Mamm Genome (2007) 18:402–411
DOI 10.1007/s00335-007-9027-3

Systems developmental biology: the use of ontologies in annotating models and in identifying gene function within and across species

Jonathan Bard

Received: 6 March 2007 / Accepted: 12 March 2007 / Published online: 14 June 2007

Abstract Systems developmental biology is an approach to the study of embryogenesis that attempts to analyze complex developmental processes through integrating the roles of their molecular, cellular, and tissue participants within a computational framework. This article discusses ways of annotating these participants using standard terms and IDs now available in public ontologies (these are areas of hierarchical knowledge formalized to be computationally accessible) for tissues, cells, and processes. Such annotations bring two types of benefit. The first comes from using standard terms: This allows linkage to other resources that use them (e.g., GXD, the gene-expression [G-E] database for mouse development). The second comes from the annotation procedure itself: This can lead to the identification of common processes that are used in very different and apparently unrelated events, even in other organisms. One implication of this is the potential for identifying the genes underpinning common developmental processes in different tissues through Boolean analysis of their G-E profiles. While it is easiest to do this for single organisms, the approach is extendable to analyzing similar processes in different organisms. Although the full computational infrastructure for such an analysis has yet to be put in place, two examples are briefly considered as illustration. First, the early development of the mouse urogenital system shows how a line of development can be graphically formalized using ontologies. Second, Boolean analysis of the G-E profiles of the mesenchyme-to-epithelium transitions that take place during mouse development suggest Lhx1, Foxc1, and Meox1 as candidate transcription factors for mediating this process.

Introduction

Up until the 1980s, most research in developmental biology involved analyzing the interactions among and within the tissues that participated in some embryologic event (e.g., limb development) and, on the basis of careful experimentation, inferring something about these interactions. A second and complementary approach was to use kinetics and other theoretical approaches to model a problem in development such as patterning. In either case, where there was more than one possible explanation of a phenomenon, it seemed obvious and sensible to give preference to the explanation that seemed the most parsimonious on grounds of natural selection. The gradual and continuing discovery of the intricacy of the signalling conversations between participating tissues, the richness of the activated molecular networks that regulate developmental change, and the complexity of the resulting processes have shown just how naïve was that original paradigm.

Over the past two decades, our ability to use a wide range of molecular technologies to investigate these regulatory networks and to collate the patterns of gene expression characterizing a particular state of differentiation has produced enormous amounts of information, often accessible from online databases (e.g., http://www.informatics.jax.org), on how development proceeds. This ability to exploit the new technologies and so to explore complex developmental events at the molecular level has enabled the field, over a period of some 20 years, to progress from a

J. Bard (✉)
School of Biomedical Sciences, Edinburgh University,
Edinburgh EH8 9XD, UK
e-mail: j.bard@ed.ac.uk

small-scale subject interesting relatively few scientists to an area of major interest and excitement across the world. One stimulus here has been the realization that mutation-derived errors in these networks underpin many human congenital abnormalities. The consequent study of these abnormalities, often using mouse models, has the dual benefit of advancing medical research and giving us a tool to pry open these networks. A second has been the realization that homologous networks do similar things in very different organisms and that we therefore have a means to explore the mechanisms of evolutional change which usually operate, as Waddington was probably the first to emphasize, through mediating changes in development (see below and Waddington 1975).

All this work has led to a wonderful increase in our understanding of developmental events, particularly those that involve signalling and those in which the activation of a transcription factor initiates a new process (for review, see Gilbert 2006). That said, it has to be admitted that, for most developmental events, there are now large amounts of molecular expression data that are hard to interpret unambiguously. Often we do not really know in a particular event which proteins are important, which are secondary, and which are background, and knockout and other experimental data can be either ambiguous or unhelpful. In one sense, the situation is worse than it was in the 1980s: Then we could appeal to parsimony via natural selection to make choices; now things are so complicated that we have no means of recognizing parsimony, and would not trust the concept anyway.

One approach to this complexity is to say that if only we had enough data, everything would become clear, but it is unlikely that anyone in the field really believes this. A second is to say that we need better and stronger intellectual frameworks than just relational databases for organizing and analyzing the new data that are pouring out of laboratories. A third is to take the view that we need not just a better framework for handling data, but better intellectual ideas. The third view is certainly right, but those ideas have yet to emerge and, in the absence of some deeply original and intuitive thinking, may well emerge from the second approach, which is hard enough at the moment and which is articulated under the general name of *systems developmental biology*. This approach is new and does not yet have any formal structure but, in general, seeks to embed the events of a particular developmental event within a computational and hierarchical framework that links tissues, cells, processes,and molecular/genomic data, and often aims to capture the results of high-throughput technology (e.g., Kimelman 2006). Perhaps the best-known example of a systems approach is the work on sea-urchin development (e.g., Ben-Tabou de Leon and Davidson 2006), which integrates tissues, genes, and networks (Longabough et al. 2005). Other important systems approaches include analyses of developmental networks (Xia et al. 2006) and the molecular basis of very early mouse development (Eviskov et al. 2004).

This article does not seek to provide a systems approach to any particular phenomenon but to consider how best to take advantage of the computational tools currently available so as to ensure that systems descriptions based on tissues, cells, and processes can be interoperable in the sense that they use a common language. This would enable them to query one another and use each other's formal knowledge (much as we can do for genes and proteins that are already linked through their IDs to their appropriate database). The key tools here are ontologies and the purpose of this article is to discuss what ontologies are, how they can be used in formalizing systems approaches to development within and across species, and what are the resulting benefits.

