BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 6. 446–456
Advance Access publication October 10, 2007

doi:10.1093/bib/bbm047

Informatics in neuroscience

Leon French and Paul Pavlidis

Submitted: 9th July 2007; Received (in revised form): 18th September 2007

Abstract

The application of informatics to neuroscience goes far beyond 'traditional' bioinformatics modalities such as DNA sequences. In this review, we describe how informatics is being used to study the nervous system at multiple levels, spanning scales from molecules to behavior. The continuing development of standards for data exchange and interoperability, together with increasing awareness and acceptance of the importance of data sharing, are among the key efforts required to advance the field.

Keywords: *computational neuroscience; databases; data integration; neuroinformatics; neuroscience; semantic web*

INTRODUCTION

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 behavior; Table 1). Where neuroinformatics involves the analysis of genes and proteins, there is extensive overlap with 'traditional' bioinformatics. However, 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.

Our aim in this review is to 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. We briefly cover the large field of nervous system modeling and simulation. In addition, neuroimaging informatics has been reviewed in detail recently [1–3]. Thus our 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.

INITIATIVES

There have been a number of initiatives designed to increase research activity in bioinformatics relating to neuroscience. The first and most prominent is the Human Brain Project (HPB, http://www.nimh.nih.gov/neuroinformatics) started in 1993, based on recommendations developed starting in 1989 [4]. 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 article [5]. The HBP has been succeeded by the NIH Blueprint for Neuroscience research as neuroinformatics is increasingly folded into the mainstream of neuroscience and informatics [6]. The Blueprint for Neuroscience research is a large collaborative NIH effort for creating resources of general utility to neuroscience research [7]. The recently established

Corresponding author. Paul Pavlidis, UBC Bioinformatics Centre (UBiC), 177, Michael Smith Laboratories, 2185 East Mall, University of British Columbia, Vancouver BC V6T1Z4, Canada. Tel: 604 827 4157; Fax: 604 608 2964; E-mail: paul@bioinformatics.ubc.ca

Leon French is a PhD candidate at the University of British Columbia, in the CIHR/MSFHR Strategic Training Program in Bioinformatics. His research interests are neuroanatomy, neuroinformatics, text mining and bioinformatics.

Paul Pavlidis, PhD, is an assistant professor of the Department of Psychiatry at the University of British Columbia. His research is in neuroinformatics and gene expression data analysis.

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 [8]. The INCF was founded in response to a report of the Organization for Economic Co-operation and Development (OECD) [9].

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 [10]. The most visible result is the Neuroscience Database gateway (NDB, http://ndg.sfn.org/). NDB organizes 178 databases into five main categories, and 15 classes. Recently, a consortium-based effort was formed to construct a successor to NDB, the Neuroscience Information Framework (http://neurogateway.org). Currently in exploratory phase, this framework will provide links to databases, findings and tools, organized and annotated using a formal controlled vocabulary, the Brain Markup Language (BrainML, described below).

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) [11]. 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 US and the UK. BIRN is discussed below in the context of several specific areas of neuroinformatics.

Table 1: Data domains in neuroscience

Levels of nervous system organization	Examples of data modalities
Organism	Behavior, physiology
Whole brain	Functional and anatomical imaging, brain region connectivity
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://bioinformatics.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 catalog	http://www.neuroinf.de/
Neuroscience Database Gateway	Resource catalog	http://ndg.sfn.org/
Neuroscience Information Framework	Resource catalog	http://neurogateway.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/

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 [12]. 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 [12–16].

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) [17] 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 [18]. BioPax has been widely adopted, and currently several pathway and interaction databases are available in BioPax format with more converted by third parties [19, 20].

