Neuroscience Informatics Platform of Shared Databases and Tools at the Australian National Neuroscience Facility

G.F. Egan[1,2,3], W. Lui[1], E.Tan[1], P-S. Ong[1,4], D. Hang[1]

[1] Howard Florey Institute, [2] Centre for Neuroscience, [3] National Neuroscience Facility, &
[4] School of Electrical Engineering, University of Melbourne, 3010 Australia

g.egan@hfi.unimelb.edu.au

Abstract. The Australian National Neuroscience Facility (NNF) has been established to provide Australian neuroscientists with access to networks of laboratories offering neuroscience consultancy, technical expertise and state-of-the-art equipment. The facility is fostering neuroscience excellence, combining science, technology, innovation, investment, creativity and the opportunity to advance our understanding and treatment of the brain and mind. Within the NNF a Neuroscience Informatics platform has being established with the objective of enhancing both the national neuroscience research capability, as well as the commercialisation opportunities for the Australian health and biotechnology industries. The Platform has developed a NeuroGrid facility consisting of computational resources and Grid middleware, internet accessible neuroimage databases, and standardised neuroimage analysis tools. A customised NeuroGrid portal is currently under development. It is envisaged that the NeuroGrid facility and software tools will provide the basis for application of Grid computing technologies to other areas of neuroscience research.

1. Introduction

Neuroscience is one of the fastest growing areas of scientific research and of industry development in the biotechnology sector. There is growing interest amongst neuroscientists to equitably share neuroscience data and analytical tools since this sharing affords the opportunity to differently re-analyze previously collected data; secondly, it encourages new neuroscience interpretations; and thirdly, it also fosters otherwise uninitiated collaborations. Sharing of neuroscience data and tools is playing an increasingly important role in human brain research and is leading to innovations in neuroscience, informatics and the diagnosis and treatment of neurological and psychiatric brain disorders [1].

Neuroscience Informatics is the use of information technology to acquire, store, organize, analyse, interpret and computationally model neuroscience data. The term is used to describe both neuroscience databases, as well as computer-based tools for using these databases. In the past decade there has been a rapid growth in the volume and complexity of neuroscience data. The limited degree to which scientific publications can describe the richness of the inter-relationships between these complex data has stimulated the development of internet accessible neuroscience databases. Internet access to these databases permits efficient re-analysis of these data as new concepts are developed, as well as the application of new quantitative modelling approaches to enable more precise investigations of the conceptual understanding that has been derived from the initial research studies.

The Australian National Neuroscience Facility (NNF) has been developed with a Major National Research Facility grant from the Australian Government and with additional supporting funds from the Victorian State Government. A key objective of the National Neuroscience Facility is to develop physical infrastructure to place Australia at the forefront of international neuroscience research, and to provide researchers with efficient access to this infrastructure. This objective is being pursued through the establishment of neuroscience technology platforms which aim to provide: (i) national resources for data storage, retrieval and analysis; (ii) scientific training; (iii) a national capability for research, development and commercialisation through local start-up companies and major international commercial alliances; and (iv) expansion of the existing operations and platform technologies from currently separate entities specialising in neuroscience and related fields into co-located facilities. Platforms have been established in the following fields: neurogenomics, neuroproteomics, cellular and electrophysiology, integrative neuroscience, neuroimaging, clinical trials, a tissue repository, and neuroinformatics. The NNF was officially opened in August, 2003.

The Neuroscience Informatics platform has been developed to address the following issues: the difficulties associated with accessing disparate and limited datasets, the lack of standardised analysis tools and experimental methods, and the lack of integrative analyses across sub-disciplines of neuroscience research. The Neuroscience Informatics platform aims at providing the necessary informatics infrastructure and expertise required by neuroscientists associated with the NNF, or who are working within other NNF platforms, particularly in the neuroimaging and neurogenomics platforms. It is envisaged that the Neuroscience Informatics Platform will become an Australian node of international neuroinformatics projects. The platform is also co-operating with a number of commercial organizations that have developed proprietary databases and are integrating behavioural, psychophysiological, neuroimaging and gene expression datasets.

