Neural Networks 21 (2008) 1070–1075

Contents lists available at ScienceDirect

Neural Networks

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

2008 Special Issue

G-Node: An integrated tool-sharing platform to support cellular and systems neurophysiology in the age of global neuroinformatics

Andreas V.M. Herz [a,b,c,*], Ralph Meier [d], Martin P. Nawrot [e,c], Willi Schiegel [c], Tiziano Zito [c]

[a] Department of Biology, Ludwig-Maximilians-Universität, Munich, Germany
[b] Bernstein Center for Computational Neuroscience, Munich, Germany
[c] Bernstein Center for Computational Neuroscience, Berlin, Germany
[d] Bernstein Center for Computational Neuroscience, Freiburg, Germany
[e] Institute of Biology - Neurobiology, Freie Universität, Berlin, Germany

ARTICLE INFO

Article history:
Received 13 December 2007
Received in revised form
30 April 2008
Accepted 29 May 2008

Keywords:
Neuroinformatics
Neurophysiology
Computational neuroscience
Data access
Data storage
Data analysis
Data sharing
Open source
Toolbox
Incf
International neuroinformatics coordination facility
National node

ABSTRACT

The global scale of neuroinformatics offers unprecedented opportunities for scientific collaborations between and among experimental and theoretical neuroscientists. To fully harvest these possibilities, a set of coordinated activities is required that will improve three key ingredients of neuroscientific research: data access, data storage, and data analysis, together with supporting activities for teaching and training. Focusing on the development of tools aiming at neurophysiological data, the newly established German Neuroinformatics Node (G-Node) aims at addressing these aspects as part of the International Neuroinformatics Coordination Facility (INCF). Based on its technical and scientific scope, the Node could play a substantial role for cellular and systems neurophysiology as well as for the neuroscience community at large.

1. Motivation

The human brain is one of the most complex biological systems. It contains more than 10^{11} nerve cells and 10^{15} synaptic connections, and its functional elements extend over more then ten orders of magnitude in space — from molecular pathways in individual synapses and neurons to the entire brain. Similarly astounding is the fact that dynamical processes that last for less than a millisecond underlie life-long memories — a span of more than ten orders of magnitude in time. Moreover, the brain utilizes multiple nonlinearities and nested feedback loops that severely limit any purely intuitive approach. Analyzing the dynamics and function of the human brain therefore continues to be a formidable challenge.

In the last decades, novel experimental methods such as patch-clamp recording and imaging techniques have changed basic as well as clinical neuroscience in a most dramatic way. Suddenly, single ion channels could be studied and functional magnetic resonance imaging is nowadays used in a routine manner. It is most likely that all these developments do not mark the end of a success story but rather represent some intermediate step. In addition, as shown by the recent discovery of adult neurogenesis (Eriksson et al., 1998; Gould et al., 1999), neuroscience's firm dogmas of today may already be overthrown tomorrow.

Despite the great success of neuroscience over the last decades, there is thus no doubt that we are far from understanding the human brain. This concerns all levels of neural processes — from the regulation of molecular pathways, the dynamics of single synapses, and the information processing of small neural networks to the orchestrated function of the entire brain. As new studies are initiated on the basis of current interpretations of available data, long-term progress in the neurosciences will crucially depend on the broad availability of high-quality data

* Corresponding author at: Department of Biology, Ludwig-Maximilians-Universität Munich, Germany.
E-mail address: herz@bio.lmu.de (A.V.M. Herz).

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.05.011

(including pre-processed data as well as the underlying raw data) and high-quality data-analysis tools (e.g. Teeters, Harris, Millman, Olshausen, and Sommer (2008)). However, many of today's commercial recording tools are based on highly individual and proprietary data formats and come only with limited and typically closed-source software tools for data mining and analysis. This shortcoming severely complicates the access, storage, analysis and sharing of neuroscientific data and thus slows down the future development of brain research.

As a central element of the German Neuroinformatics Node (G-Node, www.neuroinf.de) within the International Neuroinformatics Coordination Facility (INCF) a novel software and hardware infrastructure will therefore be developed that eases the acquisition, storage and further processing of experimental data in the spirit of the overall INCF goals (e.g. Bjaalie and Grillner (2007) and Horn and Pelt (2008)). We will focus on cellular and systems neurophysiology for a number of reasons. First, the lack of common data standards is rather severe in this field; as a consequence, successful standardization could have an enormous impact. Second, the complexity of an integrated and internationally coordinated approach to advanced data acquisition, storage and analysis goes beyond the capability of a single lab — one of the reasons for the slow progress seen so far — but can be tackled through a concerted effort. Third, without a thorough quantitative understanding of cellular and systems neurophysiology, there is no solid foundation for computational neuroscience and brain theory. In addition, the methodology and tools developed within the project could later also be used in other neuroscience areas. Especially early in the project, however, it will be helpful to focus on a specific task and research community to quickly reach critical mass.