Currently, many neuroscience databases use their own neuron, anatomical region and receptor type vocabularies, but this situation is likely to change rapidly. For example, the BIRN includes an ontology 'taskforce', and is developing BIRNLex for use in BIRN projects [21], 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. It is our hope that BIRNLex (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. There are two established nomenclatures for the rat brain [22, 23], and three for the mouse [23–25]. Thankfully, mappings exist between the terminologies [26, 27]. 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 1900 structures linked to over 7500 terms describing the human, rodent and macaque brain [26]. NeuroNames has been integrated into the Foundational Model of Anatomy, BIRNLex, and the Unified Medical Language System [28]. A web-based interface to NeuroNames is provided by BrainInfo [29]. 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 [14]. 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 is being 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 open biomedical ontologies (OBO) repository [30] reveals 61 classes across eight ontologies. The Gene Ontology also contains many neuroscience concepts such as 'hippocampus development' (biological process), 'GABA receptor activity' (molecular function) and 'axon' (cellular component). Ideally, new terminologies will meld seamlessly as possible with these existing terminologies and avoid reinventing the wheel.

NEUROSCIENCE AND THE SEMANTIC WEB

The notion of a 'semantic web', outlined in a seminal article by Berners-Lee *et al.* [31], is an updated worldwide web in which web-based resources (e.g. web pages) are made more 'computable' by including semantic as well as structural information.

This can be accomplished by extending HTML to include information that describes the data and its relations to other data. In practice, there are several key technologies that are being used to develop resources that are 'semantically enabled', primarily Resource Description Framework (RDF) [32] and the Web Ontology Language (OWL) [33]. We point readers who are unfamiliar with these concepts to the World Wide Web Consortium site [34].

The semantic web has made some of its quickest inroads in neuroscience. This is in contrast to other areas within and outside of science, where the semantic web has been slow to catch on [35]. While these efforts are still in the evaluation and prototype stages, there seems to be a critical mass of interest forming:

- Semantic Web Applications in Neuromedicine (SWAN) presents a framework for scientific discourse based on semantic web technologies. SWAN puts emphasis on community factors of scientific discourse and representing hypotheses. Pilot projects of SWAN are grounded in the Alzheimer Research forum, which includes connections to literature, drugs, antibodies and genes [36]. The knowledge model for SWAN has been formed in OWL.
- The Neurocommons project has the goal of open access and web standards for neuroscience based on semantic web technologies [37]. The project is developing an 'open source knowledge management platform' with early work on mining existing biomedical literature.
- The BIRNLex ontology (mentioned above), which covers many neuroscience concepts, is developed in OWL and strives to follow semantic web best practices and OBO Foundry biomedical ontology development principles [21].
- The SenseLab project (described below) is exploring semantic web technology for increasing neuroscience data interoperability [38].
- The Cell Centered Database group created the Subcellullar Anatomy Ontology (SAO). SAO is a nervous system subcellular anatomy ontology covering 'mesoscale' structures described in OWL (http://ccdb.ucsd.edu/sao.html).
- Recently, the W3C Semantic Web Health Care and Life Sciences Interest Group (HCLSIG) demonstrated integration of 12 databases, including several of those listed above, using semantic web technologies [39].
- Integration of gene expression and neuroimaging data using semantic mapping was recently studied by Pantazatos *et al*. [40]. While this study did not rely on semantic web technologies, it demonstrates how a common ontology can be used to join disparate data sources at a semantic level.

Further adoption of semantic web ideas in neuroscience is still a challenge. Some of the problems are technological (relating to performance, for example) but also reflect the difficulty of formally representing available knowledge. Particular sticking points are how to identify resources (such as genes) in a universally accepted way, and how to represent uncertainty (a fundamental feature of most scientific knowledge). A good overview of some practical challenges is provided by Ruttenberg *et al*. [39].

DATABASES OF MOLECULES AND CELLS

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 [38]. Within the neuronal databases (CellProbDB, 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 [41, 42].

Textpresso for Neuroscience [43] uses a text-mining approach to provide a neuroscience-focused search tool, indexing over 15 000 abstracts and full papers from the biomedical literature. 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 [44]. The developers of NeuroExtract [45] 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 [45]. 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 and then manual screening [46]. 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 18 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 BIB [47]) 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.