Ontologies of anatomical tissues and of cell types

At the core of development is the predictable production of functional and differentiated tissues from early, less well-defined tissues. It would therefore be sensible if, when one person uses, for example, the term ''E14.5 mouse left atrium'' in his systems model of heart development, another person using the same term in her model can link to that of the first. The way that such linkage is done for proteins is to use an ID from a standard database (e.g., the protein ID from Uniprot, http://www.ebi.uniprot.org), and because proteins are all amino-acid strings and hence of the same rank, they can readily be stored in the tables of relational databases.

Anatomical tissue organization, in contrast, is hierarchical in nature: The vertebrate hindlimb, for instance, is obviously partitioned into regions (thigh, knee, calf, foot), each of which has its own parts, and the concept of ''hindlimb'' would naturally be expected to include these subordinate parts, together with information about their relationship to the hindlimb and to one another. While it is obviously straightforward to assign a unique ID to a given tissue at a given developmental age, it is clear that the hierarchical organization of tissues poses some organizational problems beyond those needed for handling sequence data.

The way that such hierarchical information is most appropriately handled is through *ontologies*. These are domains of knowledge formalized in a way that allows them to be computationally accessible. In practice, ontologies are built up by linking *facts* in a hierarchical way. Here, a *fact* is a triad of the general form <term><*relationship*><term> and terms can have parents and children (e.g., the E14.5 left atrium *is part of* the E14.5 heart; the E14.5 heart *is part of* the E14.5 cardiovascular system, etc.). Although they are tedious to produce (even the simplest organ system has a great many tissues and a lot of organization), there are now *part-of* ontologies for the tissues of all the main model adult organisms and for the developmental anatomy of the mouse, zebrafish, and *Drosophila* (accessible from the Open Bio-Ontologies site, http://www.obo.sourceforge.net). Every term in these ontologies carries a standard ID of the form <abcd><ijkl>, where abcd gives a short letter code for the ontology (e.g., EMAP for mouse development) and ijkl gives the number for a specific tissue at a specific developmental age (e.g., EMAP:7917 is the ID for the E14.5 mouse left atrium, with EMAP standing for the Edinburgh Mouse Atlas Project, http://www.genex.hgu.mrc.ac.uk). It is these IDs that allow for interoperability because they represent defined concepts (or terms) that can be used anywhere, even as synonyms.

It is worth noting that such an anatomical ontology is more than just the list of the parts as it includes a great deal of knowledge about how these parts are organized into larger structures and these larger structures into organ systems (e.g., Fig. 1). Such an ontology may also include additional knowledge built on other relationships such as *derives from* (an ontology of developmental anatomy would well include lineage relationships) and type data (e.g., the femur *is a* bone). There is also no reason why a child should have only a single parent in the ontology: For example, it is equally appropriate to describe the femur as <*part of*><the skeleton> as <*part of*><the hindlimb>, and a rich ontology could well include both relationships (and this multiparenting of terms means that it would be called by the technical term *Directed Acyclic Graph*, or *DAG*). This is not the place to include a detailed discussion of how anatomical ontologies are built and used (the interested reader should consult Bard 2005), but it should be mentioned that the internal organization of an anatomy ontology is usually rather complex (the structure needs to be able to handle many relationships as well as definitions and links) and is best read in a browser program such as OBO-Edit or COBrA (Figs. 1 and 2; Aitken et al. 2004; Harris et al. 2004) that is visualized in a GUI rather than as a list on paper. There are several languages in current use for handling ontologies (the best known are OBO and OWL) and they can be translated into each other using the COBrA tool.

In the context of systems developmental biology and in addition to the appropriate anatomy ontology, there are two general ontologies that are also useful. The first is the Cell-Type Ontology (Bard et al. 2005) and the second is the Gene Ontology (Ashburner et al. 2000; Harris et al. 2004).

