Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Feeding the human brain model

Paul Tiesinga[1], Rembrandt Bakker[1,2], Sean Hill[3] and Jan G Bjaalie[4]

The goal of the Human Brain Project is to develop, during the next decade, an infrastructure capable of simulating a draft human brain model based on available experimental data. One of the key issues is therefore to integrate and make accessible the experimental data necessary to constrain and fully specify this model. The required data covers many different spatial scales, ranging from the molecular scale to the whole brain and these data are obtained using a variety of techniques whose measurements may not be directly comparable. Furthermore, these data are incomplete, and will remain so at least for the coming decade. Here we review new neuroinformatics techniques that need to be developed and applied to address these issues.

Addresses
[1] Donders Institute, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
[2] Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6), Jülich Research Centre and JARA, 52425 Jülich, Germany
[3] Blue Brain Project, Campus Biotech, Bâtiment B1, Ch. Des Mines 9, Geneva CH-1202, Switzerland
[4] Institute of Basic Medical Sciences, University of Oslo, PO Box 1105, Blindern, N-0317 Oslo, Norway

Corresponding author: Tiesinga, Paul (p.tiesinga@science.ru.nl)

Current Opinion in Neurobiology 2015, **32**:107–114

This review comes from a themed issue on **Large-scale recording technology**

Edited by **Francesco P Battaglia** and **Mark J Schnitzer**

For a complete overview see the Issue and the Editorial

Available online 26th February 2015

http://dx.doi.org/10.1016/j.conb.2015.02.003

Introduction and background

A key goal of the 1 billion euro, 10 year Human Brain Project (HBP) is to build a scaffold model of the human brain. This will enable the global community iteratively build and refine whole brain models, starting from the mouse and working towards the human brain, which is about a thousand times larger. Different teams of researchers will each deal with different sets of challenges. One set of challenges is to develop the hardware and software to make it possible to simulate such a large-scale model, store and analyze its output, and control the simulation. Another set of challenges is to fully specify the computational model that needs to be simulated, and identify key missing data, which is the topic of this review. When these challenges are met, the HBP infrastructure will provide the community with new tools to accelerate the understanding of the brain in health and disease. The HBP approach can be seen as neuroscience data integration through model building. The idea is that only by building an integrated model will neuroscientists be able to find out whether there is missing data and, if so, determine an experimental strategy to measure it directly or use a predictive neuroinformatics approach (see below) to infer it.

Simulation of the brain models will occur at different scales and levels of abstraction. The cellular-level biophysical simulation is in simplified terms the numerical integration of a set of coupled (partial) differential equations. These equations, their parameters, and their initial conditions, need to be fully specified to run the simulation. For instance, there are differential equations that describe the time evolution of the membrane potential and ionic concentrations as a function of spatial location within a neuron [1,2]. Other equations describe the molecular cascades inside the cell [3], such as those involving transcription and translation into proteins [4,5] that are, for instance, necessary to lay down memories [6]. The general form of the equations is well established, and even though there may be discussion on what level of simplification is acceptable, the main challenge is to populate the model parameters with reasonable estimates.

To obtain the necessary data a three-pronged strategy is needed. First, integrate existing data from different labs. Second, predict missing data that have not (yet) been measured. Third, increase the amount of available experimental data through new molecular neurobiology techniques and industrial neuroscience approaches (i.e. high throughput). A key example of this approach is the work by the Allen Institute for Brain Science (AIBS).

The data integration challenge

The information constraining the HBP model comes from diverse sources and is obtained using different experimental techniques. Hence, for the same basic assertion, say the likelihood of a connection between two neurons in area A and B, there are multiple sources of data, each potentially giving a different answer. These data need to be integrated. A key problem is the representation of information in such a way that they are comparable and so

that their reliability and precision can be quantified and taken into account.

