University of Edinburgh
Division of Informatics

Surfing The Mouse Atlas

4th Year Project Report
Artificial Intelligence and Computer Science

Carol Gail Sinclair

May 28, 2002

Abstract: When you're looking for the answer
how to classify all cancer,
proteins, microbes, fish and succulent legumes,
You must know a little Latin
tell a round fish from a flat one
and have memory with lots and lots of room

But, before we start alinking
we should sit back and be thinking
on our methods, clientele and on our goal,
Lest we make a mammoth bank
rarely used and rarely thanked
just consuming funds and efforts: A BLACK HOLE.

Leslie Sobin

Acknowledgements

Many, many thanks to Bonnie Webber and Duncan Davidson without whose endless help this project would have been a non-starter, and whose enthusiasm for the subject at times surpassed my own. Thanks also to those others in the Mouse Atlas team including Jonathan Bard and Richard Baldock.

I would also like to thank my fellow South Bridge inhabitees, Anna and Dave for putting up with my inane questions and considerably slowing the progression of insanity.

Contents

1. Introduction

1.1 Bioinformatics

Bioinformatics is the area comprising the analysis of biological sequence information, recovery of evolutionary patterns, prediction of gene function, biological data mining and "silicon based biology". The amount of information emerging through bioinformatics is so great the we need sophisticated tools to deal with it.

1.1.1 Anatomical Databases

The primary goal of computational molecular biology, like molecular biology itself, is to understand the meaning of the genomic information and how this information is expressed. Genetic bioinformatics is concerned with the problems of predicting the biological function of genes and gene products from their primary sequence and structure (sometimes known as functional genomics).

Many resources exist on the web that help to develop this information, including anatomical databases. These databases generally include searchable information on the embryonic development of a particular species (including human, fruit fly, zebra fish and rat) indicating where in the embryo certain genes are expressed and at what stage in development. Search within these databases can vary from the sublime (simple, see Figure 1.2) to the ridiculous (very complicated, see Figure 1.1).

1.1.2 Terminology

Terminology plays an important part in providing user satisfaction with these search facilities and heterogeneous terms (i.e words with more than one meaning) pose significant problems [8, 9, 16, 17, 18, 19, 20, 24]. This type of *semantic confusion* [19] is a major problem in bioinformatics, due to the increasing size and complexity of biological information on the web and the overwhelming amount of existing and newly created concepts and terminology. Scientific databases include an enormous amount of information, which can only be put to proper use if there is transparency and agreement in the naming convention of the concepts. Researchers in medicine and biology would therefore benefit from tools to facilitate the discovery and identification of important and relevant concepts. If data is to be integrated, exchanged, and searched efficiently any term that a biologist may use needs to be recognised by the resources developed.

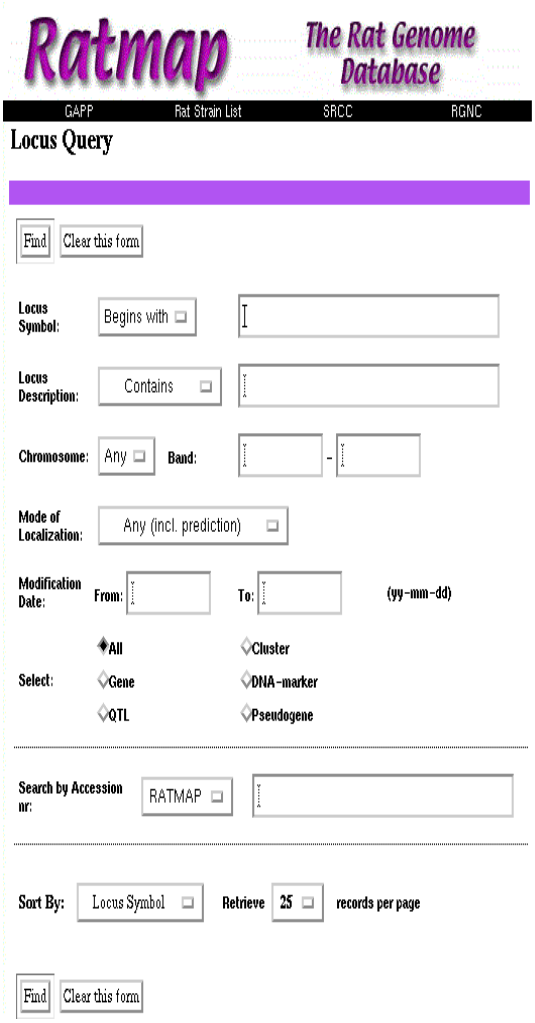

Figure 1.1: Example Query Form from The Rat Genome Database

1.2 Importance of HCI

Any community that can be expected to use a particular database can also be expected to use several other databases and interfaces dealing with a similar type of information. Because of this, the user cannot be expected to remember the most efficient way of accessing the information of interest for each interface. For example, Figure 1.1 demonstrates the amount of domain-specific prior user knowledge required for interaction with the RatMap database[1] while Figure 1.2 demonstrates that only one item of prior knowledge of anatomy is required for search to begin.

In Human Computer Interaction (HCI), one of the most important considerations is that of *usability* [7]. Usability involves three main principles, *learnability*,i.e. the ease

[1]http://ratmap.gen.gu.se/

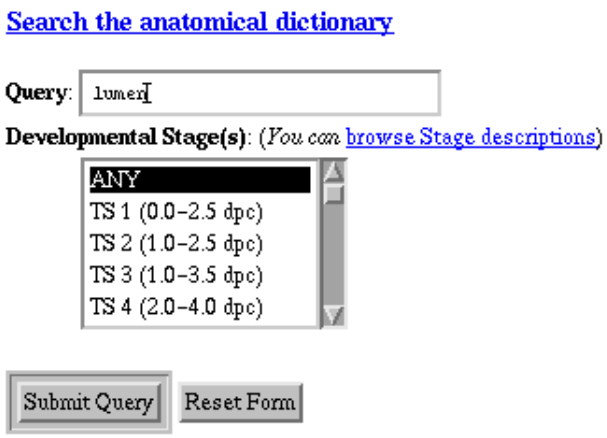

Figure 1.2: GXD Search Interface

with which new users can achieve efficient interaction, *flexibility*, i.e. the multiplicity of ways in which a user and the system can interact, and *robustness*, i.e. the level of support provided to the user in evaluating system performance and identification and correction of errors. Involved in all three of these principles is *cognitive effort*.

Cognitive effort is the amount of working memory required on the part of the user to be able to use an interface as efficiently as possible. The more that cognitive effort is reduced in an interface, the more likely it is that the user will be satisfied with his or her interaction with that interface. If the effort involved in the usage of the system is kept at a minimum, the user will tend to return to that resource time and again to find information rather than giving up and turning elsewhere. The techniques developed in this project reduce just that cognitive effort making the database in question more user friendly and easier to navigate textually.

1.3 Natural Language Processing (NLP)

Resources are time-consuming and often expensive to develop, and Language Technology (LT) rarely has the luxury of calling upon resources specially designed for the task at hand. For LT applications in developmental anatomy such as robust interfaces to anatomically-indexed gene expression data and effective text mining tools to assist in building such databases, resources already exist in the form of *anatomical nomenclatures* for several model organisms including mouse, zebra fish, drosophila and human, with others to follow. These nomenclatures have been developed by biologists for biologists, to record in a clear, intuitive and structured way the structures that can be distinguished at each stage of an embryo's development. The challenge for LT applications is to stretch them to serve other purposes as well.

1.3.1 Information Retrieval/Extraction and Data/Text Mining

Information Extraction (IE) and Information Retrieval (IR) are complementary NLP methods. They are closely related and the distinction between the two can often seem vague, but a general definition could be that IR retrieves relevant documents from a much larger set, while IE extracts relevant information from within a document or data set [9].

Two areas within IE are that of Data Mining and Text Mining. Both these techniques develop methods and tools for analysing large data sets and for searching for unexpected relationships in the data, and involve the development of combinatorial pattern matching with statistical techniques and database methods. Data Mining (also known as Knowledge Discovery in Databases (KDD)) assumes that the data to be mined is already in the form of a structured database, whereas Text Mining discovers useful knowledge from unstructured, free text [10, 14].

1.4 Dissertation Outline

This dissertation describes how one of these anatomical nomenclatures, *The Mouse Atlas Nomenclature*, with the use of the previously described NL methods, can help in the extraction of a new resource that can support a more accessible interface to anatomically-indexed data. The techniques used are not specific to the Mouse Anatomical Nomenclature, and can be applied to anatomical nomenclatures for other model organisms as well.

The remainder of this dissertation goes on to discuss certain existing problems of The Mouse Atlas Nomenclature in Chapter 2. In Chapter 3, the enhancements made to this Nomenclature via data mining are described as is the text mining of synonyms in Chapter 4 along with other work in similar areas. Chapter 5 summarises the evaluation of my project and Chapter 6 describes future work already arranged and suggests further related work.

2. Mouse Atlas Background

2.1 MRC HGU's Mouse Atlas

The *Mouse Atlas*[1], developed by researchers at the Medical Research Council's Human Genetics Unit (MRC HGU) in Edinburgh, is a 3D atlas of mouse embryo development. Anatomical structures within each of the 26 Theiler Stages[2] [22] of embryo development are labelled, and 3D reconstructions of each stage can be displayed in transverse, frontal, sagittal or arbitrary planes.

The *Mouse Atlas* is now being used to support indexing of gene expression data, allowing the results of gene expression experiments to be indexed with respect to where in the developing embryo gene expression is occurring. A database of *spatially indexed* gene expression data (the EMAGE database) is being developed at the MRC HGU. (In spatial indexing, data is associated directly with volume elements, *voxels*.) A database of *symbolically indexed* gene expression data (the *Gene Expression Database* or GXD) is being developed by the Jackson Laboratory[3] in Bar Harbor, Maine. This exploits a second resource within the *Mouse Atlas* called the *Mouse Anatomical Nomenclature*. This is a set of 26 trees of anatomical terms (one tree per Theiler Stage) structured primarily by part-whole relations (and some set-member relations). (In symbolic indexing, gene expression data is associated with a label specifying a pre-defined region of the embryo.)

2.2 Nomenclature Description

The root node of each Theiler Stage tree corresponds to the entire embryo at that stage, while all other nodes correspond to organ systems, subsystems, spatially-localised parts of subsystems or particular anatomical structures. Each node within a tree has a label (i.e. its *component term*), and while more than one node within a tree (or across trees) may have the same component term, e.g.

> CRANIAL labels both a child of GANGLION (i.e., ganglia located in the head) and a child of NERVE (i.e., nerves located in the head)

every *path* from the root node has a unique denotation, where a path is specified by the sequence of *component terms* along the path. Thus, indexing involves the pairing of a

[1] http://genex.hgu.mrc.ac.uk

[2] The 26 Theiler Stages describe and encompass the 18 day gestation period of the mouse embryo.

