Neuroscience databases: tools for exploring brain structure–function relationships

R. Kötter

*C. & O. Vogt Brain Research Institute, and Institute of Morphological Endocrinology and Histochemistry,
Heinrich-Heine-Universität, Moorenstrasse 5, D-40225 Düsseldorf, Germany (rk@hirn.uni-duesseldorf.de)*

Faster than ever, neuroscience is generating vast amounts of data that await cross-referencing, comparison, integration and interpretation in the endeavour to unravel the mechanisms of the brain. The complex, diverse and distributed nature of these data requires the development of advanced neuroinformatics methodologies for databases and associated tools that are now beginning to emerge. This paper presents an overview of current issues in the representation, integration and analysis of neuroscience data from molecular to brain systems levels, including issues of implementation, standardization, management, quality control, copyright, confidentiality and acceptance. Particular emphasis is given to integrative neuroinformatics approaches for exploring structure–function relationships in the brain.

Keywords: database technology; data representation; data mining; copyright; analysis; computer models; neuroinformatics

1. INTRODUCTION AND BACKGROUND

Following other scientific fields like particle physics and molecular biology, neuroscience is currently entering a new stage that is shaped by the advances of information technology (IT). Decades of intensive and expanding neuroscientific research have yielded enormous resources of data on many different descriptive levels of the nervous system, spanning the organizational range from genomic to behavioural levels, and the temporal range from ionic to evolutionary processes. Collectively, these data will help to disclose the structural and functional architecture of the brain and provide insights into its mechanisms. The way to this ultimate goal, however, is still very long and arduous. Beyond the generation of high-quality data it is a major task to organize the data in a meaningful way such that they deliver insights into the organization of the brain at all spatio-temporal scales. Based on these data, we can characterize the components of a system whose global functions emerge from the interactions of the components over time. A multitude of complex system analyses, involving massive amounts of data, will be necessary to provide a better understanding of brain mechanisms.

Simultaneously, the large amount and the diversity of the data required for the analysis of complex brain systems make ambitious research projects less and less manageable if the traditional way of manually extracting, comparing, and evaluating data from the literature and other data sources is continued. Already, neuroscientists find it impossible to read all published work related to their own field of interest, not to speak of related fields that might deliver useful information or ideas. However, it is necessary to keep abreast of the flood of data. To an increasing extent we rely on selections, reviews, abstracts and other mediators between data and user to form our

opinions on the state of scientific knowledge. It is obvious, however, that such mediators can obscure or miss important aspects and may lead to subjectively biased or wrong conclusions. The challenge, therefore, is to provide insights through tools that allow for a systematic, flexible and efficient user-controlled access to a multitude of data.

Two decades ago molecular biology passed beyond the manual stage of scientific practice that neuroscience is only now beginning to leave. The number of Internet accessible sequence databases has exploded since the beginning of the 1980s, leading to a large ensemble of specialized data repositories that provide online assistance to the individual researcher (for reviews see, for example, Frishman *et al.* 1998; Persidis 1999; and the annual database issues of the journal *Nucleic Acids Research* at <http://nar.oupjournals.org/content/vol29/issue1/>). More and more research projects in molecular biology rely on such databases, either when researchers analyse their own findings in the context of the myriads of other published results (e.g. by determining homologies across different species), or when they scrutinize the available data in the search for unidentified answers to their questions. Databases and their associated tools have formed the foundation for the development of bioinformatics, a key discipline in modern molecular biology (Persidis 1999; Wooley 1999; Butler 2001; Roos 2001).

Neuroscience is gradually taking a similar direction. Several efforts have been started to provide the neuroscientific community with powerful databases of various data types ranging from literature references via experimental data to meta-descriptions of data. In direct analogy to the indispensable role of bioinformatics for molecular biology, neuroinformatics is about to gain considerable importance for the investigation of complex structure–function relations in the nervous system (Kötter 1999; Beltrame & Koslow 1999; Young & Scannell 2000). A major impediment

for applying the IT developments in molecular biology to neuroscience is the tremendous complexity of neuroscientific data. They require their own and innovative ways of data representation, expert maintenance, data integration and analysis. Nevertheless, there are many experiences that neuroinformatics can glean from the field of bioinformatics. Particularly well developed in the latter field are the organizational, technological, financial, legal and sociological aspects of large-scale databasing. In addition, a considerable subset of molecular data is immediately relevant to neuroscience. Given these touching points, it is surprising that comparatively little interaction takes place between these two communities. But this is very likely to change with the growth and maturation of neuroscience databases. Already, there are a growing number of collaborations between molecular biologists and neuroscientists in tasks such as correlating brain structure and function with various molecular markers, or characterizing the effects of gene knockouts in mice at morphological, electrophysiological and behavioural levels. Bringing together all of these different aspects of information will create a formidable data source for scientific investigations into the workings of the brain.

2. SCOPE OF DATABASES AND SOFTWARE TOOLS

The current scope of neuroscience databases and associated software tools ranges from data inventories for personal use and specialized data collations by a subcommunity to large-scale database projects of general interest.

Individual and specialized data collections and research tools usually arise from the specific requirements and interests of a local research group. Many of these efforts are never published and sometimes re-created in several places; some have grown and make a considerable impact, such as software for statistical analyses or computer modelling (see, for example, statistical parametric mapping (SPM), <http://www.fil.ion.ucl.ac.uk/spm/>; general neural simulation system (GENESIS), <http://www.bbb.caltech.edu/GENESIS/>; NEURON, <http://www.neuron.yale.edu/>), and collections of structural and functional data (Felleman & Van Essen 1991; Scannell *et al.* 1995; the BrainMap project, <http://ric.uthscsa.edu/projects/brainmap.html>). Large-scale projects are usually initiated and carried out by a consortium of research groups in the context of an ambitious research programme (see the Human Genome Project, <http://www.ornl.gov/hgmis/>; the Human Brain Project (<http://www.nimh.nih.gov/neuroinformatics/index.cfm>) or the International Consortium for Brain Mapping, ICBM, Mazziotta *et al.*, this issue). The American Human Brain Project, as the outstanding example of neuroinformatics funding in neuroscience, supports a variety of research approaches that develop databases and tools that can efficiently deal with the complexity of brain data (Shepherd *et al.* 1998). This includes databases, graphical interfaces, querying approaches, information retrieval, data visualization and manipulation, data integration through the development of integrated analytical tools, synthesis, and tools for electronic collaboration. The focus is on development of structural, functional, integrative, and analytical models and simulations.

