Modelling data across labs, genomes, space and time

Jason R. Swedlow, Suzanna E. Lewis and Ilya G. Goldberg

Logical models and physical specifications provide the foundation for storage, management and analysis of complex sets of data, and describe the relationships between measured data elements and metadata — the contextual descriptors that define the primary data. Here, we use imaging applications to illustrate the purpose of the various implementations of data specifications and the requirement for open, standardized, data formats to facilitate the sharing of critical digital data and metadata.

THE DATA PROBLEM IN BIOLOGY: IT'S NOT JUST SIZE

How we long for simpler times when the body of work produced by a postdoctoral fellow was contained in a stack of bound notebooks, labelled with roman numerals down the spine, filled with dried gels, exposed autoradiography films and faded thermal printer tapes from a gamma counter. Compare this with the current cohort of students and fellows — 'notebooks' are now laptop computers with mobile hard disks filled with images, spectra, sequences and other files. While a previous generation cut its teeth on spreadsheets, familiarity with data modelling languages (such as extensible markup language, XML; and unified modelling language, UML) and analysis tools (such as R, MATLAB, Octave and Scilab) is becoming a critical skill for the next generation of students.

At first glance, it may seem that the differences between these generations are cosmetic —it's the same data, just stored on different media. However, closer inspection reveals two fundamental changes that mandate a new approach to the management and use of scientific data. The first is a massive change in the scale of the problem. Previously, a northern blot reported the expression of a single gene across a panel of samples. Now, a microarray assay allows thousands of genes or gene products to be monitored in a single experiment[1,2]. This scaling has occurred across all biological disciplines, making it impossible to productively store experimental results in anything but digital form. As the scale of these assays grows, so does the need for managing the data they produce.

The second, and more profound, change in modern biological data is dimensionality — it is now routine to obtain measurements from the same cell or tissue across space, time and spectral range. The most common example of this process is digital imaging — especially in the form of molecular assays in cells and tissues using fluorescence microscopy (so called "five-dimensional imaging")[3–6]. The definition and calibration of each dimension (for example, μm voxel^{-1}, msec image^{-1} and nm channel^{-1}) are all key parameters for interpreting and analysing images, and producing physical measurements of the dynamic behaviour and cellular constituents. With the advent of multiplexed "high-content" imaging assays, where the localization of cellular components (for example, a fluorescently labelled protein) is assayed by imaging in an array that samples different small-molecule inhibitors or siRNAs, there is a further increase in the dimensionality and scale of the biological data produced[7–10]. Analysis of multiplexed and multidimensional data requires facilities for managing not only assay results, but also the links to any other biological data that define the sequence, cell or tissue that is being studied (for example, the name of the gene and the sequence for each image or microarray spot). Clearly, systems analysis and the use of distributed tools through web services is impossible without such links.

Thus 'more data' actually means a broader sampling of the behaviour and dynamics of biological mechanisms. As systems biology matures, such multidimensional data is critical, as it forms the foundation for the quantitative models and hypotheses that describe the spatial and temporal dynamics of biological systems. However, this analysis will be impossible without the assay metadata — the contextual descriptions of the measured data. In imaging, metadata provide definitions that document the dimensionality, scale, imaging technique, contrast method (that is, what is being imaged) and other parameters that are critical for understanding what is being imaged and to interpret the data. Moreover, metadata can include critical acquisition instrument settings and calibrations that ensure validation and reproducibility. Furthermore, user annotations, or other manual measurements and filters, are often critical components of the analysis workflow. All of this contextual metadata must be accessible and, wherever possible, stored in a coherent manner, linked to the actual measurements.

SOLUTIONS FOR DATA MANAGEMENT

Definitions of some of the components that are used for managing data are listed in Table 1. Each has different properties and therefore solves different parts of the problem: ontologies are used to describe the semantics (that is, the types of data that exist and the relationships between them); exchange protocols describe how to send and receive the data across the wire, including compression techniques and error checking and recovery methods; and file formats with openly available documentation and software allow the data to be parsed from a file.