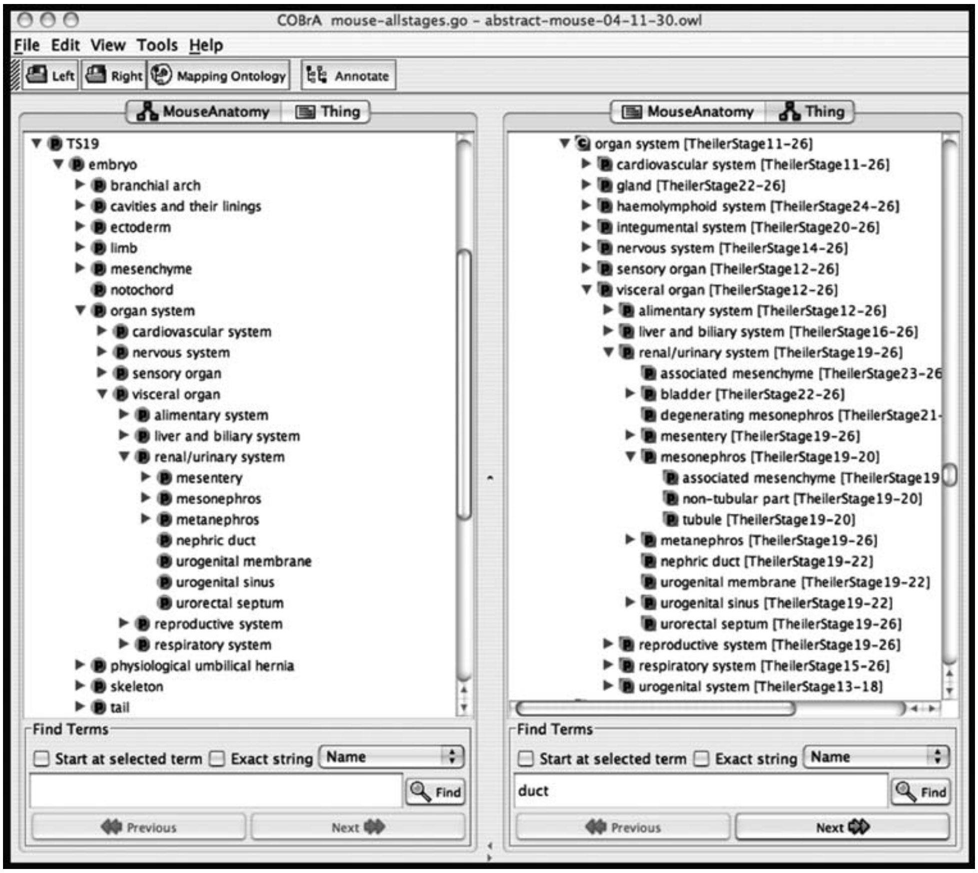
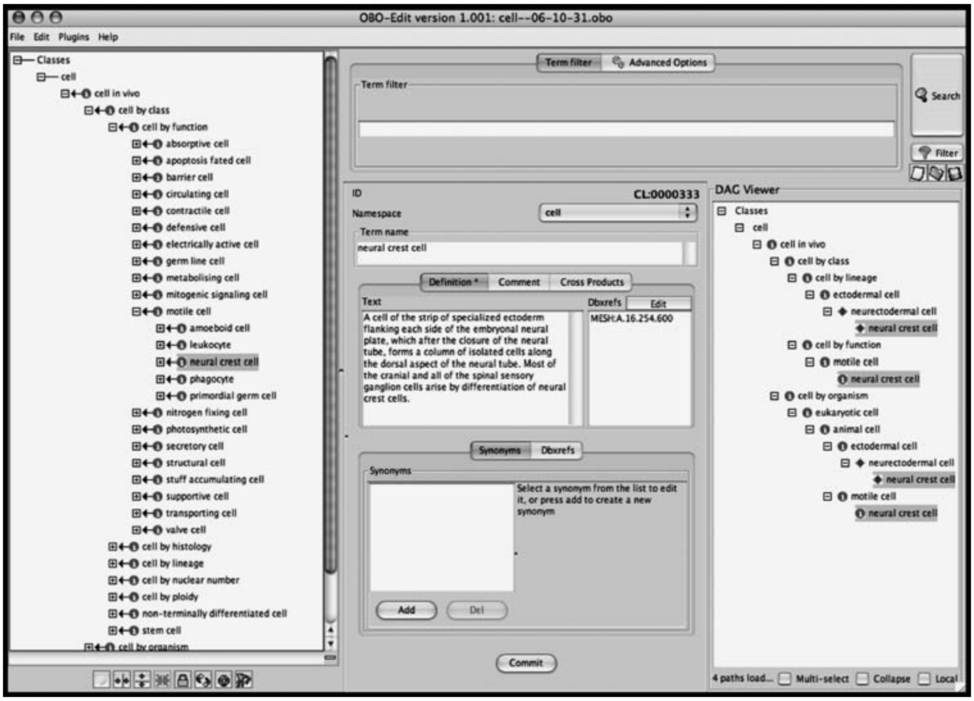


Fig. 1 The ontology of mouse developmental anatomy as displayed in the COBrA browser. The left panel shows the tissues in the metanephros of the TS 19 (E11.5 mouse). These tissues carry an EMAP ID. The right panel shows the ontology of what is called the ''abstract mouse'' and includes all the tissues with range of times at which they are present during embryogenesis. These terms carry an EMAPA ID

```
▼ TS19
  ▼ embryo
    ▶ branchial arch
    ▶ cavities and their linings
    ▶ ectoderm
    ▶ limb
    ▶ mesenchyme
      notochord
    ▼ organ system
      ▶ cardiovascular system
      ▶ nervous system
      ▶ sensory organ
      ▼ visceral organ
        ▶ alimentary system
        ▶ liver and biliary system
        ▼ renal/urinary system
          ▶ mesentery
          ▶ mesonephros
          ▶ metanephros
            nephric duct
            urogenital membrane
            urogenital sinus
            urorectal septum
        ▶ reproductive system
        ▶ respiratory system
    ▶ physiological umbilical hernia
    ▶ skeleton
    ▶ tail
```

```
▼ organ system [TheilerStage11-26]
  ▶ cardiovascular system [TheilerStage11-26]
  ▶ gland [TheilerStage22-26]
  ▶ haemolymphoid system [TheilerStage24-26]
  ▶ integumental system [TheilerStage20-26]
  ▶ nervous system [TheilerStage14-26]
  ▶ sensory organ [TheilerStage12-26]
  ▼ visceral organ [TheilerStage12-26]
    ▶ alimentary system [TheilerStage12-26]
    ▶ liver and biliary system [TheilerStage16-26]
    ▼ renal/urinary system [TheilerStage19-26]
        associated mesenchyme [TheilerStage23-26
      ▶ bladder [TheilerStage22-26]
        degenerating mesonephros [TheilerStage21-
      ▶ mesentery [TheilerStage19-26]
      ▼ mesonephros [TheilerStage19-20]
          associated mesenchyme [TheilerStage19
          non-tubular part [TheilerStage19-20]
          tubule [TheilerStage19-20]
      ▶ metanephros [TheilerStage19-26]
        nephric duct [TheilerStage19-22]
        urogenital membrane [TheilerStage19-22]
      ▶ urogenital sinus [TheilerStage19-22]
        urorectal septum [TheilerStage19-26]
    ▶ reproductive system [TheilerStage19-26]
    ▶ respiratory system [TheilerStage15-26]
    ▶ urogenital system [TheilerStage13-18]
```

duct



Term filter

ID CL:0000333

Namespace cell

Term name
neural crest cell

Text
A cell of the strip of specialized ectoderm flanking each side of the embryonal neural plate, which after the closure of the neural tube, forms a column of isolated cells along the dorsal aspect of the neural tube. Most of the cranial and all of the spinal sensory ganglion cells arise by differentiation of neural crest cells.