After the first call for feasibility studies in 1993, the Human Brain Project is now in its second round and is beginning to make an impact on the development of information technologies in neuroscience. Europe and Japan are still slow in picking up on the developments in neuroinformatics, even though individual research groups in these countries significantly contribute to progress in the field (see EU-US workshop on 'Databasing the brain', 1-2 July 2001, <http://www.nesys.uio.no/Workshop/>). However, the foundation stones are now laid and the field is now shaped. The Human Genome Project has been a precedent, showing that the scope of informatics approaches is of international scale and significance. Within less than 15 years this project has completely changed the way research in molecular biology is carried out and how we perceive it. It seems not too far-fetched to predict that neuroinformatics approaches will do the same for neuroscience research. Developments will probably repeat many of the experiences with the Human Genome Project. To review the state of the art and to propagate the insights gained, it seems timely and appropriate to provide an overview of the achievements, difficulties and prospects of neuroscience databases and associated tools in the field of neuroinformatics.

3. DATA ACQUISITION

The first prerequisite for any database is the availability of data to base it on. At first sight, the acquisition of data does not seem to pose a particular challenge; in fact the majority of database projects are born in situations where a growing amount of available data calls for more efficient ways of handling them. As examples, multi-electrode recordings in behaving animals, or functional imaging experiments in humans, create vast amounts of data that immediately require strategies for efficient storage and indexing. The various independent developments have led to many different and incompatible data formats that hinder data exchange and data integration. This lack of standardization is a major challenge in data acquisition as soon as the data are acquired from multiple external sources (see Mazziotta *et al.*, this issue; Van Horn *et al.*, this issue).

A second problem in data acquisition is incompleteness of datasets and of their accompanying information. For a start, the data presented in scientific publications are usually only a small subset of the useful data produced in the course of experiments. Besides the bias to publish positive rather than negative findings, the pressure for publication space often limits presentations to the most significant data that suit the exposed scientific argument, whereas seemingly minor, poorly understood, or merely confirmatory data are suppressed. Even with the most interesting data the accompanying description of experimental details is often sketchy and, despite the peer review process, insufficient to repeat or evaluate all aspects of the experiment without additional information. Therefore, some databases extend their scope to raw data underlying the published results or even to unpublished data (see Shepherd *et al.* 1998; Van Horn *et al.*, this issue).

Third, the amount of available data may not suffice for informative meta-analyses or, by contrast, it may exceed the data-handling capacities. Notably, the significance of

genome and protein databases results from the massive amounts of stored data arising from the addition of hundreds or thousands of new entries per day (GenBank currently contains more than 12 billion bases in more than 11 million sequence records, <http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>). This amount of input relies on the active participation of the data producers who have an interest to see their data represented in a database, often as a prerequisite for publication (see Le Novère & Changeux, this issue). In neuroscience the situation is different so far: although databases cover many different types of data (e.g. receptors, ion channels, neuron types, connections, circuits, spike trains, activations, models), most data producers have no necessity and few incentives for taking the time and the trouble to enter their data into public data repositories. The majority of data collations in neuroscience are carried out by a comparatively small group of neuroscientists for the purpose of the meta-analyses that they want to perform.

4. DATA QUALITY CONTROL

As soon as data are acquired, the issue of quality control arises. Even in small databases where a single individual scrutinizes all data it is necessary to formulate explicit criteria for the types and requirements of data that are to be represented in the database. A common and convenient procedure is to refer to external control mechanisms, such as the criterion of publication in peer-reviewed international journals. Clearly, the peer review process has its defects, and its aims are not fully congruent with the aims of specific database projects. Nevertheless, this procedure has the advantage of being generally understood and widely accepted. It is very laborious for a database project to implement its own independent review system, which is necessary if preliminary communications (published abstracts) or even unpublished data are to be included (see Stephan *et al.*, this issue, for appraisals of data quality). In addition, relying on external review keeps rule making separate from data collating, which reduces possible conflicts of interest in the selection and evaluation of data.

In any case, it is necessary to keep track of the sources of data, to verify the accuracy and completeness of the entries, and to check the appropriateness of the data representations. The latter is a particular worry where the original data have to be classified or transformed to meet the database structure (e.g. reduction of dendritic morphology to the canonical neuron concept of proximal, middle and distal dendritic regions used in NeuronDB; Mirsky *et al.* 1998, Shepherd *et al.* 1998). This is already the case if authors enter their data themselves but it becomes even more complicated if a dedicated data collator appraises the original data. Ideally, different data collators should arrive at the same representation of a data set, and the objectivity of this additional step can be evaluated (see Stephan *et al.*, this issue; Burns, this issue).

Another issue of quality control refers to the difference between raw data as presented by the data source, and interpreted data as derived by some process of evaluation or curation. The most important rule is that there must be transparency so that raw and interpreted data remain

identifiable and separable. The next question concerns the quality of the curation process. The simplest case is data annotation, which in molecular databases has been automated to some extent (see Frishman *et al.* 1998). Even if it cannot be automated completely, the rules for data curation must be specified and documented as far as possible. Important methods for data annotation and evaluation are automatic tissue classification in brain images (see Dickson *et al.*, this issue; Mazziotta *et al.*, this issue) or coordinate-independent mapping of brain data (Stephan *et al.* 2000a). There is a continuum from such methods to more sophisticated data analysis tools (see § 8).

