BIOINFORMATICS

Vol. 20 no. 2 2004, pages 259–267
DOI: 10.1093/bioinformatics/btg400

Formalization of mouse embryo anatomy

Albert Burger, Duncan Davidson and Richard Baldock*

MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, Scotland, UK

Received on February 5, 2003; revised on March 20, 2003; accepted on June 8, 2003

ABSTRACT

Motivation: The Edinburgh Mouse Atlas and Gene Expression Database project has developed a digital atlas of mouse development to provide a spatio-temporal framework for spatially mapped data such as *in situ* gene expression and cell lineage. As part of this database, a mouse embryo anatomy ontology has been created. A formalization of this anatomy is required to document its precise semantics and how it is used in the context of the Mouse Atlas.

Results: The paper describes the existing anatomy ontology and formalizes aspects of it using a predicate logic based approach. It therefore provides a guide for users of the current version of the ontology, as well as the basis for a description of the anatomy using an ontology language, such as OWL, thus enabling future work on reasoning about the Mouse Atlas in the context of an intelligent gene expression bioinformatics workflow system. The logic has been implemented in a Prolog prototype.

Availability: The Mouse Atlas is available on-line at http://genex.hgu.mrc.ac.uk

Contact: Albert.Burger@hgu.mrc.ac.uk

1 INTRODUCTION

The Edinburgh Mouse Atlas (EMAP) and Gene Expression (EMAGE) Database project (Brune *et al*., 1999; Davidson and Baldock, 2001; Davidson *et al*., 1997; Ringwald *et al*., 1994) (*Mouse Atlas**—an asterisk indicates a URL; Table 2) has developed a digital atlas of mouse development, which provides a bioinformatics framework to spatially reference biological data. The core databases contain three-dimensional (3D) grey-level reconstructions of the mouse embryo at various stages of development, a systematic nomenclature of the embryo anatomy (the anatomy ontology), and defined 3D regions (domains) of the embryo models that map the anatomy ontology on to the spatial models. Through the 3D domains, users can navigate from the spatial representation of the embryo to the ontology and vice versa. Data from an *in situ* gene expression database is spatially mapped on to the atlas allowing the users to query gene expression patterns using the 3D embryo model and/or the ontology as a reference.

As with all developments of ontologies, there is a trade-off between the effort that can be expended on its creation and the level of formalization and detail that can be achieved, the current version of the mouse anatomy ontology is relatively informal. With a view to integrate the Mouse Atlas system with other bioinformatics resources, and as part of the ongoing efforts to develop the ontology further, work has been carried out in formalizing aspects of the current representation of the anatomy, giving a more precise description of its semantics. This description will aid other researchers in making correct use of the currently available version of the ontology. It also provides the basis for further development work, such as the representation of the mouse embryo anatomy using an ontology language, e.g. *OWL**.

The work presented here is only an example of a larger effort by the bioinformatics community to produce useful biological ontologies. For example, the Gene Ontology Consortium (*GO**), has published ontologies for *molecular functions*, *biological processes* and *cellular components*. A number of other shared vocabularies for use within genomics and proteomics can be found on the open Biological Ontologies (*OBO**) web site.

In this paper, we will not discuss the respective advantages and disadvantages of different ways of representing anatomy ontologies. The purpose here is not to propose a general anatomy ontology structure, but to report on an existing anatomy ontology and its use in a real bioinformatics application. For related work on anatomy ontologies, we refer the interested reader to the relevant literature, such as work carried in the *GALEN** and *Visible Human** projects.

The remainder of this paper is organized as follows. Section 2 gives an introduction to the mouse embryo ontology developed for the Mouse Atlas. Section 3 describes the formalization of various aspects of this ontology. A prototype implementation of the formalization is outlined in Section 4. Brief comments on the use of *is–a* relationships and description logics are given in Sections 5 and 6, respectively. The paper is concluded in Section 7, summarizing the research carried out thus far and describing future work.

*To whom correspondence should be addressed.

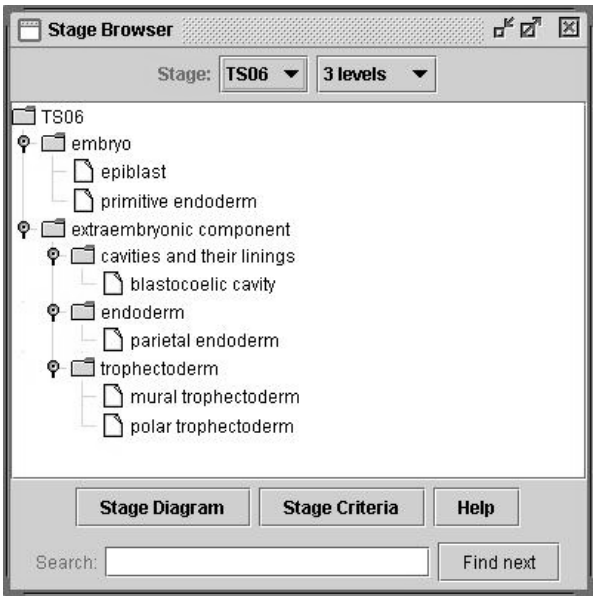

Fig. 1. Anatomy Stage Browser: showing top three levels of the mouse anatomy at Theiler stage 6.