The importance of high-quality physiological data and improved data-analysis tools for computational neuroscience is nicely illustrated by the task of deriving reliable single-neuron model parameters (see also Fig. 1). The standard strategy to deal with the key problem of cell-to-cell variability, i.e., population averaging, can be severely misleading because the dynamical characteristics of single-cell models are, in general, not a monotone function of their parameters. As a consequence, the mean behaviour within a class of models may strongly differ from that of a model with mean parameter values (Golowasch, Goldman, Abbott, & Marder, 2002) and nearly identical dynamical characteristics may be implemented by rather different parameter combinations (Goldman, Golowash, Marder, & Abbott, 2001). In addition, with increasing model complexity, the number of parameters to be estimated increases dramatically so that they have to be taken from different cells or even different preparations, further lowering the model's trustworthiness. Last but not least, models are often calibrated using *in vitro* data, yet designed to capture the neural dynamics and computations of behaving animals. All these examples underscore the need for widely available high-quality experimental data as well as improved data-analysis tools — the latter being a key goal of the G-Node project.

By directly addressing this need of a broad range of experimental and theoretical neuroscientists, the G-Node will thus support ongoing and future experiments in cellular and systems neurophysiology, encourage the standardization of data formats as well as analysis tools and thus directly facilitate the cooperation within and between different labs (Baxter, Day, Fetrow, & Reisinger, 2006). It is our hope that the G-Node, including its services as a general neuroscience community site and its data archive functionality, will also attract a large group of neuroscientists that do not yet have close links to computational approaches or neuroinformatics. To foster international cooperation, the project will be carried out in close cooperation with the INCF secretariat and interested INCF nodes. To achieve its various goals, a larger group of neurobiologists and computer scientists will work on the G-Node project. At its host site (Ludwig-Maximilians-Universität Munich), there will be a staff of five full-time scientists/software developers. In addition, a group of five to ten collaborators from Germany and beyond will join the G-Node based on part-time contracts. Finally, there are also funds to support younger scientists that want to prepare their data and tools so that they can be integrated into the G-Node. Following the open-source concept of the forerunner Neuroinformatics Portal all tools developed within the G-Node will be made freely available to the national and international community. Together, these measures will incorporate the technological opportunities of global neuroinformatics into everyday neurophysiological routine and thus help establish a new scientific culture.

