Ontology Recapitulates Phylogeny: Design, Implementation and Potential for Usage of a Comparative Anatomy Information System

Ravensara S. Travillian

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2006

Program Authorized to Offer Degree: Medical Education and Biomedical Informatics

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Ravensara S. Travillian

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

Linda Shapiro
Linda G. Shapiro

Reading Committee:

Ira J. Kalet
Ira Kalet

Billie J Swalla
Billie Swalla

John Gennari

Linda Shapiro
Linda G. Shapiro

Date: August 18, 2006

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature____________________

Date August 18, 2006

University of Washington

Abstract

Ontology Recapitulates Phylogeny: Design, Implementation and Potential for Usage of a Comparative Anatomy Information System

Ravensara S. Travillian

Chair of the Supervisory Committee:
Professor Linda G. Shapiro
Computer Science and Engineering

Building on our previous design work in the development of the Structural Difference Method (SDM) for symbolically modeling anatomical similarities and differences across species, we describe the design and implementation of the associated comparative anatomy information system (CAIS) knowledge base and query interface, and provide scenarios from the literature for its use by research scientists. Our work includes several relevant informatics contributions. The first one is the application of the structural difference method (SDM), a formalism for symbolically representing anatomical similarities and differences across species. We also present the design of the structure of a mapping between the anatomical models of two different species, and its application to information about specific structures in humans, mice, and rats. The design of the internal syntax and semantics of the query language underlies the development of a working system that allows users to submit queries about the similarities and differences between mouse, rat, and human anatomy; delivers result sets that describe those similarities and differences in symbolic terms; and serves as a prototype for the extension of the knowledge base to any number of species. We also contributed to the expansion of the domain knowledge by identifying medically-relevant structural questions for humans, mice, and rats. Finally, we carried out a preliminary validation of the application and its content by means of user questionnaires, software testing, and other feedback.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

GLOSSARY

ADAPTATION: Change to a trait or a characteristic of an organism which gives it an advantage in surviving or functioning in a particular environment. Example: the loss of the upper incisors by the sloth bear *(Melursus ursinus)* is an *adaptation* that gives it an advantage in digging and sucking ants and termites out of fallen logs for food. new: In the evolutionary sense, some heritable feature of an individual's phenotype that improves its chances of survival and reproduction in the existing environment.

ANALOGY, *ADJ.* ANALOGOUS: *Similarity* of function between *anatomical structures* in different species. Example: the "torpedo" body shapes of the tuna, the penguin, and the dolphin all developed separately from each other, but perform the *analogous* function of reducing water resistance for increased speed and maneuverability underwater. new: Body part in different species that is similar in function but not in structure that evolved in response to a similar environmental challenge.

ANATOMICAL ENTITY: Biological entity, which constitutes the structural organization of a biological organism, or is an *attribute* of that organization. Examples: `Cell`, `Heart`, `Head`, `Peritoneal cavity`, `Apex of lung`, `Anatomical term`, `Sagittal plane`.

ANATOMICAL SET: Material physical *anatomical entity* which consists of the maximum number of discontinuous members of the same class. Examples: `Set of cranial nerves`, `Ventral branches of aorta`.

ANATOMICAL STRUCTURAL ABSTRACTION (ASA): A component of the *FMA* which describes the *partitive* and spatial relationships among the *anatomical entities* in the *AT*.

ANATOMICAL STRUCTURE: Material physical *anatomical entity* which has inherent 3D shape; is generated by coordinated expression of the organism's own structural genes; its parts are spatially related to one another in patterns determined by coordinated gene expression. Examples: `Heart`, `Right ventricle`, `Mitral valve`, `Myocardium`, `Endothelium`, `Lymphocyte`, `Fibroblast`, `Thorax`, `Cardiovascular system`, `Hemoglobin`, `T cell receptor`, `Gene`.

ANATOMICAL TAXONOMY (AT): A component of the *FMA* which specifies the taxonomic relationships of *anatomical entities* and assigns them to classes according to defining *attribute*s which they share with one another and by which they can be distinguished from one another. Example: the human prostate and heart share the defining *attribute* of being *organs*, and are distinguished from each other by the defining *attributes* that the prostate is a `Lobular organ`, while the heart is a `Cavitated organ`

ANATOMICAL TRANSFORMATION ABSTRACTION (ATA): A component of the *FMA* which describes the time-dependent morphological transformations of the entities represented in the *ontology* during the human life cycle. For example, vertebrate embryos of both sexes each start out with two different types of ducts, Müllerian (*syn.* paramesonephric duct) and Wolffian (*syn.* archinephric duct, mesonephric duct). The male embryo undergoes the following transformation: the Müllerian ducts regress, and the Wolffian ducts go on to form the ureter and vas deferens as the male reproductive system develops. The female embryo undergoes a different transformation: for the most part, the Wolffian ducts regress (although parts do go on to form the ureter), and the Müllerian ducts go on to form the uterine tube, the uterus, and the upper vaginal canal. The ATA would therefore contain *entities* for all of these *anatomical structures*, so that their appearance and disappearance over time could be modeled.

ANIMAL MODEL: Any animal which is studied for medical purposes as a surrogate for another species, usually (but not always) human. Subset of *biological model.* Example:

metastasis of prostate cancer is studied in the *rat model.*

ANTERIOR PROSTATE: Synonym for *coagulating gland*, a type of rodent prostate. Not to be confused with the *ventral prostate*, which is a different rodent prostate, nor with the anterior prostate in humans, which is a shortened term for the anterior lobe of the prostate. The term *coagulating gland* is preferred, and the term *anterior prostate* is deprecated, because of the possible confusion between "anterior" and "ventral" in human anatomy.

ASA: *See Anatomical Structural Abstraction.*

AT: *See Anatomical Taxonomy.*

ATA: *See Anatomical Transformation Abstraction.*

ATTRIBUTE: Property or characteristic which describes or limits a *node* of a graph. Represented as a slot in the frame-based *Protégé* representation of the *FMA*. Examples: *bounded-by*, *has-part.*

AVES: Birds.

BASAL: In phylogenetic terms, an earlier, "default" structure or organism, from which *derived* ones diverged. Synonym of *primitive.*

BAUPLAN: Shared structural *similarity* among different species or higher taxa, based on shared evolutionary history.

BIDIRECTIONAL: A property of a function or a relation in which it returns the same result, no matter in which direction its arguments are evaluated. Synonym of *symmetric.* Example: addition is *bidirectional*, because $a + b = b + a$.

BIJECTIVE MAPPING: A *mapping* which is both *injective* and *surjective.*

BIOLOGICAL MODEL: Any biological organism which is studied for medical purposes as a surrogate for another species, usually (but not always) human. Superset of *animal model.*

BIOLOGICAL SPECIES CONCEPT: Organisms are classified in the same species if they are potentially capable of interbreeding and producing fertile offspring.

BN: *See Boundary Network.*

BOUNDARY NETWORK (BN): A component of the *ASA* which describes the relationships among *anatomical entities* that bound each other or are bound by each other. Example: the `Anterior surface of the left ventricle of the heart` is bounded by the `Line of the interventricular sulcus`, the `Left margin of the heart`, and the `Line of the left interatrial sulcus`.

BREAST: Subdivision of the pectoral part of the chest which consists of the nipple, areola, fibroglandular mass of breast, superficial fascia, and skin of breast

CANONICAL (ABSTRACTION OF ANATOMY, PHENOTYPES, ETC.): A synthesis of generalizations based on qualitative observations, and sanctioned implicitly by accepted usage among domain experts. (Source: Rosse 1998)

CARNIVORE: A meat-eating animal, as opposed to herbivores (plant-eaters), insectivores (insect-eaters), etc.

CAVITATED ORGAN: `Organ` the unshared parts of which surround one or more macroscopic anatomical spaces. Examples: `Neuraxis`, `Tooth`, `Esophagus`, `Heart`, `Long bone`, `Corpus spongiosum of penis`.

CHORDATA, CHORDATE: An organism which possesses a notochord at some stage of its development; this group includes the *vertebrates.*

COAGULATING GLAND: A type of rodent prostate. Preferred synonym for the deprecated term *anterior prostate.*

COMPARATIVE ANATOMY: The study of *corresponding anatomical entities* in different species, at all levels of organization, in order to understand the significance of those *similarities* and *differences*, and their implications for organizing the derived medical information.

COMPARATIVE GENOMICS: The study of human genetics by comparisons with model organisms such as mice, the fruit fly, and the bacterium E. coli.

COMPARATIVE MEDICINE: A medical discipline in which the *similarities* and *differences* between different species in health and disease is studied.

COMPLETE: Of a theory: having the property that every sentence that is true in all interpretations is provable in the theory. If it is also *sound*, then truth and deduction are equivalent in that theory, with the attendant implications for reasoning in the context of a knowledge base such as the FMA.

CONCEPT: The "thought or reference" vertex of Ogden and Richards' "*meaning triangle*"—a component of meaning which is the mental image a real-world object (or *referent*) invokes. Example: The same *referent* bear may evoke the *concept* "livestock-killing pest" to one individual, "good and protective mother" to a second individual, "endangered species" to a third, and so forth.

CORRESPOND, *ADJ.* CORRESPONDING, *NOUN* CORRESPONDENCE: 1. Elements from two sets or graphs that are linked by a *mapping* are said to *correspond.* 2. *Anatomical entities* from different organisms that are linked by *homology* are said to *correspond.*

DEGENERACY: The ability of entities that are structurally different to perform the same function or yield the same output. Degeneracy is a ubiquitous biological property and a feature of complexity at genetic, cellular, system, and population levels. *Cf. redundancy.* (Source: Tononi 1999, Edelman 2001)

DEGENERATE: A limiting case in which a class of object changes its nature so as to belong to another, usually simpler, class. For example, the point is a degenerate case of the circle as the radius approaches 0, and the circle is a degenerate form of an ellipse as the eccentricity approaches 0. (Source: http://mathworld.wolfram.com/Degenerate.html, accessed 26 May 2006)

DERIVED: In phylogenetic terms, a later structure or organism, which diverged from the earlier *basal* ones.

DEVELOPMENT: The process whereby a single cell becomes a differentiated organism. The process of orderly change that an individual goes through in the formation of structure.

DEVELOPMENTAL BIOLOGY: The study of how an organism develops. Developmental biology includes embryology, but is a much broader discipline.

DIFFERENCE, *ADJ.* DIFFERENT: Absence or lack of *similarity.*

DIFFERENTIA, *PL.* DIFFERENTIAE: Defining *attributes* by which classes in a taxonomy can be distinguished from one another. Example: the human prostate and heart share the defining *attribute* of being `Organs`, and are distinguished from each other by the defining *attributes* that the prostate is a `Lobular organ`, while the heart is a `Cavitated organ`. `Organ` is the *genus* in this case, and `Cavitated` and `Lobular` are the *differentiae.*

DIMENSIONAL ONTOLOGY (DO): *DO* is a type hierarchy of geometric objects and shapes, in terms of which the three networks of the ASA may be described at an abstract level. Example: *has-dimension*, *dimension*, *has-shape*, *shape*, etc.