2 MOUSE EMBRYO ANATOMY OVERVIEW

2.1 Basic anatomy trees

Following the description of mouse embryo development by Theiler (1989), the anatomy ontology is organized into 26 developmental stages, referred to as *Theiler stages* (TS1–TS26). Each stage is primarily organized as a *structural part–of* tree. Figure 1 shows part of the top three levels of the tree at TS6. (The browser shown in the figure is available on-line at the *Mouse Atlas** web site.)

The tissues represented by subnodes of a node in the tree are intended to be *non-overlapping (exclusive)* and *complete*, i.e. they describe *all distinct* parts of the parent tissue. (In this paper, the term 'tissue' is used in a very generic way, meaning both: whole anatomical structures as well as specific tissues.) For example, in Figure 1, the `trophectoderm` consists of the `mural trophectoderm` and the `polar trophectoderm`, which are *distinct* from each other and are the *only* parts of the `trophectoderm` at that stage. The requirement for the anatomy ontology to be non-overlapping and complete was a design decision by the biologists who developed it, rather than a constraint needed for its formalization. The formalization does, however, take advantage of this constraint for some of its reasoning (an example is given in Section 3.2.3).

2.2 Groups of tissues

Although most biologists will probably find the given *part–of* hierarchy intuitive and easy to understand, it does not present the only possible way of structuring a mouse embryo anatomy ontology. For example, a scientist with a special interest in the nervous system, or one with a special interest in the development of the skeleton, is likely to look for groupings of tissues that are not directly supported by the given primary tree structure.

For example, at TS19 the node `embryo` has subnodes `skeleton`—representing the entire embryo's skeleton apart from the part in the tail—and `tail` (amongst others). The `tail` node has a subnode called `skeleton`, which denotes the part of the skeleton found in the tail. There is no single node in the ontology that refers to the entire skeleton of the mouse at TS19. Hence, one may want to introduce a new group node that links to the existing two skeleton nodes and would therefore represent the entire skeleton of the mouse embryo at TS19.

To enable such alternative views of the anatomy, so-called *groups* have been introduced into the ontology. A group is essentially a new term that is being added. As with the 'primary tissues' of the trees, a group is represented as a node that has subnodes, which identify the tissues that are *part of* this new group. The sub-nodes may either be primary tissues or other groups. Although 'sibling' primary and group nodes are not necessarily exclusive—they may share common subnodes—the *completeness* assumption still holds. When a new group is added, in addition to identifying the 'sub-parts' of this new group, it is also necessary to determine what this group is part of. Note that even after the addition of group nodes, the only type of relationship between tissues thus far is that of *structural part–of*.

With the introduction of groups, the graph representing the ontology changes from a set of *rooted directed trees* to the more general form of a set of *rooted directed acyclic graphs*, rooted DAGs.

2.3 Abstract mouse

Originally invented as a schema design for the object-oriented database system used to store the anatomy, the idea of an *abstract mouse* has subsequently also proved helpful at the conceptual level. The abstract mouse is effectively a time-independent 'summary graph' of all anatomy tissues found during the development of the mouse and can be used to refer to an anatomical concept, such as `heart`, without the need to specify a particular developmental stage.

The abstract mouse is algorithmically derived from the existing stage-dependent anatomy. Essentially, the set of nodes in the DAG for the abstract mouse is the union of the sets of nodes of the 26 stage-specific DAGs. There is a link between two nodes in the abstract mouse DAG, if there is a corresponding link in any of the 26 stage DAGs.

