

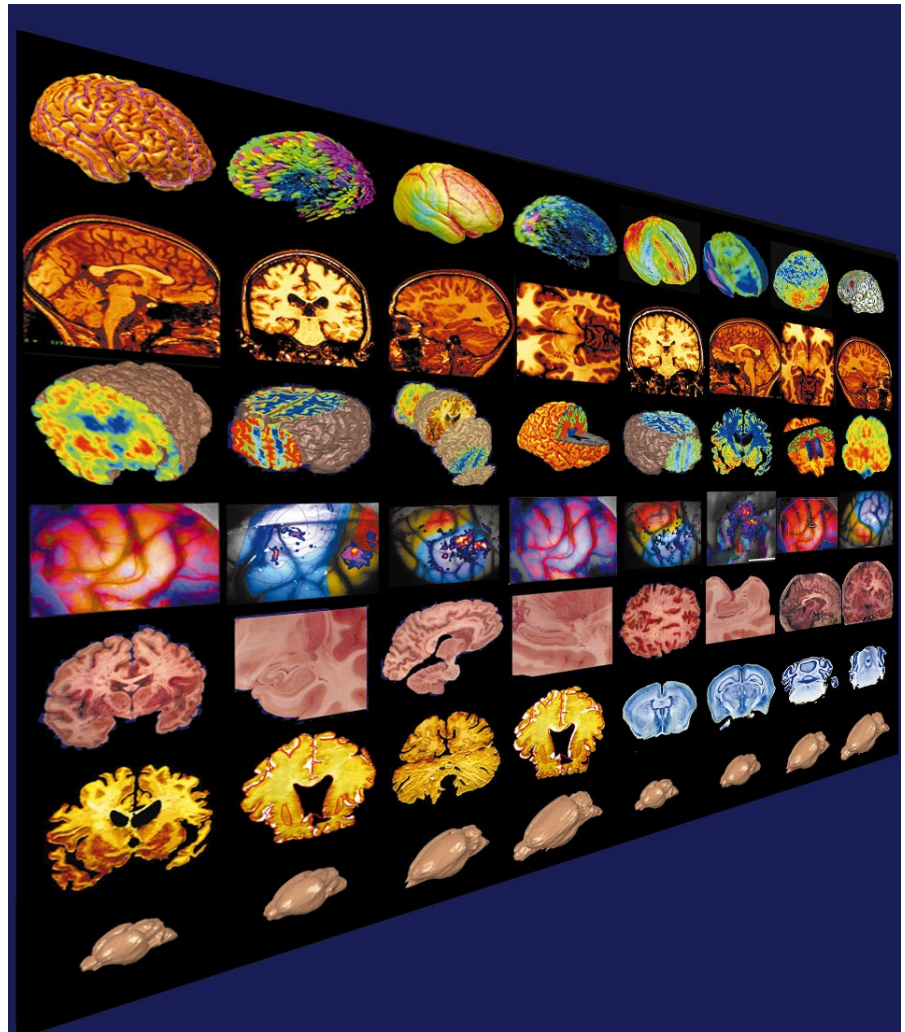
Databasing the brain

Progress in neuroscience might be faster if researchers shared their results in a network of databases. But the technical challenges are huge, and reaching a consensus on what to archive won't be easy, says Marina Chicurel.

Stephen Kuffler probably did not realize what he was starting when he established the Department of Neurobiology at Harvard Medical School in 1966. Kuffler helped found the modern discipline of neuroscience by bringing together physiologists, biochemists and anatomists to focus their efforts on the nervous system.

Today, more than 50,000 neuroscientists worldwide study everything from individual molecules to complex behaviours in species from nematode worms to humans. Generating enough data to fill more than 300 journals, they have created one of the largest, most unwieldy datasets in science. Several prominent neuroscientists are now arguing that the time has come to tame this monster. They believe that progress could be boosted by creating interoperable databases, allowing researchers to share their results and make links between data from labs around the world. "It's a topic that needs to be spotlighted," says Dennis Choi of the Washington University School of Medicine in St Louis, who is president of the Society for Neuroscience.