Dbxrefs
MESH:A.16.254.600

Synonyms
Select a synonym from the list to edit it, or press add to create a new synonym

Fig. 2 The Cell-Type Ontology visualized in the OBO-Edit browser. The left panel shows the ontology with its classes for cell types with the motile cell entry opened and neural crest cell highlighted. The central panel gives the ID and the definition, while the right panel shows the various places within the graph where the neural crest cell can be found

The former, unlike the anatomy ontologies, not only includes all the common and many of the uncommon cell types that are found across the phyla but it is essentially species-independent and so facilitates cross-species analyses and comparisons. This ontology is structured to include our knowledge of the many properties of these cell types and each is separately coded under function, morphology, ploidy, development, etc., using two relationships, *is-a* and *descends-from* (see Fig. 2). This ontology is thus a terse summary of a great deal of knowledge about cell types and their properties.

The Gene Ontology or GO is by far the best known and most used of the standard bio-ontologies (it is used for protein annotation in Uniprot). Unlike Uniprot, it does not include sequence information but focuses on the properties of proteins and includes hierarchical knowledge about (1) cellular locations, (2) molecular functions, and (3) the functional processes in which they are involved. For systems developmental biology, it is the latter that is the most important and the process hierarchy includes a wide variety of developmental processes (although they are distributed across the ontology rather than integrated under a single heading [Fig. 3]), each of which, of course, has a unique ID. Of particular interest here is the database of proteins that is linked to the GO so that a user can easily identify all the stored proteins associated with a GO term, or the GO terms associated with a chosen protein (although it should be said that keeping this database up-to-date is a major task).

One important factor about ontology terms is that they can be associated with data (usually held in a standard relational database and linked to the ontology via the appropriate IDs); examples include the proteins that satisfy the definition of a GO term (http://www.godatabase.org), the genes expressed in a particular mouse tissue at a particular time, (http://www.informatics.jax), and the micrographs associated with a pathologic state (http://www.pathbase.net). Here, the hierarchical knowledge within the ontology comes into play: If, for example, a user requires the genes associated with the developing mouse forelimb at E12.5, the response comes from searching the ontology to identify the constituent tissues in the limb and using their IDs to collect all the associated data. This can be done because this type of *part of* relationship has the property known as *upwards propagation*. This means that if a term has data associated with it, then these data can be associated with the parent (e.g., a gene expressed in the tarsus is also expressed in the hindlimb). Propagation is associated with some bio-ontology relationships (e.g., *part of, is a*) but not with others (e.g., *develops from*; one would not expect pigment cells to have the same properties as their neural-crest-cells precursors).

Using ontology terms for annotating systems models for mouse development

Ontologies, together with their linked data, provide an important online resource and have several key roles in systems developmental biology. The first is the use of well-defined terms (with their associated IDs) to standardize

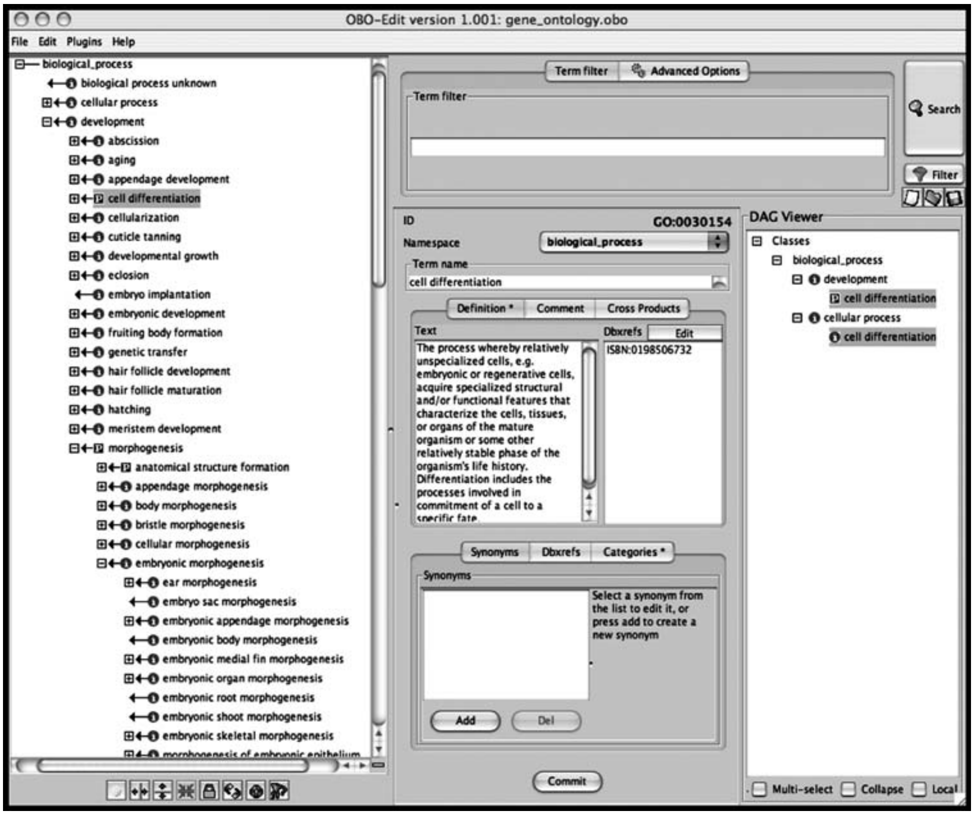


Fig. 3 The process component of the Gene Ontology (left panel) visualized in the OBO-Edit browser. The cell differentiation term is highlighted (left panel) and found in two contexts (right panel). The center panel gives the definition and the ID

ID GO:0030154

Namespace biological_process

Term name
cell differentiation

Text
The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history. Differentiation includes the processes involved in commitment of a cell to a specific fate.