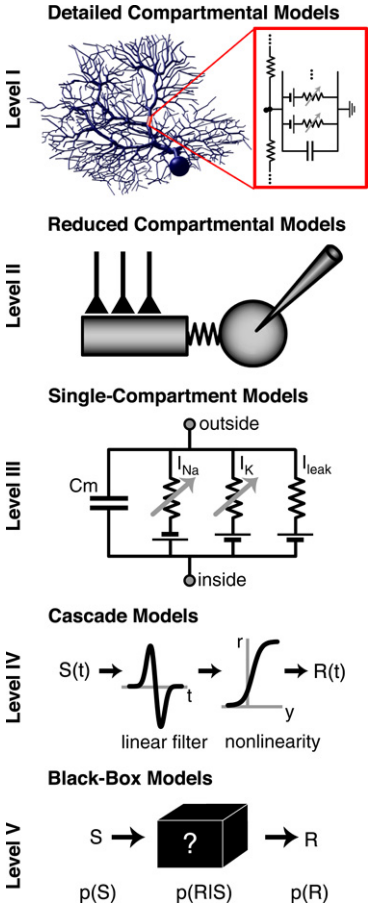

Fig. 1. Computational models of single-neuron dynamics and computations — a core component of modern cellular and systems neurophysiology — require high-quality physiological data on every level of abstraction, as demonstrated by five representative levels of single-cell modeling. **I**: Detailed compartmental model of a Purkinje cell. The dendritic tree is segmented into many electrically coupled Hodgkin–Huxley-type compartments (Level III). **II**: Two-compartment model. The dendrite receives synaptic inputs and is coupled to the soma where the neuron's response is generated. **III**: Hodgkin–Huxley model, the prototype of single-compartment models. The cell's in- and outside are separated by a capacitance C_m and ionic conductances in series with batteries describing ionic reversal potentials. Sodium and potassium conductances (g_{Na}, g_K) depend on voltage; the leak g_{leak} is fixed. **IV**: Linear–nonlinear cascade. Stimuli $S(t)$ are convolved with a filter and then fed through a nonlinearity to generate responses $R(t)$, typically time-dependent firing rates. **V**: Black-box model. Neglecting biophysical mechanisms, conditional probabilities $p(R|S)$ describe responses R for given stimuli S. With increasing model complexity (from bottom to top) more and more data are required to constrain the rapidly growing number of model parameters — many thousands for neurons with complex dendritic trees; but even on the most abstract level of black-box models, faithful experimental estimates for the relevant probability distributions require advanced experimental paradigms and data analysis. This is a huge challenge for cellular and systems neurophysiology and a major motivation for our own initiative. Global neuroinformatics offers the tools to address this challenge — within the INCF, the G-Node project aims at building an integrated tool-sharing platform to support ongoing and future international efforts (Figure redrawn from Herz et al. (2006)).

In the following sections, the four components of the project are summarized. Each component will be motivated by a short "Problem" discussion, followed by a description of the proposed "Solution and Specific Measures". As there are substantial overlaps between the four project components, the solutions will overlap as well.

2. Data access

2.1. Problem

There exists a plethora of physiological recording devices, many of which come with proprietary data formats. This makes it difficult and sometimes even impossible to share data with colleagues. In addition, the lack of standard data formats complicates the analysis of raw data by a person different from the one who originally recorded the data. Collaborations between different experimental labs as well as cooperation between theoreticians and experimentalists, which have been one key ingredient for the advancement of the neurosciences in recent years, are severely inhibited.

After recording, physiological data can often only be analyzed with the software shipped with the specific recording hardware. Such programs offer limited functionality. To carry on with advanced data analysis, or to test new analysis methods, the experimenter or data analyst is forced to bypass the software constraints through hand-made jugglery. This results in undesired fragility and unnecessary complexity of the experimental workflow, a day-to-day experience in many labs.

2.2. Solution and specific measures

- The G-Node will provide an open-source tool to allow direct access to a broad range of proprietary data formats. The tool will be initially based on NeuroShare (see http://neuroshare.sourceforge.net/) and its extensions within the FIND Matlab toolbox (Meier, Egert, Aertsen, & Nawrot, 2008). Note, however, that the scope of the NeuroShare API is not sufficient in this respect because (a) it is platform specific — Windows only — and based on a proprietary format; (b) it describes how data are to be retrieved and delivered, but does not foster any standard regarding how data are stored; (c) it does not offer a reference implementation; (d) the libraries offered by hardware manufacturers are closed-source. In order to use the libraries, users need to buy the corresponding hardware and/or commercial software. Bug reports and feature requests are usually not possible. The FIND toolbox inherits these NeuroShare-specific shortcomings with respect to its data import and requires significant expenses for the commercial Matlab software licenses.
- The G-Node will offer a platform-independent, open-source "data import" tool that will be made available in common programming languages.
- To assure long-term usability and distribution of the data, the establishment of a widely used unified data format is necessary. A unified data format has proved to be the crucial means towards standardization for many research areas (see for example the European Data Format EDF+ for EEG data analysis, Kemp & Olivan, 2003, or also http://www.edfplus.info/). Various efforts have been devoted recently in this direction for neural data (see for example the BrainML initiative by the NIH, Gardner et al. (2003), or http://brainml.org/), but no drop-in solution has been produced. Within the German Neuroinformatics Node, we intend to design such a format in collaboration with experimentalists from participating labs and in close interaction with the INCF, to tailor the format to the community needs.
- The user-oriented development and readily available implementation of the Node project will assure a widespread use of the unified data format, which will in turn provide an incentive for manufacturers of recording hardware and software to endorse this format as a standard (on a smaller scale, this happens already with NeuroShare). Note that the planned unified data format is not incompatible with NeuroShare, but complements it by specifying in detail how data should be stored, and thus makes it possible to work with existing hardware and software without inhibiting the use of newly developed tools.
- Unified export functions will be provided to basic formats (CSV, XLS, XML, binary, Matlab arrays, etc.), to databases (SQL, Oracle, ROOT, HDF5, etc.), and to the unified data format developed at the G-Node. The database back-end can be provided by the individual user. The G-Node will also offer standard database templates and "HowTo"s to allow maximal flexibility for advanced users and maximal ease of use for the novice. The design of the database template will be realized in strict collaboration with the participating labs.
- The G-Node will develop and offer an electronic lab book template (see, for example, the ELB article at Wikipedia, http://de.wikipedia.org/wiki/Laborjournal/ or a similar project at http://NeuroScholar.org). The template will be designed such that it is easily customized according to local and experiment-specific needs. It will be possible to automatically integrate the electronic lab book in the database and to store raw data together with metadata. The lab book will also feature an optional security layer, so that it can act as a full replacement or at least a helpful complement for the "paper" lab book required by law. Filling in a paper lab book tends to be cumbersome. The electronic lab book could help in this respect, for example by offering precompiled fields for static data (e.g. author, date, room temperature, etc) and, when integrated with the recording hardware, even for setup-specific variables (pressure, voltages, amplification levels, etc). The electronic lab book will be searchable and will maintain links to the experimental raw data, even when those are stored remotely. An important advantage of a standardized lab book will be that it facilitates efficient information exchange, both internally and among different labs.
- All software tools will remain open source and freely available.

