BrainConductor: An open source platform for development of neuroimaging data analysis tools

Han Liu[1*], Haixiao Du[2*], Yu Wang[2],Xiang Liu[2], Huazhong Yang[2]

[1]*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA*
[2]*Department of Electronic Engineering, Tsinghua University, Beijing, China, 100084*
**These authors contributed equally to this work*

BrainConductor project is a project parallel to Bioconductor. Like Bioconductor, BrainConductor is an open platform for sharing neuroimaging data and R-based data analysis software. BrainConductor aims at promoting interdisciplinary collaboration between worldwide neuroscientists and statistical scientists, bringing state-of-the-art statistical methodologies and toolboxes into neuroscience, and enhancing the reproducibility of remarkable research results. We describe details of our motivations, goals, methods, and the difference between BrainConductor and related neuroinformatics projects. Finally, we present some working instances to better explain advantages of BrainConductor.

In recent years, the allure and charm of human brain captivate large amounts of neuroscientists and breed several ambitious collaborative neuroscience projects (e.g., BRAIN initiative project, Human Brain Project) [1]. Non-invasive imaging techniques gain popularity nowadays in human brain studies. These techniques include structural MRI, functional MRI, diffusion tensor imaging (DTI), and Electroencephalography (EEG). They provide an opportunity to look into human brain and recognize the connectivity patterns of the complex brain network, i.e., the human connectome [2,3]. However, the processing and interpretation of imaging data are convoluted and related to multitudinous statistical issues.

Neuroimaging data processing is undergoing two essential revolutions. First, the development of imaging techniques increases both spatial and temporal resolution. Second, The number and size of datasets are growing dramatically in the community, for example, from the International Neuroimaging Data-Sharing Initiative Project (http://fcon_1000.projects.nitrc.org/indi/indi_ack.html). These transformations require not only data processing platforms with high throughput and computational capability but also efficient statistical models for high-dimensional data in the big data era. Therefore, the collaboration between statisticians and neuroscientists is of necessity for the evolution of neuroimaging domain.

The cooperative trends creates new challenges, and the foremost of all is to build a knowl-

edge hub and interdisciplinary community by connecting neuroscientists and statisticians. Neuroscientists obtain access to and make use of state-of-the-art statistical methodologies from the community, and get assistance in improving processing strategies suitable for neuroimaging data. On the other hand, such communication platform lower barriers to entry to research in neuroimaging by integrating different data sources, and simplifying the processes of data acquisition, data transformation, data management, and data sharing.

The requirement inspired one of the most prominent projects, Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) (http://www.nitrc.org/). NITRC is a resourceful repository collecting popular neuroimaging tools and data from various datasets for specific neuroscience tasks. It provides download links for several popular software packages designed for the preprocessing, analysis, and display of neuroimaging data, such as SPM [4], FSL [5], FreeSurfer [6] and AFNI [7]. NITRC also organizes index and introduction for datasets from different data sources, for example, Alzhelmer's Disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu/data-samples/), and 1000 Functional Connectome Project (http://fcon_1000.projects.nitrc.org/).

However, Most of the software packages released on NITRC are independently developed by multiple teams. They are based on different programming models and processing pipelines, and have different styles of interfaces. It is difficult to develop new packages based on existing ones, which gives rise to a large amount of duplicate work to implement similar functions especially for the data preprocessing. NITRC does not provide guidance on effective software development. Besides, investigators always need lots of training sessions to learn how to install and use these tools, let alone frequently change from them to utilize their particular functions. Moreover, neuroimaging data have special formats. Most data sources proffer raw data in the standard industrial DICOM format. The transformation to widely-used neuroimaging data format and the data preprocessing steps increase difficulties and time consumption in accessing and analysing the data especially for statisticians.

In response, we propose the BrainConductor project, an integration of high-quality and easy-access datasets, and user-friendly software for both neuroscientists and statisticians. BrainConductor is a project parallel to Bioconductor. In the past tens of years, Bioconductor [8] makes a great contribution to the progress of the human genome research because of its transparency, pursuit of reproducibility, and efficiency of development. Moreover, the open-source developing architecture of Bioconductor based on R language provides facility to take advantage of the mature statistical packages rather than re-implementing functionality. In the future neuroimaging studies, an open-source platform is of necessity to produce collaborative creation of extensible computational tools and enhance the reproducibility of research results.

