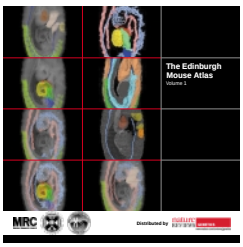
# BIOINFORMATICS BEYOND SEQUENCE: MAPPING GENE FUNCTION IN THE EMBRYO

*Duncan Davidson and Richard Baldock*

The spatio-temporal expression pattern of a gene during development is a valuable piece of information. But there is no way to compare precisely the patterns of expression of different genes, or the way the patterns are changed in a mutant. One way to solve this problem is to construct digital reference images of development (a bioinformatics framework), to which expression patterns can be mapped and stored, then compared. Such frameworks are under active development in several model systems. They will form the basis of powerful and integrated gene expression databases, which facilitate comparisons between genes, tissues and species.

COMPUTATIONAL GENETICS




A CD-ROM version of the Edinburgh Mouse Atlas — one of the resources discussed in this article — is available to all subscribers this month. For further copies, contact: ma-cdrom@hgu.mrc.ac.uk

*MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK. Correspondence to D.D. e-mail: Duncan.Davidson@hgu.mrc.ac.uk*

In the field of molecular genetics, the rapid establishment of databases and related technology has a solid foundation that depends on the fact that the nucleic-acid sequence provides a simple, one-dimensional framework to which a wealth of other data can be related (FIG. 1). However, looking beyond those data that relate directly to genomic sequence, bioinformatics is less well equipped to help understand gene function. The widespread use of electronic literature, introduction of high-throughput assays for gene expression and other large-scale projects (for example, mutagenesis screens and phenotyping projects for the mouse[1,2]) are vastly increasing the amount of digital information that is available. Advances in imaging and molecular tagging are opening up exciting new ways to visualize and follow the processes of development in time and space. But the complexity, variety and volume of information, as well as an inherent lack of rigour by comparison with molecular data[3], present serious problems. The challenge is to integrate this multidimensional information in a way that will help biologists to find meaningful relationships in the data and to formulate creative hypotheses that can be tested by observation and experiment.

One strategy to meet this challenge is to create digital frameworks that are analogous to DNA sequence but that represent other levels of biological organization, and with which different types of functional information can be associated (TABLE 1). In this article, we review the progress in devising and building bioinformatics frameworks at the organism level. Although the approach can, in principle, be applied to the adult, we confine our attention to digital frameworks for embryonic development (TABLE 2).

## Frameworks at the organism level

To understand gene function at the embryo level, developmental biologists must bring together very different kinds of information: gene expression patterns (for example, from *in situ* hybridization and microarray assays), patterns of cell proliferation, cell lineage, mutant phenotypes, changes in physiological variables with time in development, and experimental data from studies of morphogenesis and pattern formation. But almost none of this data can be accessed efficiently in a bioinformatics context. The common usage of anatomical names, for instance, is too inconsistent to provide an adequate reference framework. For example, any attempt to search a bibliographic database for papers that describe gene expression in the anterior part of the spinal cord in the mouse will quickly highlight deficiencies. Papers will be missed, and with many of those that are returned the desired comparisons between different

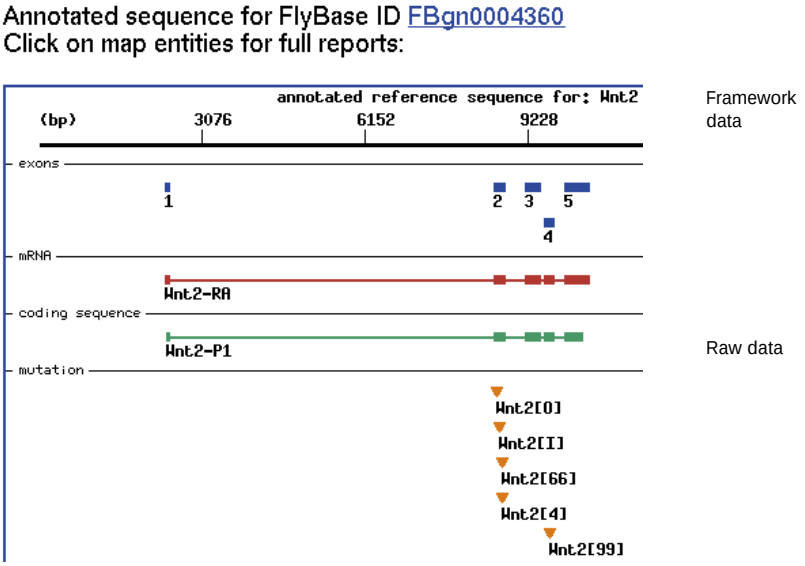



Figure 1 | **Genomic sequence as a framework.** Part of a window from FlyBase[36] that illustrates the use of the *Drosophila* genomic sequence around the *Wnt2* (Wnt oncogene analogue 2) gene as a reference framework to index, and thus integrate, different types of information, including gene structure, RNA and mutants. (Reproduced with permission from FlyBase.)

CONTROLLED VOCABULARY
A list of permitted terms that can be used to describe data: for example, radius, forelimb and bone.

ONTOLOGY
A collection of terms that describe concepts, entities and their relationships: for example, the 'radius' is part of the 'forelimb' or the 'radius' is an instance of 'bone'.

DIRECTED ACYCLIC GRAPH
A graph of relationships between objects in which any object can have more than one parent, each link is directional and cyclical relationships are prohibited.

gene expression patterns will be impossible. As data increases in diversity and complexity, and as searches aim to satisfy several conditions, a systematic reference system with which to index this information becomes essential. This is the role of a bioinformatics framework.

A biological, rather than technical, frame of reference is likely to provide the best means to capture the information that will help us to explore gene function; a suitably general biological framework is a digital description of the organism itself. The 'framework' data that constitutes such a reference description is distinct from the 'raw' data that is derived from the study at hand. For example, in reports of an *in situ* hybridization experiment, the raw data comprise the labelling pattern and the framework data comprise the list of names that is used to describe which tissues are labelled. To be used in an informatics context, the framework should be in the form of a database that can be searched and should be made widely accessible, as well as interoperable with other informatics resources. A discussion of the important technical challenges that must be addressed to achieve database interoperation, including the design of information media for data, is outside the scope of this article (see REF. 4 for a review). Our focus here is on the general nature and application of frameworks. Framework data describe the entities that comprise the system and the relationships between them, and should refer to the evidence on which these data are based. In particular, such a framework should include time, location in the embryo, and the common concepts of developmental biology, which include anatomical structure, cell type and developmental processes (such as 'gastrulation'). Indexing must also take account of quantitative aspects of the data.