While this works well at the database implementation level, it introduces a problem at the conceptual level: the same tissue concept may be represented more than once in the abstract mouse. For example, at TS12 we have `/embryo/organ system/cardiovascular system/heart/primitive heart tube/outflow tract`, whereas at TS13,

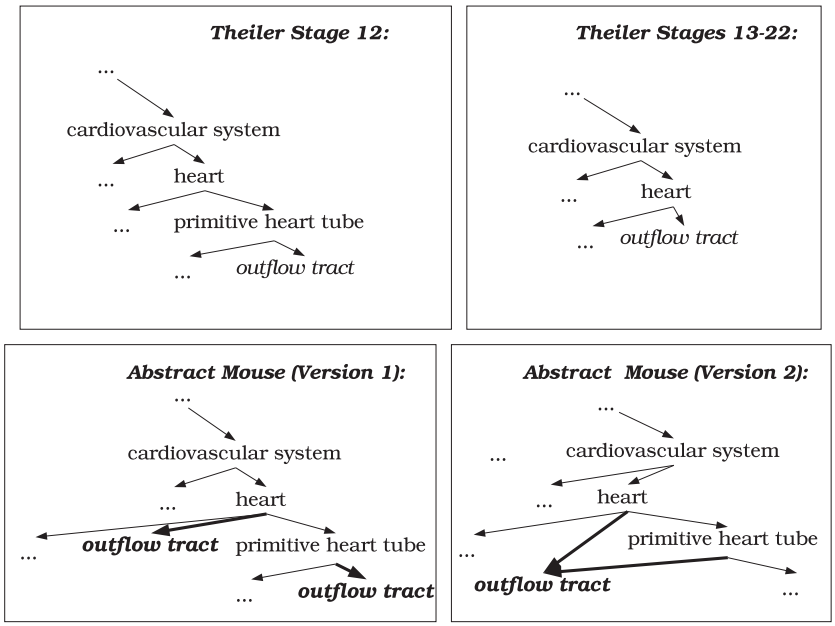

Fig. 2. Abstract Mouse: multiple representation of 'outflow tract' collapsed into a single node.

we have `/embryo/organ system/cardiovascular system/heart/outflow tract`. The abstract mouse graph contains `outflow tract` twice (Figure 2: Abstract Mouse, Version 1), each with its own *access id* (accession number).

To be able to avoid this problem, distinct nodes in the abstract mouse, which conceptually refer to the same tissue must be marked in some way. From a modelling perspective, this can be achieved through the introduction of a new relationship between tissue nodes in the abstract mouse; we shall call this new relationship *sameAs*. We note that this *sameAs* equivalence really only holds at the time-independent context of the abstract mouse, since the `outflow tract` of the `primitive heart tube` at stage TS12 is physically not identical to the `outflow tract` of the `heart` at stage TS13.

Adding *sameAs* relationships to the abstract mouse cannot be automated, i.e. algorithmically derived from the stage-based anatomy, but requires additional input from developmental biologists. Work is currently under way to extend the existing Mouse Atlas anatomy (EMAP) accordingly.

When presenting the abstract mouse to the user, it may be best to 'collapse' all nodes of a *sameAs* family into a single node, as illustrated in Figure 2 (Abstract Mouse, Version 2), although semantically this is no different from keeping the original abstract mouse graph amended with additional *sameAs* links.

2.4 Lineage data

To track the development of tissue over time, so-called *lineage data* is included in the anatomy ontology. Lineage is modelled as a *derivedFrom* link between tissue nodes in the stage-based anatomy, either within a single stage or across two stages. Lineage is a many-to-many relationship, i.e. a single tissue may be derived from a number of different tissues at an earlier time, and may become part of a number of different tissues at a later time.

The notion of tissue as a concept becomes difficult when adding lineage. At what point do two tissues linked through lineage become two different tissues conceptually? For example, most biologists would probably agree that the `heart` at TS12 and the `heart` at TS13 are referring to the same concept of *heart*. However, is the `future brain` at stage TS16 conceptually the same as the `brain` at stage TS17? These discussions are beyond the scope of this paper.

3 FORMAL REPRESENTATION

3.1 Anatomy graphs

As we have seen earlier, the anatomy is largely described in hierarchical terms. To simplify the discussion of our formalization, we will use the following terminology. A *full name* of an anatomical tissue is given as an n-tuple: $(t_0, t_1, \ldots, t_n)$. The *path name* of the tissue is $(t_0, t_1, \ldots, t_{n-1})$. The *component name* is t_n. For example, given the tissue name (using a file directory style notation):

/embryo/branchial arch/3rd arch/branchial pouch/endoderm/dorsal

its full name is:

(*embryo, branchial arch, 3rd arch, branchial pouch, endoderm, dorsal*),

its path name is:

(*embryo, branchial arch, 3rd arch, branchial pouch, endoderm*),

and its component name is:

dorsal.

Each tissue also has a unique *identifier*, a Mouse Atlas accession number. We use predicate $tissue(X, FN)$ to state that the tissue with identifier X has FN as its full name. Predicate $hasPart(X, Y)$ represents the fact that tissue Y is part of tissue X; X and Y are unique tissue identifiers.