2. Neuroscience Informatics Platform – Application to Neuroimaging

The application of neuroimaging methods to the study of human brain structure and function is continuing at an undiminished pace, particularly in the study of psychiatric disorders and higher cognitive functions. Existing MR image datasets from investigations in these fields contain enormously valuable information. The Neuroscience Informatics platform has focussed on the development of neuroimaging databases and specialized neuroimaging software tools. Existing databases within the NNF include a human neonate structural brain image database containing brain images from over 270 human subjects, and with over 1200 structural images stored [2]. Other databases developed by scientists and clinicians in the NNF include a human functional imaging database (including 180 adult subjects containing over 400 structural and over 20,000 functional MR images), and an adolescent and adult human structural MR image database (containing over 1000 structural images) used to investigate brain morphology in schizophrenic patients and individuals at risk of developing schizophrenia [3].

National and indeed international collaboration has been widely recognized [4,5,6,7] as a key requirement to develop effective informatics facilities that will enable researchers to fully reap the potential benefits of neuroimaging research investigations. In order to maximally capture these benefits, analyses of large imaging data sets need to be supported by modern data mining techniques. However, a number of significant research collaboration challenges must firstly be overcome.

The integration of database systems where the data representations are intrinsically different is problematic. Furthermore, the establishment of a standard human neuroanatomical ontology is required in order to, for example, enable unambiguous indexing of images in databases. This would it turn enable retrieval of a specific neuroanatomical segment (or brain structure) from a human brain image. Finally, important enhancements to neuroimaging analyses could be achieved by linking image databases to related subject information such as gene expression and genetic characterisation (using micro-array and/or serial analysis of gene expression) in the same subject. Nevertheless, in spite of these on-going challenges the recent rapid developments in Grid computing and the current high level of interest in e-science and e-research motivate our application of Grid technologies to neuroscience research.

3. Methods

Efficient analyses of large neuroimaging datasets requires layered information technology (IT) systems capable of data management (image database creation and curation), standardized image data processing algorithms (image registration, tissue classification, segmentation, and spatial normalization to name but a few), and standardized access to computational facilities. The NeuroGrid system aims to utilise the computational power and networking technologies of the Grid to perform standardised analyses of large volumes of neuroimaging data that are stored and organized into image databases within the NNF.

Eventually it is anticipated that data access and processing operations will be managed through seamless Grid computing facilities that are running multiple operating systems including Mac OS-X, Linux/Solaris, and Windows 2000/XP [8,9]. Importantly, the analysed or processed imaging results will be stored into the originating databases, or to alternative results databases. The NeuroGrid system therefore consists of a cluster of Mac OS-X and Linux desktop computing nodes, a Grid management software (middleware) layer, internet accessible neuroimaging data repositories and neuroimaging analysis algorithms, and a customised Grid graphical user interface (GUI) application (the NeuroGrid portal).

NeuroGrid Computing Infrastructure

The NeuroGrid computing facility currently consists of the following computing hardware: an X-serve G5, an X-serve RAID system, an X-serve G5 Cluster Node, and 12 Power Mac G5 desktop machines (Apple Computer Inc, Cuppertino, CA); and 12 PC desktop machines. The X-serve G5 is configured as the X-serve Raid controller and is operating both as the file server and database server. As demand for the database increases it is envisaged that the database server will be deployed on a separate machine.

The X-serve G5 cluster node is the controller for the internal Grid, accepting and distributing jobs to the other nodes. Ten of the Mac G5 machines are configured on a gigabit switch and are situated in a teaching laboratory. These machines are used as desktop machines and also as remote computing nodes. Two of the Mac G5 machines are used for development and testing of the facilities. The desktop PCs are also configured into he internal Grid and are accessible through the SGE. The specifications of the computing facilities are given in Table1.

Computing node	Specifications
X-serve G5	Dual 2GHz, 4GB Ram, 750GB HD, DVD+CDRW, Dual GB Ethernet
X-serve RAID system	3.5TB HD, 512 MB + 512 MB Cache, Battery Backup
X-serve G5 cluster node	Dual 2GHz, 4GB Ram, 80GB HD, Dual GB Ethernet
Power Mac G5 desktop	Dual 1.8GHz, 4GB Ram, 160GB HD, SuperDrive, Geforce FX 5200, GB Ethernet
PC desktop	Dual 2.2GHz, 1GB Ram, 160GB HD, Geforce FX 5200, GB Ethernet

Table 1. Computing specifications of the NeuroGrid facility at the National Neuroscience Facility.