But there are three big obstacles. First, reaching a consensus on what is worth including in databases. Second, the techni-



ANDREW LEE, PAUL THOMPSON & ARTHUR TOGA

cal difficulty of collating and relating such disparate types of information. And third, the reluctance of researchers who have traditionally guarded their results jealously to embrace data sharing. "It's a data-hugging community," observes Michael Arbib, director of the Brain Project at the University of Southern California in Los Angeles, which is developing a variety of neuroscience databases.

Where to begin?

Bioinformatics is most advanced in genomics and proteomics, where DNA and protein sequences are routinely archived in public databases. Indeed, many journals will only consider papers if authors make their sequences public. But linear sequence data are easy to store in this way.

"Technically, the genome and protein databases are relatively trivial," says Stephen Koslow, director of the Office of Neuroinformatics at the National Institute of Mental Health in Bethesda, Maryland. "Neuroscience data are much more complex." But in the latest issue of *Nature Neuroscience*, Koslow argues that the technical obstacles can be overcome, and makes a plea for more sharing of data¹.

Koslow and others point to projects that are showing the way forward. Many of these are supported by the Human Brain Project, a multi-agency US government initiative established in 1993 to support research into databases and related tools for neuroscientists. Bioinformatics enthusiasts note the scientific advances made possible by such databases (see 'Making the connections', opposite). And they are encouraged by the Organisation for Economic Cooperation and Development's (OECD's) recent acceptance of a proposal to set up a working group on neuroinformatics, chaired by Koslow, which aims to establish guidelines for neuroscience databases and software, and provide an Internet portal for their dissemination².

But shovelling data into a database is only useful if they can be organized in meaningful ways. And in some areas, it is still unclear whether neuroscientists know enough to sort the wheat from the chaff. Geneticists agreed long ago on the value of storing reproducibly generated DNA sequences, but not images of their sequencing gels. Many of neuroscience's subdisciplines have yet to reach a consensus on what is worth storing. In addition, techniques that neuroscientists use to generate and analyse data are often not



Technically, the genome and protein databases are relatively trivial. Neuroscience data are much more complex.

Stephen Koslow.

standardized. "Part of this process must be to confront the devils early in the game," says Choi. "The devils include the possibility of prejudice overly influencing future thought, essentially by controlling the shape of the database or the priority given to certain data over others."

Most database developers agree that flexibility is the key. Daniel Gardner of the Weill Medical College of Cornell University in New York is developing the Cortical Neuron Net Database, a repository for neurophysiological recordings from the brain's cortex. Trying to ensure that the database keeps pace with the field, Gardner's team is developing a hierarchical classification scheme that allows for new categories to be introduced without making earlier entries obsolete or restricting searches. "It's never economically possible to go back and re-classify old databases," says Bruce Schatz, an information scientist at the University of Illinois at Urbana-Champaign. "It just costs too much."

Lack of standardization is a persistent problem, and the confused nomenclature of neuroanatomy is a particular bugbear. Given that the same brain region can go by the name of the caudate nucleus or nucleus caudatus, or may be referred to as part of a larger structure known as the basal ganglia, only experts can reliably sort the data. Worse still, different neuroanatomists sometimes use the same names to refer to different structures. "The actual business of data collation is a hugely, hugely onerous thing," says Malcolm Young, a systems neuroscientist at the University of Newcastle-upon-Tyne in England. "It's more onerous in my experience than doing experimental neuroscience."

Some scientists are turning to computers for help. Researchers led by Rolf Kötter at the Heinrich Heine University in Düsseldorf, Germany, have developed an algorithm called Objective Relational Transformation (ORT)³ for their database of cortical connectivity in the macaque brain, known as CoCoMac. This algorithm transforms neuroanatomical data from one representational format, or mapping scheme, into another. By applying ORT to data in their original formats, individual database users can convert the data into any format they wish to work with. "Besides reducing costs, the approach is more objective, reproducible, transparent and correctable than human labour," says Kötter.

There is also more disagreement over the

interpretation of neuroscience data than there is for DNA sequences. Gordon Shepherd and his colleagues at Yale University in New Haven, Connecticut, are working on this problem. Their SenseLab database integrates information on the olfactory system, including the sequences of olfactory receptor proteins and the electrical properties of the neurons involved in processing smells.

SenseLab flags controversial data through annotations that describe conflicting studies and point users to the primary literature.