3. Data storage

3.1. Problem

Modern recording techniques generate rapidly growing amounts of data. In many labs, the problem of archiving and long-term storage is tackled locally with suboptimal means such as DVDs or external hard disks. As a result, old data are often practically unavailable for further analysis and are not secured against hardware failures. Not only is this in stark contrast with basic rules of good scientific practice, see, for example, the suggestions by the German Science Association, which enforces the storage of scientific data used for publications and for projects funded by DFG for ten years (http://www.dfg.de/antragstellung/gwp/index.html/); sloppy data storage also hinders the spreading of scientific knowledge and damages the public reputation of neuroscience.

3.2. Solution and specific measures

Funded through a grant from the German Federal Ministry of Education and Research, the G-Node will offer data storage and management services, initially restricted to projects and international collaborations that involve German institutions. Support will be both remote and local, and include hardware consulting, design and realizations, as well as maintenance and

backup services. Upon request, the technological concepts and software solutions will be made available on an international level.

• Remotely, the G-Node will offer a central Data Storage System, which will handle large volumes of data and will be designed such that it can also meet increasing future demands. The data upload and download for the end user will be transparent, i.e., as easy as moving files to and from an internal drive. Data can be searched and (optionally) integrated into the database developed in the "Data Access" part of the Node project. Data will be secured against hardware failures and stored for a time period chosen by the user — the storage of lab books will be indefinite. The access to the stored data will be granted by an authorization module fully configurable by the user. If different labs want to share data, all they have to do is to set up permissions accordingly. The end user does not need to perform any backups or to care about the scalability of local storage devices.

• Connected with the Data-Storage System, a data mining service will be offered for groups with limited local computation capabilities. Results can be offered for further analysis in different formats depending on the user needs.

• At a later stage of the project, services could also be offered on a pay basis to a wider group of users. This could be an interesting option not only for individual researchers or labs but also for publishers, e.g., of conference proceedings, who do not have the facilities to store and make available supplementary material.

4. Data analysis

4.1. Problem

The complexity of modern physiological data often requires a careful data analysis that goes far beyond standard statistical tests (see, for example, Brown, Kass, and Mitra (2004), Grün (submitted for publication) and Kass, Ventura, and Brown (2005)). The underlying techniques are, however, not routinely available to mainstream neuroscientists. For example, a thorough understanding of the responses of a neuron to dynamic stimuli may only be possible using a theoretical model described by a set of nonlinear differential equations. At this level of sophistication, experimentalists may lack the appropriate mathematical and computational background (and often also the time) to carry out a model-based analysis of their data — and resort to traditional methods. These methods, however, do not allow the researchers to make best use of their data. As a consequence, data are often only used to provide certain evidence for some model whereas, with proper statistical scrutiny, they could have been used to refute an alternative model and thus be far more valuable scientifically. Similarly, theoreticians frequently propose new models and analysis algorithms without having tested them in depth with real data. Both aspects lead to an almost grotesque situation: some complex but potentially highly valuable data are not studied with state-of-the-art analysis tools, and some potentially excellent tools are not brought to the attention of a general neuroscience audience but only published in theory-oriented journals. On both the experimental as well as the theoretical side, substantial resources are thus wasted.

