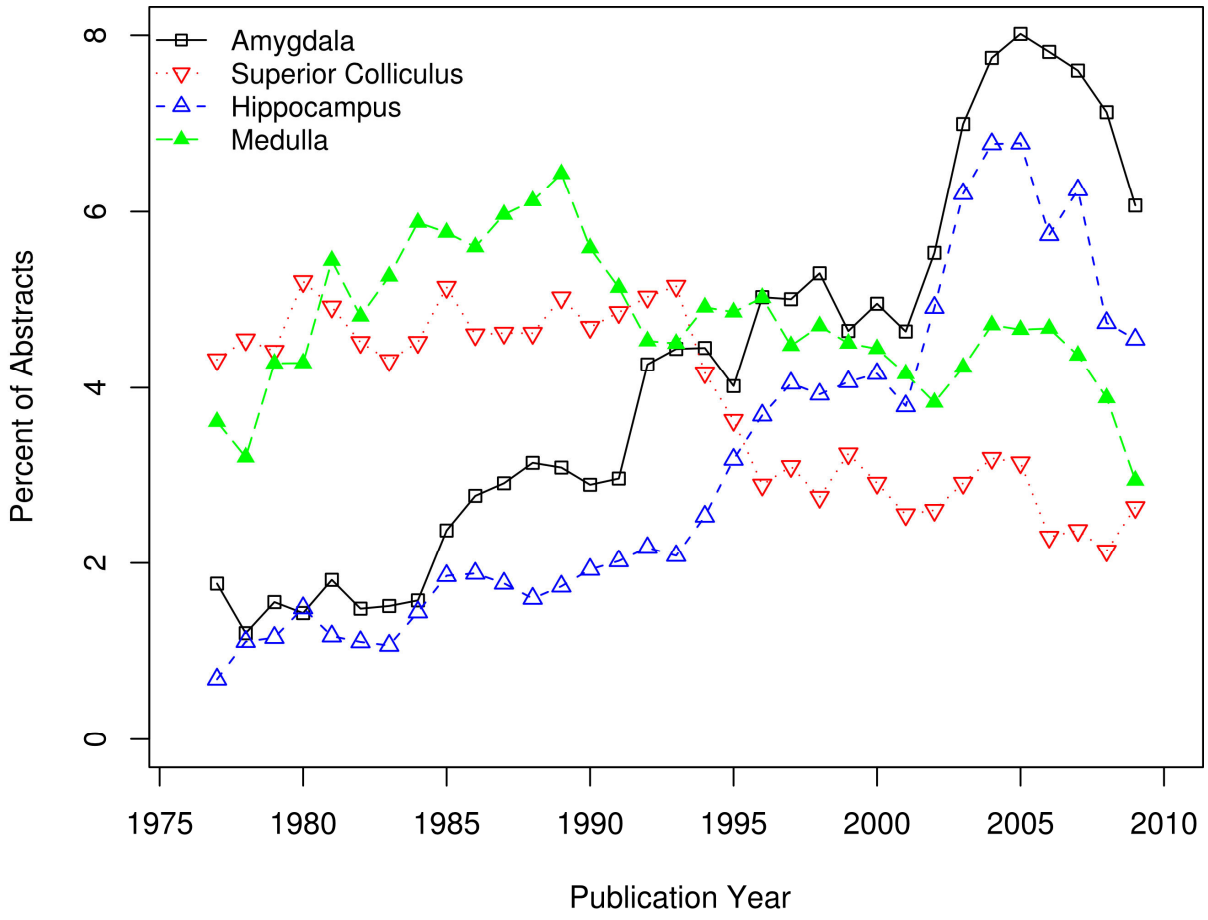

Figure 3 Trends in the proportion of yearly abstracts mentioning amygdala (black square), superior colliculus (red triangle), hippocampus (blue triangle) and medulla (green triangle)

Proportion values are smoothed by averaging the previous, current and following years.

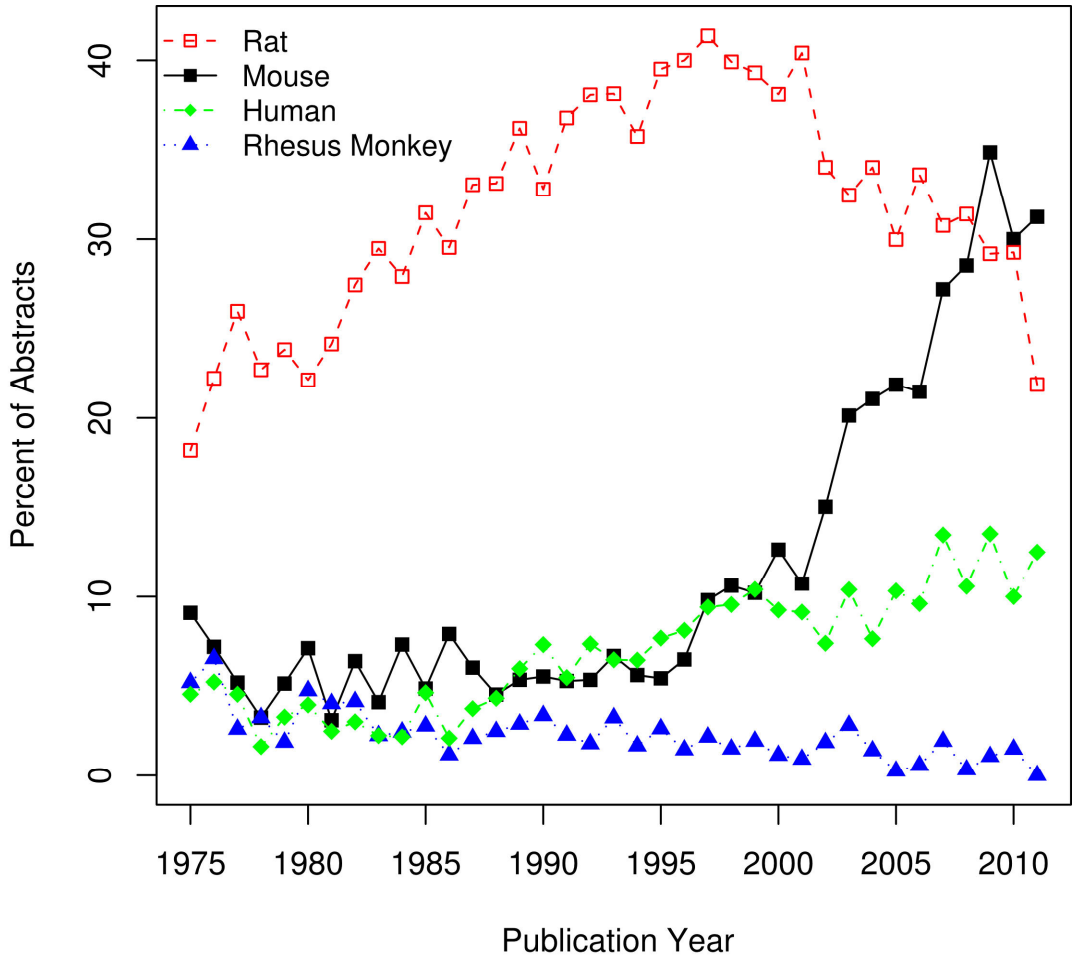

Figure 4 Changes in the proportion of yearly abstracts mentioning rat (red square), mouse (black filled square), people (green diamonds) and Rhesus monkey (blue triangle) over time

Only 32 abstracts are considered for the 2011 year. Rat, mouse and human are significantly increasing over the entire time period ($p < 0.05$). Abstracts mentioning Rhesus monkey are significantly declining ($p < 0.001$).

3.4 Discussion

Our contribution in this chapter is the development and thorough evaluation of a pipeline for mapping specific brain region concepts to free text in journal abstracts. While we achieve a high degree of coverage (64.5%) and precision (95.1%), and yield a data set of value for additional analyses, we identified many challenges and current limitations that need to be addressed. The primary problem we encountered was with the terminologies, which are not well-standardized and also, apparently, incomplete. The terminologies we used have surprisingly little overlap, despite some of them having common target organisms. This reflects the extensive variation in how neuroanatomical concepts are expressed in natural language, but the lack of harmonization across terminologies is striking. In addition, authors often mention regions that are beyond the granularity of the terminologies (for example, by adding a modifier such as "mediolateral" to a recognized term). While we presented a lossy mapping method that handles this problem, it is likely that some of these fine-grained terms should be added to the terminologies. To this end, we have contributed 136 new brain region concepts to NeuroLex (Larson et al., 2010). We selected the regions by filtering our results for specific-to-general mappings to an existing NITSFD brain region concept. We then selected terms that are co-mentioned with rhesus monkey, *macaca fasicularis*, rat, mouse or human in at least two separate abstracts, assuming that repeated use in the literature is evidence of utility. This automatically generated list of 152 region terms was reduced to 136 after manual adjustments for synonyms and conjunctions. Although this is a small first step, formalization of these mention-to-concept pairings would reduce the specific-to-general mapping rate by 2.5 percentage points. Further, these 136 mentions occur over 2,400 times in

the complete set of JCN abstracts. Because NeuroLex is presented in a wiki format, the community can review and edit these additions (http://neurolex.org). Another potential avenue for improving lexicons is the International Neuroinformatics Coordinating Facility (INCF) Program on Ontologies of Neural Structures (PONS) which seeks to establish formalized lexicons for neuroanatomy (http://www.incf.org/core/programs/pons).

In addition to pointing out gaps in the existing terminologies, our results point to a mismatch in the other direction, in that the terminologies contain numerous terms that do not appear in any JCN abstracts. Some of these are likely to be valid terms that are just not used often (for example, the rodent term “Perireunensis nucleus” never appears in any PubMed abstract; a wider web search turns up just a single mention in the accessible literature (Jacobsson et al., 2010)). An overall picture emerges of lexicons that are incomplete while simultaneously full of terms which may not be actually used in practice. Our results may thus aid the developers of lexicons and highlights the need for more work in this area.

Overall we found that our designed resolvers were precise at the task. We believe this is due in part to our avoidance of acronyms and relying on strict matches. The best resolver appears to be the Bag of Stems resolver that almost reaches the coverage of all resolvers combined while holding a low 1.9% unique term rejection rate. This agrees with previous work that tested a similar resolver for cross species mapping of thalamic atlases (Srinivas et al., 2005). The LOOM Simple Mapping Matcher, designed for a different task of ontology mapping performs worse than any other resolver. One advantage is that its one-character mismatch allowance provides some mappings our other resolvers cannot. While providing unique

mappings, the mismatch allowance leads to mapping errors such as “central” to “ventral”. Past work in the ontology mapping domain has placed LOOM at par to other more advanced methods (Ghazvinian et al., 2009). Those results suggest more complicated resolvers will not yield substantial improvements.

Through studying unmatched mentions and tuning the system we were able to improve the coverage from 47.1% to 63.5%. Unfortunately most of our added techniques resulted in modest improvements. The main contributors were the manually created mappings for unmatched mentions and the lossy editors that allowed resolution to enclosing regions (e.g. mapping “medial lateral cervical nucleus” to Lateral cervical nucleus). Our analysis of one hundred unmatched mentions suggests more advanced methods employing contextual information could be used to resolve ambiguous and co-referenced mentions. Context information has already been applied to cross-species mapping and may be adaptable to brain region mapping outside of atlases (Srinivas et al., 2005).

The most important contextual information seems to be the species of study. By applying an automated species extractor we linked the organism of study with the brain region mentions. Across the over 200 species we observed varied degrees of resolution. As expected, brain region mentions from amphibians, insects and fish had increased rejection and more specific-to-general mappings. Mammalian species like rabbit and cat performed at levels close to the average. Rat, the most common species of study in the corpus, had an above average coverage but also a high amount of specific-to-general mappings (31.9%). In comparison, the increasingly common mouse abstracts had only 13.6% specific-to-general mappings while

achieving the highest coverage (75.7%). This may reflect that the larger rat brain is commonly used for study of detailed rodent neuroanatomy that extends beyond the standard atlases. In addition the human abstracts have results similar to mouse with high coverage of mentions and few specific-to-general mappings compared to Rhesus monkey abstracts. Approximately half of the mentions are linked to species matching the target lexicons. These mentions from commonly studied species are accurately normalized, with low rejected (3.4%) and specific-to-general mappings (33.7%).

We applied the methods to an unseen set of automatically tagged region mentions from the remaining JCN abstracts. The results mirror those from within the manually annotated corpus and suggest the methods could readily be extended to larger scales. Our method appears to scale well, with over 1,000 brain region concepts appearing in the extended corpus but not the original annotated set.

To increase the value of the data set to the neuroscience community, our results have been incorporated into the NeuroLex database, where work is in progress to display, for example, time-trends of brain region mentions in the JCN alongside other information on each region (Anita Bandrowksi, personal communication). We provide a bulk version of the data suitable for third-party analyses on our website (http://www.chibi.ubc.ca/WhiteText/). As mentioned in the introduction, having brain regions mapped to abstracts is only one step in making full use of the information embedded in the literature. Future work will focus on the linking of brain region mentions to each other and to other concepts such as drugs and diseases. Our eventual goal is to provide computationally rich linkages of brain regions to diverse

neuroinformatics resources.

Chapter 4: Application and evaluation of automated methods to extract connectivity statements from free text

4.1 Introduction

The brain is a vast interconnected network. Each neuron communicates with many others with chemical and electrical synapses to integrate information. Neurons are grouped into named nuclei or layers that make diverse connections across the brain, forming pathways of information flow. This structural connectivity primarily defines its neural function and is frequently used by neuroscientists and clinicians to interpret physiological data. Examples include understanding strokes (Haines, 2004) and interpreting brain imaging results. In addition, neurologists have observed connectivity abnormalities in bipolar (Houenou et al., 2007), autistic (Koshino et al., 2005), Alzheimer's (Stam et al., 2007), and schizophrenia patients (Karlsgodt et al., 2008). A major goal of modern neuroscience is to understand the organization of the brain at all levels in as much detail as possible, and to understand how this networked organization relates to brain function and ultimately behaviour and human health (Sporns, 2011).

The characterization of the connectivity network or wiring diagram of the brain is incomplete (Crick and Jones, 1993). In part, this is due to the complexity of the brain and the difficulty in collecting data. However, we suggest that informatics technologies can be used to leverage

existing knowledge that has already been collected to make new discoveries and guide further experimentation.

In this work we are primarily concerned with “macroconnections”, or connections that can be identified between small brain regions (as opposed to microcircuitry which describes the connections among neurons per se). These macroconnections between groups of neurons are predicted to number between 25,000-100,000 (Bota et al., 2003). This suggests a high level of complexity, though comfortably placed between the more gross levels of brain organization and the microarchitecture which encompasses billions of neurons and quadrillions of synapses (Sporns et al., 2005). Furthermore, this estimated amount of macroconnections is smaller in scale than estimates of the human interactome at 650,000 interactions between 25,000 proteins (Stumpf et al., 2008).

Connectivity between brain regions can be assayed using tract tracing or electrophysiology. Tract tracing typically involves injecting a dye or other tracer (for example, horseradish peroxidase) into one brain region and following the fate of the tracer as it follows axonal pathways (Lanciego and Wouterlood, 2011). Electrophysiological methods use electrical or other stimulation in one site along with electrical recording at a second site to test the functional connectivity of regions. Using these methods a researcher can determine connections that send signals to the region (afferent) or away from the region (efferent). Over many years, thousands of connectivity studies have been performed, each of which typically elucidates at most a few connections. The presence of a deep literature on neuronal connectivity is a major motivation for this work: the data are out there, they just need to be

assembled.

Attempts to turn this huge accumulation of knowledge into an ‘omics’ scale database have been extremely limited, despite the potential value of such a resource. Previous efforts have primarily used manual reviews of the literature to laboriously generate connectivity maps for limited parts of the brain. In 1991, Felleman and Van Essen published a connectivity matrix of the macaque visual cortex covering 305 pathways between 32 areas (1991). Scannell and colleagues followed the same procedure to collate 1139 connections between 65 brain regions in feline cerebral cortex (Scannell et al., 1995). Currently, a large number of collated connections are stored in the Collations of Connectivity data on the Macaque brain database (CoCoMac) (Kotter, 2004). CoCoMac contains detailed information from 413 literature reports regarding 7007 macaque brain regions. A fourth model organism with large scale connectivity data is the rat with over 40,000 reports of connections formalized in the Brain Architecture Management System (BAMS) (Bota et al., 2005). Information is added to these databases manually, and therefore they are accurate but sparse. Currently, the only complete connectome scale database is the neuron-level wiring diagram of *C. elegans*, determined from electron micrographs (White et al., 1986).

We seek to extend and complement manual efforts with automated text mining techniques. Over ten years of efforts to recognize gene and protein mentions and their interactions inspire our work (Blaschke et al., 1999;Jensen et al., 2006). In the gene interaction task, one must extract information from sentences such as “gene A interacts with gene B” (to give a toy example). Despite the difficulty of this task, great progress has been made. A recent survey

found that performance ranges from 29-80% precision and 45-90% recall. The varied results are partially explained by the selection criteria and size of the text corpus used for training and testing (Zhou and He, 2008). A comprehensive evaluation of kernel methods for extracting protein-protein interactions detailed precision and recall values ranging from 45-70% by varying experiment design, dataset and method tested (Tikk et al., 2010). At the second Critical Assessment of Information Extraction systems in Biology (BioCreAtIvE II) the top team was able to extract normalized, directed interaction pairs from full text articles with precision of 37% and recall of 33%. The analogy to brain connectivity is very tight: we wish to extract information from sentences akin to "brain region A connects to brain region B". This related research gives us hope that the approaches applied to extracting gene interaction information can successfully mine connectivity relations.

While attempts to use text mining in neuroscience have been limited, they are instructive. The Neuroscholar project previously explored automated extraction of connectivity data from text (Burns et al., 2007). Burns et al. focus on extraction of detailed parameters and results of a tract-tracing experiment. They manually annotated 1,047 sentences from the Results sections of 21 documents with five labels that describe a tract-tracing experiment. These annotations provided the test and training examples for a conditional random field classifier that was able to label with 80% accuracy. We note that Burns et al. attempted to extract detailed information about connectivity experiments; we seek to extract much less detailed, but still valuable, information. The favourable results of Burns and colleagues' research suggest that a somewhat simplified task may yield even better results.

Given the complexity of the domain we have simplified the problem by limiting our input dataset and output results. We restrict the corpus to abstracts from the Journal of Comparative Neurology because it contains a high frequency of connection reports across many diverse brain region mentions. Abstracts were chosen over full text documents because they are enriched for high level summary statements and are more accessible. Further, we predict connectivity relations between brain region mentions that have been manually annotated instead of automatically recognized spans. In previous chapters we have previously evaluated recognition and normalization of brain region mentions and chose to isolate these steps from the evaluations. We later demonstrate and evaluate a completely automated connectivity extraction system that employs automated recognition and resolution. Finally, we test methods for extracting the presence of connectivity relations but ignore the type or direction of connectivity (afferent, efferent or bidirectional). These generalizations allow a feasible first step to more detailed studies.

We show that text mining methods can be usefully applied to brain connectivity by adapting text-mining approaches previously used to analyze protein networks. Our large manually annotated corpus allowed testing and training of various techniques possible. Beyond the corpus based evaluations we compared a large set of automatically extracted connectivity statements to an existing connectivity database.

4.2 Methods

4.2.1 Annotated data

To test and train text mining algorithms we created a large gold standard dataset. This dataset or corpus consists of abstracts manually annotated by a research assistant for connection verbs, species of study, brain region mentions, and connections between them. We annotated 1,377 abstracts for 4,529 connections and 17,585 brain region mentions. In this design each connection consists of two brain regions, text describing the connection and the associated organism. Two hundred and thirty of the abstracts have been annotated by both annotators. This corpus provides sufficient training examples for machine learning methods. Abstracts for the gold standard corpus were randomly chosen from the Journal of Comparative Neurology.

We have developed guidelines and software for the annotation process. Briefly, our main guidelines are: 1) annotate all brain region mentions whether they are part of a connection or not, 2) annotate all connections and brain regions for all organisms and organism states, 3) do not annotate mentions of white matter tracts. The General Architecture for Text Engineering (GATE) was used by annotators to highlight and connect brain region mentions in text (Cunningham et al., 2002). We have implemented software that uses GATE for abstract importing, corpus management and interannotator agreement computations. Furthermore, a GATE plug-in was created to allow annotation of connectivity relationships between two

brain region mentions.

4.2.2 Co-occurrence

To extract neuroanatomical connections as described by the abstract authors we must at least link two brain region mentions. Our first method, acting as a naïve baseline method, predicts a stated connection between every pair of brain region mentions (Jensen et al., 2006). We evaluate co-occurrence for single sentences and entire abstracts (including title).

4.2.3 Rule based

We created two simple rule based extensions of the co-occurrence technique. The first simply limits co-occurrence extraction to sentences that have a limited number of brain region mentions. The second requires presence of a connectivity related keyword (“afferent”, “efferent”, “projects”, “projection”, “pathway” or “inputs”).

4.2.4 Kernel based methods

Seven advanced kernel based methods were applied to the dataset. These methods were designed for a similar task, extraction of protein-protein interactions from biomedical literature. Each technique uses different features, parameters and kernel functions. Implementations were brought into a common evaluation framework by Tikk and colleagues (Tikk et al., 2010). The methods are categorized according to the type of features extracted from the sentences. Four syntax tree based methods use different techniques to compare the sentence parse trees (Collins and Duffy, 2001;Vishwanathan and Smola, 2002;Moschitti,

2005;Kuboyama et al., 2007). Going beyond syntax parses, the all-paths graph kernel (Airola et al., 2008) and k-band shortest path spectrum kernel (Tikk et al., 2010) employ dependency parse information. Lastly, the shallow linguistic kernel (SL) employs only shallow parsing information such as word occurrences and part-of-speech tags (Giuliano et al., 2006). We employed this framework to benchmark each of the kernel based methods on the brain region connectivity task. Of the nine methods described by Tikk et al. we were able to successfully test seven, including the three top performing kernels. For every method, the same parameter sets used by Tikk and colleagues were tested on our corpus.

4.2.5 Experiment setup

We evaluate connection extraction independently of the previously described methods for automated brain region recognition. This is done by providing the manually annotated brain region mentions to the relation extraction algorithm. Under this design the extraction task only requires correct linking of brain regions mentions.

To find the optimal method while avoiding over-fitting, method comparison and selection was performed on the 1,146 abstracts annotated only by the primary annotator. Results for the kernel methods were computed using ten-fold cross-validation. Each sentence became an input instance for the kernel methods (including article title). Sentences of an abstract were not split between training and testing sets (document level split).

4.2.6 Evaluation

Performance is measured against the number of true connectivity relations that are annotated completely within the evaluation scope. The rule and co-occurrence based methods can operate at the abstract or sentence level while the kernel methods are limited to sentence level scope. Precision is computed as the proportion of predicted relations that are correct, and recall is the proportion of true relations that are predicted by the method. The f-measure or f-measure is the harmonic mean of these two values, providing a balance of both. We also compute the area under the receiver operating curve where applicable (AUC). This measure uses a ranked list of predictions with descending classification prediction scores (scores represent distance from the discrimination hyperplane and approximate confidence in the prediction). This ranking allows computation of the true positive and false positive rates for a range of discrimination thresholds. Previous experiments have found the AUC measure to be more robust and stable than f-measure for interaction mining (Tikk et al., 2010).

4.2.7 Comparison to existing connectivity database

Normalization of brain region mentions to brain region concepts in formalized lexicons was targeted to the BAMS atlas (Swanson, 1999). BAMS was chosen because it's wealth of curated rat tract tracing studies (Bota et al., 2005). In addition, rat is the most commonly studied species in the corpus. The previously described Bag of Stems resolver was applied with all Mention editors employed, including those that create resolutions to enclosing regions (see Chapter 3). The lexical information in BAMS was expanded with synonym

information to increase normalization performance. All possible normalized parings are evaluated when a mention maps to more than one region. Connections in the BAMS connectivity matrices were up-propagated. The up-propagation procedure ensures that if there is a connection between regions A and B then all enclosing regions of A and B are also connected. Self connections extracted from literature were ignored. The LINNAEUS species tagger was employed to recognize species names in the abstracts (Gerner et al., 2010).

4.3 Results

We annotated 4,276 connectivity relations across the complete corpus of 1,377 abstracts. To gauge interannotator agreement, a second curator annotated a random subset of 231 documents. Roughly 80% of the second curator's annotations matched the main curator (79.5% recall at 82.3% precision). Unlike the automated methods that predict relations between given brain region mention spans, this evaluation required both annotators to highlight the same brain region mention spans. By removing this restriction and allowing partially matching spans, the precision and recall reach 93.9% and 91.9% respectively.

The co-occurrence based analysis reveals the proportion of brain region mention pairs that are co-mentioned and described as connected. At the abstract level 2.2% of all possible brain region pairings form connectivity statements. While many relationships can be formed between any two brain regions, co-occurrence assumes the relation is a connectivity statement. Often this is incorrect at the abstract level with precision of 2.2% at 100% recall, and a combined f-measure score of 4.3%. Within a sentence, co-occurrences between all pairs predicts connected pairs at 13.3% precision and 72.0% recall (remaining relations

span sentences). This level of recall means that over ¼ of all annotated connectivity relations are formed with regions in different sentences. Due to the difficulty in extracting connections spanning sentences, all of the below evaluations are performed at the sentence level with the relations spanning sentences excluded (under this evaluation framework sentence level co-occurrence provides 100% recall).

We tested two simple modifications of the sentence level co-occurrence technique. The first reduces co-occurrence predictions to sentences with a limited number of brain region mentions. By extracting co-occurring pairs from sentences with only two brain region mentions, precision reaches 23.1% and 17.2% recall (f-measure = 19.7%). This means that an average sentence with two brain region mentions is reporting a connection in almost one out of four cases. By varying this threshold the f-measure increases until sentences with 6 or more brain region mentions are included. We observed that some of these larger sentences merely list brain regions involved in the study and not their relationships. By limiting at 5 brain region mentions or less per sentence, co-occurrence provides 18.8% precision and 66.1% recall (f-measure = 29.3%). The second rule tested requires the sentences contain one of six connectivity related keywords (afferent, efferent, projects, projection, pathway and inputs). This keyword based rule increases recall to 17.4% and precision to 92.7% (f-measure = 29.4%). We created a new approach named “Keyword 5-threshold” by combining these two rules. This again provides improvement with f-measure reaching 34.1% (precision = 23.7%, recall = 60.8%). As expected, rule based methods increase precision at the cost of lower recall when compared to unrestricted co-occurrence.

We applied seven methods for extracting protein-protein interactions to our connectivity relation dataset. While the methods were designed for a different type of biomedical relation, they do not require any changes to extract undirected relations between other entity types. The cross-validation results on the testing dataset (1,146 abstracts) are provided in Table 11. For each method, the parameter set with the highest AUC score is shown. The parameter sets range in size and are reproduced from Tikk et al. without modification (primarily grid searches of support vector machine settings). F-measure scores for all of the seven methods outperform unrestricted co-occurrence based analysis for at least one parameter set. The simple rule based methods outperform the more complex Partial Tree and Subset Tree based methods. While all of the syntax tree based methods are outperformed by the Keyword 5-threshold approach, they provide much higher precision than recall. When ranked by AUC, the SL kernel performs best with a 58.3% f-measure and an AUC of 88.9%. The All Paths Graph and *k*-band shortest path spectrum kernel methods rank a close second and third with similar scores.

For further application we choose the SL method due to its accuracy, speed and single parameter set (global n-gram = 3 and local window = 2). Unlike the other kernel methods the SL method uses only shallow linguistic information at the local (neighbouring words) and global sentence levels to predict relationships (Giuliano et al., 2006). This information forms feature vectors that are used to train a support vector machine classifier (scalar product kernel). The performance of SL is consistent on the complete set of 1,377 abstracts with f-measure of 0.592. Figure 5 displays the resulting ROC curve (AUC = 0.899). For large-scale application, we applied the SL classifier to candidate sentences extracted from a set of 12,557

abstracts from the Journal of Comparative Neurology (covering 1975-2011). While previous evaluations used manually annotated brain region spans for input, this combination of automatic brain region recognition (see Chapter 2) and relation extraction will result in a higher error rate. To estimate the combined effects we trained the SL classifier on the manual annotations of the 1,146 test abstracts and test on the remaining manually annotated set of 231 abstracts. The automatic brain region annotations on the test set are used as input to the SL classifier instead of the manual annotations. The result is 323 of 770 predicted connections exactly matching an annotated connection (precision = 41.9%, recall = 52.4%, f-measure = 46.6%). While not obtained in a cross-validation framework, this experiment presents a good estimate of accuracy for the combined pipeline. In the set of 12,557 abstracts, application of the previously described automatic brain region recognizer provided 33,466 sentences that mention two or more brain regions. Within these sentences, SL predicted 18% of the 156,484 possible brain region pairings to be connectivity relations. Of these predicted relations, 9,676 are in an abstract that mentions rat and can be evaluated against BAMS. Figure 6 shows the progression from abstracts to predicted connectivity relationships.

Table 12 presents the ten most and least confident rat connectivity relations. Classication is approximated with the SL prediction score (distance to classifiying hyperplane), with highest values representing the cases closest to positive training examples. Two of the most confident predictions are extracted from an article title and have the same form (ranks 1 and 5). The sentences containing top predictions are shorter on average (192 characters) than the sentences with least confident predictions (282 characters); suggesting sentence complexity affects the prediction results. Of these twenty examples only two are clearly false positive

predictions (ranks 9764 and 9758) while several others point to errors in previous automated steps. An abbreviation expansion error appears in the sentence containing the relationship ranked 9762. Organism identification errors occur in two sentences that refer to connections in monkey although the associated abstracts mention rat neuroanatomy (ranks 8 and 9760). The mentions of "internal capsule" (rank 9766) and "Met-enkephalin" (rank 4) are incorrectly predicted as brain region mentions (our definition of a brain region excludes fibre tracts). We manually compared these twenty results to the BAMS system and found it difficult to map the mentioned regions to those in BAMS. For example, “retrosplenial dysgranular cortex” and “dorsal medullary reticular column” were not found in BAMS. Connections in BAMS were found for several of the relationships but between enclosing regions (ranks 9767, 7, and 5). The low confidence relationship ranked 9761 shows an incorrectly marked brain region mention: “central olfactory cortical”, in addition the enclosing sentence contains connections not found in BAMS. For example, the stated connection between the anterior olfactory nucleus and piriform cortex is correctly extracted (rank 5795, score = 0.83) but the connection is not curated in BAMS. These results from the SL method are very encouraging and motivate a larger evaluation.

We compare our results to an existing connectivity database (BAMS) to gauge accuracy of connections extracted from the unannotated set of 12,557 abstracts. Compared to the manual annotations, this is a less precise evaluation because BAMS does not cover the complete literature and is limited to rat studies (Bota et al., 2005). In addition, resolution errors resulting from linking brain region mentions to target regions in BAMS reduces accuracy (see Chapter 3 for details). For example, 12% of mentions are resolved to more than one

brain region due to ambiguous synonyms. To benchmark the BAMS evaluation metric we first tested it on manually curated connectivity relations. Our process first extracts abstracts that mention rats and resolves the brain region mentions to the BAMS lexicon. These rat connectivity relationships are then compared to the BAMS connectivity matrix. In this framework only 167 manually annotated connectivity relations are resolved with 70.5% having a connection in BAMS. In contrast, the set of 2,617 brain region pairings not annotated as connections but co-occur in sentences are connected in BAMS at 49.8%. This is not unexpected because co-occurring regions may be connected, but the author is not stating that in the sentence. In the unannotated set of rat abstracts 2,688 predicted connectivity relations are successfully resolved and 63.5% are connected in BAMS (Figure 6). For comparison, the remaining set of co-occurring brain region pairs are connected in BAMS at a rate of 51.1%. We note that the extracted relationships are between larger brain regions than those in BAMS. The average number of enclosing or parent brain regions for a connected pair in the BAMS matrix is 9.6. The literature extracted connections are shallower with an average number of enclosing regions of 7.9.