DO: *SEE DIMENSIONAL ONTOLOGY.*:

DORSOLATERAL PROSTATE (OR LOBE): A type of rodent prostate.

ECTOPIC: occurring in an abnormal location (*e.g.*, an ectopic kidney, pregnancy, or eye).

EDGE (RELATIONSHIP) DIFFERENCES: Differences in *edges* (relationships) in the graph derived from the *FMA*. Synonym of *relationship differences.*

EDGE ATTRIBUTE VALUE DIFFERENCES: Differences in the *attribute* values (*Protg slot contents*) of existing relationships between structures across species. For example, in many species of fish, one kidney (the "head kidney") migrates significantly closer to the head than does the other one. So the values in the slots of some of the spatial relationships (*e.g.*, what the head kidney is *adjacent-to*) would differ from the corresponding values for the other kidney in the same fish, as well as from the kidney in other *vertebrates* in which it does not migrate. The slot values (*i.e.*, of these spatial relationships are *edge attribute value differences.*

EDGE SET DIFFERENCES: Differences in the existence of relationships (*edges*) between *anatomical structures* across species. For example, some mammary glands of the mouse are adjacent to the inguinal ligament, which is not the case in the human. Therefore, there can be no *adjacent-to edge* between `Mammary gland` and `Inguinal ligament` in the human, and therefore this represents an *edge set difference.*

EDGE: Generally, a line connecting two *nodes* of a graph; more specifically, representing a relationship in the derived graph representation of the FMA.

ELASMOBRANCH: Cartilaginous fish such as sharks, rays, and skates, whose skeleton is *similar* in shape to other fishes, but is composed of cartilage rather than of bone.

EMBRYOLOGY: study of embryogenesis, the development of animals and plants from fertilization to birth/hatching.

EPITHELIUM, *PL.* EPITHELIA: The layer of tissue, sometimes secretory, which lines duct structures in the body.

EUTHERIAN MAMMAL: Placental mammal, as opposed to marsupials (such as the kangaroo, the koala, and the wombat) and monotremes (such as the platypus and the echidna). Examples: dog, cat, human, mouse, bear, whale.

EVO-DEVO: Study of the evolution of developmental processes. Study and examination of how changes in development can influence evolution.

EVOLUTIONARY BIOLOGY: *See also Systematics.* A biological discipline in which the evolutionary relationships of organisms across time are studied.

EVOLUTIONARY DEVELOPMENTAL BIOLOGY (EVO-DEVO): Study of the evolution of developmental processes. Study and examination of how changes in development can influence evolution.

FIRST-CLASS OBJECT: An object that can be manipulated by a computer program or mathematical operations.

FOUNDATIONAL MODEL OF ANATOMY (FMA): An *ontology* which furnishes a comprehensive set of *entities* and relationships which describe the body at all levels of structural organization.

GENOTYPE: Underlying genetic makeup of an organism. new: The specific allelic composition of a cell, either of the entire cell or more commonly for a certain gene or a set of genes. The genes that an organism possesses.

GENUS, *PL.* GENERA: Definition 1, Definition 2. Defining *attributes* which classes in a taxonomy share with one another. Example: the human prostate and heart share the defining *attribute* of being *organs*, and are distinguished from each other by the defining *attributes* that the prostate is a

BASAL: In phylogenetic terms, an earlier, "default" structure or organism, from which *derived* ones diverged. Synonym of *primitive.*

`Lobular organ`, while the heart is a `Cavitated organ`. `Organ` is the *genus* in this case, and `Cavitated` and `Lobular` are the *differentiae*

GRAPH DISTANCE: Measurement of the *difference* or *similarity* between graphs.

GRAPH IDENTITY: *See Graph isomorphism.*

GRAPH ISOMORPHISM: *See also Isomorphism, Set isomorphism.* The relationship between two graphs whose *nodes* are in a *one-to-one* and *onto* correspondence, and whose *edges* are in a *one-to-one* and *onto* correspondence as well.

GROSS: Visible to the naked eye (*i.e.*, unassisted by a microscope) (of an anatomical structure).

HAS-MEMBER: The *partitive* relationship between an *anatomical set* and the classes which constitute it.

HAS-PART: The relationship from a class to its constituents.

HERBIVORE: A plant-eating animal, as opposed to carnivores (meat-eaters), insectivores (insect-eaters), etc.

HETEROLOGS: Heterologs differ in both origin and activity. Genes that are "unique" in activity and sequence are said to be heterologous.

HISTOGENESIS: Origin of tissues.

HISTOLOGICAL: Of or pertaining to the tissue level of anatomical organization.

HOLOCEPHALAN: Member of a subset of *elasmobranchs* (cartilaginous fishes) distinguished by the shape of their head, which tapers off into a long tail. The only living holocephalans are the chimaeras, or ratfish (*Hydrolagus colliei*).

HOMOLOGS: Homologs have common origins but may or may not have common activity. Genes that share an arbitrary threshold level of similarity determined by alignment of matching bases are termed homologous. Homology is a qualitative term that describes a relationship between genes and is based upon the quantitative similarity. Similarity is a quantitative term that defines the degree of sequence match between two compared sequences. Homology implies that the compared sequences diverged in evolution from a common origin. Homologous sequences are termed homologs and this term may be applied to both genes and proteins. Homologs look similar to each other and appear to share common ancestry but they may or may not display the same activity.

HOMOLOGY, *ADJ.* HOMOLOGOUS: Similarity of ancestral origin between *anatomical structures* in different species, or the relationship between two *anatomical structures* which can be traced back in time to the same structure, or clearly-related structures, in a common ancestor. Homology can be *serial* (such as the homology between a vertebrate's thoracic and lumbar vertebrae), *sexual* (such as the homology between the female ovary and the male testis), or *taxic* (across species or other phylogenetic groups). Examples: human and mouse hearts, mammalian ear ossicles and reptile jaw bones. new: Similarity in DNA or protein sequences between individuals of the same species or among different species.

HOMOMORPHISM, *ADJ.* HOMOMORPHIC: The relationship between two graphs which contain a structure-preserving partial *mapping.*

HOMOPLASY, *ADJ.* HOMOPLASTIC: *Similarity* of appearance between *anatomical structures* in different species. Example: ichthyosaurs (sea-going dinosaurs) look very much like dolphins, even though they are as extinct reptiles only very distantly related to contemporary highly-specialized mammals. Their body shape is *homoplastic* to that of the dolphins.

IDENTICAL: Perfect similarity or *isomorphism.*

INDOLENT: Slow to spread (describing cancer).

INHERITANCE (SUBSUMPTION) HIERARCHY: The organization of classes into superclasses and subclasses. The relationship from superclass to subclass is *subsumes*; the relationship from subclass to superclass is *is-a.*

INJECTIVE MAPPING: A *mapping* from set A to set B so that every element of A is mapped to a unique element of B. Synonym of *one-to-one.*

IS-A: The relationship from subclasses to superclasses in an *inheritance hierarchy.*

ISOMORPHISM, *ADJ.* ISOMORPHIC: The similarity relationship between two structures whose *mapping* at the level of organization under study is *one-to-one* and *onto.* For the *SDM*, the term *isomorphism* always implies *graph isomorphism.*

LACTIFEROUS DUCT TREE: The part of the *lactiferous gland* which consists of a wall and a lumen, and branches into smaller subtrees, terminating in the lactiferous acini. Example: There are variable numbers of *lactiferous duct trees* opening on to each nipple in different species.

LACTIFEROUS GLAND: `Lobular organ` which consists of a `Lactiferous duct tree` and the `Set of lactiferous acini` that are connected to the duct tree

LEAST-ERROR MATCHING: The method developed by Shapiro and Haralick which yields the best match (*relational distance*, smallest graph distance) between two graphs

LOBULAR ORGAN: `Parenchymatous organ` the `Stroma` of which subdivides the `Parenchyma` into `Lobes`, `Segments`, `Lobules`, and `Acini`. Examples: `Lung`, `Liver`, `Lactiferous gland`, `Testis`.

MAMMARY GLAND, *SYN.* SET OF LACTIFEROUS GLANDS: *Anatomical set* which consists of all the *lactiferous glands* of one breast. Examples: There are only two canonical instances, right *mammary gland* and left *mammary gland.* Note that in the human, there are multiple *lactiferous duct trees* per *mammary gland*; this is not true for species such as the mouse, which has one *lactiferous duct tree* per *mammary gland.*

MAPPING: For two different sets of parts representing comparable structures in two different species, a *mapping* is a specification of the *correspondences.*

MEANING TRIANGLE, *SYN.* SEMANTIC TRIANGLE, SEMIOTIC TRIANGLE: A representation of the elements of meaning (*concept*, *symbol*, and *referent*) and their interaction in contributing to overall meaning

METAKNOWLEDGE (MK): A component of the *FMA* which comprises the principles and sets of rules according to which the relationships are represented in the model's other three component abstractions. Example: an *anatomical entity* which *bounds* another *anatomical entity* always possesses one dimension less than the *anatomical entity* it bounds. The anterior surface of the left ventricle of the heart is a plane (2-dimensional), and each of the lines bounding it (the line of the interventricular sulcus, the left margin of the heart, and the line of the left interatrial sulcus) is 1-dimensional.

METAMODEL: A representation of a *model*, including rules for modifying the model represented. Example: the mouse *FMA* is a *model*, while the rodent *FMA* is a *metamodel* containing rules (*metaknowledge* such as “rodents’ teeth grow continuously throughout the animal’s lifespan”). These rules apply to the mouse *model*, the rat *model*, the hamster *model*, and so forth, all of which, along with the appropriate rules, constitute the rodent *metamodel.*

MK: *See Metaknowledge.*

MODEL: Definition 1, Definition 2. A simplified representation of a real-world object. Example: the *FMA* and its derived graph are *models* of *anatomical structures.* 2. A biological organism which is studied as a surrogate for another biological organism. Example: vertebrate embryology is often studied in the zebrafish *model.*

MONOTONIC INHERITANCE: A form of inheritance of *attributes* from superclasses to subclasses in an *inheritance hierarchy* where the *attributes* from the superclass are directly inherited by the subclass. Additional *attributes* can be acquired by the subclass, but they cannot cancel the *attributes* inherited from the superclass. Example: as a class, `Mammals` have fur (an *attribute* represented here by the notation *have fur?* $= T$). `Ferrets` are `Mammals`, and they inherit the *have-fur?* $= T$ *attribute* value from their superclass `Mammals`. Therefore, `Ferrets` inherit fur monotonically from `Mammals`. Additionally, `Ferrets` have `Musk glands`, which they did not inherit from the superclass `Mammals`, but because the glands do not cancel any `Mammal` *attributes*, they do not change the `Ferrets'` *monotonic inheritance* of *have fur?* $= T$.