Other database architects are designing filters to let users decide how to weight the data. At the University of Southern California, Gully Burns is constructing a database on neural connectivity in the rat brain called NeuroScholar. Future users will be able to score studies based on attributes such as the journal they appeared in or the techniques the investigators used. They will then be able to combine the weighted data from several studies to assess the strengths of hypotheses they might wish to test.

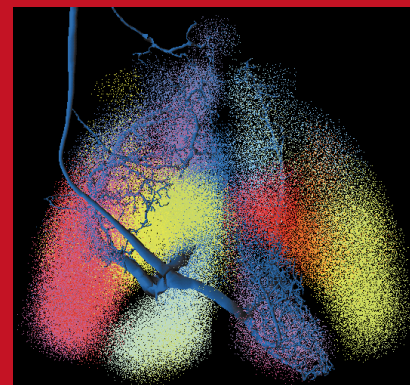
Other researchers agree that approaches such as this, which allow users to impose their own organizational schemes on the data are, in theory, the best solution. James

Making the connections

Why bother to build neuroscience databases? Ask a bioinformatics enthusiast, and he or she will fire back a handful of success stories that provide a glimpse of the future.

The International Consortium for Brain Mapping, which has built a database of brain images from 7,000 people⁸, is one example. As human brains vary with age and between individuals, it is impossible to create an absolute atlas of structure and function. But by using sophisticated algorithms to align a large numbers of images, the consortium has generated atlases that show the probability that any given region of the brain is involved in the performance of a certain task, or is disrupted by a particular disease. Atlases that highlight regions commonly affected by Alzheimer's disease and stroke, for example, are becoming powerful tools for tracking the effects of candidate therapies in clinical trials. John Mazziotta of the University of California at Los Angeles, who heads the consortium, argues that the database promises to become even more powerful as researchers begin to match up the brain images with additional information such as genomic, neurochemical and behavioural data.

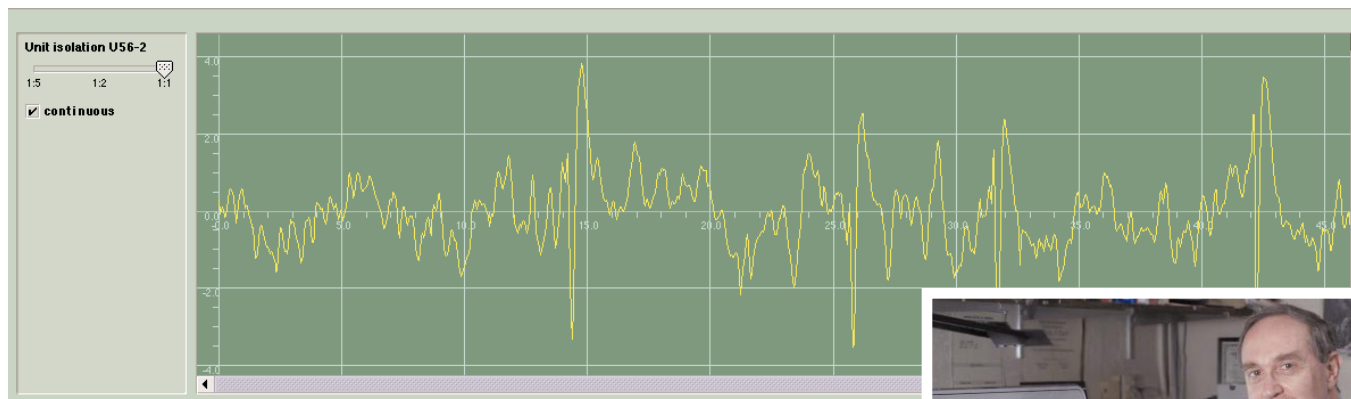
Meanwhile, Malcolm Young at the University of Newcastle-upon-Tyne is gaining new insight into how the brain is hooked up by mathematically sorting through thousands of anatomical studies. It started in 1990, when Young set eyes on a global diagram of the hundreds of pathways that interconnect the visual cortex, drawn up by David Van Essen of the Washington University School of Medicine in St Louis⁹. He soon realized that by applying mathematical analyses to the collated data, he could mine it for valuable nuggets of information¹⁰. Since then, Young has developed additional methods, successfully predicted new relationships between structure and function in the brain, and is helping to spawn a new field of research in which the analysis of anatomical



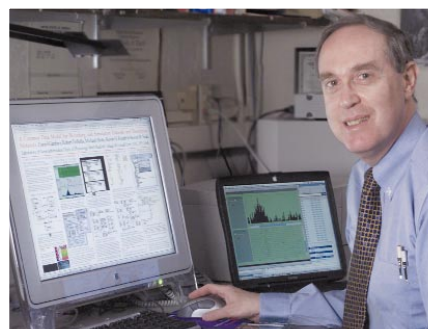
Plugged in: a NeuroSys image of a neuron's inputs. Each colour represents the sensing of air motion from a particular direction.

data is helping researchers to localize neurons with specific electrophysiological properties¹¹.