Table 1 | Examples of frameworks at different levels

| Level | Example resource | Type of information |
|---|---|---|
| Genome | GenBank | DNA sequence |
| | SWISS-PROT | Protein sequence |
| Protein structure | CATH | Protein structure classification |
| Metabolic pathway | KEGG | Molecular pathway ontology |
| | EcoCyc | Molecular pathway ontology |
| Organism | FlyBase — controlled vocabulary | Ontology of names for body parts in *Drosophila* |
| | Normal table of *Xenopus laevis* | Stages of development |
| Species | NCBI taxonomy | Taxonomic vocabulary |

Links to the resources mentioned are provided online.

For each model organism, the preferred temporal reference is a standard series of discrete, defined stages of development (TABLE 2), although experimental results are often reported with respect to time elapsed since fertilization. In addition, panels of carefully defined, and widely available, molecular assays can be used to mark key developmental events. Cell lineage and tissue derivation provide an alternative view of developmental time that, in some cases, for example in *Caenorhabditis elegans* (FIG. 2), is consistent and detailed.

Text is an appropriate medium to relate data to concepts. The volume of data and the opportunities presented by database technology have led to a formalization of textual terms to create CONTROLLED VOCABULARIES and ONTOLOGIES (TABLE 2; FIG. 2). The flexibility of text allows diverse information to be indexed, and incomplete or low-resolution information to be dealt with pragmatically. Hierarchical ontologies allow data with different levels of resolution to be indexed. However, a single hierarchy is insufficient to describe the relationships between anatomical components. For example, ontologies that are realized as DIRECTED ACYCLIC GRAPHS (DAGs), which allow any component to have more than one 'parent', are more satisfactory in this respect. FlyBase has successfully indexed data that relate to gene function for many thousands of genes in *Drosophila*, using ontologies and controlled vocabularies that are supplemented where necessary with free text.

The Gene Ontology (GO) framework[5] is an especially interesting example of a flexible and rapidly developing textual framework for the molecular functions of gene products, their sub-cellular localization and the biological processes in which they function in the embryo or adult. The GO Consortium, initially a collaboration between the Mouse Genome Informatics, FlyBase and *Saccharomyces* Genome databases, is pioneering the description of biological processes and helping to establish standards for the community-wide organization of frameworks and for the integration of data from different model organisms.

Textual frameworks are limited, however, in their ability to describe development. Space and time are essential ingredients of developmental processes: cells and tissues move relative to one another and interact through diffusible signal molecules to induce new programmes of development. Images are the appropriate medium for spatial and temporal information and, as we

Table 2 | Examples of proposed and existing frameworks at the organism level

| Model organism | Temporal (development) | Spatial | Textual anatomy | Mutant phenotype | Gene function |
|---|---|---|---|---|---|
| *Saccharomyces cerevisiae* | – | – | – | – | Gene Ontology (GO) framework |
| *Caenorhabditis elegans* | REF. 38 WormBase | – | REF. 38 WormBase | WormBase | – |
| *Drosophila melanogaster* | REF. 39 FlyBase | REF. 19 | FlyBase: anatomy | REF. 25 FlyBase | GO framework |
| *Xenopus laevis* | REF. 40 Stages of embryonic development | – | – | – | – |
| *Danio rerio* | REF. 41 Zebrafish development | REF. 18 | Zebrafish: information server | – | – |
| *Gallus gallus* | REF. 42 Chick development | – | – | – | – |
| *Mus musculus* | REFS 14,43 EMAP | µMRI Mouse Atlas: REF. 16 and EMAP | Embryo: REF. 13 and EMAP: anatomy Adult: MGI anatomy | REF. 26 Dysmorphic Human-Mouse Homology | GO framework |
| *Human* | REF. 44 Developmental Anatomy Centre | – | REF. 45 Developmental anatomy | London Dysmorphology Database Dysmorphic Human-Mouse Homology | |

Resources (see links) and references are shown as appropriate. Some of the references refer to continuing work.

have mentioned, the widespread use of digital imaging in developmental biology is creating exciting new opportunities in this direction. For example, rapid methods to generate three-dimensional reconstructions of gene expression data that can be stored in a database have recently been described[6,7]. Clearly, there is a need to make image data accessible to bioinformatics. The problem is that the spatial information in independent images is not easy to search. One solution is to store images in a database and annotate them using text, including anatomical terms. This approach is taken, for example, by the Mouse Genome Informatics Gene Expression Database — the GXD[8]. However, searching remains limited by the textual framework. Moreover, even when the appropriate raw data can be retrieved from the database, comparative analyses of different images is difficult if the data are too incomplete, or the patterns too complex, to allow visual comparisons even in a small data set.

These difficulties can, in principle, be overcome by mapping data into a common spatial framework, thus making it accessible to spatial searches. Recent advances in image processing and computer graphics have begun to make spatial mapping possible for biological applications[9,10]. Different kinds of framework can be used, including diagrams and different projections of the embryo. Three-dimensional models of the developing embryo provide an especially powerful tool because mapped data can also be compared visually in the appropriate developmental context so that relationships can be explored more easily than, for example, in textual descriptions or tables. This approach might eventually be amenable to semi-automatic application because, in contrast to textual annotation, little previous analysis is required to index data. Current work discussed below has begun to establish the feasibility of this approach and to provide resources that are of practical use to developmental biologists.

THEILER STAGE
Stages of development of the mouse embryo determined by particular developmental events: for example, eye closure or appearance of digits.

## Atlas frameworks for development

There are numerous instructive graphical atlases and other resources that help researchers to identify anatomical structures and review developmental processes. We concentrate attention here, however, on those atlases that plan to make framework data accessible for bioinformatics applications (TABLE 2).

*The Edinburgh Mouse Atlas Project*. The Edinburgh Mouse Atlas Project (EMAP) atlas[11,12] combines textual and spatio-temporal approaches (FIGS 3, 4). A stage-by-stage ontology of anatomical names[13] is linked to a temporal series of three-dimensional digital model embryos based on serial histological sections. The principal named structures are delineated in the models, anchoring anatomical terms (and data associated with them, for example, tissue 'lineage') to the spatial framework. The atlas is based on the definitive books by Theiler[14] and Kaufman[15]; where possible, the same specimens have been used as are illustrated in Kaufman's book and the anatomical nomenclature is based on this text. EMAP is an ongoing project that has benefited from the advice of community-funded networks and of the GXD team at the Jackson Laboratory, who are developing a related textual anatomical ontology for the adult mouse (the GXD). The EMAP atlas at present contains completed models of

a

| Cell | Cell Group | Life Stage |
|---|---|---|
| AB.alppap | pharyngeal isthmus | 1.5-fold embryo |
| AB.alppapa | pharyngeal marginal cells | 2-fold embryo |
| AB.alppapp | pharyngeal motor neurons | 3-fold embryo |
| AB.alppp | pharyngeal muscle | 50-70 cell embryo |
| AB.alpppa | pharyngeal neurons | 90 cell embryo |
| AB.alpppaa | pharyngeal processes | adult |
| AB.alppppap | pharyngo-intestinal valve | adult hermaphrodite |
| AB.alppppp | pharynx | adult male |

