Brain Research Bulletin 119 (2015) 145–149

Contents lists available at ScienceDirect

Brain Research Bulletin

journal homepage: www.elsevier.com/locate/brainresbull

Research report

Why sharing matters for electrophysiological data analysis

Leslie S. Smith

Computing Science and Mathematics, University of Stirling, Stirling FK9 4LA, UK

ARTICLE INFO

Article history:
Received 13 March 2015
Received in revised form 17 June 2015
Accepted 29 June 2015
Available online 3 July 2015

Keywords:
Electrophysiology
Data sharing
Analysis tool sharing
Repository
Neuroinformatics

ABSTRACT

We present the case for the sharing of electrophysiological datasets and tools for their analysis. Some of the problems, both sociological and technical, associated with improving the sharing of data and analysis tools are discussed. The work that has been done to try to improve data and code sharing in the electrophysiology area is reviewed. The sharing aspects of the current large projects in brain research are considered.

1. Background

Data sharing is a major issue in many areas of science, and has been for many years. Arzberger et al. (2004) from the Organisation for Economic Cooperation and Development (OECD) state that "publicly funded research data should be openly available to the maximum extent possible". They go on to discuss the issues and mechanisms for achieving this, and their work was a major factor in the founding of the International Neuroinformatics Co-ordinating Facility (INCF), whose aims include "to foster scientific interaction through information flow within our global network" (http://www.incf.org/about).

Sharing data raises different issues in different sciences. Unlike some sciences (e.g. astronomy, genomics) where the underlying data is the same in each experiment, in neuroscience, the underlying data from each experiment will be different. This is not a matter of experimental noise obscuring measurements, but of real variation in the underlying data because repeating experiments precisely is impossible. Data sharing for electrophysiologists (and indeed, for other areas of neuroscience) is needed in order to enable studies across the datasets gathered over multiple experiments by many groups, as well as to allow cross validation of their results. Analysis code sharing (particularly open source code sharing) allows others to check the validity of analyses, as well as enabling them to apply identical analyses to multiple datasets, which is critical for cross validation.

Electrophysiologists work in a variety of environments, from University research, to research, development and testing in companies, to clinical applications. Their work will have different aims, as well as different rules covering sharing of data and of analysis code. There are many different types of systems used for electrophysiology, ranging from those that record from single electrodes either intracellularly or extracellularly, to systems that record from multiple electrodes (generally extracellularly), to EEG systems. These systems sometimes are used to record (relatively) slowly varying potentials (local field potentials), and sometimes they are used to seek out action potentials from neurons near to the electrodes. These are, and are likely to remain, areas of rapid technological change as new measurement and analysis techniques are developed. Sometimes the equipment that is used for this work is built by the researchers themselves, but mostly it is manufactured and sold by a number of companies.

Generators and users of electrophysiology data largely work independently or in small groups, sometimes collaborating with colleagues who are geographically remote. Equipment manufacturers generally provide rather more than the recording equipment: they provide complete suites of software tools that can be used for particular types of analyses of the datasets. As the equipment becomes more complex, and as higher and higher density electrode arrays are created, the data sizes get larger and larger, and there is an ever greater need for appropriate analysis tools. Some of the lower-level software tools (basic recording, simple filtering, for example) used are common across domains, but others, particularly at higher levels, very much reflect the application for which the data is being collected. Clinicians might be interested in activity levels, perhaps seeking foci of ictal episodes, or pre-ictal states;

E-mail address: l.s.smith@cs.stir.ac.uk

http://dx.doi.org/10.1016/j.brainresbull.2015.06.009
0361-9230/

researchers might be interested in patterns of activity elicited by particular stimulation to a lab animal; neuropharmacologists might be interested in the modulation of activity in a brain slice after application of some specific agonist. In short, there is a large variety of types of data, and data analysis tools that are often application specific. Often the users of the analysis tools are not able to check the precise functionality of these tools, either because the source code is not available to them, or because they are not programmers.

So why share data? Why share analysis tools? Why does it matter?

2. Scientific and clinical reasons

Scientific experiments are supposed to be repeatable. Clinical decisions are supposed to be taken on the grounds of evidence. Scientists and clinicians are supposed to be able to defend their results and decisions, by appealing to established evidence. If the data and the techniques for analysing the data are not shared, then it becomes difficult to show that the science underlying the evidence is good, or the clinical decisions appropriate. This is true for any experimental science, or any science based decision-making. Neurophysiological datasets and their analysis tools are no exception to this.