[3] http://www.informatics.jax.org

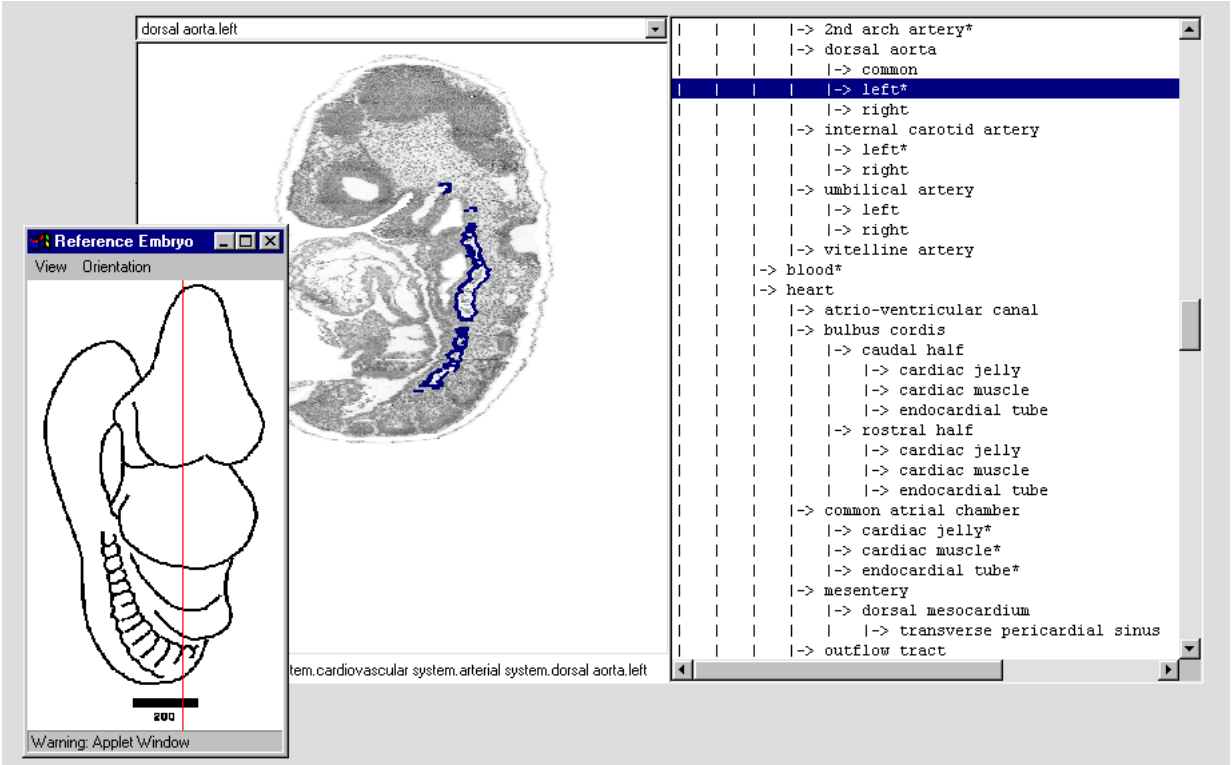

Figure 2.1: Screen shot of Mouse Atlas interface, displaying a Theiler Stage 14 embryo.

path specification such as

> EMBRYO.ORGAN SYSTEM.NERVOUS SYSTEM.CENTRAL NERVOUS SYSTEM.NERVE.CRANIAL.TRIGEMINAL V[4]

(i.e., the trigeminal, or fifth cranial, nerve) with gene expression data and other information such as the type of assay, bibliographic citation, type of probe, etc., and the last date that the entry was modified.

To access gene expression data, both spatial and symbolic access is possible. An elegant spatial interface is being completed at the MRC HGU, that pairs an active window containing a view of the embryo stage of interest, with a window containing the corresponding Nomenclature tree[5]. Clicking at a point in the embryo view highlights the most specific corresponding node of the Nomenclature being displayed (i.e., subtrees of a node can be either hidden or exploded). Similarly, clicking on a term in the Nomenclature highlights the corresponding structure within the embryo along the plane currently being displayed. A screen-shot from the interface is shown in Figure 2.1.

[4]Full stop is used to separate component names along a path.

[5]http://genex.hgu.mrc.ac.uk/Resources/GXDQuery1

On the left of the figure is an outline frontal drawing of the embryo, on which a sagittal section plane is marked in red. The centre panel shows a digital, sagittal section through the volumetric embryo model with the delineated left dorsal aorta coloured blue. The corresponding component term is highlighted on the Mouse Anatomical Nomenclature on the right. Users can access the gene expression data on the highlighted structure by another mouse click.

This project aims at improving symbolic, Natural Language access to gene expression data. Currently, the Mouse Anatomical Nomenclature provides such access in two different ways.[6] (1) Users can do a tree-walk through the Nomenclature for a given stage, to find the anatomical structure whose associated gene expression data is of interest to them, or (2) they can enter a term to be matched against individual *component names* within a stage (or across all stages), with all possible substring matches returned for the user to choose among.

Problems exist with both forms of symbolic access. Navigating through a tree is tedious. An additional problem arises from anatomy being forced into a tree-structure that it doesn't have, leading to different parts of the same structure being realised as distantly related leaves – for example, part of the endocardial tube corresponds to the path EMBRYO.ORGAN SYSTEM.CARDIOVASCULAR SYSTEM.HEART.COMMON ATRIAL CHAMBER.ENDOCARDIAL TUBE (i.e., where that part is located), while another part corresponds to the path EMBRYO.ORGAN SYSTEM.CARDIOVASCULAR SYSTEM.HEART.OUTFLOW TRACT.ENDOCARDIAL TUBE. So tree-walk access, in addition to being tedious, doesn't by itself guarantee the user that s/he has found all sections of the anatomical structure of interest.

2.3 Precision and Recall

Access by sub-string matching on individual component terms in the Mouse Atlas Nomenclature has problems of both *recall and precision* (P and R) [11]. P and R is a common pair of metrics used in Natural Language Processing, and specifically IE and IR. Precision measures the fraction of relevant information retrieved with all information retrieved while recall measures the fraction of retrieved information with the total amount of relevant information available. To illustrate precision and recall, consider an algorithm that was written to retrieve all even numbers from the following input:

1 2 3 4 5 6 7 8 9 0

If the information then retrieved was

1 2 4 6 7 8 9 0

[6] http://www.informatics.jax.org/menus/expression_menu

then this would exhibit 100% (i.e 5/5) recall but only 62.5% (i.e. 5/8) precision as, although all the even numbers were retrieved, so also were irrelevant (i.e. odd) numbers. Conversely, if the output of the algorithm was

2 4 6 0

then this displays 100% (i.e. 4/4) precision but only 80% (i.e. 4/5) recall as all the numbers that were retrieved were even but there were more even numbers that could have been identified. In other words, precision measures the efficiency of the retrieval method, while recall measures the breadth of the search for relevant information.

The aim of any IE or IR technique is to attain as close to both 100% recall and 100% precision as possible.

2.3.1 Existing P and R

Searching the Nomenclature, a user may enter a string that matches nothing (0% recall) for one of two reasons:

- the string is not a component term or a substring within a component term – e.g., while HEART is a common synonym for the modifier "cardiac" (and a component term in its own right) and CARDIAC MUSCLE is a component term, the string "heart muscle" yields no match;
- the string spans multiple component terms (i.e. with respect to the Nomenclature, it involves a sub-path rather than a single node) – e.g., while GLAND is a component term, and PITUITARY is the component term of one of its children (i.e., a member of the set of glands), the string "pituitary gland" does not yield a match.

The opposite may happen as well: 100% recall with low precision. Sub-string matching, especially when a particular stage isn't specified, can yield too many matches for a user to quickly search through, for the one or more anatomical structures of interest. For example, a search on "hindbrain" yields 22 matches, while "mesenchyme" yields as many as 1056 matches.

The goal of this project has been to provide better symbolic, Natural Language access to anatomical structures. In particular, this project has tried to (1) reduce the amount of effort that a user needs to expend in finding anatomical structures of interest; (2) better organise the results of searching for anatomical structures; (3) improve *recall*, to reduce the number of times that no match is found; and (4) improve *precision* over that which is possible using substring matching on individual component names.[7]

[7]When the project began, substring matching was done *within* words, further reducing precision. Thus "ear" retrieved matches on HEART, COCHLEAR DUCT, etc. This has since been fixed.

This also needed to be done in a manner that reduced the amount of effort required of the HGU's Duncan Davidson, who kindly agreed to help with this project on anatomical matters. Thus, the project was carried out using a combination of the Mouse Anatomical Nomenclature, a host of Perl scripts were written to extract particular types of information from the Nomenclature, and Language Technology tools (i.e. a part-of-speech (POS) tagger and noun phrase chunker [5]) applied to textbook information on developmental anatomy, to produce an improved interface to the *Mouse Atlas* and eventually, the *GXD*.

3. Existing Nomenclature Terms

3.1 Nomenclature Representation

As mentioned earlier, the Mouse Atlas Nomenclature consists of 26 Theiler Stage trees of mouse embryo development. For example, Figure 3.1 is a partial tree for Theiler Stage 19.

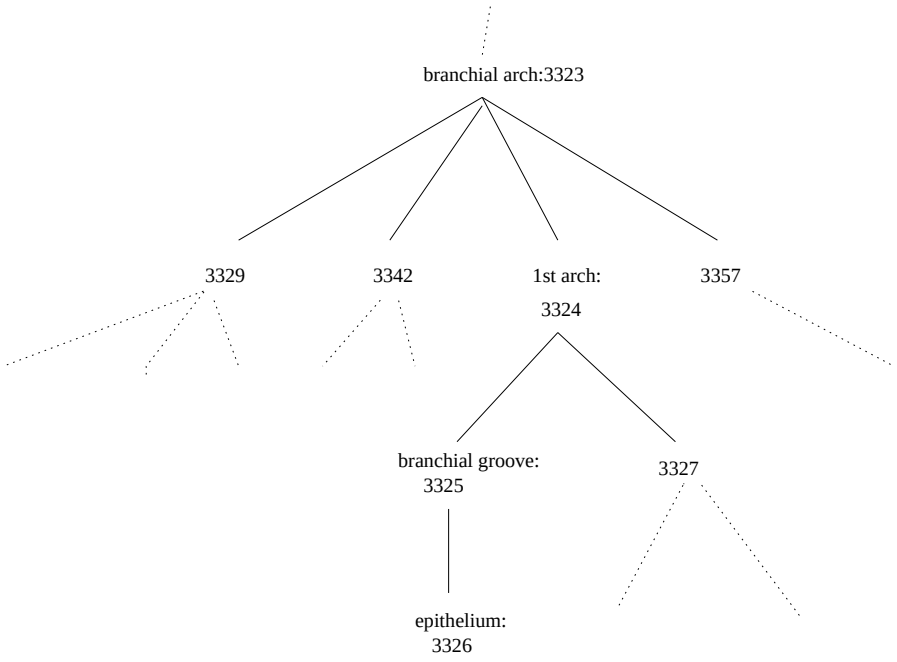

Figure 3.1: Partial tree representation of Theiler Stage 19

The hierarchical properties of these stages is represented in XML and these 26 XML files were analysed and manipulated to extract the properties of anatomical structures via Perl scripts. Figure 3.2 is the XML representation of the above tree.