Bringing together electrophysiological and anatomical data on more than 200 neurons, Gwen Jacobs at Montana State University in Bozeman is discovering how networks of neurons encode sensory information. Neuroscientists are often limited to studying small numbers of cells at a time, rarely seeing the big picture of how larger neural networks transmit complex messages. But by focusing on a simple system and setting up NeuroSys, a network of databases that allows her to integrate data from multiple neurons, Jacobs is overcoming this limitation. "It came out of the frustration of trying to hold ten different variables in your head at the same time," she says. Jacobs has collected data on the behaviour and shape of neurons that are involved in detecting air motion in crickets. Plugging her findings into NeuroSys, she has discovered that the coding system used to specify the direction that a puff of air has come from seems to depend on the branching patterns of a small set of key neurons¹².



Virtual oscilloscope: Gardner's invention lets researchers view neurophysiological recordings.



Bower at the California Institute of Technology in Pasadena, for example, is developing model-based database systems whose organization relies almost exclusively on the user. Bower's GENESIS Database holds data that range from the characteristics of ion channels in cell membranes to the details of the connections between neurons. But these data only become usefully organized when a user creates a model of a neuron or network of neurons.

Super models

For example, a user may devise a model of a Purkinje cell — a type of neuron found in the cerebellum — to predict how electrical signals from its projections, or dendrites, will be transmitted to the cell body. The model will then sort relevant data from the database, and help the researcher spot inconsistencies. If the model is not compatible with known information, or if data are contradictory, the model will alert the researcher to the problem. And because the models are formal representations of hypotheses, they provide a way for neuroscientists to communicate and compare their ideas quantitatively.

More problems are posed by data that vary in three dimensions. Like a crumpled beach ball, the human brain's surface is covered with folds, serving as natural landmarks for neuroanatomists comparing results. But

the distribution of folds varies between individuals, as does their location relative to functional brain areas. Some researchers have resorted to shrinking or expanding their images to align them with different-sized brains. But brains often vary in more complicated ways than size. So several researchers have developed more sophisticated warping algorithms. David Van Essen of the Washington University School of Medicine, for example, has created algorithms that flatten the surface of the brain. These reduce the alignment problem to two dimensions, and have been incorporated into Van Essen's software toolkit and database of cortical structure and function, the Surface Management System.

Neurophysiologists, meanwhile, generate large amounts of time-series data as they record the electrical behaviour of neurons. Michael Gabriel and his colleagues at the University of Illinois have developed a time series data protocol which has been incorporated into the Neuronal Time Series Analysis Workbench, a set of informatics tools designed to help researchers who work with such data. Like Kötter's ORT, the protocol is associated with translational filters that allow users to work with data in a variety of formats. Another tool for neurophysiologists is a virtual oscilloscope, devised by Gardner at Cornell⁴. This allows electro-physiologists to see extended datasets — as opposed to the small, static excerpts that appear in published articles — as well as artificial traces predicted from simulations.

Neuroscientists often have to wrestle with huge computer files. An image of a single section of monkey brain can take up 80 megabytes, says Edward Jones of the University of California at Davis, who is building brain atlases of such high resolution that they include individual neurons. A typical atlas of a monkey brain includes 1,500 sections, so his files often fill several gigabytes. Functional magnetic resonance imaging (fMRI) datasets pose similar problems. Having access to servers that can handle such monstrous files and ship them across the Internet is problematic. But computer

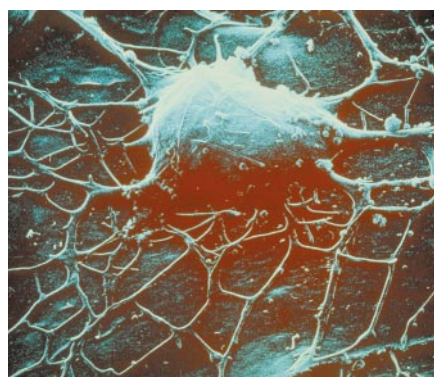
experts such as Schatz say that computer power and Internet bandwidth should soon cease to be limiting factors.