Platform Based On R Language. Our ultimate goal for Brainconductor project is to identify biomarkers and help early diagnosis for a variety of neuropsychiatric disorders. More specifically, we are interested in problems related to data management, mining and analysis associated with noninvasive neuroimaging technologies. This orientation requires a programming environment which has good capabilities of improving quality and increasing efficiency for software development, and stimulating interdisciplinary communication and cooperation.

These concerns motivate the selection of R language as the basic programming modal for Brainconductor project. R is a very good prototyping language with good numerical capabilities, so that one can quickly prototype new statistical computational methods. R has a packaging protocol that different software modules can be developed individually and distributed with clear notions. The packaging system of R enables a world-wide collaborative construction of the Comprehensive R Archive Network (CRAN, cran.r-project.org) with a wide range of high-quality and well-documented statistical and visualization software packages. These hundreds of packages are independently developed for specific objectives, but the objective-oriented programming style of R guarantees the robust interoperability of packages for more complicated applications. All of these characteristics of R would decrease development efforts and release time for reliable software for neuroimaging data analysis. Besides, R also provides support for high-performance and parallel computing (ref), and access to different on-line databases and to web resources. Finally, the community of R is consisted of thousands of active users and developers including biologists, mathematicians, and engineers. The community provides a natural bridge to connect the neuroscientists and statisticians. This is much in line with the intention of BrainConductor projects and is the most important motivation of our selecting R.

Hierarchical Data structure. The BrainConductor project began with significant investment in the infrastructure construction for software development by formulating the standard for general data structures of neuroimaging data. A general data structure unifies the interface port and increases the reusability and interoperability of software packages. The code written for the analysis of a dataset can be adapted to another similar dataset since they have the same structure. A researcher doesn't need to modify the code interface or even write code from scratch.

We define an NIdata class based on the NIfTI format in our Brainbase package as the general data structure. NIfTI is a product of the Data Format Working Group (DFWG) from the Neuroimaging Informatics Technology Initiative project. An NIdata contains both the header information about the data acquisition parameters and data part from a NIfTI file. Based on R matrix, data.frame and list structures, NIdata class facilitate programmers and analysts in processing neuroimaging data, and constructing data subsets of interest based on customized masks. The design of NIdata class also ensures the compatibility, because it is straightforward from the NIfTI data for-

mat. Brainbase package provides functions to read in the header information and high-dimensional data array from binary Neuroimaging data files into NIdata class objects. Software developers can implement algorithms operating directly on the NIdata objects and ignore the basic structures. While users can create NIdata instances using Brainbase package and utilize software packages released on BrainConductor. With the assistance of NIdata, users including neuroimaging data analysts are bridged to the developers mostly consisted of the statistical researchers.

However, we found that Neuroimaging data files often comprise a large amount of redundant information. Datasets in different modalities have their specific regions of interest that can be defined by a customized mask. For example, the analysis of functional MRI data mainly focuses on the gray matter, while the processing of diffusion weighted MRI always focuses on the white matter region. Therefore, a hierarchical data structure would help save storage and memory space. Human Connectome Project (HCP) had the same concern and defined the CIFTI file format and grayordinates, a combined cortical surface and subcortical volume coordinate system [9].

To be continued... Other strategies for saving storage resources and how much can be saved.

Introduction of existing R packages in CRAN Medical Imaging task view. The conversion from DICOM to NIfTI is not well implemented. We based on mature software dcm2nii. The conversion results visualization compared to results of oro.nifti or fmri packages.

BrainConductor provides data packages with well-preprocessed neuroimaging data from popular data sources, for example, ... This feature will appeal to more statistical scientists to take brain imaging data analysis as their applications.

A brief Conclusion at last.

Methods

Data structure and standards. For MRI data (e.g. resting-state fMRI, Diffusion Tensor MRI), the most popular data format is DICOM, and Nifti (or ANALYZE). DICOM is the standard industrial format for raw data directly collected from an imaging device. The DICOM format is very broad and very sophisticated. In brief, each .dcm suffix file contains a number of attributes, including not only the image pixel data but also large amounts of meta-data information about the subject, imaging devices and settings during data acquisition.