5. DATA REPRESENTATION

The issue of data representation is central to most databasing approaches in neuroscience, and it poses the biggest challenges. Compared with the complexity of most neuroscience data, the representation of DNA, RNA and proteins is acknowledged to be comparatively simple: the latter can be reduced to nucleotide and amino-acid sequences, respectively, whose representation requires only linear strings of a small set of letters. This simple and method-independent representation is feasible because the code for nucleic acids and proteins has been deciphered. Equally universal and simple codes are not known for the fundamental objects of neuroscience, such as neuronal cell types and their activity patterns, cortical columns and layers, not even for macroscopic entities such as brain areas and nuclei (see below). The present state of research recognizes multiple and partially incompatible classificatory schemes, which cannot be interpreted without reference to the methods for their delineation. It may be compared with the evaluation of DNA microarrays for studying gene expression, which requires supporting information defining the type of chip used, the nature of the tissue samples, the precise procedures used to hybridize the sample RNA to the DNA spot, etc. (Anonymous 2001). At this stage of knowledge it is often unclear what aspect of the data can be attributed to a fundamental biological property, what is related to methodical limitations and what is a plain artefact. Such ambiguities have a long history and are notorious in attempts to create systematic ontologies of neuroscientific objects (see classifications used by the NeuroNames system: <http://braininfo.rprc.washington.edu/aboutindex.html>; by learned societies and by databasing approaches, e.g. <http://www.sfn.org/cfa/themes.html>; Gardner *et al.*, this issue; Mazziotta *et al.*, this issue; the Ion Channel Network, <http://www.le.ac.uk/csn>; and Conley 1998).

Consequently, there are two fundamentally different approaches to describing the localization of brain data: the first approach references neuroanatomical entities such as neuronal cell types, columns, layers or areas as mentioned above. This approach would be ideal if these entities could be identified in each individual brain, and if their fundamental roles could be unambiguously specified. The alternative approach avoids these ambiguities with the specification of a three-dimensional reference system. These two concurrent approaches are exemplified by the connectivity databases CoCoMac (Stephan *et al.*, this issue) and XANAT (Press *et al.*, this issue), respectively. Three-dimensional coordinates are the usual

reference frame of image databases (Roland & Zilles 1996; Dickson *et al.*, this issue; Mazziotto *et al.*, this issue; Van Horn *et al.*, this issue); for interindividual comparisons they depend crucially on adequate warping methods to transform individual brain images to the same reference system (see Mazziotto *et al.*, this issue).

Apart from interindividual variations among brains the context dependence of functional data is a particular challenge, even within the same individual. The complexity of experimental paradigms makes it difficult to develop adequate parameterizations to represent, classify and compare functional data. Examples are provided by NeuronDB (Mirsky *et al.* 1998), BrainMap (<http://ric.uthscsa.edu/projects/brainmap.html>) for positron emission tomography (PET), or Van Horn *et al.* (this issue) for functional magnetic resonance imaging (fMRI). It is well known that small parameter variations can have large effects on the observed activation patterns in functional imaging, which are often hard to quantify or control (Casey *et al.* 1998). Unexplained variations occur even in the same subject under apparently the same experimental condition (McGonigle *et al.* 2000). Notwithstanding these challenges, databasing efforts are in no way limited to the comparatively simple arena of neuroanatomy; they are also prominent in human brain mapping, where large datasets, good computational infrastructure and international comparison create a favourable environment for innovative developments.

Assuming that an adequate representation is feasible, it is often necessary to decide between several alternative representations: a fundamental decision concerns the relationship between the number of observed datasets and the number of their representations in a database. If the datasets contain mutually congruent information, it may be tempting to create one summary representation in the database. This approach reduces storage demands but impedes future analyses of features in the individual datasets that were not obvious at the time of their collation. For example, new analysis techniques may make it possible to quantify the degree of congruence between datasets more precisely, or seemingly minor and thus neglected aspects of the datasets may gain more importance at a later stage. Thus, if recovery of individual datasets is an issue, then a one-to-one relationship between datasets and separable database representations is necessary (ignoring identical datasets). This storage-intensive approach is required for look-up databases (e.g. Entrez PubMed, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>) and for analyses of variation among datasets (see Press *et al.*, this issue). All major collations of connectivity data have made a transition from earlier many-to-one representations to one-to-one representations (see Stephan *et al.*, this issue). The situation is different with molecular sequence data where it is argued that the step of representing multiple datasets by one database entry does not reduce flexibility but actually weeds out errors and method-related noise (see Le Novère & Changeux, this issue). Nevertheless, even then a pointer to the original datasets in other 'raw' databases has been retained to permit detailed reference.

A far more compact data representation is the extraction of a small set of parameters that are sufficient to characterize one or many full datasets. For example,

Ascoli *et al.* (this issue) describe the use of various morphological parameters in combination with stochastic techniques to generate full datasets of dendritic morphology that are statistically indistinguishable from the original data. Despite this statistical faithfulness, a neuroanatomist immediately distinguishes the examples of artificially generated neuronal morphologies provided by Ascoli *et al.* from the appearance of real neurons of the respective cell types. This points to the limitation that this compact type of representation discards some of the original information. Whether this loss is significant or not depends on what one considers being relevant properties of dendrites: what an anatomist considers a significant deviation from typical morphology may be functionally irrelevant, and vice versa.

Computational models of neuronal and brain function are probably the most compact form of representation of biological data (see Conley 1998). Such models parameterize not only morphological but also some functional properties of brain structures. A small number of parameters can be used to generate not only neurons and networks, but also to simulate the dynamic behaviour of these structures in health and disease (see e.g. Reggia *et al.* 1999). A neuroinformatics extension of this approach is to organize computational models and their components in the form of a database (see ModelDB, <http://senselab.med.yale.edu/senselab/ModelDB/>; Modelers' Workspace, <http://www.bbb.caltech.edu/hbp/mws-overview/mws-overview.html>; Goddard *et al.*, this issue). Even the design of (analogue) very large-scale integration (VLSI) hardware implementations of neuronal models may profit from a systematic databasing approach to models of neuronal circuitry (see Breslin & O'Leakie, this issue).