We suspect that more recent reports of connectivity are of higher quality when compared to the BAMS database. Guidance is provided by a study of different eras of tract tracing techniques that revealed large improvements in accuracy (Bota et al., 2003). Bota and colleagues found that limbic system connections observed through axon degeneration (Nauta, 1952) experiments are 60% accurate. In contrast, newer methods first applied in 1987 to exploit axonal transport are more accurate with over 90% considered valid. By splitting our corpus into documents published before and after 1987 we tested for a similar signal that

separates eras of experimental techniques. In agreement with the manually quantified trend, we observe an increase from 59.4% to 65.6% in the rate of connectivity statements validated in BAMS ($p = 0.00071$, hypergeometric test). We note the specificity of regions involved in the connections also increases while resolution rate is unchanged.

Although we extract large sets of relations the number of unique resolved brain regions in the unannotated set is only 433 regions. In comparison 633 unique regions are connected in BAMS. Further, each unique predicted connectivity relation occurs more than twice on average in our text mined set.

We further evaluate the results in the form of connectivity matrices that count the number of connections extracted for each brain region pair. In this framework, 54.7% of the predicted connections from the unannotated set of rat abstracts are connected in BAMS. From a recall perspective, 3.2% of BAMS connections are connected in the literature based matrix. By thresholding the literature matrix to two or more relation mentions, precision reaches 65.9% while recall drops to 1.4% (Table 13). This accuracy is near the 67.5% precision of the hand annotated set of connections. Precision gradually increases as the threshold increases, eventually reaching 100% for nine connections that are extracted at least 12 times. Further, we note the specificity of the connections increase with the average number of enclosing regions reaching 10.2 when thresholded at 12 occurrences. The region pairs not predicted to form connectivity relations have precision of 33.7% and recall of 9.3%. Again, this level of precision results from co-mentioned regions that are connected in BAMS but the author is not specifying that in the sentence. Further, the higher recall value results from the much

larger set of pairings (6079 compared to 1286 SL predicted parings). From a co-occurrence perspective we found that brain regions that co-occur in eight or more sentences recall 1.6% of the BAMS connections at 66.4% precision. Interestingly, this naive co-occurrence based method performs at par to the SL method that extracts direct connectivity statements. As the threshold is increased from 8 co-occurrences precision continues to gain, suggesting a large number of co-occurring mentions can predict connectivity as well as a few connectivity statements (Table 13).

Table 11 Training set cross-validation results

Precision, recall, f-measure and AUC values are averaged across the ten cross-validation runs. The "Parameter Sets" column gives the size of parameter sets tested. For a given method, results are shown only for the parameter set with the highest scoring AUC value (F-measure when AUC is not applicable).

Kernel	Precision	Recall	F-measure	AUC	Parser Type	Parameter Sets
Co-occurrence	13.3%	100.0%	23.5%		none	1
Subset Tree Kernel	44.2%	20.8%	28.1%	74.8%	syntax	12
Co-occurrence 5-threshold	18.8%	66.1%	29.3%		none	25
Partial Tree Kernel	43.3%	23.1%	29.8%	75.2%	syntax	12
Keyword Co-occurrence	17.4%	92.7%	29.4%		none	1
Spectrum Tree Kernel	37.4%	26.1%	30.2%	72.9%	syntax	21
Subtree Kernel	40.7%	25.2%	30.8%	74.6%	syntax	12
Keyword 5-threshold	23.7%	60.8%	34.1%		none	25
k-band Shortest Path Spectrum	46.8%	70.5%	55.8%	86.7%	dependency	288
Shallow Linguistic Kernel (SL)	50.3%	70.1%	58.3%	88.9%	part-of speech tagger	1
All-paths Graph Kernel	60.4%	57.9%	58.4%	88.4%	dependency	4

Table 12 Top and bottom predicted relations ranked by SL classification score

Each predicted relationship represents a single row (most sentences have many predicted relationships). Brain region mentions that participate in the extracted relationships are marked in bold text.

Rank	Sentence	Score	Reference
1	**Trigeminal** projections to **hypoglossal and facial motor nuclei** in the rat.	3.47	(Pinganaud et al., 1999)
2	The **cortical** projections to **retrosplenial dysgranular cortex**(Rdg) originate primarily in the infraradiata, retrosplenial, postsubicular, and areas 17 and 18b cortices.	3.34	(van Groen and Wyss, 1992)
3	The **thalamic** projections to **retrosplenial dysgranular cortex**(Rdg) originate in the anterior (primarily the anteromedial), lateral (primarily the laterodorsal), and reuniens nuclei.	3.33	(van Groen and Wyss, 1992)
4	Our results indicate that the **centromedial amygdala** receives **Met-enkephalin**(ENK) afferents, as indicated by the presence of mu-opioid receptor(MOR) , delta-opioid receptor(DOR) , and Met-enkephalin(ENK) fibers in the central(CEA) and medial(MEA) , originating primarily from the bed nucleus of the stria terminalis (BST) and from other amygdaloid nuclei.	3.32	(Poulin et al., 2006)
5	**Thalamic** projections to **retrosplenial cortex** in the rat.	3.28	(Sripanidkulchai and Wyss, 1986)
6	The **thalamic** projections to **retrosplenial granular a cortex**(Rga) originate mainly in the anterodorsal (AD) and laterodorsal (LD) nuclei with sparse projections arising in the anteroventral (AV) and reuniens nuclei.	3.28	(van Groen and Wyss, 1990)
7	Finally, reciprocal projections from the **hypothalamus** to the **intergeniculate leaflet**(IGL) arise from neurons in the retrochiasmatic area, suprachiasmatic nucleus(SCN) , and adjacent anterior hypothalamus.	3.27	(Card and Moore, 1989)
8	The **amygdala** projects to **orbitofrontal cortex**(OFC) by both a direct amygdalocortical (AC) pathway and an indirect pathway through mediodorsal thalamus.	3.27	(Miyashita et al., 2007)
9	The **rostral part of the medial accessory olive** projects to **zebrin-positive areas**, in particular to the P4+ band of the anterior lobe and lobule VI and to the P5+ band of the posterior lobe, indicating that C2 has two noncontiguous representations in the SL and crus 1.	3.21	(Pijpers et al., 2005)
10	**Cortical** projections to **retrosplenial granular a cortex**(Rga) originate in the ipsilateral area infraradiata, the retrosplenial agranular and granular b cortices, the ventral subiculum, and the contralateral retrosplenial granular a cortex(Rga) .	3.19	(van Groen and Wyss, 1990)

Rank	Sentence	Score	Reference
.....	9747 relationships		
9758	In the dorsal horn, terminals or preterminal axons were found in the dorsal horn marginal zone (lamina I), the substantia gelatinosa (lamina II), the **nucleus proprius** (laminae III and IV--the most consistent projection), **Clarke's column** (lamina VI), and the dorsal gray commissure.	9.08E-004	(Nunez et al., 1986)
9759	In addition, tracer injections into anteromedial(AM) , ventromedial(VM) , and **ventrolateral**(VL) revealed dense clusters of labeled neurons in layer VI of the **medial agranular (Agm) zone**, which corresponds to the MI whisker region.	8.98E-004	(Alloway et al., 2008)
9760	Additionally, the **pontine** indoleamine-containing cells in M. mulatta extended laterally through the tegmentum such that they were often adjacent to catecholamine-containing neurons of the **locus coeruleus complex**.	7.84E-004	(Schofield and Everitt, 1981)
9761	The anterior olfactory nucleus (AON) is a **central olfactory cortical** structure that has heavy reciprocal connections with both the olfactory bulb (OB) and **piriform cortex**.	7.23E-004	(Illig and Eudy, 2009)
9762	**Amygdala** infusion labeled neurons in the endopiriform nucleus, temporal cortex, piriform cortex, paralimbic cortex, hippocampus, subiculum, ento recombinant human(rh) inal cortex, amygdala, basal forebrain, thalamus, hypothalamus, substantia nigra, pars compacta, raphe, and **pontine parabrachial nuclei**.	6.99E-004	(Sobreviela et al., 1996)
9763	The sparse reciprocal connections to the other **amygdaloid nuclei** suggest that the central nucleus does not regulate the other amygdaloid regions but, rather, executes the responses evoked by the other amygdaloid nuclei that innervate the **central nucleus**.	5.46E-004	(Jolkkonen and Pitkanen, 1998)
9764	The majority of the Endomorphin 1(EM1) / Fluoro-Gold(FG) and endomorphin 2(EM2) / Fluoro-Gold(FG) double-labeled neurons in the **hypothalamus** were distributed in the **dorsomedial nucleus**, areas between the dorsomedial and ventromedial nucleus, and arcuate nucleus; a few were also seen in the ventromedial, periventricular, and posterior nucleus.	4.36E-004	(Chen et al., 2008)
9765	Projections from the **dorsal medullary reticular column**(DMRC) are largely bilateral and are distributed preferentially to the **ventral subdivision of MoV**, to the dorsal and intermediate subdivisions of VII, and to both the dorsal and the ventral subdivision of XII.	2.91E-004	(Cunningham and Sawchenko, 2000)
9766	Two additional large projections leave the medial forebrain bundle in the hypothalamus; the ansa peduncularis-ventral amygdaloid bundle system turns laterally through the **internal capsule** into the striatal complex, **amygdala** and the external capsule to reach lateral and posterior cortex, and another system of fibers turns medially to innervate medial hypothalamus and median eminence and form a contrelateral projection via the supraoptic commissures.	2.87E-004	(Moore et al., 1978)

Rank	Sentence	Score	Reference
9767	In animals with injected horseradish peroxidase(HRP) confined within the main bulb, perikarya retrogradely labeled with the protein in the **ipsilateral forebrain** were observed in the **anterior prepyriform cortex horizontal limb of the nucleus of the diagonal band**, and far lateral preoptic and rostral lateral hypothalamic areas.	3.36E-005	(Broadwell and Jacobowitz, 1976)

Table 13 Aggregate connectivity results from several methods and relation sets

Anatomical depth is the combined number of enclosing parent neuroanatomical structures for each brain region forming the pair. Threshold values represent the minimum number of occurrences required for a relationship to be considered a connection.

Relation Set	Method	Threshold	Anatomical Depth	Connections	Precision	Recall	F-Measure
Positive annotated	Curation	1	8.7	200	67.50%	0.61%	1.22%
Negative annotated	Curation	1	8.7	1606	41.91%	3.06%	5.71%
Positive predictions	SL Kernel	1	8.4	1286	54.70%	3.20%	6.05%
Positive predictions	SL Kernel	2	8.4	454	65.90%	1.40%	2.74%
Positive predictions	SL Kernel	12	10.2	9	100.00%	0.04%	0.08%
All pairings	Co-occurrence	1	8.3	6474	34.00%	10.01%	15.47%
All pairings	Co-occurrence	2	8.3	2865	44.96%	5.86%	10.37%
All pairings	Co-occurrence	8	8.2	515	66.41%	1.56%	3.04%
All pairings	Co-occurrence	16	8.4	189	71.43%	0.61%	1.22%

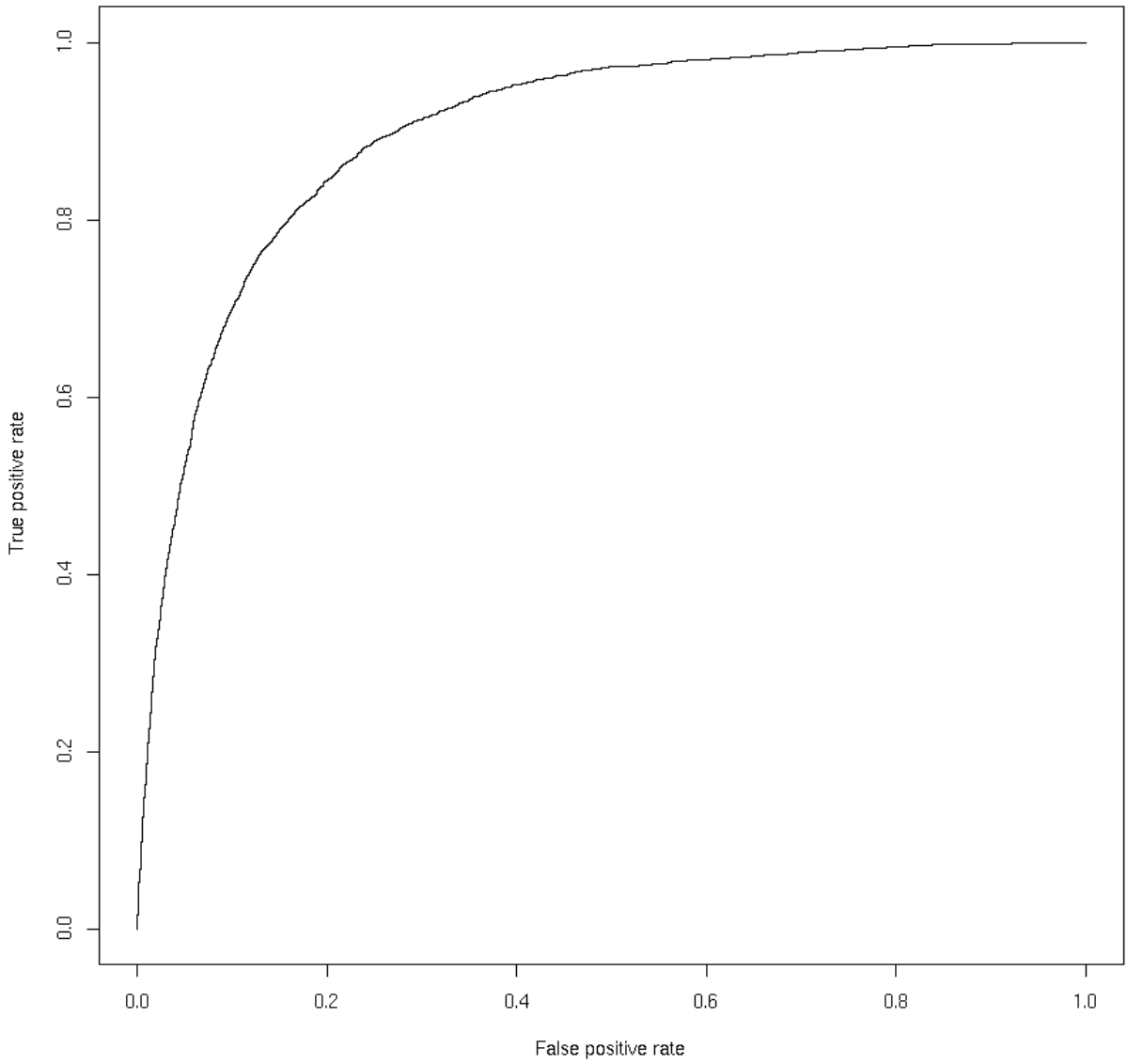

Figure 5 Summary ROC curve for the SL method (AUC = 0.899)

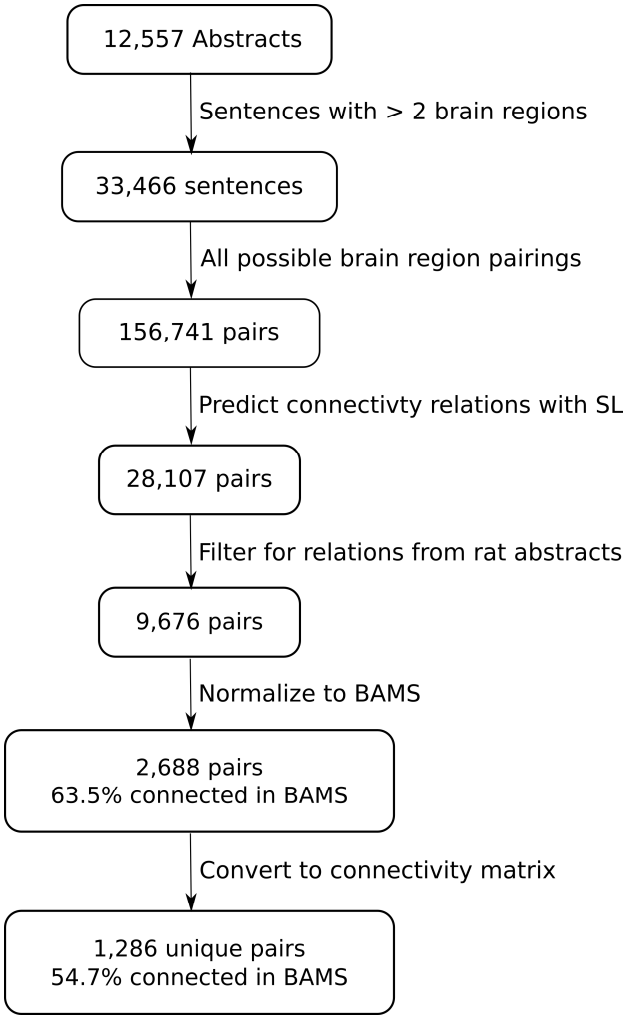

Figure 6 Flow chart depicting the processing steps

4.4 Discussion

We demonstrate a complete system for extracting connectivity statements from biomedical abstracts. The method provides high recall of manually annotated connectivity relations described in single sentences. Precision of predicted relations is 63.5% when evaluated against an independent source of rat connectivity. This result compares well to the 70.5% precision of manually annotated connectivity statements. By processing a large dataset

we found precision increases with the recency and frequency of the extracted relationships.

A limitation of our work is that we assumed the connectivity statements are bidirectional although most of the relationships we extracted have a direction described. In addition, some reports of disconnection are described (region A does not project to region B). Extracting this information by extracting keywords such as “afferent”, “not” or “input” will require future work. These relationship modifiers are manually annotated in the corpus and can be used to design more complex rules.

Our methods that focused on the sentence level cannot extract the large number of relations that span sentences. When these connections are taken into account the SL parser provides only 51.7% recall of annotated connections. Application of advanced natural language processing techniques may be necessary to bridge the sentences (e.g. anaphora resolution).

The comparison of seven cutting edge kernel based approaches mirrored the previous results from the protein interaction relationship extraction domain (Tikk et al., 2010). Several of the kernel methods have lower performance than our simple rule based technique. Effort spent crafting more complex rules may yield higher precision at the cost of lower recall. The top three kernel methods (SL kernel, All-paths graph, k-band Shortest path spectrum kernel) all have similar accuracy (AUC and f-measure scores) but vary in precision and recall. This difference suggests higher performance may be achieved by combining the methods.

Our results suggest a larger set of input abstracts will yield a larger number of precise connections. The largest possible extension set is Medline with over 10 million abstracts and

120 million sentences. Tikk and colleagues calculated that the SL parser could process all of Medline in 141 days (Tikk et al., 2010). A two step process may reduce runtime and increase accuracy by first identifying abstracts with connectivity statements and then extracting the specific connections with SL. Another targeted approach is to extend the analysis to other journals that often publish tract tracing studies (e.g. Journal of Chemical Neuroanatomy).

In natural language processing, it has been observed that simple statistical models (e.g. co-occurrence) outperform more complex models based on less data (Halevy et al., 2009). Indeed, in our corpora we found that brain region pairs with many co-mentions tend to be connected. In our evaluations this simple technique produces a larger set of potential connections with reasonable precision. Although this will produce a larger set of results than the SL method, it does not target connections that can be directly curated in light of experimental evidence because the co-mentions may or may not describe connectivity. Further, such co-occurrences may result from region proximity or popularity that may influence research attention in both the literature and in BAMS. However, such co-occurrence networks show valuable areas of focus when combined with co-mentions of genes and diseases (Hayasaka et al.).

Under one third of the extracted relationships were successfully mapped to a brain region concept pair in a standardized lexicon. This has been studied in previous chapters but for relationship extraction the resolution rate is greatly reduced as both pairs of a connectivity relation must be mapped. Further, it appears that regions forming connectivity relations are harder to resolve on average. For this work we managed to double the resolution rate to the

BAMS lexicon by adding synonyms. Additional work to improve the lexicons will lead to better resolution of connectivity statements, allowing validation and linking to other resources.

For our evaluation to an outside database we focused on BAMS (Bota et al., 2005). While rat is the most frequent mentioned organism, other evaluations could compare the connectivity results to the Collations of Connectivity data on the Macaque brain (CoCoMac) (Kotter, 2004) or the Avian Brain Circuitry Database (ABCD) (Schrott and Kabai, 2008). Beyond evaluation, our dataset and method can provide a large set of extracted connectivity relationships for other species specific databases.

In conclusion, we provide the first application of large-scale text mining to neuroanatomical connectivity extraction. We demonstrated that machine learning tools designed for extraction of protein-protein interactions are generalizable to mining brain region connections. From an information retrieval perspective, our large set of uncurated connections can aid neuroscientists in forming hypotheses and models. Future work will be aimed at further evaluating and disseminating the results before extending the analysis.

Chapter 5: Large-scale analysis of gene expression and connectivity in the rodent brain: insights through data integration[4]

5.1 Introduction

Understanding gene function requires the analysis of interactions among them, and ultimately unraveling the function of the genome will require comprehending how all of the parts interoperate in complex networks. An analogous situation exists for the brain and its regional connectome (Bota et al., 2003;Sporns et al., 2005;Lichtman and Sanes, 2008;Biswal et al., 2010;Sporns, 2011). Given the relationships between these two systems (genome and connectome), as well as the fact they are both complex networks, it is natural to ask how analysis of one can inform understanding of the other. Indeed, the integrated analyses of the connectome with other modalities will be critical to understanding brain function. In this chapter our modality of interest is gene expression, for which extensive information exists.

It is obvious that the connectome is related to the genome. Axon pathfinding, target recognition, synapse formation and plasticity are tightly controlled by gene expression (Ressler et al., 2002;Polleux et al., 2007). The function of synapses requires the coordinated

[4] A version of this chapter has been published. French L, Tan PPC and Pavlidis P (2011) Large-scale analysis of gene expression and connectivity in the rodent brain: insights through data integration. Frontiers in Neuroinformatics. doi:10.3389/fninf.2011.00012

expression of genes directing the synthesis of neurotransmitters in the presynaptic cell and of receptors in the postsynaptic cell. Because high throughput experimental technologies for studying the genome are well developed, in many ways our understanding of gene expression and gene networks is better than for the connectome (though this situation is changing rapidly). This allows the collection of large data sets describing gene expression patterns at high levels of resolution. It is increasingly feasible to use this molecular level information to elucidate neuroanatomy.

Analysis of connectomes with transcription data began with the nematode *C. elegans* because neuron-level connectivity and gene expression levels are known. (White et al., 1986;Harris et al., 2010). Neuron-level gene expression data in *C. elegans* is not available for all genes, but there is enough to perform reasonably large-scale analyses. The earliest study integrated the connection and expression profiles of 280 neurons and 292 genes (Varadan et al., 2006). Varadan and colleagues employed a systems-based approach to discover logical gene expression based rules that predict connectivity. Within the resulting gene modules they found high levels of "multivariate synergy", suggesting statistically interacting genes were more important than single genes. The authors extracted several gene sets that correlate expression in pre and post-synaptic neurons to presence of gap and chemical synapses. Interestingly, gene sets which contained the most information about the formation of synapses included cell adhesion molecules, transcription factors and axon guidance cues.

Kaufman et al. performed a similar analysis (Kaufman et al., 2006). They found a more general statistical relationship between gene expression and connectivity. Their analysis

employed a co-variation correlation assay, also known as a Mantel test. The Mantel test correlates similarity or distance measures across common objects (in this case, neurons). The Mantel correlations found by Kaufman et al. were up to 0.18. This signal, while statistically significant, is not strong enough to allow prediction of connectivity from gene expression. Using an optimization method, Kaufman et al. identified a set of 15 genes whose expression patterns carried the most information about connectivity. Similar to the results of Varadan et al. (2006), they found that a statistically significant number of these were previously linked to synaptogenesis, neuron type, axon guidance and development.

A third *C. elegans* study, by Baruch et al. (2008) focused on finding relationships between gene expression and certain aspects of synapse formation (Baruch et al., 2008). They used expression profiles to model the type of synapse (e.g., electrical or chemical) between connected neurons. Like Varadan et al. (2006) they employed a machine learning method to find gene expression-based logical rules, and the genes found to be most predictive of connection type often had known functional roles in neural development.

Similar analyses are starting to appear for the mammalian brain, though in terms of data the situation is the opposite of that for the worm: gene expression is more fully described than connectivity. Dong et al. (2009) provided a fascinating glimpse into the relationships between brain wiring and gene expression in the mammalian brain (Dong et al., 2009). They studied the Allen Mouse Brain Atlas (ABA) for spatial gene expression profiles that segmented the hippocampal field CA1 along its longitudinal axis. Nine of the genes that segmented the CA1 field had concordant expression patterns in the lateral septal nucleus,

apparently reflecting the patterns of projections between the respective dorsal and ventral aspects of the two regions. Dong et al. (2009) were able to interpret the CA1 segmentation from the perspective of brain function and connectivity. They noted that the ventral half is linked to goal-oriented and autonomic response while the dorsal half plays roles in navigation.

A limitation of previous studies integrating gene expression and connectivity is the challenge of interpreting the patterns observed in terms of other parameters such as cellular composition of different brain regions. In the current chapter, we extend our earlier work, starting with a directed search for expression patterns of interest. We hypothesized that expression patterns that strongly distinguish brain regions from each other might be functionally relevant and potentially related to connectivity. We were specifically interested in gene pairs with expression patterns showing strong negative correlations across multiple brain regions. We then use connectivity data as well as information on cell-type-specific gene expression to further dissect and ascribe biological meaning to the patterns we identified. In addition to identifying a novel pattern of gene expression in the mouse brain, our analysis serves as a demonstration of how a complex gene expression pattern can be dissected using multiple data types including connectivity.

5.2 Materials and Methods

5.2.1 Neuroanatomical connectivity data

For neuroanatomical connectivity knowledge, we used the Brain Architecture Management system (BAMS). BAMS contains extensive information about neural circuitry curated from neuroanatomical atlases and tract tracing experiments (Bota et al., 2005;Bota and Swanson, 2010). The version of the BAMS database we use contains 7,308 structural connections between 961 rat brain regions and is accessible via bulk download (http://brancusi.usc.edu/bkms/xml/swanson-98.xml). Instead of parsing the original XML we used a converted semantic web version created by John Barkley (http://sw.neurocommons.org/2007/kb-sources/bams-from-swanson-98-4-23-07.owl). The BAMS system stores information on projection strength, number of reports, report citations and absence of connections but it is not available in the database version we obtained. However, directions of the neuroanatomical connections are known, allowing splitting of our analysis between incoming and outgoing connection profiles.

The BAMS curators comprehensively studied the bed nuclei of the stria terminalis (BNST) and indicate that its connection matrix is considered complete (Bota and Swanson, 2010). We were concerned that this unusually well-studied region would bias our results, as it has more known connections than the other regions (we considered regions that lack a documented connection to be unconnected). For example, it has over seven times the average number of outgoing connections. To reduce this bias in the dataset, we removed connection

information for the BNST and its subparts. We do not suspect the quality of these connections but wished to prevent one well-characterized region from being overrepresented. We believe the complete connectivity matrix of the BNST will be valuable for future focused analysis.

5.2.2 Gene expression data

We employed the expression energy quantifications of the ABA images. For each image the expression energy of every voxel is defined as the product of expression area and expression intensity (Ng et al., 2009). Pixels are averaged within voxels and brain regions to provide a single expression energy value for each brain region. To reduce computation time and filter genes of low and constant expression values we restricted our analysis to genes for which ABA has expression patterns in coronal sections. This set of 4261 image series (3976 genes) were assayed by ABA in the coronal plane because they showed marked regional expression patterns in the sagittal plane (Ng et al., 2009). Most “housekeeping” genes which tend to have widespread expression are not present in the set. Some genes were represented by more than one imageseries (that is, there are replicate data sets in the Allen Atlas), which were kept separate in our analysis. To create a single expression profile for a set of genes we averaged the expression values per region.

For analysis of expression data alone, we used 150 non-overlapping ABA regions. When connectivity data was used the regions were limited to those for which we had connectivity data: 112 regions for outgoing, 141 for incoming connectivity and 142 resulting from joining

the two.

5.2.3 Neuroanatomical matching and selecting

The names of brain regions are formalized in hierarchies both in BAMS (Swanson, 1999;Bota and Swanson, 2008) and the ABA data (Dong, 2007), but the schemes are not identical. In addition, the BAMS dataset contains information at a finer neuroanatomical resolution than ABA. To maximize the use of connectivity information, we created connection profiles of coarser scale by using an up-propagation procedure. Up-propagation maps the brain region to its parent region until the desired level in the neuroanatomical hierarchy is reached. This procedure was applied to all connection pairs in BAMS. For example, a connection between region A and region B will be expanded to the set of all possible connections between the neuroanatomical parents of both region A and region B. To prevent enrichment of up-propagated connections we kept regions that had zero connections to the ABA mapped regions.

Although the two datasets are at the brain region level, the organisms differ. The rat brain with a wealth of neuroanatomical information is bigger and for some regions like the cerebellum, more complex. In contrast, genetics and molecular research is more commonly performed on the smaller mouse brain. For this work we considered neuroanatomical differences between the mouse and rat to be minor at the level of granularity we used (Swanson, 2003); for example, the Paxinos mouse atlas was guided by several rat brain atlases (Paxinos and Franklin, 2008), and brain regions names largely coincide between the two. These common names allowed quick lexical mapping for most of the regions. To

join the two data types we mapped nomenclatures manually. We used primarily a region's name, then secondarily its parent region and spatial borders to pair brain regions. The mappings for the Allen Brain regions are provided in Appendix B.

The neuroanatomical atlases from ABA (Dong, 2007) and BAMS (Swanson, 2004) provide information on which brain regions are neuroanatomical children or parts of others. These relations create correlations in the gene expression profiles and the connectivity data (due to up-propagation). To negate this effect we used only 149 of 207 Allen brain regions for the primary region list. These remaining regions have no neuroanatomical subparts in the ABA dataset.

The Allen Atlas provides a differing grouping of regions than the BAMS hierarchy. The superior colliculus is one example. The ABA divides its regions into motor and sensory areas, while the BAMS atlas groups the regions into optic, gray and white layers. Differences were resolved by creating "virtual regions" in the BAMS atlas space that contained the corresponding subregions of the Allen Atlas. The connectivity profiles of the mapped regions were joined using a logical OR operation to provide the virtual region's BAMS connections. For example the superior colliculus sensory related virtual region has all of the BAMS connections of the zonal, optic and superficial gray layers. In addition to the superior colliculus, virtual regions were created for the pallidum medial region and nucleus ambiguus.

After mapping of brain regions, the ABA data is an x (number of regions in the ABA) by y (number of genes) matrix, and the BAMS connectivity data is a square w (number of regions in BAMS) by w (region) matrix (Figure 13). The two matrices are not directly

comparable because the number of regions in BAMS is greater than those in ABA ($w>x$). Rather than discarding all information from regions which lack expression information, we use the x by w submatrix of the BAMS data. Thus each of the x regions has a y-dimensional expression vector and a w-dimensional connectivity vector. This maximizes the use of connection information, but we note that the connectivity profiles include information from regions for which we lack expression information.

