# Bringing fly brains in line

Wolf Huetteroth & Scott Waddell

Software for fast and accurate alignment of brain images is used to generate a partial brain atlas for *Drosophila melanogaster* and should enable circuit mapping.



Comparative brain anatomy is an old field that has undergone a major technology-driven boom in the last decade. The synergy of confocal microscopy with ever-improving and more affordable computer performance has led to the construction of several three-dimensional 'standard brains', formed by averaging brain images taken from multiple individual animals[1]. In this issue of *Nature Methods*, Peng *et al.*[2] use brains from *Drosophila melanogaster* to demonstrate the utility of their latest computational tool, BrainAligner, to align large numbers of brain images with high quality and speed (**Fig. 1**).

Early attempts to generate standard brains often required the sample brain images to be manually annotated before registration, which is both hugely time-consuming and prone to human error. Computer algorithms were subsequently developed to automatically extract common features from collections of images, but they required enormous computational power and time, and the standard brains generated were of variable quality.

The freely available BrainAligner software combines and optimizes some of these pre-existing methods to provide fast and accurate automatic registration of brain images and should be useful to most investigators interested in neural circuit construction. In addition, the potentially high-throughput nature of BrainAligner allows one to efficiently handle large amounts of anatomical data, which is essential for grand-scale anatomy based screening.

In the paper published in this issue, Peng *et al.*[2] use BrainAligner to construct a 'standard' or 'target' fly brain by statistically averaging images from 295 fly brains labeled with an antibody that recognizes a general neural process marker (the nc82 antibody that labels the presynaptic protein bruchpilot). They then manually annotated the target fly brain to add 172 conserved brain compartment positions as 'landmarks'. A partial brain atlas could then be constructed by automatically aligning each of 470 discrete expression patterns from 2,954 sample brains, into the target brain, guided by the landmarks in the nc82-stained neural tissue.

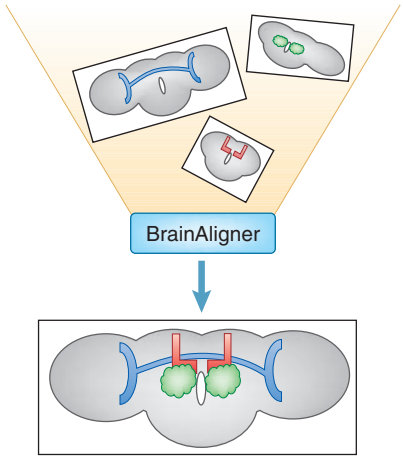


**Figure 1** | Schematic showing that images of various parts of the fruit fly brain can be aligned with BrainAligner software to generate a neural map.

BrainAligner selectively aligns each sample using only the most reliable landmarks, allowing it to maintain accuracy despite variable nc82 staining quality, partial tissue damage and image distortion. Dropping landmarks that fall outside the normal statistical variance led to quantifiably less error-prone warping of samples (expressed as the percentage of landmarks used for each alignment).

The authors improved alignment speed 50-fold by implementing hierarchical interpolation to generate the final warping field. In other words, the pixel resolution in each dimension (that is, the voxel resolution) of the image stack was first reduced or 'downsampled' by four, and after warping, the original resolution was approximated by interpolating the remaining voxels. This allowed two three-dimensional image stacks of 1,024 × 1,024 × 256 voxels to be aligned in about 40 minutes on a standard computer. Therefore, using BrainAligner and our desktop computer it would take 90 days to repeat the 3,248 alignments performed by Peng *et al.*[2]—a remarkable 12-year improvement in processing time over previous methods.

What might your average neurobiology researcher use BrainAligner for? An obvious application would be an automatic search of brain images for enhancer trap lines that express in overlapping brain regions. Although neural cell body position is highly variable, BrainAligner can find similar neurons using their primary neural tracts.

One might also wish to construct one's own atlas of the entire brain or of neurons comprising individual regions of the brain. This can be done by double-staining brains with nc82 and aligning the new images with the target brain constructed by Peng *et al.*[2] or by constructing one's own target brain—with landmarks generated using nc82 or another robust antibody—and then aligning new samples using that same set of reference points. The key is that once a specified 'target' brain with reference label is established, one can align images to it so long as the program can retrieve some of the landmarks in the target. If the standard developed by Peng *et al.*[2] was adopted by the entire community, we could all in principle log our expression patterns to a common reference.

Identification of neural connectivity would be a big bonus of such a brain atlas project, but realistically it can only indicate neurons that are putative synaptic partners. Nevertheless, narrowing down to such putative partners for functional validation would expedite the process of circuit mapping. Although a map of physical connections

Wolf Huetteroth and Scott Waddell are in the Department of Neurobiology, University of Massachusetts Medical School, Worcester, Massachusetts, USA.
e-mail: scott.waddell@umassmed.edu

can be useful, a full understanding of circuit function requires additional types of knowledge, for instance, about the neurotransmitters involved, the electrical properties of component neurons and the influence of modulatory systems.