CONNECTIVITY

Brain connectivity can be thought of as a property of neurons (cell A connects to cell B) or of anatomical regions (inferior olive projects to the cerebellum). Measuring connectivity has a long history in neuroscience, and efforts to create exhaustive maps and databases are not new [48]. However, due to the difficulty of collecting connectivity data, the only complete nervous system connectivity map is for *Caenorhabditis elegans* [49]. Clearly, having a good-quality map of human brain connectivity would serve as a cornerstone for understanding brain function and structure.

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 [50, 51]. As an example of an ambitious modeling project that will need connectivity information, the Blue Brain project envisions computational modeling of the entire brain [52].

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 [53]. CoCoMac is a searchable database of connectivity data from over 400 literature reports in the Macaque monkey [54]. A related database, CoCoDat, contains detailed microcircuitry reports [55]. Finally, the complete wiring diagram of the *C. elegans* nervous system can be downloaded from http://www.wormatlas.org/ [56].

An interesting experimental project to populate connectivity databases using natural language processing is part of the Neuroscholar project [57, 58]. Neuroscholar is able to classify text with respect to several experimental parameters of interest in tract-tracing studies with 80% precision [57]. Another part of Neuroscholar project, NeuARt II digitizes analog atlases to create a flexible brain-mapping infrastructure [59]. As currently implemented, Neuroscholar is designed to operate with human supervision, as an assist to 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 [38]. BAMS is also involved in integration efforts, and provides links between neuron and cell-associated molecules to brain regions.

COMPUTATIONAL NEUROSCIENCE

We define computational neuroscience as the development and application of computational models of nervous systems or their components; it is the 'systems biology' of neuroscience. This is a very broad area that we cannot hope to do justice here, so we primarily point readers to further resources covering models of single neurons [38], networks [60], and sensory/information processing [61]. More behaviorally focused research includes analysis of working memory [62], visual attention [50], sensorimotor transformations [63] and object recognition [64].

Modeling efforts are often categorized as 'top-down' or as 'bottom-up' [65]. In the bottom-up approach, detailed information about the system (for example, the connectivity diagram of a neuronal network) is used to construct a model. In top-down analysis, the inputs and outputs of the system are considered first (for example, a behavior or neuronal firing pattern) to infer computational strategies that the underlying system might use. The aims in both approaches are to generate testable hypotheses about the function of neuronal systems.

Our ability to model nervous systems is limited by the state of our knowledge of the systems, especially in the bottom-up approach. There are very few neuronal systems that are understood at the level, which is now commonly expected in systems biology analyses of biochemical networks [65], but the increasing trend to take '-omics' approaches to neuroscience is bound to have an impact [66, 67]. In addition, coupling existing large-scale 'omics' data sources and semantically rich but narrower neuroinformatics resources will also enable more detailed computational models of the nervous system. The power of modeling when a system is well understood is demonstrated by classic work on circuit dynamics in crustaceans [68]. A more explicit 'systems biology' approach has been used to search for computational modules in the *C. elegans* wiring diagram [69], and neuronal models are increasingly ambitious [52].

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). There is extensive interest in making imaging data sharable and comparable for the purposes of archiving and meta-analyze, 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 [1–3], so we only give the briefest possible overview of this area.

Several repositories and databases of structural and fMRI images exist, for example fMRIDC [21]. Some systems provide extensive additional analysis tools. The Surface Management System Database [70], Brainmap [71] and the Brede database [72] allow visualization of brain locations and searches based on a reference coordinate system [73]. The Brede database also 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 [73], opening possibilities for meta-analysis and uses text mining to link articles to brain activation studies [72].

A specialized form of MRI, diffusion tensor imaging (DTI), can be used to generate connectivity maps ('tractography') of living human brains [74, 75]. For example, DTI has been used to describe connections between the thalamus and cortex [76]. 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 [77]. Although a few DTI data sets are available online [78, 79], to our knowledge there are no databases of connectivity derived from DTI.