Each node in the Nomenclature has associated with it a *component name*, an *id* number as well as a *startEmbryoStage* and *stopEmbryoStage*, which denotes which Theiler stage the structure first and last appears respectively, and *printName*, which gives the tree path of that structure in *parent.child* order. Optional XML tags for each node are *abbreviation* and *synonyms* for that structure, any *children* the structure may have in the tree and the single *parent* node that the structure originated from. The *abbreviation* tags are always present with each node, however no information as yet been inserted

```
.....
⟨component name=”branchial arch” id=”3323”⟩
  ⟨printName⟩embryo.branchial arch⟨/printName⟩
  ⟨abbreviation⟩⟨/abbreviation⟩
  ⟨synonyms⟩pharyngeal arch⟨/synonyms⟩
  ⟨childrenId⟩3329⟨/childrenId⟩
  ⟨childrenId⟩3342⟨/childrenId⟩
  ⟨childrenId⟩3324⟨/childrenId⟩
  ⟨childrenId⟩3357⟨/childrenId⟩
  ⟨startEmbryoStage⟩12⟨/startEmbryoStage⟩
  ⟨stopEmbryoStage⟩19⟨/stopEmbryoStage⟩
  ⟨parentId⟩3322⟨/parentId⟩
    ⟨component name=”1st arch” id=”3324”⟩
    ⟨printName⟩embryo.branchial arch.1st arch⟨/printName⟩
    ⟨abbreviation⟩⟨/abbreviation⟩
    ⟨childrenId⟩3327⟨/childrenId⟩
    ⟨childrenId⟩3325⟨/childrenId⟩
    ⟨startEmbryoStage⟩12⟨/startEmbryoStage⟩
    ⟨stopEmbryoStage⟩19⟨/stopEmbryoStage⟩
    ⟨parentId⟩3323⟨/parentId⟩
      ⟨component name=”branchial groove” id=”3325”⟩
      ⟨printName⟩embryo.branchial arch.1st arch.branchial groove⟨/printName⟩
      ⟨abbreviation⟩⟨/abbreviation⟩
      ⟨childrenId⟩3326⟨/childrenId⟩
      ⟨startEmbryoStage⟩12⟨/startEmbryoStage⟩
      ⟨stopEmbryoStage⟩19⟨/stopEmbryoStage⟩
      ⟨parentId⟩3324⟨/parentId⟩
        ⟨component name=”epithelium” id=”3326”⟩
        ⟨printName⟩embryo.branchial arch.1st arch.branchial groove.epithelium⟨/printName⟩
        ⟨abbreviation⟩⟨/abbreviation⟩
        ⟨startEmbryoStage⟩19⟨/startEmbryoStage⟩
        ⟨stopEmbryoStage⟩19⟨/stopEmbryoStage⟩
        ⟨parentId⟩3325⟨/parentId⟩
        ⟨/component⟩
```

Figure 3.2: Partial XML representation of Theiler Stage 19

in this manner. A valid abbreviation for the APICAL ECTODERMAL RIDGE would be AER but this is not stated in the Nomenclature as an abbreviation but rather is represented as a synonym. The 235 synonyms stated in the Nomenclature, as identified by pattern matching in Perl scripts, are not currently being utilised in search but they could provide a significant improvement with regards to user satisfaction in search, as discussed later inChapter 4.

3.1.1 Number of Anatomical Structures

In order to determine whether each anatomical structure in the Nomenclature had a unique designator, it also had to be be known how many distinct anatomical structures a user could specify within the Nomenclature. First it was required to discover how

many nodes there were within these trees. This involved extracting all the lines in the XML code that specified the *printName,* i.e the paths specification of that particular end node. This resulted in the identification of 13727 nodes across all 26 trees[1], suggesting that there was a maximum of 13727 anatomical structures represented in the Nomenclature.

Each *component name* is accompanied by an *id* in the XML code and this could be thought of as denoting the component's unique identifier. This is essentially true as there are indeed 13727 different identifiers, however the same structure present in more than one stage has more than one distinct *id*. For example, the *organ system* in Theiler Stage 16 has the *id* number 1676, while in Theiler Stage 17, what is essentially the same *organ system* has the *id* number 2220.

To find out the minimum number of distinct anatomical structures within the Nomenclature, each represented *component name* was also extracted and condensed by removing duplicates. This resulted in 1416 individual component names that were used in the naming convention within the Nomenclature. These 1416 terms either suggest that there are only 1416 anatomical structures (e.g. bones, organs, tissues) within the mouse or that the Nomenclature exhibits some "term ambiguity". The former, on consultation with an expert, was found to be highly unlikely, and so the task required was to identify the ambiguous terms, locate alternative names for them that could be considered unique and also discover how many actual distinct anatomical terms were represented, (i.e. somewhere between 1416 and 13727).

3.2 Reducing Cognitive Effort

As already noted, *component terms* are not unique designators for anatomical structures: the only unique designators in the Mouse Nomenclature are *path specifications* (i.e the XML *printName*). Thus technically, the only way a user can specify an anatomical structure of interest is to enter the entire path name (in the GXD interface) or to find it through navigating the tree down from the root (in the Mouse Atlas interface).

However, it was possible to identify developmentally valid notions of uniqueness with respect to which some of the 1416 component terms associated with the 13727 nodes in the 26 Theiler Stage trees of the Mouse Anatomical Nomenclature could be taken to be unique.

[1]The size of these trees ranges from 3 nodes in stage 2, during early development, to 1739 nodes immediately pre-birth in stage 26, with the average size being 528 nodes.

3.2.1 Path Uniqueness

The first such notion of uniqueness could be associated with an anatomical structure that develops by some Theiler Stage *j* and then persists under the same name through subsequent stages. In the Mouse Anatomical Nomenclature, this situation corresponded to path specifications that differed only in their root node (which designates the embryo at the corresponding Theiler Stage). If the component term at the terminating node did not occur elsewhere in the Nomenclature outside this path specification, then this component name could be classified as unique.

Every path specification was examined using Perl scripts to locate identical paths and the stages they were present in. In doing this, 1019 component terms were found to be unique in this sense, including EYELID[2], CARDIOGENIC PLATE, CRANIUM. That is, EYELID occurs in Theiler stages 21 to 26 but only ever with the path specification:

EMBRYO.SENSORY ORGAN.EYE.EYELID;

CARDIOGENIC PLATE occurs in Theiler stages 11 and 12, but only with the preceding path

EMBRYO. ORGAN SYSTEM.CARDIOVASCULAR SYSTEM.HEART.
CARDIOGENIC PLATE;

and CRANIUM only ever occurs in stages 20 to 26 with the specification

EMBRYO.SKELETON.CRANIUM.

Although 397 component terms still remained ambiguous, this notion of uniqueness actually meant that approximately 8500 nodes were covered, with just over 5000 still remaining. These 1019 component terms could potentially be used to access gene expression data associated with some or all of the Theiler Stages through which the uniquely designated structure exists, except for one problem that will be described in Section 3.4.2.

3.2.2 Lineage Uniqueness

The second notion of uniqueness was an extension of the first. Before anatomical structures are fully formed, they tend to be referred to by names that denote the same anatomical structure but also convey that it is not fully formed. An example of this is the FUTURE FOREBRAIN, which develops into the FOREBRAIN. Such component terms can be linked with the component term of the structure they develop into, treating the two together as a unique designator across the extended sequence of Theiler Stages. A user seeking gene expression data for the FOREBRAIN without specifying

[2]The fact that there are actually two eyelids is discussed in Section 3.2.3

a particular Theiler Stage, could then select from stages 15-16, which contain the FUTURE FOREBRAIN, as well as from stages 17-26, which contain the FOREBRAIN. This type of pattern can be identified by looking at the children of the particular nodes. For example, the lineage connection between FUTURE FOREBRAIN and FOREBRAIN was identified on analysis of the (two) path specifications for the term DIENCEPHALON:

(1)EMBRYO.ORGAN SYSTEM.NERVOUS SYSTEM.
CENTRAL NERVOUS SYSTEM.FUTURE BRAIN.FUTURE FOREBRAIN.
DIENCEPHALON

in Theiler stages 15 and 16 and;

(2)EMBRYO.ORGAN SYSTEM.NERVOUS SYSTEM.
CENTRAL NERVOUS SYSTEM.BRAIN.FOREBRAIN.DIENCEPHALON

in Theiler stages 17 through 26.

With these two path specifications it could then be said that the two DIENCEPHALON nodes in question were indeed the same structure and therefore the term, DIENCEPHALON, could be considered a unique designator for that structure (i.e. *lineage unique*).

But there is an additional complication here, in that the anatomical structures in which these structures are developing, are themselves developing and changing. Thus the tree specifications of two anatomical structures whose component terms should be taken to co-designate in this lineage sense, may no longer only differ in their root node and their potentially co-designating leaf terms - such trees may also differ simply in the particular component term associated with a non-terminal node – e.g. FUTURE FOREBRAIN above is part of FUTURE BRAIN in stages 15-16, while FOREBRAIN is part of BRAIN in stages 17-26.

Less obvious lineage terms also manifested themselves on further analysis of the path specifications and, with that, the identification of new patterns. On further examination of the term DIENCEPHALON, a further path specification is identified:

(3)EMBRYO.ORGAN SYSTEM.NERVOUS SYSTEM.
CENTRAL NERVOUS SYSTEM.FUTURE BRAIN.PROSCENCEPHALON.
FUTURE DIENCEPHALON

in Theiler stage 14.

Since the pattern of terms, *future X = X*, had already been identified as to denote equivalent terms, FUTURE DIENCEPHALON and DIENCEPHALON could now be considered to represent the same anatomical structure, albeit in different stages of development. Now PROSCENCEPHALON needed to be looked at. Of the above paths, (3) has the same preceding path as (1), i.e. EMBRYO.ORGAN SYSTEM.NERVOUS SYSTEM.CENTRAL NERVOUS SYSTEM.FUTURE BRAIN and has equivalent leaf nodes, as mentioned above. On examination of the stages the two paths appear in, it can be seen that path (3) occurs immediately previous to path (1). PROSCENCEPHALON could then be considered as a

candidate for lineage uniqueness with FOREBRAIN. These cases, once identified, were then simply verified by domain expert Davidson that the intermediate structures were themselves in a lineage relation.

However, the tree paths may also differ in *length*, with the path in the earlier stage tree being longer than that in the later stage tree. This is because the earlier stage specifies the tissue from which the structure is developing from – for example, MESENCHYME in

EMBRYO.LIMB.HINDLIMB.LEG.LOWER LEG.MESENCHYME.FIBULA

in Stage 23, becoming

EMBRYO.LIMB.HINDLIMB.LEG.LOWER LEG.FIBULA

in Stages 24 to 26. To recognise such cases, analysis was required as to which nodes contribute to differences in path length and decide whether two component terms co-specify on that basis. Using pattern matching in Perl scripts[3], any pairs (or more) of path specifications which seemed to exhibit these types of patterns were searched for and then amateur and expert manual verification took place in order to discard the obvious non-lineage patterns and discuss possible ones. A helpful guide in this task was the examination of the stages that each path occurred in, and if a pair of candidate lineage terms did not occur in adjacent stages then they were unlikely to be co-specifying terms. However, any candidate pairs that were temporally adjacent required expert verification.

For example, it could be thought that the two BULBAR CUSHIONs below may denote the same structure since the BULBO-VENTRICULAR REGION in path (4) seems to only appear immediately previous to path (5)

(4)EMBRYO.ORGAN SYSTEM.CARDIOVASCULAR SYSTEM.HEART.
BULBO-VENTRICULAR REGION.BULBAR CUSHION

in Theiler Stage 21 and

(5)EMBRYO.ORGAN SYSTEM.CARDIOVASCULAR SYSTEM.HEART.
BULBAR CUSHION

in Theiler Stages 22 through 26. However, on expert advice the pattern of *bulbo-ventricular region.X = X* was considered not to be viable in a lineage sense.

Another example of an unsuccessful candidate lineage equivalent pair is;

EMBRYO.ORGAN SYSTEM.CARDIOVASCULAR SYSTEM.HEART.ATRIUM.
COMMON ATRIAL CHAMBER.RIGHT PART.VALVE

in Theiler Stages 17 and 18, and then

[3]See Appendix C.1

EMBRYO.ORGAN SYSTEM.CARDIOVASCULAR SYSTEM.HEART.ATRIUM.
RIGHT ATRIUM.VALVE

in Theiler Stages 19 through 26. The sub-path COMMON ATRIAL CHAMBER.RIGHT PART was not considered to be an developmentally earlier version of the RIGHT ATRIUM.

The most common lineage patterns were those involving *future X* and *mesenchyme.X* but others included

- *forelimb bud.X = forelimb.X*, e.g.
  - EMBRYO.LIMB.FORELIMB BUD.ARM Theiler Stage 19
  - EMBRYO.LIMB.FORELIMB.ARM Theiler Stages 19-26
  - i.e. each time the term ARM is encountered it represents the same structure.
- *foregut X = X*, e.g.
  - EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.ALIMENTARY SYSTEM.GUT. FOREGUT.GLAND Theiler Stages 14-23
  - EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.ALIMENTARY SYSTEM.GUT. GLAND Theiler Stages 24-26
  - i.e. the two GLANDs stated here represent the same structure.
- *renal/urinary system.X = urogenital system.X*, e.g.
  - EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.UROGENITAL SYSTEM. NEPHRIC DUCT Theiler Stages 14-18
  - EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.RENAL/URINARY SYSTEM. NEPHRIC DUCT Theiler Stages 19-26
  - i.e. NEPHRIC DUCT is a unique designator.