Dbxrefs Edit
ISBN:0198506732

Synonyms
Select a synonym from the list to edit it, or press add to create a new synonym

□ Multi-select □ Collapse □ Local

annotation, the second is for linkage to databases that store data associated with the terms, and the third is to facilitate the identification of similar terms in very different contexts. These ideas are explored here and in the next section and, while the approach is applicable to the development of any organism and also across organisms (see Discussion), the examples focus on the mouse. This is because our knowledge about its development is now so deep that it is often possible not only to describe how any tissue develops morphologically over time (Kaufman and Bard 1999) but to identify the processes and changes in cell type that underpin each time slice of a tissue's development (see http://www.xspan.org). In addition, the mouse community is fortunate to have access to substantial online informatics resources that are available from The Jackson Laboratory. In the context of this article, the most important of these is GXD, a database of gene-expression (G-E) data for the developing mouse in which expression data are annotated with (and hence searchable by) tissue name, developmental stage, and GO IDs, as well as other genetic identities.

Although development is complicated, it can be seen as the operation on tissues of relatively few core processes that involve

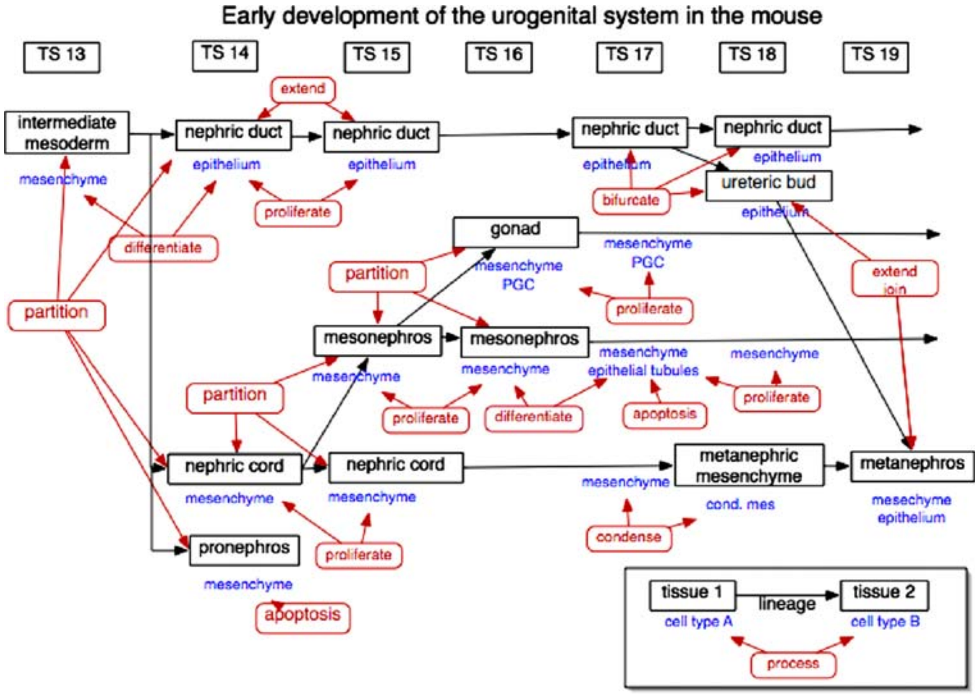
- Patterning – this sets up future events in groups of cells
- Proliferation and apoptosis – the basis of growth and shaping
- Cell differentiation – changing a cell's phenotype
- Morphogenesis – the generation of spatial organization (e.g., via movement)

with each process having subprocesses. Such processes are archived in the Gene Ontology (see above and Fig. 3).

If the formation of a system is to be modeled, then the first step is to lay out its normal pattern of development graphically. Much of the stage-by-stage lineage data for mouse embryogenesis is available in text format (Kaufman and Bard 1999) and can be linked to the tissues (with their IDs) in the ontology of mouse developmental anatomy (Bard et al. 1998) and hence with GXD. Staging of mouse embryos is based on the appearance of standard external identifying features as embryogenesis proceeds; Theiler staging for the mouse gives, in essence, two stages a day when things are going rapidly (E6–E12.5) and one stage a day when the appearance of new features is slower (E1–E5 and E12.5 onward). Annotating the tissue names is straightforward because each tissue at each stage has a unique ID accessible from the ontology of mouse developmental anatomy (e.g., Fig. 1).