b

embryonic/larval foregut ; FBbt:00001862 ~ stomodeum % foregut
hypopharyngeal lobe ; FBbt:00001863 ~ intercalary segment
<keyhole structure ; FBbt:00001864
<alveolar cell ; FBbt:00001865 % cell
<embryonic/larval pharynx ; FBbt:00001866
<pharyngeal bars ; FBbt:00001867 ; synonym: T-bars
<embryonic/larval epipharynx ; FBbt:00001868 < labral segment
<embryonic/larval hypopharynx ; FBbt:00001869 < intercalary segment
<embryonic/larval oesophagus ; FBbt:00001870 ; synonym: larval esophagus ~ stomodeum % oesophagus
<embryonic/larval proventriculus ; FBbt:00001871 ~ stomodeum

c

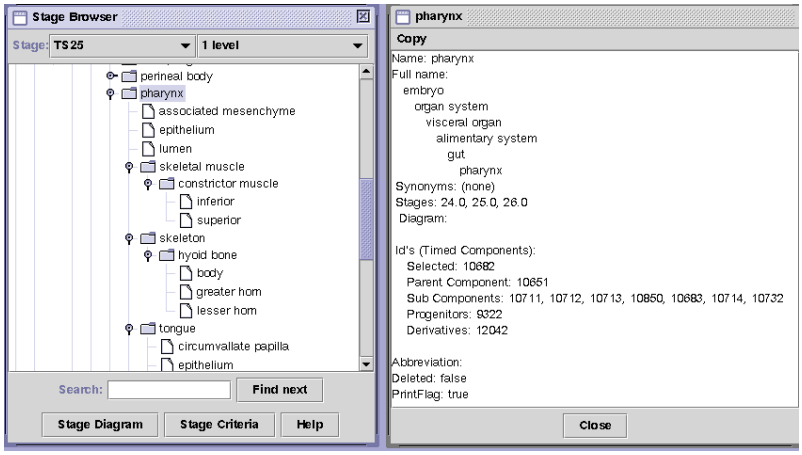



Figure 2 | **Textual anatomical frameworks. a** | Cell lineage, cell groups and life stages frameworks for the nematode *Caenorhabditis elegans* shown here in the WormBase gene expression search interface[37]. **b** | Part of the ontology for anatomy from FlyBase. Unique identifiers are shown as FBbt numbers. <, 'part of its parent term' relations; %, 'instance of' relations; and ~, 'derived from'. Synonyms are also shown. **c** | Part of the ontology for the mouse embryo from the EMAP. The left window shows the hierarchical tree, with the pharynx highlighted. The right window shows the properties of the pharynx from the database, including the unique identifier, parent terms and sub-components. These three examples also highlight some of the difficulties of comparing the anatomy of different organisms, where similar terms might refer to very different structures. (Reproduced from EMAP and with permission from FlyBase and WormBase.)

XML FORMAT
The eXtensible Markup Language (XML) is related to HyperText Markup Language (HTML) and is an emerging standard for structuring documents.

OBJECT-ORIENTATED
The term used to describe a programming design in which all data are regarded as parts of objects that hold both the data and the procedures or methods that can be applied to that data. This encapsulation enforces good programming practice and makes code more portable.

early post-implantation embryos from THEILER STAGES (TS) 7–14 and further models are planned or in progress. The textual ontology is available for all stages and future plans include additions in response to community requirements, conversion of the hierarchy to a DAG and the addition of data on tissue lineage and cell types. (For the current status of the EMAP atlas and graphical gene expression database, see the EMAP web site.) The textual ontology of the atlas is available in XML FORMAT and is in use to annotate gene expression in the GXD. The atlas framework is implemented as an OBJECT-ORIENTATED, CORBA-compliant database with JAVA-based interfaces. One of the aims of EMAP is to make available generic tools to build similar frameworks for other species. Specific features of the atlas and instructions for using it online and on CD have recently been reviewed[16].

*MRI atlas of mouse development.* High-resolution magnetic resonance imaging (μMRI)[17] provides one alternative to using histological sections as a means to build an atlas. This method is being used in the μMRI Mouse Atlas project by Russell Jacobs and colleagues at Caltech to build three-dimensional digital model mouse embryos that can be viewed interactively as a series of sections through the three-dimensional image. Principal anatomical domains have been delineated and named in a model of the 13.5-day-old embryo. Even μMRI achieves much lower resolution than can be achieved with histological material, and structures that are visualized using MRI do not correspond in detail with those visible using conventional optical methods. However, MRI has the advantage that it does not disrupt the embryo and can rapidly produce three-dimensional, digital models that accurately represent its shape.

*Atlas frameworks in other organisms.* A digital graphical atlas has been proposed for *Drosophila* development. This is based on serial optical sections with detailed delineation at the cellular level[19]; current effort is focused on building three-dimensional models of the larval brain (the FlyBrain project) that contain landmark features such as boundaries between compartments and principal axon tracts. An atlas of zebrafish development is being constructed[18] (FIG. 5), and a project has been initiated by the University of Newcastle in collaboration with EMAP to build an atlas of the embryonic human brain at two stages of development. Neither of these resources is yet available. Like the EMAP, these atlases will include a textual anatomical framework linked to three-dimensional model embryos that have been reconstructed from images of serial sections. An atlas of human development is also being prepared using MRI techniques (The Multidimensional Human Embryo), but its use as framework data for bioinformatics applications has not been proposed.

Digital atlases of the adult brain are being developed for several species. Although these adult atlases fall outside the scope of this review, the methods and approaches are closely related (see REF. 20 for a review). These include brain atlases for humans[21] (multimodal brain atlases), mice (Mouse Brain Library, Brain Molecular Anatomy Project and Mouse Brain Atlas) and rats[22].