MORPHOLOGICAL SPECIES CONCEPT: Classification of organisms as being in the same species if they appear identical by morphological (anatomical) criteria.

MORPHOLOGY: The study of the interaction of anatomical form and function. Example: the “torpedo” form of the tuna’s body serves the function of reducing water resistance for long-distance swimming.

MURINE: Adjectival form of "mouse": of or pertaining to or describing a mouse. Example: *murine* models of cancer.

NATURAL SELECTION: The process in nature whereby one genotype leaves more offspring than another genotype because of superior life history attributes (fitness) such as survival or fecundity.

NODE (STRUCTURE) DIFFERENCES: *Differences* between the *nodes* in the derived *FMA* graphs representing *anatomical entities* in the source species and the *corresponding* entities in the target species, reflecting nonexistence or a different distribution of existence of *homologous anatomical entities* across species. Synonym of *structure differences.*

NODE ATTRIBUTE DIFFERENCES: *Differences* in the existence of an *attribute* between two *corresponding* structures in the source and target species in other words, the structure exists in each species, but it occupies a different place in the AT, and thus, the slots required for a *sound* and *complete* description of the structure—its *attributes*—differ across species. For example, *has-member* (which is a specialization of the *partonomic* relationship constrained in the FMA to `Anatomical sets`) is an *attribute* of the node `Set of mouse prostates`. In this partonomic scheme, `Anatomical set` is made up of member `Organs`. In the human, the `Prostate` is a single *Organ.* The class `Organ`, however, lacks the *attribute has-member*, and therefore a *node attribute difference* exists between the `Prostates` of the two species.

NODE ATTRIBUTE VALUE DIFFERENCES: *Differences* in values of *corresponding attributes* shared between *corresponding nodes* of two species in other words, the structure exists in both species, but there is some *difference* in the values of its *attributes* from one species to the other. For example, an *isomorphism* exists between the mouse (and rat) and human `Stomachs` at the levels of whole `Organ` and `Organ part`. The *difference* between mouse and human emerges in the attribute values for the node `Mucosa`, which is only *glandular* for the human, but *glandular* and *non-glandular* for the mouse.

NODE SET DIFFERENCES: *Differences* between the number of *nodes* in the derived graph representing *anatomical entities* in the source species and the *corresponding* entities in the target species, reflecting nonexistence or a different distribution of existence of *homologous anatomical entities* across species. *Node set differences* may be 1-null, null-l, 1-to-many, many-to-one, or many-to-many. Example: the human `Prostate` organ maps to different mouse `Prostate` *organs*: the `Ventral prostate`, the `Right` and `Left coagulating glands`, and the `Right` and `Left dorsolateral prostates`.

NODE: Generally, a point on a graph connected to other points on the graph by one or more *edges*; more specifically, representing an *anatomical entity* in the derived graph representation of the *FMA*. Synonym of *vertex.*

NON-MONOTONIC INHERITANCE: A form of inheritance of *attributes* from superclasses to subclasses in an *inheritance hierarchy* where the *attributes* from the superclass can be cancelled or overridden by another value in the subclasses. Example: as a class, `Mammals` do not fly (an *attribute* value represented here by the notation *does-fly?* = *F*). However, `Bats` (a subclass of `Mammals`) do fly, so they have overridden the FALSE value of the *does-fly? attribute* they inherited from the superclass `Mammals`. Yet, `Baby bats` (a subclass of `Bats`) override the TRUE value of the *does-fly? attribute* they inherited from their superclass `Bats`, as `Baby bats` do not fly. Therefore, `Bats` inherit non-monotonically from `Mammals`, and `Baby bats` inherit non-monotonically from `Bats`.

ONE-TO-ONE MAPPING: A *mapping* from set A to set B so that every element of A is mapped to a unique element of B. Synonym of *injective.*

ONTO MAPPING: A *mapping* from set A to set B so that every element of B is mapped to a unique element of A. Synonym of *surjective.*

ONTOGENETIC: Pertaining to the development of an individual, as distinguished from *phylogenetic*, or pertaining to the development of a related group.

ONTOLOGY: An explicit specification of a *conceptualization*

ORGAN COMPONENT: Organ part, which has a definable shape, bounded predominantly by *bonafide* boundaries and is countable. Examples: Lobe of lung, Osteon, Acinus, Submucosa, Anterior leaflet of mitral valve, Capsule of kidney, Cortical bone, Muscle fasciculus.

ORGAN PART: Anatomical structure, which consists of two or more types of tissues that form a defined structural aggregate in an Organ. Examples: Osteon, Cortical bone, Neck of femur, Bronchopulmonary segment, Left lobe of liver, Anterior right side of heart, Interventricular branch of left coronary artery, Right atrium, Mitral valve, Head of pancreas.

ORGAN: *Anatomical structure*, which consists of the maximal set of *organ parts* so connected to one another that together they constitute a unit of macroscopic anatomy, structurally distinct from other such units. Examples: Femur, Biceps, Liver, Heart, Skin, Tracheobronchial tree, Sciatic nerve, Ovary.

ORTHOLOGS: Orthologs are homologs produced by speciation. When speciation follows duplication and one homolog sorts with one species and the other copy the other species, subsequent divergence of the duplicated sequence is associated with one or the other species. Such species specific homologs are termed orthologous. Thus, orthologs are homologs from duplication that precedes speciation, followed by divergence of sequence but not activity in separate species. Orthologs have homologous origin and homologous activity. (Source: Fitch 1970, Popovici 2001)

PARALOGS: Paralogs are homologs produced by gene duplication. Homologous genes produced by gene duplication are termed paralogous. Paralogous genes are homologous genes that result from divergent evolution from a common ancestral gene. Paralogous implies that gene duplication and divergence occurred within the same organism/species and divergence of sequence led to divergence of activity. Paralogs have homologous origin but heterologous activities. (Source: Fitch 1970, Popovici 2001)

PART-OF: The relationship from the constituents of a class to the class itself.

PART-OF NETWORK (PN): A network which consists of a number of subnets describing *partonomies* between different classes of *anatomical entities*

PARTONOMY, *ADJ.* PARTONOMIC, PARTITIVE: A hierarchy which describes the relationships between classes and their constituents (as opposed to their subclasses).

PHENOTYPE: The observable attributes of an organism.

PHENOTYPE: (1) The form taken by some character (or group of characters) in a specific individual. (2) The detectable outward manifestations of a specific genotype. (3) The observable attributes of an organism.

PHYLOGENETIC: Pertaining to the development of a related group, as distinguished from *ontogenetic*, or pertaining to the development of an individual.

PN: *See Part-of network.*

PRIMITIVE: In phylogenetic terms, an earlier, "default" structure or organism, from which *derived* ones diverged. Synonym of *basal.*

PROCYONID: Related to or pertaining to raccoons.

PROTÉGÉ: A frame-based development environment for knowledge-based systems

REDUNDANCY: The ability of entities that are structurally identical to perform the same function or yield the same output. See *redundancy.* (Source: Tononi 1999, Edelman 2001)

REFERENT: The "*referent*" vertex of Ogden and Richards "*meaning triangle*"—a component of meaning which is the real-world object referred to. Example: The individual of any of the various *ursid* species which is referred to by the English *symbol* "bear", the French *symbol* "ours", the Greek *symbol* "*άρκτος*", the Navajo *symbol* "shash", and so forth.

RELATED: Structures are said to be *related* when there exists an evolutionary inheritance relationship between the structures being compared.

RELATIONAL CONSTRAINT: In this context, the requirement of a *graph isomorphism* that all edges, as well as all *nodes*, must be *one-to-one* and *onto* between the two compared graphs.

RELATIONAL DISTANCE: The best match (smallest *graph distance*) between two graphs, derived via the *least-error matching* method of Shapiro and Haralick

RELATIONAL HOMOMORPHISM: A structure-preserving function that maps the *nodes* of one graph to those of a second graph in a way that preserves the interrelationships among the *nodes.*

RELATIONSHIP DIFFERENCES: *Differences* in *edges* or relationships in the graph derived from the *FMA*. Synonym of *edge differences.*

RUMINANT: A sub-order of mammals that chew the cud (*i.e.*, *rumination*), and have a complex stomach and an even number of toes. Examples: cow, llama.

RUMINATION: Regurgitation and chewing of previously-swallowed food.

SAN: *See Spatial Association Network.*

SDM: Syn. of Structural Difference Method. Definition.

SELECTION: Differential survival and expression (or endurance) of traits which are better suited for a given environment by conferring some advantage on an organism.

SEMANTIC TRIANGLE, *SYN.* MEANING TRIANGLE, SEMIOTIC TRIANGLE: Definition.

SEMIOTIC TRIANGLE, *SYN.* MEANING TRIANGLE, SEMANTIC TRIANGLE: A representation of the elements of meaning (*concept*, *symbol*, and *referent*) and their interaction in contributing to overall meaning

SERIAL HOMOLOGY: *Homology* among consecutive *similar anatomical structures* in an organism. Example: the *homology* among a vertebrate's thoracic and lumbar vertebrae.

SET ISOMORPHISM: The relationship between two sets whose members are in a *one-to-one* and *onto* correspondence.

SEXUAL HOMOLOGY: *Homology* among consecutive *similar anatomical structures* between male and female organisms. Example: ovary and gonad.

SIMILARITY, *ADJ.* SIMILAR: The *concept* that objects under comparison resemble each other in some way, usually (but not always) visual.

SNOMED: Systematized NOmenclature of MEDicine, an initiative by the College of American Pathologists to systematically integrate medical terminologies

SOMITE: An embryological *anatomical structure* that arises from a germ layer and later develops into a segmented structure in the adult. Example: each vertebra develops from the caudal half of the preceding *somite*, plus the cranial half of the following *somite.*

SOUND: Of a theory: having the property that every provable sentence is true in all interpretations. If it is also *complete*, then truth and deduction are equivalent in that theory, with the attendant implications for reasoning in the context of a knowledge base such as the *FMA*

SPATIAL ASSOCIATION NETWORK (SAN): A network which consists of a number of subnets describing location, orientation, and connectivity relations between different classes of *anatomical entities*

STRUCTURE DIFFERENCES: Synonym of *node differences.*

SUB-GRAPH ISOMORPHISM: The relationship between two graphs, one of which contains a sub-graph which is *isomorphic* to the entire other graph.

SUBSUMES: The relationship from superclasses to subclasses in an *inheritance hierarchy.*

SUBSUMPTION HIERARCHY: *See Inheritance hierarchy.*

SURJECTIVE MAPPING: A *mapping* from set A to set B so that every element of B is mapped to a unique element of A. Synonym of *onto.*

SYMBOL, *ADJ.* SYMBOLIC, *ADV.* SYMBOLICALLY: The "*symbol*" vertex of Ogden and Richards "*meaning triangle*"–a component of meaning which is the written or verbal or signed string a real-world object invokes. Example: A *referent* individual of any of the various *ursid* species is referred to by the English *symbol* "bear", the French *symbol* "ours", the Greek *symbol* "ἀρκτος", the Navajo *symbol* "shash", and so forth. Synonym of *term*

SYMBOLIC MODEL: A *model* made up of symbolic (i.e., textual) information.