Jason R. Swedlow is in the Division of Gene Regulation and Expression, Wellcome Trust Biocentre, Faculty of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK. Suzanna E. Lewis is in the Life Sciences Division, 1 Cyclotron Road MS-64R0121, Lawrence Berkeley National Laboratory, Berkeley, CA 94729, USA. Illya G. Goldberg is in the Image Informatics and Computational Biology Unit, Laboratory of Genetics, National Institute on Aging, National Institutes of Health, 333 Cassell Drive, Suite 4000, Baltimore, MD 21224, USA. e-mail: jason@lifesci.dundee.ac.uk

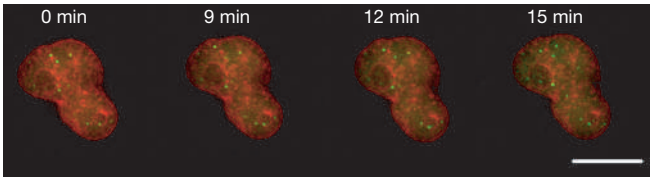

Figure 1 HeLa cell expressing GFP–coilin (green) and histone H2B–YFP (red). The image shows a series of timepoints as two-dimensional projections from a five-dimensional image of a living cell expressing histone H2B and coilin fusion proteins[18]. Some critical metadata necessary for interpreting this image include: pixel size = 0.12 μm; *z*-section spacing = 0.5 μm; interval between individual timepoints = 3 mins; fluorescence exposure time = 0.1 s. To be used by a computer, all of these metadata, and many more, must be expressed in a defined form (for example, http://openmicroscopy.org/XMLschemas/OME/latest/ome_xsd/; ref. 13)

An example of the potential power of these approaches is the Open Biological Ontologies project (OBO; http://obo.sourceforge.net/)[11], which provides resources for controlled vocabularies across different biological and medical domains. These projects were initiated to solve the immediate needs of individual genome projects, but led to data-sharing solutions that are useful across biology. One of the OBO ontologies is the Gene Ontology (GO; http://geneontology.org/), the major goal of which is to define, as much as possible, major molecular pathways and biological processes by describing the types and relationships that are shared by the majority of biological systems. Central to the definition of individual elements are relationships like 'is_a', 'part_of', etc. (see http://obo.sourceforge.net/relationship/). For example, in the Sequence Ontology (SO; http://song.sourceforge.net), 'mRNA' 'is_a' 'processed_transcript' and 'processed_transcript' 'is_a' 'transcript'. These types of relationships allow a computer to conclude that the mRNA sequence is derived from the gene sequence. Although this is obvious to a person, it must be explicitly specified to a computer. The goal of OBO is to provide computable definitions that can be used across all biological systems to describe and understand biological structures and processes.

USER, KNOW THY DATA

In general, assay data and metadata are acquired and then stored for subsequent visualization or analysis. These data cannot be stored as an undefined series of values — they must follow a defined semantic and syntactic structure, or a model that describes data location, type and layout. A semantic model defines the meaning of the data — what it is, where it comes from, its units and potentially any processing or transformations applied to it. A syntactic model describes the relationships between different data. These are organizational tools that enable use and understanding.

Fortunately, it is possible to express semantic and syntactic models in a form that can be parsed directly by a computer, so that the computer can use the data the model describes while preserving its meaning. Unfortunately, computer-parseable semantic models are not used as often as they should be, which is a major cause of the current data crunch. This computer-parseable model of the data can be expressed in a variety of physical forms — a formatted file or a database are two common examples (Table 1 and see also Kersey *et al.* in this issue). Knowledge of this model and its physical representation allow access, use and, hopefully, understanding.

As an example, the time-lapse image shown in Fig. 1 can be processed by iterative deconvolution only with knowledge of the objective lens and fluorescence wavelength used for the imaging. Tracking and velocity measurements require knowledge of the time interval between images — preferably the real times, and not just those programmed into data acquisition software. Finally, the image has multiple channels, so the definition of these channels (red, histone H2B–YFP; green, GFP–coilin) is necessary for the interpretation of the molecular meaning of the image. Each element of the workflow (acquisition → illumination correction → deconvolution → segmentation → tracking) requires a defined model of the inputs and output for each step.