*Combining data from different organisms.* The development of independent bioinformatics frameworks for several model organisms calls attention to the need for interoperability. To achieve this will require the application of standards for data and data exchange across the developmental biology community. The reward is likely to be a series of stimulating interactions between bioinformatics and comparative biology. On the one hand, the need to formalize relationships between frameworks for different organisms will focus attention on the relative meanings of terms and concepts (FIG. 2), and on the relationships between spatio-temporal domains in different species. On the other hand, the facility to compare data from different organisms (for example, composite expression patterns of groups of functionally

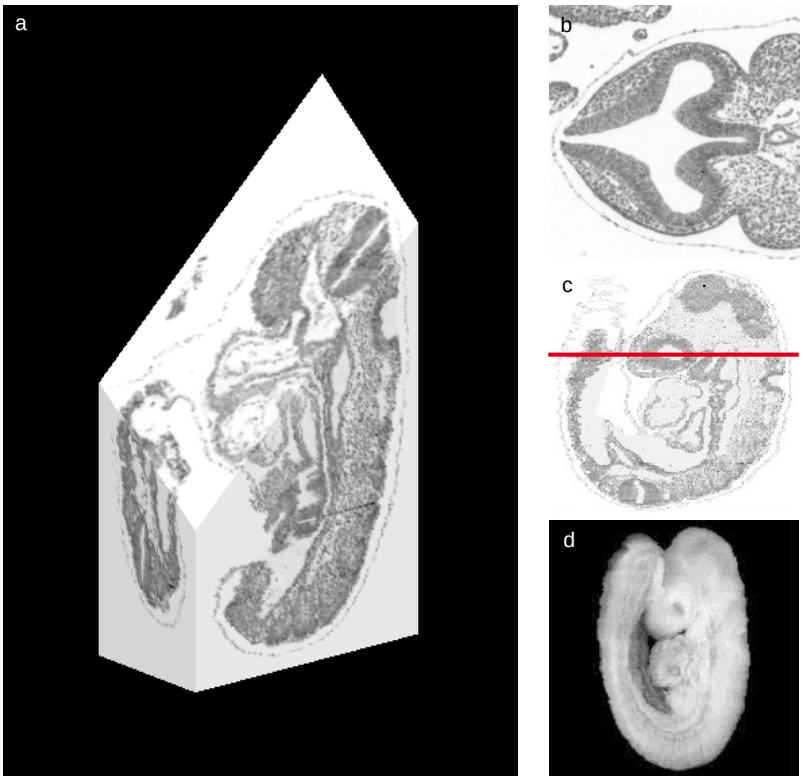



Figure 3 | **EMAP model embryos. a** | Each model embryo is constructed from images of serial transverse histological sections. As shown here for the Theiler stage 14 (embryonic day 9) model embryo, the volume can be digitally re-sectioned in any plane to show the histology along the cut faces. **b** | Part of one of the original images from which the model embryo was built. Section thickness (7 μm) limits the resolution of the three-dimensional model. In this model, each VOXEL represents a cuboid of 4 μm × 4 μm × 7 μm. In the higher-resolution original images, pixel size is 1.3 μm × 1.3 μm. **c** | A digital section through the model with the original plane of section marked (red line). **d** | In addition to section views, the same model can be viewed as a whole embryo. Whole-embryo views provided on CD and online can be rotated to visualize the model in three dimensions.

related genes) can be expected to illuminate the relationships between development and evolution[23]. A single, pan-species framework at the organism level can be envisaged, comprising a distributed system of linked databases, each pertaining to a different species. Such a distributed system might also include complementary frameworks for the same species. The nomenclature translations and spatio-temporal transformations that comprise these interspecies links are themselves framework data and could be contained, along with pointers to supporting evidence, in separate ontologies or 'translation' databases. This explicit representation of the relationships between species frameworks means that spurious links arising from use of the same word (for example, 'vein' in mammals or in the fly wing) are avoided and that, conversely, corresponding components/tissues with very different names can be linked.

## Compromises and limitations

Although the broad vision for the future is becoming clear and is certainly exciting, for the moment the techniques for building and using digital, spatio-temporal frameworks of the type described above are in their infancy and much remains to be done. The temporal, spatial and anatomical resolution of the framework must be sufficient to index spatio-temporal and quantitative relationships in the raw data and to relate these data to the morphological and histological development of the embryo. With finite resources, there must also be a compromise between the number of models in the temporal series and the spatial and anatomical resolution of each model (FIGS 3,4). The number of models and the subdivision of delineated anatomical domains can be progressively increased, but spatial resolution is likely to remain limiting.

The choice of image material for building the models is crucial. Voxel images, which provide a three-dimensional 'photographic' representation of stained, histological structure (FIGS 3,5) are more generally useful than wire-frame or surface models, which are limited to outlines of selected components, unless the position of each cell is marked[19]. Moreover, grey-level voxel data can index quantitative relationships. Voxel size is a measure of spatial resolution. The resolution of models in the EMAP atlas, for example, shows individual tissues, including epithelial sheets, and also gives a coherent overview of the embryo that will allow the distribution of gene products in different regions to be mapped. This is important to allow the study of the pleiotropic effects of gene expression and to integrate data from different organ systems. The Zebrafish Atlas will have similar resolution[18]. In the longer term, high-resolution models of particular organs could be added and linked to a general atlas framework. These might be at approximately single-cell resolution, such as models of early post-implantation (TS 7–12) in the EMAP atlas.

In the atlas frameworks described above, each model is based on data from a single specimen and represents the topology and the approximate MORPHOMETRIC characteristics of the embryo. However, the shape and size of individual specimens are subject to natural and experimental variation, and so the models are of limited value as a source of detailed morphometric measurements, for example to compare wild-type with mutant embryos. Ultimately, each model must accommodate information about variation between specimens. This might come from data on unsectioned embryos, or even live specimens using MRI. Methods to represent variation are likely to follow the pioneering approaches that are being developed for the human brain, where three-dimensional to three-dimenional image transformation is used to build probabilistic atlases that record and display variation in different parts of the structure[20].

## Using atlas frameworks in bioinformatics

To help address the problems of bioinformatics at the organism level, framework data can be deployed in various ways (summarized in general terms in FIG. 6). These are discussed more specifically in the following sections, using the EMAP atlas as our main example.

*A reference framework.* The simplest use of an atlas is as a bench tool to identify and name parts of the embryo and to explore their morphological development[16]. Each spatial model in the EMAP atlas can be viewed interactively

CORBA
Common object request broker architecture is an industrial standard protocol for making objects, both data and methods, accessible over the Internet for remote access and invocation. If a database provides a CORBA interface then it is CORBA-compliant.

JAVA
An object-orientated language developed by Sun Microsystems that has become a standard for programs delivered as part of World Wide Web documents.

VOXEL
A digital image is represented as an array of image values, either as a grey-level or colour. A voxel is the term for one value in a three-dimensional array for a three-dimensional image.

MORPHOMETRIC
A measure of shape.