SYMMETRIC: A property of a function or a relation in which it returns the same result, no matter in which direction its arguments are evaluated. Synonym of *bidirectional.* Example: addition is *symmetric*, because $a + b = b + a$.

SYSTEMATICS: *See also Evolutionary biology.* A biological discipline in which the study of organisms and their evolutionary relationships to each other is used for the purpose of description and classification of those organisms.

TAXIC HOMOLOGY: *Homology* or relatedness of structures across taxa, or *phylogenetic* groups. Example: rat lung and human lung.

TERM: The "*symbol*" vertex of Ogden and Richards "*meaning triangle*" a component of meaning which is the written or verbal or signed string a real-world object invokes. Example: A *referent* individual of any of the various *ursid* species is referred to by the English *term* "bear", the French *term* "ours", the Greek *term* "*άρκτος*", the Navajo *term* "shash", and so forth. Synonym of *symbol*

THEILER STAGE: A developmental stage in the embryonic mouse, after the staging system published by Karl Theiler

TRANSITIVITY, *ADJ.* TRANSITIVE: A property of relations or functions in which if a relation exists between the first and second element, and between the second and third element, then the same relation exists between the first and the third element. Example: the *is-a* relationship is transitive, and so from the propositions "Baby bats are bats" and "Bats are mammals", it can be deduced that "Baby bats are mammals"'.

ULTRASTRUCTURAL: Pertaining to the smallest (sub-microscopic) structural elements of a cell.

URSID: Related to or pertaining to bears.

VENTRAL PROSTATE (OR LOBE): A type of rodent prostate. Not to be confused with *anterior prostate*, which is a deprecated *term* for a different rodent prostate, the *coagulating gland.*

VERTEBRATE: An animal whose nerve cord is surrounded by a backbone. The main groups of vertebrate animals are the fishes, amphibians, reptiles, birds, and mammals.

VERTEX: Generally, a point on a graph connected to other points on the graph by one or more *edges*: more specifically, representing an *anatomical entity* in the derived graph representation of the FMA. Synonym of *node.*

VESTIGIAL: The trait in an *anatomical structure* of being significantly smaller or incompletely developed when compared to the corresponding structure in another species, or at another stage of individual development.

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to the following individuals:

To Linda Shapiro, my advisor and mentor, for all her academic and professional guidance over the years.

To Ira Kalet, Peter Tarczy-Hornoch, Sherrilynne Fuller, David Masuda, David Chou, and Fred Wolf for giving me the opportunity to pursue my right livelihood in medical informatics.

To John Gennari and Robin McEntire, for thoughtful and provocative questions.

To Cornelius Rosse, for teaching me how to think about anatomy and informatics.

To my biology teachers Billie Swalla, Karen Petersen, John Wingfield, and Marilyn Ramenofsky, for imparting the joy and rigor of learning about animals and their adaptations.

To Onard Mejino, August Agoncillo, Franz Calvo, Todd Detwiler, Richard Martin, Dan Cook, Franz Calvo, and Kurt Rickard, for the SIG lab and iLab discussions from which I have greatly benefitted.

To Kremena Diatchka, Tejinder K. Judge, and Katarzyna Wilamowska, for their programming skills and their design suggestions.

To Robert Cardiff, Geoffrey Cunha, Anne Donjacour, Claire Yang, Robert Vessella, and Janna Quinn, for their time and patience in teaching me about rodent anatomy.

To Cheryl Frederick, Florence Patten, Edwina, Suntil, Marcella, Ting, Honey Bear, Bandau, Juwita, and Scruffy, for giving me the opportunity to put theory into practice in a real-life, intensive biology application.

To Laurel Rees, Tristan Burch, Marie Goines, and Jeremiah Jester, for making this dissertation possible by keeping the machines up and running, despite my deadlines.

To PZ Myers and John Wilkins, for so very many vital and exciting ideas and their detailed explication.

To Georges Moustaki, Leonard Cohen, and Johnny Cash, for countless hours of companionship in developing and writing up this research.

To the University of Washington University Initiatives Fund, the National Library of Medicine, and the Student Technology Fee grants for the funding which made this work possible.

To the experimental animals.

And, finally, to Iain, for his patience, tolerance, good humor, and support.

Chapter 1

INTRODUCTION

A comparative anatomy information system is a computer system that allows users to compare *canonical*[1] *phenotypes* of *corresponding* (*i.e.*, *homologous*) anatomical structures across medically-relevant species at varying levels of detail, and returns responses to queries about those comparisons. The need for such a system is due to the importance of *animal models* in *comparative medicine and genomics*, as well as out of the explosion in the quantity of data to be managed. The system we have developed is an initial attempt to address some of the informatics issues involved in meeting these needs.

This dissertation describes the design, implementation, and potential use of a comparative anatomy information system (CAIS). This system is based on the structural difference method (SDM) formalism for symbolically representing the similarities and differences between homologous anatomical structures across different species [125]. The anatomical structures of the species to be compared, as well as the mappings between species, are modeled on templates from the Foundational Model of Anatomy (FMA) knowledge base and implemented in the Protégé-2000 ontology editor and knowledge-based framework [46]. A graphical user interface (GUI) allows users to issue queries that retrieve information concerning the similarities and differences between the species being examined. Queries from diverse information sources, including domain experts, peer-reviewed articles, and reference books, have been used to test the system and to illustrate its potential use in comparative anatomy studies.

[1] Words defined in the glossary are italicized the first time they appear in the text.

1.1 Background and Significance

The amount of anatomical and associated medical information emerging from animal modeling in comparative medicine and comparative genomics is increasing at an exponential rate ([1], [8], [10], [13], [21], [30], [31], [33], [52], [56], [57], [62], [65], [67], [76], [79], [93], [127]). Consequently, innovative techniques in evaluating, organizing, and managing that information for researchers and clinicians are imperative to develop. The increasing need for extrapolating information from one species to another has been highlighted by contemporary research in bioinformatics, genomics, proteomics, and animal models of human disease, as well as other fields [125]. Additionally, the urgency of finding ways to organize and manage the volume of data has been remarked upon by many observers, especially in light of the identification and characterization of genomic sequences across species [35]. Information systems have been and continue to be an important tool in this task.

At the same time that the amount of information generated is increasing so rapidly, traditional barriers between scientific domains are being blurred. As medical research becomes more interdisciplinary, researchers from traditional biomedical disciplines (anatomy, embryology, etc.) join forces with scientists from newer disciplines (*e.g.*, molecular biology, genomics) and clinicians in the attempt to translate the discoveries from bench science into clinical applications that can realize effective treatments for patients. Additionally, new players are becoming involved in the increasing need to answer what Ceusters terms "medically-relevant questions" [18], because these discoveries have improved and expanded options for medical care [120]. Accordingly, the audience for information has expanded to include, among others, patients and policy makers [89], and information systems dealing with this type of data must be flexible enough to accommodate the various needs of these different groups of users. Therefore, in addition to rigorous attention to the quality of the anatomical information involved, such a system must be flexible and extensible enough to accommodate different information views, depending on the needs of the user, whether a bench scientist, a clinician, a student, or a patient.

In contrast to the vast amount of theoretical knowledge about mechanisms of disease gained as a result of studying animal models, the low rate of success in translating this knowledge into treatments for cancer [85], Parkinson's disease [69], and other conditions stands as a major disappointment. One of the reasons is that animal models, in which these conditions are studied, are similar to, but not identical to, humans, and understanding the anatomical differences involved is crucial to understanding how much the knowledge from animal models can be turned into translational treatments.

Therefore, in addition to the amount of information being generated, and the information needs of different users, the quality of that information is also an issue that must be addressed. A fundamental principle of animal modeling of human disease is that animal models, while not identical to humans, are *genotypically* similar enough that the results from animal experimentation can be leveraged into applications in human medicine. This principle rests on two assumptions: first, that the differences in phenotypes among the species involved is not as medically relevant as the degree of similarity of the genotypes [11], and second, that the correspondences among those different phenotypes are well-enough understood for the principled application of those findings.

However, these assumptions need to be examined to determine their validity. Certainly in the case of the second assumption, there is reason to believe that the lack of standardization of anatomical knowledge and histopathological preparation techniques has had an impact on the validity of experimental results from animal models. Suwa states in so many words that "Because sampling of the paired lobes (ventral, dorsal, lateral, and anterior) of the mouse prostate has often been inconsistent, comparisons among different investigations have lacked validity. The absence of site identification for prostatic lesions has made reported incidences relatively nonspecific" ([117], [118]). Providing a robust reference ontology [103] as an anatomical baseline standard against which protocols such as Suwa's proposed techniques could be measured can only improve the quality of the information gained from these experiments.

To this end, we agree with Ceusters' assertion: he proposes the development of formal logical and linguistic tools for the development and quality-assurance process, both for these large terminologies [19], and for the relationships developed among those terms for the ontology. This is consistent with Spackman, as well, who asserts that applying these principles and constraints makes the ontology more consistent and useful [113]. In this way—

by formally and rigorously defining the principles on which our ontology is founded, and on which correspondences are assigned—we can address the content of those correspondences, as well as the possibility of progressive conceptual changes in those entities in light of new, more reliable, data from experiments [16].

The cross-species model proposed in this dissertation provides a formalized ontological framework for the analysis of structural phenotype comparison, with the foundational principles as well as application of these principles to real-life queries on animal models. Such an information system will support formal reasoning about the comparisons of the structural phenotypes involved [106], and provides a structure on which the quantity of information involved can be organized. The possibility of establishing and validating structural correspondences between different structural phenotypes has tremendous potential for addressing both issues by improving the quality, management, and dissemination of information about animal models of human disease and comparative genomics.

Due to their medical importance, a subset of the cancer sites identified by the Mouse Models of Human Cancer Consortium (MMHCC) comprised the primary subject matter for the development of the information system. For example, prostate cancer alone is responsible for approximately 200,000 new diagnoses, approximately 40,000 deaths, and expenditures in excess of $13B each year in the United States [110], and breast cancer exacts a comparable toll [22]. Those are only two of the site cancers for which the MMHCC concentrates on developing mouse models; the other sites are gastrointestinal tract (gastrointestinal cancer); blood and lymphoid tissues (hematopoietic cancer); lung (lung cancer); brain and spinal cord (nervous system cancer); ovary (ovarian cancer); skin (skin cancer and melanoma); uterus, cervix, and vaginal vault (cervical and gynecological cancer); mouth and nasal cavity (oral cancer); fat, blood vessels, nerves, bones, muscles, deep skin tissues, and cartilage (sarcoma) [86].