Modelling is more than an exercise in definitions and semantics — it provides a mechanism for communicating data. Invariably, complex datasets must be analysed by multiple software tools. Rather than rewrite software tools so that different representations can be understood, it is much easier to have a single approach that all tools can use. This facilitates analysis within a laboratory and opens the possibility of sharing large datasets between groups and across communities, and ensures that legacy data is not lost once the software that reads it is outdated or is not supported (a common occurrence in the current environment of rapidly evolving operating systems). This approach of sharing data using open, specified formats has been a key component of the success of the genome sequence databases and bioinformatics that are now a part of the daily life of every biologist, and must be extended for all research disciplines. In systems biology, the systems biology markup language (SMBL)[12] is a computer-readable format, derived from XML, for representing models of biochemical reaction networks. SBML is designed to be an exchange format, to enable different software packages to communicate with the same data (see http://sbml.org for a list of software packages that support SBML). This type of capability accelerates discovery, as it allows the use of chains of individual tools and ensures that software modules for reading and writing data can be efficiently reused.

STANDARDS AND BIOLOGICAL DISCOVERY

A key requirement for data sharing is standardization: agreement on the types and definitions of structures and processes, and the formats used to access and share data. Projects like OBO and MGED (http://www.mged.org/Mission/) have made great progress towards these goals. If every term is defined and readable, in a standardized format, then any person and software can understand and use the data. This is the strategy used in the specifications of shareable image file formats, which describe light microscopy data, known as OME–XML and OME–TIFF files[13]. Rather than write critical metadata either as binary code or free-form text, these formats use XML to provide standardized descriptors of light microscopy data and acquisition systems, while supporting local customization.

In practice, the terms that are used to define biological data inevitably reflect current understanding. Standardization using traditional modelling and databases ensures these data can be is universally accessible and understood. However, an existing model or specification may not fully describe all biological systems or new assay methods. In addition, understanding and assay methodology are linked, and systems analysis will be driven by new multiplexed, multidimensional assays. Definitions of new assay methods and deeper understanding of biological systems require new data specifications. New modes of microscopy and mass spectrometry are constantly being developed and the data these new systems generate (and their limits and errors) all evolve. It is thus axiomatic that any data standard, whether used in a single laboratory or across a

Table 1 Examples of data-model implementations for multidimensional biological data

Type	Description	Example(s)	Links and references
Requirements	Free form written description — often a list of minimal data elements	MIAME	http://www.mged.org/Workgroups/MIAME/miame.html (ref. 24)
Protocol	A set of rules governing data exchange between computers	DAS	http://biodas.org/ (ref. 25)
Data model	The information that will be contained in a database, including the types of data and their relationships.	Chado	http://www.gmod.org/?q=node/6
Ontology	Defines vocabulary and logical relationships, producing hierarchies and allowing reasoning	Gene ontology, sequence ontology, MGED ontology	http://obo.sourceforge.net/ http://www.geneontology.org/ (ref. 11) http://song.sourceforge.net http://mged.sourceforge.net/ontologies/index.php (ref. 26)
Database	One or more large structured sets of persistent data that are usually associated with software to update and query the data.	OME, PSLID	http://openmicroscopy.org (ref. 27) http://murphylab.web.cmu.edu/services/PSLID/ (ref. 28) See also Kersey, P. *et al.* in this issue
File format	The structure of a computer document that specifies how information is organized	SBML, DICOM, OME–XML, OME–TIFF	http://sbml.org/ (ref. 12) http://medical.nema.org/; http://openmicroscopy.org/formats/ (ref. 13)

whole community, must be dynamic and responsive. Standardization implies a foundation of well-established shared terms and relationships with mechanisms to adapt to new methods and discoveries.