For annotating changes in the state of cellular differentiation, there are two options. The first is to use an appropriate GO term, but because the GO includes less than 50 cell types, it cannot (and was not intended) to do justice to developmental anatomy. A better option, there-

Fig. 4 A systems model of mouse urogenital development during the periods TS 13–19 (E8.5–E11.5) showing the tissues present at each time interval, the cell types, and the processes driving change, all of which have unique ontology IDs (for further details see http://www.bioontology.org/wiki/images/1/1a/CARO-UG-development-JB.pdf)



fore, is to use the higher-level GO term ‘‘cell differentiation’’ and combine it with two terms from the Cell-Type Ontology. The annotation would thus be

$$<\text{cell type }1> < \textit{cell differentiation} > <\text{cell type }2>$$

In the case of metanephric mesenchyme being induced to form nephron epithelium, the annotation is

$$<\text{mesenchyme}> <\textit{cell differentiation}> <\text{duct epithelium}>$$

or, using the appropriate IDs from the two ontologies:

$$<\mathrm{CL}:0000134> <\mathit{GO}:0030154> <\mathrm{CL}:0000068>$$

Where the state of a tissue changes between two Theiler periods, one can annotate the developmental change that drives this transition (this is not the usual way in which development is considered!) with the appropriate GO process terms. In this way, the final graph has, superimposed on the lineage flow of developmental anatomy, the appropriate differentiation and process terms that drive the development of each tissue. Underpinning each of these transitions is the appropriate ontology ID, so that the final graph is set up to be complete, formal, and interoperable.

Figure 4 illustrates the result of annotating in this way the development of the mouse urogenital system over Theiler stages 13–19 (E8.5–E11.5), where classical descriptive embryology has, first, shown that the intermediate mesoderm differentiates to give the nephric duct, mesonephros (from which develops the gonad) and the metanephros, and, second, given the cell types associated with each tissue. The graph also includes the results of experimental work that has clarified the processes that push development from one stage to the next. For clarity, the figure excludes the IDs but they are all readily available from the appropriate ontologies.

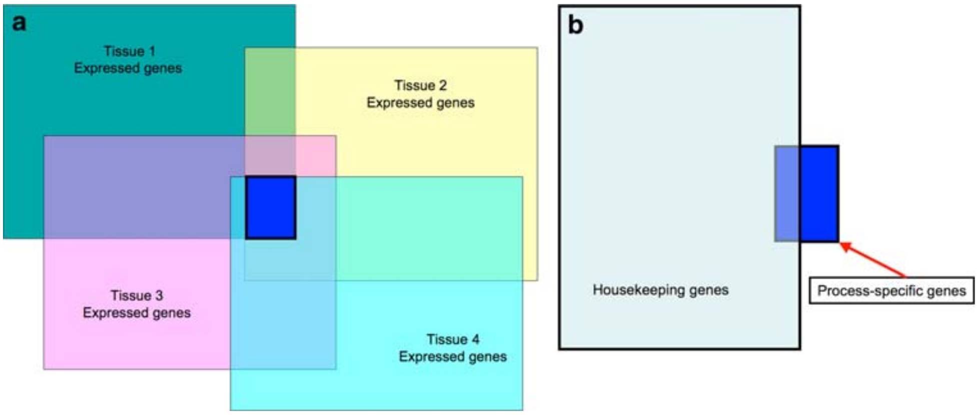
There is one immediate use of this model that derives from annotating terms with standard ontology IDs. The graph as it stands has no molecular data, but all the current gene-expression information associated with a particular developing mouse tissue at a given Theiler stage is computationally accessible from GXD through ID interoperability. GXD genes also carry GO IDs which enables searches to be quite sophisticated. It is straightforward, in principle at least, to use these GO IDs to identify, for example, signals and receptors for tissues that signal to one another (Bard 2002) or transcription factors that are synthesized at a particular stage and ready for a future event.

In short, ontology annotations of developmental systems are not only the key to interoperability and standardization of systems models, they give rich searching possibilities.

The genes underpinning common processes

There is a further bonus from such annotations: As development proceeds, the same developmental processes are used in very different contexts within one organism. This similarity goes beyond the differentiation of the same cell type from different tissues (e.g., neurons can differentiate directly from neuroepithelium and indirectly after

Fig. 5 Diagrammatic representation of the Boolean subtractions of (left) the gene-expression patterns of four tissues that undergo the same developmental process; this gives a mixture of process genes and common (housekeeping) genes. A second subtraction (right) eliminates housekeeping genes



the migration of cells originally from the neural crest or from epithelial placodes). Obvious examples are the branching of epithelial tubules (in glands and in the vascular system), epithelial folding in its many forms, the forming of mesenchymal condensations (the first step in the development of muscles, bones, and cartilage), and the initiation of movement (in tissues as different as neural crest cell, primary germ cells, neurons, and gastrulating epiblast cells), pigmentation (retinal epithelium, neural crest cells). Indeed, such processes are common to development across the phyla.

Consider the hypothesis that each of these processes can be viewed as a ‘‘motor’’ driven by the activation of particular set of transcription factors (TFs). If this hypothesis is correct, then each set of tissues that are about to participate in a particular event should express those TFs and they should be present in the appropriate G-E profile in the associated database (they may also be missing, but they should not have been shown to be absent). If so, then the overlap of the G-E profiles of tissues about to initiate a particular process should include (1) those proteins involved in initiating that process and (2) housekeeping proteins common to all (or at least most) cells. If these housekeeping genes can be excluded, such Boolean analysis should yield key proteins involved in that process. A similar analysis of the G-E profiles for those tissues immediately after they have initiated a particular process should yield those proteins involved in that process.

Any analysis along the lines suggested makes several assumptions beyond that of common TFs underpinning common processes. First, the time resolution of the G-E database has to be fine enough to discriminate between the period of a tissue’s *competence* to undergo a process and the process itself. In the case of mouse development for which the database archives expression by Theiler staging, this means time slices of 24 hours for early and late mouse development and 12 hours over the period E6–E12; this is probably adequate. Second, the database needs to contain enough data on the expression of all relevant genes. This latter criterion is unlikely to be met, even for GXD. While this rich resource currently includes some 250,000 expression results for about 7500 genes (information courtesy of Dr. Martin Ringwald), the data are not uniformly distributed across tissues or time slots. There is thus an element of chance as to whether the database holds information about the expression of a gene in a particular tissue at a given time.