Again, when these patterns are encountered, the component names, providing they are not involved in any other initial tree paths, can be classified as being unique, further reducing the number of names to be disambiguated. 44 of these lineage patterns have been identified, which extends to the further elimination of 119 (lineage unique) terms from the set of ambiguous component terms.

3.2.3 Group Uniqueness

The third notion of uniqueness that appeared useful for reducing cognitive effort in specifying anatomical structures whose gene expression data is of interest, can be called *group uniqueness*. The original motivation for the Mouse Atlas Nomenclature was not to necessarily include the notion of groups. However, during the design of the

Nomenclature it seems that some sense of grouping was unavoidable. For example, although the component term TOOTH appears in different path specifications, one or more of whose internal nodes specify anatomical structures that are not themselves equivalent

EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.ALIMENTARY SYSTEM.
ORAL REGION.**LOWER** JAW.TOOTH

EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.ALIMENTARY SYSTEM.
ORAL REGION.**UPPER** JAW.TOOTH

(i.e., LOWER JAW does not co-specify with UPPER JAW), they possess some sub-parts with the same anatomical/developmental properties that, in the context of gene expression, could be considered the same.

Again, I developed Perl scripts to identify candidate group patterns (in a similar way as that to identify lineage patterns), which Davidson has then been able to review and identify those cases which should be considered "group unique". This has picked up an additional 60 component terms that could then be removed from the set of ambiguous terms.

Component terms which could also be considered groups are ones that occur in left-right pairs, e.g. FORELIMB, EYELID and FEMUR. However, genetically, each of these pairs of structures are considered to exhibit the same gene expressions and so, in the Mouse Atlas Nomenclature, they are treated as the same entity. Other structures which exhibit different genetic sequences in the left and right parts are specified, e.g. the LEFT and RIGHT ATRIA.

3.2.4 Term Generation

Turning now to component terms that are not unique in any of the senses discussed so far, it still did not appear to be the case that a user needed to enter an *entire* path specification to refer to its associated anatomical structure. In many cases, a *sub-path* specification of two, or in some cases, three component terms appeared sufficient to uniquely specify an anatomical structure of interest.

The question was what Natural Language phrases these multi-component terms correspond to, since it is such phrases that would be used in an interface, not the sequence of component terms. The formation of appropriate phrases, without demanding that our domain expert explicitly list them was investigated. In the process, three different phrasal patterns for two-component sub-paths were found: **(1)** In cases where the child and parent nodes are in a part-whole relation and both are realised as nouns – e.g., a child with component term CAPSULE descending from a parent LENS, or a parent CORTEX or a parent OVARY,

EMBRYO.ORGAN SYSTEM.SENSORY ORGAN.EYE.LENS.CAPSULE

EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.RENAL/URINARY SYSTEM.
METANEPHROS.EXCRETORY COMPONENT.CORTEX.CAPSULE

EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.REPRODUCTIVE SYSTEM.
FEMALE.OVARY.CAPSULE

the multi-component term can be realised as a phrase CHILD OF PARENT[4], generating the three uniquely specifying phrases "capsule of lens", "capsule of cortex" and "capsule of ovary". Alternatively, a natural phrase of the form PARENT CHILD, (i.e. "lens capsule", "cortex capsule" and "ovary capsule" can also be constructed as a natural way of describing the anatomical structure that the path denotes.

(2) In cases where the child and parent nodes are in a part-whole relation, but the component term associated with the child is an adjective such as LEFT, UPPER or ANTERIOR, then the pattern CHILD PARENT can be used to form an appropriate phrase. For example, the path specification

EMBRYO.ORGAN SYSTEM.CARDIOVASCULAR SYSTEM.VENOUS SYSTEM.
VENA CAVA.INFERIOR

can be accessed by the phrase "inferior vena cava".

(3) In cases where the child and parent nodes are in a type-token relation, as in the case of

EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.ALIMENTARY SYSTEM.
ORAL REGION.GLAND.PITUITARY

again the pattern CHILD PARENT can be used to form an appropriate phrase – for example "pituitary gland" in this case. Phrases thus formed from multi-component sub-paths may again be unique with respect to an interval of Theiler Stages, or with respect to lineage within the Theiler Stages, or with respect to a group.

So again, I developed a set of Perl scripts that enumerated all sub-paths from nodes with a non-unique associated component term to the root of their corresponding Theiler Stage tree (i.e., paths being specified in child-parent order), in order to determine which two-component path specifications had a unique denotation and, of the remainder, which three-component path specifications did (e.g. CHILD of PARENT of GRANDPARENT).

With this process, many of the remaining nodes of the trees within the Nomenclature were disambiguated. More specifically, another 105 component terms are covered via the 2-component terms and 54 via the 3-component terms. However, some component terms remained ambiguous as the final three children of their associated path specifications did not correspond to unique designators, leaving 59 path specifications for

[4]See Appendix A.1

which it was still necessary to find unique designators via other methods. It is unlikely that four or more terms would be used in natural language for these remaining path specifications: [2] found that 88% of the noun phrases in relevant MedLine abstracts contained either one or two modifiers.

3.3 Increasing Precision

The introduction of phrases based on more than one component term within the Mouse Anatomy Nomenclature can significantly reduce the number of irrelevant matches obtained over searches that only allow a match within a single component term.

To continue our example with CAPSULE from Section 3.2, the current situation is that no results will be found if "cortex capsule" is entered as a search query. If the user then simply searches for "capsule" across all stages, 31 instances will be returned. However, although all sub-paths leading to CORTEX.CAPSULE are returned, the other 87% of the results are irrelevant to the user's intention. If the multi-component terms were included as uniquely designating replacements of existing terms, 100% recall would be maintained, while increasing precision to 100% percent.

As there is some systematicity involving parent and child component terms, these terms can be automatically generated. Within the Nomenclature, we can use the same set of phrases as noted in Section 3.2.4 for multi-component terms, i.e.

- the child being a descriptor of the parent e.g. superior vena cava
- the child being a part of the parent e.g. ovary capsule or capsule of ovary
- the child being a member of the parent set e.g. pituitary gland.

Of course, recognising which tree path belongs to which of the patterns above required the expert help of Davidson, but once identified, new terms can be generated that are more likely to be used naturally to refer to the relevant anatomical components. Once implemented within the Nomenclature these new terms would ensure that any amount of recall would be of high precision.

Allowing all patterns to be acceptable as component terms, again, significantly increases our chances of obtaining only relevant results to our search. Entering all these patterns (i.e. CHILD OF PARENT, PARENT CHILD, CHILD PARENT and similar patterns involving GRANDPARENTS) as synonyms for existing terms would be one way of enabling this. The advantage of synonyms is more fully described in Chapter 4.

3.4 Other Nomenclature Issues

3.4.1 Context-Dependent Terms

The only problem with using many of the component terms within an interface for gene expression data is that, while unique with respect to the Nomenclature, they do not appear meaningful in everyday terms without taking the context (i.e preceding tree path) into account – for example, while LOOP uniquely denotes the anatomical structure

> EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.ALIMENTARY SYSTEM.
> GUT.MIDGUT.LOOP

a biologist would not simply use the phrase "loop" to refer to the the loop of the midgut. Similarly, while DISTAL uniquely designates

> EMBRYO.LIMB.FORELIMB.JOINT.RADIO-ULNARJOINT.DISTAL

in the Nomenclature, "distal" is neither an every day nor a technical term on its own for the joint of the radius and ulna bones of the forelimb furthest from the shoulder. Some of the unique generated terms of Section 3.2.4 also inherited this problem (e.g. BARE AREA of RIGHT).

As it is extremely unlikely for these terms to be used on their own in a search query, they needed to be augmented to make them more expressive. In these cases, Davidson provided alternative, more meaningful component names to be used in the interface, some of which will replace the existing terms in the Nomenclature, and some of which will be included as synonyms.

3.4.2 Term and Path Inconsistencies

Throughout the life of this project, many unexpected inconsistencies within the existing Nomenclature were discovered outwith the area of term ambiguity. These ranged from trivial term inconsistencies, i.e. spelling mistakes, to path structure inconsistencies as follows

- The HGU decided at the onset of the Nomenclature that each structure name should be singular, however, HAMSTRING occurs both as a plural and a singular term.
- The terms PRE-CARTILAGE CONDENSATION and CARTILAGE CONDENSATION frequently occur within the Nomenclature modified by the relevant term (e.g. VERTEBRAL). However in one case, with the FABELLA, the above two terms are represented as abbreviations, i.e. PCC and CC respectively.

- The cartilage associated with the larynx is represented in 2 different ways, i.e.
  EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.RESPIRATORY SYSTEM.
  LARYNGEAL CARTILAGE
  in Theiler Stages 21 and 22, while in stages 23 to 26, its path specification was
  EMBRYO.ORGAN SYSTEM.VISCERAL ORGAN.RESPIRATORY SYSTEM.
  CARTILAGE.LARYNGEAL.

- In Theiler Stages 21 and 22 there is the path specification
  EMBRYO.ORGAN SYSTEM.NERVOUS SYSTEM.PERIPHERAL
  NERVOUS SYSTEM.SPINAL.NERVE PLEXUS.LUMBO-SACRAL PLEXUS.
  SCIATIC
  while in stages 23 to 26, there is the equivalent path specification
  EMBRYO.ORGAN SYSTEM.NERVOUS SYSTEM.PERIPHERAL
  NERVOUS SYSTEM.SPINAL.NERVE PLEXUS.LUMBO-SACRAL PLEXUS.
  SCIATIC NERVE.

- The (unique) path specification for the LARYNGO-TRACHEAL GROOVE appears in Theiler Stages 15 and 17 only. Did it not exist in Stage 16?

- Another example of where LEFT and RIGHT distinction is relevant is with the RECURRENT LARYNGEAL BRANCHes. However they are represented contrary to the usual part-whole convention, i.e.
  EMBRYO. ... .NERVE.VAGAL X NERVE TRUNK.
  LEFT RECURRENT LARYNGEAL.RECURRENT LARYNGEAL BRANCH
  RIGHT RECURRENT LARYNGEAL BRANCH.RECURRENT LARYNGEAL BRANCH.

- An extra unnecessary term has been included in Stage 13 in the path specification
  EMBRYO.BRANCHIAL ARCH.1ST ARCH.MESENCHYME. HEAD MESENCHYME.
  MESENCHYME DERIVED FROM HEAD MESODERM
  whereas the path should be the same as
  EMBRYO.BRANCHIAL ARCH.1ST ARCH.MESENCHYME.MESENCHYME
  DERIVED FROM HEAD MESODERM
  in stages 12 and 14.

All the above examples, once pointed out to Davidson, were deemed to be inconsistent with the intended design of the Nomenclature and noted for amendment.

Another inconsistency, but one that apparently would not be considered for amendment, is that concerning the vertebrae. The typical path specifications for the cervical vertebrae are

EMBRYO.SKELETON.AXIAL SKELETON.CERVICAL REGION.
(1) INTERVERTEBRAL DISC.C1
(2) VERTEBRA.C1
(3) VERTEBRAL CARTILAGE CONDENSATION.C1
(4) VERTEBRAL PRE-CARTILAGE CONDENSATION.C1

C1 would normally be thought of as a particular vertebra and so path (2) could possibly be thought of denoting a member, C1, of the group VERTEBRA. However, the C1 here should be taken as the area containing the C1 vertebra as, for example, path (4) could not be taken as C1 being part of the VERTEBRAL PRE-CARTILAGE CONDENSATION or a member of its group. This is contrary to the part-whole convention of the Nomenclature. The relationships as explained to me between these structures would possibly be better represented as C1 being the parent of INTERVERTEBRAL DISC, VERTEBRAL CARTILAGE CONDENSATION, VERTEBRAL PRE-CARTILAGE CONDENSATION and also possibly VERTEBRA, if not considered a group.