BALANCING LANGUAGE AND STANDARDS

Natural language (the language people use to talk to each other) is the most expressive and effective way to represent new information. In a computer, natural language is stored as free-form textual annotation. The meaning of free-form text cannot be reliably interpreted by a computer and therefore cannot be queried and shared. How do we progress from free-form textual annotations, to semi-structured annotations, to fully computable representations? Ideally, this would be done by researchers themselves, without the need for data-modelling expertise. Faced with a new observation, a researcher defines a new category and associates it with the data. The advantage over free-text is that when the same observation is made again, the category is already defined and the exemplar is available for comparison. Clearly, this approach captures a user's evolving impressions and supports the further understanding that is driven by new data and experiments. It also allows sharing of data, a user's custom annotations, and even the history of the evolution of those annotations. Implementing tools that use this type of dynamic model and annotation approach is challenging, but systems are now available to support custom annotation of whole images (for example, http://openmicroscopy.org/custom-annotations/; ref. 14). As these annotation structures mature, they can be shared with small groups of interested users and adapted to serve a wider community to ultimately mature and become a community-wide standard.

FROM MANUAL ANNOTATION TO AUTOMATED IMAGE ANALYSIS

Automated object identification and tracking are now commonly used in the analysis of imaging data and provide critical insight into the underlying mechanics of the movements of molecules and organelles in living cells[15–18]. A full study of complex cellular processes not only requires tracking of the centre of mass of an object and measurement of its volume or velocity, but also requires identification of cellular phenotypes and structures. As the nature of cell movements and physiological changes are only now being studied, the specific phenotypes that occur are not easily automatically characterised. The development of automated classification tools that can recognize and distinguish different cellular phenotypes holds great promise[19,20], but these require examples of the specific phenotypes (often called training sets). Unsupervised methods are available that find classes of objects and phenotypes without training sets, but these still require a person to manually evaluate the results to determine those that are of interest.

Timelapse imaging is increasingly used to study the movements of cells within tissues to monitor changes in cell physiology during development[21,22]. The use of automated classification for timelapse data has recently been achieved in images of cultured cells progressing through the cell cycle[7]. More complicated phenotypes, involving large numbers of cells and tissues, ultimately may also be subjected to automatic classification. Supervised methods will certainly need training sets containing user-assigned annotations, but unsupervised methods will also require user based annotation of sets of phenotypes. In either case, tools that support custom annotations help serve as an essential foundation for new image-analysis methods.

THE NEED FOR OPEN, STANDARDIZED SPECIFICATIONS

Biomedical research has increasingly turned to commercial manufacturers to provide the sophisticated digital data-acquisition systems that drive discovery. Invariably, these systems are supplied with proprietary software that manages data acquisition and analysis. In these cases, the data model underlying the proprietary software is often not publicly available, making it difficult to access the acquired data with third-party software. This access is critical: the nature of all sciences, including basic and clinical biomedical research, is the ability to ask a new, previously unimagined question — "What if…?". Addressing a novel question may entail a new software tool or sending data to a distant collaborator who cannot access the proprietary software. Existing tools cannot anticipate the new question or discovery.

As a result, a natural tension has developed between commercial and academic scientists, who need open, standardized data formats, and commercial providers, who need to capitalize on their image-analysis technology. Definition and use of standardized data formats has been achieved in medical imaging (DICOM; http://medical.nema.org/) and

flow cytometry (FCS[23]). Similar standards have not yet been accepted for biological microscopy, high content screening and mass spectrometry, although proposals have been made by the Open Microscopy Environment (http://www.openmicroscopy.org/formats/), and the Mass Spectrometry Standards Working Group (http://psidev.sourceforge.net/ms/). One way these efforts can succeed is if customers, and funding agencies, condition purchases on the delivery of open, specified and standardized data formats.

QUERYING COMPLEX MULTIDIMENSIONAL DATA

How will it be possible to precisely express a biological experiment including reagents, cDNA sequences, cell lines, protocols, acquisition systems, and the raw and processed data? Currently, laboratory-management-data systems to seamlessly manage this workflow are still developing, with the major challenge being the balance between standardization and adaptability described above. Search and querying is an absolute requirement that serves the need to explore recorded data for new links and findings. Biological ontologies enable queries across a range of organisms, processes or molecules. Every scientist has had the experience of recalling old results and reinterpreting their meaning in the light of new findings. The dominance of relational database technology is, in part, due to the reliability of the queries that report back an exact listing of records that match the search criteria (for example, "select all images owned by Jason of HeLa cells with GFP–CENP-A where interkinetochore distance of kinetochore pairs at metaphase is >800 nm"). The limitation of this approach is the rigid boundaries imposed by the highly structured data models. The popularity of web search engines like Google, which apply a query across all web pages, spanning many sites and types of documents, is due to their breadth and comprehensiveness. The problem with unstructured queries (for example, "Jason HeLa GFP–CENP-A >800") is the underlying fuzziness — the value 800 may match any number of metadata values, so the query produces a longer and much more confusing list of results. Sometimes this is beneficial and the results of a Google search often provide useful surprises. This comparison illustrates what is really needed — a hybrid of structured and fuzzy queries.