However, DICOM datasets are redundant, ascribed to the storage of massive numbers of small files, because DICOM stores each image slice as a separate file. ANALYZE and NIfTI-1

formats are more widely employed in the neuroimaging community. An ANALYZE format document is composed of one "hdr" file and one "img" file. The former contains information about the acquisition settings, while the "img" file contain the image data. NIfTI was released as an extension of the ANALYZE format. The NIfTI data format merges the header and image information of ANALYZE document into one file (.nii) and enables extending of the header information. The NIfTI format has alleviated problems with data storage and sharing across diverse centers, and became one of the most popular neuroimaging format recently. More details about NIfTI format can be found on the website http://medical.nema.org/standard.html and http://nifti.nimh.nih.gov/nifti-1/.

BrainConductor proffers functions to access all these data format, and also provides data packages in normal data structures such as data.frame, list and array. The data packages can be easily accessed by R and contains multiple features extracted from various methods (e.g. the correlation matrix, DTI tracking map, graph theoretical metrics). Users organize their experimental results into two basic parts (descriptions and data) and release the packaged data results in BrainConductor repository for the the reproducibility of remarkable work and further study with the released data features.

Using BrainConductor. (put in supplementary). The current release of BrainConductor is a test version 1.0; we require R version to be above 3.1.1. Users of older R versions must update their installation to start with BrainConductor. Download the latest release of R, then download and install basic packages of BrainConductor by starting R and entering the commands

```
> source("http://10.8.7.219/packages/BrainCo/BCoinstall.R")
> BCoInstall()
```

The BCoinstall.R script installs BrainCoSetup package. BCoInstall is a function of BrainCoSetup package to install core packages if called by default arguments. To install specific packages, e.g., "fmri" and "AnalyzeFMRI", call the BCoInstall function with

```
> BCoInstall(c("fmri","AnalyzeFMRI"))
```

BCoInstall acquiescently installs the core packages in the MedicalImaging task view on CRAN. Users can suppress the default installation with

```
> BCoInstall(installmedicalimgTV = FALSE)
```

For details of installing a CRAN task view, please see the help document of R package "ctv".

BCoInstall also updates outdated R packages with a prompt. Users can suppress the prompt easily using the argument ask = FALSE.

In some cases, underlying alterations in the operating system, especially in Linux system, require recompiling all installed packages. Users can start a new R session and enter

```
> source("http://Domain/BCoinstall.R")
> pkgs <- rownames(installed.packages())
> BCoInstall(pkgs, type="source")
```

Users can check packages that are either outdated or too new for their BrainConductor version with

```
> library(BrainCoSetup)
> BCoValid()
```

The output provides possible solutions to identified problems, and the help page ?BCoValid shows detailed arguments and behaviours of the function.

1. Kandel, E. R., Markram, H., Matthews, P. M., Yuste, R. & Koch, C. Neuroscience thinks big (and collaboratively). *Nature Reviews Neuroscience* **14**, 659–664 (2013).

2. Sporns, O., Tononi, G. & Kötter, R. The human connectome: a structural description of the human brain. *PLoS Comput Biol* **1**, e42 (2005).

3. Sporns, O. The human connectome: a complex network. *Annals of the New York Academy of Sciences* **1224**, 109–125 (2011).

4. Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J. & Nichols, T. E. *Statistical parametric mapping: the analysis of functional brain images: the analysis of functional brain images* (Academic press, 2011).

5. Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. Fsl. *Neuroimage* **62**, 782–790 (2012).

6. Fischl, B. Freesurfer. *Neuroimage* **62**, 774–781 (2012).

7. Cox, R. W. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research* **29**, 162–173 (1996).

8. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).

9. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the human connectome project. *Neuroimage* **80**, 105–124 (2013).

Acknowledgements Put acknowledgements here.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to Han Liu, Ph.D., Sherred Hall 224, Princeton University, Princeton, NJ 08544 USA (email: hanliu@princeton.edu), and Yu Wang, Ph.D., Room 4-303, Rohm Building, E.E. Dept., Tsinghua University, Beijing, China 100084 (email: yu-wang@tsinghua.edu.cn)

Figure 1 Each figure legend should begin with a brief title for the whole figure and continue with a short description of each panel and the symbols used. For contributions with methods sections, legends should not contain any details of methods, or exceed 100 words (fewer than 500 words in total for the whole paper). In contributions without methods sections, legends should be fewer than 300 words (800 words or fewer in total for the whole paper).