3.5 Improving the Display of Search Results

3.5.1 Current Result Display

Currently, within the interface to the Gene Expression Database, one can search for an anatomical structure of interest within a single tree or across all stages. A query across all stages results in a list of all stages with a matching component, and associated with each stage is one or more path specification terminating at a matching component name. This does not easily enable the user to locate the specific entity they are interested in. If the term is not unique, then the results contain all possible anatomical structures the query could represent.

For example, the result of string matching on the phrase “lumen” is partially illustrated in Figure 3.3. If the full search results were shown this would take up 189 lines of text (ignoring white space) which is equivalent to 6 html pages.

One of Jacob Nielsen’s frequently quoted “Top 10 Mistakes in Web Design” [15] is that a designer should avoid endlessly scrolling pages, which this example would certainly exhibit. In 1996, it was found that only 10% of users would scroll down a page to find the item of interest to them. Although this percentage has since increased, it is still an issue to be considered. The LUMEN example is not the largest set of search results that can be elicited from the GXD, and so a far greater number of lines of text can be expected from other search terms (e.g. the 1056 senses of MESENCHYME involved across 16 stages).

Locating the entity of interest amongst all these tree paths, where each sense represents a different tree path, would be an arduous task if the query term occurred many times within the Nomenclature in different tree paths. Even if the term was unique, the same entry could be repeated across multiple stages, leading to a visual search problem.

AD Browser - query results

---

4 Theiler stage 12 term(s) matching query "**lumen**":

future spinal cord;neural tube;neural **lumen**
foregut diverticulum;**lumen**
hindgut diverticulum;**lumen**
midgut;**lumen**

---

5 Theiler stage 13 term(s) matching query "**lumen**":

future spinal cord;neural tube;neural **lumen**
foregut diverticulum;**lumen**
hindgut diverticulum;**lumen**
midgut;**lumen**
foregut-midgut junction;**lumen**

---

7 Theiler stage 14 term(s) matching query "**lumen**":

future spinal cord;neural tube;neural **lumen** (future spinal canal, spinal canal)
hindgut diverticulum;**lumen**
midgut;**lumen**
foregut-midgut junction;**lumen**
rest of foregut;**lumen**
foregut;pharyngeal region;**lumen**
otic pit;**lumen**

---

10 Theiler stage 15 term(s) matching query "**lumen**":

optic recess (**lumen** of optic stalk)
future spinal cord;neural tube;neural **lumen** (future spinal canal, spinal canal)
hindgut diverticulum;**lumen**
midgut;**lumen**
pharynx;**lumen**
foregut-midgut junction;**lumen**
hindgut;**lumen**
rest of foregut;**lumen**
foregut;oesophageal region;**lumen**
otic pit;**lumen**

Figure 3.3: Partial search results for search string *lumen*

3.5.2 Proposed Result Display

An alternative, cleaner way of presenting search results would be to take the matching component terms as the primary display key and then associate it with a list of stages where its corresponding path specification occurs. That is, each “sense” of the search term is only stated once with its corresponding Theiler stages. For non-unique search queries such as the example above, this would result in Figure 3.4.

Here, if the full results were shown, this would result in 72 lines of text, as opposed to the current 189 lines, since there are 36 senses of the term LUMEN.

AD Browser - query results

future spinal cord;neural tube;neural **lumen** matches query **"lumen"** in:
Theiler stages 12, 13, 14, 15, ...

foregut diverticulum;**lumen** matches query **"lumen"** in:
Theiler stages 12, 13

hindgut diverticulum;**lumen** matches query **"lumen"** in:
Theiler stages 12, 13, 14, 15, ...

midgut;**lumen** matches query **"lumen"** in:
Theiler stages 12, 13, 14, 15, ...

foregut-midgut junction;**lumen** matches query **"lumen"** in:
Theiler stages 13, 14, 15, ...

rest of foregut;**lumen** matches query **"lumen"** in:
Theiler stages 14, 15, ...

otic pit;**lumen** matches query **"lumen"** in:
Theiler stages 14, 15, ...

optic recess;**lumen** matches query **"lumen"** in:
Theiler stages 15, ...

pharynx;**lumen** matches query **"lumen"** in:
Theiler stages 15, ...

foregut;oesophageal region;**lumen** matches query **"lumen"** in:
Theiler stages 15, ...

foregut;pharyngeal region;**lumen** matches query **"lumen"** in:
Theiler stages 14

Figure 3.4: Alternative display of search results for search string LUMEN

On the other hand, it is possible, although it does not frequently occur, for the current display to be more concise than the proposed display. For example, the ALVEOLUS appears in 5 path specifications across only 2 stages. The current display would then result in 10 lines of text (i.e. 2 stages x 5 senses), with the proposed display resulting in 50% more lines of text (i.e. (1 sense + 2 stages) x 5 senses = 15 lines).

The alternative display still has to be verified that it better facilitates users finding the structure(s) and stage(s) of interest to them.

4. Increasing Recall

Even when every anatomical part in the Nomenclature is designated uniquely, there is also still the issue concerning the variation in how people refer to these parts. The addition of synonyms into the search facility would not only increase recall but also support substitutivity (a principle of *usability*) in that equivalent values of input (i.e. synonyms) can be arbitrarily substituted for each other.

4.1 Extracting New Terms

While the Mouse Anatomical Nomenclature was designed to specify every anatomical structure within the developing mouse embryo, it does not contain all the terms that developmental biologists might use to refer to anatomical entities. Although some synonyms have been explicitly recorded in the Nomenclature, no attempt has been made to be exhaustive. In order to increase the *recall* of user searches for anatomical structures, the aim was to increase the number and range of synonyms for elements of the Nomenclature, by semi-automatically analysing texts likely to contain terms related to the developmental anatomy of the mouse.

To demonstrate the potential value of this approach, a manual review was carried out of the short textual descriptions that accompany each Theiler stage within the Mouse Atlas and highlights the main features of the stage – for example, this short description accompanies the schematic of Theiler Stage 20 (Figure 4.1):

> The handplate (anterior footplate) is no longer circular but develops angles which correspond to the future digits. The posterior footplate is also distinguishable from the lower part of the leg. It is possible to see the pigmentation of the pigmented layer of the retina through the transparent cornea. The tongue and brain vesicles are clearly visible.

All the noun phrases (NPs) in these descriptions that could potentially refer to an anatomical structure were collected and within the 1380 words comprising the descriptions, 25 anatomical terms were discovered that were not included in the Nomenclature either as component terms or as synonyms for component terms. From the above extract, *anterior footplate, posterior footplate* and *pigmented layer (of the retina)* were not specified anywhere in the Nomenclature. Since the same people developed these textual descriptions as developed the Nomenclature, it shows how difficult it is to record all terms used for anatomical structures without systematic effort.

To support such a systematic effort, text analysis software was applied to a textbook on developmental anatomy, including a tokenizer, part-of-speech tagger and NP chunker,

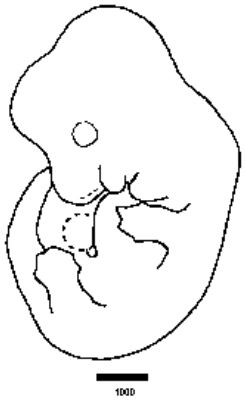

Figure 4.1: Schematic of a Theiler Stage 20 embryo

the latter two being Hidden Markov Model (HMM) techniques from the Language Technology Group (LTG)[1] in Edinburgh, as well as additional Perl scripts – in order to identify noun phrases, from which were extracted those most likely to refer to an anatomical structure. The latter were then discussed with the domain expert, Davidson.

The text analysed was the chapter from [13] that describes the heart. In order to use LTG's POS tagger, the text first had to be tokenised. This involved (1) splitting punctuation from adjoining words, (2) converting double quotes to doubled single forward and backward quotes and (3) splitting verb contractions and the possessive *'s* from the component morphemes.

For example

1. how are you? → how are you ?
2. "quote" → `` quote ''
3. won't → wo n't
   Nigel's → Nigel 's

I wrote a Perl script to enable any input text to be tokenised in this manner[2]. This tokenised text was then input into the HMM tagger and chunker and the tagged and chunked output was saved to file.

For example, a part of the original text was

> The heart starts to develop in the anterior part of the ventral region of the embryo soon after gastrulation is complete (E7.5-8). Two endothelial tubes form from a plexus of endothelial cells believed to be of lateral plate (splanchnopleuric) mesodermal origin which straddle the midline in

[1] http://www.ltg.ed.ac.uk

[2] Verb contractions were not robustly taken care of as it was deemed less important considering the task was to extract noun phrases

> the region subjacent to the intraembryonic coelomic cavity and these then aggregate to produce a single heart tube that is soon surrounded by a myocardial 'mantle' layer. This comes from the cardiogenic plate that differentiated from the lining of the ventral part of the intraembryonic coelomic cavity.

Part of Speech tagging is a commonly used technique in many areas of Natural Language Processing (NLP); for example, information extraction and retrieval, spelling correction, text to speech systems as well as terminology extraction and mining. The main goal of POS tagging is to assign appropriate morpho-syntactic categories to words with respect to their context. Depending on the tagging technique used, ambiguity (e.g. words that can be classified as either a noun or a verb) may be handled by a combination of lexical and local contextual constraints or by a random choice between the options. Similarly, unrecognised words may be analysed and tagged according to any similarities to known words such as suffixes or just simply always tagged as a noun. Techniques behind POS tagging can be rule based, stochastic and neural network models, any of which can then be supervised or unsupervised. LTG's HMM tagger is an unsupervised, stochastic model which involves tag sequence (e.g. *determiner-noun-verb*) probabilities and word frequency measurements.

Text chunking consists of grouping (i.e. *chunking*) the set of tagged words into non-overlapping phrases and the NP chunker deals with part of this task by, as the name suggests, identifying chunks that constitute a noun phrase.

After tagging and chunking the text input was transformed as below[3]:

> ⟨/S⟩[[The_DT heart_NN]] ((starts_VBZ)) ((to_TO develop_VB)) in_IN [[the_DT anterior_JJ part_NN]] of_IN [[the_DT ventral_JJ region_NN]] of_IN [[the_DT embryo_NN]] soon_RB after_IN [[gastrulation_NN]] ((is_VBZ)) complete_JJ (_(E7_SYM._.⟨/S⟩ ⟨S⟩5_LS -_: [[8_CD]])_) ._.⟨/S⟩ ⟨S⟩[[Two_CD endothelial_JJ tubes_NNS]] ((form_VBP)) from_IN [[a_DT plexus_NN]] of_IN [[endothelial_JJ cells_NNS]] ((believed_VBD)) ((to_TO be_VB)) of_IN [[lateral_JJ plate_NN]] (_(splanchnopleuric_JJ)_) [[mesodermal_NN origin_NN]] [[which_WDT]] ((straddle_VBP

[3]The tag symbols after each word relate to

CC co-ordinating conjunction

CD cardinal number

DT determiner

IN preposition/subordinating conjunction

JJ adjective

NN singular or mass noun

NNS plural noun

RB adverb

TO "to"

VB uninflected verb

VBD past tense verb

VBG present participle verb

VBN past participle

VBP singular present verb (non 3rd person)

VBZ third person singular present verb

WDT "wh" determiner

)) [[the_DT midline_NN]] in_IN [[the_DT region_NN subjacent_NN]] to_TO [[the_DT intraembryonic_JJ coelomic_JJ cavity_NN]] and_CC [[these_DT then_JJ aggregate_NN]] ((to_TO produce_VB)) [[a_DT single_JJ heart_NN tube_NN]] [[that_WDT]] ((is_VBZ soon_RB surrounded_VBN)) by_IN a_DT myocardial_JJ '_" [[mantle_NN]] '_" [[layer_NN]] ._.⟨/S⟩ ⟨S⟩[[This_DT]] ((comes_VBZ)) from_IN [[the_DT cardiogenic_JJ plate_NN]] [[that_WDT]] ((differentiated_VBD)) from_IN [[the_DT lining_VBG]] of_IN [[the_DT ventral_JJ part_NN]] of_IN [[the_DT intraembryonic_JJ coelomic_JJ cavity_NN]] ._.⟨/S⟩

Since noun phrases were the area of interest, each separately tagged entity was separated out according to their syntactic category. That is, all noun groups (i.e. those enclosed within [[ ]] parentheses) were kept separate from all verb groups [4](i.e. those enclosed within (( )) parentheses) which were similarly separated from all other syntactic categories (i.e. outwith parentheses).