5.2.4 Statistical analysis

To compare expression energy to spatial location and connectivity degree we compute Spearman rank correlation coefficients (ρ). Statistical significance was established by resampling 1,000 gene sets of the same size to generate empirical null distributions. This provides the probability that an equally sized gene set randomly chosen from the set of all genes scores a higher correlation. We used linear regression for computing partial correlation coefficients. Principal component analysis was performed after rescaling the gene profiles to a common mean and variance. We employed the complete-linkage agglomeration method for hierarchical clustering with the Euclidean distance function.

5.2.5 Cell-type enriched gene lists

Cell type enriched gene sets were extracted from the “The Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes” (Cahoy et al., 2008). The database contains gene expression profiles of cell-type purified mature mouse forebrain samples. Mouse gene symbols were extracted from supplementary tables S4-S6 of Cahoy et al. (2008) . These

tables provide lists of astrocyte, neuron, and oligodendrocyte enriched genes. After removing genes that are not in the ABA coronal gene set, 716 astrocyte, 831 neuron and 571 oligodendrocyte enriched genes remain.

5.2.6 Gene Ontology enrichment

We used the ErmineJ software to extract overrepresented Gene Ontology (GO) groups (Ashburner et al., 2000;Gillis et al., 2010). The set of 3976 coronal genes formed the background gene list for the over-representation analysis. GO groups were limited to the biological process division and required 5-300 annotated genes.

5.2.7 Ortholog assignment

For each gene we extracted its homologous sequences from the HomoloGene database (build version 64) (Wheeler et al., 2007). HomoloGene groups were used to convert the mouse gene identifiers to genes from *S. cerevisiae* (yeast), *C. elegans* (worm), and *D. melanogaster* (fly).

5.3 Results

To identify genes showing strong negatively correlated expression patterns with other genes, we ranked all pairs of genes in the data set by their Spearman correlations across 150 ABA brain regions, and considered pairs with the strongest negative correlations. By filtering gene-gene correlations at a maximum Spearman's rank correlation coefficient (ρ) of -0.72 we selected the 456 most anti-correlated gene pairs. We choose this stringent but arbitrary threshold because we wanted a small list that could be manually examined for interesting

relationships, though our findings proved to hold for other reasonable selection thresholds.

Our first observation was that this list of 456 pairs includes only 102 different genes, indicating there would be strong positive correlations present within this set, rather than numerous distinct patterns. Hierarchical clustering and visualization of the expression patterns of these genes (Figure 7) shows that the original 456 inversely correlated patterns are essentially one inverse relationship corresponding to two gene expression profiles. Visualization of all gene-gene correlations within the set demonstrates this relationship with a clear bimodal distribution with peaks at -0.6 and 0.7 (Figure 8). To further examine the inverse relationship we use clustering to divide the data into two sets: pattern NE (43 image series, 40 genes, Table 14) and pattern OE (68 image series, 62 genes, Table 15). This choice of names will be clarified later in our results. Figure 9 shows expression energy images in the sagittal plane for a pattern NE (CamK2a) and OE gene (S100b). The average profiles of these patterns are strongly negatively correlated (Spearman's rank correlation (ρ) = -0.88). Given the strength of this pattern, although it only includes a small fraction of the genes studied, we asked if it might correspond to patterns uncovered by principal component analysis (PCA). We found the pattern NE and OE genes are strongly separable in PC2 (Figure 10) and the mean loadings in PC1 differ significantly (p-value < 0.001). Thus these patterns correspond to major trends in the data.

Inspection of the gene names and symbols suggested that pattern NE was enriched for neuron-associated genes such as calcium/calmodulin-dependent protein kinase II alpha (Camk2a) (Ouimet et al., 1984) and calbindin-28K (Calb1) (Pfeiffer et al., 1989). In contrast,

several glial cell markers appear in the pattern OE list: carbonic anhydrase II (Car2) (Ghandour et al., 1979;Ghandour et al., 1980), S100b (Ghandour et al., 1981;Rosengren et al., 1986) and glutamine synthetase (Glul) (Wu et al., 2005). Also, one neuron marker, neurofilament high molecular weight (Nefh) appears in the pattern OE list (Letournel et al., 2006). We note that none of the ABA regions are white matter tracts (most are small nuclei), so the pattern does not reflect a simple contrast between grey and white matter.

Gene Ontology (GO) enrichment analysis allowed us to objectively quantify these trends. The GO provides extensive annotations of genes that allow testing for enrichment of specific functions, subcellular localizations or processes. By looking for annotations overrepresented in patterns NE or OE we find several interesting groups, though none reach significance after multiple test correction. For pattern NE the top ranked groups include “regulation of transport” (GO:0051049, p-value=8.3×10^{-5}) and “regulation of neurotransmitter secretion” (GO:0046928, p-value=0.0035). Pattern OE is enriched for groups such as “potassium ion transport” (GO:0006813, p-value=0.0047), “cellular ion homeostasis” (GO:0006873, p-value=0.013) and “regulation of membrane potential” (GO:0042391, p-value=0.0015).

By linking homologous sequences we quantified how evolutionary recent the pattern NE and OE genes are. Surprisingly, only three of the pattern NE genes had a homolog in yeast, worm or fly genomes (7.5%, p-value=0.00023, hypergeometric test). The pattern OE group had 23 (37%, p=0.067) of earlier origin, slightly more than the faction seen in the entire coronal gene set (32%). Both sets had about the expected number of detected orthologs in the human genome.

We used a third bioinformatics approach to test whether these two patterns might reflect differences in cellular populations, using the Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes (Cahoy et al., 2008). Figure 7 and Figure 11 show that pattern NE is enriched for genes identified by Cahoy et al. as being neuron enriched (p-value=0.0016, hypergeometric test). In contrast, pattern OE has half the number of expected number of neuron-enriched genes (p-value=0.015). For the Cahoy oligodendrocyte genes the opposite pattern appears, with 29 genes in pattern OE (p-value<0.0001). Genes from the Cahoy "astrocyte" gene set were represented approximately equally in both sets at the expected proportions. Similar results were obtained by using the lists of oligodendrocyte and neuron enriched gene sets from the ABA (Lein et al., 2007). These strong cell type signals led us to label the two gene sets as neuron enriched (NE) and oligodendrocyte enriched (OE).

The results presented thus far are limited to information obtained at the gene level. While the two profiles seem to have a relationship to cell type, we wanted to test if they provide information about higher-level brain structure. Our next analysis stage incorporated information on spatial locations within the brain and connectivity.

We first summarized patterns NE and OE as the average of the expression patterns of the gene sets. While pattern OE has slightly lower expression levels on average, the two patterns have similar variance. This expression pattern across regions was found to be significantly correlated with the anterior-posterior axis: regions that have high pattern OE expression tend to be at the posterior end of the brain (Spearman's ρ=0.81), with the opposite true of pattern NE (ρ = -0.76). Regions in the posterior end of the brain had fewer connections (ρ = 0.55).

Accordingly we found that the expression patterns correlated with the number of connections the regions have. For incoming connectivity degree the Spearman correlations are 0.49 and -0.54 for pattern NE and OE respectively (141 brain regions). For the 112 regions that have at least one report of an outgoing connection the correlations are 0.32 and -0.44 for pattern NE and OE respectively. Joining the incoming and outgoing connections provides 142 brain regions with correlations of 0.48 (pattern NE) and -0.59 (pattern OE). This means that higher expression of pattern NE is found in "hub-like" regions with many connections, and high expression of pattern OE is observed in "relay-like" regions with few connections. The relationship is shown in Figure 12 with regions of high connectivity degree with low pattern OE expression and high pattern NE expression. All of the above correlations are significant at $p < 0.001$. It is important to note that the entire coronal gene set has substantial correlations of expression levels to anterior-posterior axis ($\rho = 0.29$), incoming ($\rho = -0.19$), outgoing connection degree ($\rho = -0.25$). This spatial correlation reflects a bias in the coronal set gene selection, which favoured genes expressed in the cortex and hippocampus (Ng et al., 2009). Against this baseline, the anterior-posterior expression gradient of the pattern NE and pattern OE genes is still very high.

Because of the known relationship between spatial location in the brain and patterns of connectivity, we sought to correct for this in our analysis of the NE and OE patterns, using partial correlations. We found that the correlations with incoming connectivity degree are still significant after correction for anterior-posterior location, with correlations of 0.20 (pattern NE) and -0.30 (pattern OE). Similarly, the outgoing degree correlations were still significant, though reduced in magnitude: 0.07 (pattern NE, p-value=0.001) and -0.30

(pattern OE). Correlations to the combined degree across 142 regions are 0.16 (pattern NE) and -0.35 (pattern OE; all of the above correlations are significant at $p < 0.001$ unless otherwise noted). A similar analysis carried out using the full Cahoy "neuron" and "oligodendrocyte" lists show similar trends, albeit much weaker than patterns NE and OE. Expression of the Cahoy astrocyte-enriched genes is not significantly correlated with connectivity degree or anterior-posterior axis ($p > 0.1$).

Table 14 Pattern NE gene symbols and names

Gene Symbol	Name
6720401G13Rik	RIKEN cDNA 6720401G13 gene
Calb1	calbindin-28K
Camk2a	calcium/calmodulin-dependent protein kinase II alpha
Camkv	CaM kinase-like vesicle-associated
Cenpf	centromere protein F
Cox6a2	cytochrome c oxidase, subunit VI a, polypeptide 2
Cpne2	copine II
Cpne7	copine VII
Cyln2	cytoplasmic linker 2
Dusp6	dual specificity phosphatase 6
E2f1	E2F transcription factor 1
Egr3	early growth response 3
Fos	FBJ osteosarcoma oncogene
Gria1	glutamate receptor, ionotropic, AMPA1 (alpha 1)
Gria2	glutamate receptor, ionotropic, AMPA2 (alpha 2)
Grik5	glutamate receptor, ionotropic, kainate 5 (gamma 2)
Heatr5b	HEAT repeat containing 5B
Hpcal4	hippocalcin-like 4
Itm2c	integral membrane protein 2C
Kalrn	kalirin, RhoGEF kinase
Ly6h	lymphocyte antigen 6 complex, locus H
Mef2c	myocyte enhancer factor 2C
Mef2d	myocyte enhancer factor 2D
Nnat	neuronatin
Ntrk2	neurotrophic tyrosine kinase, receptor, type 2
Ogt	O-linked N-acetylglucosamine (GlcNAc) transferase (UDP-N-acetylglucosamine:polypeptide-N-acetylglucosaminyl transferase)

Gene Symbol	Name
Pdgfra	platelet derived growth factor receptor, alpha polypeptide
Pea15	phosphoprotein enriched in astrocytes 15
Pkia	protein kinase inhibitor, alpha
Ppap2b	phosphatidic acid phosphatase type 2B
Prkcc	protein kinase C, gamma
Psg16	pregnancy specific glycoprotein 16
Ptprz1	protein tyrosine phosphatase, receptor type Z, polypeptide 1
Rtn4rl1	reticulon 4 receptor-like 1
Shisa9	shisa homolog 9 (Xenopus laevis)
Sirpa	signal-regulatory protein alpha
Slc27a1	solute carrier family 27 (fatty acid transporter), member 1
Tiam1	T-cell lymphoma invasion and metastasis 1
Tnrc4	trinucleotide repeat containing 4
Unc84a	unc-84 homolog A (C. elegans)

Table 15 Pattern OE gene symbols and names

Gene Symbol	Name
3632451O06Rik	RIKEN cDNA 3632451O06 gene
Acyp2	acylphosphatase 2, muscle type
Adssl1	adenylosuccinate synthetase like 1
Ankrd34b	ankyrin repeat domain 34B
Arhgef10	Rho guanine nucleotide exchange factor (GEF) 10
Armc2	armadillo repeat containing 2
Aspa	aspartoacylase (aminoacylase) 2
B630019K06Rik	RIKEN cDNA B630019K06 gene
Bcat1	branched chain aminotransferase 1, cytosolic
Cables2	Cdk5 and Abl enzyme substrate 2
Car2	carbonic anhydrase 2
Cldn11	claudin 11
Cnp1	cyclic nucleotide phosphodiesterase 1
Cnp1	cyclic nucleotide phosphodiesterase 1
Cryab	crystallin, alpha B
Cyp27a1	cytochrome P450, family 27, subfamily a, polypeptide 1
Daam2	dishevelled associated activator of morphogenesis 2
Ddt	D-dopachrome tautomerase
Dip2a	DIP2 disco-interacting protein 2 homolog A (Drosophila)
Elovl5	ELOVL family member 5, elongation of long chain fatty acids (yeast)
Endod1	endonuclease domain containing 1
Enpp2	ectonucleotide pyrophosphatase/phosphodiesterase 2
Fa2h	fatty acid 2-hydroxylase
Fts	fused toes
Galnt6	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6
Gatm	glycine amidinotransferase (L-arginine:glycine amidinotransferase)
Glra1	glycine receptor, alpha 1 subunit
Glul	glutamate-ammonia ligase (glutamine synthetase)
Gprc5b	G protein-coupled receptor, family C, group 5, member B
Hcn2	hyperpolarization-activated, cyclic nucleotide-gated K+ 2
Kcng4	potassium voltage-gated channel, subfamily G, member 4
Kctd9	potassium channel tetramerisation domain containing 9
Klk6	kallikrein 6
Lgi3	leucine-rich repeat LGI family, member 3
Limk1	LIM-domain containing, protein kinase
Map2k6	mitogen activated protein kinase kinase 6
Mmel1	membrane metallo-endopeptidase-like 1

Gene Symbol	Name
Nefh	neurofilament, heavy polypeptide
Nifun	NifU-like N-terminal domain containing
Nrg1	neuregulin 1
Pacs2	phosphofurin acidic cluster sorting protein 2
Plekhb1	pleckstrin homology domain containing, family B (evectins) member 1
Plp1	proteolipid protein (myelin) 1
Pnkd	paroxysmal nonkinesiogenic dyskinesia
Prune2	prune homolog 2 (Drosophila)
Pvalb	parvalbumin
Qdpr	quininoid dihydropteridine reductase
Rnd2	Rho family GTPase 2
Rnf13	ring finger protein 13
S100a16	S100 calcium binding protein A16
S100b	S100 protein, beta polypeptide, neural
Scn1a	sodium channel, voltage-gated, type I, alpha
Sema7a	sema domain, immunoglobulin domain (Ig), and GPI membrane anchor, (semaphorin) 7A
Serpinb1c	serine (or cysteine) peptidase inhibitor, clade B, member 1c
Sgpp2	sphingosine-1-phosphate phosphotase 2
Slc12a2	solute carrier family 12, member 2
Slc39a14	solute carrier family 39 (zinc transporter), member 14
Slc44a1	solute carrier family 44, member 1
Slc4a2	solute carrier family 4 (anion exchanger), member 2
Slc6a5	solute carrier family 6 (neurotransmitter transporter, glycine), member 5
Syt2	synaptotagmin II
Vamp1	vesicle-associated membrane protein 1
Zfyve9	zinc finger, FYVE domain containing 9

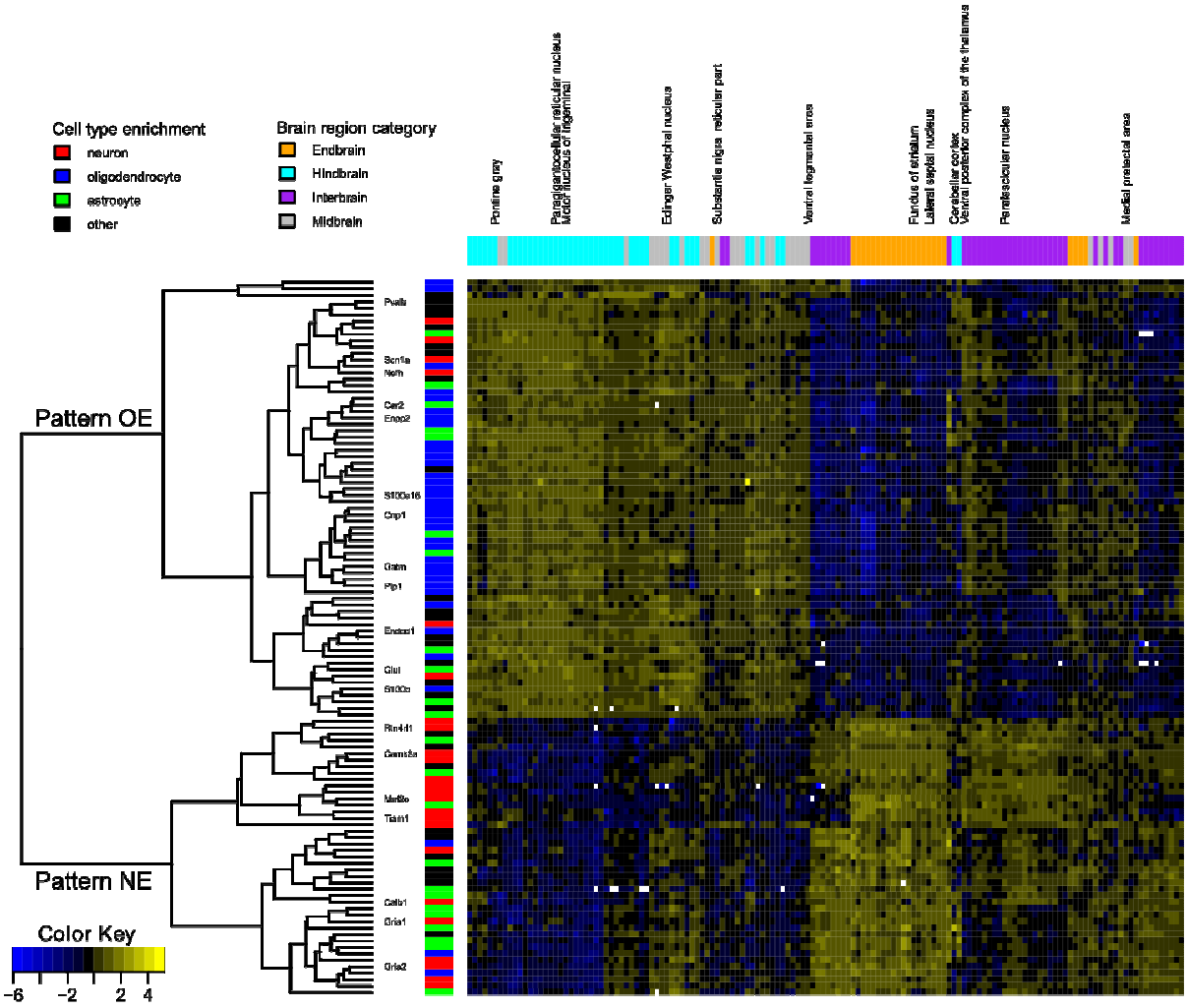

Figure 7 Expression patterns of genes involved in the top 456 negative expression correlations

Normalized expression is colour coded, ranging from blue (low) to yellow (high) and in white for missing values. Genes mentioned in the chapter are labelled. Gene membership in the transcriptome database for astrocytes (green), neurons (red), and oligodendrocytes (blue) is marked (Cahoy et al., 2008). The dendrogram shows the split between pattern NE and pattern OE. Brain regions are coloured as orange for endbrain, cyan for hindbrain, purple for interbrain and grey for midbrain. Expression data for each gene was normalized to mean zero

and variance one for contrast.

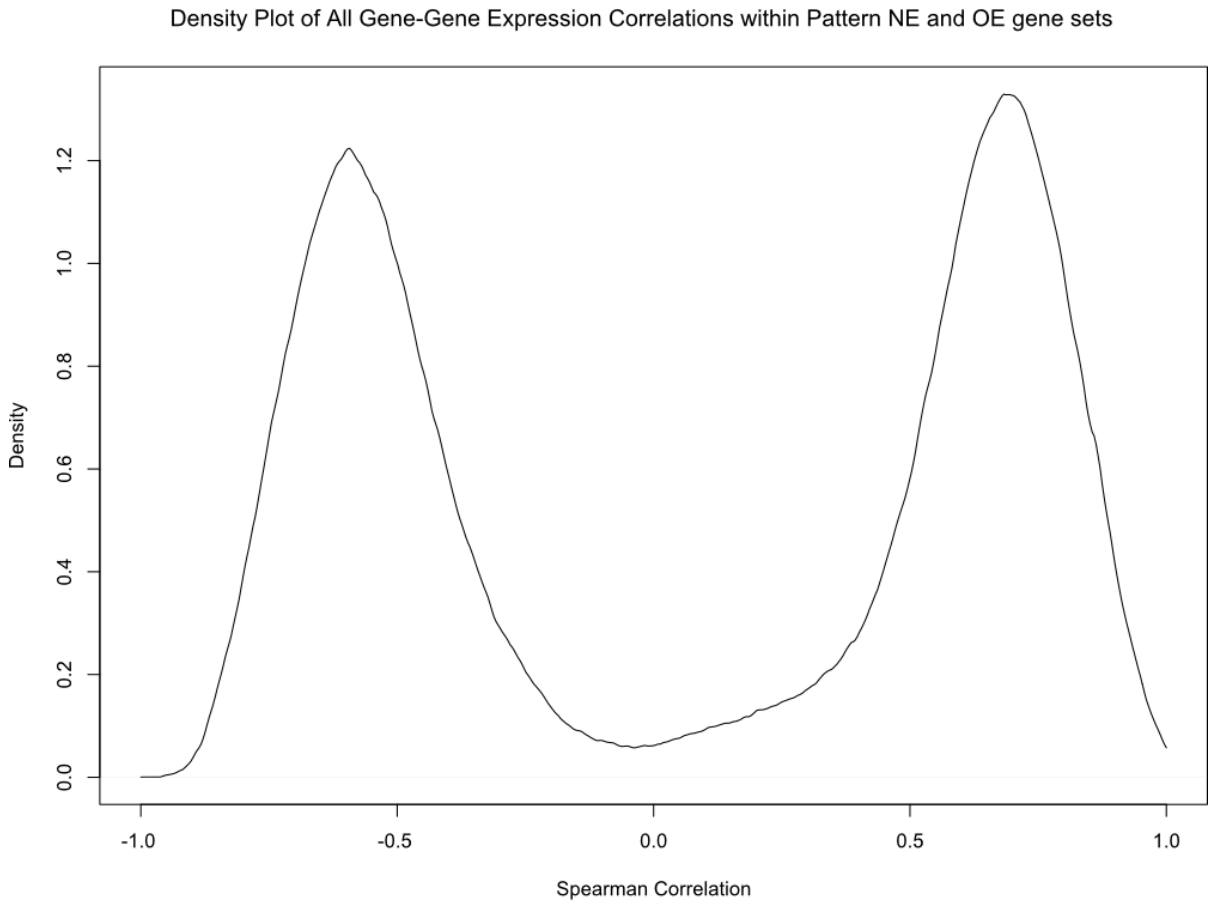

Figure 8 Density plot of expression correlations within pattern NE and OE genes

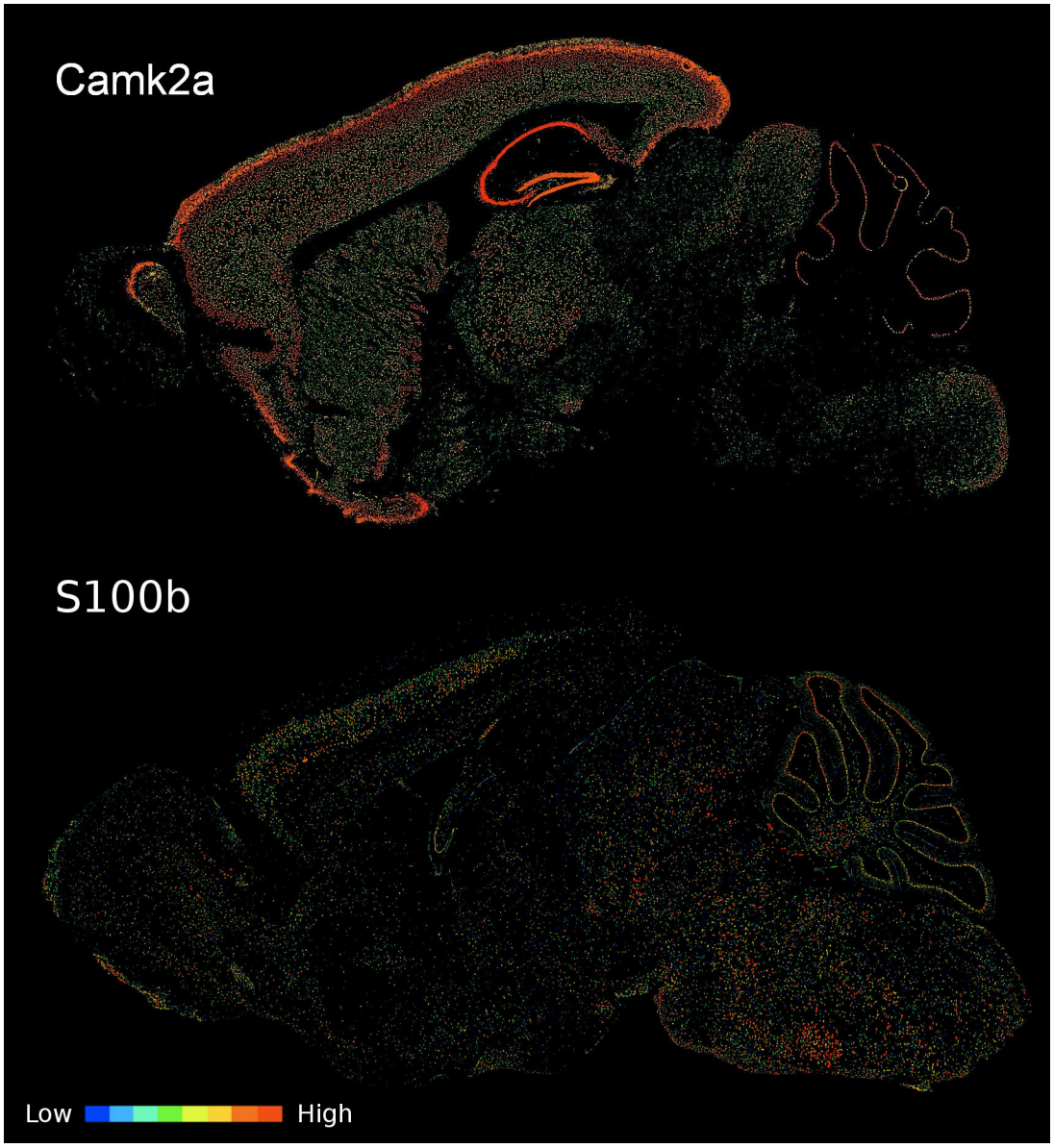

Figure 9 Sagittal expression energy images of a pattern NE and OE gene

CamK2a displays pattern NE (image series 79360274) and S100b shows pattern OE (image series 924). Images were downloaded from the ABA web site (http://www.brain-map.org). While all expression information for the analysis is from coronal assays, we selected a

sagittal view to better show interregional variability in a single section.

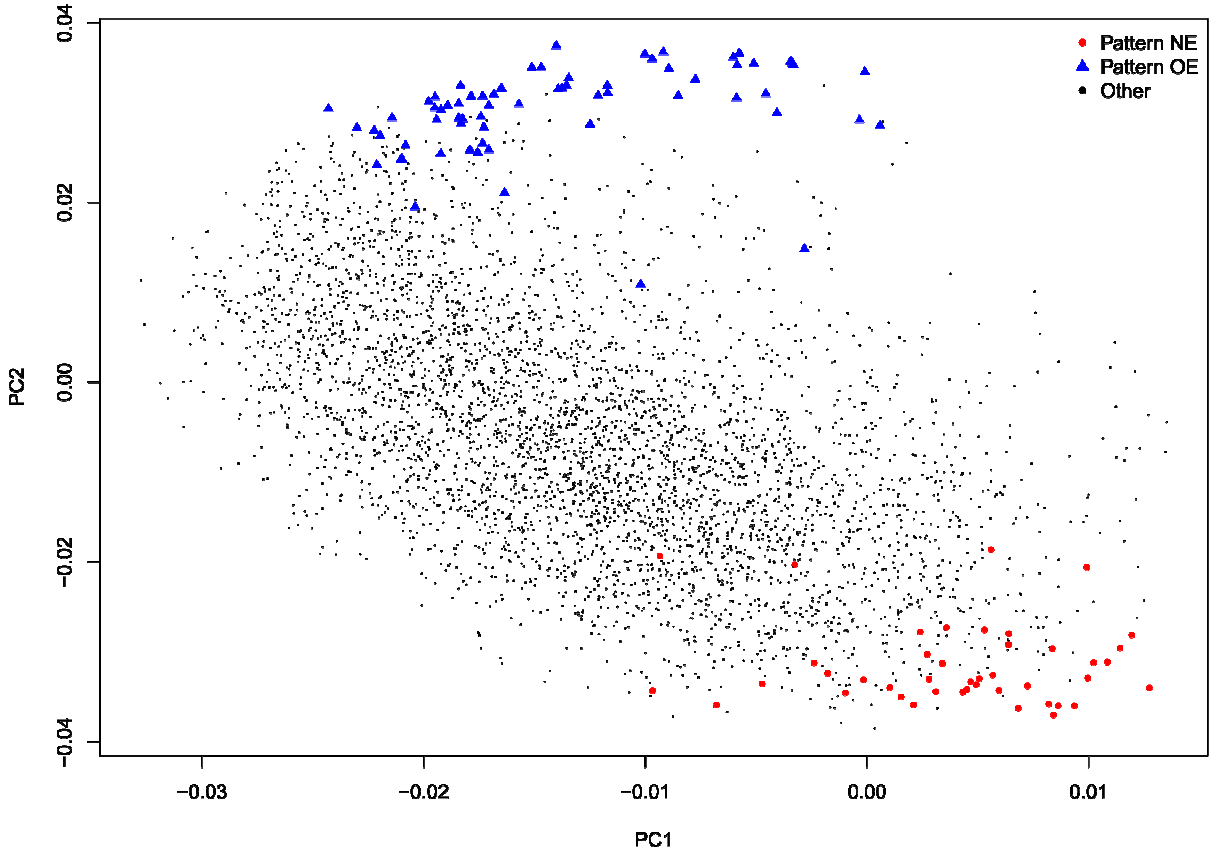

Figure 10 Principal components analysis. Gene loadings for pattern NE (red circles), pattern OE (blue triangles) and all other genes (small black circles) are plotted

The first two principal components, PC1 (16.4 % of the variance) and PC2 (11.8 % of the variance) separate the two patterns.