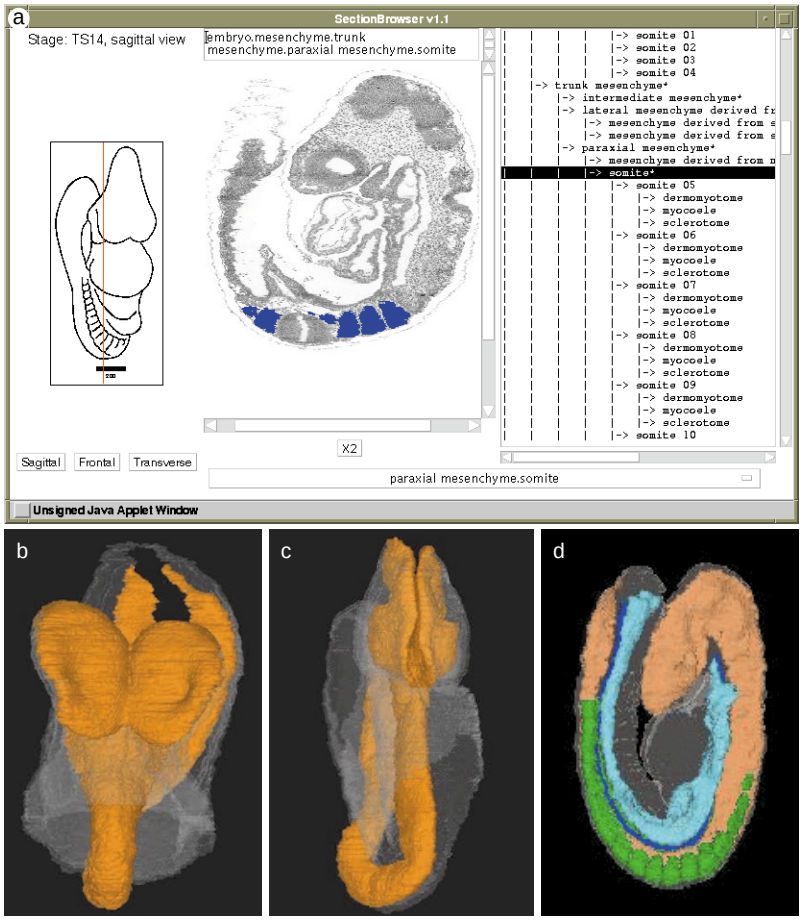



Figure 4 | **Integrating anatomical names with development in time and space in an atlas framework.** Model embryos in the EMAP atlas can be viewed in section or as three-dimensional movies that highlight particular anatomical structures to show their spatial relationships and developmental changes. **a** | Selected parts of the embryo (in this case, somites in a Theiler stage 14 embryo) can be identified, highlighted and named in different planes of section. **b** | Theiler stage 12 embryo (embryonic day (E)8.0). (See movie 1 online.) **c** | Theiler stage 13 (E8.5) embryo. (See movie 2 online.) Only two tissues have been highlighted in **b** and **c**; the remaining delineated tissue domains and the underlying histological structure are hidden from view (orange, neural tissue; translucent, surface ectoderm). **d** | Whole-embryo views show selected landmark structures in the Theiler stage 14 embryo as an aid to navigation (translucent, surface of the embryo; orange, neural tissue; light blue, gut; dark blue, notochord; green, somites).

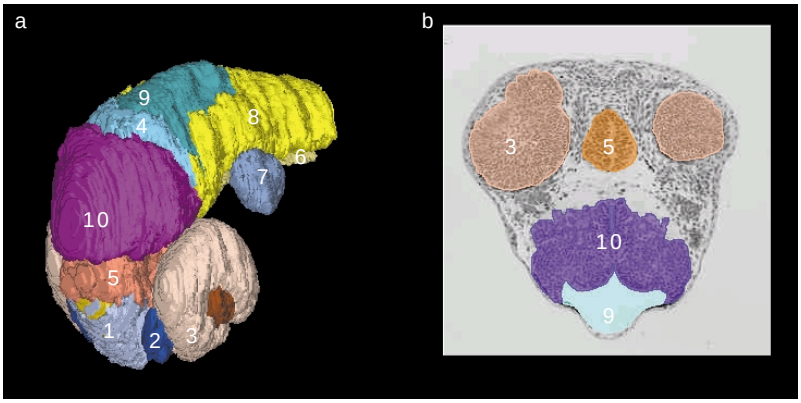
as a 'whole-mount' preparation with coloured landmark structures to help the researcher form a mental map in which to find and name smaller structures (FIGS 3, 4). Views of successive stages allow one to follow the development of those parts that have been delineated in the models. Each model can also be viewed in a series of pre-calculated, digital histological sections or, using more advanced tools, in any arbitrary plane of section, to obtain an approximate match with the plane of section of experimental material. The presentation of the spatial models is closely integrated with the ontology of anatomical names, thus allowing structures to be identified. A similar tool for identifying named parts is also provided by FlyBase, based on two-dimensional images of parts of the fly (see links). These tools are designed to help researchers analyse their own experimental material. They can also help them to prepare material for publication, for example, by building a list of anatomical names for keywords or tables of results. The EMAP atlas can thus be used to construct a textual description of expression for submission to GXD.

As more framework data are added, the reference function of the atlas will expand. For example, the expression of well-documented, widely available molecular markers might be included in the framework. The resulting 'molecular anatomy' will provide an additional analytical tool.

*Interaction with other data sources.* Information in different databases can be indexed using a common atlas framework. Independent interfaces can combine and analyse this diverse information, making use of the key advantage of the atlas in placing data in the context of the developing embryo. For example, the EMAP interface can help researchers to frame a query to the GXD and interpret the images of raw data returned from the query in terms of the anatomy and development of the embryo (EMAP:GXD Query interface). Where databases contain information that is spatially mapped to the atlas framework, such interfaces might also deploy tools for complex spatial analysis and visualization.

*A tool for local analysis.* The real value of an atlas as a bioinformatics framework depends on mapping raw data to the spatial coordinates of the model. This is also an area where much remains to be done. The process of mapping *in situ* gene expression patterns illustrates the issues involved (FIG. 7).

It is important to remember that many of the difficulties that limit the resolution of digital mapping are similar to those that hamper everyday attempts to compare images in published papers or in the laboratory. Differences in experimental methods, incomplete data, and differences in developmental stage and shape between specimens, all contribute to the difficulty of making comparisons. The temporal and spatial resolution of the framework and the accuracy of the methods used to match framework data to raw data are also important. Present methods, which are based on mapping two-dimensional images of sections or whole-mount preparations to equivalent two-dimensional images from the model, underperform when the shapes of specimen and model greatly differ. Where the difference between specimen and atlas model is extreme, mapping can be unreliable, so the approach must be used with care, and methods to measure divergence between raw and framework data might become important in helping to assess the validity of mapping. The standardization of experimental data and the incorporation, in the framework, of information about variation are likely to increase mapping accuracy (see REF. 24 for general guidance on preparing gene expression data from *in situ* hybridization experiments for database entry). Routine methods for three-dimensional to three-dimensional image transformations applied to embryonic material will be an important advance. Not only will this aid the mapping of experimental data to the framework, but also, importantly, will allow construction of temporal sequences by interpolation between models in a series, as well as mapping between models of different organisms.




1 Telencephalon
2 Olfactory placode
3 Optic cup
4 Cerebellum
5 Diencephalon
6 Notochord
7 Otic vesicle
8 Myelencephalon
9 4th ventricle
10 Mesencephalon

Figure 5 | **The zebrafish atlas. a** | Model of the head region of the zebrafish. **b** | One slice of the model with the annotated anatomy in colour. The material shown is from a pilot experiment using serial sections of the head region of a 48-hour post-fertilization zebrafish embryo. (Reproduced with permission from Fons Verbeek.)