[4]The verb phrases were singled out for the possible future analysis of verbs commonly occurring either immediately previous or following anatomical terms. See Section 4.2

To continue this example the 3 relevant files would then be

NPs	VPs	Other
The heart	starts	in
the anterior part	to develop	of
the ventral region	is	soon after
the embryo	form	complete
gastrulation	believed	from
Two endothelial tubes	to be	in
a plexus	splanchnopleuric	to
endothelial cells	straddle	and
lateral plate	to produce	that
mesodermal origin	is soon surrounded	by a myocardial
which	comes	from
the midline	differentiated	
the region subjacent		
the intraembryonic coelomic cavity		
these then aggregate		
a single heart tube		
mantle		
layer		
This		
the cardiogenic plate		
that		
the lining		
the ventral part		

As can be seen by the above table, the tagger did not classify everything correctly. For example, *these then aggregate* is classified as a noun phrase, where the tagger presumably has decided upon the noun form of *aggregate* with *these then* acting as adjectives. However, in context, *aggregate* is actually acting as verb.

Although an HMM approach to POS tagging and chunking is one of the more efficient tagging methods of those mentioned above, since it is supervised, it does initially require pre-tagged corpora for training. The best performance will result when the tagger is tested on the same genre of text it was trained on. Because neither the POS tagger nor the chunker were specially trained for this genre of technical text[5], their performance was rather weak.

Chunker output from the chapter on the heart, produced 5547 phrases, of which 2465 were considered to be NPs. 2.4% (i.e. 74) of the 3082 claimed non-NPs were obvious false negatives, where such “obvious” phrases were ones other than those which could

[5]The training data for LTG’s tagger and chunker came from editions of the Wall Street Journal

be taken to be noun or verb phrases e.g. *figure* and *term*. Most of this percentage involved the terms *vena cava*, *septum primum/secundum* and *ductus arteriosus/venous*. Of the terms classified as NPs, 3.7% (i.e. 92) were found to be false positives. Most of these errors involved words that could be classed as either verbs or nouns and which were adjacent to true NPs, and then classified as plural nouns but which, in context, were acting as third person singular present verbs, e.g. *the ostium secundum forms*.

Of the true NPs, duplicates, and irrelevant and/or non-anatomical phrases and modifiers were semi-automatically removed so that only probable anatomical terms remained. Such terms included pronouns, anaphora, numbers, initials, authors names, and plural terms whose singular form was also present. A Perl script was created to remove the commonly occurring irrelevant terms such as *that, his, two, another, several*. Plurals were dealt with by automatically singularising the NPs as far as possible and then removing the plural form if a singular entry was found. This involved identifying phrases with endings such as *s, es, ies* and inserting their appropriate singular form. The remaining NPs were then manually analysed for anomalies which may have been missed through automation, such as rare plural inflections (e.g. *vena cavae*) and infrequent irrelevancies. Matches with existing Nomenclature terms were automatically removed through pattern matching with Perl and again plurals were taken into account.

Of the 451 relevant NPs, 82 were found to be exact matches for component terms and 8 were matches for the synonyms in the Nomenclature. These were also removed, leaving 361 possible anatomical terms not found in the Nomenclature. This set was further reduced by only considering NPs headed by or modified by a frequent head or modifier from within the set of component terms. Here, frequent meant $\geq$3 times. For example, *carotid, fibrous* and *endocardial* were found to be frequent modifiers, while *artery, septum* and *tissue* were found to be frequent head nouns. These frequently headed or modified terms were identified automatically and of the 361 remaining NPs from the heart chapter, 115 shared a high frequency head noun with terms already in the Nomenclature, while 105 shared a high frequency modifier[6] with terms in the Nomenclature. These 220 NPs were considered *probable* anatomical terms, with the remaining 141 being *possible* anatomical terms. I then discussed these two sets of probable and possible NPs with Davidson on whether they denoted anatomical entities of interest to the Mouse Atlas and its Nomenclature - see Section 5.2.

4.2 Related Work

While the work described in this section uses only the resources provided by the Mouse Atlas itself and one of the books it was based on [13], the original intention was to also use the National Library of Medicine's MedLine and UMLS (Unified Medical Lan-

[6]See Appendix B.1

guage System) [12] resources to identify synonyms for the Nomenclature path specifications. MedLine is a searchable database containing abstracts and citations from more than 4,600 biomedical publications. UMLS contains three knowledge sources, one of which, the Metathesaurus, contains over 1.5 million medical terms. The Metathesauraus has associated with each term semantic considerations such as synonyms (similar in meaning), hyponyms (more specific), hypernyms (more general). Bodenreider, Rindflesch and Burgun, 2002, [2] also uses these resources to extend biomedical terminology with a similar methodology as this project's work. That is, Bodenreider et al use the modifiers and heads of noun phrases found in the UMLS to identify new noun phrases from MedLine that could then be included in the Metathesaurus[7]. For example, *pancreatic bronchogenic cyst* was found in a MedLine abstract and could be considered for incorporation into the Metathesaurus because the phrases *bronchogenic cyst* and *pancreatic haemorrage* were existing terms in the Metathesaurus. The use of known domain relevant head nouns for term identification has also been used in [16].

Demetriou and Gaizaukas [6] have also used domain specific (i.e. protein structures) journal abstracts for automatic term acquisition but with initially only a small number of known (protein) terms and without the need for any prior syntactic or semantic knowledge (i.e. tagging). Their algorithm involves an iterative acquisition of seed terms (protein names) and their contextual patterns. The initial terms are used to identify frequent contextual patterns which in turn are used to identify other possible protein names and the process is repeated until no other new patterns or seeds are encountered. The initial part of this algorithm is a similar methodology, but not technique, to the reasoning behind the retention of verb phrases found in the heart chapter in this project. The context of these VPs were to be analysed with respect to frequently occurring immediate predecessors and ancestors which may then have denoted anatomical terms. This would also constitute a contextual pattern similar to that of Demetriou and Gaizaukas, e.g. $\langle$X$\rangle$ *encodes a* $\langle$Y$\rangle$. Blaschke, Andrade, Ouzounis and Valencia [1] and Craven and Kumlein [4] used similar NL properties to identify protein-protein interactions from relevant MedLine abstracts.

The above two techniques were integrated by Rindflesch, Hunter and Aronson [18], who combined the identification of noun phrases from MedLine abstracts with the further restriction of only considering NPs associated with forms of a certain verb, *bind,* in order to extract information about molecular binding affinities.

Abbreviated anatomical terms truncated due to context (e.g. *atrial branch* in the text following *anterior atrial branch of right coronary artery*) has also been researched in work by Sneiderman, Rindflesch and Bean [21] with respect to hospital patient records.

[7]Bodenreider et al's work was done in parallel/simultaneously to this project and the two exhibit a remarkably similar methodology. This project, however, restricted itself to a far smaller body of text and known biological terms to work with.

5. Evaluation

Throughout the life of this project it seemed necessary to do some manual analysis of the terminology involved. Now that I know what type of language and free text phenomenon that would be looked for, and also now that I have experience with pattern matching with Perl, I believe I could make all resources developed more fully automated.

5.1 Existing Terms

The 44 lineage patterns will not be exhaustive as there may be ancestral lines that do not exhibit any systematicity or perhaps patterns that I have missed. Indeed, on analysis of the remaining 59 terms, the term TELENCEPHALIC PART OF INTERVENTRICULAR FORAMEN would appear to be a unique designator for the 2 path specifications in which it appears:

> EMBRYO.ORGAN SYSTEM.NERVOUS SYSTEM.CENTRAL NERVOUS SYSTEM.
> FUTURE BRAIN.FUTURE FOREBRAIN.TELENCEPHALON.TELENCEPHALIC
> VESICLE.TELENCEPHALIC PART OF INTERVENTRICULAR FORAMEN

and

> EMBRYO.ORGAN SYSTEM.NERVOUS SYSTEM.CENTRAL NERVOUS SYSTEM.
> BRAIN.FOREBRAIN.TELENCEPHALON.LATERAL VENTRICLE.TELENCEPHALIC
> PART OF INTERVENTRICULAR FORAMEN

The pattern *lateral vesicle.X = telencephalic vesicle.X* had been verified as being a valid lineage pattern by Davidson. However, on searching for this pattern, the twice occurring pattern *future X = X* was not also taken into account. This suggests that a more combinatorial approach to pattern matching is required.

The terms FUTURE SPINAL CORD and SPINAL CORD also appear in the remaining 59 terms. This would seem like an oversight, but the former appears in more path specifications than the latter and so the pattern matching alone could not allow these terms to be considered equivalent.

57 of the 60 terms considered *group unique* were simply the alphanumeric names of vertebrae while the other 3 were related to the teeth and so this meant that there seemed to be only 2 types of groups within the Nomenclature. However, this low return of *group unique* terms is understandable given that the initial design of the Nomenclature was not intended to include groups (See Section 6.1.2).

Equivalence Class	Number Unique	%age Covered	Remaining Ambiguous
Original Component Names	1416		1416
Identical Tree Paths	1019	72	397
Lineage Unique	119	30	278
Group Unique	60	22	218
2-Term Compound Names	105	48	113
3-Term Compound Names	54	48	**59**

Table 5.1: Summary of component term disambiguation

Summaries of the results of the work done on the existing terms in the Mouse Atlas Nomenclature are shown in Table 5.1 and Table 5.2. These show that the remaining 59 ambiguous terms[1] occur across 2019 nodes. That is, there are still between 59 and 2019 anatomical structures that require unique identifiers. To this end, these terms and nodes will require more expert (biological) analysis than even my new found anatomical knowledge can muster.

With respect to an enhanced interface to the gene expression data, it is now possible to take the results of these analyses and use them to provide a potentially more effective way of searching for relevant anatomical structures within the Nomenclature and displaying the results.

5.2 Heart Terms

The extraction of terms from the Heart chapter, after analysis with Davidson, was found to have identified 28 possible synonyms of existing terms as well as 38 terms that were not included in any way in the Nomenclature but that were deemed appropriate for inclusion. 18 of these synonyms came from the *probable* set of terms, (split evenly between the frequent modifier terms and frequent head terms) as did 32 of the *new*

[1] See Appendix A.2

Equivalence Class	Nodes Covered	%age Covered	Remaining Nodes
Original Component Names	13727		13727
Identical Tree Paths	8604	63	5123
Lineage Unique	251	5	4872
Group Unique	736	15	4136
2-Term Compound Names	1251	30	2885
3-Term Compound Names	866	30	**2019**

Table 5.2: Summary of node disambiguation

terms (split 20 and 12 respectively). This corresponds to 75% of the synonyms found and 82% of the new terms identified were obtained through the frequency measurements showing that initial analysis of existing terms provides a more efficient method for anatomical entity recognition. Fractionally under 40% of the *possible* set also seemed to denote anatomical structures, which corresponds to 85% recall of anatomical terms from the text using frequent modifiers and heads. However the majority were truncated versions of structures previously referred to, so this figure could increase on elimination of contextually referring terms. The *possible* set did, however, glean 6 synonyms and 6 new terms.