Let predicates $pName(FN, PN)$ and $cName(FN, CN)$ represent the fact that PN and CN are the *path name* and *component name* of the *full name* FN, respectively. The following constraints must hold for all primary anatomy trees, i.e. without the addition of groups:

(1) A full name uniquely identifies a tissue, i.e. there are no two tissues with the same full name.

$$tissue(X, FNx) \wedge tissue(Y, FNy) \wedge FNx = FNy \rightarrow X = Y.$$

(2) The full name of a node is the path name of all its immediate sub-part nodes:

$$hasPart(X, Y) \wedge tissue(X, FNx) \wedge tissue(Y, FNy) \wedge pName(FNy, PNy) \rightarrow FNx = PNy.$$

We say tissue X is a *super-part* of tissue Z if there exists a $hasPart$ path from X to Z, e.g. $hasPart(X, Y)$ and $hasPart(Y, Z)$. X is a *sub-part* of tissue Y, if Y is a super-part of X. Formally, we define $superPart(X, Z)$ and $subPart(X, Z)$ recursively as follows:

$$hasPart(X, Z) \vee (hasPart(X, Y) \wedge superPart(Y, Z)) \rightarrow superPart(X, Z)$$

$$superPart(Z, X) \rightarrow subPart(X, Z).$$

As previously discussed, there is a need to complement the primary anatomy hierarchies with *groups*. In general, primary and group nodes can be treated equally. Hence, the same predicates *tissue* and *hasPart* are used to represent groups in the ontology. However, for some of the reasoning we need to be able to distinguish them.

The predicate $primary(X)$ is true, if X is a primary node. Predicate $group(X)$ is true, if X is a group node. All nodes are either primary or group, but not both:

$$primary(X) \wedge group(X) \rightarrow \perp$$

Unless explicitly stated otherwise [using predicates $primary(X)$ and $group(X)$], rules concerning anatomical tissues and their properties, including the propagation of properties as described in Section 3.2.3, apply to all tissues, primary as well as groups.

There are a number of constraints that groups must adhere to; too many to list them all, so we will only give one example, the definition of a *minimal group*.

Assume tissue a has parts b and c, and someone wishes to create a new group tissue g that consists of b, c and d (d is not part of a). An obvious way to achieve this would be to add $hasPart(g, b)$, $hasPart(g, c)$and $hasPart(g, d)$. However, we would like to keep the graph *minimal*, i.e. place *hasPart* links at the highest appropriate level. In our example, instead of adding $hasPart(g, b)$ and $hasPart(g, c)$ we should add $hasPart(g, a)$. We generalize this idea into the *minimal group* constraint.

Before giving a formal definition for this constraint, the concept of *shared parts* is introduced. Predicate $sharedParts(X, Y)$ states that X and Y have at least one common part:

$$hasPart(X, Z) \wedge hasPart(Y, Z) \rightarrow sharedParts(X, Y)$$

DEFINITION 1. *Group G is* minimal, *if for every tissue T it shares some part with, at least one of the parts of T is not also a part of G:*

$$\forall T \cdot sharedParts(T, G) \cdot \exists X \cdot hasPart(T, X) \wedge \neg hasPart(G, X) \rightarrow minGroup(G)$$

Being able to formulate constraints such as the one above, is one of the benefits of formalizing an ontology. We have also formalized aspects of the abstract mouse and lineage data, but omit the details here to keep the paper consise.

3.2 Properties of tissues

3.2.1 Tissues and their properties As discussed above, the anatomy serves as a framework for other biological data, i.e. it allows us to index that data using anatomical tissues as keys. To capture this association between tissues and other data, we introduce the concept of *properties*. We use:

$pos_e(P, T)$ to state that there is experimental evidence that tissue T has property P, and

$neg_e(P, T)$ to state that there is experimental evidence that tissue T does *not* have property P.

For example, let $genex(X)$ denote the property that gene X is expressed, we can then write:

$pos_e(genex(X), T)$ to state that gene X has been found to be expressed in tissue T, and

$neg_e(genex(Y), T$ to state that gene Y has been found not to be expressed in tissue T.

More generally, we write

$pos(P, T)$ to state that it is known that tissue T has property P, and

$neg(P, T)$ to state that it is known that tissue T does not have property P.

A tissue is known to have, or not to have, a certain property either because of direct experimental evidence:

$$pos_e(P, T) \rightarrow pos(P, T), \quad \text{and}$$
$$neg_e(P, T) \rightarrow neg(P, T),$$

or because of the propagation of properties (see below).