6. SYSTEM IMPLEMENTATION

The implementation of databases has wide-ranging aspects, of which some are genuine computer science research topics themselves. It is beyond the scope of this paper to survey this rapidly developing field. Notwithstanding the complexity and the rapid development of this field, even ambitious database projects can be carried out using the existing technology and most require only standard products, which minimizes the overhead associated with their implementation.

Standard desktop software, such as Microsoft Access, nowadays offers more functionality than the average home-grown database project requires. This includes a relational design for efficient representation of heterogeneous datasets, structured query language (SQL) data access, creation of graphical interfaces and (static) hypertext mark up language (HTML) pages, as well as replication technology for multi-user environments. The more professional Microsoft SQL server product comprises enhanced database management and integrity with the support of transactions and stored procedures. On UNIX platforms, MySQL and PostgreSQL are widespread multi-user-capable systems with the additional advantages of being available at no cost and in source code. All of these are very common in business applications and can be used for the dynamic generation of HTML pages, for example with common gateway interface (CGI), active server pages (ASP) or PHP: hypertext

preprocessor scripts. Commercial database systems can be licensed from Oracle, Sybase, Informix, IBM (DB2), and others. Besides the costs for the licences some database projects in neuroscience are limited by requirements for efficient storage and processing, e.g. of large datasets (mainly image data), for flexible representation of heterogeneous data types, and for object-oriented rather than relational data representation (see Mirsky *et al.* 1998; Goddard *et al.*, this issue). Finally, development cycles and portability of database implementations can be enhanced through reusable software components and platform-independent programming languages such as Java (see Gardner *et al.*, this issue; Burns, this issue).

The exchange of data across the Internet would be facilitated if the database structure allowed a data representation that was already in an Internet-suitable format, for example the emerging extensible mark-up language (XML) standard. While genuine XML databases are being developed, at present, data are converted several times to facilitate their way from one database system across the Internet to another (see Gardner *et al.*, this issue). In addition, the exchange of data requires a common dictionary. A universal mark-up language for neuroscience data could efficiently deal with this semantic aspect (see Goddard *et al.*, this issue). Finally, there are performance issues when the size of data or the number of database users becomes very large as in the case of the Entrez database, which processes more than 50 000 queries per day, or the protein database Swiss-Prot, which has more than 200 000 users worldwide. Since the bandwidth of the Internet has not grown in parallel with the amount of data passed via it, strategies for efficient data transfer and presentation can become crucial. The acceptance of databases depends to a considerable extent on their response times; therefore, it is worthwhile considering issues such as server- or client-side processing, dynamic IP addresses and installation of mirror sites.

Apart from the databasing of neuroscience data, their analyses often reach a degree of complexity that requires special provisions for efficient data representation, scalable and interactive processing, and extended fail-safe computations. Examples are the optimizations carried out by the CANTOR package within the portable UNIX programming system (PUPS) environment (O'Neill & Hilgetag, this issue) or the objective relational transformation approach to the mapping of connectivity data (Stephan *et al.* 2000a).

7. USER INTERFACE AND DOCUMENTATION

The requirements for presentation and documentation of a database depend on whether it serves a personal use, a small group of collaborators, the neuroscience community or the public. Even if the database is developed and used by a single person, this individual should already see a need for documentation since programming details easily fall into oblivion and criteria for data collation may drift over time. As more people become involved, written documentation becomes essential. Public database presentation calls for additional explanation and an intuitive interface, particularly if the user is left alone with the material. It has become so easy to present data online via the Internet that this can be taken as the default case of

interest. Superseding the download of data files in individual text or binary format by file transfer protocol (FTP), it is now common to present HTML-formatted data on the Internet. The complexity of these presentations can range from simple viewing of text or tables to dynamically created graphical interfaces with data-associated programs and specialized up- and download facilities. The choice of the presentation will depend on the type of data, the ease of implementation and the purpose of the interface. As an example, neuroscience data can be retrieved and presented as text referring to the structures of interest. The same data could be presented graphically for enhanced display of spatial or temporal properties, or for interactive selection. This graphical presentation can be more intuitive, but it usually involves a larger programming effort and may require corresponding viewing software through plug-in programs or software applets on the user side. The idea of creating reusable graphical software components (e.g. 'biowidgets', <http://www.cbil.upenn.edu/bioWidgets/>) has not yet had much impact on neuroscience databases. Graphics, although user-friendly, may also be misleading where they obscure the reference system of the data (see §5): they may imply a morphology that is not present in the data when drawing individual data on a general template or when using a precise coordinate system for representing approximate locations (see the atlas of Talairach & Tournoux (1989) and its online version at <http://ric.uthscsa.edu/projects/talairachdaemon.html>).

Probably the best-known biomedical database with relevance to neuroscience is the literature database PubMed. This database comprises original text and descriptive data from published journal articles. Despite the uniformity and simplicity of the data the Internet interface is fairly complex. For example, both free text and advanced field-delimited searches are possible; free text is automatically converted into a detailed query that is hidden from the common user. The resulting output pages contain in addition to the literature-related links a large number of display options. If we now imagine a specialized database for more complex data than journal articles with sophisticated tools generating large datasets in various versions then it becomes clear that database management is not a minor or a trivial issue (see Dickson *et al.*, this issue). Reliable data management, intuitive presentation and concise documentation are required to tell the user what the data are, how to obtain the desired subsets of them, and how to interpret the retrieved output.

Data upload is another challenge. Here it must be verified who the contributor is and that the contributed data have the correct contents, structure and format. To cover a variety of situations the interface may differentiate between required and optional data fields and accept different data formats. In some database implementations, the terminology and the required structure of the data differ for technical reasons from neuroscientific customs. In the case of the CoCoMac database (Stephan *et al.*, this issue) the manual, which specifies the rationale of appropriate data entry, comprises more than 50 written pages, putting high demands on the data collator. Therefore, it is not surprising that most database projects in neuroscience do not envisage data entry by the data producer but rely

on trained data collators to incorporate new data. This contrasts strongly with genome and protein databases, where the large number of daily new entries could not possibly be cross-checked by experts (see Le Novère & Changeux, this issue). An intermediate approach is taken in NeuronDB, where the user can submit data to the database manager, who will review and add them to the database if appropriate (Mirsky *et al.* 1998).