However, a far greater number of other terms were also identified that did not refer to unique entities relevant to the Nomenclature. Some of these were *group* types that were either too general for inclusion or were in consideration for further implementation in the Mouse Atlas[2].

The probable set still contained several terms that already existed in the Nomenclature but that were modified by terms not already removed as described above, e.g. *single straight primitive heart tube*, where PRIMITIVE HEART TUBE was an existing component term. A more rigorous analysis of the NPs chunked and candidate words for removal would significantly reduce this number. Using the frequency measurement did, however, obtain 100% precision in both the modifier and head cases, in that all retrieved *probable* terms were actual anatomical terms. This suggests that it could be

[2]See Section 6.1.2

effective to reduce the frequency limit to increase recall, without significantly reducing precision.

The Nose and Mouth chapter of [13] was also used to extract anatomical terms. Of the 2617 noun phrases extracted, 272 *probable* terms (110 frequently modified terms, 162 frequently headed) and 193 other *possible* anatomical terms were elicited using the same process of removal of duplicates and irrelevant terms. The *probable* set again exhibits 100% precision while 61 terms in the *possible* set appear to be valid anatomical structures. The frequency measurement therefore exhibits 87% recall in this case. Again, a significant amount of the *possible* NPs were contextually referring terms. The nose and mouth terms have not been expertly analysed with respect to synonyms and new terms suitable for inclusion in the Nomenclature. With both chapters, the tagger handled hyphens (which occur frequently in biology) incorrectly which would need to be looked at before future use.

These results (regarding synonym extraction) may at first seem disappointing but they are perhaps not unexpected since the authors of the text was significantly involved in the development of the Mouse Atlas Nomenclature and can be expected to be consistent to some degree with their anatomical terminology. Analysis of text by anatomical expert authors not involved in the Mouse Atlas would most likely produce a far higher number of useful new synonyms.

6. Future Work

6.1 Mouse Atlas Nomenclature

6.1.1 Incorporation into Search

Currently, if an existing synonym in the Nomenclature is input as a search term, no match is found and this needs to be addressed. Although the input term may not be part of the tree representation visible to the user, the system should take the user to the synonym's visible equivalent. Once this is facilitated, any identified additional synonyms also need to be incorporated into the search engine to enable the increase in recall.

An addition to this, to further improve recall, is to also include modifier and head noun synonyms. For example, a structure presently within the Nomenclature is the CARDIAC MUSCLE. Any search for its equivalent, *heart muscle*, would result in no match. However if the system was structured so that when there was no whole-term match, it looked for partial synonyms, then the search may succeed, i.e. if *heart* was known as a synonym for *cardiac* then the term *heart muscle* could be searched for and result in a match for CARDIAC MUSCLE.

A multi-modifier term may also be input as a search string, where some of the modifiers are known terms or synonyms, and some are not. In this case, the most specific match could be returned. For example, search term *1modifier 2modifier 3modifier headnoun* may result in a partial match where the *headnoun* is unknown, where one of the modifiers is unknown or where two of the modifiers are unknown. The last of these matches would be discarded as it too general a match. If weighting (i.e. preference) is given to a head noun match, then the former would be chosen as the result of search.

These are possible changes which I may be involved in during my placement with HGU this summer.

6.1.2 Related Additions

The Nomenclature structure (i.e. that of trees) does not fully represent the true relationships between anatomical structures as certain structures can be thought of as being part of more than one structure and can also be considered a member of different structure groups. To emulate this the MRC's HGU are considering in the future to re-implement the Nomenclature as more of a rooted Directed Acyclic Graph (DAG)

where each node can have more than one parent but cannot have its children being a parent of its ancestors (i.e. cycles).

The team at the HGU are also currently implementing the sense of lineage into the Mouse Atlas so that gene types can be compared across stages. Cell types are also being added. These are not yet available online.

6.2 Named Entity Recogniser

This project has developed a method to mine for anatomical terms, that can be applied to additional texts, ideally after re-training the POS-tagger and chunker to better reflect the types of texts to be dealt with. The fact that 100% precision was returned in the extraction of anatomical terms from the Heart chapter gives encouragement to the idea that an anatomical Named Entity Recogniser (NER) can be developed. A general NER searches through text for entities such as person names, organisation names, locations, dates, times and numbers, i.e. the who, where, when and how much in a sentence. Zhou and Su [23] are currently working on a general NER using an HMM-based tagger and chunker. An anatomical NER, as you would expect, would be more specific and search for anatomical entities.

6.3 Other Anatomical Databases

The flood of data on genomics has led to the emergence of a new inter-species field where organisation, analysis and processing of this data has to be done efficiently. The biological databases are of many different kinds and more and more are emerging with the advent of new technology in the respective fields. Although the different databases are related to each other, the link is not always clear. The challenge lies in understanding the link between these databases to glean new biological information.

One result of this work has been to catch both structural and terminological inconsistencies in the *Mouse Anatomical Nomenclature* because output from my Perl scripts has made it very easy for biologists to see differences from one branch to another or one tree to another in the Nomenclature. On an initial Perl analysis of the nomenclature of the Gene Ontology (GO) [3], I found that there were 128 terms that were represented multiple times but with different identification numbers. The majority of the ontology portrays consistent structural and terminological representations but these 128 terms and their specifications would need to be investigated to ensure that GO was fully consistent. GO is one of the largest and most developed biological ontologies in existence, and so if it can exhibit term inconsistency then it is likely that most other anatomical databases will do the same.

Similar nomenclatures of developmental anatomy exist for other model organisms, including drosophila, zebra fish and human. These too will be used to index gene expression data for these organisms, eventually supporting cross-species comparison of gene expression patterns and further understanding of development.

Interoperability between these anatomical databases is an important future development so that genetic similarities between species can be identified. That is, it may be a possibility that the genetic sequence of a drosophila leg is equivalent to a fin of the zebra fish. A new project, XSPAN (Cross-Species Anatomy Network)[1] has recently been launched by the University of Edinburgh and Heriot-Watt University to develop just such interoperability between the anatomy of different species. The more uniform each database, the better understanding of genomic parallels across species can be obtained.

[1] http://hw-news.ac.uk

Bibliography

[1] C. Blaschke, M. Andrade, C. Ouzounis, and A. Valencia. Automatic extraction of biological information from scientific text: Protein-protein interactions. In *International Conference on Intelligent Systems for Molecular Biology. Heidelberg, 1999 (In press)*, 1999.

[2] Olivier Bodenreider, Thomas Rindflesch, and Anita Burgun. Unsupervised, corpus-based method for extending a biomedical terminology. In *Proceedings of ACL'02: the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[3] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genet*, 25:25–29, 2000.

[4] M. Craven and J. Kumlien. Constructing biological knowledge-bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86, Germany, 1999.

[5] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

[6] George Demetriou and Robert Gaizauskas. Automatically augmenting terminological lexicons from untagged text. In *Proceedings of 2nd International Conference on Language Resources and Evaluation*, 2000.

[7] Alan Dix, Janet Finlay, Gregory Abowd, and Russell Beale. *Human-Computer Interaction*. Prentice-Hall, 1997.

[8] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing '98 (PSB'98), Jan. 1998.*, 1998.

[9] R. Gaizauskas and A. Robertson. Coupling information retrieval and information extraction: A new text technology for gathering information from the web. In *Proceedings of the 5th RIAO Computer-Assisted Information Searching on Internet, pages 356–370, Montreal, Canada, June 25–27*, 1997.

[10] M. Hearst. Untangling text data mining. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics.*, 1999.

[11] David A. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Research and Development in Information Retrieval*, pages 329–338, 1993.

[12] B.L. Humphreys, D.A.B Lindberg, H.M Schoolman, and G.O Barnett. *The Unified Medical Language System: An informatics research collaboration*. 1998.

[13] Matthew H. Kaufman and Jonathan Bard. *The anatomical basis of mouse development*. Academic Press, 1999.

[14] Un Yong Nahm and Raymond J. Mooney. A mutually beneficial integration of data mining and information extraction. In *AAAI/IAAI*, pages 627–632, 2000.

[15] Jakob Nielsen. Top ten mistakes of web design, 1996.

[16] Chikashi Nobata, Nigel Collier, and Jun ichi Tsujii. Automatic term identification and classification in biology texts. In *Proceedings of natural language Pacific Rim Symposium*, 1999.

[17] Obstacles of Nomenclature. *Nature*, 389(6646), 1997.

[18] T. Rindflesch, L. Hunter, and A. Aronson. Mining molecular binding terminology from biomedical text. In *Proceedings of the AMIA Annual Symposium*, 1999.

[19] Steffen Schulze-Kremer. Ontologies for molecular biology. In *3rd Pacific Symposium on Biocomputing*, pages 705–716, 1998.

[20] Gail Sinclair, Bonnie Webber, and Duncan Davidson. Enhanced free text access to anatomically-indexed data. In *Proceedings of ACL'02: the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[21] Charles A. Sneiderman, Thomas C. Rindflesch, and Carol A. Bean. Identification of anatomical terminology in medical text. In *Journal of the American Medical Informatics Association*, pages 428–432, 1998.

[22] K. Theiler. *The House Mouse: Atlas of Mouse Development*. Springer Verlag, New York, 1989.

[23] Guodong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of ACL'02: the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[24] P Zweigenbaum, B Habert, A Nazarenko, and J Bouaud. Tuning an existing nomenclature for specific domain corpora: a syntax-based similarity method. *Journal of the American Medical Informatics Association*, 5(suppl):1110, 1998.

Appendix A. Existing Terms

A.1 Child of Parent Pattern

Alternative compound names for existing terms created using the pattern 'child of parent'.