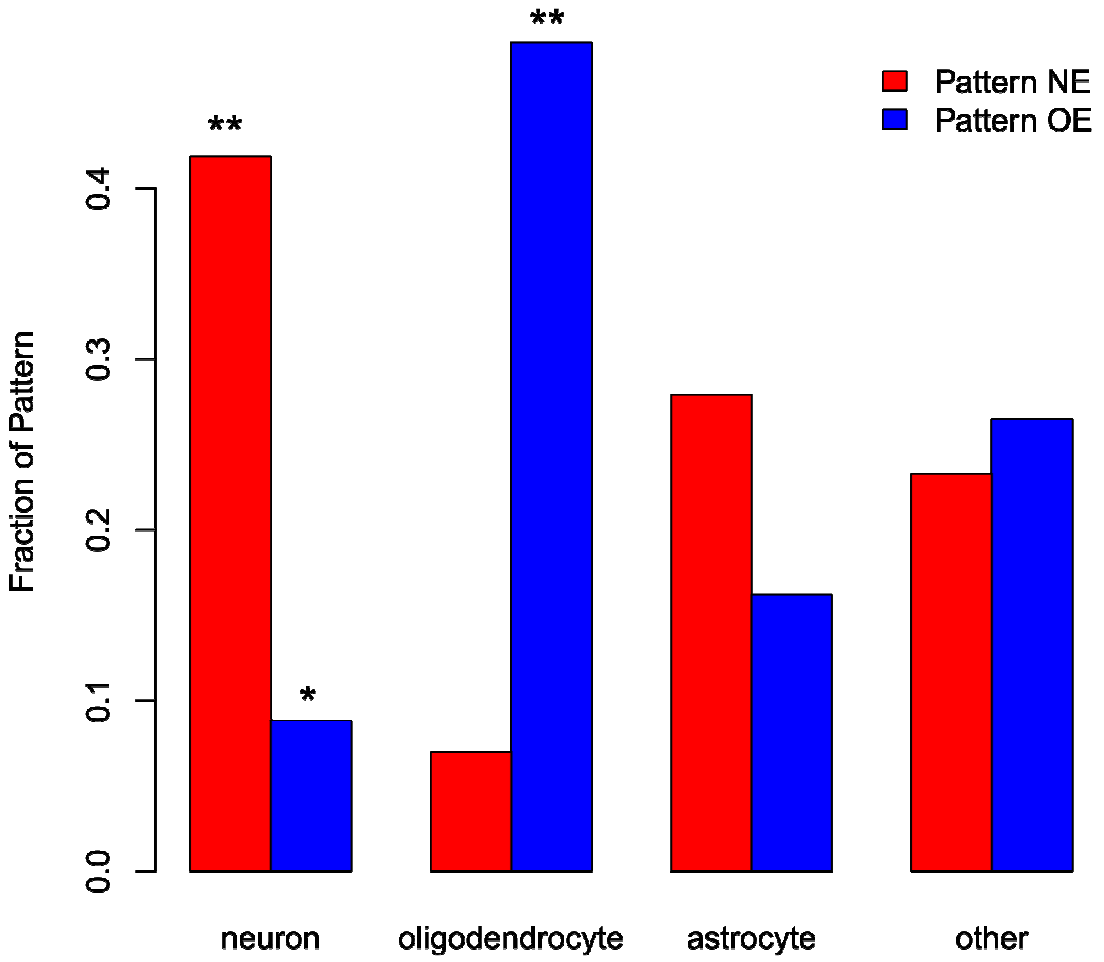

Figure 11 Fraction of cell type enriched genes appearing in the two patterns

P-values below 0.05 are marked by * and below 0.005 with **. Neuron enriched genes are overrepresented in the NE list and underrepresented in the OE list. Oligodendrocyte genes are overrepresented in the OE list but not significantly underrepresented in the NE list.

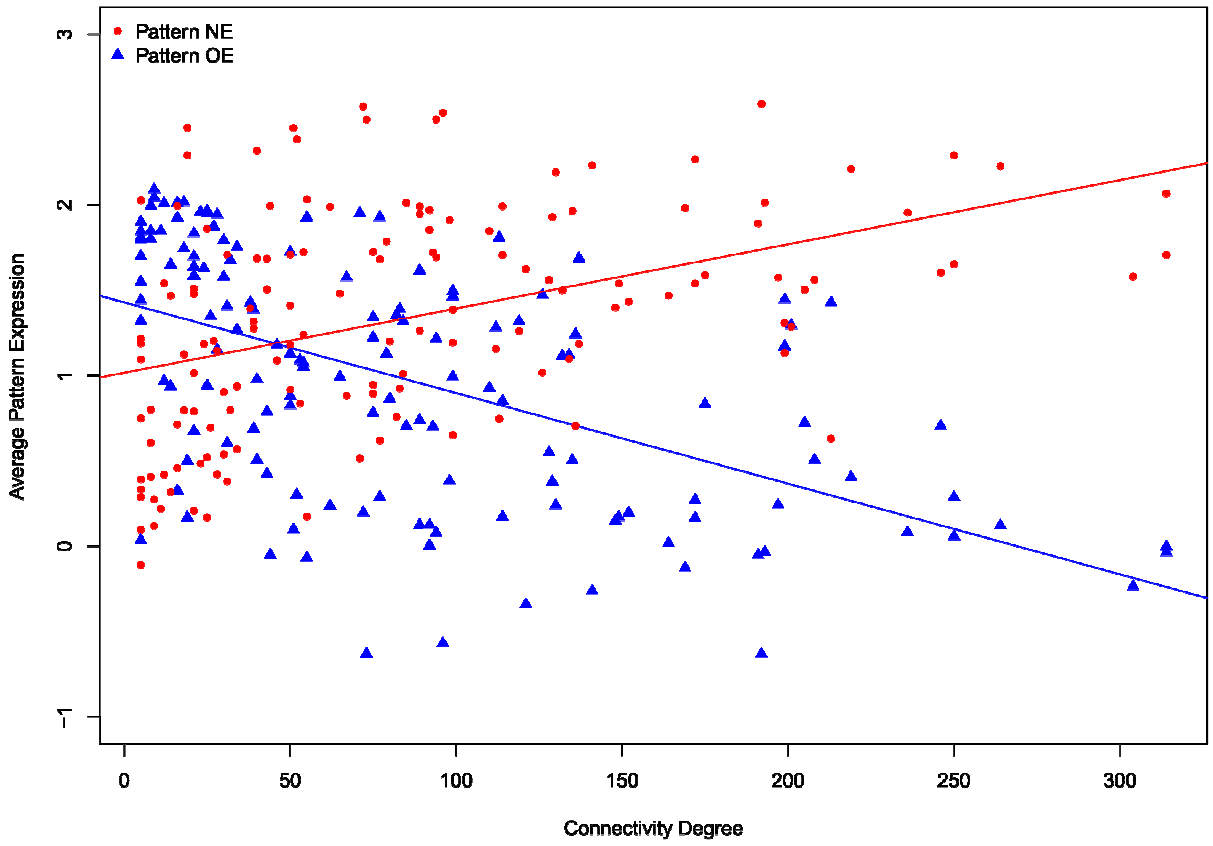

Figure 12 Relationships between degree and expression patterns

Connectivity degree is plotted against average pattern NE (red circles) and OE (blue triangles) expression levels for each brain region. Degree for the 142 regions is the sum of both incoming and outgoing connections.

5.4 Discussion

In this chapter we have shown how a complex expression pattern in the rodent brain can be dissected in terms of genes, cell types, spatial location and connectivity. To our knowledge, the expression patterns we identified have not been previously described. However, previous

work has uncovered possible links between neuroanatomy, gene expression and cell type. Using a voxel-based PCA on a subset of the ABA data, Bohland et al. noted that the two most separable structures, the striatum and cerebellum, contain a relatively large number of GABAergic inhibitory neurons (Bohland et al., 2009b). There are a number of differences between the analysis of Bohland et al. and ours, including the use of voxels vs. brain regions and the choice of genes analyzed, so it is not easy to compare them (indeed it appears the components in the two PCAs are not equivalent), but it is likely that at least some of the highly weighted genes in the pattern identified by Bohland et al. are genes in the pattern we found. A second study has examined a link between expression and connectivity for two specific brain regions (Ng et al., 2009). Using the Anatomic Gene Expression Atlas (AGEA) Ng et al. visualized correlated expression profiles of the parafascicular nucleus and the ventral posterior complex. The ventral posterior complex is a "relay nucleus" and has fewer connections than the hub-like parafascicular nucleus. The AGEA visualization demonstrated that the regions have diverse expression correlation maps that might reflect their diverse function (Ng et al., 2009). In agreement with this result, in our analysis the highly-connected parafascicular nucleus has high expression of the neuron-enriched pattern NE compared to the ventral posterior complex. For the oligodendrocyte enriched pattern OE the opposite is true. Our results are consistent with the idea that degrees of connectivity might be reflected in expression pattern.

Patterns NE and OE are suggestive of differences in the relative proportion of neuronal and glial cell populations in the brain regions in which they are expressed. We further hypothesize that the correlations these patterns have with connectivity might be explained in

terms of highly-connected regions having more neurons, and concomitantly fewer oligodendrocytes. However, we could not rigorously test these ideas here because measurements of glia-to-neuron ratios across many brain structures do not appear to be readily available. More detail about the nature of connectivity supported by the pattern NE and OE regions could also provide insight; in particular the connectivity data we used does not detail if the connections are highly myelinated, inhibitory or excitatory. We also found that the pattern NE genes have a more recent evolutionary origin, while the pattern OE genes tend to be more ancient. This agrees with past work that found evolutionary expansion and regional variation of synaptic genes that are expressed primarily in neurons (Pocklington et al., 2006;Emes et al., 2008).

We note that the connectivity data we employ does not form a complete connectome. The connectivity data we use lacks information about connections that have been shown not to exist. In addition, many brain region pairs have not been studied in a curated tract tracing experiment and may or may not be connected. Of these three cases only one (connected but not known) would increase connectivity degree of a region. Large increases in connectivity degree will affect our results but small changes in connectivity degree are unlikely to change the correlations because we measure Spearman's rank correlation coefficient. However, we expect additional connectivity data for regions with few reported connections will allow deeper analysis. Further, use of the BAMS connectivity data requires pooling of the underlying voxel based gene expression data into brain regions. This limits our results to less than half of the brain by volume but prevents large regions from dominating the analysis. A larger analysis at the voxel level may result in more robust inverse correlations. However,

associations to connection degree could not be performed because voxel level connectivity data is limited for mouse (Moldrich et al., 2010).

Our analysis required the integration of several complex data sets, illustrating several methodological problems that hinder such efforts. Mapping between anatomical atlases presents a significant challenge in linking transcriptomics to connectomics. While genomics has mostly sorted out how to reference specific genes (Gerstein et al., 2007), it is much harder to identify and delineate a specific brain region (Bohland et al., 2009a;Hawrylycz et al., 2011). In *C. elegans* the stable number of neurons allows each one to be given a unique identifier, but in more complex organisms even within a specific atlas it can be hard to map brain regions across atlases. For example, in the BAMS database we found differences between the 1998 atlas and 2004 rat brain atlases (Swanson, 1999;, 2004). Although mappings between the two atlases are formalized and accessible, only 60% of the regions have mappings (Swanson and Bota, 2010). CoCoMac, a tract tracing database of Macaque connectivity has many conflicting atlases and like BAMS it provides information on equal, overlapping, and enclosing brain regions (Stephan et al., 2000;Kotter, 2004;Kotter and Wanke, 2005). Using CoCoMac, Modha and Singh were able to merge the 379 parcellation schemes and over 16,000 mapping relations to create the largest wiring diagram for the Macaque brain (Modha and Singh, 2010). These formalized brain maps will play an important role in future multimodal analyses of the nervous system. Overall, limitations in our ability to interpret these results stress the need for highly detailed neuroinformatics databases of many modalities (Akil et al., 2011).

In conclusion, we identified a novel expression pattern in the rodent brain that correlates with patterns of connectivity and measures of cellular composition. Future work will be aimed at further dissecting these and other patterns, including the potential relationships they may have with behavioural mutations in mice or neuropsychiatric disorders in humans.

Chapter 6: Relationships between gene expression and brain wiring in the adult rodent brain[5]

6.1 Introduction

In the last chapter we studied relationships between gene expression, cell types and number of connections. This chapter extends the analysis of gene expression to examine which connections the regions make. These "macroconnections" between neuroanatomically-defined brain regions are thought to number between 25,000-100,000 in the mammalian brain (Bota et al., 2003), forming a complex network. Knowledge of the "connectome" is used to diagnose neurological disorders such as ischemic stroke, to interpret brain imaging results and to computationally model the brain. There is also growing evidence of connectivity abnormalities in disorders such as autism and schizophrenia (Lawrie et al., 2002;Geschwind and Levitt, 2007;Just et al., 2007).

As reviewed in Chapter 5, the most comprehensive studies of connectivity have been done in the worm *Caenorhabditis elegans* (at the level of single neurons) and the macaque monkey

[5] A version of this chapter has been published. French L, Pavlidis P (2011) Relationships between Gene Expression and Brain Wiring in the Adult Rodent Brain. PLoS Computational Biology 7(1): e1001049. doi:10.1371/journal.pcbi.1001049

(White et al., 1986;Kotter, 2004). Recent work has begun plumbing the properties of these networks, examining node degree distribution (Hugues and Olivier, 2007), network motifs (Sporns and Kotter, 2004), and modularity (Hilgetag and Kaiser, 2004). It has been shown that anatomical neighbours tend to be connected (Scannell et al., 1995), and there is evidence that wiring cost partially explains network structure (Costa Lda et al., 2007;Perez-Escudero and de Polavieja, 2007). There is also increasing interest in the integration of neuronal connectivity and information about genes. This is in part driven by the fact that many genes show spatially-restricted or varying expression in the nervous system, but in many cases the reasons for the expression patterns are not clear (Su et al., 2002;Zapala et al., 2005;Lein et al., 2007;Bohland et al., 2009b). Please refer to Section 5.1 for review of literature describing relationships between gene expression and connectivity.

In this chapter we examine gene expression patterns and macroconnectivity in the adult rodent brain, using data from the Allen Brain Atlas (Lein et al., 2007) and the Brain Architecture Management System (Bota et al., 2005;Bota and Swanson, 2008). Unlike Chapter 5 that studied the number of connections, we analyze gene expression patterns in the context of specific connections. Our results suggest that in the mammalian brain, as in *Caenorhabditis elegans*, there is a correlation between gene expression and connectivity, and the relevant genes are enriched for involvement in neuronal development and axon guidance.

6.2 Materials and Methods

Data and methods were based on those used in Chapter 5. Specifically, the connectivity data is exactly the same while the gene expression gene set has expanded from 3976 regionally enriched genes to a more complete set of 17,530.

6.2.1 Neuroanatomical connectivity data

For neuroanatomical connectivity knowledge, we used the Brain Architecture Management system (BAMS). BAMS contains extensive information about neural circuitry curated from neuroanatomical atlases and tract tracing experiments (Bota et al., 2005;Bota and Swanson, 2010). The version of the BAMS database we use contains 7,308 structural connections between 961 rat brain regions and is accessible via bulk download (http://brancusi.usc.edu/bkms/xml/swanson-98.xml). Instead of parsing the original XML we used a converted semantic web version created by John Barkley (http://sw.neurocommons.org/2007/kb-sources/bams-from-swanson-98-4-23-07.owl). The BAMS system stores information on projection strength, number of reports, report citations and absence of connections but it is not available in the database version we obtained. However, directions of the neuroanatomical connections are known, allowing splitting of our analysis between incoming and outgoing connection profiles.

The BAMS curators comprehensively studied the bed nuclei of the stria terminalis (BNST)

and indicate that its connection matrix is considered complete (Bota and Swanson, 2010). We were concerned that this unusually well-studied region would bias our results, as it has more known connections than the other regions (we considered regions that lack a documented connection to be unconnected). For example, it has over seven times the average number of outgoing connections. To reduce this bias in the dataset, we removed connection information for the BNST and its subparts. We do not suspect the quality of these connections but wished to prevent one well-characterized region from being overrepresented. We believe the complete connectivity matrix of the BNST will be valuable for future focused analysis.

6.2.2 Gene expression data

We considered using gene expression profiles from SAGE and microarray experiments, but spatial resolution was too low. Therefore we used high-resolution colourmetric *in situ* hybridization (ISH) measurements produced by the ABA (Lein et al., 2007). The complete expression matrix from the ABA (kindly provided by the Allen Institute for Brain Research) consists of 5,380,137 entries formed by 25,991 ISH image series and 207 brain regions. In many cases a gene was assayed more than once, using a different probe or plane of sectioning. The ABA provides values for expression "energy", "level" and "density" across a region. Because level and density had a large fraction of data missing (~40%) we choose to use expression energy (3% missing). Expression energy is defined as the sum of expressing pixel intensities normalized by the number of pixels in a region. The natural logarithm of expression energy values formed our gene expression matrix. Genes that do not have detectable expression in the ABA were removed. The list of non-expressing genes list was

provided in Lein et al. as supplementary data (Lein et al., 2007). After removing the non-expressing genes the final gene expression profiles contain 22,771 image series representing 17,530 genes.

6.2.3 Neuroanatomical matching and selecting

The names of brain regions are formalized in hierarchies both in BAMS (Swanson, 1999;Bota and Swanson, 2008) and the ABA data (Dong, 2007), but the schemes are not identical. In addition, the BAMS dataset contains information at a finer neuroanatomical resolution than ABA. To maximize the use of connectivity information, we created connection profiles of coarser scale by using an up-propagation procedure. Up-propagation maps the brain region to its parent region until the desired level in the neuroanatomical hierarchy is reached. This procedure was applied to all connection pairs in BAMS. For example, a connection between region A and region B will be expanded to the set of all possible connections between the neuroanatomical parents of both region A and region B. To prevent enrichment of up-propagated connections we kept regions that had zero connections to the ABA mapped regions.

Although the two datasets are at the brain region level, the organisms differ. The rat brain with a wealth of neuroanatomical information is bigger and for some regions like the cerebellum, more complex. In contrast, genetics and molecular research is more commonly performed on the smaller mouse brain. For this work we considered neuroanatomical differences between the mouse and rat to be minor at the level of granularity we used (Swanson, 2003); for example, the Paxinos mouse atlas was guided by several rat brain

atlases (Paxinos and Franklin, 2008), and brain regions names largely coincide between the two. These common names allowed quick lexical mapping for most of the regions. To join the two data types we mapped nomenclatures manually. We used primarily a region's name, then secondarily its parent region and spatial borders to pair brain regions. The mappings for the Allen Brain regions are provided in Appendix B.

The neuroanatomical atlases from ABA (Dong, 2007) and BAMS (Swanson, 2004) provide information on which brain regions are neuroanatomical children or parts of others. These relations create correlations in the gene expression profiles and the connectivity data (due to up-propagation). To negate this effect we used only 149 of 207 Allen brain regions for the primary region list. These remaining regions have no neuroanatomical subparts in the ABA dataset.

The Allen Atlas provides a differing grouping of regions than the BAMS hierarchy. The superior colliculus is one example. The ABA divides its regions into motor and sensory areas, while the BAMS atlas groups the regions into optic, gray and white layers. Differences were resolved by creating "virtual regions" in the BAMS atlas space that contained the corresponding subregions of the Allen Atlas. The connectivity profiles of the mapped regions were joined using a logical OR operation to provide the virtual region's BAMS connections. For example the superior colliculus sensory related virtual region has all of the BAMS connections of the zonal, optic and superficial gray layers. In addition to the superior colliculus, virtual regions were created for the pallidum medial region and nucleus ambiguus.

After mapping of brain regions, the ABA data is an x (number of regions in the ABA) by y (number of genes) matrix, and the BAMS connectivity data is a square w (number of regions in BAMS) by w (region) matrix (Figure 13). The two matrices are not directly comparable because the number of regions in BAMS is greater than those in ABA ($w>x$). Rather than discarding all information from regions which lack expression information, we use the x by w submatrix of the BAMS data. Thus each of the x regions has a y-dimensional expression vector and a w-dimensional connectivity vector. This maximizes the use of connection information, but we note that the connectivity profiles include information from regions for which we lack expression information.

6.2.4 Statistical tests

Correlations between gene expression values and connection degree were computed using Spearman's rank correlation coefficient (ρ). Connection degree for each brain region is the sum of its propagated incoming and outgoing connections. Significance of the correlation was corrected for multiple testing using the Bonferroni method.

Mantel test: To test the hypothesis that there is a statistical relationship between connectivity and gene expression profiles, we apply the Mantel test (Mantel, 1967). The Mantel test is similar to methods previously applied to *Caenorhabditis elegans* data (Kaufman et al., 2006). The Mantel test uses correlation at two levels to measure the relationship between the connectivity and gene expression profiles. First, Pearson correlation for the connectivity and gene expression profiles is computed for each pair of brain regions, resulting in a distance or similarity matrix (Figure 13). The upper triangles of the

similarity or distance matrices are then converted to linear vectors. The Pearson correlation of these two vectors is then computed to provide dependence between the connectivity and gene expression profiles for all brain region pairings. The statistical significance is determined from an empirical null distribution. We performed the same analytic procedures used on the 'real' data 1,000 or more times using shuffled data. To keep the distribution of the gene expression and connectivity values constant, we shuffle the brain region labels. Significance is determined by counting the number of shuffled datasets that score higher than the non-shuffled result. Mantel correlograms were created using the "mantel.correlog" R library developed by Pierre Legendre (http://www.bio.umontreal.ca/legendre/).

Spatial and nomenclature distance matrices: To create the spatial distance profiles we computed Euclidean distance between a given region's centroid and all others, using the Allen Brain Atlas programming interface (API). Further, we created another measure of brain region proximity using the neuroanatomical part-of hierarchy. Similarity between two regions in the nomenclature profile is simply the number of shared neuroanatomical parents. Using these distance matrices we then performed the Mantel test using the spatial, nomenclature and connectivity profiles. Further we applied the partial Mantel test to determine if the correlation between connectivity and expression is still significant after controlling for these proximity measures (Smouse et al., 1986;Legendre and Fortin, 1989). Akin to performing a partial correlation, the partial Mantel test uses the residuals of a regression fitted to the distance matrix.

6.2.5 Gene ranking and enrichment

We generate a ranked list of genes so that a gene's rank is proportional to its contribution to the connectivity correlation score. To achieve this we reduce the number of genes in the expression profiles while maximizing the Mantel test correlation score. Since it is not feasible to compute all possible subsets of the image sets, we approximate an optimal candidate list of genes. Again, we take guidance from Kaufman et al. (Kaufman et al., 2006) and use a greedy backward elimination algorithm with the Mantel test. Each iteration of the algorithm involves ranking each gene by its contribution to the global correlation, removing the least informative gene, and repeating the test on the remainder. For the connectivity gene rankings we optimized a partial Mantel correlation that modelled proximity in the connection matrix but not the expression correlations (due to computational constraints).

For functional enrichment analysis we employed the ErmineJ software to explore the roles of the candidate genes (Ashburner et al., 2000;Gillis et al., 2010). Overrepresentation analysis was used on the set of genes removed after correlation reached a maximum. To increase resolution of the genes, NCBI identifiers were used instead of gene symbols. Gene Ontology (GO) groups included in the analysis required 5 to 200 measured gene members and were limited to the biological process division. Benjamini-Hochberg false discovery rate was used to control for testing multiple GO groups (Benjamini and Hochberg, 1995). GO groups were sorted by corrected p-value to determine rankings.

6.3 Results

We obtained data sets of macroconnectivity in the rat brain and gene expression data on mouse (see Materials and Methods and Figure 13). By carefully mapping brain regions across them, we identified 142 distinct (non-overlapping) brain regions in common (the "common" regions; see Materials and Methods). In total these regions account for nearly half of the volume of the brain. A notable omission is many regions of the neocortex, which is not sub-parcellated in our data set.

The expression data set, which is filtered to remove unexpressed genes (see Materials and Methods) consists of the expression levels of 17,530 genes in the 142 regions. Because many genes were assayed more than once in the Allen Atlas (independent "image series" in their terminology), there are 22,771 rows in the expression data matrix. The connectivity data consists of the connectivity profiles of 942 regions with the 142 common regions (Figure 13). In this binary matrix, a value of 1 at index (*i*,*j*) indicates a connection exists between region *i* and region *j*. In most of our analyses, we considered the directionality of connectivity. Of the 142 common regions, 112 have efferent (outgoing) connections, and 141 have afferent (incoming) connections; there are 5216 outgoing connections and 6110 incoming connections. Our results are based on various direct and indirect comparisons of the connectivity and expression data matrices or their corresponding correlation matrices.

We began our study with some relatively simple analyses designed to explore the relationship between connectivity, gene expression and other parameters such as spatial

distribution and size of brain regions.

We first tested the simple hypothesis that regions which are connected might have more similar expression patterns. This is in effect a more global search for patterns like the ones identified by Dong et al. (Dong et al.) (note that the CA1 subregions studied by Dong et al. were not represented in our data). To do this we compared the distribution of correlations in expression profiles for regions which are connected to the distribution for regions that are not connected (Figure 14). We found that on average, regions that are connected (ignoring directionality; 456 connected pairs among the 142 regions) have more similar expression profiles than the 8,187 non-connected region pairs (0.79±0.06 for connected; 0.76±0.06 for unconnected; p-value $< 2.2\times10^{-16}$, t-test). This is an initial indication that structural connectivity and gene expression are related.

We found that the size of a region is significantly correlated with its connection degree (Spearman's rank correlation, $\rho = 0.22$). We also noted that the more posterior the region, the fewer connections it has ($\rho = 0.55$). Regions containing motor neurons that project long axons to the spinal cord or muscles were found to have significantly fewer connections (they also tend to be in posterior locations; p-value $= 1.32\times10^{-6}$, Wilcoxon–Mann–Whitney test).

While the above analyses suggest some interesting generic patterns relating connectivity to expression and other parameters, they are not able to expose more complex relationships. Like Kauffman et al. (Kaufman et al., 2006) and Varadan et al. (Varadan et al., 2006), we hypothesized that expression patterns carry information about specific neural connectivity patterns involving multiple regions. To test the global correlation between expression and

connectivity profiles we used the Mantel test. Unlike the test used above to examine the relationship between pair-wise connectivity and expression patterns (using the direct connectivity matrix), here we are asking if the similarity of the connectivity profiles of two regions is related to the similarity of the expression profiles of the two regions, regardless of whether those two regions are themselves connected. In this analysis we are comparing the correlation matrices for the expression data set and the connectivity data (Figure 13).

A key finding is that, as in *Caenorhabditis elegans* (at the level of individual neurons), we find that brain regions that have similar connectivity patterns tend to have similar patterns of gene expression. The Mantel correlation ("correlation of correlations") between expression and incoming connectivity patterns (141 regions) is 0.248 (p-value < 0.0001). Using the outgoing connectivity profiles for 112 regions yielded a correlation of 0.226 (p-value < 0.0001). This relationship holds separately for some of the five major neuroanatomical divisions in the Allen reference atlas. For outgoing profiles the Mantel test is significant at p-value < 0.001 for the interbrain (r = 0.42), cerebrum (r = 0.30) and hindbrain (r = 0.21) divisions but not midbrain or cerebellar divisions. For incoming connectivity only the cerebrum (r = 0.29) and interbrain (r = 0.34) divisions have significant Mantel correlations with expression. Again, we note that unlike our observation of similar expression profiles among connected regions, here we are comparing connectivity patterns of regions, which does not require that the regions be connected to each other.

One factor in this analysis is that regions which are near each other tend to be connected (Scannell et al., 1995) and also might be expected to have higher correlations in expression

patterns (because nearby regions will tend to be of the same embryonic origin, for example). This will tend to obscure the degree to which expression is specifically correlated with connectivity (and in turn obscure the degree to which expression is specifically correlated with location). We assessed the overall degree of spatial autocorrelation by performing the Mantel test as above, but comparing expression or connectivity to a matrix representing physical distance or, alternatively, nomenclature distance (relationships in the nested hierarchy of brain regions). As expected, the Mantel test results are all significant (Figure 15). The connection data ($r = 0.32$; p-value < 0.001, Mantel test) appears to be less spatially autocorrelated than expression ($r = 0.49$; p-value < 0.001, Mantel test).

We visualized the spatial correlation structure with Mantel correlograms (Figure 17). The Mantel correlogram displays the correlation between a data matrix and a matrix formed by grouping region pairs into distance classes. The correlogram will not be flat if it is possible to predict the distance class of a pair based on connectivity or expression correlations alone. As shown in Figure 17, there is indeed an effect of distance on the correlation between connectivity and expression. We therefore attempted to correct our analysis for the effect of spatial autocorrelation, using regression. We calculated regressions between the distance and expression or connectivity correlations for all region pairs. The residuals of these regressions provide proximity-controlled correlations. As shown in Figure 17, an improvement in the correction is obtained when using log-transformed distances.

Using the log-transformed distance matrix from above, we can control for spatial autocorrelations by applying the partial Mantel test (Smouse et al., 1986;Legendre and

Fortin, 1989). The partial Mantel test applies the same regression mentioned above to both the connectivity and expression similarity matrices. Then a standard Mantel test is calculated between the two spatially-corrected residual matrices. We found that after correction, the partial Mantel test between connectivity and expression remains significant, indicating the relationship is not entirely due to neighbourhood effects. However as expected the correlations are lower. Using the spatial correction, the correlation between incoming connectivity and expression is 0.109 (p-value = 0.008, Mantel test), for outgoing it is 0.126 (p-value = 0.001, Mantel test). As a further confirmation for the effectiveness of the correction based on spatial distance, we found that the correlation between nomenclature distance and expression or connectivity correlation drops substantially, though the correlations are still significant (Mantel correlation -0.089 for expression, p-value = 0.006; 0.11 for connectivity, p-value < 0.001). This incomplete correction is perhaps not surprising as the nomenclature hierarchy reflects connectivity as well as spatial location.

The above tests use expression information for all expressed genes in the Allen Brain Atlas, but we expect that many genes will not contribute any information on connectivity. To find the most informative genes, we applied a greedy algorithm that identifies subsets of the data which maximize the correlation between connectivity and expression patterns (see Materials and Methods). Figure 19 displays the change in the Mantel correlations as genes are iteratively removed. As shown in Table 16, this yields much smaller sets of genes (357 and 433 for outgoing and incoming, respectively) and much higher Mantel correlations (0.56 and 0.65 for outgoing and incoming connectivity respectively). As a control, we performed the same procedure on multiple shufflings of the expression data, yielding a maximum

correlation across ten runs of r = 0.42 and r = 0.51 for outgoing and incoming respectively. We also carried out the same procedure for the spatial correlations instead of connectivity, yielding a "spatial proximity" list of 401 genes and a Mantel correlation of 0.934. Eighty-five image series (89 genes) were found to overlap between the lists for incoming and outgoing connectivity, which is not surprising because there is a fair amount of reciprocal connectivity. Twenty-one image series (31 genes) overlap across the spatial proximity list and one or both of the connectivity gene sets, suggesting that for the most part, different genes provide information about connectivity and proximity. The top twenty image series for the rankings are provided in Table 17. If we consider just the top 20 genes, the Mantel correlations are 0.516 (incoming), 0.460 (outgoing) and 0.590 (proximity). As an additional control, we found that the correlations obtained for the optimized gene sets are robust to the completeness of the connectivity network (tested by, for example, randomly removing brain regions and recomputing the Mantel correlations). Thus, while the connectivity map of the rodent brain is incomplete, the correlations with expression appear robust.

We next examined the expression patterns of the optimized gene lists in more detail. It was of interest to determine, for example, if all the genes had similar expression patterns, which would suggest a single overwhelming signal in the data. A hierarchical clustering and visualization of the expression patterns of the optimized gene sets suggested that the patterns are in fact diverse. This is supported by a comparison of the distributions of gene-gene correlations within the optimized outgoing list, which are on average slightly lower than the full data set (0.10±0.21 for top outgoing genes; 0.15±0.21 for all genes; p-value < 0.0001, t-test, Figure 16). This suggests that many different gene expression patterns are contributing

to the overall correlation between connectivity and gene expression.