ELECTROPHYSIOLOGY

Electrophysiology refers to the analysis and interpretation of biologically generated electrical signals. Electrophysiological time-series data extends from high sample rate recordings of individual ion channels, through multielectrode and multichannel optical recordings of cellular and circuitry activity. Furthermore, multi-unit non-invasive recordings are preformed on nerve, muscle and whole brain activity. Electrophysiological methods are at the core of much of neuroscience, as electrical signals are the primary mode by which information communication and processing occurs in the nervous system. Unfortunately, electrophysiologic data comprises a wide variety of large and complex data sets, and there is no widely accepted standard way for data to be stored or described. Another problem is that the potential benefits for re-use or sharing of electrophysiological data has not been absorbed by neuroscientists. As in other areas of research where data sharing is or is becoming established, the benefits to the researcher need to be made clear, and the field is in need of good use cases and leadership by example.

Presumably due to these difficulties, there are currently few efforts to standardize and database physiological data. Two developing large-scale projects, CARMEN [80] and Neurobase [81] are potential solutions. The largest structured database we are aware of is Neurodatabase.org, which contains data from eleven experiments [82]. Beyond this some researchers put their data on a personal website for download, as a supplement to published articles (for example, electroencephalography data found at http://sccn.ucsd.edu/~arno/indexeeg.html). The Neurodatabase data sets are well annotated and structured according to the BrainML standard [14]. Demonstrating the advantages of a structured design, Neurodatabase provides cross-references to BrainInfo for brain regions and NeuronDB for neuron types. This provides some glimpses of the potential for how electrophysiological data could be integrated into multi-modal informatics-driven research.

GENETICS AND GENE EXPRESSION

The relationship between genes and behavior (the output of the nervous system) has long been appreciated. However, it is generally much easier to analyze genes than behavior, and the links between the two 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 behavior (and its disorders, such as psychosis) 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/) [83]. GeneNetwork uses RNA profiling data from recombinant inbred mice, which have been extensively phenotyped (behaviorally 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 behavior. For example, Korostynski *et al.* [84] used GeneNetwork to help identify candidate genes for variation in opioid preference between different mouse lines. Similarly, Kempermann *et al.* [85] examined how genetic variation in adult neuronal proliferation in the hippocampus covaries with gene expression Additional applications can be found referenced on the GeneNetwork website. We also note that the same mouse lines that are used in GeneNetwork are being analyzed as part of a BIRN project, adding additional dimensions to the available reference phenotypes [86].

Understanding differences in the genes expressed in different brain regions and neurons have 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).

Arguably the best known of the atlases is the Allen Brain Atlas (ABA), which contains high-resolution colorimetric *in situ* RNA hybridization data for most of the known mouse genes, in the adult brain [87]. The ABA is primarily accessible via a sophisticated web-based graphical interface [88]. The ABA is beginning to introduce search tools that allow searching for genes by similarity of expression patterns (NeuroBLAST), and makes extensive summarized data on expression patterns available for download in XML format. ABA has also contributed a nomenclature for mouse brain anatomy. 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 [89]. Additional information and comparison of these and other atlases are given in Sunkin [90].

RNA expression profiling using microarrays or sequence-based approaches (SAGE) 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 [91]. 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 [92] and ArrayExpress [93], which together contain hundreds of brain-related expression studies. A thorough review of expression studies in the brain [94] identified 448 papers as of June 2004, of which less than one in five had data available online. A more recent review we performed (Wan and Pavlidis, in press) identified about 400 brain-related studies with public data in GEO and ArrayExpress.

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, which is developed in our lab, 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 [95–97]. Integrating this type of analysis with spatially resolved atlases will be an important area of activity [90]. Expression data from microarrays can be compared to *insitu* data such as the ABA, in order to aid interpretation [98].

CONCLUSION

In a recent editorial [10], the president of the Society for Neuroscience, David Van Essen, identified 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?’ [10]. 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 a bioinformatician can already explore a wealth of neuroscience information stored within general and domain-specific bioinformatics resources at multiple scales from molecules to behavior. More scientists are embracing the concepts of open access publishing, open source software and community building via sharing knowledge over the Internet. The power of computers and networks continues to increase unabated. We feel neuroinformatics is reaching a critical mass, and look forward to the next few years of developments as an exciting time.