A growing number of databases offer facilities for automated data search and retrieval. This strategy short-circuits interactive Internet dialogues and provides opportunities for more flexible and more detailed interactions. As a simple example, authors frequently use the automatic retrieval capability of the PubMed database to create links to the abstracts of their journal articles, including formatting instructions for their display (see <http://www.ncbi.nlm.nih.gov/entrez/query/static/linking.html>). This strategy can be extended to the retrieval of selected datasets from multiple databases. If the format of the results is known (for example through the use of a predefined mark-up language), then automatic processing of the retrieved data for seamless integration into a common portal becomes a possibility (see § 11).

8. TOOLS FOR DATA ANALYSIS

Current DNA sequence and protein databases are formidable data repositories. Databasing, however, is not a goal in itself. To paraphrase a statement by Hamming (1962): the purpose of databasing is insight, not data. Beyond the storage and retrieval of data in the sense of a look-up table the challenge is to get out of databases new information about the comparative and interactive properties of the elements described by the data (Roland & Zilles 1996; Conley 1998; Shepherd *et al.* 1998; Wooley 1999). This additional value requires the development of sophisticated tools that support analysis and integration of data. Simple examples of such tools in bioinformatics are the basic local alignment search tool (BLAST) and others that detect pattern similarities (e.g. homologies) in sequence data and that efficiently correlate similarities in sequence with known functions. Efforts are under way to develop programmes for making more sophisticated inferences, such as the prediction of genes within sequence data (Genscan, <http://genes.mit.edu/GENSCAN.html>) and the prediction from the linear amino-acid sequence of the protein's conformation or even of its biological actions. In neuroscience, too, already the correlation of different types of data, such as brain morphology and DNA genotype (see Mazziotto *et al.*, this issue) or connectivity and receptor distributions (Kötter *et al.* 2001a), can lead to new inferences. The neuroinformatics challenge, however, is to develop tools that will allow more sophisticated analyses, e.g. for recognizing organizational patterns in connectivity data (e.g. Young *et al.* 1995; O'Neill & Hilgetag, this issue) or for making inferences on activation patterns (e.g. Friston *et al.* 1995). Eventually, we need not only high-quality data in accessible databases, but a federation of databases (see § 11) and innovative inferential approaches that can compare and evaluate data from heterogeneous sources across data modalities and species (see Burns, this issue; Bota & Arbib 2001; Kötter *et al.* 2001a). This could enable us to

address questions systematically such as how properties of brain structures vary within a brain, across individuals or across species.

Closely linked with databasing is computer simulation as perhaps the most far-reaching approach to multi-level integration of structural and functional data. In molecular biology, an unprecedented increase in computational studies of gene regulatory networks (Hasty *et al.* 2001) is accompanied by massive public and private funding. For more than a decade specially designed neuronal simulation environments such as GENESIS and NEURON facilitate the implementation of biophysical models for detailed studies of the interplay of morphology, intrinsic and synaptically activated currents, intracellular calcium, etc., in the generation of neuronal firing patterns. Besides building more specific and more detailed models, multi-level efforts aim at an integration of biophysical and biochemical models of neuronal information processing (Kötter & Schirok 1999) and the linking of detailed micro-circuit representations with more simplified network models (Kötter *et al.* 2001b). The investigation of multi-level systems will reach a new stage if flexible and integrated object-oriented modelling tools become available (NEOSIM; see Goddard *et al.*, this issue). A panoply of general as well as special purpose modelling tools with applications ranging from biochemical pathways to brain systems is now easily available (for a listing of modelling tools and databases see <http://www.hirnforschung.net/cneuro>).

As a growing amount of data becomes available electronically, computational tools extend and refine the scientific work of sorting, analysing and integrating data. In fact, data mining is an attractive commercial application. Vice versa, however, databases and their associated tools also act back on the scientific community and change the patterns of research. Literature database searches have become a regular part of scientific work, and a growing proportion of high-profile research includes a neuroinformatics component (Butler 2001). Future developments will include data-mining tools for non-trivial extraction of implicit, previously unknown, and potentially useful information from neuroscience data. To facilitate the access and to reduce the time spent on data mining it will be useful to develop intelligent data-gathering systems ('knowledge robots') that alert the user in an individualized fashion. The use of such tools is primarily for informing scientific enquiries into the organization and functioning of the brain. In the long term, data mining may gain an important role in the search for causes of neurological and psychiatric disease and the design of effective therapeutic approaches.

9. BUDGET AND MAINTENANCE

As a database project grows it involves more and more people, leading to a differentiation between database developers, administrators, data collators and general users, etc. At this point, new managerial strategies have to be developed, which may aim at purpose-designed development grants (most projects described in this theme issue), integration into scientific library systems (e.g. PubMed, Human Genome Project), or conversion into the public domain (e.g. XANAT) or into a commercial

enterprise (Celera). Whereas financial support is comparatively easy to obtain for developing new databases, funding their maintenance can become a problem, even for exceptionally important databases. Swiss-Prot, when facing imminent closure for lack of continued funding, started to charge corporate users an annual fee (Persidis 1999).

We still know very little about the life cycles of databases. How many database projects are started per year? how fast do they grow? How many are stopped and for what reasons? What are the conditions for their success and longevity? Most databases are planned as open-ended projects and grow with the number of newly generated data in the field. Even if the type of data collated undergoes a shift, general-purpose modular software tools for storage, query, display or analysis of neuroscience data can always be adapted to changing requirements, increasing their permanence (see Gardner *et al.*, this issue). Some data collations come to closure with the approximate exhaustion of the available data sources (CoCoMac collation of strychnine neuronographic data, Stephan *et al.* 2000b) or after extensive exploitation of the collected material (Scannell *et al.* 1995, 1999).