We selected five of these sites (prostate, breast/mammary gland, lung, ovary, and cervix) to model for our information system. We built on our foundational work in rodent mammary gland and prostate symbolic model development and comparison [125] to continue development of rodent anatomical models, including leveraging the work on mouse structures as templates for the corresponding rat structures with particular attention to the documented similarities and differences between the two rodent species. Our research design involved close collaboration with colleagues in biological structure and structural informatics, computer science, and comparative vertebrate embryology, who contributed domain content, assisted in development of the system, and evaluated its usefulness and accuracy.

In addition to organizing and managing information on the comparative anatomy of different structural phenotypes across species, the proposed information system will serve as a resource for improving the quality of available structural information by clarifying ambiguities and establishing an anatomical baseline for comparison and correlation. By developing the mouse and rat models on this small scale, we hope to not only provide a resource that will be useful for diverse groups of users, but also to provide a methodology that will create an incentive for domain experts in other laboratory animals to contribute content. We hypothesize that the development of these robust models will eventually pave the way for meta-model development, in which not only the data about the species under consideration is included, but also the rules, principles, methods, and axioms underlying those species models can be incorporated.

1.2 Contributions

In this dissertation, we describe a comparative anatomy information system for querying similarities and differences across species, the knowledge base it operates upon, the method it uses for determining the answer to the queries, and the user interface it employs to present the results. The relevant informatics contributions of our work include:

- the application of the structural difference method (SDM), a formalism for symbolically representing anatomical similarities and differences across species;

- the design of the structure of a mapping between the anatomical models of two different species, and its application to information about specific structures in humans, mice, and rats;

- the design of the internal syntax and semantics of the query language;

- the development of a working system that:
  - allows users to submit queries about the similarities and differences between mouse, rat, and human anatomy;
  - delivers result sets that describe those similarities and differences in symbolic terms;
  - serves as a prototype for the extension of the knowledge base to any number of species;
- the expansion of the domain knowledge by identifying medically-relevant structural questions for the human, the mouse, and the rat;
- the validation of the application and its content by means of user questionnaires, software testing, and other feedback.

1.3 Outline of this Dissertation

In this dissertation, the problem of comparing anatomical structures across species is outlined, an approach to symbolically representing similarity and differences in corresponding structures is developed, and the design and implementation of a system based on that approach is described. It is organized in the following way:

- *Chapter 1: Introduction*—an overview of the background and significance of the problem we address, and the contributions of our work;
- *Chapter 2: Related Literature*—a review of the background to our proposed system, previous work in the area (including more detail on the FMA, set and graph matching, model matching, and comparative anatomy), and comparison to our system;
- *Chapter 3: Comparative Anatomy and the Structural Difference Method*—a description of our method for symbolically describing and classifying the similarities and differences between anatomical structures across species, and a description of how this meets the information needs of different types of users of the system;

- *Chapter 4: Design of the Comparative Anatomy Information System (CAIS)*—a description of the design of anatomical mappings and other design considerations, as well as implementation of the knowledge base in Protégé-2000;
- *Chapter 5: Interface and Sample Queries*—a description of the system's interface, and a detailed discussion of the components and their significance to the user, with examples of queries that can be executed by the system;
- *Chapter 6: Data and Results*—a review of the selection of the data and the methods used to acquire it, and the results of a set of queries representative of real-world comparative anatomy problems;
- *Chapter 7: Putting the Biology in Bioinformatics: Conclusions and Future Work*—a summary of our completed work and its contributions and a preview of future work.
- *Glossary*—definitions of the significant terms used in our work;
- *Appendix A*—the questionnaire we sent to comparative anatomy domain experts;
- *Appendix B*—the domain experts' responses to the questionnaire.

1.4 Conventions and Notations

The first significant appearance of a term that is defined in the Glossary is indicated by *italics*: "*Coagulating gland* is the preferred synonym for *anterior prostate* in rodents."

Names of slots are also in *italics*: "The `Line of the left atrioventricular sulcus` *bounds* the `Anterior surface of the left ventricle of the heart`."

Classes and entities (nodes) in the Foundational Model of Anatomy (FMA) or other derived graphs are indicated by `monospaced text` and an initial capital letter, while anatomical terms used in the general discussion appear in standard text. Thus, "the `Left atrium of the heart (mouse)` maps to the `Left atrium of the heart (human)`", while "Like the human heart, the mouse heart is divided into four chambers, one of which is the left atrium."

Classes in CAIS are always indicated by a species' common name in parentheses after the anatomical structure [*e.g.*, `Right coagulating gland (mouse)`, `Lung (human)`], while classes in the FMA have no species' common name (*e.g.*, `Lung`).

Minor typographical or grammatical errors in the responses from researchers to our questionnaire were corrected before publication. None of the corrections had any effect on the meaning of the response.

Chapter 2
RELATED LITERATURE

This chapter provides a review of previous work in the areas that our system is based on. The background for our method is drawn from comparative anatomy, knowledge representation and modeling, and graph theory. Brief overviews of each of these domains follow.

2.1 Comparative anatomy

Comparative anatomy is the study of corresponding anatomical entities in different species. Its name is an umbrella term that covers many different subspecialties, users, and information needs. As a result, the detail of information available is anisotropically distributed, which creates fragmentation of the information resources available. The information differences can be classified along six different axes—user, purpose, species under study, anatomic specialty, level of abstraction, and granularity of information—in order to better understand what information is available in how much detail for what species, and what gaps remain in compiling adequate information to construct an anatomical model. In order to address these questions, however, first we review some fundamental comparative anatomy concepts.

2.1.1 Basic concepts in comparative anatomy: similarity and relatedness

Figure 2.1 shows the number of species of different kinds of life, to our best ability to determine. Of the approximately 1.6 million species shown in the figure, it is almost impossible to find any that do not have some degree of comparative medical interest, although some are obviously more immediately relevant than others for particular problems. The determination of which structures *correspond* across species is non-trivial, and our method does not derive those correspondences, but rather it models what anatomical consensus has deemed to be corresponding. The concept of "corresponding" is related to, but not synonymous with, the concept of "similar". Traditionally, comparative anatomy has recognized three kinds of similarity at a macro level—*homology* or similarity of ancestral origin, *analogy* or similarity of function, and *homoplasy* or similarity of appearance—all of which are orthogonal to each other.

At this point, it is useful to briefly explain what we mean by "similar" and "related", without formally defining them. The colloquial English, intuitively-understood sense of the word means that objects under comparison resemble each other in some way, usually visual. In other words, without any further refinement, "similar" in the way it is normally used in conversation is roughly equivalent to "homoplastic". This use of "similar" does not imply any evolutionary or inheritance relationship (nor does it rule one out), so we may say, for example, that bird wings are "similar" to bat wings, because they superficially look alike. Additionally, they are analogous, since they are both used for flight. But since wings evolved separately in bats and in birds, and since the superficial structures of the wings attach to the body at different places and use different bones of the animal's "hand" to support the structures, they are not considered evolutionarily "related" as **wings**. They are, however, related to each other as **forelimbs**, just as they are related to the forelimbs of any other vertebrate species that has forelimbs and hindlimbs, such as mammals, amphibians, or reptiles. As we will see over and over again, this example illustrates the importance of specifying the level of organization at which the structures are being compared.

By "related", we mean that there is an evolutionary inheritance relationship between the structures being compared. In other words, the structure evolved before the species diverged from each other, so both species inherited the "related" structure (or, in some cases, both inherited an earlier loss of a structure). An example is the mammalian lung, which evolved before the different kinds of mammals split off. So all mammals have related lungs, which happen to be very similar across species. Another example is that of the forelimb, mentioned above—because it developed in vertebrates long before birds and bats evolved, birds and bats consequently share related forelimbs, if not related wings.

After species diverge from each other, a great deal of change can occur on either or both sides, so related structures can undergo a lot of modification. The fact that structures are related (or homologous) does not necessarily imply that they appear similar (or homoplastic), or function similarly (or analogously). In fact they can appear so dissimilar that

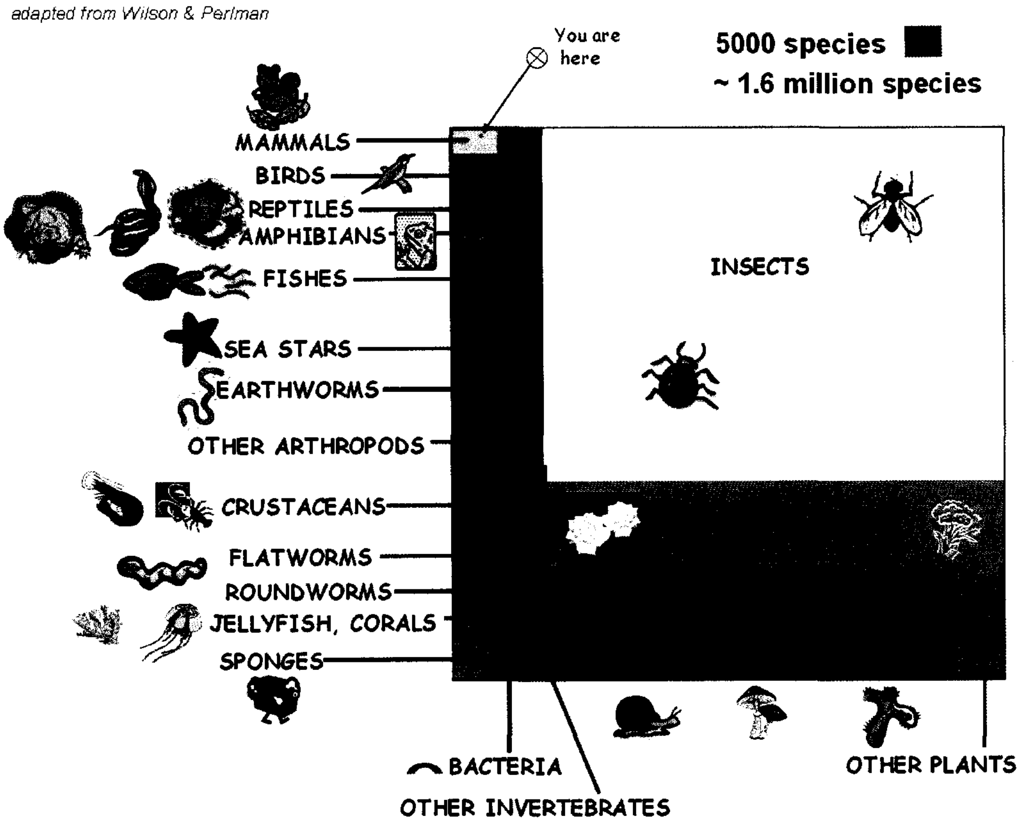


Figure 2.1: The scope of living species of biomedical interest (adapted from Wilson and Perlman's *Diversity of Life* CD).

researchers mistake them for unrelated structures. For a long time, this was the case with the eye. In *Drosophila* (fruit fly), squid, and vertebrates, the eye appears so different that it was assumed that eyes had evolved on at least three separate occasions. But recent genetic expression experiments have shown that eye development is controlled by homologous genes in each of the species in question, and that, therefore, despite superficial differences, eyes are indeed related in species as diverse as vertebrates, squid, and flies. Even this contention, however, is not uncontroversial. It has been suggested, for example, that although the same genetic expression is involved across the orders, that perhaps the homology lies not at the organ level of eye, but rather at the level of "photoreceptive visual organ", and that the eyes are indeed only analogous as eyes. Our method does not resolve these issues, but is flexible enough to model the domain experts' current consensus on what constitutes homology, and to remodel those comparisons should the consensus change [48].