The emergence of open-source software tools that offer a search engine "core", which combines free-text indexing with an awareness of relational data structures (see http://lucene.apache.org/; http://www.opensymphony.com/compass/content/hibernate.html), may make it feasible to implement a hybrid search-and-querying engine for experimental biological data. Using these types of tools, it will be possible to support the requirements for flexible searching and querying over structured and semi-structured data, both locally and globally.

CONCLUSIONS

Modelling of biological data generates powerful facilities for data management and analysis, and is increasingly important for systems biology. We have focused our discussion on applications in imaging, but the points made are applicable to all forms of large-scale multidimensional data acquisition. The various data representations we have discussed provide the critical metadata that defines the types and relationships needed for data analysis. The defined structure of data models and their physical implementations must be shared and as open as possible, to enable scientists to explore new approaches and methods of analysis.

Table 2 Glossary

Controlled vocabulary	A defined set of words used to describe specified objects and/or events. Often this vocabulary will be expressed in a defined form (for example, XML; see below).
Data model	The product of a design process that aims to identify and organize the data into logical units, with specified relationships, for use in a database.
DICOM	A defined file format for medical imaging; see http://medical.nema.org/
FCS	An industry standard for flow cytometry data; see http://www.isac-net.org/
Free text	Text-based data expressed in a natural language (for example, English or French). The meaning of the text is not accessible to a computer, although it can find exact matches to words and phrases.
Metadata	The "data about the data". Descriptive and contextual information about the acquired data.
Multiplexed assay	An assay that has been converted such that multiple parallel measurements can be executed simultaneously, often in an arrayed format.
OBO	Open Biological Ontologies; http://obo.sourceforge.net/
OME	Open Microscopy Environment; http://openmicroscopy.org/
Ontology	A controlled vocabulary that defines terms and their relationships for a specific knowledge domain. Although ontology and data model can be used interchangeably, an ontology emphasises the vocabulary used for the entities and their relationships, whereas a data model emphasises its structure.
SBML	Systems biology markup language (SBML; http://sbml.org/) is derived from XML and is a format used for exchange of the data types used in systems biology.
Time-lapse assay	An experiment where measurements are made at defined intervals (the "lapsed" time).
Training set	A defined set of data, usually used as input for a machine learning algorithm, for the purpose of calculating a set of decision rules that enable automated recognition of different types of structures or events.
XML	Extensible markup language (XML; http://www.xml.org/) is a form of structured text language or "markup language" designed for specifying and describing complex sets of data. XML is readable by both humans and software, making it an ideal tool for specifying data models and ontologies. Similar in syntax to HTML, XML is a language for semantic markup, whereas HTML is a language for display markup.

ACKNOWLEDGEMENTS

Work in the authors' laboratories is supported by grants from the Wellcome Trust, the Biotechnology and Biological Sciences Research Council (BBSRC) and Cancer Research UK (J.R.S.) and the National Institutes of Health (I.G.G. and S.E.L.). J.R.S. is a Wellcome Trust Senior Research Fellow.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing finanial interests.