3.2.2 Negation of knowledge about properties The negation of the *pos* predicate does not imply that the corresponding *neg* predicate is true. For example, $\neg pos(genex(X), T)$ simply states that whether gene X is expressed in tissue T is not known. It does not mean that we know that gene X is not expressed in tissue T. Similarly, the negation of the *neg* predicate does not imply that the corresponding *pos* predicate is true. In general, the following holds for any property P:

$$\neg pos(P, T) \not\rightarrow neg(P, T), \quad \text{and}$$
$$\neg neg(P, T) \not\rightarrow pos(P, T).$$

However, if we know that a particular property P is true for tissue T, we also know that it cannot at the same time be the case that the property is not true, and vice versa. Therefore,

$$pos(P, T) \rightarrow \neg neg(P, T), \quad \text{and}$$
$$neg(P, T) \rightarrow \neg pos(P, T).$$

3.2.3 Propagation of properties If a tissue has a certain property, for example that it expresses gene X, then we know that the same property holds for its parent tissues. However, the fact that a property holds for some tissue does not imply that it holds for all its sub-part tissues. Formally, we say:

$$pos(P, Y) \wedge hasPart(X, Y) \rightarrow pos(P, X), \quad \text{and}$$
$$pos(P, X) \wedge hasPart(X, Y) \not\rightarrow pos(P, Y).$$

If we know that a certain tissue does not have a particular property, then it must be the case that none of its sub-part tissues have this property. However, it does not imply that its parent tissue will also not have this property:

$$neg(P, X) \wedge hasPart(X, Y) \rightarrow neg(P, Y), \quad \text{and}$$
$$neg(P, Y) \wedge hasPart(X, Y) \not\rightarrow neg(P, X).$$

In summary, 'positive knowledge' about a tissue's properties propagates up the anatomy hierarchy, whereas 'negative knowledge' about a tissue's properties propagates down the anatomy hierarchy.

However, because of the *completeness* condition of sub-nodes, if it is known that all immediate sub-part tissues of tissue X do not have a certain property P, then neither can X. In case X is a primary tissue node, it is sufficient if all the immediate primary sub-part tissues do not have property P for this also to be true for X.

To simplify the formalization of this propagation rule, we introduce predicates $notallneg(P, X)$ and $notallprimaryneg(P, X)$. The former is true if at least for one of the immediate sub-part nodes of X $neg(P, Y)$ does not hold. The latter is true if X is a primary node and for at least one of its immediate primary sub-part nodes $neg(P, Y)$ does not hold:

$$hasPart(X, Y) \wedge \neg neg(P, Y)$$
$$\rightarrow notallneg(P, X)$$
$$hasPart(X, Y) \wedge primary(Y) \wedge \neg neg(P, Y)$$
$$\rightarrow notallprimaryneg(P, X)$$

We can now add the following two propagation rules for negative knowledge:

$$hasPart(X, Y) \wedge neg(P, Y) \wedge \neg notallneg(P, X)$$
$$\rightarrow neg(P, X)$$
$$hasPart(X, Y) \wedge primary(X) \wedge primary(Y) \wedge$$
$$neg(P, Y) \wedge \neg notallprimaryneg(P, X)$$
$$\rightarrow neg(P, X)$$

3.2.4 Maybe properties Let us assume that gene G is expressed in tissue X $[pos(genex(G), X)]$ and that Y is a sub-part tissue of X $[subPart(Y, X)]$, then we do not know for certain whether or not gene G is expressed in tissue Y (no downward propagation of positive knowledge).

However, we may want to interpret the fact that gene G is expressed in one of Y's super-part tissues as some indication that it may also be expressed in Y itself. Of course, in reality this may not be the case, but in the absence of any other firm knowledge, this assumption can still be useful. A similar argument holds for negative knowledge and upwards propagation of properties.

We therefore introduce the notion of *maybe* properties. Predicates *maybe_pos* and *maybe_neg* are defined as follows:

$$hasPart(X, Y) \wedge pos_e(P, X) \wedge \neg neg(P, Y)$$
$$\rightarrow maybe\_pos(P, Y), \quad \text{and}$$
$$hasPart(X, Y) \wedge neg_e(P, Y) \wedge \neg pos(P, X)$$
$$\rightarrow maybe\_neg(P, X).$$