So structures under comparison can be dissimilar and unrelated (*e.g.*, human lungs and human kidneys)[1], similar and unrelated (*e.g.*, bat wings and bird wings), dissimilar and related (*e.g.*, fly eyes, squid eyes, and vertebrate eyes), or similar and related (*e.g.*, dog livers and human livers). Although there is no technical reason why our method could not be used to compare any anatomical entities, in practice, comparisons of homologous structures are considered the only sound basis for making inferences from the source species to the target species, and so we confine the scope of our study to similarities and differences in homologous structures, as defined by anatomists. It is this homology that we refer to as "corresponding". These types of comparisons of related structures are the basis for animal models of disease, and for the translation to other species of the information that emerges from such models.

2.1.2 Levels of abstraction and the vertebrate Bauplan

The reason that medical knowledge can be leveraged across species at all is due to the fundamental structural similarity, or *Bauplan*, of mammals in particular, and vertebrates in general. The fact that fundamental aspects of the basic structure are so similar across the subphylum *Vertebrata*, and that there are such specific differences among the species within the subphylum, account for both the ability to apply knowledge across those species and for the difficulty of doing so in a consistent, predictable manner. These similarities and differences across species will occur at every level of organization, and will be accounted for in our method.

For example, despite species-specific differences in relative size and shape, the skulls of cats, dogs, bears, and humans share a great deal of similarity at the abstract level. They are all recognizable in isolation as "skulls", even when the exact species of the animal remains

[1] At least, they are unrelated at the organ level of organization. However, it may make sense to compare their branching epithelia to determine whether the genetic mechanisms that regulate the branching are related. See [27] for more information.

unknown. Figure 2.2 is clearly a skull of some sort, even without the specific information that it belonged to a panda.

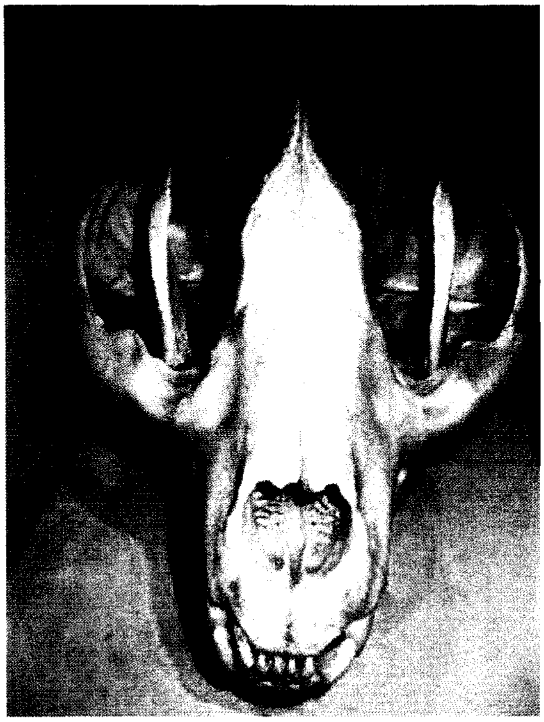


Figure 2.2: Skull of giant panda, National Zoo, Washington, DC.

The human hand, the mouse's paw, the horse's hoof, the seal's flipper, and the bird's wing have many specific, concrete differences, yet when observed at a higher level of abstraction, they are very similar in their structure: they are all the terminal segment of the forelimb of a vertebrate, and all originate from limb buds in the embryo and develop in the same way. So, when viewed as "hand", "paw", "hoof", "flipper", and "wing", the emphasis is on the differences; when viewed as "terminal end of vertebrate forelimb", they share a great deal of similarity.

This interplay between similarity and difference at the gross anatomical level is reflected at higher levels of organization as well: for example, at the organism level, these species look very different from one another, yet they all have a vertebral column, four limbs, a body divided into head, cervical (neck), thoracic (chest), and abdominal regions, etc. The differences are concrete, visible, and obvious; the similarities are abstract and less obvious. Therefore, the differences appear to be more numerous than the similarities, when the opposite is actually true. Despite the superficial visible differences, humans and other vertebrates (especially other mammals) are more similar than they are different, and this inherent similarity is the basis for the ability to make cross-species medical comparisons. It is worth noting that, no matter how similar two anatomical entities are across species at the gross level, or the histological level, or even at the level of resolution that can be viewed through an electron microscope, there will always be ultrastructural elements that are species-specific. For example, mouse and human mammary gland tissue may be indistinguishable from each other through the microscope, yet in the walls of the cells of those tissues are immunohistochemical antigens that recognize what species the tissue is, and will provoke a large immune reaction if transplanted into another species. Similarly, no matter how different two structures under comparison are at any given level, they will always be isomorphic at the level of `Anatomical entity`. For these reasons, there will never be perfect similarity (= identity) nor perfect difference at every level of comparison for two structures. A related point is that similarity is not transitive—structures can be similar at one level of organization, yet very different at another.

The reason that animals share such fundamental high-level similarity is due to the highly-conserved nature of the genes that regulate the establishment of the vertebrate *Bauplan* during its embryonic development ([24], [95]). For example, homologues of the set of homeobox genes that regulate the development of the mouse embryo into head, neck, chest, abdominal, and tail regions control the development of the human embryo. Even more surprisingly, they can be found in flies, worms, and other *basal* animals, as well. It is this similarity of highly-conserved genes across the animal kingdom that makes them the object of study in the databases described above, and which makes the question of comparing anatomy across species so important ([2], [7], [15], [25], [26], [41], [43], [42], [45], [63], [64], [75], [111]).

Despite the predominant similarities in structure across species, however, in this thesis we will be focusing on how the differences can be represented symbolically. The reason for this emphasis is that once similarity has been established at some level, there is not a lot of detail that attaches to that similarity. Multiple kinds of differences, however, can occur at multiple levels of organization and classification, and must be accounted for in more detail for a sound and complete representation. Therefore, despite the fact that in reality many more similarities than differences will be encountered in cross-species comparisons, the various kinds of differences and their classification and representation will be emphasized in this work.

Now that we have reviewed the basic concepts of similarity, difference, and correspondence in comparative anatomy, we address how different users have differing information needs in that domain.

2.1.3 The history of comparative anatomy, groups of users, and information needs

Much of the classical work in comparative anatomy has been written by evolutionary biologists or systematists, whose focus is on change over time in organisms and organ systems with an emphasis on function. Often for the sake of comparison, they tend to work with a greater number of species, but they write for their audiences in less detail (or granularity) than human physicians or surgeons do about structural attributes of organs for any one species[2]. Because they are greatly interested in the similarities in order to trace points where species diverge from each other, the published literature has tended historically to focus on higher levels of abstraction and less granularity. A great deal of the research has traditionally been devoted to the question of evolution, and so the systematists look at high-level changes across large taxonomic groups as adaptations to specific environments for evidence of or nuances to the larger evolutionary issue. For example, Hildebrand's discussion of the gall bladder [49] states: "The organ is always present in carnivores. It is lacking in the adult lamprey, several teleosts [fish], and in certain herbivores distributed in five families of birds and six orders of mammals." In exactly which species the gall bladder is lacking—essential information for modeling the anatomy of a particular species—is not the

[2] However, there are many exceptions to this generalization, and it should not be ignored that some of the finest, most detailed work for particular species has been carried out by systematists. Often, the detailed anatomical information lies in the primary literature, while the textbooks or popular literature are confined to the higher-level points. An important related issue which needs to be addressed, but which lies outside the scope of this thesis, is the risk of loss of huge quantities of valuable detailed anatomical information from these primary sources, which have gone out of print before ever having been made digitally available.

important point for him in this book, but rather the important point is the association of the existence of the structure to whether the animal is a carnivore or an herbivore, and the entailed vertebrate evolutionary issues across families and orders upon which this variation in existence sheds light. In order to get the detailed attribute information, on the other hand, one would have to hunt down the primary literature, if it even remains available.

As a result of the systematists' emphasis on adaptation and selection, they have often tended to focus not on anatomy *per se*, but on the closely-related discipline of morphology, or the study of the interplay between form and function (*cf.* Hildebrand's description of the gall bladder is not of the structure in isolation, but is rather in relation to whether different species are herbivores or carnivores, where the gall bladder provides an adaptive advantage in digestive physiology). One of the most celebrated examples is Davis' study of the giant panda (*Ailuropoda melanoleuca*), which resolved the issue of whether the panda was more closely related to raccoons or to bears ([77]). Based on feeding behavior, a small minority of scientists (the behaviorists) argued that the giant panda was a close relative of the lesser (or red) panda (*Ailurus fulgens*), and therefore, like the red panda, was a *procyonid* (closely related to raccoons). By examining the anatomical structures of the giant panda at a high level (anatomy), and by relating those structures to adaptations for the panda's diet of bamboo (morphology), Davis was able to show that, despite a superficial resemblance to the red panda—no doubt reinforced by the name—the giant panda is in structural terms indeed a bear, whose adaptations in structure were functional responses to its dietary niche, rather than evolutionary relatedness to procyonids. Although a more famous example than most, this one is representative of the types of problems with which systematists often concern themselves—morphology, rather than anatomy proper—and the published literature reflects this emphasis, which makes getting details of the pure anatomy often somewhat more complicated.

Lately in the systematist literature, there has been a new emphasis on molecular zoology, in order to trace phylogenetic distance ([50], [51], [68], [84], and [121] are representative of the genre), and to tie molecular signatures to the development or disappearance of specific structures in different species. However, this kind of information is found at very low levels of organization—the intermediate levels of anatomical detail, where the attribute differences between structures live, often does not have immediately useful tie-ins with the features under study. But the larger questions of the dynamic tension between form and function, pioneered by the systematists, continue to inform the debate in comparative anatomy.

Veterinary information users, on the other hand, tend to work at the same level of detail as human anatomists, but the information available tends to be constrained to economically or sentimentally important species, such as dogs and cats, or cows, horses, pigs, and sheep. The standard reference source for veterinary terminology, *Nomina Anatomica Veterinaria (NAV)* [88], confines itself to the above species (although in the section on neuroanatomy only, they introduce a primate species to increase the level of complexity that they are able to name). *NAV* is a partonomy written in Latin only; there is no translation or definition of the terms, although some discussion of interspecies subtleties and refinements takes place in the footnotes. An example of terms for parts of the face is shown in Figure 2.3.