There is a huge and active literature on the analysis tools, whether for electrophysiology or other forms of neuroscience dataset (for example, Lewicki (1998), Harris et al. (2000), Hulata et al. (2002), Takahashi et al. (2003), Pouzat et al. (2004), Mizuseki et al. (2014), Rossant et al. (2015), to name but a few purely in the area of spike sorting). However, simply noting that a particular type of analysis (based on some equations in a paper, for example) has been carried out is not the same as sharing the analysis tools. Virtually all analysis tools have a number of parameters that are critical to their application, and appropriate values for these parameters need to be known in order to replicate a piece of analysis. In addition, as analysis tools become more sophisticated, it becomes very easy indeed to implement something close to, but not quite exactly the same as the analysis technique intended. This becomes important when one is comparing the results of analysis across different datasets: slight differences in algorithm can give rise to considerable differences in results, for example, in the thresholds used in threshold setting in spike detection Mtetwa and Smith (2006), or in the way in which one interacts with a spike sorter such as spikeclus (see http://www.vis.caltech.edu/ rodri/Wave-clus/Wave_clus_home.htm). If the actual tools are shared, for example using one of the many internet open source facilities (e.g. repositories for open-source code such as github, sourceforge) then one can (i) use exactly the same tool and (ii) check that the tool being used does exactly what you believe it to do. One can try out different tools which are intended to provide the same results, and see whether they really do give the same results. In addition, there may be issues of *software rot* due to languages and packages not being supported any more on current systems. Topalidou et al. (2015) provide a good example.

The major advantage of data sharing is that others can work with more data than would otherwise be available to them from their own laboratory and collaborators. Different tools implementing different algorithms (or even different tools implementing what is allegedly the same algorithm, perhaps even based on the same equations) can be used on the datasets, so that the results of earlier analysis can be confirmed. Given appropriate metadata (which often entails contacting the originators of the dataset or of the code) novel cross-analyses can be carried out (see below, and (Eglen et al., 2014)). Further, there is a real need for different types of dataset and a variety of analysis tools to be made available for educational and training purposes, so that analysts-in-training can have examples of real datasets and analysis tools to work with.

3. Data and tool-sharing technologies

Data sharing in electrophysiology is in its infancy. There are many reasons for this, some technical, and some sociological, discussed in Section 5. Yet at the same time, there is a growing push from both funding agencies and journals for data and analysis tools to be public, so that experimental results may be validated.

The Neuroscience Information Framework (https://www.neuinfo.org) provides links into a very large amount of material on the web that relates to neuroscience. Its searching mechanism is sophisticated, and ontology based, making it a very effective start point. Of course, quite a lot of data may be held behind a portal (as is the case with CARMEN http://www.carmen.org.uk: see Section 6 for more details), but at the very least, this system will find the portal. The International Neuroinformatics Co-ordinating Forum (INCF) keeps a list of re-usable Neuroscience (including Neurophysiology) resources at http://incf.org/resources/research-tools. There is a curated set of over 700 processing resources at the NITRC website (http://www.nitrc.org), with quite a wide variety of tools for MR imaging, EEG, ECoG, MEG, and others. In addition, the INCF's Electrophysiology Task Force maintains a list of resources for data sharing in (Neuroimaging and) Electrophysiology at http://tinyurl.com/d7f35qb. The Open Source Brain group (http://opensourcebrain.org) are creating repositories primarily of models and technology in the Neuroscience modelling area.

There are a small number of electrophysiological datasets directly available over the internet. About 30 datasets (as at June 3 2015) are available directly from http://crcns.org, a website jointly funded by the US NSF and NIH. These are mostly datasets that have resulted in specific papers, and are aimed particularly at the Computational Neuroscience community. Because these datasets generally have had full papers written about them, they are normally well described, and include sufficient metadata to make them reasonably easily re-usable. They have been used to test out specific tools, and have been important in quite a large number of papers (see http://crcns.org/publications); they are also useful for educational purposes. The UK's CARMEN project contains a variety of electrophysiological recordings. This was designed as a portal based system that users could share data (and make it public as well), as well as running services and workflows. It is discussed further in Section 6. The Japanese Brainliner project (http://brainliner.jp) provides public access to some EEG, EMG, and ECoG datasets, with appropriate metadata. In addition it also provides public domain MATLAB software for reading a variety of data formats based on the Neuroshare portal. The German Neuroinformatics Node (http://g-node.org) has a small amount of publicly available data, but rather more in the way of public-domain software, including a REST interface to a portal based system, and a small variety of other software, including their neuroscience interchange format, nix. EEGbase (https://eegdatabase.kiv.zcu.cz) contains a variety of EEG and ERP datasets, along with some behavioural stimulation files. The UCSD website http://headit-beta.uscd.edu/studies provides access to a number of well-documented EEG studies. neurodatabase.org allegedly stores some raw datasets, but it appears to be currently inactive. The NeuroElectro database (http://neuroelectro.org) is different, in that it is a curated database that stores 23 electrophysiological properties of 233 different types of neurons (as at June 3 2015), but not the raw data from which these were calculated.

There are undoubtedly other resources that contain useful data, metadata, and analysis tools.