1. Shyamsundar, R. *et al.* A DNA microarray survey of gene expression in normal human tissues. *Genome Biol.* **6**, R22 (2005).
2. Janes, K. A. *et al.* The response of human epithelial cells to TNF involves an inducible autocrine cascade. *Cell* **124,** 1225–1239 (2006).
3. Wouters, F. S., Verveer, P. J. & Bastiaens, P. I. Imaging biochemistry inside cells. *Trends Cell Biol.* **11,** 203–211 (2001).
4. Andrews, P. D., Harper, I. S. & Swedlow, J. R. To 5D and beyond: Quantitative fluorescence microscopy in the postgenomic era. *Traffic* **3,** 29–36 (2002).
5. Lippincott-Schwartz, J., Snapp, E. & Kenworthy, A. Studying protein dynamics in living cells. *Nature Rev. Mol. Cell Biol.* **2,** 444–456 (2001).
6. Gerlich, D. & Ellenberg, J. 4D imaging to assay complex dynamics in live specimens. *Nature Cell Biol.* **Suppl**, S14–S19 (2003).
7. Neumann, B. *et al.* High-throughput RNAi screening by time-lapse imaging of live human cells. *Nature Methods* **3,** 385–390 (2006).
8. Nybakken, K., Vokes, S. A., Lin, T. Y., McMahon, A. P. & Perrimon, N. A genome-wide RNA interference screen in *Drosophila melanogaster* cells for new components of the Hh signaling pathway. *Nature Genet* **37,** 1323–1332 (2005).
9. Kittler, R. *et al.* An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. *Nature* **432,** 1036–1040 (2004).
10. Perlman, Z. E. *et al.* Multidimensional drug profiling by automated microscopy. *Science* **306,** 1194–1198 (2004).
11. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet* **25,** 25–29 (2000).
12. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19,** 524–531 (2003).
13. Goldberg, I. G. *et al.* The open microscopy environment (OME) data model and XML file: open pools for informatics and quantitative analysis in biological imaging. *Genome Biol.* **6**, R47 (2005).
14. Johnston, J., Nagaraja, A., Hochheiser, H. & Goldberg, I. A flexible framework for web interfaces to image databases: supporting user-gefined ontologies and links to external databases. *3rd IEEE Intl Symp. on Biomedical Imaging: Macro to Nano* 1380–1383 (2006).
15. Thomann, D., Rines, D. R., Sorger, P. K. & Danuser, G. Automatic fluorescent tag detection in 3D with super-resolution: application to the analysis of chromosome movement. *J. Microsc.* **208,** 49–64 (2002).
16. Ponti, A., Machacek, M., Gupton, S. L., Waterman-Storer, C. M. & Danuser, G. Two distinct actin networks drive the protrusion of migrating cells. *Science* **305,** 1782–1786 (2004).
17. Dufour, A. *et al.* Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces. *IEEE Trans. Image Process* **14,** 1396–1410 (2005).
18. Platani, M., Goldberg, I., Lamond, A. I. & Swedlow, J. R. Cajal body dynamics and association with chromatin are ATP-dependent. *Nature Cell Biol* **4,** 502–508 (2002).
19. Roques, E. J. S. & Murphy, R. F. Objective evaluation of differences in protein subcellular distribution. *Traffic* **3,** 61–65 (2002).
20. Conrad, C. *et al.* Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res.* **14,** 1130–1136 (2004).
21. Chuai, M. *et al.* Cell movement during chick primitive streak formation. *Dev. Biol.* **296,** 137–149 (2006).
22. Forouhar, A. S. *et al.* The embryonic vertebrate heart tube is a dynamic suction pump. *Science* **312,** 751–753 (2006).
23. Seamer, L. C. *et al.* Proposed new data file standard for flow cytometry, version FCS 3.0. *Cytometry* **28,** 118–122 (1997).
24. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genet* **29,** 365–371 (2001).
25. Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R. & Stein, L. The distributed annotation system. *BMC Bioinformatics* **2,** 7 (2001).
26. Whetzel, P. L. *et al.* The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics* **22,** 866–873 (2006).
27. Swedlow, J. R., Goldberg, I., Brauner, E. & Sorger, P. K. Informatics and quantitative analysis in biological imaging. *Science* **300,** 100–102 (2003).
28. Huang, K., Lin, J., Gajnak, J. A. & Murphy, R. F. Image content-based retrieval and automated interpretation of fluorescence microscope images via the protein subcellular location image database. *Proc. IEEE Symp. Biomed. Imaging* 325–328 (2002).