A group of researchers working in the same lab knows the experimental settings used when measurements are made. These measurements are stored in data files, most likely without documentation of what the units are, or what preprocessing (filtering) has been done. Information regarding sampling rates and other parameters might be stored separately (for instance, written in a lab book) or assumed to be the default setting specified in the protocol. The data will be subsequently analyzed using tools bundled with the experimental equipment or lab-specific user-written code. It will be stored in another file, without explicitly mentioning these post-processing steps and their parameters. This set up makes sense within the lab, but it means that the data are practically useless outside the lab, because the metadata — the additional information about the experiments — is missing. Further, standardization of the metadata is required to assure correct interpretation. Hence, an ontology of the concepts describing the data and measurement parameters needs to be developed. There are ontologies for neuroscience concepts, collected in the Neuroscience Information Framework (NIF) [7], but they do not yet cover the necessary experimental and analysis protocols in a comprehensive way. When, for instance, axonal projection patterns are determined using tracing experiments, the protocol can be structured in different modules, each of which can be characterized by well-defined parameters out of an ontology or controlled vocabulary (Figure 1a). When the data are subsequently stored in a database and referenced relative to a standard atlas, this information can be easily accessed by other researchers [8] (Figure 1b).

Predictive neuroinformatics

Predictive neuroinformatics aims to fill in missing data based on existing data and general principles. It naturally builds on methods developed in other fields where similar problems have been encountered. For instance, during clinical trials, for a given subject, sometimes only a part of the measurements are conducted. This leads to incomplete data, with missing entries for particular subjects, which need to be filled out using so-called imputation techniques [9]. One can consider these data as a matrix, where each row represents a subject and each column a feature. Microarray experiments can also be expressed as a matrix, where the rows represent conditions and the columns the expression level for each gene [10], or for genome wide association studies (GWAS) where the genotypes of a single nucleotide polymorphism (SNP) are placed in a matrix, with the row representing the subject and the column the SNP location [11]. The empty elements in matrix can be filled by replacing each empty element by the mean value of the corresponding non-empty elements, but more advanced techniques are available [10,12,13]. A recent approach, directly applicable to neuroscience, is matrix completion [14,15]. This approach assumes that the data matrix is comprised of a sum of a low-rank matrix and a sparse matrix [16]. The solution for this problem resembles the well-known LASSO procedure for regression [17], in which an L1 (i.e. absolute value) penalty term on the regression weights is added to the L2 loss function (i.e. squared difference). This shrinks the weights to zero, with those smaller than a specific value made exactly zero. In the matrix version, the low rank part is found by applying a singular value decomposition (SVD) and shrinking the singular values to obtain, after reconstitution of the components with the new singular values and after a number of iterations, the best fitting low rank matrix [18,19].

The wiring diagram of the nervous system in *Caenorhabditis elegans* has been determined [20,21], and the gene expression profile of each cell has also been measured and made publically available (http://www.wormbase.org). Analyses reveal that the expression profile can successfully predict the absence or presence of a synapse [22,23]. For the mouse, gene expression patterns [24] and meso-scale connectivity [25••] have been made publicly available by AIBS. (Note: the AIBS meso-connectome would be more appropriately referred to as macro-scale connectivity since currently it is at the level of brain areas rather than cell populations.) When a similar analysis was applied at the level of brain areas rather than individual cells, it revealed that the connectivity pattern was to a large extent predictable by areal patterns of expression [26,27]. This means that when gene expression patterns are available, but connectivity data are not, the connectivity can still be predicted.

Experimental sources of data to constrain the model

Brain areas

To build the brain model, the different brain areas need to be defined (a parcellation), their typical size determined, as well as the density of each cell type within that area and the distribution across substructures (layers, subnuclei). A number of brain parcellations have been proposed, each based on different criteria, such as cytoarchitecture or the density with which receptors are expressed [28]. Recently, progress has been made with data-driven approaches to define brain areas. As mentioned above, AIBS has made available an atlas of transcription patterns in the mouse brain [24]. Each voxel was characterized by a vector, which contained the strength of expression of each of the analyzed genes. Nearby voxels, likely belonging to the same area, should also have a similar expression pattern. When a clustering procedure was applied to these vectors a parcellation emerges [29••] (Figure 2d), which when 60 or more clusters are sought, resembles parcellations based on cytoarchitecture in microscopic sections (Nissl stains) or MRI contrast [30,31] (Figure 2e).