That should assist neuroscientists in tackling perhaps the most difficult challenge of all: linking their various databases into a seamless federation. Ideally, says Schatz, users should be able to navigate between databanks without noticing borders, as if interacting with a single, cohesive database. But this has proved difficult, even in fields with much simpler datasets. It took years to connect the European Molecular Biology Laboratory's DNA sequence database with the SWISSPROT protein sequence database, because researchers could not agree on nomenclature.

Learning to share

Many database developers believe the best approach is to standardize the communication between databases, rather than standardizing the data they hold. One helpful software trick is to 'wrap' data with universal labels that describe their composition and format so they can be understood by different types of software, running on different computers — the HTML code that underlies web pages is one such wrapper. Gardner is developing an interface to link different databases using the Biophysical Description Markup Language, a wrapper that he designed specifically to label neuroscience data, based on a code called XML. Schatz believes another key tool will be concept-switching, a search strategy that relies on analysing the contextual relationships between phrases to identify underlying concepts^{5,6}. Concept-switching algorithms, for instance, would identify the term 'nucleus caudatus' used in one database as occurring in similar contexts as 'basal ganglia' in another, and link them together.

But the technical issues are not the only



Networking: databases allow users to model the ways that neurons link together.

SPR

ones that need to be ironed out. In a highly competitive field, in which data are often hard-won, many researchers are reluctant to release their data to be exploited by others. Some neuroscientists, such as Peter Fox at the University of Texas Health Science Center in San Antonio, also argue that the complexity of the data poses problems. When Fox sent primary brain-imaging data to a collaborating lab, he says he had to be in almost continuous communication to explain how he had calibrated his machines, subtracted noise and displayed the data. If primary data were freely available, Fox fears that some researchers might be led to the wrong conclusions. "The submission of truly raw data would lead to a nightmare of misinterpretation," he says.

The sensitivity of the data-sharing issue was underlined last month, with the opening of a new repository for fMRI data at Dartmouth College in Hanover, New Hampshire. A letter sent to leading researchers in the field announcing the National fMRI Data Center by its director Michael Gazzaniga raised fears that, in the future, journals might require primary data relating to papers on fMRI to be deposited in the database⁷. Gazzaniga sent the letter in his capacity as editor of the *Journal of Cognitive Neuroscience*, which had decided to adopt such a policy. That prompted several dozen fMRI specialists to sign a letter arguing against mandatory data submission, which was sent to the data centre's financial backers and to the editors of 14 leading journals.

Although it may not be universally applicable, Yale's Shepherd has devised a way of encouraging neuroscientists to share



Flat pack: Van Essen irons out the brain's wrinkles to make mapping and comparisons easier.

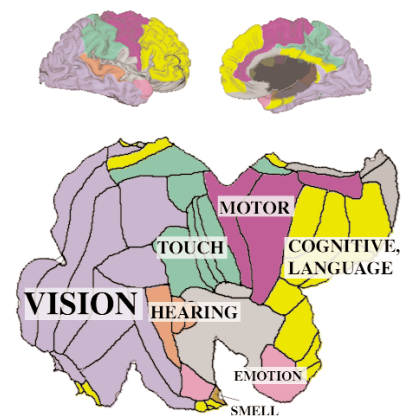
without losing control of their data. The olfactory receptor database within Shepherd's SenseLab allows users to deposit unpublished sequences, which are then kept hidden from other users' view. When a search for sequence homologies finds a match with one of these unpublished sequences, the database provides the searcher with the contact information for the researchers who submitted the sequences. In this way, researchers can avoid revealing their primary data, yet benefit from identifying potential collaborators.

The task ahead

Koslow accepts that it may take some time before neuroscientists adjust to a culture of data sharing. But he argues that the complexity of data do not pose insurmountable problems. For example, he says, details of the experimental conditions and variables needed to interpret primary data could be incorporated into databases. Koslow believes that, ultimately, neuroscientists must be allowed to come to a consensus on the desirability of data sharing, and that attempts by funding agencies or journals to force the pace could be counterproductive.