Additionally, there are surgical atlases for those animals (particularly dogs and horses), which gives attribute information in some detail, but mice and rats have not traditionally been species that veterinarians have concerned themselves with treating, and thus such detailed centralized anatomical reference sources are not readily available for those rodents. Much of the information for mice, as well as for other species, does not exist in traditional atlas form, but rather is distributed across published journal articles, and there is no independent verification that different investigators mean the same thing by the same terms in these articles. For example, some investigators differentiate the dorsal and lateral prostates (*e.g.*, [100]); others regard the dorsolateral prostate as one organ (*e.g.*, [83]). Sometimes these structures are referred to as organs; other times as lobes (constituent parts of a lobular organ). Even the most widely recommended atlas for rodents ([94]) does not give much detail beyond the organ level, although such detail would be valuable in resolving these issues and discrepancies.

Surgeons have traditionally used pig, sheep, and dog organs for practicing their techniques, and in that way, have probably paid more attention to the attributes of structures that have significance for pathological transformation, such as adjacencies, innervation, blood-supply and lymph-supply, etc. However, their focus is on practicing for human surgery, not on recording comparative anatomy discoveries, and so this source has often

```
Facies
  Oculus
    Palpebra superior
    Palpebra inferior
    Rima palpebrarum
    Bulbus oculi
    Sulcus infrapalpebralis
  Nasus
    Dorsum nasi
    Apex nasi
    Ala nasi
    Naris
    Planum nasale
    Planum nasolabiale
    Rostrum
    Planum rostrale
  Os
    Labium maxillare
    Labium mandibulare
    Rima oris
    Cavum oris
    Lingua
    Fauces
    Bucca (Mala)
    Mentum
    Sulcus mentolabialis
```

Figure 2.3: Hierarchy of terms for parts of the face from *Nomina Anatomica Veterinaria.*

not produced much information, organized and published in a systematic way for other species. Surgeons such as Narath ([82]), who dissected hundreds of lungs of different species of animals, recorded their observations as part of hypotheses about human development, but the raw data on which these hypotheses were based is extremely difficult to obtain, if it still exists at all.

In contrast to systematics, comparative medicine *per se* is a relatively new discipline, but the amount of information emerging from it is exploding at an unprecedented rate. Practitioners of comparative medicine work on the structures themselves, in any species that is of interest in animal modeling of human disease. However, they do not necessarily need to know all the names of the other structures nearby, which is essential for modeling the spatial and other relationships in the ASA, a component of the FMA which will be explained in more detail below. The reason for this is that they are not medically or surgically treating the animal they study in the same way that a veterinarian or physician would, and so do not have the same need for detailed knowledge of the names and spatial relationships of the nearby blood vessels. They often dissect the animal, and so they focus on the structure or pathology of interest rather than on learning the names of surrounding structures. This approach serves their needs well, but it means that they are not as knowledgeable about the structural attributes and relationships among surrounding structures as one might expect. Additionally, the quantity of molecular biology information often tends to detract from focusing on certain anatomical details, such as attributes, in favor of gross differences in the entities (structures) themselves.

As we have seen above, the choice of species for a particular anatomical problem often depends on the user's information needs, and that, in turn, influences how much and what type of information is available for a particular species. We have seen that information on economically important species has emerged from the needs of veterinarians and veterinary surgeons, while human surgeons have often assembled information from practice on species such as the pig and dog, due to their similarity to humans.

In addition to information needs, logistics and tradition drive the choice of experimental animal, and thus the distribution of readily-available information. Dogfish sharks (*Squalus acanthias*), frogs (*Rana spp.*), and cats (*Felis cattus*) have been popular choices for classroom dissections due to their availability, and that in turn has led to the development of a great deal of published anatomical information, although the direct relevance to specific medical problems is not always obvious. The growth of animal modeling of disease and genomics research has increased the importance of mice and rats as experimental animals; species that certainly had been studied previously, but not to the extent that they currently are. Yet that has not translated into the development of centralized, easily-available anatomical information, as will be discussed in further detail in the sections on specific mouse resources.

We wish to develop sound and complete representations for the anatomical structures we are modeling. However, there is no single set of users who have compiled this information already for their own needs. We therefore have to generate much of our own data, in order to come up with a meaningful model, because if we continue to work at the high level of abstraction of much of the current comparative anatomy literature, we tend to skew toward a false similarity. It is in the mid-level attributes that the most differences emerge, and from which our method can be most rigorously validated. For this reason, we need data that is the union of the needs of the groups of comparative anatomists identified above. This requirement makes development of the models more difficult, but has the benefit that, once they are fully developed, they can serve the information needs of many different groups of users simultaneously, through the use of views [28].

2.2 Knowledge representation and modeling

> "Leonardo da Vinci's famous sketch of a human fetus in the uterus, shown below [in Figure 2.4], is intriguing because he clearly gave it a cotyledonary placenta as is seen in ruminants. The reason for this mistake is not known, but the level of detail presented indicates that he was very familiar with the ruminant placenta." [http://arbl.cvmbs.colostate.edu/hbooks/pathphys/reprod/placenta/leonardo.html]

This sketch by da Vinci is a symbol for the importance of getting the comparative anatomical information right to start with, since a model is only as good as its underlying information. Da Vinci carefully, lovingly, and studiously crafted an elaborate rendition of the human fetus in the womb. The problem is, working from animal models, he inadvertently gave the uterus a cotyledonary placenta (as in the cow in Figure 2.5. The human placenta, by contrast, is discoid, like the brown bear placenta in Figure 2.5 (and not bidiscoid as is the case for the more closely-related rhesus monkey, and what one might therefore expect in the absence of more specific information). (Image source: [37]

Leonardo da Vinci's sketch is thus an object lesson in getting the comparative anatomy information underlying the model right from the start. We do so in the following way:

- correspondences are drawn not from superficial homoplasy or analogy or term resemblances, but from the genetics and embryology underlying the structures (to the

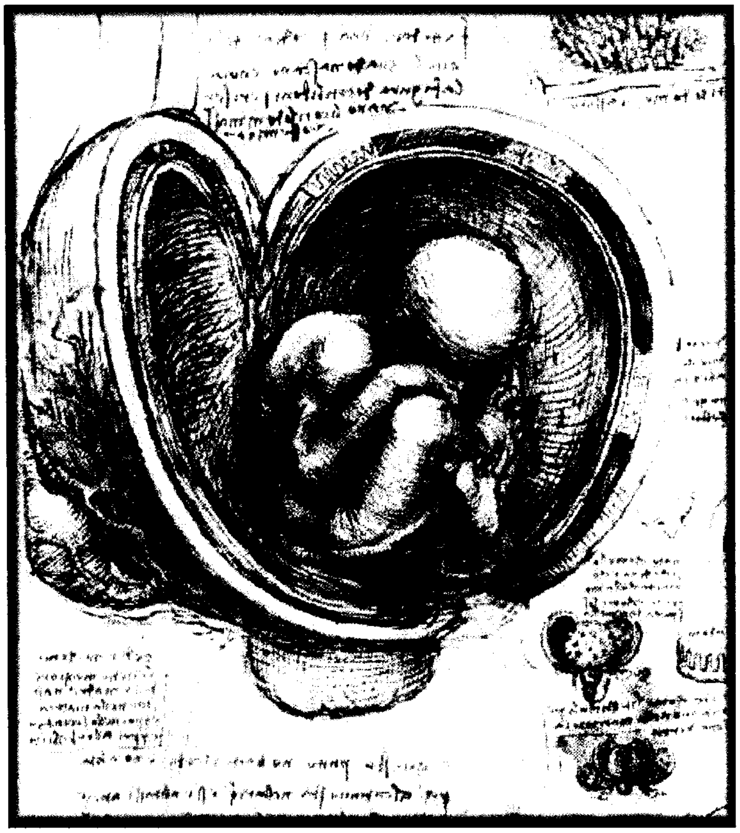


Figure 2.4: The importance of getting the anatomy right, inadvertently illustrated by Leonardo da Vinci.

degree that that information is known and available);

- developmental biology and evolutionary biology are essential to the proper understanding of that underlying genetics and embryology (sources); hence, the unavoidable necessity of dealing at some level with the ATA and Mk;

- in order to correctly render the underlying anatomy, we employ Smith and Rosse's approach of "biological reality (refs) and Perl's principled modeling, with the underlying premise that "formalization improves conceptualization" (Rosse).

The title of this dissertation is a tribute to two seminal ideas in biology. Underlying our whole modeling approach is Dobzhansky's "Nothing in biology makes sense except in the light of evolution"—it is the underlying evolutionary history of the structures we are comparing that renders our model sound and complete by focusing on homology, rather than the misleading analogies and homoplasies. The second seminal idea is from Haeckel, and the title is a play on his observation that "ontogeny recapitulates phylogeny", or the individual embryo of any species passes through developmental stages that reflect the history of the species (*e.g.*, the tail of the human embryo, which is later lost). Although in its original naïve formation, it was flawed and missed important nuance, it was still an important step in recognizing the phylogenetic connectedness of the different species, which directly leads to animal models and comparative medicine. In order to fully integrate biology and informatics—to "put the biology in bioinformatics"—such an understanding of how evolutionary and developmental biology inform our modeling efforts is crucial to a sound and complete comparative anatomical representation.

Other work on symbolically modeling the mouse

Although there is a great deal of data emerging from the mouse model, and consequently a large incentive to organize that data, there has not been much done in the way of constructing a sound and complete symbolic model for the mouse. A few attempts have been made, but they embody the fragmented state of current knowledge, and replicate problems in the literature.

2.2.1 Introduction to the Foundational Model of Anatomy (FMA)

As previously mentioned, the first step in our approach is the collection of information about the biological model from domain experts and secondary literature. Once that information has been gathered and organized, the next step is to structure it into an appropriate symbolic model, and for that purpose, we used the existing models of the homologous human structures in the Foundational Model of Anatomy (FMA) as a template.

The FMA is a symbolic model of the physical organization of the human body. More specifically, it is an ontology which furnishes a comprehensive set of entities and relationships which describe the human body at all levels of structural organization. At the highest level of abstraction, it consists of the following components:

$$\text{FMA} = \{\text{AT, ASA, ATA, Mk}\}\text{, where} \tag{2.1}$$

$$\text{AT} = \text{Anatomical taxonomy} \tag{2.2}$$

$$\text{ASA} = \text{Anatomical structural abstraction} \tag{2.3}$$

$$\text{ATA} = \text{Anatomical transformation abstraction} \tag{2.4}$$

$$\text{Mk} = \text{Metaknowledge (principles, rules, and axioms)} \tag{2.5}$$

The *AT* component is a class hierarchy of entities that describes the body at levels of organization from organism down through organ and cell to macromolecule, based on the *is-a* relationship ([101]). Extending it to the mouse involved ascertaining the important entities and terms involved. The AT's emphasis on entities, rather than terminology, serves us well when deciding what structures to correlate; this will be discussed in more detail below.