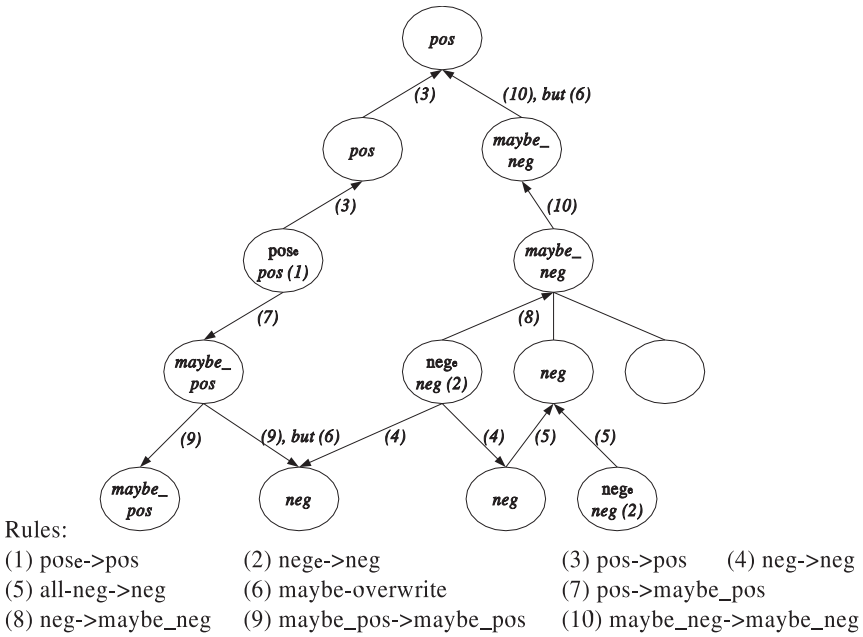

Fig. 3. Propagation of properties: the example illustrates how knowledge is propagated up and down the *part–of* hierarchy. Where appropriate, arrows indicate direction of propagation and labels show which rules have been applied.

In the absence of any contrary information, *maybe* properties can also propagate along the hierarchies:

$$hasPart(X,Y) \wedge maybe\_pos(P,X) \wedge \neg neg(P,Y) \rightarrow maybe\_pos(P,Y), \quad \text{and}$$

$$hasPart(X,Y) \wedge maybe\_neg(P,Y) \wedge \neg pos(P,X) \rightarrow maybe\_neg(P,X).$$

Figure 3 shows a summary of the property propagation rules that can be applied. Only the three nodes marked either pos_e or neg_e have been annotated with properties obtained from actual experiments. All but one of the remaining nodes were associated with properties following the propagation rules described above.

3.3 Queries

Given the representation of anatomy tissues and their properties, we can now formulate the answers to a number of basic queries. We use gene-expression as an example property. The queries and the logical expressions to answer them are given in Table 1. The first four queries are equivalent to typical questions biologists put to the Mouse Atlas using the EMAGE interface (accessible on-line). The *may_be* queries reflect an extension of the current system that is implemented in the Mouse Atlas database server, and will soon be available on-line through EMAGE.

More complicated queries can of course also be constructed. For example, to answer the query '*Find all genes which are co-expressed in tissue* t_1*, but which are not co-expressed in tissue* t_2.', we could write:

$$pos(genex(G1),t1) \wedge pos(genex(G2),t1) \wedge \neg(pos(genex(G1),t2) \wedge pos(genex(G2),t2))$$

Please note that the second (negated) part of this expression simply states that there is currently no evidence that $G1$ and

Table 1. Queries

Input	Query statement in logic
Which genes are definitely expressed in tissue T?	
T	$pos(genex(G),T)$
Which genes are definitely not expressed in tissue T?	
T	$neg(genex(G),T)$
In which tissues is gene G definitely expressed?	
G	$pos(genex(G),T)$
In which tissues is gene g definitely not expressed?	
G	$neg(genex(G),T)$
Which genes are maybe (but not definitely) expressed in tissue T?	
T	$maybe\_pos(genex(G),T) \wedge \neg pos(genex(G),T)$
Which genes are maybe not expressed in tissue T?	
T	$maybe\_neg(genex(G),T) \wedge \neg neg(genex(G),T)$
In which tissues is gene G maybe (but not definitely) expressed?	
G	$maybe\_pos(genex(G),T) \wedge \neg pos(genex(G),T)$
In which tissues is gene G maybe not expressed?	
G	$maybe\_neg(genex(G),T) \wedge \neg neg(genex(G),T)$

The table shows a number of basic queries typically used in EMAP/EMAGE, and the logic expressions used to answer these queries.

$G2$ are co-expressed in tissue $t2$. A more strict interpretation of the query might be that there is evidence that the co-expression for $t2$ does not hold, in which case we should rewrite the solution as:

$$pos(genex(G1), t1) \wedge pos(genex(G2), t1) \wedge (neg(genex(G1), t2) \vee neg(genex(G2), t2))$$

It is exactly this semantic precision when dealing with such queries that is a key benefit of formalizing the mouse anatomy and associated data.