Figure 20 shows the expression patterns for two genes that rank high in the “outgoing” gene list, overlaid on schematics of the connectivity data. In Figure 20A, we show the pattern for Pcp2 (Purkinje cell protein 2; Figure 20A). Although Pcp2’s function is unknown, it is almost exclusively expressed in the projection neurons of the cerebellar cortex (Purkinje cells). We did not expect this specific expression pattern to carry information about connectivity because no other regions express Pcp2. However, the connections of the cerebellar cortex are also unique and specific: of the 112 outgoing regions, 69 place the cerebellar cortex in the bottom tenth percentile of similar regions based on proximity controlled connectivity. As a result, the optimization procedure finds that Pcp2's expression pattern marks the cerebellar cortex's unique connectivity profile. Figure 20B shows the expression pattern of Pgrmc1 (Progesterone membrane component 1), a gene that may play roles in axon guidance (Runko and Kaprielian, 2002;, 2004). In contrast to Pcp2, which is expressed in only one brain region, expression of Pgrmc1 in two regions is correlated with a connection between them (Figure 18). Thus, clusters of highly connected regions tend to show higher levels of Pgrmc1 expression (Figure 20B). While the strong relationships shown in Figure 20 are not representative of the data set as whole, they serve to illustrate how expression patterns can contain information on connectivity.

One concern about using high-throughput *in situ* hybridization data might be the potential for artifacts. While all of the image series we used had passed the Allen Brain Atlas project’s (ABA) own quality control criteria, we did note occasional spatial artifacts such as dust or

bubbles, though there was no indication such problems were more common in the genes we ranked highly. In addition, while there is good evidence that the ABA data are reliable, with a high quantitative and qualitative agreement with other data (Lee et al., 2008;Jones et al., 2009), there are genes (~6% in ABA) for which ABA has disparities (Jones et al., 2009) and a few of those genes show up in our results (at approximately the expected proportion; see Dataset S1). To help address these concerns, we extracted a higher-confidence subset of results by considering genes measured more than once in the Allen Brain Atlas. These "duplicate" image series vary primarily by the RNA probe sequence used and the plane of section (sagittal vs. coronal), and it seems unlikely that results which are concordant across image series would be due to expression analysis artifacts. Seventeen genes in our top outgoing connectivity list have two concordant image series. In the case of incoming connectivity, 16 of the genes on our list are represented by at least two image series (Rprm has three, and Calb2 has four of its 20 image series across the atlas). We refer to these as the "high-confidence" lists.

The next stage of our analysis was to consider in greater detail the types of genes which are correlated with connectivity. We accomplished this through a combination of Gene Ontology (GO) annotation enrichment analysis and manual review of the literature relating to the genes, particularly those on our high-confidence lists. We specifically hypothesized that genes that play roles in neural development might be found, as suggested by previous work on *Caenorhabditis elegans* (Kaufman et al., 2006;Varadan et al., 2006).

In agreement with this hypothesis, our Gene Ontology analysis of the "outgoing" list

revealed significant enrichment in categories related to neuronal development (Table 18; note that many of the top groups have overlapping gene members. No GO terms were significant for the “incoming” or “proximity” lists). A manual examination of the connectivity top gene lists makes it clear that this is due to the presence of many different genes that play a variety of roles in neuronal development, but axon guidance was a prominent theme. Our lists contain a total of 14 members of three major axon guidance families (Semaphorin, Ephrin, and Slit families) (Chilton, 2006) (Table 19). These gene families express cell-surface or secreted proteins that function to provide guidance signals to growing axons. This was most striking for the Semaphorin family, with ligands, receptors and co-receptors appearing in the incoming or outgoing top gene lists (Table 19). Six of the 17 genes from the high-confidence “outgoing” list function in neuronal development and axon guidance. Two of these six, Gpc3 and Hs6st2 encode a heparan sulfate proteoglycan and a heparan sulfate sulfotransferase respectively. Two additional heparan sulfotransferases, Hs3st1 and Hs6st1 appear with one image series on outgoing top gene list. Heparan sulfate proteoglycans are membrane proteins that have been linked to neurogenesis, axon guidance and synaptogenesis (Yamaguchi, 2001). Hs6st2 has been specifically linked to retinal axon targeting in Xenopus (Irie et al., 2002). Another gene on the high-confidence list is the L1 cell adhesion molecule (L1cam), a recognition molecule involved in neuron migration and differentiation (De Angelis et al., 2002). Vesicle-associated membrane-protein (Vamp2) is another gene connected to connectivity through two image series; in addition Vamp1 occurs once in the outgoing list. Recently Vamp2 has been linked to attractive axon guidance but not repulsion in chick growth cones (Tojima et al., 2007). Neurturin is another high-ranking gene with two image sets linked to outgoing and one linked to incoming. Neurturin is well known to promote

neuronal survival and induce neurite outgrowth (Yan et al., 2004). Lastly, Serinc5 is enriched in white matter and Inuzuka et al. (Inuzuka et al., 2005) suggest its major role is to provide serine molecules for myelin sheath formation.

In the case of genes correlated with patterns of incoming connectivity, 4 of the 16 of the genes on our high confidence list have previously suggested roles in brain connectivity. Neurensin-1 shows up with two image series and is known to be involved in neurite extension (Nagata et al., 2006). Recently, Stat5a has been labelled a key effector molecule in the mammalian CNS, affecting axon guidance in the spinal cord and cortex (Markham et al., 2007). Thirdly, Uchl1 is mutated in the GAD mouse strain that presents axon targeting and genesis defects (Miura et al., 1993). Finally, ciliary neurotrophic factor receptor (Cntfr) appears twice on the top ranked list and is known to promote neuron survival and plays important roles in nervous system regeneration and development (Ip et al., 1993;Miotke et al., 2007).

Another trend we notice from the GO results is that groups of genes with negative regulatory roles are much more prominent than the corresponding “positive” groups (e.g., “negative regulation of neurogenesis”) though these groups are not statistically significant after multiple test correction. The high ranking of these terms (which share members) is due to 11 genes: Hdac5, Notch3, Nrp1, Cd24a, Cit, Apc, Nr2e1, Ptk2, Gpc3, and Runx2. The “negative” aspect of the function of these genes varies but all have roles in neuronal development and/or plasticity. For example Nrp1 is a coreceptor for semaphorins and triggers inhibition of axonal growth (Chedotal et al., 1998), while Hdac5 is a histone

deacetylase whose activity is associated with repressed chromatin conformations that are altered after addictive stimuli (Renthal et al., 2007).

We also conducted a search among our high-confidence list for genes whose homologs are implicated in human disorders of the nervous system. We found evidence for such a role for five of the 30 genes. Prominent among the five is L1Cam, defects in which cause several brain disorders including partial agenesis of the corpus callosum (Gu et al., 1996). Two genes in the high confidence lists have been linked to heritable forms of Parkinson's disease (alpha-synuclein (Snca) (Polymeropoulos et al., 1997) and Uchl1 (Ragland et al., 2009)). Finally, two genes have been linked to autistic spectrum disorder (ASD). The human homolog of Cadps2 has been linked to autism and lies in the 7q autism susceptibility locus (AUTS1) (International Molecular Genetic Study of Autism Consortium, 1998;Sadakata et al., 2007). Another, Btg3 is in a genetic locus linked to autistic children characterized by a history of developmental regression (Molloy et al., 2005). By examining our expanded list of genes, we found several more of our connectivity linked genes are in AUTS1 and have been studied in the context of autism: Reln (Persico et al., 2001), Mest (Kwack et al., 2008), Ptprz1 (Bonora et al., 2005), Dpp6 (Marshall et al., 2008) and En2 (Kuemerle et al., 2007). To further explore the potential connection between our results and autism, we downloaded all autism candidate genes from the AutDB database (Basu et al., 2009). Of those genes, 163 were available in our dataset, and 17 appear in at least one of the connectivity linked lists (14 for incoming connectivity and Nrp2, Cadps2, Ntrk1,and Apc appear in both incoming and outgoing lists). The probability of this occurring by chance is 0.00029 (hypergeometric test; considering the incoming list alone the p-value is 5.43×10^{-5}). In contrast, the proximity-

ranked list contains only 5 genes in the AutDB set (p-value = 0.32).

Table 16 Peak correlation and size of optimized Mantel tests

Name	Peak Correlation	Size (image series)
Incoming	0.645	452
Outgoing	0.564	374
Proximity	0.934	420

Table 17 Top twenty genes for proximity and proximity-controlled incoming and outgoing Mantel tests

Incoming			Outgoing			Proximity		
Rank	Symbol	Imageset	Rank	Symbol	Imageset	Rank	Name	Imageset
1	Nrp2	80514091	1	Pgrmc1	797	1	Nup37	68795447
1	D4st1	74657927	1	Slc25a37	68445000	1	Klrg1	69735903
3	Acadvl	227161	3	Pcp2	77413702	1	Dnahc1	73520818
4	Pgrmc1	797	4	Galr1	80514053	1	Pus1	532760
5	8030411F24Rik	74580853	5	1700054O13Rik	69117086	1	Mm.359340	71209910
6	Gda	70276867	6	Plk1s1	70295882	1	Tm2d3	77414123
7	Mdfi	275690	7	Alpk3	71574473	1	LOC433436	73636096
8	3110082D06Rik	74581400	8	Lrrn6c	72128919	8	Gba2	68844337
9	Lyzs	68191492	9	Gm47	70565879	9	Prrg2	276063
10	Atad2b	71496393	10	Cpne5	544709	10	Ccdc137	1979
11	Slc5a2	68632936	11	Nmbr	77332086	11	Col5a3	74272917
12	Dbnl	74819497	12	Trim52	70205626	12	Kcnk2	75147764
13	Dmp1	74511936	13	AI427122	71495698	13	Comt	68301371

Incoming			Outgoing			Proximity		
Rank	Symbol	Imageset	Rank	Symbol	Imageset	Rank	Name	Imageset
14	Gata3	73931427	14	Slc44a4	68321886	14	Bcl2l12	71064289
15	Rgs9	73521819	15	Nrp2	80514091	15	Mtif2	68341663
16	En2	69288944	16	Anxa3	69526665	16	Eomes	80516770
17	Wisp2	68523207	17	A930033C23Rik*	74300717	17	Gcnt1	68546476
18	Cypt3	80474702	18	Tac2	77279001	18	LOC433088	70722898
19	F2rl1	199391	19	C1qtnf9	70228041	19	Mrpl45	70919854
20	1700018L24Rik	74634791	20	Kirrel1	71613657	20	Gda	70276867

Table 18 Top twenty GO groups enriched in the proximity controlled outgoing ranked gene list

Name	ID	Group Size	Hits	P-value	Corrected P-value
neuron projection development	GO:0031175	186	16	0.00000	4.63E-003
cell morphogenesis involved in differentiation	GO:0000904	183	13	0.00013	0.05
cell projection morphogenesis	GO:0048858	157	12	0.00012	0.06

Name	ID	Group Size	Hits	P-value	Corrected P-value
cell part morphogenesis	GO:0032990	166	12	0.00020	0.06
cell migration	GO:0016477	189	13	0.00018	0.06
axonogenesis	GO:0007409	145	12	0.00005	0.07
cell morphogenesis involved in neuron differentiation	GO:0048667	157	12	0.00012	0.07
neuron projection morphogenesis	GO:0048812	154	12	0.00010	0.08
positive regulation of secretion	GO:0051047	27	4	0.00227	0.45
negative regulation of cell communication	GO:0010648	150	9	0.00438	0.45
heparan sulfate proteoglycan biosynthetic process	GO:0015012	5	2	0.00420	0.45
lymphocyte differentiation	GO:0030098	83	7	0.00177	0.47
leukocyte activation	GO:0045321	161	9	0.00691	0.47
B cell differentiation	GO:0030183	33	4	0.00480	0.48

Name	ID	Group Size	Hits	P-value	Corrected P-value
positive regulation of cell-cell adhesion	GO:0022409	5	2	0.00420	0.48
negative regulation of neuron differentiation	GO:0045665	28	4	0.00260	0.48
regulation of neuron differentiation	GO:0045664	82	6	0.00746	0.48
epithelial cell development	GO:0002064	19	3	0.00689	0.48
central nervous system neuron axonogenesis	GO:0021955	13	3	0.00223	0.48
lymphocyte activation	GO:0046649	140	8	0.00936	0.48

Table 19 Members of three canonical axon guidance families appearing in our connectivity and proximity top genes lists

Name	Connectivity	Proximity
Semaphorins and receptors	Sema3a, Sema6a, Nrp1, Nrp2, Plxna2, Plxnb2	Sema3a
Ephrin/Eph	Ephb1, Epha7, Epha8	Efna1, Epha7
Slit/Robo	Slit1	Slitrk4

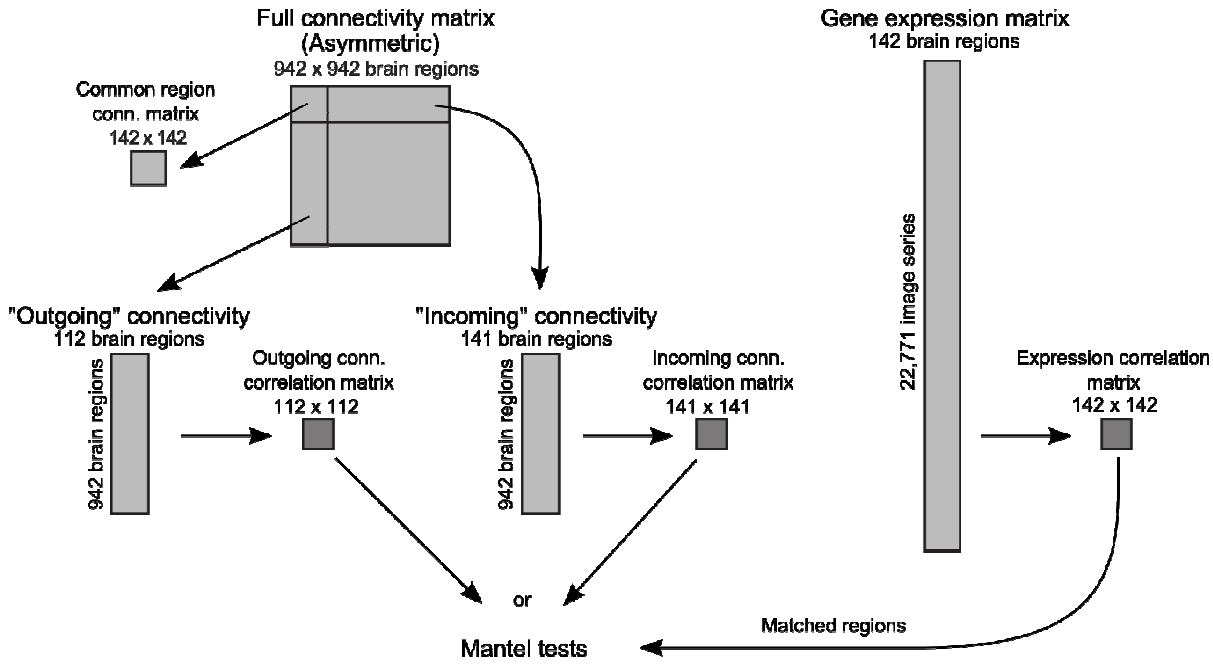

Figure 13 Datasets and correlation matrices used in this chapter

Matrices are shown schematically as shaded boxes; arrows indicate steps in the workflow. For example, from the full connectivity matrix we extracted submatrices of “outgoing” or “incoming” connectivity, and compared their correlation matrices with the correlation matrix of the brain region expression patterns.

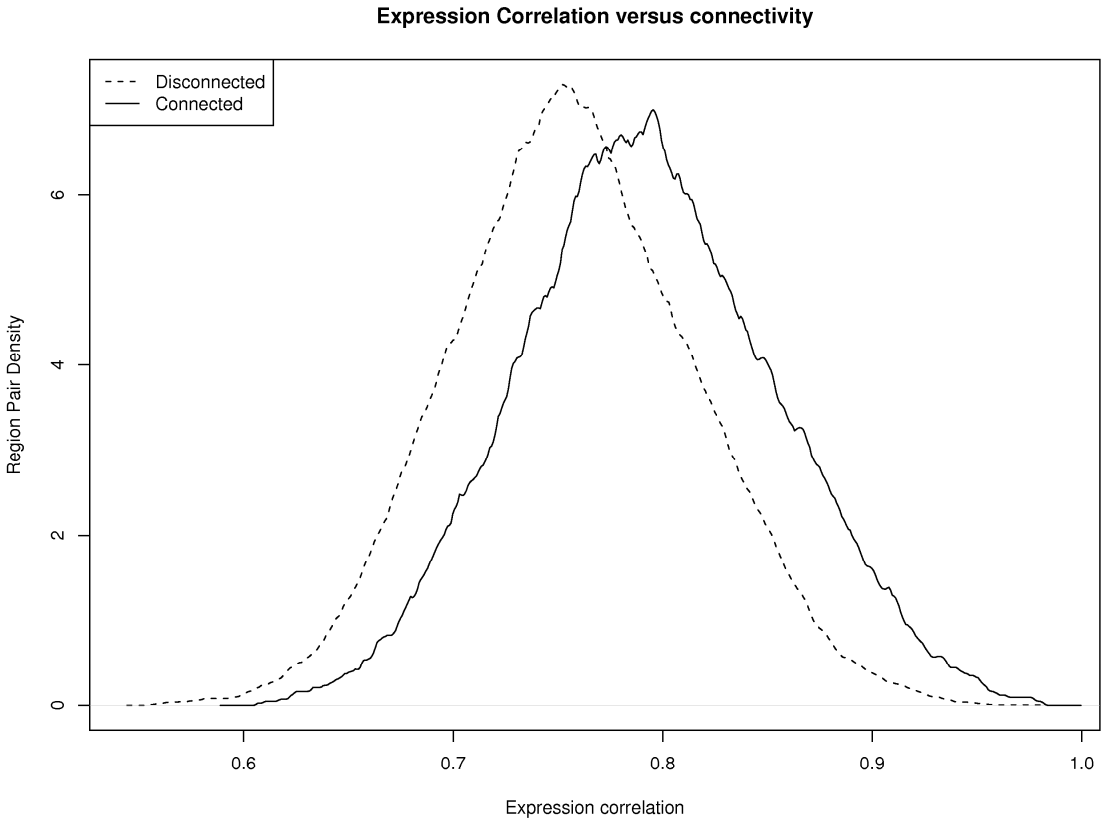

Figure 14 Density plot of expression correlation between region pairs

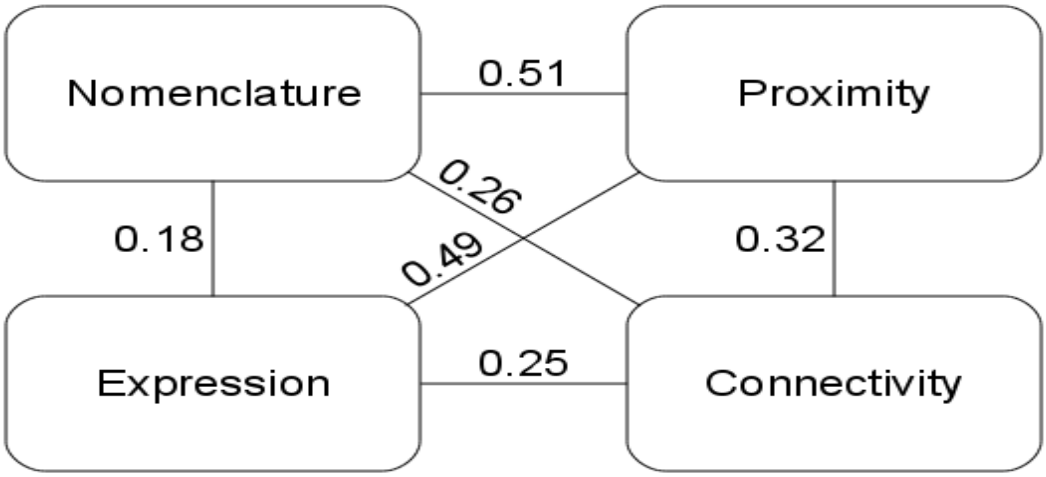

Figure 15 Mantel correlations between different matrices

"Nomenclature" and "Proximity" refer to the two different measures of spatial distance that we used (see Materials and Methods). The 141 regions with incoming connectivity information were used to generate the correlations for this figure.

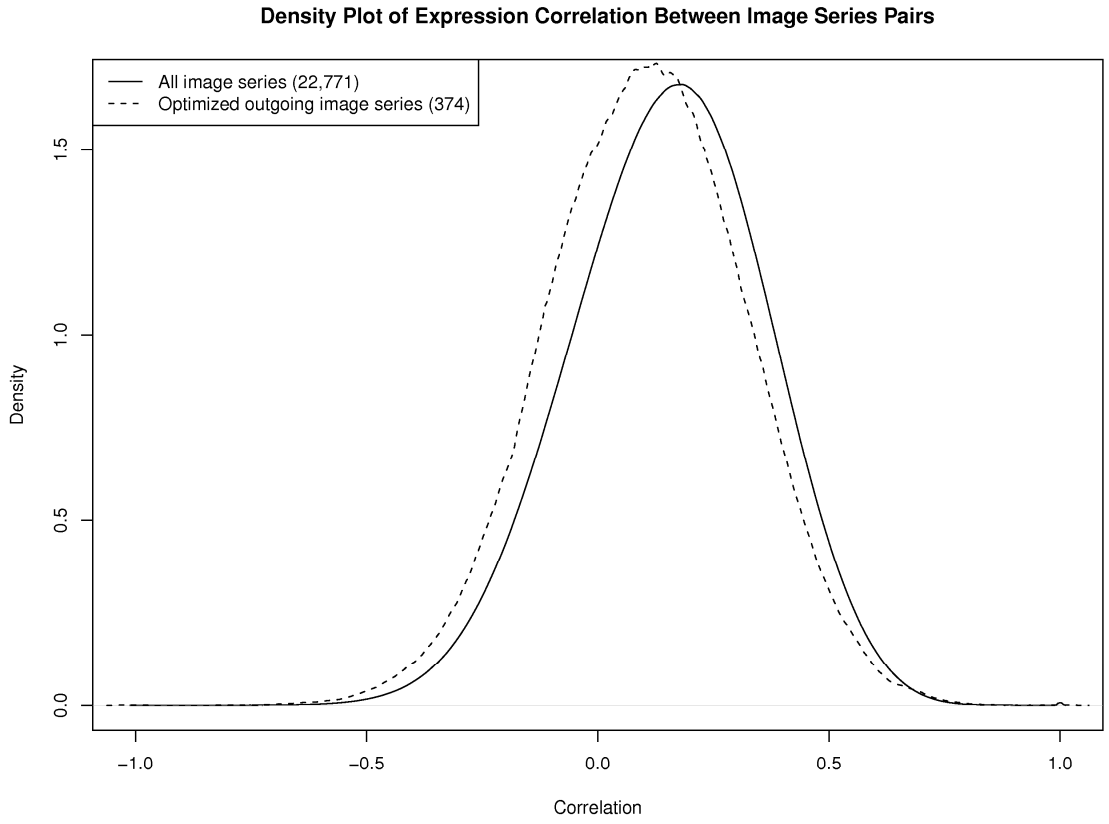

Figure 16 Density plot of gene-to-gene correlations

Gene to gene correlations were computed within the “outgoing” gene list and all genes.

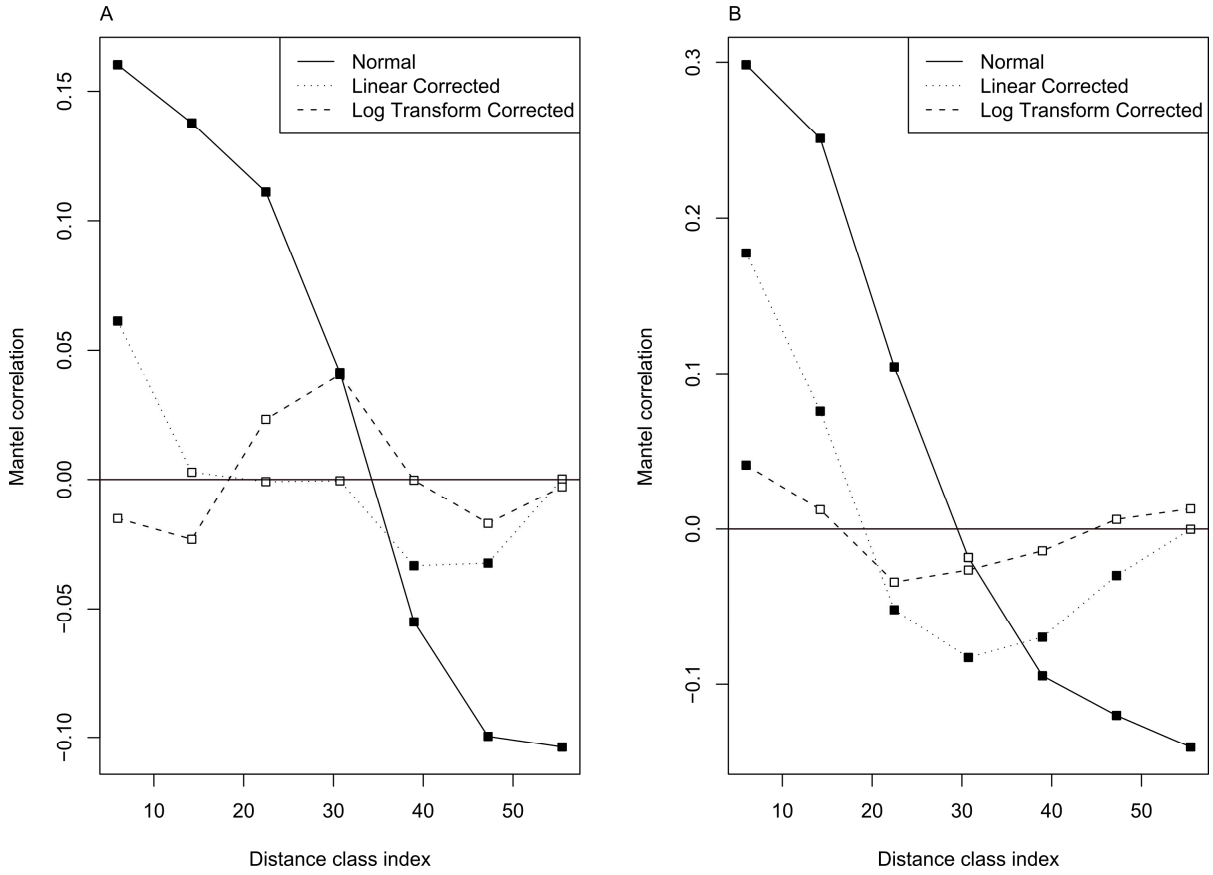

Figure 17 Connectivity (A) and expression (B) Mantel correlograms for uncorrected, linear and log transform corrected spatial distance matrices

Filled squares mark distance classes with significant spatial correlation after multiple test correction.

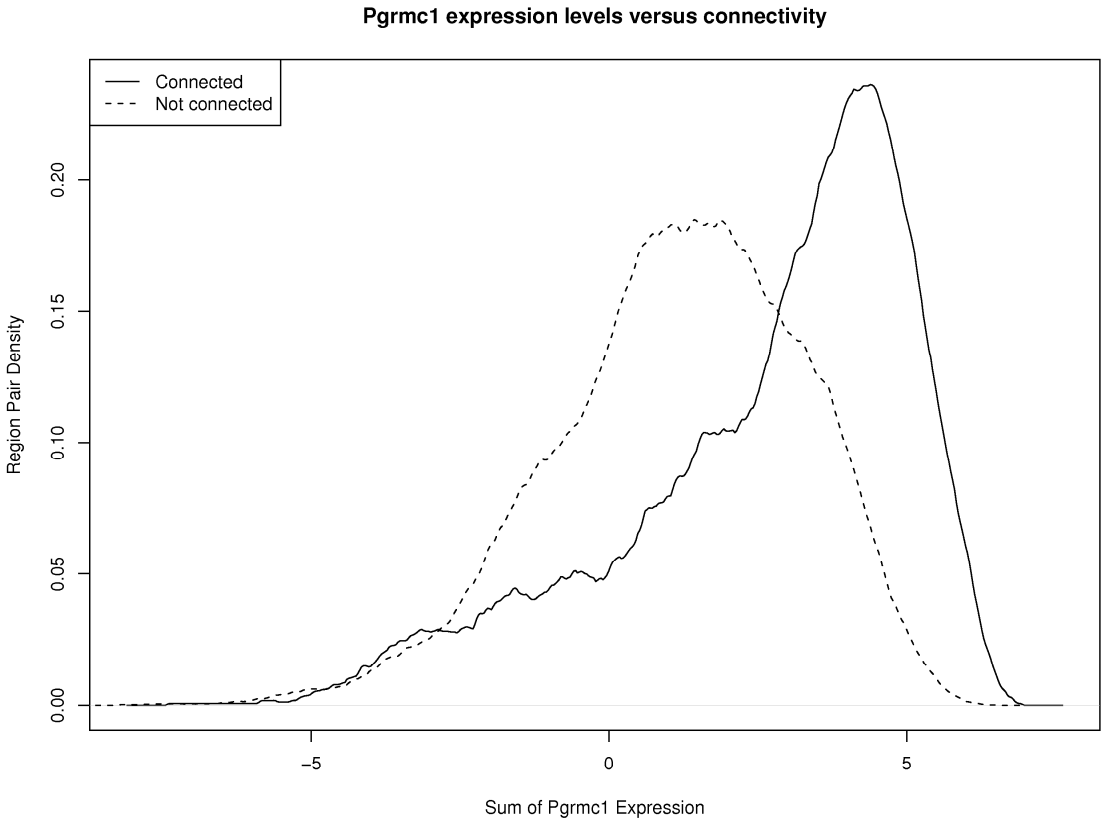

Figure 18 Pgrmc1 expression levels versus connectivity

For each region pair this plot shows the sum of the two regions' expression in the context of their connectivity.

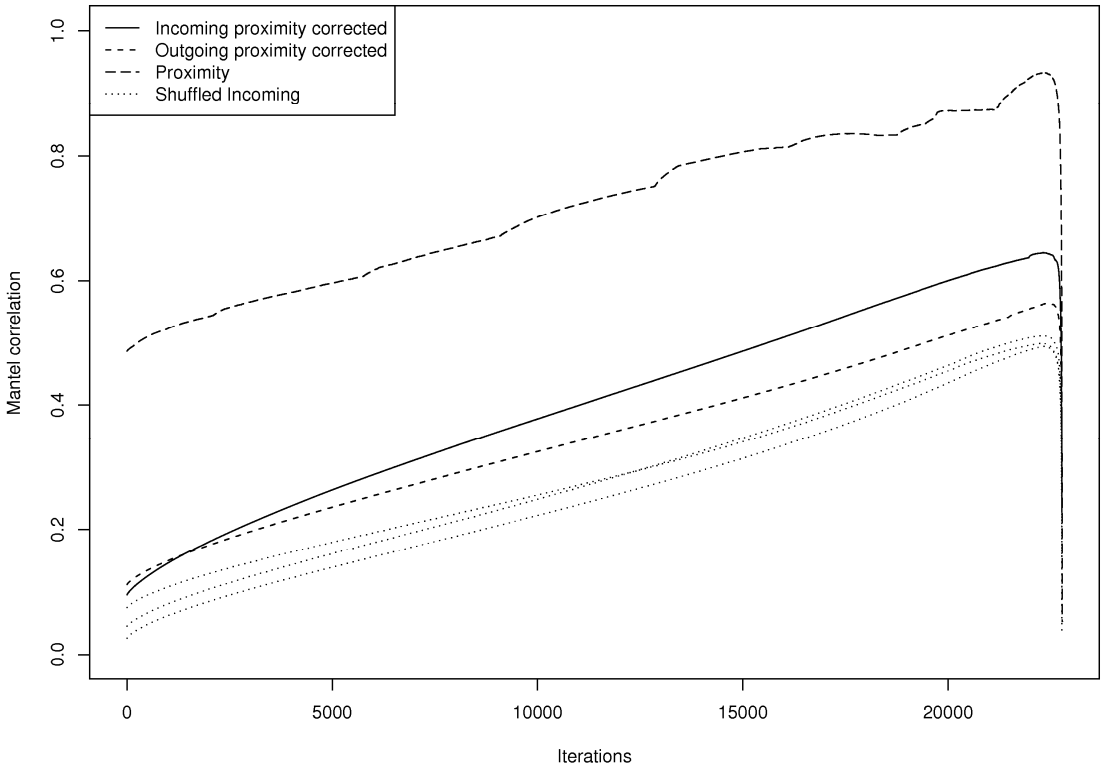

Figure 19 Optimization of Mantel correlation by iteratively removing image series

Each curve documents the correlation across iterations (as genes are greedily removed).