In spite of these limitations, mapping data to the framework provides very real advantages. Even without entering mapped data into a database, mapping is a useful tool for comparing small numbers of gene expression patterns and aids the analysis of individual gene expression results. Work with prototype methods (FIG. 7) shows that mapped data can reveal genuine details of expression that pass unnoticed in the course of simple visual analysis, particularly in serial sections or replicate data. Where the aim is to analyse a single set of experimental data, mapping the atlas onto raw data is useful as it brings the knowledge and information in the framework to bear on the data in their original form.

The trade-off between precision and time is crucial. The aim of mapping is to annotate significant aspects of the pattern with enough accuracy to enable useful comparisons and thus generate testable hypotheses. It is, therefore, important to store at least representative images of the original data as well as to record the mapping parameters. The resolution required to explore relationships between two patterns might be greater than is needed to index data for general searches that can be complemented by visual scans of returned results. With finite resources it will be important to optimize the balance between mapping any one pattern in detail and indexing large quantities of data with lower resolution. Further developments in pattern recognition and image processing are needed in order to apply spatial mapping to large volumes of data for bioinformatics purposes, but manual methods such as those illustrated in FIG. 7 provide a platform for the development of automated methods.

Other information related to gene function, such as cell proliferation, apoptosis or cell lineage, can be mapped using methods similar to those discussed above. Mapping mutant phenotype to a framework is more difficult[25]. At least a temporary, pragmatic solution is urgently needed to integrate data on adults and embryos from clinical studies in man (OMIM) and experimental studies in the mouse (TBASE), including large-scale mutagenesis screens. It will also be important to integrate databases that contain this data with gene expression databases and tissue lineage. Standard anatomical vocabularies can be used to index the tissues affected in a mutant. But to describe the nature of the effect, which might be entirely new, presents a challenge[26]. The aim must be to find ways to index mutant phenotype at the organism level that facilitate bioinformatics approaches to exploring the functions of genes and pathways. The development of terms that describe the features of mutant phenotypes is one area where this problem is being actively addressed (such as in the Dysmorphic Human–Mouse Homology Database). Spatial models in atlas frameworks might also be used to index this data in the future, for example by associating original images of mutant phenotypes with the spatial region affected, linking these with gene expression patterns in the wild type and mutant.

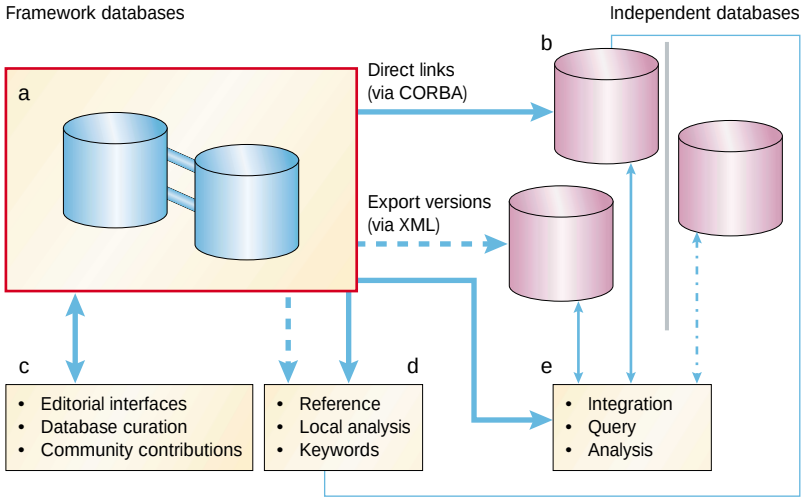



Figure 6 | **The deployment of framework data.** Framework data, contained in the red frame, is represented as a 'distributed' system, with linked databases (blue cylinders). Blue arrows show the flow of framework data — broken lines represent versioned output (see below), solid lines represent continuous access. Components of framework data should be **a** | mutually accessible, **b** | available for incorporation into the data models of independent databases, such as gene expression and bibliographic databases (pink cylinders) and **c** | accessible to the community. Direct access to the data might be provided through the use of standard protocols (for example, CORBA) or versions might be downloaded using standards as defined, for example, in XML. **d** | Presentation of framework data in an interactive bench tool aids analysis of new data and provides a source of reference keywords and images. **e** | The framework can also be used to construct independent interfaces that access diverse databases. The vertical grey line separates a database that does not use the framework from those that do.

*Databases.* Mapped data can be stored in databases that use the framework as part of the data model. The textual part of the framework can be used on its own, as is the case for the GXD, or the full framework can be used. The possibilities range from a private, local database (a simple list of files shared between collaborators, each file recording a single pattern) to a large-scale community database with remote submission to an editorial office, and facilities for textual and spatial searching through the Web. For example, the GXD and the EMAP are collaborating to build interoperable gene expression databases that can be used as a single resource (The Mouse Gene Expression Information Resource) to hold both textually and spatially mapped data[27].