4 PROTOTYPE

4.1 Prolog implementation

A prototype implementation for the logic discussed in this paper has been developed using Prolog (Bratko, 2000). The underlying knowledge base uses the following Prolog predicates:

`tissue(TID,[N1,N2,N3,...]).`
to capture tissue names and IDs; TID is the ID for the tissue with the name represented by the list $[N1, N2, N3, \ldots]$;

`hasPart(TID1,TID2).`
to capture that the tissue with ID $TID2$ is a *part of* the tissue with ID $TID1$;

`pos_e(genex(G),TID).`
as discussed in Section 3.2;

`neg_e(genex(G),TID).`
as discussed in Section 3.2;

For example, let tissue */embryo/branchial arch/3rd arch/branchial pouch/endoderm* have a part called *dorsal* (tissue IDs 123 and 124, respectively) and let there be experimental evidence that gene *msx1* is expressed in the *dorsal*, our knowledge base would contain the following:

```
tissue(123,['embryo','branchial arch',
            '3rd arch','branchial pouch',
            'endoderm']).
tissue(124,['embryo','branchial arch',
            '3rd arch','branchial pouch',
            'endoderm','dorsal']).
hasPart(123,124).
pos_e(genex(msx1),124).
```

The rules presented in previous sections can also easily be translated into the corresponding Prolog code. For example, the propagation rule for positive knowledge can be written as:

```
pos(P,T) :- pos_e(P,T).
pos(P,X) :- hasPart(X,Y), pos(P,Y).
```

Similarly, the propagation of positive *maybe* properties can be written as:

```
maybe_pos(P,Y) :- hasPart(X,Y),
                  pos_e(P,X),
                  \+ neg(P,Y).
maybe_pos(P,Y) :- hasPart(X,Y),
                  maybe_pos(P,X),
                  \+ neg(P,Y).
```

The propagation of negative knowledge has similar implementations.

The question: '*Which genes are expressed in the /embryo/branchial arch/3rd arch/branchial pouch/endoderm*?' can be expressed in the following Prolog query:

```
tissue(ID,['embryo','branchial arch',
 '3rd arch','branchial pouch',
 'endoderm']),
pos(genex(G),ID).
```

Assuming the above knowledge base, the Prolog interpreter would find 123 as the ID for the given tissue and then find $msx1$ as the gene expressed in that tissue (using positive knowledge propagation).

4.2 Evaluation

The purpose of the logic presented is to give a precise semantic description of the Mouse Atlas anatomy ontology. The Prolog prototype provides an independent—from the actual Mouse Atlas system—implementation of this logic. Hence, the Prolog implementation is used to verify that the logic described provides an accurate description of the relevant parts of the actual system. Therefore, the evaluation of the logic was carried out by answering a variety of queries using the actual Mouse Atlas system as well as the Prolog prototype and checking that the results obtained were consistent. The first set of queries only dealt with the anatomy itself, e.g. finding tissues that match certain strings, and finding super-parts and sub-parts of tissues. Additional queries included gene-expression data examples and propagation of properties. For the Prolog queries, the complete anatomy ontology was exported from the actual system in the form of the above predicates. A small gene-expression knowledge base was created manually based on actual data in the Mouse Atlas. Although we have not carried out an exhaustive test suite, covering all data in the Mouse Atlas, based on the numerous examples that were successfully tested, we are confident that the logic presented in this paper is an accurate description of the Mouse Atlas anatomy ontology.

4.3 Availability

A Prolog knowledge base, using the predicates described in this section, representing the complete Mouse Atlas anatomy ontology is available on-line from the Mouse Atlas web site (Table 2 for URL).

Table 2. Cited websites

Name	URL
DAML + OIL	www.daml.org
Mouse Atlas	genex.hgu.mrc.ac.uk
GALEN	www.opengalen.org
OWL	www.w3.org
GO	www.geneontology.org
OBO	obo.sourceforge.net
Visible human	www.nlm.nih.gov/research/visible/visible_human.html

5 ON THE USE OF *IS–A* RELATIONSHIPS

Many ontologies make use of so-called *is–a* relationships, which indicate that a concept is of a certain type or a kind of other concept. For example, one could say that a `cardiac muscle` is a kind of `muscle`. The current version of the Mouse Atlas ontology does not use *is–a* relationships, but only uses *part–of*—using predicate `hasPart()` in our formalization. One might argue that some of these *part–of* relationships might more accurately be modelled as *is–a* types. In order to decide when to use *part–of* or *is–a*, it may be helpful to look at the irrespective property propagation rules.