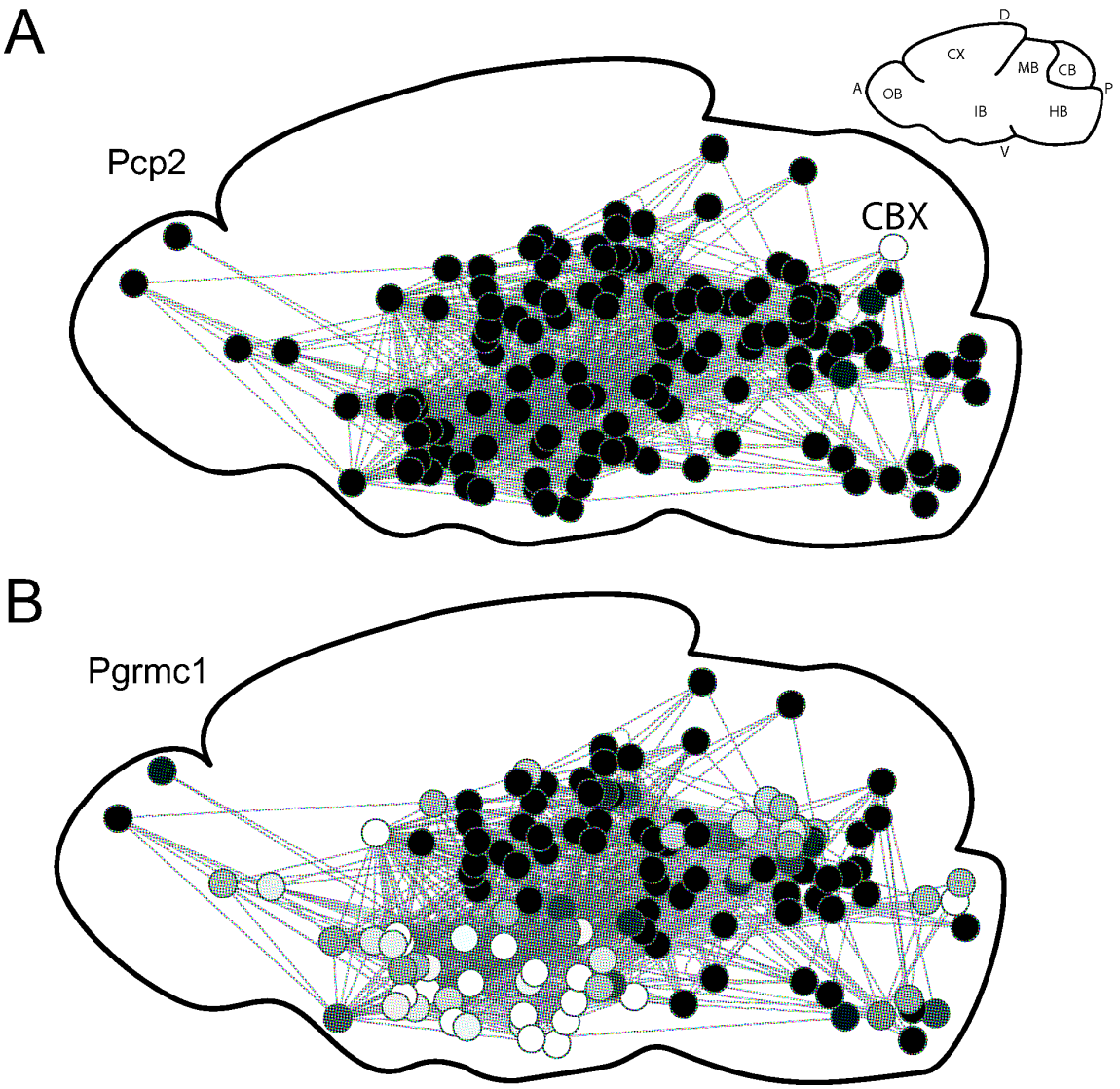

Figure 20 Connectivity in the context of Pcp2 (A) and Pgrmc1 (B) expression

The connectivity map is a 2-D projection of the network on the saggital plane. Each node represents a brain region (placed at the center of the region as measured in the Allen reference atlas). Expression levels are depicted as shades of grey, with lighter shades indicating higher expression. Pcp2 expression is restricted to the cerebellar cortex (CBX), while Pgrmc1 tends to be expressed highly in both regions of connected pairs. The small

inset brain diagram provides orientation (anterior (A), dorsal (D), ventral (V) and posterior (P)) and the locations of the olfactory bulb (OB), cortex (CX), interbrain (IB), midbrain (MB), hindbrain (HB) and cerebellum (CB).

6.4 Discussion

Our analysis revealed a number of interesting relationships between gene expression and patterns of connectivity in the adult mammalian brain. Our key finding is that genes whose expression patterns carry information on connectivity are enriched for genes involved in neural development, and axon guidance in particular. While our results are based on analysis of the brains of rodents, it is of potential importance that many of the genes we identify have human homologs implicated in disorders of the nervous system including ASD. Because there is an increasing interest in the idea that ASD and other disorders are in part due to abnormalities in connectivity (Belmonte et al., 2004;Geschwind and Levitt, 2007), and given the heritability of many such disorders, the relationship between gene expression and connectivity is pertinent. The enrichment of homologs of autism candidate genes in our results suggests that these patterns could be relevant to the understanding of behavior in autism and potentially avenues for treatment.

After our results appeared in literature, a similar study replicating the finding of correlations between gene expression and connectivity appeared (Wolf et al., 2011). Wolf and colleagues showed that machine learning methods could be used to predict connectivity from gene expression patterns in a statistically significant manner, for approximately one half of tested brain regions. Their analysis found that genes known to be associated with schizophrenia,

autism and attention deficit disorder are enriched in their gene sets that predict connectivity. Although the authors did not perform correction for the effect of spatial autocorrelation, they tested the robustness of the connectivity data and the quality of the expression images from the Allen Brain Atlas.

Interestingly, a previous focused examination of the correlation between expression and connectivity for two brain regions identified some of the same genes we did. Dong et al. (Dong et al., 2009) examined correlations between genes that are differentially expressed between the dorsal and ventral hippocampus (which we were not able to treat as separate regions in our analysis). For nine of their genes, they observed matching expression patterns in a connected brain region, the lateral septal nucleus. Three of these nine genes appear on our connectivity correlation lists (Gpc3, Man1a, Wfs1); this is unlikely to occur by chance (p-value = 0.0045, hypergeometric test). In contrast, none of the nine appear on the proximity gene list.

We stress that because what we observe are correlations, it is difficult to ascribe a definite mechanism or meaning to the patterns. In addition, in absolute terms the Mantel test correlations may seem low when we considered all genes. However, we do obtain a correlation of 0.65 between gene expression patterns and proximity-controlled incoming connectivity after gene selection. We also point out that at the neuron to neuron level in *Caenorhabditis elegans*, Kaufman et al. (Kaufman et al., 2006) reported statistically significant correlations of 0.075 and 0.176 between expression and incoming and outgoing connectivity, respectively. Thus the patterns we observe in the adult mammalian brain are at

least as strong as those observed in previous studies. An obvious question is whether the signals we observe are strong enough to predict patterns of connectivity. Unfortunately, while the signals we observe are statistically significant, they are not strong enough to allow prediction of connections based on expression patterns. Kaufman et al. (Kaufman et al., 2006) attempted this with their data and achieved very low accuracy. Using similar data, Baruch et al. (Baruch et al., 2008) attained statistically significant results in predicting the direction of connectivity between neurons known to be connected or which share a common synaptic partner. Using advanced imaging techniques on human subjects, Honey et al. (Honey et al., 2009) attempted to predict diffusion tensor imaging (DTI) based cortical connectivity from fMRI functional connectivity. By setting thresholds on functional connectivity, they achieved an AUC value of 0.79 that could predict only ~6% of inferred DTI connections (Honey et al., 2009). Despite these limitations, our results suggest some underlying models that in turn provide some testable hypotheses.

Many of the genes we find to be associated with connectivity patterns in the adult are thought to be primarily active in the developing brain, when large-scale connectivity is determined. The reasons for expression of these genes in the adult brain is not fully understood, though there is evidence in some cases that they continue to play roles in the maintenance or tuning of neuronal connectivity at finer scales (Zapala et al., 2005;Murray et al., 2007). There is even less known about why the genes show regionally restricted patterns in the adult brain. Our results are the first to link the expression signatures of some of these genes to macroscopic connectivity. Our results have at least two possible biological interpretations. One is that the expression patterns in adulthood are a “residue” of the developmental pattern

that reflects processes occurring when connectivity is laid down, but that the adult expression pattern is not causally related to connectivity at the scale we studied. An alternative is that the expression patterns in adulthood are functionally relevant with respect to connectivity, perhaps in modulating activity in certain pathways. The patterns we identified could be used to design experiments to distinguish between these alternatives.

The connectivity linked gene sets differ from the Pattern NE and OE gene lists presented previously in Chapter 5. Those genes were selected on the basis of spatial anti-correlation and we observed relationships to connection degree, not the connectivity patterns. Detailed comparison reveals thirteen of the pattern NE genes and five pattern OE genes overlap with the connectivity-optimized gene sets. Using the methods of this Chapter, the pooled Pattern NE and OE gene list does not contain significant information about connectivity patterns.

While we have provided evidence for a relationship between connectivity and gene expression in the mammalian brain, our analysis is surely hindered by the incompleteness of connectivity and expression information. There are many brain regions for which we had expression data but no connectivity. While some of these regions might never have been studied, there are many reports in the literature that are not included in the current connectivity databases. Advances in the generation of connectivity information from new experiments or from more complete use of existing reports will be essential. The availability of additional expression data would also improve our ability to interpret the patterns we observe. In particular, having detailed information on gene expression patterns during development, and their relationships to the developing projection patterns in the brain, could permit stronger inference of causal relationships. A final limitation is that the structural

connections we use cannot be easily linked to specific states or functions of the brain. Because of this we could only interpret our results in the context of gene function information. It would be of interest to employ functional connectivity data to link gene expression to more dynamic and task specific states of the brain, especially in the context of genetic variation.

Chapter 7: Conclusion

7.1 Summary

My thesis was focused on applying bioinformatics techniques to neuroanatomical connectivity. The first objective was to create a database of neuroanatomical connectivity from neuroscience literature. Our informatics approach to extracting connectivity statements details unparalleled resources and evaluations for neuroscience text mining. The second was to examine relationships between neuroanatomy and gene expression. This second objective resulted in the discovery of several novel patterns that provide new insight into global brain architecture.

7.1.1 Extraction of connectivity statements from text

To address the fragmented nature of neuroanatomical connectivity reports we annotated a set of 1,377 abstracts for brain region mentions and connectivity relations. Using this corpus I developed and evaluated state of the art technologies for three tasks required for extracting connectivity relationships between brain region pairs. Our results and evaluations provide the most critical assessment of text mining for neuroscience to date.

Our text mining system differs from related work of Burns and colleagues. Burns et al. created a system to automatically label detailed variables that describe connectivity experiments (in full text articles) (Burns et al., 2007). In contrast, our work focuses on

summary statements in abstracts to extract brain region mentions and relationships between them. Both sets of results support the value and feasibility of automatically extracting connectivity information from natural language text.

Chapter 2 covers the first task of recognizing mentions of brain regions in free text. From our analysis we suspect a large amount of error is due to conjunctions, previously unseen words and brain regions of less commonly studied organisms. We found context windows, lemmatization and abbreviation expansion to be the most informative techniques. We implemented a conditional random field classifier that was able to label brain region mentions at 76% recall and 81% precision. This performance is much higher than naive dictionary-based methods. Although textual features derived from the neuroscience domain did increase performance, we found that most of the knowledge needed to extract brain region mentions can be learned from a large set of examples. To reduce lexical variation and link the brain region mentions to existing databases we normalize brain region mentions to standardized identifiers in five existing neuroanatomical lexicons (Chapter 3). Based on the analysis of the manually annotated corpus, we estimate mentions are mapped at 95% precision and 63% recall. Our results provide insights into the patterns of publication on brain regions and species of study in the Journal of Comparative Neurology, but also point to important challenges in the standardization of neuroanatomical nomenclatures. We find that many terms in the formal terminologies never appear in our corpus, while conversely; many terms authors use are not reflected in the terminologies. To improve the terminologies we deposited 136 unrecognized brain regions into the Neuroscience Lexicon (NeuroLex).

Chapter 4 builds on Chapters 2 and 3 by extracting connectivity relationships between brain region mentions. I tested several methods on our annotated corpus in a cross-validation framework. Of these methods, the shallow linguistic kernel recalled 50% of the sentence level connectivity statements at 70% precision. The all-paths graph and k-band shortest path spectrum kernels provided similar performance. Due to its speed and simplicity we applied the shallow linguistic kernel to 12,557 abstracts, resulting in 28,107 connectivity relationships. We compared a normalized subset of 2,688 relationships to BAMS (Bota et al., 2005). The extracted connections were connected in BAMS at a rate of 63.5%, compared to 51.1% for co-occurring brain region pairs. By aggregating the data into a connectivity matrix form we found that precision can be increased at the cost of recall by requiring connections to occur more than once across the corpus.

7.1.2 Relationships between neuroanatomy and gene expression

In Chapters 5 and 6 we examine complex patterns of gene expression in the rodent brain in the context of regional brain connectivity and differences in cellular populations. We utilized a large data set of the rat brain "connectome" from the Brain Architecture Management System (Bota et al., 2005) and used statistical approaches to relate the data to the gene expression signatures in 142 anatomical regions from the Allen Brain Atlas (Lein et al., 2007). In Chapter 5 we identified two novel patterns of mouse brain gene expression showing a strong degree of anti-correlation, and relate this to multiple data modalities including connectivity. We found that these signatures are associated with differences in expression of neuronal and oligodendrocyte markers, suggesting they reflect regional

differences in cellular populations. We also find that the expression level of these genes is correlated with connectivity degree, with regions expressing the neuron-enriched pattern having more connections with other regions. Chapter 6 goes beyond the number of connections per region to discover relationships between gene expression and specific neuroanatomical connections. Our analysis shows that adult gene expression signatures have a statistically significant relationship to connectivity. In particular, brain regions that have similar expression profiles tend to have similar connectivity profiles, and this effect is not entirely attributable to spatial correlations. In addition, brain regions which are connected have more similar expression patterns. Using a simple optimization approach, we identified a set of genes most correlated with neuroanatomical connectivity, and find that this set is enriched for genes involved in neuronal development and axon guidance. Further, a number of the genes have been implicated in neurodevelopmental disorders such as autistic spectrum disorder. Our results have the potential to shed light on the role of gene expression patterns in influencing neuronal activity and connectivity, with potential applications to our understanding of brain disorders.

Our results in Chapter 6 answer questions first posed when the ABA data was first described (Lein et al., 2007). The interest of the field in these questions is confirmed by the work of Wolf and colleagues, who replicated our essential findings (Wolf et al., 2011). Using the same sources of connectivity and gene expression they support our findings by employing a classification framework that predicts connectivity from gene expression data. In addition they corroborate our finding that autism associated genes carry significant information about

connectivity.

7.2 Conclusions

We describe and apply a system for large scale automatic extraction of connectivity knowledge. By analyzing over 13,000 abstracts we found the neuroscience literature contains a wide diversity of terms, organisms of study, and brain region descriptions. Unfortunately, this diversity far exceeds that of the existing formalized neuroanatomical lexicons and our manually curated dataset. While this limits the automatic resolution of brain region mentions, we were able to implement several methods that improve automatic extraction. Our results suggest it is feasible to generate a useful database of connectivity statements from neuroscience abstracts.

Our text mining approach provides a novel collection of data to the growing list of neuroinformatics resources. In Chapters 5 and 6, we demonstrate the value of similar large scale neuroscience datasets by integrating a large connectivity database with a brain-wide gene expression atlas. The results show that the wealth of formalized knowledge at the gene level provides valuable insight into neuroanatomy at the brain region level. Specifically, there is a relationship between patterns of gene expression and connectivity in the adult rodent brain. These relationships are linked to cell-type expression signatures that provide new insight into brain architecture. Although we observe only correlations, we used our methods to prioritize specific genes that can be targeted by experimental manipulations to reveal causality. Several of these genes are already associated with disorders involving

abnormal brain development and connectivity.

A final conclusion is that large collections of neuroinformatics data, when combined, provide new insight into global brain architecture. Further application of computational tools to process, integrate, analyze and interpret large heterogeneous neuroscience data will improve our understanding of the brain and its complexity.

7.3 Future Research Directions

Several clear avenues of research can extend the WhiteText project. For example, a large number of abstracts describing neuroanatomical connectivity are available outside of the Journal of Comparative Neurology. By applying our pipeline of brain region recognition, resolution and relationship extraction we can aggregate a much larger set of connectivity statements. Before expansion, we first seek to make our current set of over 28,000 predicted connectivity statements more accessible to neuroscientists for building models and refining hypotheses. These mined statements can also provide literature based context and validation for connections revealed by mouse and human connectome projects. Creation of an information retrieval portal that allows searching of the data is a future objective. Before this data is released we plan detailed evaluations of the predicted relationships to fully measure the accuracy of the complete system. For decreasing amounts of predicted relationships we will evaluate in the context of the sentence, abstract, full text paper and complete literature (tests if the connection has been refuted by other results). Our contribution to the community has already begun with our additions to the NeuroLex resource. In addition to new brain

region concepts extracted from literature, our data provides extensive data on synonyms, frequency of use and co-occurrences that can be used to improve lexical resources.

For the ABAMS project several directions may elucidate the relationships between gene expression and connectivity. Going beyond patterns OE and NE found in Chapter 5, we observe further clustering into patterns as the number of anticorrelated gene pairs grows. We suspect these additional gene clusters are linked to differing populations of neuron and oligodendrocyte types.

Exploration of the relationships described in Chapter 5 and 6 in other stages or organisms are attractive future directions for the ABAMS project. The availability of spatially registered developmental mouse brain expression data would improve our ability to interpret the patterns we observe. In particular, having detailed information on gene expression patterns during development, and their relationships to the developing projection patterns in the brain, could permit stronger inference of causal relationships.

A limitation of the ABAMS findings is that the structural connections we use cannot be easily linked to specific states or functions of the brain. This restricted our interpretation of results to functional information associated with genes. It would be of interest to employ functional connectivity data to link gene expression to more dynamic and task specific states of the brain, especially in the context of genetic variation. The literature provides a possible source of functional data that we can extract with the methods created for the WhiteText project. For example, co-occurrences between extracted brain regions (functional connectivity) and terms like “addiction” or “memory” (functional activation) can provide

a large dataset of functional associations that can be analyzed in light of gene expression patterns.

While we have provided evidence for a relationship between connectivity and gene expression in the mammalian brain, our analysis is surely hindered by the incompleteness of connectivity information. There are many brain regions for which we had expression data but no connectivity. While some of these regions might never have been studied, there are many reports in the literature that are not included in the current connectivity databases. Our advances in extracting connectivity reports from the biomedical literature can address this need. Our new connectivity database backed by the literature can be directly examined for relationships to gene expression. Unfortunately, such an analysis would yield uncertain results given the error rates observed for connectivity statement extraction. Again, further evaluation of the large set of extracted connectivity relations will help refine the dataset.

Although I provide some insight into the complexity of the brain my work is limited to rodent neuroanatomy. It is my hope that the work conducted for this dissertation will guide similar studies of the human brain. This hope is supported by two large projects that are undertaking the immense tasks of characterizing the human connectome and transcriptome. The Human Connectome Project is using MRI technologies to map brain wiring in over 1,000 subjects and the Allen Institute for Brain Science has released gene expression data covering almost 1,000 brain sites in two normal adult donors. Our methods are immediately applicable to these data of the human brain and may provide significant insight into human neuroanatomy. For example, application of the methods used in Chapter 5 could provide

brainwide estimates of neuron to glia ratios. Further, studies of relationships between gene expression and connectivity that are focused on the human brain may inform new therapies for connectivity related disorders.

References

(2006). *The BioPAX Working Group, BioPAX Biological Pathways Exchange Language. Level 2, Version 1.0 Documentation.* [Online]. Available: http://www.biopax.org [Accessed July 9 2007].

Aarnio, V., Paananen, J., and Wong, G. (2005). Analysis of microarray studies performed in the neurosciences. *J Mol Neurosci* 27**,** 261-268.

Airola, A., Pyysalo, S., Bjorne, J., Pahikkala, T., Ginter, F., and Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics* 9 Suppl 11**,** S2.

Akil, H., Martone, M.E., and Van Essen, D.C. (2011). Challenges and opportunities in mining neuroscience data. *Science* 331**,** 708-712.

Alloway, K.D., Olson, M.L., and Smith, J.B. (2008). Contralateral corticothalamic projections from MI whisker cortex: potential route for modulating hemispheric interactions. *J Comp Neurol* 510**,** 100-116.

Amari, S., Beltrame, F., Bjaalie, J.G., Dalkara, T., De Schutter, E., Egan, G.F., Goddard, N.H., Gonzalez, C., Grillner, S., Herz, A., Hoffmann, K.P., Jaaskelainen, I., Koslow, S.H., Lee, S.Y., Matthiessen, L., Miller, P.L., Da Silva, F.M., Novak, M., Ravindranath, V., Ritz, R., Ruotsalainen, U., Sebestra, V., Subramaniam, S., Tang, Y., Toga, A.W., Usui, S., Van Pelt, J., Verschure, P., Willshaw, D., and Wrobel, A. (2002). Neuroinformatics: the integration of shared databases and tools towards integrative neuroscience. *J Integr Neurosci* 1**,** 117-128.

Ao, H., and Takagi, T. (2005). ALICE: an algorithm to extract abbreviations from MEDLINE. *J Am Med Inform Assoc* 12**,** 576-586.

Aronson, A.R. (2006). MetaMap: Mapping Text to the UMLS Metathesaurus, http://skr.nlm.nih.gov/papers/references/metamap06.pdf.

Aronson, A.R., and Lang, F.M. (2010). An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 17**,** 229-236.

Ascoli, G.A. (2006). The ups and downs of neuroscience shares. *Neuroinformatics* 4**,** 213-216.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25**,** 25-29.

Assou, S., Le Carrour, T., Tondeur, S., Strom, S., Gabelle, A., Marty, S., Nadal, L., Pantesco, V., Reme, T., Hugnot, J.P., Gasca, S., Hovatta, O., Hamamah, S., Klein, B., and De Vos, J. (2007). A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells* 25**,** 961-973.

Baitaluk, M., Sedova, M., Ray, A., and Gupta, A. (2006). BiologicalNetworks: visualization and analysis tool for systems biology. *Nucleic Acids Res* 34**,** W466-471.

Baruch, L., Itzkovitz, S., Golan-Mashiach, M., Shapiro, E., and Segal, E. (2008). Using expression profiles of Caenorhabditis elegans neurons to identify genes that mediate synaptic connectivity. *PLoS Comput Biol* 4**,** e1000120.

Basu, S.N., Kollu, R., and Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. *Nucleic Acids Res* 37**,** D832-836.

Baughman, R.W., Farkas, R., Guzman, M., and Huerta, M.F. (2006). The National Institutes of Health Blueprint for Neuroscience Research. *J Neurosci* 26**,** 10329-10331.

Behrens, T.E., Johansen-Berg, H., Woolrich, M.W., Smith, S.M., Wheeler-Kingshott, C.A., Boulby, P.A., Barker, G.J., Sillery, E.L., Sheehan, K., Ciccarelli, O., Thompson, A.J., Brady, J.M., and Matthews, P.M. (2003). Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nat Neurosci* 6**,** 750-757.

Belmonte, M.K., Allen, G., Beckel-Mitchener, A., Boulanger, L.M., Carper, R.A., and Webb, S.J. (2004). Autism and abnormal development of brain connectivity. *J Neurosci* 24**,** 9228-9231.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B* 57**,** 289-300.

Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kotter, R., Li, S.J., Lin, C.P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., Mcmahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.F., Zhang, H.Y., Castellanos, F.X., and Milham, M.P. (2010). Toward discovery science of human brain function. *Proc Natl Acad Sci U S A* 107**,** 4734-4739.

Blaschke, C., Andrade, M.A., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol***,** 60-67.

Bohland, J.W., Bokil, H., Allen, C.B., and Mitra, P.P. (2009a). The brain atlas concordance problem: quantitative comparison of anatomical parcellations. *PLoS One* 4**,** e7200.

Bohland, J.W., Bokil, H., Pathak, S.D., Lee, C.K., Ng, L., Lau, C., Kuan, C., Hawrylycz, M., and Mitra, P.P. (2009b). Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. *Methods* 50**,** 105-112.

Bonora, E., Lamb, J.A., Barnby, G., Sykes, N., Moberly, T., Beyer, K.S., Klauck, S.M., Poustka, F., Bacchelli, E., Blasi, F., Maestrini, E., Battaglia, A., Haracopos, D., Pedersen, L., Isager, T., Eriksen, G., Viskum, B., Sorensen, E.U., Brondum-Nielsen, K., Cotterill, R., Engeland, H., Jonge, M., Kemner, C., Steggehuis, K., Scherpenisse, M., Rutter, M., Bolton, P.F., Parr, J.R., Poustka, A., Bailey, A.J., and Monaco, A.P. (2005). Mutation screening and association analysis of six candidate genes for autism on chromosome 7q. *Eur J Hum Genet* 13**,** 198-207.

Bota, M., Dong, H.W., and Swanson, L.W. (2003). From gene networks to brain networks. *Nat Neurosci* 6**,** 795-799.

Bota, M., Dong, H.W., and Swanson, L.W. (2005). Brain architecture management system. *Neuroinformatics* 3**,** 15-48.

Bota, M., and Swanson, L.W. (2008). BAMS Neuroanatomical Ontology: Design and Implementation. *Front Neuroinformatics* 2**,** 2.

Bota, M., and Swanson, L.W. (2010). Collating and Curating Neuroanatomical Nomenclatures: Principles and Use of the Brain Architecture Knowledge Management System (BAMS). *Front Neuroinformatics* 4**,** 3.

Bowden, D.M., Dubach, M., and Park, J. (2007). Creating neuroscience ontologies. *Methods Mol Biol* 401**,** 67-87.

Bowden, D.M., and Dubach, M.F. (2002). "BrainInfo. An Online Interactive Brain Atlas and Nomenclature," in *Neuroscience Databases,* ed. K. R. (Dusseldorf: Kluwer Academic Press), 259–274.

Bowden, D.M., and Dubach, M.F. (2003). NeuroNames 2002. *Neuroinformatics* 1**,** 43-59.

Brette, R., Rudolph, M., Carnevale, T., Hines, M., Beeman, D., Bower, J.M., Diesmann, M., Morrison, A., Goodman, P.H., Harris, F.C., Jr., Zirpe, M., Natschlager, T., Pecevski, D., Ermentrout, B., Djurfeldt, M., Lansner, A., Rochel, O., Vieville, T., Muller, E., Davison, A.P., El Boustani, S., and Destexhe, A. (2007). Simulation of networks of spiking neurons: A review of tools and strategies. *J Comput Neurosci*.

Brinkley, J.F., and Rosse, C. (2002). Imaging and the Human Brain Project: a review. *Methods Inf Med* 41**,** 245-260.

Broadwell, R.D., and Jacobowitz, D.M. (1976). Olfactory relationships of the telencephalon and diencephalon in the rabbit. III. The ipsilateral centrifugal fibers to the olfactory bulbar and retrobulbar formations. *J Comp Neurol* 170**,** 321-345.

Bug, W.J., Ascoli, G.A., Grethe, J.S., Gupta, A., Fennema-Notestine, C., Laird, A.R., Larson, S.D., Rubin, D., Shepherd, G.M., Turner, J.A., and Martone, M.E. (2008). The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics* 6**,** 175-194.

Burns, G., Feng, D., and Ehovy, E. (2007). "Intelligent Approaches to Mining the Primary Research Literature: Techniques, Systems, and Examples," in *Computational Intelligence in Bioinformatics,* eds. A. Kelemen, A. Abraham & Y. Chen. Springer-Verlag, Germany).

Burns, G.A., and Cheng, W.C. (2006). Tools for knowledge acquisition within the NeuroScholar system and their application to anatomical tract-tracing data. *J Biomed Discov Collab* 1**,** 10.

Burns, G.A., Cheng, W.C., Thompson, R.H., and Swanson, L.W. (2006). The NeuARt II system: a viewing tool for neuroanatomical data based on published neuroanatomical atlases. *BMC Bioinformatics* 7**,** 531.

Buyko, E., Tomanek, K., and Hahn, U. (Year). "Resolution of coordination ellipses in biological named entities using conditional random fields", in: *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*), 163-171.

Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L., Christopherson, K.S., Xing, Y., Lubischer, J.L., Krieg, P.A., Krupenko, S.A., Thompson, W.J., and Barres, B.A. (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci* 28**,** 264-278.

Card, J.P., and Moore, R.Y. (1989). Organization of lateral geniculate-hypothalamic connections in the rat. *J Comp Neurol* 284**,** 135-147.

Chedotal, A., Del Rio, J.A., Ruiz, M., He, Z., Borrell, V., De Castro, F., Ezan, F., Goodman, C.S., Tessier-Lavigne, M., Sotelo, C., and Soriano, E. (1998). Semaphorins III and IV repel hippocampal axons via two distinct receptors. *Development* 125**,** 4313-4323.

Chen, B.L., Hall, D.H., and Chklovskii, D.B. (2006). Wiring optimization can relate neuronal structure and function. *Proc Natl Acad Sci U S A* 103**,** 4723-4728.

Chen, T., Hui, R., Wang, X.L., Zhang, T., Dong, Y.X., and Li, Y.Q. (2008). Origins of endomorphin-immunoreactive fibers and terminals in different columns of the periaqueductal gray in the rat. *J Comp Neurol* 509**,** 72-87.

Chilton, J.K. (2006). Molecular mechanisms of axon guidance. *Dev Biol* 292**,** 13-24.

Collins, M., and Duffy, N. (2001). "Convolution kernels for natural language", in: *Proc. of Neural Information Processing Systems (NIPS'01).* (Vancouver, BC, Canada).