Key Points

- As more neuroscience-specific data is rapidly becoming available on the Internet, there has been a surge in interest in developing purpose-built resources and tools. Some themes include the use of text mining and an interest in using semantic web technologies to ease interoperability.
- Efforts to harmonize, standardize and integrate neuroscience data are critical, and development of new ontologies and data standards are needed to push the field forward.

Acknowledgements

The authors thank Jesse Gillis for input on computational neuroscience, and Bill Bug for suggestions on the manuscript. L.F. is supported by Natural Sciences and Engineering Research Council of Canada. P.P. is supported by a career award from the Michael Smith Foundation for Health Research and a Human Brain Project grant from the National Institutes of Health (GM076990).

References

1. Nielsen FA, Christensen MS, Madsen KH, *et al.* fMRI neuroinformatics. *IEEE Eng Med Biol Mag* 2006;**25**: 112–9.
2. Brinkley JF, Rosse C. Imaging and the Human Brain Project: a review. *Methods Inf Med* 2002;**41**:245–60.
3. Toga AW. Neuroimage databases: the good, the bad and the ugly. *Nat Rev Neurosci* 2002;**3**:302–9.
4. Shepherd GM, Mirsky JS, Healy MD, *et al.* The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci* 1998;**21**:460–8.
5. Koslow SH. Discovery and integrative neuroscience. *Clin EEG Neurosci* 2005;**36**:55–63.
6. Huerta MF, Liu Y, Glanzman DL. A view of the digital landscape for neuroscience at NIH. *Neuroinformatics* 2006;**4**: 131–8.
7. Baughman RW, Farkas R, Guzman M, *et al.* The National Institutes of Health Blueprint for Neuroscience Research. *J Neurosci* 2006;**26**:10329–31.
8. Amari S, Beltrame F, Bjaalie JG, *et al.* Neuroinformatics: the integration of shared databases and tools towards integrative neuroscience. *J Integr Neurosci* 2002;**1**:117–28.
9. The Global Science Forum Neuroinformatics Working Group. Report on Neuroinformatics, Organisation for Economic Co-operation and Development, 2002.
10. Van Essen D. Neuroinformatics – What's in It for You? *Neurosci Q* 2007;1–5.
11. Martone ME, Gupta A, Ellisman MH. E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nat Neurosci* 2004;**7**:467–72.
12. Insel TR, Volkow ND, Li TK, *et al.* Neuroscience networks: data-sharing in an information age. *PLoS Biol* 2003;**1**:E17.
13. Eckersley P, Egan GF, Amari S, *et al.* Neuroscience data and tool sharing: a legal and policy framework for neuroinformatics. *Neuroinformatics* 2003;**1**: 149–65.
14. Gardner D, Abato M, Knuth KH, *et al.* Dynamic publication model for neurophysiology databases. *Philos Trans R Soc Lond B Biol Sci* 2001;**356**:1229–47.
15. Ascoli GA. The ups and downs of neuroscience shares. *Neuroinformatics* 2006;**4**:213–6.
16. Koslow SH. Should the neuroscience community make a paradigm shift to sharing primary data? *Nat Neurosci* 2000;**3**:863–5.
17. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