A preliminary analysis of the G-E data associated with some process widely undergone during mouse development should thus be based on the following lines:

1. Collate all the G-E data for each tissue in the 24 hours leading up to the process (a reasonable estimate of the period of competence); this list should include the *infrastructure* proteins for establishing that process.
2. Collate all the G-E data for each tissue during the stage at which the process is initiated, and probably the following one; this list should include all the *process* genes.
3. Identify the overlaps of the G-E patterns for steps 1 and 2. Given that GXD is incomplete, this probably means, in the first instance, including any protein expressed before or after the process in more than one tissue (Fig. 5), and particularly any of the latter whose expression is initiated just before the process is initiated (these are candidate genes for being activated by the TFs) .
4. Remove any of these proteins that can be identified as a housekeeping gene (this can be done from a standard list or perhaps from the G-E overlap of very different tissues); this will give the candidate *infrastructure* and *process* genes.
5. Analyze these candidate populations to see which genes (a) are heavily represented and (b) seem important (e.g., TFs); this may involve Bayesian statistical analysis.

The analysis as a whole should provide a set of candidate genes for the process of interest (provided that the group of tissues as a whole is undergoing no more than one common development-specific process). However, the

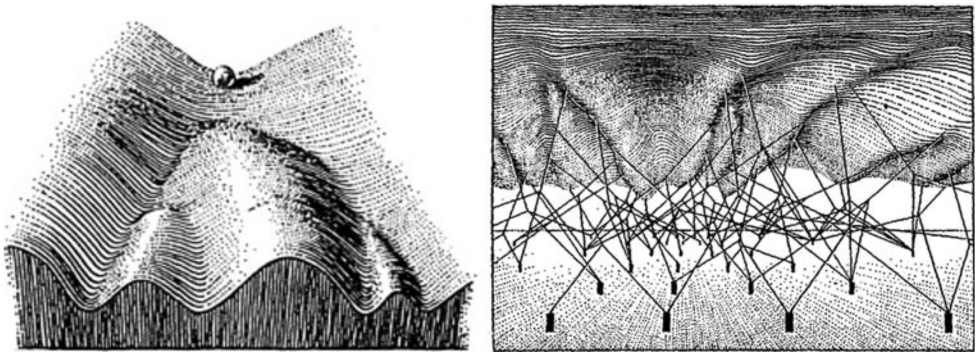


Fig. 6 **Left** Waddington's original drawing of the epigenetic landscape showing the developmental trajectory of a cell being shaped by its tissue environment. **Right** A later drawing showing that this environment is itself underpinned by complex genetic interactions (Waddington 1940, 1957, with permission from Cambridge University Press)

process is quite lengthy and, given that GXD may include several hundred expressed genes for even a single tissue at a specific Theiler stage, can really only be properly handled through a substantial computational infrastructure that has yet to be put in place. Indeed, the more expressed genes that can be associated with the set of tissues, the more reliable will be the analysis. The situation will get better but more complicated once GXD includes microarray and other high-throughput data.

Fortunately, there is a relatively simple shortcut that can be used for a quick exploration of the approach, and which takes advantage of the GO annotations in GXD. If one merely restricts one's searches to (1) the periods of competence of tissues about to initiate a process and (2) genes with a GO transcription factor ID, the output should be restricted to those TFs associated with the initiation of that process. As an example, consider the mesenchyme-to-epithelium transition that takes place many times during development. A preliminary examination (full details will be published elsewhere) of the gene-expression profiles in GXD shows that there are substantial entries for the formation of blood vessels in the early mouse heart, the differentiation of heart endocardium, the metanephric ducts, the mesenchyme that forms the mesonephric ducts, and the early stages of somite development. If the search is restricted, using GO IDs, to TFs in the participating tissues in the two Theiler stages before these transitions take place, the data show that three TFs, Lhx1, Foxc1, and Meox1, are present in all these tissues (apart from a couple where there is incomplete data). A further inspection of the complete distributions of these genes shows that their expression (insofar as it is fully represented in GXD) is highly restricted over space and time, and because they do not in general overlap one another, they cannot be considered as housekeeping genes; their coexpression is hence unlikely to be a coincidence.

The TFs Lhx1, Foxc1, and Meox1 are thus, as a set, good candidates for collectively initiating a mesenchyme-to-epithelium transition, although they seem not to have been previously identified as fulfilling this role. It is therefore a prediction that this set be expressed in other tissues undergoing a mesenchyme-to epithelial transition. Examples that might be worth investigating here include the stromal fibroblasts in the cornea that become the corneal endothelium and the splanchnopleure mesoderm that forms mesothelium (GXD currently includes no relevant expression data for these tissues). If the prediction were confirmed, it would be worth investigating which proteins were synthesized following their activation.

Discussion

Although systems developmental biology is thought of as a relatively new subject, its basis lies in an idea that Waddington originally had in the late 1930s and that he expressed graphically as *the epigenetic landscape* (Fig. 6). In its original form (left), the picture was of a ball rolling down the valleys of a complex hillside. In its final form (right), this picture showed that the topology of this surface was shaped through complex linkage to a set of pegs on an underlying flat surface. The meaning of this metaphor is that the developmental trajectory of a developing cell over time (the rolling of the ball down the valleys) is shaped by its environment (the undulating surface), with the form of the landscape being determined by the interacting properties of many genes (the pegs and their ties). Local development was thus viewed as a gene-based interaction between a particular group of cells and its environment and the system had to be viewed as a whole. This was a startlingly original view to hold more than 50 years ago, at a time when the scientific community almost uniformly held the view that all genes did was code for enzymes and such trivialities as eye color (Van Speybroeck 2002). Its value as a metaphor was shown by Waddington's use of it to describe evolution: Small changes (mutations) in genes led to changes in the landscape and hence to altered patterns of development and so to novel organisms (see Waddington 1975).