Costa Lda, F., Kaiser, M., and Hilgetag, C.C. (2007). Predicting the connectivity of primate cortical networks from topological and spatial node properties. *BMC Syst Biol* 1**,** 16.

Crasto, C., Marenco, L., Miller, P., and Shepherd, G. (2002). Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. *Nucleic Acids Res* 30**,** 354-360.

Crasto, C.J., Marenco, L.N., Liu, N., Morse, T.M., Cheung, K.H., Lai, P.C., Bahl, G., Masiar, P., Lam, H.Y., Lim, E., Chen, H., Nadkarni, P., Migliore, M., Miller, P.L., and Shepherd, G.M. (2007a). SenseLab: new developments in disseminating neuroscience information. *Brief Bioinform* 8**,** 150-162.

Crasto, C.J., Marenco, L.N., Migliore, M., Mao, B., Nadkarni, P.M., Miller, P., and Shepherd, G.M. (2003). Text mining neuroscience journal articles to populate neuroscience databases. *Neuroinformatics* 1**,** 215-237.

Crasto, C.J., Masiar, P., and Miller, P.L. (2007b). NeuroExtract: facilitating neuroscience-oriented retrieval from broadly-focused bioscience databases using text-based query mediation. *J Am Med Inform Assoc* 14**,** 355-360.

Crick, F., and Jones, E. (1993). Backwardness of human neuroanatomy. *Nature* 361**,** 109-110.

Cunningham, E.T., Jr., and Sawchenko, P.E. (2000). Dorsal medullary pathways subserving oromotor reflexes in the rat: implications for the central neural control of swallowing. *J Comp Neurol* 417**,** 448-466.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). A framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)***,** 168-175.

De Angelis, E., Watkins, A., Schafer, M., Brummendorf, T., and Kenwrick, S. (2002). Disease-associated mutations in L1 CAM interfere with ligand interactions and cell-surface expression. *Hum Mol Genet* 11**,** 1-12.

Destexhe, A., and Contreras, D. (2006). Neuronal computations with stochastic network states. *Science* 314**,** 85-90.

Dong, H.W. (2007). *The Allen Atlas: A Digital Brain Atlas of C57BL/6J Male Mouse.* Hoboken, NJ: Wiley.

Dong, H.W., Swanson, L.W., Chen, L., Fanselow, M.S., and Toga, A.W. (2009). Genomic-

anatomic evidence for distinct functional domains in hippocampal field CA1. *Proc Natl Acad Sci U S A* 106**,** 11794-11799.

Dyhrfjeld-Johnsen, J., Maier, J., Schubert, D., Staiger, J., Luhmann, H.J., Stephan, K.E., and Kotter, R. (2005). CoCoDat: a database system for organizing and selecting quantitative data on single neurons and neuronal microcircuitry. *J Neurosci Methods* 141**,** 291-308.

Eckersley, P., Egan, G.F., Amari, S., Beltrame, F., Bennett, R., Bjaalie, J.G., Dalkara, T., De Schutter, E., Gonzalez, C., Grillner, S., Herz, A., Hoffmann, K.P., Jaaskelainen, I.P., Koslow, S.H., Lee, S.Y., Matthiessen, L., Miller, P.L., Da Silva, F.M., Novak, M., Ravindranath, V., Ritz, R., Ruotsalainen, U., Subramaniam, S., Toga, A.W., Usui, S., Van Pelt, J., Verschure, P., Willshaw, D., Wrobel, A., and Tang, Y. (2003). Neuroscience data and tool sharing: a legal and policy framework for neuroinformatics. *Neuroinformatics* 1**,** 149-165.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30**,** 207-210.

Emes, R.D., Pocklington, A.J., Anderson, C.N., Bayes, A., Collins, M.O., Vickers, C.A., Croning, M.D., Malik, B.R., Choudhary, J.S., Armstrong, J.D., and Grant, S.G. (2008). Evolutionary expansion and anatomical specialization of synapse proteome complexity. *Nat Neurosci* 11**,** 799-806.

Evans, A.C. (2006). The NIH MRI study of normal brain development. *Neuroimage* 30**,** 184-202.

Felleman, D.J., and Van Essen, D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1**,** 1-47.

Gabbott, P.L., Warner, T.A., Jays, P.R., Salway, P., and Busby, S.J. (2005). Prefrontal cortex in the rat: projections to subcortical autonomic, motor, and limbic centers. *J Comp Neurol* 492**,** 145-177.

Gardner, D., Abato, M., Knuth, K.H., Debellis, R., and Erde, S.M. (2001). Dynamic publication model for neurophysiology databases. *Philos Trans R Soc Lond B Biol Sci* 356**,** 1229-1247.

Gardner, D., Akil, H., Ascoli, G.A., Bowden, D.M., Bug, W., Donohue, D.E., Goldberg, D.H., Grafstein, B., Grethe, J.S., Gupta, A., Halavi, M., Kennedy, D.N., Marenco, L., Martone, M.E., Miller, P.L., Muller, H.M., Robert, A., Shepherd, G.M., Sternberg, P.W., Van Essen, D.C., and Williams, R.W. (2008). The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics* 6**,** 149-160.

Gerner, M., Nenadic, G., and Bergman, C.M. (2010). LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics* 11**,** 85.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res* 17**,** 669-681.

Geschwind, D.H., and Levitt, P. (2007). Autism spectrum disorders: developmental disconnection syndromes. *Curr Opin Neurobiol* 17**,** 103-111.

Ghandour, M.S., Langley, O.K., Labourdette, G., Vincendon, G., and Gombos, G. (1981). Specific and artefactual cellular localizations of S 100 protein: an astrocyte marker in rat cerebellum. *Dev Neurosci* 4**,** 66-78.

Ghandour, M.S., Langley, O.K., Vincendon, G., and Gombos, G. (1979). Double labeling immunohistochemical technique provides evidence of the specificity of glial cell markers. *J Histochem Cytochem* 27**,** 1634-1637.

Ghandour, M.S., Langley, O.K., Vincendon, G., Gombos, G., Filippi, D., Limozin, N., Dalmasso, D., and Laurent, G. (1980). Immunochemical and immunohistochemical study of carbonic anhydrase II in adult rat cerebellum: a marker for oligodendrocytes. *Neuroscience* 5**,** 559-571.

Ghazvinian, A., Noy, N.F., and Musen, M.A. (2009). Creating mappings for ontologies in biomedicine: simple methods work. *AMIA Annu Symp Proc* 2009**,** 198-202.

Gillis, J., Mistry, M., and Pavlidis, P. (2010). Gene function analysis in complex data sets using ErmineJ. *Nat Protoc* 5**,** 1148-1159.

Giuliano, C., Lavelli, A., and Romano, L. (Year). "Exploiting shallow linguistic information for relation extraction from biomedical literature", in: *Proc. of the 11st Conf. of the European Chapter of the Association for Computational Linguistics (EACL'06)*), 401–408.

Gong, S., Zheng, C., Doughty, M.L., Losos, K., Didkovsky, N., Schambra, U.B., Nowak, N.J., Joyner, A., Leblanc, G., Hatten, M.E., and Heintz, N. (2003). A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425**,** 917-925.

Group, G.S.F.N.W. (2002). "Report on Neuroinformatics, Organisation for Economic Co-operation and Development", (ed.) G.S.F.N.W. Group.).

Gu, S.M., Orth, U., Veske, A., Enders, H., Klunder, K., Schlosser, M., Engel, W., Schwinger, E., and Gal, A. (1996). Five novel mutations in the L1CAM gene in families with X linked hydrocephalus. *J Med Genet* 33**,** 103-106.

Haines, D.E. (2004). *Neuroanatomy: An Atlas of Structures, Sections, and Systems.*

Halevy, A., Norvig, P., and Pereira, F. (2009). The Unreasonable Effectiveness of Data. *Intelligent Systems, IEEE* 24**,** 8-12.

Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Muller, H.M., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E.M., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L.D., Spieth, J., and Sternberg, P.W. (2010). WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res* 38**,** D463-467.

Hawrylycz, M., Baldock, R.A., Burger, A., Hashikawa, T., Johnson, G.A., Martone, M., Ng, L., Lau, C., Larson, S.D., Nissanov, J., Puelles, L., Ruffins, S., Verbeek, F., Zaslavsky, I., and Boline, J. (2011). Digital atlasing and standardization in the mouse brain. *PLoS Comput Biol* 7**,** e1001065.

Hayasaka, S., Hugenschmidt, C.E., and Laurienti, P.J. (2011). A network of genes, genetic disorders, and brain areas. *PLoS One* 6**,** e20907.

Helmer, K.G., Ambite, J.L., Ames, J., Ananthakrishnan, R., Burns, G., Chervenak, A.L., Foster, I., Liming, L., Keator, D., Macciardi, F., Madduri, R., Navarro, J.P., Potkin, S., Rosen, B., Ruffins, S., Schuler, R., Turner, J.A., Toga, A., Williams, C., and Kesselman, C. (2011). Enabling collaborative research using the Biomedical Informatics Research Network (BIRN). *J Am Med Inform Assoc* 18**,** 416-422.

Hermoye, L., Saint-Martin, C., Cosnard, G., Lee, S.K., Kim, J., Nassogne, M.C., Menten, R., Clapuyt, P., Donohue, P.K., Hua, K., Wakana, S., Jiang, H., Van Zijl, P.C., and Mori, S. (2006). Pediatric diffusion tensor imaging: normal database and observation of the white matter maturation in early childhood. *Neuroimage* 29**,** 493-504.

Hilgetag, C.C., and Kaiser, M. (2004). Clustered organization of cortical connectivity. *Neuroinformatics* 2**,** 353-360.

Hochheiser, H., and Yanowitz, J. (2007). If I only had a brain: exploring mouse brain images in the Allen Brain Atlas. *Biol Cell* 99**,** 403-409.

Hof, P.R., Young, W.G., Bloom, F.E., Belichenko, P.V., and Celio, M.R. (2000). *Comparative Cytoarchitectonic Atlas of the C57BL/6 and 129/Sv Mouse Brains.* Elsevier.

Hole, W.T., and Srinivasan, S. (2003). Adding NeuroNames to the UMLS Metathesaurus. *Neuroinformatics* 1**,** 61-63.

Honey, C.J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.P., Meuli, R., and Hagmann, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc Natl Acad Sci U S A* 106**,** 2035-2040.

Houenou, J., Wessa, M., Douaud, G., Leboyer, M., Chanraud, S., Perrin, M., Poupon, C., Martinot, J.L., and Paillere-Martinot, M.L. (2007). Increased white matter connectivity in euthymic bipolar patients: diffusion tensor tractography between the subgenual cingulate and the amygdalo-hippocampal complex. *Mol Psychiatry* 12**,** 1001-1010.

Hsu, C.N., Chang, Y.M., Kuo, C.J., Lin, Y.S., Huang, H.S., and Chung, I.F. (2008). Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics* 24**,** i286-294.

Huerta, M.F., Liu, Y., and Glanzman, D.L. (2006). A view of the digital landscape for neuroscience at NIH. *Neuroinformatics* 4**,** 131-138.

Hugues, B., and Olivier, T. (2007). Modeling self-developing biological neural networks. *Neurocomputing* 70**,** 2723-2734.

Illig, K.R., and Eudy, J.D. (2009). Contralateral projections of the rat anterior olfactory nucleus. *J Comp Neurol* 512**,** 115-123.

Insel, T.R., Volkow, N.D., Li, T.K., Battey, J.F., Jr., and Landis, S.C. (2003). Neuroscience networks: data-sharing in an information age. *PLoS Biol* 1**,** E17.

International Molecular Genetic Study of Autism Consortium (1998). A full genome screen for autism with evidence for linkage to a region on chromosome 7q. International Molecular Genetic Study of Autism Consortium. *Hum Mol Genet* 7**,** 571-578.

Inuzuka, M., Hayakawa, M., and Ingi, T. (2005). Serinc, an activity-regulated protein family, incorporates serine into membrane lipid synthesis. *J Biol Chem* 280**,** 35776-35783.

Ip, N.Y., Mcclain, J., Barrezueta, N.X., Aldrich, T.H., Pan, L., Li, Y., Wiegand, S.J., Friedman, B., Davis, S., and Yancopoulos, G.D. (1993). The alpha component of the CNTF receptor is required for signaling and defines potential CNTF targets in the adult and during development. *Neuron* 10**,** 89-102.

Irie, A., Yates, E.A., Turnbull, J.E., and Holt, C.E. (2002). Specific heparan sulfate structures involved in retinal axon targeting. *Development* 129**,** 61-70.

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat Rev Neurosci* 2**,** 194-203.

Jacobsson, J.A., Stephansson, O., and Fredriksson, R. (2010). C6ORF192 forms a unique evolutionary branch among solute carriers (SLC16, SLC17, and SLC18) and is abundantly expressed in several brain regions. *J Mol Neurosci* 41**,** 230-242.

Jensen, L.J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* 7**,** 119-129.

Jolkkonen, E., and Pitkanen, A. (1998). Intrinsic connections of the rat amygdaloid complex: projections originating in the central nucleus. *J Comp Neurol* 395**,** 53-72.

Jones, A.R., Overly, C.C., and Sunkin, S.M. (2009). The Allen Brain Atlas: 5 years and beyond. *Nat Rev Neurosci* 10**,** 821-828.

Jonquet, C., Shah, N.H., and Musen, M.A. (Year). "The Open Biomedical Annotator", in: *AMIA Summit on Translational Bioinformatics*), 56–60.

Just, M.A., Cherkassky, V.L., Keller, T.A., Kana, R.K., and Minshew, N.J. (2007). Functional and anatomical cortical underconnectivity in autism: evidence from an FMRI study of an executive function task and corpus callosum morphometry. *Cereb Cortex* 17**,** 951-961.

Karlsgodt, K.H., Van Erp, T.G., Poldrack, R.A., Bearden, C.E., Nuechterlein, K.H., and Cannon, T.D. (2008). Diffusion tensor imaging of the superior longitudinal fasciculus and working memory in recent-onset schizophrenia. *Biol Psychiatry* 63**,** 512-518.

Kaufman, A., Dror, G., Meilijson, I., and Ruppin, E. (2006). Gene expression of Caenorhabditis elegans neurons carries information on their synaptic connectivity. *PLoS Comput Biol* 2**,** e167.

Kim, J.D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus--semantically annotated corpus for bio-textmining. *Bioinformatics* 19 Suppl 1**,** i180-182.

Koehler, J. (2005). Special Issue on Text Mining. *Brief Bioinform* 6**,** 220-312.

Korostynski, M., Kaminska-Chowaniec, D., Piechota, M., and Przewlocki, R. (2006). Gene expression profiling in the striatum of inbred mouse strains with distinct opioid-related phenotypes. *BMC Genomics* 7**,** 146.

Koshino, H., Carpenter, P.A., Minshew, N.J., Cherkassky, V.L., Keller, T.A., and Just, M.A. (2005). Functional connectivity in an fMRI working memory task in high-functioning autism. *Neuroimage* 24**,** 810-821.

Koslow, S.H. (2000). Should the neuroscience community make a paradigm shift to sharing primary data? *Nat Neurosci* 3**,** 863-865.

Koslow, S.H. (2005). Discovery and integrative neuroscience. *Clin EEG Neurosci* 36**,** 55-63.

Kotecha N, B.K., Lu W, Shah N (Year). "Pathway Knowledge Base: Integrating BioPAX Compliant Data Sources, HCLS Workshop", in: *5th International Semantic Web Conference*).

Kotter, R. (2004). Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database. *Neuroinformatics* 2**,** 127-144.

Kotter, R., and Wanke, E. (2005). Mapping brains without coordinates. *Philos Trans R Soc Lond B Biol Sci* 360**,** 751-766.

Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L., and Valencia, A. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol* 9 Suppl 2**,** S1.

Kubicki, M., Mccarley, R., Westin, C.F., Park, H.J., Maier, S., Kikinis, R., Jolesz, F.A., and Shenton, M.E. (2007). A review of diffusion tensor imaging studies in schizophrenia. *J Psychiatr Res* 41**,** 15-30.

Kuboyama, T., Hirata, K., Kashima, H., Aoki-Kinoshita, K., and Yasuda, H. (2007). A spectrum tree kernel. *Information and Media Technologies 2***,** 292–299.

Kuemerle, B., Gulden, F., Cherosky, N., Williams, E., and Herrup, K. (2007). The mouse Engrailed genes: a window into autism. *Behav Brain Res* 176**,** 121-132.

Kwack, K., Lee, K.L., Kim, M., Nam, M., Bang, H.J., Yang, J.W., Choe, K.S., Kim, S.K., Hong, M.S., Chung, J.H., and Kim, H.G. (2008). Positive association between the mesoderm specific transcript gene and autism spectrum disorder in a Korean male population. *The FASEB Journal* 22**,** 906.908.

Lafferty, J., Mccallum, A., and Pereira, F. (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", in: *Proceedings of the Eighteenth International Conference on Machine Learning.*).

Laird, A.R., Lancaster, J.L., and Fox, P.T. (2005). BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics* 3**,** 65-78.

Lanciego, J.L., and Wouterlood, F.G. (2011). A half century of experimental neuroanatomical tracing. *J Chem Neuroanat*.

Larson, S., Iman, F., Bakker, R., Pham, L., and Martone, M. (2010). "A multi-scale parts list for the brain: community-based ontology curation for neuroinformatics with NeuroLex.org", in: *Neuroinformatics 2010.* (Kobe, Japan).

Lawrie, S.M., Buechel, C., Whalley, H.C., Frith, C.D., Friston, K.J., and Johnstone, E.C. (2002). Reduced frontotemporal functional connectivity in schizophrenia associated with auditory hallucinations. *Biol Psychiatry* 51**,** 1008-1011.

Le Bihan, D., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., and Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. *J Magn Reson Imaging* 13**,** 534-546.

Lee, C.K., Sunkin, S.M., Kuan, C., Thompson, C.L., Pathak, S., Ng, L., Lau, C., Fischer, S., Mortrud, M., Slaughterbeck, C., Jones, A., Lein, E., and Hawrylycz, M. (2008). Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data. *Genome Biol* 9**,** R23.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Res* 14**,** 1085-1094.

Legendre, P., and Fortin, M.J. (1989). Spatial pattern and ecological analysis *Plant Ecology* 80**,** 107-138.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., Chen, L., Chen, L., Chen, T.M., Chin, M.C., Chong, J., Crook, B.E., Czaplinska, A., Dang, C.N., Datta, S., Dee, N.R., Desaki, A.L., Desta, T., Diep, E., Dolbeare, T.A., Donelan, M.J., Dong, H.W., Dougherty, J.G., Duncan, B.J., Ebbert, A.J., Eichele, G., Estin, L.K., Faber, C., Facer, B.A., Fields, R., Fischer, S.R., Fliss, T.P., Frensley, C., Gates, S.N., Glattfelder, K.J., Halverson, K.R., Hart, M.R., Hohmann, J.G., Howell, M.P., Jeung, D.P., Johnson, R.A., Karr, P.T., Kawal, R., Kidney, J.M., Knapik, R.H., Kuan, C.L., Lake, J.H., Laramee, A.R., Larsen, K.D., Lau, C., Lemon, T.A., Liang, A.J., Liu, Y., Luong, L.T., Michaels, J., Morgan, J.J., Morgan, R.J., Mortrud, M.T., Mosqueda, N.F., Ng, L.L., Ng, R., Orta, G.J., Overly, C.C., Pak, T.H., Parry, S.E., Pathak, S.D., Pearson, O.C., Puchalski, R.B., Riley, Z.L., Rockett, H.R., Rowland, S.A., Royall, J.J., Ruiz, M.J., Sarno, N.R., Schaffnit, K., Shapovalova, N.V., Sivisay, T., Slaughterbeck, C.R., Smith, S.C., Smith, K.A., Smith, B.I., Sodt, A.J., Stewart, N.N., Stumpf, K.R.,

Sunkin, S.M., Sutram, M., Tam, A., Teemer, C.D., Thaller, C., Thompson, C.L., Varnam, L.R., Visel, A., Whitlock, R.M., Wohnoutka, P.E., Wolkey, C.K., Wong, V.Y., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445**,** 168-176.

Leitner, F., Krallinger, M., Rodriguez-Penagos, C., Hakenberg, J., Plake, C., Kuo, C.J., Hsu, C.N., Tsai, R.T., Hung, H.C., Lau, W.W., Johnson, C.A., Saetre, R., Yoshida, K., Chen, Y.H., Kim, S., Shin, S.Y., Zhang, B.T., Baumgartner, W.A., Jr., Hunter, L., Haddow, B., Matthews, M., Wang, X., Ruch, P., Ehrler, F., Ozgur, A., Erkan, G., Radev, D.R., Krauthammer, M., Luong, T., Hoffmann, R., Sander, C., and Valencia, A. (2008). Introducing meta-services for biomedical information extraction. *Genome Biol* 9 Suppl 2**,** S6.

Letournel, F., Bocquet, A., Perrot, R., Dechaume, A., Guinut, F., Eyer, J., and Barthelaix, A. (2006). Neurofilament high molecular weight-green fluorescent protein fusion is normally expressed in neurons and transported in axons: a neuronal marker to investigate the biology of neurofilaments. *Neuroscience* 137**,** 103-111.

Lichtman, J.W., and Sanes, J.R. (2008). Ome sweet ome: what can the genome tell us about the connectome? *Curr Opin Neurobiol* 18**,** 346-353.

Lovins, J.B. (1968). Development of a Stemming Algorithm. *Mechanical translation and computational linguistics***,** 22-31.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res* 27**,** 209-220.

Marcus, D.S., Harwell, J., Olsen, T., Hodge, M., Glasser, M.F., Prior, F., Jenkinson, M., Laumann, T., Curtiss, S.W., and Van Essen, D.C. (2011). Informatics and data mining tools and strategies for the human connectome project. *Front Neuroinform* 5**,** 4.

Markham, K., Schuurmans, C., and Weiss, S. (2007). STAT5A/B activity is required in the developing forebrain and spinal cord. *Mol Cell Neurosci* 35**,** 272-282.

Markram, H. (2006). The blue brain project. *Nat Rev Neurosci* 7**,** 153-160.

Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., Thiruvahindrapduram, B., Fiebig, A., Schreiber, S., Friedman, J., Ketelaars, C.E., Vos, Y.J., Ficicioglu, C., Kirkpatrick, S., Nicolson, R., Sloman, L., Summers, A., Gibbons, C.A., Teebi, A., Chitayat, D., Weksberg, R., Thompson, A., Vardy, C., Crosbie, V., Luscombe, S., Baatjes, R., Zwaigenbaum, L., Roberts, W., Fernandez, B., Szatmari, P., and Scherer, S.W. (2008). Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 82**,** 477-488.

Martone, M.E., Gupta, A., and Ellisman, M.H. (2004). E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nat Neurosci* 7**,** 467-472.

Mccallum, A. (2002). *MALLET: A Machine Learning for Language Toolkit.* ***http://mallet.cs.umass.edu*** [Online]. Available: http://mallet.cs.umass.edu [Accessed].

Mccallum, A. (2003). "Efficiently Inducing Features of Conditional Random Fields", in: *Conference on Uncertainty in Artificial Intelligence.* (Acapulco, Mexico).

Miotke, J.A., Maclennan, A.J., and Meyer, R.L. (2007). Immunohistochemical localization of CNTFRalpha in adult mouse retina and optic nerve following intraorbital nerve

crush: evidence for the axonal loss of a trophic factor receptor after injury. *J Comp Neurol* 500**,** 384-400.
Mirnics, K., and Pevsner, J. (2004). Progress in the use of microarray technology to study the neurobiology of disease. *Nature Neuroscience* 7**,** 434-439.
Miura, H., Oda, K., Endo, C., Yamazaki, K., Shibasaki, H., and Kikuchi, T. (1993). Progressive degeneration of motor nerve terminals in GAD mutant mouse with hereditary sensory axonopathy. *Neuropathol Appl Neurobiol* 19**,** 41-51.
Miyashita, T., Ichinohe, N., and Rockland, K.S. (2007). Differential modes of termination of amygdalothalamic and amygdalocortical projections in the monkey. *J Comp Neurol* 502**,** 309-324.
Modha, D.S., and Singh, R. (2010). Network architecture of the long-distance pathways in the macaque brain. *Proc Natl Acad Sci U S A* 107**,** 13485-13490.
Moldrich, R.X., Pannek, K., Hoch, R., Rubenstein, J.L., Kurniawan, N.D., and Richards, L.J. (2010). Comparative mouse brain tractography of diffusion magnetic resonance imaging. *Neuroimage* 51**,** 1027-1036.
Molloy, C.A., Keddache, M., and Martin, L.J. (2005). Evidence for linkage on 21q and 7q in a subset of autism characterized by developmental regression. *Mol Psychiatry* 10**,** 741-746.
Moore, R.Y., Halaris, A.E., and Jones, B.E. (1978). Serotonin neurons of the midbrain raphe: ascending projections. *J Comp Neurol* 180**,** 417-438.
Moschitti, A. (2005). Efficient convolution kernels for dependency and constituent syntactic trees. *Proc. of The 17th European Conf. on Machine Learning***,** 318–329.
Muller, H.M., Kenny, E.E., and Sternberg, P.W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2**,** e309.
Muller, H.M., Rangarajan, A., Teal, T.K., and Sternberg, P.W. (2008). Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers. *Neuroinformatics* 6**,** 195-204.
Murray, K.D., Choudary, P.V., and Jones, E.G. (2007). Nucleus- and cell-specific gene expression in monkey thalamus. *Proc Natl Acad Sci U S A* 104**,** 1989-1994.
Nagata, K., Suzuki, H., Niiya-Kato, A., Kinoshita, S., Taketani, S., and Araki, M. (2006). Neurensin-1 expression in the mouse retina during postnatal development and in the cultured retinal neurons. *Brain Res* 1081**,** 65-71.
Nauta, W.J. (1952). Selective silver impregnation of degenerating axons in the central nervous system. *Stain Technol* 27**,** 175-179.
Ng, L., Bernard, A., Lau, C., Overly, C.C., Dong, H.W., Kuan, C., Pathak, S., Sunkin, S.M., Dang, C., Bohland, J.W., Bokil, H., Mitra, P.P., Puelles, L., Hohmann, J., Anderson, D.J., Lein, E.S., Jones, A.R., and Hawrylycz, M. (2009). An anatomic gene expression atlas of the adult mouse brain. *Nat Neurosci* 12**,** 356-362.
Nielsen, F.A. (2003). "The Brede database: a small database for functional neuroimaging", in: *9th International Conference on Functional Mapping of the Human Brain.* (New York, NY).
Nielsen, F.A., Christensen, M.S., Madsen, K.H., Lund, T.E., and Hansen, L.K. (2006). fMRI neuroinformatics. *IEEE Eng Med Biol Mag* 25**,** 112-119.
Nielsen, F.A., and Hansen, L.K. (2002). Modeling of activation data in the BrainMap database: detection of outliers. *Hum Brain Mapp* 15**,** 146-156.

Nielsen, F.A., and Hansen, L.K. (2004). Finding related functional neuroimaging volumes. *Artif Intell Med* 30**,** 141-151.
Nielsen, F.A., Hansen, L.K., and Balslev, D. (2004). Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics* 2**,** 369-380.
Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., and Musen, M.A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 37**,** W170-173.
Nunez, R., Gross, G.H., and Sachs, B.D. (1986). Origin and central projections of rat dorsal penile nerve: possible direct projection to autonomic and somatic neurons by primary afferents of nonmuscle origin. *J Comp Neurol* 247**,** 417-429.
Ouimet, C.C., Mcguinness, T.L., and Greengard, P. (1984). Immunocytochemical localization of calcium/calmodulin-dependent protein kinase II in rat brain. *Proc Natl Acad Sci U S A* 81**,** 5604-5608.
Pan, F., Chiu, C.H., Pulapura, S., Mehan, M.R., Nunez-Iglesias, J., Zhang, K., Kamath, K., Waterman, M.S., Finch, C.E., and Zhou, X.J. (2007). Gene Aging Nexus: a web database and data mining platform for microarray data on aging. *Nucleic Acids Res* 35**,** D756-759.
Parker, G.J. (2004). Analysis of MR diffusion weighted images. *Br J Radiol* 77 Spec No 2**,** S176-185.
Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., and Brazma, A. (2007). ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35**,** D747-750.
Paxinos, G., and Franklin, K.B.J. (2001). *The Mouse Brain in Stereotaxic Coordinates.* San Diego: Academic Press.
Paxinos, G., and Franklin, K.B.J. (2008). *The Mouse Brain in Stereotaxic Coordinates.* San Diego: Academic Press.
Paxinos, G., and Watson, C. (2007). *The Rat Brain in Stereotaxic Coordinates.* Academic Press.
Perez-Escudero, A., and De Polavieja, G.G. (2007). Optimally wired subnetwork determines neuroanatomy of Caenorhabditis elegans. *Proc Natl Acad Sci U S A* 104**,** 17180-17185.
Persico, A.M., D'agruma, L., Maiorano, N., Totaro, A., Militerni, R., Bravaccio, C., Wassink, T.H., Schneider, C., Melmed, R., Trillo, S., Montecchi, F., Palermo, M., Pascucci, T., Puglisi-Allegra, S., Reichelt, K.L., Conciatori, M., Marino, R., Quattrocchi, C.C., Baldi, A., Zelante, L., Gasparini, P., and Keller, F. (2001). Reelin gene alleles and haplotypes as a factor predisposing to autistic disorder. *Mol Psychiatry* 6**,** 150-159.
Pfeiffer, B., Norman, A.W., and Hamprecht, B. (1989). Immunocytochemical characterization of neuron-rich rat brain primary cultures: calbindin D28K as marker of a neuronal subpopulation. *Brain Res* 476**,** 120-128.
Pijpers, A., Voogd, J., and Ruigrok, T.J. (2005). Topography of olivo-cortico-nuclear modules in the intermediate cerebellum of the rat. *J Comp Neurol* 492**,** 193-213.
Pinganaud, G., Bernat, I., Buisseret, P., and Buisseret-Delmas, C. (1999). Trigeminal

projections to hypoglossal and facial motor nuclei in the rat. *J Comp Neurol* 415**,** 91-104.

Pocklington, A.J., Cumiskey, M., Armstrong, J.D., and Grant, S.G. (2006). The proteomes of neurotransmitter receptor complexes form modular networks with distributed functionality underlying plasticity and behaviour. *Mol Syst Biol* 2**,** 2006 0023.

Polleux, F., Ince-Dunn, G., and Ghosh, A. (2007). Transcriptional regulation of vertebrate axon guidance and synapse formation. *Nat Rev Neurosci* 8**,** 331-340.

Polymeropoulos, M.H., Lavedan, C., Leroy, E., Ide, S.E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., Stenroos, E.S., Chandrasekharappa, S., Athanassiadou, A., Papapetropoulos, T., Johnson, W.G., Lazzarini, A.M., Duvoisin, R.C., Di Iorio, G., Golbe, L.I., and Nussbaum, R.L. (1997). Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* 276**,** 2045-2047.

Poulin, J.F., Chevalier, B., Laforest, S., and Drolet, G. (2006). Enkephalinergic afferents of the centromedial amygdala in the rat. *J Comp Neurol* 496**,** 859-876.

Ragland, M., Hutter, C., Zabetian, C., and Edwards, K. (2009). Association between the ubiquitin carboxyl-terminal esterase L1 gene (UCHL1) S18Y variant and Parkinson's Disease: a HuGE review and meta-analysis. *Am J Epidemiol* 170**,** 1344-1357.