18. The BioPAX Working Group. *BioPAX Biological Pathways Exchange Language*. Level 2, Version 1.0 Documentation. http://www.biopax.org (9 July 2007, date last accessed).
19. Baitaluk M, Sedova M, Ray A, *et al.* BiologicalNetworks: visualization and analysis tool for systems biology. *Nucleic Acids Res* 2006;**34**:W466–71.
20. Kotecha N BK, Lu W, Shah N. *Pathway Knowledge Base: Integrating BioPAX Compliant Data Sources, HCLS Workshop, 5th International Semantic Web Conference*, Athens, GA, USA, 2006.
21. Van Horn JD, Grafton ST, Rockmore D, *et al.* Sharing neuroimaging studies of human cognition. *Nat Neurosci* 2004;**7**:473–81.
22. Swanson LW. *Brain Maps: Structure of the Rat Brain*. Amsterdam: Elsevier, 1999.
23. Paxinos GW, C. *The Rat Brain in Stereotaxic Coordinates*. London: Academic Press, 2007.
24. Hof Pry, WG, Bloom FE, Belichenko PV, *et al. Comparative Cytoarchitectonic Atlas of the C57BL/6 and 129/Sv Mouse Brains*. Amsterdam: Elsevier, 2000.
25. Dong HW. *The Allen Atlas: A Digital Brain Atlas of C57BL/6J Male Mouse*. Hoboken, NJ: Wiley, 2007.
26. Bowden DM, Dubach MF. NeuroNames 2002. *Neuroinformatics* 2003;**1**:43–59.
27. Stephan KE, Zilles K, Kotter R. Coordinate-independent mapping of structural and functional data by objective relational transformation (ORT). *Philos Trans R Soc Lond B Biol Sci* 2000;**355**:37–54.
28. Hole WT, Srinivasan S. Adding NeuroNames to the UMLS metathesaurus. *Neuroinformatics* 2003;**1**:61–3.
29. BrainInfo. http://braininfo.rprc.washington.edu/ (9 July 2007, date last accessed).
30. Rubin DL, Lewis SE, Mungall CJ, *et al.* National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *Omics* 2006;**10**:185–98.
31. Berners-Lee T HJ, Lassila O. The semantic web. *Sci. Am* 2001;**May**:34–43.
32. W3C. RDF Primer. http://www.w3.org/TR/rdf-primer/ (9 July 2007, date last accessed).
33. W3C. OWL Web Ontology Language Overview. http://www.w3.org/TR/owl-features/ (9 July 2007, date last accessed).
34. W3C Semantic Web Activity. http://www.w3.org/2001/sw/ (9 July 2007, date last accessed).
35. Good BM, Wilkinson MD. The life sciences semantic web is full of creeps!. *Brief Bioinform* 2006;**7**: 275–86.
36. Lam HY, Marenco L, Clark T, *et al.* AlzPharm: integration of neurodegeneration data using RDF. *BMC Bioinformatics* 2007;**8**(suppl. 3):S4.
37. The Neurocommons. http://sciencecommons.org/projects/data/ (9 July 2007, date last accessed).
38. Crasto CJ, Marenco LN, Liu N, *et al.* SenseLab: new developments in disseminating neuroscience information. *Brief Bioinform* 2007;**8**:150–62.
39. Ruttenberg A, Clark T, Bug W, *et al.* Advancing translational research with the semantic web. *BMC Bioinformatics* 2007;**8**(suppl. 3):S2.
40. Pantazatos SP, Li J, Pavlids P, *et al.* 'Cross-scale mapping of gene expression to neuroimaging datasets via semantic decomposition'. *MedInfo* 2007, p335, Brisbane, Australia.
41. Crasto C, Marenco L, Miller P, *et al.* Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. *Nucleic Acids Res* 2002;**30**:354–60.