Grid Management Software

The requirements for the selection of a suitable Grid middleware application include: ease of setup, maintainence and operation; availability for a heterogeneous computing environment (Mac OS-X and Linux platforms); ability to monitor the available resources at each computing node; load balancing for efficient process distribution across the computing nodes; parallel computation capability; job scheduling and batch tools; and multi-cluster support for future internet based collaboration.

An X-Grid middleware software application (Apple Computer Inc., X-grid beta release 2003) has been tested using the NeuroGrid facilities. The initial testing showed that the X-Grid application has the following disadvantages compared to other middleware applications including the Sun Grid Engine (SGE), Portable Batch System (PBS) and Load Sharing Facility (LSF) middleware applications: inaccurate or non-existent load balancing across computing nodes; remote applications only executable as user "*nobody*" creating file permission problems; applications not executable with options; and applications executed remotely in the *temp* directory, thus creating difficulties with working directories and write permissions in the *temp* directory.

A detailed investigation of alternative Grid management software included: PBS Pro, a commercially available software which is powerful but costly; LSF, a powerful free application, but hard to configure; and SGE, also a powerful but less complex and free software, but with very little support. After evaluation the SGE middleware was selected and has been implemented as an alternative middleware layer, due to its superior performance in the following criteria: ease of installation and initial setup as well as usability; access to the work load list of each active computing node; capability to receive command line jobs; capability to monitor progress of submitted jobs; and the capability to specify both the input and output data directory. The SGE has performed satisfactorily since being configured for use with the NeuroGrid computing hardware.

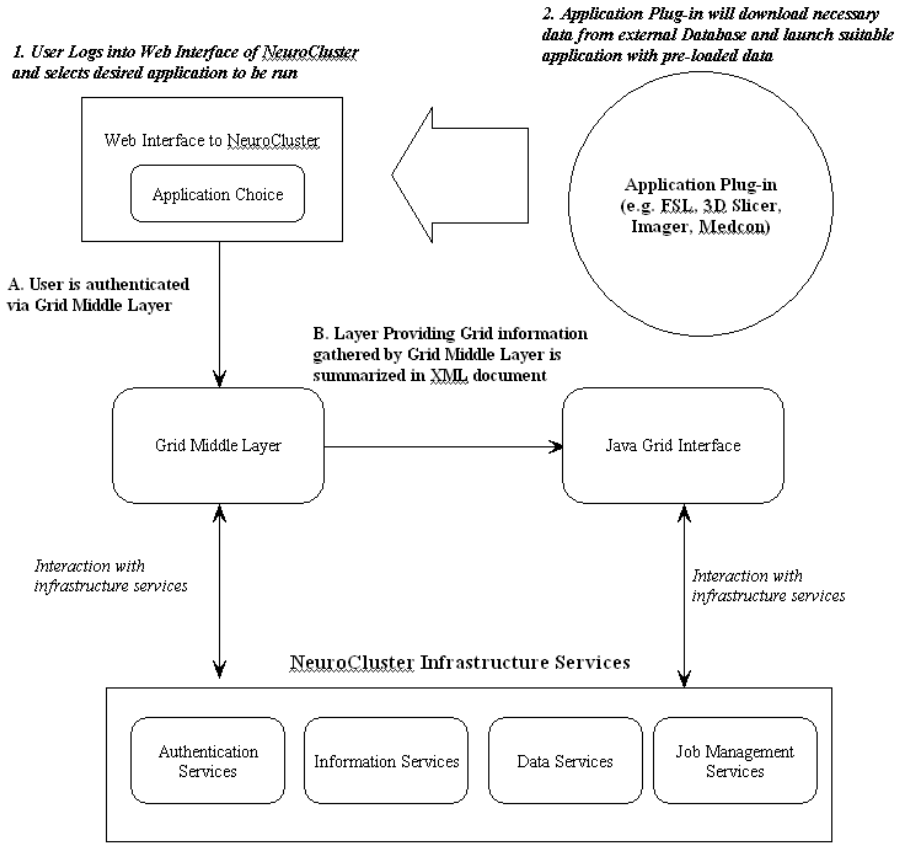

Figure 1. Inter-operability of an internet accessible database system with the NeuroGrid facility. The Grid middleware layer generates information to be packaged as an XML document. The web interface gives the user the functionality to access the system through the internet. It also allows for submission of jobs and accesses to the NeuroCluster Infrastructure Services for job processing. The NeuroGrid database system is not shown in the diagram.