10. LEGAL AND ETHICAL ISSUES

The first person to be credited is the data producer, without whom no data would be available for databasing. There is a general view that specific reports of data including the text and figures should be properly copy-righted by the publisher, whereas the underlying data themselves should be identified with the data producer, who remains free to use or distribute them (see Gardner *et al.*, this issue). In this regard, two challenges have been identified in bioinformatics that could potentially restrict the advancement of research (Roos 2001).

- (i) The first concerns questions related to the appropriate use of data released before publication, which is the policy adopted by the Human Genome Project. Some proprietary status is granted to these republication releases, and second-party publications based on such data have to add a significant degree of value as expected from a scientific work.
- (ii) The second challenge concerns restrictions on the reposting of published data. Private companies, like Celera, can place restrictions on, for example, the incorporation of their data into existing databases and thereby prevent re-analyses and data mining. Fortunately, such limitations are currently not severe due to the existence of the public genome databases. In neuroscience we do not yet see much private investment into the large-scale generation data beyond the molecular level.

Having said this, it is also important to give adequate credit to the database collators for their work in developing and populating available databases. Since neuroscience data tend to be very complex, this work involves far more than a technical job or a trivial meta-analysis. The issue of proprietary rights was highlighted in a recent debate on the sharing of functional brain imaging data (Aldhous 2000; Koslow 2000). While free

availability of full data relating to published studies increases transparency and serves the research community, authors are concerned that full publication might compromise their ability to reap the full benefit of the data that they produced. The problem is complicated by the conflict of interest that occurs when journals make publication dependent on submission of data to a database that potentially serves the personal interests of members of the editorial board. In this respect, there is much to be learnt from molecular biology, where mandatory submission of research data to a public database has itself been established as a useful procedure. The provision is that authors can temporarily put their data on hold and decide whether they want to disclose them to interested colleagues. This procedure balances the interests of the authors and the public, and provides mutual benefits. The progress made with this issue is indicated by the removal during the course of the review process of the eponym 'national' from the fMRI database project (Van Horn *et al.*, this issue). This could reflect a change of perspective from a restricted top-down imposition to an internationally open initiative for collecting fMRI data.

As another aspect, free availability of full fMRI data sets could compromise anonymity of the subjects, since the data might be used to reconstruct individual faces (Aldhous 2000). In response to this concern more rigorous efforts are now made to ensure anonymity of the subjects and to inform them specifically of the submission to public databases (see Van Horn *et al.*, this issue).

11. COMPARISONS AND FEDERATION

Neuroscience databases collectively cover a vast spatial and temporal range of investigated phenomena. Each individual database, however, focuses on a specific aspect of it with perhaps the widest range being covered by the ICBM project, which at present is not publicly available (Mazziotta *et al.*, this issue). The construction of the various databases did not follow a predefined plan but was investigator-driven, with the consequence of gaps and partial overlaps between the databases. Altogether, there is comparatively little competition between alternative approaches as a consequence of the limited data coverage of each database and the large number of problems, for which no optimal solution has yet been found. In addition, we lack criteria for assessing the quality of databases objectively.

An explicit goal of neuroinformatics, however, is the interlinking ('federation') of various approaches to facilitate integrative investigations and data mining. This includes interoperability between different databases as well as between databases and software tools for analysis and simulation. A useful development towards the goal of database federation is the common object request broker architecture (CORBA), a low-level application framework providing mechanisms by which objects transparently make requests and receive responses in heterogeneous environments transcending different hardware and software platforms. As an alternative to the hypertext transfer protocol (HTTP), it is currently not very widely used.

A recent and very promising development is the use of XML to define standards for data exchange using the HTTP protocol. XML-based standardized dictionaries of

Table 1. Survey of topics covered by the papers of this theme issue.

(The number of crosses gives a rough impression of the space devoted to discussing the corresponding topic ranging from: empty space, not covered; via x, mentioned; to xxx, extensively discussed. Note that this grading is approximate, that it does not evaluate the contents of the statements, and that a single cross may result from a short note effectively covering everything that had to be said.)

author(s)	data acqui- sition	data quality control	data represent- ation	system implemen- tation	interface/ docu- mentation	associated tools/ analysis	budget / mainte- nance	legal/ ethical issues	compari- sons / federation	impact/ significance
Ascoli <i>et al.</i>	xx	x	xxx	xx	x	xxx			x	xx
Breslin & O'Leurskie	x	x	xxx	x	xx	x				xxx
Burns	x	x	xxx	xx	xx	xx			xx	x
Dickson <i>et al.</i>	xx		xxx	xx	xxx	xxx	x		xx	x
Gardner <i>et al.</i>	x		xxx	xxx	xxx	x		xx	x	x
Goddard <i>et al.</i>			xxx	xxx	x	xx			xxx	xx
Le Novère & Changeux	x	xx	xxx	xx	xx	x	x	x	xxx	xx
Mazziotta <i>et al.</i>	xxx	xx	xx	x	xx	xxx	xx	x	xxx	xxx
O'Neill & Hilgetag			xx	xxx	xx	xxx			xx	xx
Press <i>et al.</i>	xx	xx	xx	x	xx	xxx			xx	x
Stephan <i>et al.</i>	xx	xx	xxx	xx	x	xxx	x		xx	x
Van Horn <i>et al.</i>	xxx	x	xxx	xxx	xx	xxx	x	xxx	x	xxx

neuroscience entities (including NeuroML, see Goddard *et al.*, this issue) are currently being developed and follow a trend in bioinformatics (see <http://www.bioml.org>), where even commercial companies now introduce their own data formats for batch ordering of products by big customers. Practical advantages of the use of predefined mark-up languages are the automatic validation of data requests and the opportunity for computer-integrated data processing. XML standards for data exchange are being implemented by several neuroscience databases (see Gardner *et al.*, this issue; Goddard *et al.*, this issue; Stephan *et al.*, this issue; Burns *et al.* 2001). These standards will support the development of integrative workbenches (see the NCSA Biology Workbench project, <http://bioweb.ncsa.uiuc.edu>). Currently, perhaps the best working example of a database federation in neuroscience is provided by the BrainInfo database with its dynamic links to other online databases (e.g. <http://braininfo.rprc.washington.edu/scripts/hiercentraldirectory.asp?ID=164>).