42. Crasto CJ, Marenco LN, Migliore M, *et al*. Text mining neuroscience journal articles to populate neuroscience databases. *Neuroinformatics* 2003;**1**:215–37.
43. Textpresso for NeuroScience. http://www.textpresso.org/neuroscience/ (9 July 2007, date last accessed).
44. Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2004;**2**:e309.
45. Crasto CJ, Masiar P, Miller PL. NeuroExtract: facilitating neuroscience-oriented retrieval from broadly-focused bioscience databases using text-based query mediation. *J Am Med Inform Assoc* 2007;**14**:355–60.
46. Zhang W, Zhang Y, Zheng H, *et al*. SynDB: a Synapse protein DataBase based on synapse ontology. *Nucleic Acids Res* 2007;**35**:D737–41.
47. Koehler J. Special issue on text mining. *Brief Bioinform* 2005;**6**:220–312.
48. Sporns O, Tononi G, Kotter R. The human connectome: a structural description of the human brain. *PLoS Comput Biol* 2005;**1**:e42.
49. White JG, Southgate E, Thomson JN, *et al*. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci* 1986;**314**:1–340.
50. Itti L, Koch C. Computational modelling of visual attention. *Nat Rev Neurosci* 2001;**2**:194–203.
51. Serre T, Kreiman G, Kouh M, *et al*. A quantitative theory of immediate visual recognition. Computational Neuroscience: Theoretical Insights into Brain Function. *Prog Brain Res* 2007;**165**.
52. Markram H. The blue brain project. *Nat Rev Neurosci* 2006;**7**:153–60.
53. Bota M, Dong HW, Swanson LW. From gene networks to brain networks. *Nat Neurosci* 2003;**6**:795–9.
54. Kotter R. Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database. *Neuroinformatics* 2004;**2**:127–44.
55. Dyhrfjeld-Johnsen J, Maier J, Schubert D, *et al*. CoCoDat: a database system for organizing and selecting quantitative data on single neurons and neuronal microcircuitry. *J Neurosci Methods* 2005;**141**:291–308.
56. Chen BL, Hall DH, Chklovskii DB. Wiring optimization can relate neuronal structure and function. *Proc Natl Acad Sci USA* 2006;**103**:4723–8.
57. Burns G, Feng D, EHovy E. Intelligent approaches to mining the primary research literature: techniques, systems, and examples. In: Kelemen A, Abraham A, Chen Y (eds). *Computational Intelligence in Bioinformatics*. Germany: Springer-Verlag, 2007.
58. Burns GA, Cheng WC. Tools for knowledge acquisition within the NeuroScholar system and their application to anatomical tract-tracing data. *J Biomed Discov Collab* 2006;**1**:10.
59. Burns GA, Cheng WC, Thompson RH, *et al*. The NeuARt II system: a viewing tool for neuroanatomical data based on published neuroanatomical atlases. *BMC Bioinformatics* 2006;**7**:531.
60. Brette R, Rudolph M, Carnevale T, *et al*. Simulation of networks of spiking neurons: A review of tools and strategies. *J Comput Neurosci* 2007, in press.
61. Destexhe A, Contreras D. Neuronal computations with stochastic network states. *Science* 2006;**314**:85–90.
62. Durstewitz D, Seamans JK, Sejnowski TJ. Neurocomputational models of working memory. *Nat Neurosci* 2000;**3**(Suppl):1184–91.
63. Pouget A, Snyder LH. Computational approaches to sensorimotor transformations. *Nat Neurosci* 2000;**3**(Suppl):1192–8.
64. Riesenhuber M, Poggio T. Models of object recognition. *Nat Neurosci* 2000;**3**(Suppl):1199–204.
65. Bruggeman FJ, Westerhoff HV. The nature of systems biology. *Trends Microbiol* 2007;**15**:45–50.
66. Boguski MS, Jones AR. Neurogenomics: at the intersection of neurobiology and genome sciences. *Nat Neurosci* 2004;**7**:429–33.
67. Choudhary J, Grant SG. Proteomics in postgenomic neuroscience: the end of the beginning. *Nat Neurosci* 2004;**7**:440–5.
68. Marder E, Bucher D. Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. *Annu Rev Physiol* 2007;**69**:291–316.
69. Reigl M, Alon U, Chklovskii DB. Search for computational modules in the C. elegans brain. *BMC Biol* 2004;**2**:25.
70. The Surface Management System Database. http://sumsdb.wustl.edu:8081/sums (9 July 2007, date last accessed).
71. Laird AR, Lancaster JL, Fox PT. BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics* 2005;**3**:65–78.
72. Nielsen FA, Hansen LK, Balslev D. Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics* 2004;**2**:369–80.
73. Nielsen FA, Hansen LK. Finding related functional neuroimaging volumes. *Artif Intell Med* 2004;**30**:141–51.
74. Le Bihan D, Mangin JF, Poupon C, *et al*. Diffusion tensor imaging: concepts and applications. *J Magn Reson Imaging* 2001;**13**:534–46.
75. Parker GJ. Analysis of MR diffusion weighted images. *Br J Radiol* 2004;**77**(Spec No 2):S176–85.
76. Behrens TE, Johansen-Berg H, Woolrich MW, *et al*. Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nat Neurosci* 2003;**6**:750–57.
77. Kubicki M, McCarley R, Westin CF, *et al*. A review of diffusion tensor imaging studies in schizophrenia. *J Psychiatr Res* 2007;**41**:15–30.
78. Evans AC. The NIH MRI study of normal brain development. *Neuroimage* 2006;**30**:184–202.
79. Hermoye L, Saint-Martin C, Cosnard G, *et al*. Pediatric diffusion tensor imaging: normal database and observation of the white matter maturation in early childhood. *Neuroimage* 2006;**29**:493–504.
80. Code Analysis, Repository and Modelling for e-Neuroscience. http://bioinf.ncl.ac.uk/drupal/ (15 August 2007, date last accessed).
81. Christian B. NeuroBase: An Information System for Managing Distributed Knowledge and Data Bases