Neuroimage Database

The key capabilities of the NeuroGrid image database include: the storage of primary and secondary data (raw and processed images); the storage of multiple data types (demographic data, MR images, histology images); and the use of standard neuroimage formats (NIfTI [10], Analyze). The key design features of the image databases include: protection of the confidentiality of subject data through the removal of identifying header information; adoption of secure methods for accessing the system; the ability to remotely access the database for data sharing; adoption of open source database tools (PostgreSQL) and applications; and a remote user data entry and upload capability (via php).

Further development of the NeuroGrid database includes the creation of tables and tools for data to be re-entered into the database. Specifically, the second phase of the database includes the storage of multiple instances of an image; that is, storage of modified image results from successive processing operations performed on an input image. These image results, together with the details of the processing algorithms used to produce the results, are linked to the input image. Subsequent database queries of the original image will provide an index to processed results of the image. The third phase of the database involves development of a GUI for the

semi-automated re-entry of processed image results into the database. A data viewer and data curation GUI will require users to verify the veracity of image analysis results before populating the database with the results.

Neuroimage Analysis Tools

A key objective of the design of NeuroGrid has been to rapidly provide users access to existing well established image analysis tools. This objective has been achieved by re-coding a number of the graphical user interface (GUI) based existing FSL image analysis scripts [11]. The tk/tcl GUIs for the brain extraction tool (BET), linear registration tool (FLIRT) and the automated segmentation tool (FAST) have been re-coded as Java applications. A Java application has been developed to link the output of these applications (after user selection of analysis data sources and processing options) to the SGE middleware. Currently work is aimed at re-coding Java applications of the functional easy analysis tool (FEAT) and FSL diffusion analysis tool (FDT) GUIs for integration with the SGE middleware.

The NeuroGrid software suite also includes the following software applications: FSL 3.2, the native implementation of the tk/tcl tools operating in the X-11 windowing environment; Matlab 6.5 (R13) mathematical modelling tools and associated toolboxes; SPM 2, a library of image analysis functions developed by the Functional Imaging Laboratory (FIL, University College, London, UK); Slicer 2.1, a set of image processing tools developed by the MIT Artificial Intelligence Lab, (MIT, Boston); X-Grid Blast (Beta), an implementation of the BLAST software for genomic database searching optimised for use on Apple G5 computing clusters (Apple Computing Inc, Cuppertino, CA); as well as the desktop applications Microsoft Office 2004 and Adobe Creative Suites v1.1.

NeuroGrid Portal

The NeuroGrid user interface, or portal, is being developed to provide a workflow interface for users. The portal will provide direct access to the image database where a user will initially select the data for analysis. A series of processing steps will then be selected from a list of processing algorithms available on the web services directory. This directory will also maintain a database of the machines on the Grid which have executables available for each processing operation. Depending on the operations selected, a list of the available machines is generated for the execution of the job.

The user is also required to select a destination for the output of each step of the job schedule. For intermediate output steps, the user can select to have the job execution halted subject to verification of the intermediate processing result by the user. In these cases the user will typically use an image viewer to view the intermediate result, and choose to continue the job schedule or abort the job. The full NeuroGrid facility will provide an automated means for user to re-enter the processing final results into the originating database, or another customised image database. The selection of the output data destination will require the user to select from a list of possible databases accessible to the portal. The NeuroGrid portal is being developed as a Java application and requires a substantial programming effort.

4. Discussion

Future developments include the parallelizing of codes to gain full advantage of the increased computational power available from the Grid architecture. A number of the FSL processing algorithms are directly parallelizable at the job execution level where multiple datasets are selected with an identical set of processing operations selected for each dataset. These jobs can then be directly shared to multiple machines via the Grid middleware.