We are now beginning to catch up with Waddington's thinking. Systems models are starting to be produced that aim to integrate the complexity of the molecular, cellular, and tissue details that underpin development using the computational resources that are now available. There will be many more such models, and, given the richness and complexity of development, they are bound to overlap. It is important that such overlaps allow interoperability, and a key point made in this article is that the community should not only use, as it already does, the terms and IDs for the gene and protein databases, but also incorporate the terms and IDs for cells, tissues, and processes that are to be found in standard bio-ontologies. This is partly for interoperability across models, but also to allow direct linkage to such databases as those handling G-E data.

This article also points out that there is an additional bonus from using these IDs, i.e., where the same process is used in the development of different tissues, the linking of tissue IDs to their associated gene-expression profiles can, in principle, lead through Boolean analysis to the identification of candidate genes associated with the initiation and execution of this process. The databases are currently populated with genes whose roles are still unclear so this computational approach complements experimental approaches because it enables small groups of genes to be linked to the initiation and execution of processes. This contrasts with the analysis of individual genes whose roles can be analyzed using, for example, transgenic technology and high-throughput technology that picks up a large numbers of genes but yields little about their function.

A further point to be made is that the type of computational approach to the identification of gene function given here allows us to test the hypothesis that common processes are underpinned by common TFs. If such a search yielded several candidate TFs that are found to be associated with the initiation of a process in some tissues but have yet to be found in all tissues, it suggests that the expression of these TFs should be further examined in these other tissues. A lack of expression there would cast doubt on or at least narrow the extent of the hypothesis. This approach to systems biology thus provides assays for testing our ideas.

In this article, the focus has been on formalizing mouse development and analyzing the molecular underpinnings of its underlying processes because it is for this organism that the associated expression database has the finest spatial and temporal granularity. It should of course be pointed out that the approach is equally applicable to other organisms and even across organisms that share equivalent developmental processes. In the first instance, the mouse can be used as a model for identifying process-associated genes. Where a similar process occurs in other organisms, the homologs will be candidate genes for that process (and the XSPAN facility, http://www.xspan.org, will be helpful in identifying equivalent tissues in model organisms). A further tool under development that may be useful in this context is CARO, the Common Anatomy Reference Ontology (http://www.obo.sourceforge.net/cgi-bin/detail.cgi?caro, Haendel et al. 2007) which aims to provide interoperability across species-specific anatomy ontologies. In the longer term, one can envision the construction of complex systems models that span organisms and that employ the full richness of the computational resources that are available.

At a slightly deeper level, what distinguishes systems developmental biology from other approaches to unpicking the complexities of development is the formalization of the events of embryogenesis. This in turn enable tissues, cells, and expressed genes to be linked in a way that lends to computational as well as other forms of analysis. The exercise of formalizing embryogenesis encourages the biologist to think in ways that complement other more traditional approaches.

Acknowledgment The author thanks Stuart Aitken for the many discussions about ontologies and for his computational help.

References

Aitken S, Korf R, Bard J (2004) COBrA: a bio-ontology editor. Bioinformatics. 21:825–826

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Bard JBL (2002) Growth and death in the developing mouse kidney: signals receptors and conversations. BioEssays 24:72–82

Bard JBL (2005) Anatomics: the intersection of anatomy and bioinformatics. J Anat 206:1–16

Bard JBL, Kaufman MA, Dubreuil C, Brune RM, Burger A, et al. (1998) An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. Mech. Dev 74:111–120

Bard J, Rhee SY, Ashburner M (2005) An ontology for cell types. Genome Biol 6:R21

Ben-Tabou de Leon S, Davidson EH (2006) Deciphering the underlying mechanism of specification and differentiation: the sea urchin gene regulatory network. SciSTKE 2006(361):pe47

Evsikov AV, de Vries WN, Peaston AE, Radford EE, Fancher KS, et al. (2004) Systems biology of the 2-cell mouse embryo. Cytogenet Genome Res 105:240–250

Gilbert SF (2006) Developmental Biology, 8th ed. (Sunderland, MA: Sinauer Associates)

Haendel MA, Neuhaus F, Osumi-Sutherland DS, Mabee PM, Mejino JLV, Mungall CJ, Smith B (2007) CARO – the common anatomy reference ontology. In: Burger A, Davidson D, Baldock R (eds) Anatomy ontologies for bioinformatics, principles and practice. Springer verlag, Heidelberg. In press

Harris MA, Clark J, Ireland A, Lomax J, Ashberner M, et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261

Kaufman MH, Bard JBL (1999) The Anatomical Basis of Mouse Development. (London: Academic Press)

Kimelman D (2006) Mesoderm induction: from caps to chips. Nat Rev Genet 7:360–372

Longabaugh WJ, Davidson EH, Bolouri H (2005) Computational representation of developmental genetic regulatory networks. Dev Biol 283:1–16

Van Speybroeck L (2002) From epigenesis to epigenetics. Ann N Y Acad Sci 981:61–81

Waddington CH (1940) Organisers and genes (Cambridge: Cambridge University Press)

Waddington CH (1957) The strategy of the genes. (London: Allen & Unwin)

Waddington CH (1975) Evolution of an evolutionist. (Edinburgh: Edinburgh University Press)

Xia K, Xue H, Dong D, Zhu S, Wang J, et al. (2006). Identification of the proliferation/differentiation switch in the cellular network of multicellular organisms. PLoS Comput Biol 2:e145