12. IMPACT AND SIGNIFICANCE

Given the recent excitement about neuroscience databases and neuroinformatics, it is well to ask about the impact and significance of these projects. At the start of the Human Brain Project was the concern that the devel-

opment of new technologies for databases and associated tools outstripped the ability of most neuroscientists to use these technologies (Shepherd *et al.* 1998). Specific funding of 'enabling technologies' should put neuroscientists into a position where they could make more efficient use of their data. A recent survey of the state of neuroscience databases (Chicurel 2000) concluded that the impact of neuroinformatics and the Human Brain Project has so far been modest. This is not surprising in view of the huge conceptual difficulties that have to be overcome to make neuroscience data comparable and interoperable. Even in bioinformatics, where unprecedented amounts of data are now available, and many new companies and research positions in bioinformatics are being created, these are risky investments into the future with the prospect of a long-term pay-off. It is well to remember that in times of a 'rush' those that have the most secure pay-off are those who sell the equipment. Once such developments take off, however, there will be a severe shortage of adequately trained people, as is now seen in bioinformatics.

For the time being, most experimental neuroscientists are not as excited as the media. There are many reasons for this:

- (i) databases are seen as mere data repositories that do not necessarily create insights;

- (ii) meta-studies on the basis of existing data still have a reputation of being simple second-rate work, if not outright plagiarism;
- (iii) few available databases have already sufficient amounts of properly curated data that would be of wide use to people other than those who created the database;
- (iv) federation of databases is still in its infancy;
- (v) accessing available databases is often slow and tedious—they are not intuitive enough and too frequently require computer knowledge, additional software installation, or organizational overheads;
- (vi) threats of compulsory data sharing and poorly worked out confidentiality issues do not help to increase the confidence in databases.

Thus, there are a number of practical issues that impede immediate progress, and these are not easily offset by the long-term prospects of more systematic planning of experiments, the formulation of more integrative research questions and a shift to performance of meta-studies and virtual experiments based on systematic data collations (Chicurel 2000; Butler 2001). Eventually there is, in fact, a reasonable concern that indiscriminate use of databases may even contravene scientific progress: databases might simply serve to find experimental backing for basically any scientific hypothesis. Some statements will almost always exist 'out there' in the realms of large databases that seemingly support it. Thus, as access to a wide range of data becomes easier it will be necessary to develop a more sophisticated culture of systematically retrieving, comparing, evaluating and communicating data. Analogous developments take place in other fields even where they are not driven by information technology (e.g. preparing, maintaining and promoting the accessibility of systematic reviews in the Cochrane collaboration, <http://www.cochrane.org/>).

13. FUTURE DIRECTIONS

There is much to do to improve neuroscience databases so that they can make a significant impact on our understanding of the organization and the workings of the brain. As discussed above this is not simply a matter of better data and more innovative neuroinformatics technology; it essentially also comprises issues of acceptance in the neuroscience community to make the paradigm shift that has already taken place in particle physics and molecular biology. Among the many aspects presented and discussed in this theme issue (see table 1 for a survey of what the articles are concerned with) the following topics may be emphasized:

- (i) enhance experimentalists' awareness of the added benefits that they may gain when presenting their data to databases and foster cooperation between neuroscientists and computer scientists through focused workshops, educational courses, competence centres, interdisciplinary grants and training;
- (ii) develop strategies for systematic data collation, analysis and modelling in key and problem areas, e.g. drug effects on cellular machinery; interactions between biochemical and electrophysiological processes; mechanisms of neuronal development and growth, spatio-temporal imaging and analyses of brain activity; parameterization of behavioural variables;
- (iii) establish strict but suitable standards for data quality and data format as well as for database and software interfaces in neuroscience (e.g. based on the XML standard);
- (iv) build integrated simulation tools with flexible plugin components for multi-level computer models: figuratively speaking, to 'build cars' instead of 'reinventing wheels'.

A substantial number of neuroscience databases and associated tools are now available and an increasing number of them are accessible via the Internet. Is there any particular dataset that should be preferentially collected and made available? Clearly, there is no single dataset in neuroscience that has equally unanimous prominence as the sequence of the human genome in molecular biology, and thus the Human Brain Project does not parallel the glamour of the Human Genome Project. Having said this, it is obvious that the genome sequence is only the start or the 'blueprint' of what we need to know about cellular mechanisms (Wooley 1999) and that the topics of large-scale molecular projects are to a large extent driven by commercial interests and expectations of prospective markets (Persidis 1999). In neuroscience, current commercial expectations are comparatively small or remote, focusing on the development of more specific neurotrophic drugs and of intelligent technologies for man-machine interfaces or robots. Notwithstanding, there have been prominent thoughts that the next 'big science' project might occur in neuroscience:

'The Human Genome Project... will lead the way to perhaps the determination of the complete wiring diagram of the human brain, perhaps the next large-scale biology project following in the footsteps of human genomics.' (Evans 1998, p. 1290.)

'If sufficient funding were provided there is no reason why the whole of the large-scale connectivity of the macaque monkey should not be worked out...' (Crick 1999, p. 2022.)

Klaas Stephan and Claus Hilgetag provided helpful comments on this manuscript; the German Research Council, the Wellcome Trust and the EU Thematic Network on Computational Neuroscience and Neuroinformatics were instrumental in its preparation. I am grateful for the enthusiasm of the contributors to this theme issue, the thoughtful comments of the referees, and the unfailing support by the editors and the staff of the *Philosophical Transactions* at The Royal Society.

REFERENCES

- Aldhous, P. 2000 Prospect of data sharing gives brain mappers a headache. *Nature* **406**, 445–446.
- Anonymous 2001 Free and public expression. *Nature* **410**, 851.
- Beltrame, F. & Koslow, S. H. 1999 Neuroinformatics as a mega-science issue. *IEEE Trans. Inf. Technol. Biomed.* **3**, 239–240.
- Bota, M. & Arbib, M. A. 2001 The NeuroHomology database. In *Computing the brain: a guide to neuroinformatics* (ed. M. A. Arbib & J. Grethe), pp. 337–351. San Diego, CA: Academic Press.