Will BrainAligner become the software of choice for the community? History suggests that will depend on more than alignment quality and speed. BrainAligner is freeware, and it is well integrated with the V3D and AtlasViewer freeware developed by the same group[3]. Price is therefore not an issue, but documentation and technical support, platform dependence (BrainAligner is currently available in Macintosh and Linux formats) and the availability of updates could be. Furthermore, there are other promising ventures in various states of development such as Flybrain@Stanford[4], BrainGazer[5] and FlyCircuit[6] that have similar goals, so only time will tell. Let the bidding begin!

**COMPETING FINANCIAL INTERESTS**
The authors declare no competing financial interests.

1. Ito, K. *Front. Syst. Neurosci.* **4**, 26 (2010).
2. Peng, H. *et al. Nat. Methods* **8**, 493–498 (2011).
3. Peng, H., Ruan, Z., Long, F., Simpson, J.H. & Myers, E.W. *Nat. Biotechnol.* **28**, 348–353 (2010).
4. Jefferis, G.S. *et al. Cell* **128**, 1187–1203 (2007).
5. Bruckner, S. *et al. IEEE Trans. Vis. Comput. Graph.* **15**, 1497–1504 (2009).
6. Chiang, A.S. *et al. Curr. Biol.* **21**, 1–11 (2011).

# Channeling the data deluge

**Jason R Swedlow, Gianluigi Zanetti & Christoph Best**

With vast increases in biological data generation, mechanisms for data storage and analysis have become limiting. A data structure, semantically typed data hypercubes (SDCubes), that combines hierarchical data format version 5 (HDF5) and extensible markup language (XML) file formats, now permits the flexible storage, annotation and retrieval of large and heterogenous datasets.

Biological research laboratories were once occupied by scientists whose main tools of the trade were the pipette, lab notebook, calculator and pen. Twenty-five years of automation and feats of engineering have revolutionized biology into a data-centric science, the best example being certainly the genome projects whose output is now the foundation of essentially all modern biological experiments. These projects were undertaken in a relatively few central facilities, which—after some negotiation—agreed to release their data within one day of collection using standardized formats. Today, most modern labs have access to sophisticated data generation and analysis systems that routinely generate similar amounts of data each day, all of which must be processed and analyzed to reveal biological understanding. In stark contrast to genomics, these data are produced locally by many individual scientists, but the overall scale and heterogeneity of these experimental efforts create a barrier to easy standardization: a data format that suffices for one lab will very likely only partially address the needs of another. When experimental design and outcome drive the data formats, straightforward standardization becomes nearly impossible. This priority is correct; scientific achievement should drive data formats and not vice versa.

Heterogeneity, however, comes with a considerable cost. Data generated in one lab cannot be analyzed by researchers in another, and data analyzed using one software tool often cannot be analyzed with another tool (even in a single lab). Reverse-engineering data formats is slow, time-consuming, error-prone and certainly scales poorly with the diversity of experiments. At the same time, although scientific data formats do not themselves enable discovery, they are a powerful enabling technology. Without the Genbank and Protein Data Bank (PDB) repositories, much of today's research would be impossible.

Seen in this context, weaning bench scientists from storing their data in randomly formatted spreadsheet files seems not only useful but scientifically valuable. This issue has not been lost on funding agencies such as the US National Science Foundation with its Office of Cyberinfrastructure, the UK Research Councils with their eScience program and the European Union, which funds several 'e-Infrastructures' in its FP7 program and recently commissioned a high-level expert group to report on the handling of scientific data[1].

Scientific data formats always involve a tradeoff between simplicity and flexibility. Some of the most useful formats, for example, the comma-separated values (CSV) spreadsheet or PDB and Genbank files, have a simple, line-oriented structure that is easy to process without extensive programming. But there are limits to these structures that force the use of awkward workarounds (for example, splitting a PDB file to accommodate more than 99,999 atoms in a ribosome). In this issue, Millard *et al.*[2] show that by leveraging well-established computer science tools and high-performance computing it is possible to build a simple data storage system that can efficiently and flexibly manage data coming from high-throughput imaging.

One tool they use is the hierarchical data format version 5 (HDF5), which works as a flexible vessel to efficiently store large arrays of numerical data along with textual metadata within a single file structure. HDF5 was first developed by the US National Center for Supercomputing Applications in the early 1990s as a flexible and efficient file format for large numerical datasets arising mainly in high-performance computing. An HDF5 file provides the flexibility of a file system: a single file can hold many different types of data, and arbitrary access to data elements within large matrices and datasets is supported. HDF5 is a sophisticated technology, but many open-source tools are now available that provide easy cross-platform access, making HDF5 a tool that can be used easily across scientific disciplines. As the needs for data have grown across the sciences, so has the readiness to accept the complexities of HDF5 for its flexibility and efficiency.

However, HDF5 has only limited capabilities to express nonnumerical information, such as metadata and experimental setups. Millard

JJason R. Swedlow is at the Wellcome Trust Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee, Scotland, UK. Gianluigi Zanetti is at the Center for Advanced Studies, Research and Development in Sardinia, Pula, Italy. Christoph Best is at Google UK Ltd., London, UK.
e-mail: jason@lifesci.dundee.ac.uk