A Parcellation toolbox of image analysis algorithms has been developed as in-house cortical parcellation codes, based on the Matlab image processing toolboxes. These tools are being coded into c/c++ algorithms for maximally efficient execution speeds, and will be incorporated into the image analysis tools available via NeuroGrid. The analysis of high resolution images (exceeding 250 MB per image) where cortical parcellation image analysis is possible, will require further automation of the image processing protocol developed using the FSL tools BET, FAST, FLIRT and the Parcellation toolbox.

Other future improvements of the computational speeds include the parallelizing of source code for the analysis algorithms, such as a re-code of FSL using the Mac Vector Engine (a native library for the Apple Mac G5 architecture). This could potentially increase the execution speeds by a factor of 5. Furthermore, recompilations of existing tools using the IBM-compiler is also expected to produce significant gains in execution speeds.

Future database development plans for the NeuroGrid system include access through the Internet via Java applets and a web portal. Users of the system will be able to retrieve data from external databases and select the desired processing application to apply to the images. These data will use the computing resources on the NeuroGrid to perform the computations.

5. Conclusions

The Australian National Neuroscience Facility (NNF) is now providing Australian neuroscientists and commercial organizations with access to networks of laboratories throughout the country. The facility is fostering neuroscience excellence by combining science, technology, innovation, investment, and creativity; and thus jointly providing the opportunity to advance our understanding and treatment of the brain and mind. The Neuroscience Informatics Platform is developing informatics resources for Australian neuroscience researchers, in particular for the neuroimaging research community. A NeuroGrid facility including computational resources and Grid middleware, internet accessible neuroimage databases, and standardised neuroimage analysis tools have been developed. A customised NeuroGrid portal is currently under development. It is envisaged that the facility and software tools will provide the basis for application of Grid computing technologies to other areas of neuroscience research.

Acknowledgements

This research was supported by a Major National Research Facility grant to Neurosciences Victoria, and by the National Neuroscience Facility, Melbourne.

References

1. Amari S-I, et al, "Collaborative Neuroscience: Neuroinformatics for Sharing Data and Tools. A Communiqué from the Neuroinformatics Working Group of the Global Science Forum", J of Integrative Neuroscience, 1 (2002) 117-128.
2. Hunt RW, Warfield SK, Wang H, Keane M, Volpe JJ, Inder TE. Assessment of the impact of the removal of cerebrospinal fluid on cerebral tissue volumes by advanced volumetric 3D-MRI in post-hemorrhagic hydrocephalus in a premature infant. J Neurol Neurosurg Psychiatry 2003 74(5):658-660
3. Pantelis C, Velakoulis D, McGorry PD, Wood SJ, Suckling J, Phillips LJ, Yung AR, Bullmore ET, Brewer W, Soulsby B, Desmond P, McGuire P (2003), "Neuroanatomical abnormalities before and after onset of psychosis: a cross-sectional and longitudinal MRI comparison", Lancet 361:270-1.
4. www.nbirn.net/Publications
5. Mazziotta JC, et al. (1997) Atlases of the human brain. In, Neuroinformatics: An overview of the Human Brain Project, (S.H. Koslow & M.F. Huerta, eds.). Lawrence Erlbaum Associates, Washington.
6. Koslow, SH (2000) Commentary: Should the neuroscience community make a paradigm shift to sharing primary data? Nature Neuroscience 3, 863-865.
7. Toga AW. (2002) Neuroimage databases: the good, the bad and the ugly. Nature Reviews Neuroscience. 3, 302-309.
8. Buyya R & Venugpoal S (2004) The Gridbus toolkit for service oriented grid and utility computing: an overview and status report. Proceedings of the First IEEE International Workshop on Grid Economics and Business Models (GECON 2004, April 23, 2004, Seoul, Korea), 19-36, ISBN 0-7803-8525-X, IEEE Press, New Jersey, USA.
9. Buyya R, Date S, Mizuno-Matsumoto Y, Venugopal S, & Abramson D (2004) Neuroscience instrumentation and distributed analysis of brain activity data: A case for eScience on global Grids, Journal of Concurrency and Computation: Practice and Experience, Wiley Press, USA in press.
10. For information on the NifTI format see http://nifti.nimh.nih.gov
11. For FSL tools see www.fsl.fmrib.oxford.ac.uk