When it comes to building the basic infrastructure, however, Koslow is keen to push the pace. But experts disagree on how to move forward. Koslow wants to expand the existing approach of funding a wide variety of small projects, and then building links between those that prove most successful — and is pushing this model in the OECD working group. Schatz is in favour of more centralization. In his view, small projects based in individual labs, often relying on graduate students who are amateurs in informatics, are likely to fail. He thinks the big problems, such as how to form database federations, have to be tackled by experts working in dedicated informatics centres. Based on his experience in the early days of genomics, Schatz proposes getting professional informaticians to build a network of databases focusing on a particular organism and a set of well-studied brain regions. He

Functional Subdivisions of Human Cerebral Cortex



Van Essen & Drury, 1999

also believes the field needs definite targets against which to measure progress.

As the experts debate the way forward, some neuroscientists are sceptical about investing heavily in databases. Says Rodney Douglas, a computational neuroscientist at the Institute of Neuroinformatics* in Zurich: "Is it really a good thing for us to spend a lot of money on cataloguing everything in sight?" Douglas argues that the priority in neuroscience should be to understand how nervous systems deal with information. "The nervous system doesn't collect a lot of data, it collects relevant data, and I think we can learn from that," he says.

The best answer to the sceptics would be to demonstrate the practical advantages that databases can bring. But Koslow, who chairs the Human Brain Project's coordinating committee, concedes that the project's overall impact has so far been modest. The challenge for the bioinformaticians is to produce some tools that can make a real difference to working neuroscientists. "I became convinced early on of the need to build databases," says Van Essen. "But I must admit I didn't realize how challenging a task that would be."

Marina Chicurel is a freelance writer in Santa Cruz.

For an online debate on data sharing in neuroscience, see www.nature.com

1. Koslow, S. H. *Nature Neurosci.* **3**, 863–866 (2000).
2. Smaglik, P. *Nature* **405**, 603 (2000).
3. Stephan, K. E., Zilles, K. & Kötter, R. *Phil. Trans. R. Soc. Lond. B* **355**, 37–54 (2000).
4. Gardner, D., Abato, M., Knuth, K. H., DeBellis, R. & Erde, S. M. *Phil. Trans. R. Soc. Lond. B* (in the press).
5. Butler, D. *Nature* **405**, 112–115 (2000).
6. Schatz, B. R. *et al. IEEE Computer* **32**, 51–59 (1999).
7. Aldhous, P. *Nature* **406**, 445–446 (2000).
8. Mazziotta, J. C. *et al. Neuroimage* **2**, 89–101 (1995).
9. Felleman, D. J. & Van Essen, D. C. *Cereb. Cortex* **1**, 1–47 (1991).
10. Young, M. P. *Nature* **358**, 152–154 (1992).
11. Young, M. P. & Scannell, J. W. *Phil. Trans. R. Soc. Lond. B* **355**, 3–6 (2000).
12. Jacobs, G. A. & Theunissen, F. E. *J. Neurosci.* **20**, 2934–2943 (2000).

*In the United States, the term neuroinformatics generally refers to the application of databases and related tools to neuroscience. In Europe, it is sometimes used to describe the discipline of computational neuroscience, which models the behaviour of neurons and neural networks.

Web links:

Human Brain Project

▶ <http://www.nimh.nih.gov/neuroinformatics/index.cfm>

University of Southern California Brain Project

▶ <http://www-hbp.usc.edu>

Cortical Neuron Net Database

▶ <http://cortex.med.cornell.edu>

CoCoMac

▶ <http://www.cocomac.org>

SenseLab

▶ <http://ycmi.med.yale.edu/senselab>

NeuroScholar

▶ http://www-hbp.usc.edu/Projects/neuroScholar_Connx.htm

GENESIS Database

▶ <http://www.bbb.caltech.edu/hbp>

Surface Management System

▶ <http://stp.wustl.edu/sums>

Neuronal Time Series Analysis Workbench

▶ <http://soma.npa.uiuc.edu/isnpa/isnpa.html>

International Brain Mapping Consortium

▶ <http://www.ionu.ucla.edu/ICBM/index.html>

NeuroSys

▶ <http://nervana.montana.edu/projects/neurosyst>