accessory XI of cranial
accessory XI of nerve
aditus of larynx
alar plate of lateral wall (x3)
allantois of extraembryonic component
allantois of mesoderm
alveolar duct of alveolar system (x5)
alveolar sulcus of lower jaw
alveolar sulcus of mandible primordium
alveolar sulcus of maxillary process
alveolar sulcus of upper jaw
alveolar system of bronchiole (x5)
alveolus of alveolar system (x5)
annulus fibrosus of C1
annulus fibrosus of C2
annulus fibrosus of C3
annulus fibrosus of C4
annulus fibrosus of C5
annulus fibrosus of C6
annulus fibrosus of C7
annulus fibrosus of L1
annulus fibrosus of L2
annulus fibrosus of L3
annulus fibrosus of L4
annulus fibrosus of L5
annulus fibrosus of L6
anterior of cardinal vein
anterior of cerebral artery
anterior of communicating artery
anterior of lens fibres
anterior of naris
anterior of spinal artery
apical ectodermal ridge of ectoderm (x4)
arachnoid mater of meninges (x5)
arterial system of cardiovascular system (x2)
associated mesenchyme of anal canal
associated mesenchyme of carina tracheae
associated mesenchyme of caudal part (x2)
associated mesenchyme of cochlear duct
associated mesenchyme of colon
associated mesenchyme of crus commune
associated mesenchyme of ductus reuniens
associated mesenchyme of duodenum
associated mesenchyme of endolymphatic appendage (x2)
associated mesenchyme of endolymphatic duct
associated mesenchyme of endolymphatic sac
associated mesenchyme of foregut diverticulum
associated mesenchyme of foregut-midgut junction
associated mesenchyme of fundus (x2)
associated mesenchyme of gastro-oesophageal junction (x2)
associated mesenchyme of glandular region (x2)
associated mesenchyme of gonadal component
associated mesenchyme of hindgut
associated mesenchyme of hindgut derived large intestine

associated mesenchyme of hindgut diverticulum
associated mesenchyme of jejunum (x2)
associated mesenchyme of labyrinth
associated mesenchyme of lateral semicircular canal
associated mesenchyme of left lung
associated mesenchyme of left lung rudiment
associated mesenchyme of lobar bronchus (x7)
associated mesenchyme of loop
associated mesenchyme of main bronchus
auricular region of left atrium
auricular region of right atrium
axial skeleton of skeleton (x2)
bare area of left
bare area of liver
bare area of right
basal cistern of subarachnoid space (x2)
basal plate of lateral wall (x3)
blood island of mesenchyme
blood island of mesoderm
body of hyoid bone (x2)
body of pancreas (x2)
brain of central nervous system
brain of venous system
branchial groove of 1st arch
branchial groove of 2nd arch
branchial groove of 3rd arch
branchial groove of 4th arch
branchial membrane of 1st arch
branchial membrane of 2nd arch
branchial membrane of 3rd arch
branchial membrane of 4th arch
branchial pouch of 1st arch
branchial pouch of 2nd arch
branchial pouch of 3rd arch
branchial pouch of 4th arch
bronchiole of lobar bronchus (x5)
bulbar cushion of bulbo-ventricular region
bulbar cushion of heart
bulbar ridge of bulbo-ventricular region
bulbar ridge of ventricle
capsule of cortex
capsule of lens
capsule of ovary
cardiac jelly of caudal half (x2)
cardiac jelly of common atrial chamber (x3)
cardiac jelly of early primitive heart tube
cardiac jelly of left part
cardiac jelly of outflow tract (x2)
cardiac jelly of primitive ventricle (x2)
cardiac jelly of right part
cardiac jelly of rostral half (x2)
cardiac muscle of auricular region (x2)
cardiac muscle of caudal half (x2)
cardiac muscle of common atrial chamber (x3)
cardiac muscle of early primitive heart tube
cardiac muscle of interventricular septum
cardiac muscle of left atrium
cardiac muscle of left auricular region
cardiac muscle of left part
cardiac muscle of left ventricle
cardiac muscle of primitive ventricle (x2)
cardiac muscle of right atrium
cardiac muscle of right auricular region
cardiac muscle of right part
cardiac muscle of right ventricle
cardiac muscle of rostral half (x2)

A.2 Remaining 59 Component Terms

alveolar duct
alveolar system
alveolus
anterior
associated mesenchyme
basal plate
body
cardiac jelly
cardiac muscle
cisterna chiasmatica
cochlear component
dental lamina
dental papilla
dermis
dorsal mesentery
dorsal pancreatic duct
enamel organ
endocardial tube
epidermis
epithelium
floor plate
floorplate
future spinal cord
glandular mucous membrane
head
lateral wall
left
lobe
lumen
mantle layer
marginal layer
mesenchyme
mesenchyme derived from head mesoderm
mesenchyme derived from neural crest
neural crest
neural fold
neural lumen
neural luminal occlusion
neural plate
neural tube
parenchyma
parietal
phalanx
phalanx cartilage condensation
phalanx pre-cartilage condensation
right
roof
roof plate
skeletal muscle
spinal cord
tail
telencephalic part of interventricular foramen
umbilical artery
umbilical vein
vascular element
ventricular layer
vestibular component
visceral
vitelline vein

Appendix B. Retrieved NPs from the Heart Chapter

B.1 Frequent Modifiers

Terms from the Heart chapter thought to be probably anatomical as each contains a modifier that is frequently encountered in the Mouse Atlas Nomenclature.

accessory hemiazygos
anterior veins
branchial arches
common atrium
common cardinal veins
common outflow channel
dorsal aortae
dorsal vein
dorsal wall
endocardial cushion region
endocardial heart
endocardial heart tubes
fibrous band
future ascending aorta
future pericardial region
future pulmonary trunk exit
future right ventricle
hepatic course
hepatic inferior vena cava
inferior caval stream
inferior vena cava
inferior venae cavae
inner endocardium
intermediate myocardial mantle tissue
internal iliac arteries
interventricular septation
lateral plate
lateral splanchnic folds
left 4th branchial arch artery
left 6th branchial arch artery
left and right anterior veins
left anterior cardinal vein
left arteries
left atria
left atrial chamber
left brachiocephalic
left common cardinal vein
left endocardial tubes
left forelimb
left heart
left posterior cardinal vein
left primitive cardiac tube
left subclavian artery
left superior vena cava
left superior venae cavae
left umbilical arteries
left umbilical vein
left upper limb
ligamentum teres
ligamentum venosum
lower limbs
median umbilical ligament
neural crest cells
neural innervation
posterior cardinal veins
posterior wall
primitive aortic sinuses
primitive atrium
primitive cardiac rudiment
primitive circulation
primitive foregut
primitive gut

primitive heart
primitive lungs
primitive pharynx
primitive tubular endothelial
primitive venous drainage
primitive venous system
primitive ventricule
primitive yolk sac
principal arterial supply
pulmonary circuit
pulmonary circulation
pulmonary spiral septum
right anterior cardinal veins
right aortic sinus
right arteries
right atria
right atrial wall
right atrio
right cavae
right common cardinal vein
right forelimb
right posterior cardinal vein
right sinus venosus
right subcardinal vein
right subclavian arteries
right superior vena cava
right umbilical veins
superior cavae
superior intercostal vein
superior vena cava
thoracic cavity
tunica media
umbilical cord
umbilical link
upper limb
vagus nerves
ventral midline
ventral part
ventral region
ventral surface
visceral afferent nerve fibres
visceral pericardial cells
visceral pleura

Appendix C. Perl Code

C.1 Sample Lineage Pattern Matching Code

```
open(PATHS,"paths-nostages");
open(EQUIV,"+>equivpaths");
while($path=<PATHS>) {
#search for patterns "future X.future Y" = "X.Y"
#                                         = "future X.Y"
#                                         = "X.future Y"
      if($path=~/^(.*)\.future (.*)\.future (.*)\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            $var3=$3;
            $var4=$4;
            open(COPY,"copypaths");
#The COPY file is an exact replica of the PATHS file and is opened and
#closed in each ‘‘if’’ statement so that all its contents are searched
#each time.  A separate copy is required as a file can not be compared
#to itself unless in an array - which would require additional
#variables and computation.
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\.$var3\.$var4\n/) {
                   print EQUIV "$path => $copy\n\n";}
               if($copy=~/^$var1\.$var2\.future $var3\.$var4\n/){
                   print EQUIV "$path => $copy\n\n";}}
               if($copy=~/^$var1\.future $var2\.$var3\.$var4\n/){
                   print EQUIV "$path => $copy\n\n";}
            close(COPY);}
#search for patterns "future X.Y" = "X.Y"
#                                  = "X.future Y"
      if($path=~/^(.*)\.future (.*)\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            $var3=$3;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\.$var3\n/) {
                   print EQUIV "$path => $copy\n\n";}
               if($copy=~/^$var1\.$var2\.future $var3\n/){
```

```
                print EQUIV "$path => $copy\n\n";}}
            close(COPY);}
#search for pattern "X.mesenchyme.Y" = "X.Y"
      if($path=~/^(.*)\.mesenchyme\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\n/) {
                   print EQUIV "$path => $copy\n\n";}
                     }
            close(COPY);}
#search for pattern "primitive X.Y" = "X.Y"
#                                   = "X.primitive Y"
      if($path=~/^(.*)\.primitive (.*)\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            $var3=$3;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\.$var3\n/) {
                   print EQUIV "$path => $copy\n\n";}
               if($copy=~/^$var1\.$var2\.primitive $var3\n/){
                   print EQUIV "$path => $copy\n\n";}}
            close(COPY);}
#search for pattern "primitive X.primitive Y" = "X.Y"
#                                             = "primitive X.Y"
#                                             = "X.primitive Y"
      if($path=~/^(.*)\.primitive (.*)\.primitive (.*)\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            $var3=$3;
            $var4=$4;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\.$var3\.$var4\n/) {
                   print EQUIV "$path => $copy\n\n";}
               if($copy=~/^$var1\.$var2\.primitive $var3\.$var4\n/){
                   print EQUIV "$path => $copy\n\n";}
               if($copy=~/^$var1\.primitive $var2\.$var3\.$var4\n/){
                   print EQUIV "$path => $copy\n\n";}}
            close(COPY);}
#search for pattern "proscencephalon.future X" = "future forebrain.X"
```

```
      if($path=~/^(.*)\.prosencephalon\.future (.*)\n/) {
            $var1=$1;
            $var2=$2;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.future forebrain\.$var2\n/) {
                   print EQUIV "$path => $copy\n\n";}
                }
            close(COPY);}
#search for pattern "X.foregut.Y" = "X.Y"
      if($path=~/^(.*)\.foregut\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\n/) {
                   print EQUIV "$path => $copy\n\n";}
                     }
                 close(COPY);}
#search for pattern "X.hindgut.Y" = "X.Y"
      if($path=~/^(.*)\.hindgut\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\n/) {
                   print EQUIV "$path => $copy\n\n";}
                     }
            close(COPY);}
#search for pattern "X.gland.Y" = "X.Y"
      if($path=~/^(.*)\.gland\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\n/) {
                   print EQUIV "$path => $copy\n\n";}
                     }
            close(COPY);}
#search for pattern "X component" = "X primordium"
#                                  = "X cavity"
#                                  = "X process"
      if($path=~/^(.*)\.(.*) component\.(.*)\n/) {
```

```
            $var1=$1;
            $var2=$2;
            $var3=$3;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2 primordium\.$var3\n/) {
                   print EQUIV "$path => $copy\n\n";}
               if($copy=~/^$var1\.$var2 cavity\.$var3\n/) {
                   print EQUIV "$path => $copy\n\n";}
               if($copy=~/^$var1\.$var2 process\.$var3\n/) {
                   print EQUIV "$path => $copy\n\n";}}
            close(COPY);}
#search for pattern "X.foregut-midgut junction.Y" = "X.Y"
      if($path=~/^(.*)\.foregut-midgut junction\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\n/) {
                   print EQUIV "$path => $copy\n\n";}
                     }
            close(COPY);}
#search for pattern "X.primitive heart tube.Y" = "X.Y"
      if($path=~/^(.*)\.primitive heart tube\.(.*)\n/) {
            $var1=$1;
            $var2=$2;
            open(COPY,"copypaths");
            while($copy=<COPY>) {
               if($copy=~/^$var1\.$var2\n/) {
                   print EQUIV "$path => $copy\n\n";}
                     }
            close(COPY);}
close(PATHS);
close(EQUIV);
```

C.2 Code to Tokenise Text

```
open(TEXT,"chapter");
open(FILE,"+>intermediate1");
while(<TEXT>) {

#put a space either side of any non-character e.g apostophe
        s/(\W)/ $1 /g;

#convert double quotes to single forward and back quotes
        s/\s{2,}"/ ''/g;
        s/^ "/ ''/;
        s/ "/ ``/g;

#replace control character ^M which occurred from email
#forwarding of text with newline
        s/\cM/\n/;

        print FILE;}
close(TEXT);

open(FILE1,"intermediate1");
open(FILE2,"+>intermediate2");
while(<FILE1>){

#replace any double spaces which occurred from the above
#processing with a single space
        s/  / /g;
        print FILE3;

}
close(FILE1);
close(FILE2);

open(FILE2,"intermediate2");
open(TOKENISED,"+>tokenisedfile");
while(<FILE2>){

#remove any blank lines that may have occurred from
#the above processing
        s/^ \n//;
        print TOKENISED;
```

```
}
close(FILE2);
close(TOKENISED);
```