4.2. Solution and specific measures

• The G-Node will offer a public and well-annotated code repository for the analysis and modeling of neurophysiological data. The repository will encourage theoreticians to submit their model implementations and experimentalists to test novel analysis methods. To ease broad access, no specific programming language will be requested for code contributions. To assure the quality of available material, standard practices of open-source development will be enforced, such as (a) quality management and transparent code review: a researcher who submits code suggests an external reviewer, and both are listed to demonstrate their responsibility; (b) proper credit assignment: every contributor of a tool offered by the Node will be explicitly mentioned, together with relevant original publications of the respective method. In addition, contributors will be encouraged through the Node to publish their tools in method-oriented journals. This will underscore the usefulness of the tools and further increase the standing and publicity of the project; (c) extensive documentation and official guidelines; (d) documented examples and test data.

• To make the algorithms and their usage as attractive and transparent as possible for the end user, this repository will be supplemented by a web portal that contains discussion forums, "Wiki"s, mailing lists, tutorials, "HowTo"s, FAQ, etc.

• Within the G-Node, a toolbox will be designed and implemented that offers a general framework for data analysis. This toolbox extends the repository — essentially an unstructured "big bag" of peer-reviewed tools for data analysis — to a coherent and tightly cross-linked data-analysis package. The term "Toolbox" is used here in its widest possible meaning — i.e. it comprises a collection of quite different routines. All these routines will interact with a central interface to access the data. Individual users can then focus on their core scientific tasks while the toolbox takes care of the implementation, performance, and data handling details. Users will be able to submit bug fixes, patches, feature requests, and new algorithms. The long-term goal is to implement toolboxes in different programming languages, possibly based on existing software (see, e.g., http://mdp-toolkit.sourceforge.net/ and http://chronux.org) but with common API. To allow us to incorporate existing analysis code we will therefore provide interfaces to different programming languages/interpreters. The toolboxes will be based on the export/import functions and unified formats specified in "Data Access". We believe that in the long run Python will become the most commonly used programming language for analysis of neural data in the future. Currently, Matlab plays this role and will thus stay in use with a large community for the near future. An excellent working example of seamless integration of different programming languages are the Boost Libraries (http://www.boost.org/), highly optimized C++ routines, which can be efficiently called from Python.

• The toolbox and the code repository will remain open source and freely available.

• The success of the toolbox will depend not only on its scientific quality but also on the ease with which it can be used by a general neuroscience audience (Hidetoshi, Takuto, Ryohei, & Yoichi, 2007). Here, a crucial aspect is the "look and feel" of the G-Node Web Portal. Accordingly, a prerequisite for the portal's success is an attractive graphical layout. More importantly, the portal must also provide all information in an optimally user-friendly way and allow both experts and traditional scientists to interact as easily as possible with the tools and services provided by the Node. Last but not least, the portal should also be designed such that it facilitates the communication between interested scientists and the Node team as much as possible.

5. Teaching and training

5.1. Problem

During regular courses, neuroscience students are rarely exposed to neuroinformatics, computer science or advanced data analysis; some senior neuroscientists may not even be aware of recent developments in these fields. It is therefore mandatory to complement the infrastructure of the G-Node with parallel teaching and training activities. To give just one example: The introduction of "good practices" for experimental

Herz, A. V. M., Gollisch, T., Machens, C. K., & Jaeger, D. (2006). Modeling single-neuron dynamics and computations: A balance of detail and abstraction. *Science*, *314*, 80–85.

Hidetoshi, I., Takuto, N., Ryohei, K., & Yoichi, S. (2007). Development and application of CMS-based databasemodules for neuroinformatics. *Neurocomputing*, *70*, 2122–2128.

Horn, van J., & Pelt, van J. (2008). *1st INCF workshop on sustainability of neuroscience databases*. Available at http://www.incf.org/documents/workshop-reports/incfworkshop-1st-SustainabilityDatabases.pdf.

Kass, R. E., Ventura, V., & Brown, E. (2005). Statistical issues in the analysis of neuronal data. *Journal of Neurophysiology*, *94*, 8–25.

Kemp, R., & Olivan, J. (2003). European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data. *Clinical Neurophysiology*, *114*, 1755–1761.

Meier, R., Egert, U., Aertsen, A., & Nawrot, M.P. (2008). FIND — a unified framework for neural data analysis, *Neural Network* (Special issue on Neuroinformatics). doi:10.1016/j.neunet.2008.06.019.

Teeters, J. L., Harris, K., Millman, J. M., Olshausen, B. A., & Sommer, F. (2008). Data sharing for computational neuroscience. *Neuroinformatics*. doi:10.1007/s12021-008-9009-y.

Yamaji, K., Sakai, H., Okumura, Y., & Usui, S. (2007). Customizable neuroinformatics database system: XooNIps and its application to the pupil platform. *Computers in Biology and Medicine*, *37*, 1036–1041.