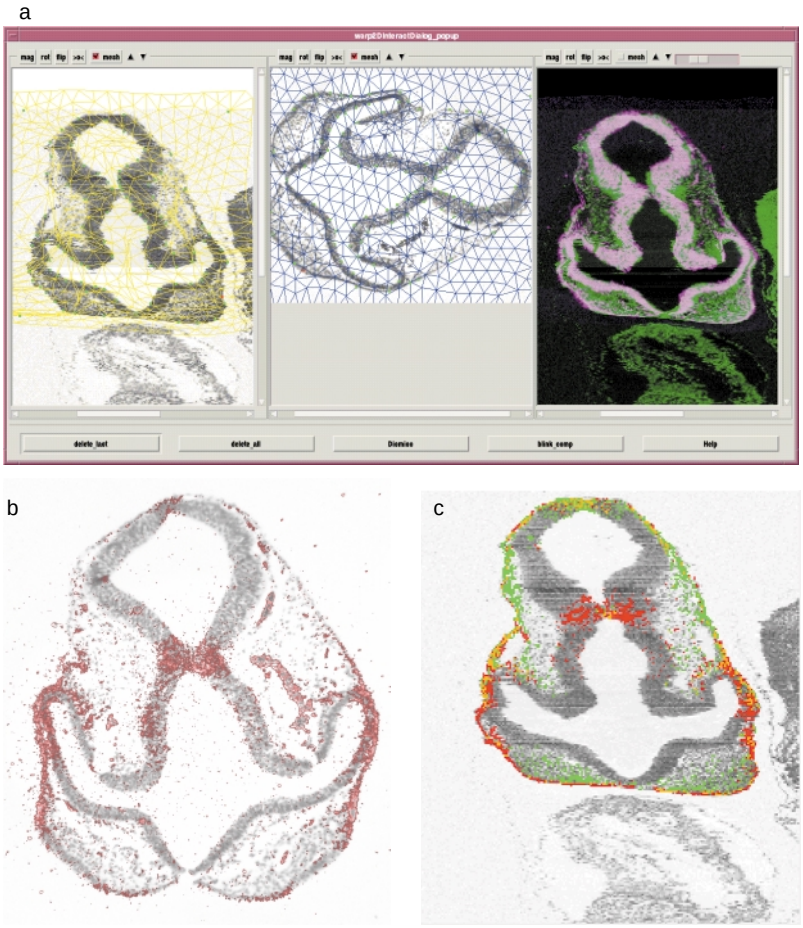



Figure 7 | **Mapping gene expression data to a framework.** The process of mapping gene expression patterns from an experimental specimen to a reference model in the EMAP. **a** | Mapping the pattern by image transformation (warping). The specimen containing raw data is matched to a model of the corresponding stage in the atlas framework. A section plane in the model (left panel) is selected that best corresponds to the specimen (middle panel). Corresponding positions are identified in the two images and used to derive an image transformation that enables all positions in the specimen to be approximately mapped to the atlas model (right panel: green, model; lilac, transformed specimen image). The result can then be edited to ensure that the mapped data corresponds as closely as desired to the original raw data. **b** | Raw data: the *Msx2* (homeobox Msh-like 2) gene expression pattern. **c** | Patterns returned from a database query: two mapped patterns compared. The comparison is between *Msx2* (red) and *Msx1* (homeobox Msh-like 1) (green).

Databases that use the full atlas framework can be interrogated using textual or spatial queries. A spatial 'query domain' might be a region drawn into the atlas framework or the domain of expression of a selected gene. Because data in the database might have been indexed using text, space or both, the resolution of the information returned depends on the resolution of query and data-input method. Pattern-recognition and image-processing methods that are used to query spatial databases are only beginning to be developed. So, although only simple queries are possible now (for example, 'Which genes are expressed in a specified three-dimensional space?'), we can expect that in the future more advanced methods, analogous to present-day sequence analyses, will detect complex relationships between patterns, such as different degrees of similarity and complementary relationships.

## Future directions

The first generation of atlas frameworks present stage-by-stage views of development. Because our aim is to understand how one stage or state is transformed into the next, it will be important to extend frameworks to represent development as a continuous process that is accessible to bioinformatics. One approach might be to annotate processes using an ontology that captures information about the temporal relationships between steps. Temporal ontologies that have been devised for other fields[28] might be applicable here. Another approach is to construct a 'movie' of the developing embryo by interpolation between successive models and to use this as a continuous spatio-temporal atlas framework to which successive steps in a process might be mapped. The success of this approach will depend on future work to define common spatial coordinates for different models in the developmental series and on developing algorithms for interpolation that give a useful match with actual changes in embryo shape.

By focusing attention on bioinformatics frameworks at the organism level, we have emphasized the practical value of a horizontal division of gene function data, but ignored the importance of information relating to the functions of gene products at other levels of biological organization[29]. Frameworks based on horizontal divisions are useful to index phenomena, such as gene expression patterns. But mechanisms are often better represented by vertical links between different levels of organization. Here, frameworks at the organism level must be linked to a different reference frame, the set of biological processes that relate to a particular 'function', however that is defined. Such a 'functional unit' might embrace genes, their products, molecular pathways[30–32], cells and the organism[5]. Extending the principles discussed in this review, one can identify frameworks at different levels of organization (TABLE 1) and envisage a bioinformatics system that is comprised of horizontal layers of framework data linked by data that represent causal relationships. There are significant gaps in the levels represented in TABLE 1, notably the cell level and those above the species, for instance the phylogenetic tree (see REF. 33 for a discussion of some of the issues involved). Canonical features of the cell cycle might form the basis of one kind of framework data at this level, but frameworks relating to other aspects of cell and subcellular properties and behaviour present an important challenge (see REF. 34 for one interesting approach).

One can anticipate that, in the future, bioinformatics will include simulation models not only as exploratory tools, but also as framework data to encode complex causal relationships and as part of the armoury for hypothesis generation for complex systems[35]. It might be possible, for example, to map new data to a computational model of a process, thereby indexing the data to the appropriate context in relation to the structure and behaviour of the model. We can speculate that such a tool might one day make accessible to bioinformatics not only

data on parameters of processes in normal development, but also information about the stability of the system as these values vary in evolution and disease.

## Conclusion

A framework approach that is based on the underlying biology provides a powerful strategy for integrating information at different levels of gene function. Atlas frameworks aim to provide such a context at the organism level. As these resources become widely available and are used by an increasing number of databases, further community involvement will become crucial. Contributions by specialists will encapsulate knowledge in these frameworks and the development of standards for software and data will widen their application. The enormous amount of data that is becoming available must not only be organized, but also used synergistically. The work involved in building frameworks and indexing data, although daunting, is small by comparison with the benefits to be gained. The real challenge is to develop the infrastructure in the biomedical community that can catalyse this process.

## Links

DATABASE LINKS Wnt2 | Msx2 | Msx1

FURTHER INFORMATION FlyBase | FlyBase vocabulary | Gene Ontology (GO) Consortium | Mouse Genome Informatics Database | *Saccharomyces* Genome Database | Mouse Genome Informatics Gene Expression Database | The Edinburgh Mouse Atlas Project | EMAP: anatomy | Jackson Laboratory | μMRI Mouse Atlas | Zebrafish development | The Multidimensional Human Embryo | FlyBrain | Multimodal brain atlases | Mouse Brain Library | Brain Molecular Anatomy Project | Mouse Brain Atlas | Rat Brain Atlas | Zebrafish Atlas | Two-dimensional images of parts of the fly | EMAP:GXD Query interface | OMIM | TBASE | Dysmorphic Human–Mouse Homology Database | The Mouse Gene Expression Information Resource | Window from FlyBase | WormBase | FlyBase: anatomy | Fons Verbeek's lab | GenBank | SWISS-PROT | CATH — Protein Structure Classification | KEGG — Kyoto Encyclopedia of Genes and Genomes | EcoCyc — Encyclopedia of *E. coli* Genes and Metabolism | The stages of *Xenopus* embryonic development | NCBI taxonomy | Chick development | MGI: anatomy | Human Developmental Anatomy Centre | Human developmental anatomy