Renthal, W., Maze, I., Krishnan, V., Covington, H.E., 3rd, Xiao, G., Kumar, A., Russo, S.J., Graham, A., Tsankova, N., Kippin, T.E., Kerstetter, K.A., Neve, R.L., Haggarty, S.J., Mckinsey, T.A., Bassel-Duby, R., Olson, E.N., and Nestler, E.J. (2007). Histone deacetylase 5 epigenetically controls behavioral adaptations to chronic emotional stimuli. *Neuron* 56**,** 517-529.

Ressler, K.J., Paschall, G., Zhou, X.L., and Davis, M. (2002). Regulation of synaptic plasticity genes during consolidation of fear conditioning. *J Neurosci* 22**,** 7892-7902.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A.M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6**,** 1-6.

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat Neurosci* 3 Suppl**,** 1199-1204.

Rosengren, L.E., Kjellstrand, P., Aurell, A., and Haglid, K.G. (1986). Irreversible effects of dichloromethane on the brain after long term exposure: a quantitative study of DNA and the glial cell marker proteins S-100 and GFA. *Br J Ind Med* 43**,** 291-299.

Runko, E., and Kaprielian, Z. (2002). Expression of Vema in the developing mouse spinal cord and optic chiasm. *J Comp Neurol* 451**,** 289-299.

Runko, E., and Kaprielian, Z. (2004). Caenorhabditis elegans VEM-1, a novel membrane protein, regulates the guidance of ventral nerve cord-associated axons. *J Neurosci* 24**,** 9015-9026.

Sadakata, T., Washida, M., Iwayama, Y., Shoji, S., Sato, Y., Ohkura, T., Katoh-Semba, R., Nakajima, M., Sekine, Y., Tanaka, M., Nakamura, K., Iwata, Y., Tsuchiya, K.J., Mori, N., Detera-Wadleigh, S.D., Ichikawa, H., Itohara, S., Yoshikawa, T., and Furuichi, T. (2007). Autistic-like phenotypes in Cadps2-knockout mice and aberrant CADPS2 splicing in autistic patients. *J Clin Invest* 117**,** 931-943.

Saric, J., Jensen, L.J., Ouzounova, R., Rojas, I., and Bork, P. (2006). Extraction of regulatory gene/protein networks from Medline. *Bioinformatics* 22**,** 645-650.

Scannell, J.W., Blakemore, C., and Young, M.P. (1995). Analysis of connectivity in the cat cerebral cortex. *J Neurosci* 15**,** 1463-1483.
Schmid, H. (Year). "Probabilistic Part-of-Speech Tagging Using Decision Trees", in: *International Conference on New Methods in Language Processing*).
Schofield, S.P., and Everitt, B.J. (1981). The organization of indoleamine neurons in the brain of the rhesus monkey (Macaca mulatta). *J Comp Neurol* 197**,** 369-383.
Schrott, A., and Kabai, P. (2008). ABCD: a functional database for the avian brain. *J Neurosci Methods* 167**,** 393-395.
Schwartz, A.S., and Hearst, M.A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput***,** 451-462.
Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007). "A quantitative theory of immediate visual recognition," in *Computational Neuroscience: Theoretical Insights into Brain Function, Progress in Brain Research, Volume 165*.).
Shepherd, G.M., Mirsky, J.S., Healy, M.D., Singer, M.S., Skoufos, E., Hines, M.S., Nadkarni, P.M., and Miller, P.L. (1998). The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci* 21**,** 460-468.
Smouse, P.E., Long, J.C., and Sokal, R.R. (1986). Multiple regression and correlation extensions of the Mantel Test of matrix correspondence. *Systematic Zoology* 35**,** 627-632.
Sobreviela, T., Pagcatipunan, M., Kroin, J.S., and Mufson, E.J. (1996). Retrograde transport of brain-derived neurotrophic factor (BDNF) following infusion in neo- and limbic cortex in rat: relationship to BDNF mRNA expressing neurons. *J Comp Neurol* 375**,** 417-444.
Sporns, O. (2011). The human connectome: a complex network. *Ann N Y Acad Sci*.
Sporns, O., and Kotter, R. (2004). Motifs in brain networks. *PLoS Biol* 2**,** e369.
Sporns, O., Tononi, G., and Kotter, R. (2005). The human connectome: A structural description of the human brain. *PLoS Comput Biol* 1**,** e42.
Srinivas, P.R., Gusfield, D., Mason, O., Gertz, M., Hogarth, M., Stone, J., Jones, E.G., and Gorin, F.A. (2003). Neuroanatomical term generation and comparison between two terminologies. *Neuroinformatics* 1**,** 177-192.
Srinivas, P.R., Wei, S.H., Cristianini, N., Jones, E.G., and Gorin, F.A. (2005). Comparison of vector space model methodologies to reconcile cross-species neuroanatomical concepts. *Neuroinformatics* 3**,** 115-131.
Sripanidkulchai, K., and Wyss, J.M. (1986). Thalamic projections to retrosplenial cortex in the rat. *J Comp Neurol* 254**,** 143-165.
Stam, C.J., Jones, B.F., Nolte, G., Breakspear, M., and Scheltens, P. (2007). Small-world networks and functional connectivity in Alzheimer's disease. *Cereb Cortex* 17**,** 92-99.
Stephan, K.E., Zilles, K., and Kotter, R. (2000). Coordinate-independent mapping of structural and functional data by objective relational transformation (ORT). *Philos Trans R Soc Lond B Biol Sci* 355**,** 37-54.
Stumpf, M.P., Thorne, T., De Silva, E., Stewart, R., An, H.J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proc Natl Acad Sci U S A* 105**,** 6959-6964.
Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., Patapoutian, A., Hampton, G.M., Schultz,

P.G., and Hogenesch, J.B. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* 99**,** 4465-4470.
Sunkin, S.M. (2006). Towards the integration of spatially and temporally resolved murine gene expression databases. *Trends Genet* 22**,** 211-217.
Swanson, L.W. (1999). *Brain Maps: Structure of the Rat Brain.* Elsevier.
Swanson, L.W. (2003). *Brain Architecture, Understanding the Basic Plan.* New York: Oxford University Press.
Swanson, L.W. (2004). *Brain Maps, Third Edition: Structure of the Rat Brain.* Oxford: Academic Press.
Swanson, L.W., and Bota, M. (2010). Foundational model of structural connectivity in the nervous system with a schema for wiring diagrams, connectome, and basic plan architecture. *Proc Natl Acad Sci U S A* 107**,** 20610-20617.
Tanabe, L., Xie, N., Thom, L.H., Matten, W., and Wilbur, W.J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics* 6 Suppl 1**,** S3.
Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., and Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol* 6**,** e1000837.
Toga, A.W. (2002). Neuroimage databases: the good, the bad and the ugly. *Nat Rev Neurosci* 3**,** 302-309.
Tojima, T., Akiyama, H., Itofusa, R., Li, Y., Katayama, H., Miyawaki, A., and Kamiguchi, H. (2007). Attractive axon guidance involves asymmetric membrane transport and exocytosis in the growth cone. *Nat Neurosci* 10**,** 58-66.
Van Essen, D. (2007). "Neuroinformatics - What's in It for You?" in: *Neuroscience Quarterly*).
Van Essen, D.C. (2009). Lost in localization--but found with foci?! *Neuroimage* 48**,** 14-17.
Van Groen, T., and Wyss, J.M. (1990). Connections of the retrosplenial granular a cortex in the rat. *J Comp Neurol* 300**,** 593-606.
Van Groen, T., and Wyss, J.M. (1992). Connections of the retrosplenial dysgranular cortex in the rat. *J Comp Neurol* 315**,** 200-216.
Van Horn, J.D., Grafton, S.T., Rockmore, D., and Gazzaniga, M.S. (2004). Sharing neuroimaging studies of human cognition. *Nat Neurosci* 7**,** 473-481.
Van Horn, J.D., and Toga, A.W. (2009). Is it time to re-prioritize neuroimaging databases and digital repositories? *Neuroimage* 47**,** 1720-1734.
Varadan, V., Miller, D.M., 3rd, and Anastassiou, D. (2006). Computational inference of the molecular logic for synaptic connectivity in C. elegans. *Bioinformatics* 22**,** e497-506.
Vishwanathan, S., and Smola, A. (2002). Fast kernels for string and tree matching. *Proc. of Neural Information Processing Systems (NIPS'02)***,** 569–576.
W3c (2004). *RDF Primer* [Online]. Available: http://www.w3.org/TR/rdf-primer/ [Accessed Feb. 24 2011].
Wain, H.M., Lush, M.J., Ducluzeau, F., Khodiyar, V.K., and Povey, S. (2004). Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res* 32**,** D255-257.
Wallach, H.M. (2004). "Conditional Random Fields: An Introduction". Department of Computer and Information Science, University of Pennsylvania).
Wan, X., and Pavlidis, P. (2007). Sharing and reusing gene expression profiling data in neuroscience. *Neuroinformatics* 5**,** 161-175.

Wang, J., Williams, R.W., and Manly, K.F. (2003). WebQTL: web-based complex trait analysis. *Neuroinformatics* 1**,** 299-308.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L., and Yaschenko, E. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 35**,** D5-12.

White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci* 314**,** 1-340.

Wolf, L., Goldberg, C., Manor, N., Sharan, R., and Ruppin, E. (2011). Gene Expression in the Rodent Brain is Associated with Its Regional Connectivity. *PLoS Comput Biol* 7**,** e1002040.

Wu, Y., Zhang, A.Q., and Yew, D.T. (2005). Age related changes of various markers of astrocytes in senescence-accelerated mice hippocampus. *Neurochem Int* 46**,** 565-574.

Xuan, W., Watson, S., and Meng, F. (2007). Tagging Sentence Boundaries in Biomedical Literature. *Lecture Notes in Computer Science* 4394**,** 186-195.

Yamaguchi, Y. (2001). Heparan sulfate proteoglycans in the nervous system: their diverse roles in neurogenesis, axon guidance, and synaptogenesis. *Semin Cell Dev Biol* 12**,** 99-106.

Yan, H., Bergner, A.J., Enomoto, H., Milbrandt, J., Newgreen, D.F., and Young, H.M. (2004). Neural cells in the esophagus respond to glial cell line-derived neurotrophic factor and neurturin, and are RET-dependent. *Dev Biol* 272**,** 118-133.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods* 8**,** 665-670.

Zapala, M.A., Hovatta, I., Ellison, J.A., Wodicka, L., Del Rio, J.A., Tennant, R., Tynan, W., Broide, R.S., Helton, R., Stoveken, B.S., Winrow, C., Lockhart, D.J., Reilly, J.F., Young, W.G., Bloom, F.E., Lockhart, D.J., and Barlow, C. (2005). Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci U S A* 102**,** 10357-10362.

Zhang, W., Zhang, Y., Zheng, H., Zhang, C., Xiong, W., Olyarchuk, J.G., Walker, M., Xu, W., Zhao, M., Zhao, S., Zhou, Z., and Wei, L. (2007). SynDB: a Synapse protein DataBase based on synapse ontology. *Nucleic Acids Res* 35**,** D737-741.

Zhou, D., and He, Y. (2008). Extracting interactions between proteins from the literature. *J Biomed Inform* 41**,** 393-407.

Appendices

Appendix A Evaluation of 100 Unmatched Mentions

Mention	Frequency	Comment
dorsomedial nuclei	2	ambiguous
hippocampus region	1	unique variant
nucleus of the electrosensory lateral line lobe	1	species, mormyrid
ventral lamella of the principal olive	1	species, cat
pontine region	1	coreference, “same pontine region”
preoptic nucleus of the hypothalamus	2	species, fish
primary gustatory sensory nuclei	1	species, fish
magnocellular lobe	1	species, octopus
rostrocaudal length of the geniculate complex	1	unique variant, “length”
vitreal retina	1	May not be brain region
thoracic intermediolateral cell column	1	spinal cord region
posterior region of the ventroposterior complex	1	species, monkey
dorsolateral column	3	spinal cord region
pars lateralis of the bed nucleus of the stria terminalis	1	too specific, can't find in atlas
dorsal parts of the dorsomedial, posterior hypothalamic nuclei	1	too specific, can't find in atlas
subdivision of parietal cortex	1	ambiguous, coreference “each subdivision”
sensory fibers of the facial nerve	1	tract not region
midrostrocaudal levels of the lateral nucleus	1	too specific, can't find in atlas
brainstem auditory nuclei	2	unique grouping of regions
vestibular	2	tract not region, “vestibular efferents”
periamygdaloid cortical region	1	unique synonym – exist in neurolex
frontal lateral neostriatum	1	species, bird
subplate zone of the pre- and parasubiculum	1	developmental term
audal superior temporal plane	1	bad annotation
periventricular stratum of the optic tectum	1	species, fish, unique modifier (stratum)
posterior half of the avcn	1	unexpanded abbreviation
second somatosensory area	4	can't find in altases
vestibular receptor organs	1	May not be brain region
ipsilateral inferior oblique muscle	1	bad annotation, not brain region
external medial and external lateral para-brachial nuclei	1	might be the hypen
anterior ectosylvian	2	species, cat

Mention	Frequency	Comment
nucleus reticularis medullaris ventralis	1	species, cat
dorsal superficial, dorsal, and suprageniculate nuclei	1	species, cat
forebrain sites	1	unique variant and coreference
dorsomedial, dorsal, and lateral cortices	1	species, lizard
cortical loci	1	nonspecific
dorsal horn of the medulla	4	species, cat
ventral horn in the spinal cord	1	Unique synonym, should match “anterior gray column” in neuronames
efferent vestibular nuclei	3	species, fish
tangential and superior vestibular nuclei	1	species, bird
medial and ventral lateral parabrachial subnuclei	1	unique variant, “sub” seems to cause trouble
interstitial vestibular region	1	species, bird
vmpo	3	unexpanded abbreviation
neurohypophyseal	1	bad annotation, references tract
external lateral and waist subnuclei	1	Unique synonym, should match “waist part of the parabrachial nucleus” in neuornames
insular and cingulate cortices	1	unique variant
antennal nerve	1	species, insect
nucleus ventrolateralis of torus semicircularis	1	species, fish
lenticular fasciculus caudodorsally	1	bad annotation, tract
parabelt auditory cortex	2	not in atlas, is defined in abstract
intercollicular region	1	strange construct, “levels of intercollicular region”
posterior ventrolateral	1	coreference, should be thalamic
subnuclear groups of the nucleus of the tractus solitarius	1	ambiguous
area ventralis telencephali pars lateralis	1	species, fish
thalamocortical relay nucleus	1	nonspecific, type of nucleus
ventral portion of globus pallidus	1	Unique synonym, should match “Ventral pallidum”
parietal insular cortex	1	too specific, can't find in atlas
rubral	2	cell group, neurons from red nucleus
cerebellar area	1	Unique synonym, should match “cerebellum”
follicle-sinus complexes	5	bad annotation, skin part
smii	2	unexpanded abbreviation
ateral septum	1	bad annotation, missed first letter
tectal cortex	2	species, bird
medial and posterior thalamic regions	1	unique variant
ventromedial subnucleus	1	coreference, should be of the hypoglossal nucleus
pallidostriatal	4	bad annotation, tract and cell group

Mention	Frequency	Comment
left bulb	1	coreference, should be the olfactory bulb
a7	2	ambiguous, cell group
auditory strip	1	not in atlas
medullary-spinal cord	1	unique term, “medullary-spinal cord junction”
c6 root	1	spinal cord region
superior and inferior parts of area 46	2	not sure, area 46 of broadmann should work
6th abdominal ganglion	1	Bad annotation, not a brain region
lateral suprasylvian visual area of cortex	1	species, cat
temporal posterior inferior area	1	species, shrew
vestibulospinal neurons	1	Bad annotation, cell group
cerebellorecipient and retinorecipient pulvinar nucleus(pul) areas	1	coreference, based on abstract
cortical areas mt	1	unexpanded abbreviation, unkown synonym – MT = middle temporal
lateral archistriatum intermedium	1	species, bird
perigeniculate nucleus	1	species, cat
ipsilateral medial rectus muscle	1	Bad annotation, muscle
n. dorsolateralis medialis	1	coreference, thalamic region
ventrolateral (vl) or ventral posterolateral (vpl) thalamic nuclei	1	not sure, ventrolateral thalamic nuclei may match
ventral dentate	1	coreference, ventral parts of dentate nucleus of the cerebellum
intermediate gray the intermediolateral nucleus in thoracic and upper lumbar segments	1	spinal cord region
posterior zone nuclei	1	species, frog
sn-vta	1	Annotation error, cell group, unexpanded abbreviation, cell group
medial to lateral, termed medial, centromedial, centrolateral, and lateral segments	1	coreference, regions of PLLL
10m	2	area 10m of Carmichael? Should match neuronames
latero-medial axis in the entorhinal cortex	1	unique variant
pviiin	3	annotation error, fibre tract
dorsal cap	4	coreference, dorsal cap of Kooy
subcortical medullary zone	1	coreference, part of cerebellum
telencephalic nucleus olfactoretinalis	1	species, fish
commissural	9	Bad annotation, tract descriptor
13l	2	area 13l of Carmichael? Should match neuronames
rhombencephalic	1	should match rhombencephalon
vision-related cortex	1	Unique synonym of area TE a
frontoparietal isocortex	1	developmental term

Mention	Frequency	Comment
areas teav	1	too specific, area TE a exists but not teav

Appendix B Mappings between the Allen and Swanson Atlases

ABA name	mapped BAMS (Swanson 98)
Abducens nucleus	Abducens nucleus
Accessory olfactory bulb	Accessory olfactory bulb
Ammon's Horn	Ammon Horn
Anterior amygdalar area	Anterior amygdaloid area
Anterior group of the dorsal thalamus	Anterior group of the dorsal thalamus
Anterior hypothalamic nucleus	Anterior hypothalamic nucleus
Anterior olfactory nucleus	Anterior olfactory nucleus
Anterior pretectal nucleus	Anterior pretectal nucleus
Anterior tegmental nucleus	Anterior tegmental nucleus
Anterodorsal nucleus	Anterodorsal nucleus of the thalamus
Anterodorsal preoptic nucleus	Anterodorsal preoptic nucleus
Anteromedial nucleus	Anteromedial nucleus of thalamus
Anteroventral nucleus of thalamus	Anteroventral nucleus of thalamus
Anteroventral periventricular nucleus	Anteroventral periventricular nucleus
Anteroventral preoptic nucleus	Anteroventral preoptic nucleus
Arcuate hypothalamic nucleus	Arcuate nucleus of the hypothalamus
Area postrema	Area postrema
Barrington's nucleus	Barrington nucleus
Basic cell groups and regions	Brain
Bed nuclei of the stria terminalis	Bed nuclei of the stria terminalis
Bed nucleus of the anterior commissure	Bed nucleus of the anterior commissure
Brain stem	Brainstem
Caudoputamen	Caudoputamen
Central amygdalar nucleus	Central nucleus of amygdala
Central lateral nucleus of the thalamus	Central lateral nucleus of the thalamus
Central linear nucleus raphé	Central linear nucleus raphe
Central medial nucleus of the thalamus	Central medial nucleus of the thalamus
Cerebellar cortex	Cerebellar cortex
Cerebellar nuclei	Deep cerebellar nuclei
Cerebellum	Cerebellum
Cerebral cortex	Cerebral cortex
Cerebral nuclei	Basal Nuclei
Cerebrum	Cerebrum
Cochlear nuclei	Cochlear nuclei
Cortical amygdalar area	Cortical nucleus of the amygdala
Cortical plate	Cerebral cortex, layers1-6a [cortical plate]
Cuneate nucleus	Cuneate nucleus
Cuneiform nucleus	Cuneiform nucleus
Dentate gyrus	Dentate gyrus

ABA name	mapped BAMS (Swanson 98)
Dentate nucleus	Dentate nucleus
Dorsal column nuclei	Dorsal column nuclei
Dorsal motor nucleus of the vagus nerve	Dorsal motor nucleus of the vagus nerve
Dorsal nucleus raphé	Dorsal nucleus raphe
Dorsal part of the lateral geniculate complex	Dorsal part of the lateral geniculate complex
Dorsal premammillary nucleus	Dorsal premammillary nucleus
Dorsal tegmental nucleus	Dorsal tegmental nucleus
Dorsomedial nucleus of the hypothalamus	Dorsomedial nucleus of the hypothalamus
Edinger-Westphal nucleus	Edinger-Westphal nucleus
Epithalamus	Epithalamus
External cuneate nucleus	External cuneate nucleus
Facial motor nucleus	Facial nucleus
Fastigial nucleus	Fastigial nucleus
Field CA1 pyramidal layer	Field CA1 pyramidal layer
Field CA3 pyramidal layer	Field CA3 pyramidal layer
Fundus of striatum	Fundus of the striatum
Geniculate group_ dorsal thalamus	Geniculate group of the dorsal thalamus
Geniculate group_ ventral thalamus	Geniculate group of the ventral thalamus
Gracile nucleus	Gracile nucleus
Hindbrain	
Hippocampal formation	Hippocampal formation
Hippocampal region	Hippocampal region
Hypoglossal nucleus	Hypoglossal nucleus
Hypothalamic lateral zone	Lateral hypothalamic area
Hypothalamic medial zone	Medial zone of the hypothalamus
Hypothalamus	Hypothalamus
Inferior colliculus	Inferior colliculus
Inferior olivary complex	Inferior olivary complex
Inferior salivatory nucleus	Inferior salivatory nucleus
Interanterodorsal nucleus of the thalamus	Interanterodorsal nucleus of the thalamus
Interanteromedial nucleus of the thalamus	Interanteromedial nucleus of the thalamus
Interbrain	Interbrain
Intergeniculate leaflet of the lateral geniculate complex	Intergeniculate leaflet of the lateral geniculate complex
Intermediodorsal nucleus of the thalamus	Intermediodorsal nucleus of the thalamus
Interpeduncular nucleus	Interpeduncular nucleus
Interposed nucleus	Interposed nucleus
Interstitial nucleus of Cajal	Interstitial nucleus of Cajal
Intralaminar nuclei of the dorsal thalamus	Intralaminar nuclei of the dorsal thalamus
Lateral dorsal nucleus of thalamus	Lateral dorsal nucleus of thalamus
Lateral group of the dorsal thalamus	Lateral group of the dorsal thalamus
Lateral habenula	Lateral habenula
Lateral mammillary nucleus	Lateral mammillary nucleus

ABA name	mapped BAMS (Swanson 98)
Lateral posterior nucleus of the thalamus	Lateral posterior nucleus of the thalamus
Lateral reticular nucleus	Lateral reticular nucleus
Lateral septal complex	Lateral septal complex
Lateral septal nucleus	Lateral septal nucleus
Lateral vestibular nucleus	Lateral vestibular nucleus
Linear nucleus of the medulla	Linear nucleus of the medulla
Locus ceruleus	Locus coeruleus
Magnocellular nucleus	Magnocellular preoptic nucleus
Magnocellular reticular nucleus	Magnocellular reticular nucleus
Main olfactory bulb	Main olfactory bulb
Mammillary body	Mammillary body
Medial amygdalar nucleus	Medial nucleus of the amygdala
Medial geniculate complex	Medial geniculate complex
Medial group of the dorsal thalamus	Medial group of the dorsal thalamus
Medial habenula	Medial habenula
Medial mammillary nucleus	Medial mammillary nucleus
Medial preoptic nucleus	Medial preoptic nucleus
Medial pretectal area	Medial pretectal area
Medial vestibular nucleus	Medial vestibular nucleus
Median preoptic nucleus	Median preoptic nucleus
Mediodorsal nucleus of thalamus	Mediodorsal nucleus of the thalamus
Medulla	
Medulla_ behavioral state related	
Medulla_ motor related	
Medulla_ sensory related	
Midbrain	
Midbrain raphé nuclei	Dorsal nucleus raphe
Midbrain raphé nuclei	Interfascicular nucleus raphe
Midbrain raphé nuclei	Rostral linear nucleus raphe
Midbrain raphé nuclei	Central linear nucleus raphe
Midbrain raphé nuclei	Superior central nucleus raphe
Midbrain reticular nucleus_ magnocellular part_ general	
Midbrain reticular nucleus_ retrorubral area	Retrorubral area
Midbrain trigeminal nucleus	Mesencephalic nucleus of the trigeminal
Midbrain_ behavioral state related	
Midbrain_ motor related	Midbrain-Hindbrain, Motor
Midbrain_ sensory related	Midbrain-Hindbrain, Sensory
Motor nucleus of trigeminal	Motor nucleus of the trigeminal
Nucleus accumbens	Nucleus accumbens
Nucleus ambiguus	Nucleus ambiguus, ventral division
Nucleus ambiguus	Nucleus ambiguus dorsal division
Nucleus incertus	Nucleus incertus

ABA name	mapped BAMS (Swanson 98)
Nucleus of the brachium of the inferior colliculus	Nucleus of the brachium of the inferior colliculus
Nucleus of the lateral lemniscus	Nucleus of the lateral lemniscus
Nucleus of the lateral olfactory tract	Nucleus of the lateral olfactory tract
Nucleus of the optic tract	Nucleus of the optic tract
Nucleus of the posterior commissure	Nucleus of the posterior commissure
Nucleus of the solitary tract	Nucleus of the solitary tract
Nucleus raphé magnus	Nucleus raphe magnus
Nucleus raphé obscurus	Nucleus raphe obscurus
Nucleus raphé pontis	Nucleus raphe pontis
Nucleus sagulum	Nucleus sagulum
Nucleus x	Nucleus x
Nucleus y	Nucleus y
Oculomotor nucleus	Oculomotor nucleus
Olfactory areas	Olfactory areas
Olfactory tubercle	Olfactory tubercle
Olivary pretectal nucleus	Olivary pretectal nucleus
Pallidum	Pallidum
Pallidum_ caudal region	Bed nucleus of the anterior commissure
Pallidum_ caudal region	Bed nuclei of the stria terminalis
Pallidum_ dorsal region	Pallidum dorsal region
Pallidum_ medial region	Triangular nucleus of the septum
Pallidum_ medial region	Medial septal complex
Pallidum_ ventral region	Magnocellular preoptic nucleus
Pallidum_ ventral region	Substantia innominata
Parabigeminal nucleus	Parageminal nucleus
Parabrachial nucleus	Parabrachial nucleus
Paracentral nucleus	Paracentral nucleus of the thalamus
Parafascicular nucleus	Parafascicular nucleus
Paragigantocellular reticular nucleus	Paragigantocellular reticular nucleus
Parapyramidal nucleus	Parapyramidal nucleus
Parasolitary nucleus	Parasolitary nucleus
Parastrial nucleus	Parastrial nucleus
Paraventricular hypothalamic nucleus	Paraventricular nucleus of the hypothalamus
Pedunculopontine nucleus	Pedunculopontine nucleus
Periaqueductal gray	Periaqueductal gray
Peripeduncular nucleus	Peripeduncular nucleus
Perireunensis nucleus	Perireunensis nucleus
Periventricular region	
Periventricular zone	Periventricular zone of the hypothalamus
Piriform area	Piriform area
Piriform-amygdalar area	Piriform-amygdaloid area
Pons	

ABA name	mapped BAMS (Swanson 98)
Pons_ behavioral state related	
Pons_ motor related	
Pons_ sensory related	
Pontine central gray	Pontine central gray
Pontine gray	Pontine gray
Posterior hypothalamic nucleus	Posterior hypothalamic nucleus
Posterodorsal preoptic nucleus	Posterodorsal preoptic nucleus
Postpiriform transition area	Postpiriform transition area
Pretectal region	Pretectal region
Principal sensory nucleus of the trigeminal	Principal sensory nucleus of the trigeminal
Red Nucleus	Red nucleus
Reticular nucleus of the thalamus	Reticular nucleus of the thalamus
Retrohippocampal region	Retrohippocampal region
Septofimbrial nucleus	Septofimbrial nucleus
Spinal nucleus of the trigeminal_ caudal part	Spinal nucleus of the trigeminal caudal part
Spinal nucleus of the trigeminal_ interpolar part	Spinal nucleus of the trigeminal interpolar part
Spinal nucleus of the trigeminal_ oral part	Spinal nucleus of the trigeminal oral part
Striatum	Striatum
Striatum dorsal region	Striatum dorsal region
Striatum ventral region	Striatum ventral region
Striatum-like amygdalar nuclei	Striatum caudal (amygdalar) region
Subceruleus nucleus	Subcoeruleus nucleus
Subiculum	Subiculum
Sublaterodorsal nucleus	Sublaterodorsal nucleus
Substantia innominata	Substantia innominata
Substantia nigra_ compact part	Substantia nigra compact part
Substantia nigra_ reticular part	Substantia nigra reticular part
Subthalamic nucleus	Subthalamic nucleus
Superior central nucleus raphé	Superior central nucleus raphe
Superior colliculus_ motor related	Superior colliculus intermediate deep gray layer
Superior colliculus_ motor related	Superior colliculus intermediate gray layer
Superior colliculus_ motor related	Superior colliculus intermediate white layer
Superior colliculus_ motor related	Superior colliculus intermediate deep white layer
Superior colliculus_ sensory related	Superior colliculus zonal layer
Superior colliculus_ sensory related	Superior colliculus optic layer
Superior colliculus_ sensory related	Superior colliculus superficial gray layer
Superior olivary complex	Superior olivary complex
Suprachiasmatic nucleus	Suprachiasmatic nucleus
Supragenicular nucleus	Supragenicular nucleus
Supragenual nucleus	Supargenual nucleus
Supramammillary nucleus	Supramammillary nucleus
Supratrigeminal nucleus	Supratrigeminal nucleus

ABA name	mapped BAMS (Swanson 98)
Taenia tecta	Taenia tecta
Tegmental reticular nucleus	Tegmental reticular nucleus
Thalamus	Thalamus
Thalamus_ polymodal association cortex related	Anterior group of the dorsal thalamus
Thalamus_ polymodal association cortex related	Epithalamus
Thalamus_ polymodal association cortex related	Geniculate group of the ventral thalamus
Thalamus_ polymodal association cortex related	Intralaminar nuclei of the dorsal thalamus
Thalamus_ polymodal association cortex related	Lateral group of the dorsal thalamus
Thalamus_ polymodal association cortex related	Medial group of the dorsal thalamus
Thalamus_ polymodal association cortex related	Midline group of the dorsal thalamus
Thalamus_ polymodal association cortex related	Reticular nucleus of the thalamus
Thalamus_ sensory-motor cortex related	Geniculate group of the dorsal thalamus
Thalamus_ sensory-motor cortex related	Peripeduncular nucleus
Thalamus_ sensory-motor cortex related	
Thalamus_ sensory-motor cortex related	Subparafascicular nucleus
Thalamus_ sensory-motor cortex related	Ventral group of the dorsal thalamus
Trochlear nucleus	Trochlear nucleus
Tuberal nucleus	Tuberal nucleus
Ventral group of the dorsal thalamus	Ventral group of the dorsal thalamus
Ventral medial nucleus of the thalamus	Ventral medial nucleus of the thalamus
Ventral part of the lateral geniculate complex	Ventral part of the lateral geniculate complex
Ventral posterior complex of the thalamus	Ventral posterior complex of the thalamus
Ventral premammillary nucleus	Ventral premammillary nucleus
Ventral tegmental area	Ventral tegmental area
Ventral tegmental nucleus	Ventral tegmental nucleus
Ventromedial hypothalamic nucleus	Ventromedial nucleus of the hypothalamus
Vestibular nuclei	Vestibular nuclei
Zona incerta	Zona incerta