in Neuroimaging. http://www.irisa.fr/visages/demo/Neurobase/index.html (15 August 2007, date last accessed).

82. Gardner D. Neurodatabase.org: networking the microelectrode. *Nat Neurosci* 2004;**7**:486–7.

83. Wang J, Williams RW, Manly KF. WebQTL: web-based complex trait analysis. *Neuroinformatics* 2003;**1**: 299–308.

84. Korostynski M, Kaminska-Chowaniec D, Piechota M, *et al.* Gene expression profiling in the striatum of inbred mouse strains with distinct opioid-related phenotypes. *BMC Genomics* 2006;**7**:146.

85. Kempermann G, Chesler EJ, Lu L, *et al.* Natural variation and genetic covariance in adult hippocampal neurogenesis. *Proc Natl Acad Sci USA* 2006;**103**:780–5.

86. Mouse BIRN. http://www.nbirn.net/research/testbeds/mouse/ (9 July 2007, date last accessed).

87. Lein ES, Hawrylycz MJ, Ao N, *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 2007;**445**: 168–76.

88. Hochheiser H, Yanowitz J. If I only had a brain: exploring mouse brain images in the Allen Brain Atlas. *Biol Cell* 2007; **99**:403–9.

89. Gong S, Zheng C, Doughty ML, *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 2003;**425**:917–25.

90. Sunkin SM. Towards the integration of spatially and temporally resolved murine gene expression databases. *Trends Genet* 2006;**22**:211–7.

91. Mirnics K, Pevsner J. Progress in the use of microarray technology to study the neurobiology of disease. *Nat Neurosci* 2004;**7**:434–9.

92. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**:207–10.

93. Parkinson H, Kapushesky M, Shojatalab M, *et al.* ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 2007;**35**: D747–50.

94. Aarnio V, Paananen J, Wong G. Analysis of microarray studies performed in the neurosciences. *J Mol Neurosci* 2005; **27**:261–8.

95. Assou S, Le Carrour T, Tondeur S, *et al.* A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells* 2007;**25**:961–73.

96. Rhodes DR, Yu J, Shanker K, *et al.* ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 2004;**6**:1–6.

97. Pan F, Chiu CH, Pulapura S, *et al.* Gene Aging Nexus: a web database and data mining platform for microarray data on aging. *Nucleic Acids Res* 2007;**35**:D756–9.

98. Ponomarev I, Maiya R, Harnett MT, *et al.* Transcriptional signatures of cellular plasticity in mice lacking the alpha 1 subunit of GABA(A) receptors. *J Neurosci* 2006;**26**:5673–83.

99. Wan X, Pavlidis P. Sharing and reusing gene expression profiling data in neuroscience. *Neuroinformatics* 2007, in press. doi: 10.1007/s12021-007-0012-5.