1. Nadeau, J. H. Muta-genetics or muta-genomics: the feasibility of large-scale mutagenesis and phenotyping programs. *Mamm. Genome* **11**, 603–607 (2000).
**A thoughtful and provocative discussion of systematic approaches to capturing gene function data at the organism level.**
2. Paigen, K. & Eppig, J. T. A mouse phenome project. *Mamm. Genome* **11**, 715–717 (2000).
3. Bray, D. Reductionism for biochemists: how to survive the protein jungle. *Trends Biochem. Sci.* **22**, 325–326 (1997).
4. Achard, F., Vaysseix, G. & Barillot, E. XML, bioinformatics and data integration. *Bioinformatics* **17**, 115–125 (2001).
5. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
**Gene Ontology is an excellent example of a dynamic bioinformatic framework that aims to unify our understanding of gene function across species and different levels of biological organization.**
6. Streicher, J. *et al.* Computer-based three-dimensional visualization of developmental gene expression. *Nature Genet.* **25**, 147–152 (2000).
7. Hecksher-Sorensen, J. & Sharpe, J. 3D confocal reconstruction of gene expression in mouse. *Mech. Dev.* **100**, 59–63 (2001).
8. Ringwald, M., Eppig, J. T., Kadin, J. A., Richardson, J. E. & the Gene Expression Database Group. GXD: a gene expression database for the laboratory mouse: current status and recent enhancements. *Nucleic Acids Res.* **28**, 115–119 (2000).
9. Lester, H. & Arridge, S. R. A survey of hierarchical non-linear medical image registration. *Pattern Recognition* **32**, 129–149 (1999).
10. Toga, A. W. & Thompson, P. M. The role of image registration in brain mapping. *Image Vision Comput.* **19**, 3–24 (2001).
11. Baldock, R., Bard, J., Kaufman, M. H. & Davidson, D. A real mouse for your computer. *BioEssays* **14**, 501–502 (1992).
12. Brune, R. M. *et al.* A three-dimensional model of the mouse at embryonic day 9. *Dev. Biol.* **216**, 457–468 (1999).
13. Bard, J. *et al.* An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech. Dev.* **74**, 111–120 (1998).
14. Theiler, K. *The House Mouse. Atlas of Embryonic Development* (Springer, New York, 1989).
15. Kaufman, M. H. *The Atlas of Mouse Development* (Academic, London, 1992).
16. Davidson, D., Bard, J., Kaufman, M. H. & Baldock, R. The Mouse Atlas Database: a community resource for mouse development. *Trends Genet.* **17**, 49–51 (2001).
17. Louie, A. Y. *et al. In vivo* visualization of gene expression using magnetic resonance imaging. *Nature Biotechnol.* **18**, 321–325 (2000).
18. Verbeek, F. J. *et al.* A standard atlas of zebrafish development for projection of experimental data. *Proc. Int. Soc. Opt. Engng* **3964**, 242–252 (2000).
19. Hartenstein, V., Lee, A. & Toga, A. W. A graphical digital database of *Drosophila* embryogenesis. *Trends Genet.* **11**, 51–58 (1995).
**An excellent discussion of the issues involved in building an atlas database framework for *Drosophila.***
20. Toga, A. W. *Brain Warping* (Academic, San Diego, 1999).
**A collection of authoritative papers on methods to combine and compare three-dimensional images of mammalian brains.**
21. Toga, A. W. & Thompson, P. in *Brain Warping* (ed. Toga, A. W.) 1–26 (Academic, San Diego, 1999).
22. Toga, A. W., Santori, E. M., Hazani, R. & Ambach, K. A digital map of rat brain. *Brain Res. Bull.* **38**, 77–85 (1995).
23. Raff, R. A. Evo-devo: the evolution of a new discipline. *Nature Rev. Genet.* **1**, 74–79 (2000).
24. Davidson, D. *et al.* in *In situ Hybridization, a Practical Approach* (ed. Wilkinson, D.) 190–214 (IRL, Oxford, 1998).
25. Drysdale, R. Phenotypic data in FlyBase. *Briefings Bioinform.* **2**, 68–80 (2001).
**A good, practical description of how FlyBase addresses the problems of describing mutant phenotype in a bioinformatic context.**
26. Eppig, J. T. Algorithms for mutant sorting. *Mamm. Genome* **11**, 584–589 (2001).
**A discussion of work to incorporate phenotypic descriptors into the textual bioinformatics framework for the mouse. Compare with reference 25.**
27. Ringwald, M. *et al.* A database for mouse development. *Science* **265**, 2033–2034 (1994).
28. Allen, J. F. Maintaining knowledge about temporal intervals. *Commun. Assoc. Comput. Machin.* **26**, 832–843 (1983).
29. Vidal, M. A biological atlas of functional maps. *Cell* **104**, 333–339 (2001).
30. van Helden, J. *et al.* Biological function from the network of molecules and interactions. *Briefings Bioinform.* **2**, 81–93 (2001).
31. Karp, P. D. An ontology for biological function based on molecular interactions. *Bioinformatics* **16**, 269–285 (2000).
32. Rzhetsky, A. *et al.* A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics* **16**, 1120–1128 (2000).
33. Hillis, D. M. & Holder, M. T. in *New Technologies for Life Sciences: a Trends Guide* (ed. Wood, R.) 47–50 (Elsevier Science, London, 2000).
34. Tomita, M. *et al.* E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72–84 (1999).
35. von Dassow, G., Meir, E., Munro, E. & Odell, G. M. The segment polarity network is a robust developmental module. *Nature* **406**, 188–192 (2000).
36. The FlyBase Consortium. The FlyBase Database of the *Drosophila* Genome Projects and community literature. *Nucleic Acids Res.* **27**, 85–88 (1999).
37. Stein, L., Sternberg, P., Durbin, R., Theirry-Mieg, J. & Spieth, J. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* **29**, 82–86 (2001).
38. Sulston, J. E., Scherienberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
39. Campos-Ortega, J. A. & Hartenstein, V. *The Embryonic Development of* Drosophila melanogaster (Springer, Berlin, 1985).
40. Nieuwkoop, P. D. & Faber, J. *Normal Table of* Xenopus laevis (North Holland, Amsterdam, 1967).
41. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev. Dynamics* **203**, 253–310 (1995).
42. Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *J. Morphol.* **88**, 49–93 (1951).
43. Downes, K. M. & Davies, T. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development* **118**, 1255–1266 (1993).
44. O'Rahilly, R. & Muller, F. *Developmental Stages in Human Embryos Including a Revision of Streeter's 'Horizons' and a Survey of the Carnegie Collection* (Carnegie Institute, Washington DC, 1987).
45. Anonymous. Nomina Anatomica, Nomina Histologica, Nomina Embryologica (Churchill Livingston, Edinburgh, 1992).



## Acknowledgements

The Edinburgh Mouse Atlas Project (EMAP) is a collaboration between D.D. and R.B. at the MRC Human Genetics Unit and M. Kaufman and J. Bard at the Department of Biomedical Sciences, University of Edinburgh. EMAP has been advised by networks with grants from the European Science Foundation and the Wellcome Trust. We thank the FlyBase and the WormBase consortiums for permission to use material from these databases and V. Hartenstein, R. Jacobs, F. Verbeek and T. Strachan for sharing information about work in progress.