Unlike for *hasPart* relationships, in the case of *is–a*, positive as well as negative knowledge is inherited by its sub-nodes, i.e. downward propagated. For example, if `somite 6` *is–a* `somite`, then if `somite` expresses gene *G*, then so do all its sub-nodes, i.e. gene *G* is also expressed in `somite 6`. If, however, `somite 6` is *part–of* a bigger `somite` structure, and we know that gene *G* is expressed in `somite`, this does not imply that *G* is also expressed in `somite 6`.

Hence, the decision of whether to use *is–a* or *part–of* relationships, and the associated propagation of knowledge about tissues, is primarily a matter of requirements, rather than one of computational technology. The model should be able to reflect the actual knowledge obtained from biological experimentation. The mapping of gene expression data on to the Mouse Atlas is consistent with the use of *part–of* relationships in its anatomy ontology.

6 ON THE USE OF DESCRIPTION LOGICS

There is currently a lot of interest in the use of *Description Logic*(DL) based languages for ontologies, specifically, *DAML + OIL**, and its successor *OWL** (Web Ontology Language). DLs are generally a subset of languages such as Prolog, but computationally more efficient. We are currently looking at the use of OWL for the description of the mouse anatomy ontology. The interest here lies in the representation of the anatomy itself as well as how best to achieve the type of reasoning described in this paper. A detailed discussion of OWL for the mouse anatomy ontology, however, is beyond the scope of this document and will be described in another paper.

7 CONCLUSION AND FUTURE WORK

The EMAP and EMAGE project has developed a mouse embryo anatomy ontology for use in a digital atlas of mouse development. This anatomy ontology is used in a spatio-temporal framework for spatially mapped data such as *in-situ* gene-expression.

In this paper, a formal description of the ontology is given using predicate-based logic. The primary purpose of this formalization is to make explicit the semantics of the current anatomy and its use for gene-expression data. However, it also serves as the basis for a review of the ontology and its representation in a description logic-based language, such as DAML + OIL or OWL.

The formalization presented captures the structural *part–of* hierarchy of the anatomy and cell-lineage information. It explicitly distinguishes between not knowing whether a tissue has a certain property, e.g. gene-expression, and knowing that a tissue definitely does not have a particular property. It also formalizes the propagation of properties along the *part–of* hierarchy. Finally, it supports so-called *maybe properties*, for which there is no definite but some circumstantial evidence.

The logic has been implemented in a Prolog prototype system. The prototype allows the testing of the logic and, since independently implemented from the actual Mouse Atlas system, the verification of some of the publicly accessible EMAP and EMAGE applications.

A Prolog version of the complete EMAP anatomy ontology is available on-line at the Mouse Atlas web site. Plans are under way also to export the EMAGE data set into a Prolog knowledge base to facilitate additional experiments with the prototype. Also, work is progressing on the use of OWL for describing mouse anatomy. Finally, future work is aimed at extending the logical representation of knowledge about mouse embryo anatomy and gene-expression data in support of intelligent workflow systems.

ACKNOWLEDGEMENTS

The development of the EMAP atlas and EMAGE database has been undertaken as a core-funded MRC programme at the MRC Human Genetics Unit in collaboration with Matthew Kaufman and Jonathan Bard at the University of Edinburgh. We would like to thank our MRC colleagues from the Mouse Atlas project as well as members of the (BBSRC funded) XSPAN project team from Heriot-Watt University and the University of Edinburgh for their input and comments on earlier drafts of this paper.

REFERENCES

Bratko,I. (2000) *Prolog Programming for Artificial Intelligence*. Pearson Higher Education Longman.

Brune,R., Bard,J., Dubreuil,C., Guest,E., Hill,W., Kaufman,M., Stark,M., Davidson,D. and Baldock,R. (1999) A

three-dimensional model of the mouse at embryonic day 9. *Dev. Biol.*, **216**, 457–468.

Davidson,D. and Baldock,R. (2001) Bioinformatics beyond sequence: mapping gene function in the embryo. *Nat. Rev. Genet.*, **2**, 409–418.

Davidson,D., Bard,J., Brune,R., Burger,A., Dubreuil,C., Hill,W., Kaufman,M., Quinn,J., Stark,M. and Baldock,R. (1997) The mouse atlas and graphical gene-expression database. *Sem. Cell Dev. Biol.*, **8**, 509–517.

Ringwald,M., Baldock,R., Bard,J., Kaufman,M., Eppig,J., Richardson,J., Nadeau,J. and Davidson, D. (1994) A database for mouse development. *Science*, **265**, 2033–2034.

Theiler,K. (1989) *The House Mouse: Atlas of Embryonic Development*. Springer-Verlag.