- Burns, G. A. P. C., Stephan, K. E., Ludäscher, B., Gupta, A. & Kötter, R. 2001 Towards a federated neuroscientific knowledge management system using brain atlases. *Neurocomputing* **38–40**, 1633–1641.
- Butler, D. 2001 Genomics: are you ready for the revolution? *Nature* **409**, 758–760.
- Casey, B. J. (and 11 others) 1998 Reproducibility of fMRI results across four institutions using a spatial working memory task. *NeuroImage* **8**, 249–261.
- Chicurel, M. 2000 Databasing the brain. *Nature* **406**, 822–825.
- Conley, E. C. 1998 Internet information on ion channels: issues of access and organization. *Methods Enzymol.* **294**, 704–731.
- Crick, F. 1999 The impact of molecular biology on neuroscience. *Phil. Trans. R. Soc. Lond. B* **354**, 2021–2025.
- Evans, G. 1998 The human genome project: applications in the diagnosis and treatment of neurologic disease. *Arch. Neurol.* **55**, 1287–1290.
- Felleman, D. J. & Van Essen, D. C. 1991 Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47.
- Frishman, D., Heumann, K., Lesk, A. & Mewes, H.-W. 1998 Comprehensive, comprehensible, distributed and intelligent databases: current status. *Bioinformatics* **14**, 551–561.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D. & Frackowiak, R. S. J. 1995 Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210.
- Hamming, R. W. 1962 *Numerical methods for scientists and engineers*. New York: McGraw-Hill.
- Hasty, J., McMillen, D., Isaacs, F. & Collins, J. J. 2001 Computational studies of gene regulatory networks: in numero molecular biology. *Nature Rev. Genet.* **2**, 268–279.
- Koslow, S. H. 2000 Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neurosci.* **3**, 863–866.
- Kötter, R. 1999 Trends in European computational neuroscience. *Rev. Neurosci.* **10**, 175–179.
- Kötter, R. & Schirok, D. 1999 Towards an integration of biochemical and biophysical models of information processing: a case study in the nigro-striatal system. *Rev. Neurosci.* **10**, 247–266.
- Kötter, R., Stephan, K. E., Palomero-Gallagher, N., Geyer, S., Schleicher, A. & Zilles, K. 2001a Multimodal characterization of cortical areas by multivariate analyses of receptor binding and connectivity data. *Anat. Embryol. (Berl.)*. (In the press.)
- Kötter, R., Nielsen, P., D-Johnsen, J., Sommer, F. T. & Northoff, G. 2001b Multi-level neuron and network modeling in computational neuroanatomy. In *Computational neuroanatomy: principles and methods* (ed. G. Ascoli). Totowa, NJ: Humana. (In the press.)
- McGonigle, D. J., Howseman, A. M., Athwal, B. S., Friston, K. J., Frackowiak, R. S. J. & Holmes, A. P. 2000 Variability in fMRI: an examination of intersession differences. *NeuroImage* **11**, 708–734.
- Mirsky, J. S., Nadkarni, P. M., Healy, M. D., Miller, P. L. & Shepherd, G. M. 1998 Database tools for integrating and searching membrane property data correlated with neuronal morphology. *J. Neurosci. Methods* **82**, 105–121.
- Persidis, A. 1999 Bioinformatics. *Nature Biotechnol.* **17**, 828–830.
- Reggia, J. A., Rupp, E. & Glanzman, D. 1999 *Disorders of brain, behavior and cognition: the neurocomputational perspective*. Amsterdam: Elsevier.
- Roland, P. & Zilles, K. 1996 The developing European computerized human brain database for all imaging modalities. *NeuroImage* **4**, S39–S47.
- Roos, D. S. 2001 Bioinformatics. Trying to swim in a sea of data. *Science* **291**, 1260–1261.
- Scannell, J. W., Blakemore, C. & Young, M. P. 1995 Analysis of connectivity in the cat cerebral cortex. *J. Neurosci.* **15**, 1463–1483.
- Scannell, J. W., Burns, G. A. P. C., Hilgetag, C. C., O’Neil, M. A. & Young, M. P. 1999 The connective organization of the cortico-thalamic system of the cat. *Cereb. Cortex* **9**, 277–299.
- Shepherd, G. M., Mirsky, J. S., Healy, M. D., Singer, M. S., Skoufos, E., Hines, M. S., Nadkarni, P. M. & Miller, P. L. 1998 The human brain project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci.* **21**, 460–468.
- Stephan, K. E., Zilles, K. & Kötter, R. 2000a Coordinate-independent mapping of structural and functional cortical data by objective relational transformation (ORT). *Phil. Trans. R. Soc. Lond. B* **355**, 37–54.
- Stephan, K. E., Hilgetag, C. C., Burns, G. A. P. C., O’Neill, M. A., Young, M. P. & Kötter, R. 2000b Computational analysis of functional connectivity between areas of primate cerebral cortex. *Phil. Trans. R. Soc. Lond. B* **355**, 111–126.
- Talairach, J. & Tournoux, P. 1989 *Co-planar stereotaxic atlas of the human brain*. Stuttgart: Thieme.
- Wooley, J. C. 1999 Trends in computational biology: a summary based on a RECOMB plenary lecture, 1999. *J. Comput. Biol.* **6**, 459–474.
- Young, M. P. & Scannell, J. W. 2000 Brain structure–function relationships: advances from neuroinformatics. *Phil. Trans. R. Soc. Lond. B* **355**, 3–6.
- Young, M. P., Scannell, J. W., O’Neill, M. A., Hilgetag, C. C., Burns, G. & Blakemore, C. 1995 Non-metric multidimensional scaling in the analysis of neuroanatomical connection data and the organization of the primate cortical visual system. *Phil. Trans. R. Soc. Lond. B* **348**, 281–308.