The *ASA* describes the structural relationships among anatomical entities in the canonical or standard adult of the species under study. It consists of the following components:

$$\text{ASA} = \{\text{DO, BN, PN, SAN}\}\text{, where} \tag{2.6}$$

$$\text{DO} = \text{Dimensional ontology} \tag{2.7}$$

$$\text{BN} = \text{Boundary network} \tag{2.8}$$

$$\text{PN} = \text{Part-of network} \tag{2.9}$$

$$\text{SAN} = \text{Spatial association network} \tag{2.10}$$

These components serve to describe the shape, connections, boundaries, location, and orientation of the structures under study, as well as describing units of organization in terms of their component parts. This is where many of the medically-important differences in the structures we are studying will be found.

While *sensu strictu*, the *ATA* and *Metaknowledge* (Mk) are outside of the scope of our information system, nevertheless in order to properly represent homology, these components are unavoidably involved, and so we treat them briefly here. The ATA spells out the "relationships that describe the morphological transformation of anatomical entities during pre- and postnatal development" ([101]). Although the ATA per se is outside the scope of this paper, it should be noted that while the ATA component of the human FMA is currently constrained to the modeling of embryology (normal development), the study of transformational processes in animal models often goes far beyond the study of normal development. The study of transformation in animal models encompasses such disciplines as teratology (*e.g.*, birth defects in zebrafish and amphibians in response to chemicals in the environment), physiology (*e.g.*, how bears preserve muscle and bone mass and regulate excretory functions during hibernation without experiencing the loss of structure and function a human would exhibit after extended periods of immobility), pathology (*e.g.*, cancer growth and metastasis in mice as a model for human disease), and pharmaceutics/pharmacology (*e.g.*, drug-induced changes in structure in various species). The ATA offers the promise of a methodology for modeling these domains as well as standard normal embryology, although, as stated, such applications lie far outside the scope of this thesis. However, our method would certainly be extensible in this domain.

Metaknowledge (Mk) is knowledge about knowledge—it includes the rules, principles, and axioms underlying the anatomical knowledge represented in the model. It is outside of the scope of our mouse model, but will become important with dealing with metamodels, such as the rodent, mammal, or vertebrate metamodels.

The FMA was originally developed to represent human anatomy. However, the common features of the vertebrate *Bauplan*, whose establishment during embryonic development is regulated by a highly-conserved group of structural genes, and the inclusion in the FMA of high-level, abstract classes which correspond to the *Bauplan*, enable the extension of the FMA to non-human species, and the resulting ability to compare corresponding structures across species. Additionally, the FMA's emphasis on the concept vertex of Ogden and Richards' *semantic triangle* ([87]), rather than on the terms vertex (where most terminologies concentrate), permits resolution of the inconsistent terminology problems referred to earlier—for example, we promoted the term `Coagulating gland`, but included the term `Anterior prostate` as a deprecated synonym. Or we can include `Dorsolateral lobe of (mouse) prostate` as a synonym for `Dorsolateral prostate`. In that way, users can freely use either term without fear of losing or compromising information as a result.

In developing hierarchies for the mouse prostate and mammary gland, we extended the existing human FMA to create mouse organ templates; we then used those templates to map structures at levels of organization from the organ down to the cell, in order to determine where the similarities and differences lay. Additionally, because the mouse anatomical symbolic model is based on the FMA, our comparison will have to deal with differences between the structures themselves at various levels of organization, but will not need to deal with model or meta-model conflicts.

An add-in to the basic Protégé interface to the FMA is *Emily*, a query engine for the FMA, focused on supporting queries on the relationships among anatomical entities. We will build on previous work on *Emily* ([29]) as a basis for our query engine.

The Jackson Laboratory has attempted to develop terminology hierarchies for mouse anatomy and for mammalian phenotypes. This is an important goal, because so many different databases exist. The Jackson Laboratory Mouse Genome Informatics web page ([55]) serves as a portal to bring a great deal of diverse information together, and is user-friendly and intuitively organized by views, such as "genes" or "alleles" or "tumor biology". The list in Figure 2.6 is a representation of body spaces at the embryologic Theiler stage 28 (TS28) in the mouse. An attempt at cross-species comparison is implicit in their Mammalian Phenotypes page, as represented by small ventral prostate in Figure 2.7.

Yet despite the worthiness of the goal and the ambitiousness of the project which they have attempted, there are problems with their hierarchies. In the case of the condition `Small ventral prostate`, following the links gets the user to the representation in Figure 2.8.

Although the dorsolateral and ventral lobes are represented there, the coagulating and ampullary glands are missing. While it is currently a matter of debate whether the ampullary glands are to be regarded as prostates, there is no question that the coagulating glands are prostates, and the fact that they are not represented is a serious content omission.

Additionally, the ventral lobe of the prostate is clearly distinguished from the dorsolateral lobe, yet the definition of small ventral prostate is reduced size of lateral [sic] lobe of the prostate both a term inconsistency and a concept inconsistency in the relationship between the phenotype and its definition.

There are other issues with the hierarchy as well. In their representations, it is clear that the part-of and is-a relationships are mixed–on their Web pages, they indicate which relationship is which with a colored superscript marker before the term-a fact which invalidates the inheritance hierarchy. For example, transitivity is inherent in is-a, but because of the variety of part-of relationships, "the transitivity of part-of relations cannot be granted in general"([47]). Mixing them in the hierarchy in this way thus limits the kind of reasoning that can be performed on the entities and relationships in this hierarchy.

The representation of body spaces at TS 28 exhibits the same confusion in the hierarchy between part-of and is-a relationships. Additionally, the criteria for part-of is not clear perhaps not every embryologist would agree that the body is part-of the embryological Theiler stage TS 28, as this hierarchy maintains. This relationship between these entities is a question for the domain experts to resolve, and for the model to represent according to their consensus.

Some of the is-a relationships in the Gland abnormalities phenotype (Figure 2.9) are similarly not universally agreed-upon: abnormal sex gland secretion is-a abnormal sex gland seems to be a dubious assertion, as does the same relationship for absence of sex glands (although perhaps dealing with the concepts in terms of the noun abnormality rather than the adjective "abnormal" plus the noun for the concept would be sufficient to clear it up). More puzzling is the relationship that glands : no defect detected is-a gland abnormality, as in Figure 2.9.

However, despite the problems in their implementation, it is important to acknowledge that they have tackled some difficult problems, such as reconciling very disparate databases, and bringing them together in one place for easy comparison. One of those resources that they incorporate into their portal is the anatomical nomenclature from the Edinburgh Mouse Atlas Project ([36]) (EMAP 2004), which has had a long-term collaboration with the Jackson Laboratory on anatomical nomenclature.

EMAP attempts to address some of the problems in the literature, and tries to be consistent in terminology and relationships. They correctly identified problems with using only Theiler's criteria to distinguish phases of early development, and have combined it with cell and somite numbers, as well as Downs and Davies characteristics ([36], [34]). They represent stages as a range, in order to account for individual variations in development, and this in itself is enough to be an important aid to the field. Additionally, they link the terms to pictures, providing a useful resource. They attempt to standardize the terms, which is useful in itself, and they offer to work with other terminology standards to facilitate translation between terminologies, which enables data exchange. The user interface is friendly and permits viewing of different Theiler stages, as well as different levels of granularity within a stage. Figure 2.10 shows a representative sample of their ontology.

However, there are some problems with this resource. It suffers from the confusion between part-of and is-a hierarchies described above. Additionally, embryological structures appear and disappear between stages, and if the structure the user is interested in does not appear in the stage being viewed, there is no easy way to search for it. Because it only represents embryological structures, and many structures (such as the prostate) develop primarily after birth, it is of limited use for those postnatal structures or for comparing to the adult. Additionally, it is limited to the mouse—although they try to link it to their human model and other cross-species comparisons are non-existent.

Wilcke's veterinary standards group at the University of Virginia is working to develop a veterinary model that can be reconciled with SNOMED, but they have encountered the anthropocentrism that is inherent in the human-based systems. By creating a parent organ approach, they overlay the animal knowledge onto the existing human counterpart, and thus attempt to side-step the anthropocentric focus of SNOMED (Figure 2.11). Their goals are stated as follows: 1) context-independent definitions; 2) logical and true relationships; 3) rapid and easy addition of variations ([131]).

Although they consistently use the term "analogous" when they mean "homologous", their approach that "analogous [sic] structures should be grouped under a parent that defines their similarities" has a great deal to recommend it. However, the combinatorics of having a separate entity for each organ for each species makes an already computationallyintensive problem into a prohibitive one. A pairwise comparison of every attribute and every relationship for every structure in every species is potentially on the order $O(fn^2)$, where f is the number of structures involved and n the number of species. Creating a child for every structure by species increases the computational effort to approximately $0(f^n)$ for the entire model. In Chapter 4, we will discuss how our approach combines the advantages of Wilcke's approach with the minimization of extra entities.

Other efforts have extended to attempting to symbolically model mouse pathology, but have the same problem that modelers of human pathology encounter: there is no firm agreement on what constitutes pathology. So in addition to any inconsistencies within a model, the problem of model and metamodel conflicts comes up. Additionally, the same inconsistencies as in the other models described above are present—lack of standardization of vocabulary, confusion of is-a and part-of, and so forth.

Despite the problems enumerated above, which are to be expected at the beginning of attempting a truly original task—that of creating symbolic models for cross-species comparisons—all of these symbolic models are first steps toward an important goal. However, in order to have a fully sound, complete, and logical representation of animal models of anatomy, the human needs to be displaced as a reference model, in favor of a vertebrate-based representation of structure. The FMA, which will be discussed in more detail below, has the necessary qualities to serve as the basis for a sound and complete pan-vertebrate metamodel, and avoids the problems discussed above. In the introduction to the FMA below, and in Chapter 4, we will discuss at greater length how these problems are avoided.

Mammalian herbivore stomachs and non-quantitative distance

An interesting example of the kind of interspecies comparison that we have discussed is demonstrated by the different expressions of the mammalian herbivore stomach. There are many different species of herbivores, and they have developed a number of different adaptations to the niche. In *The Mammalian Herbivore Stomach* ([66]); Peter Langer arranges the species by what he terms "levels of differentiation", rendered in Figure 2.12[3].

[3] Legend: *Ailuropoda* = giant panda; *Homo* = human; *Sus* = pig; *Sirenia* = manatee and dugong; *Hippopotamidae* = hippopotamus; *Bradypodidae* = sloth; *Tayassuidae* = babirusa (wild pig); *Macropodidae*

What he has touched upon in this arrangement is the possibility of a non-quantitative or non-numeric distance measure—in other words, a symbolic distance measure. It is clear from the arrangement in Figure 2.12 that in this representation, the human stomach is more like the pig stomach than it is like the panda stomach, or that the manatee stomach is more like the sloth stomach than it is like the human stomach.

If this is a valid representation of anatomical distance, then Langer has hit upon a very powerful technique for deciding which animal model is more appropriate for which disease/organ system, or for determining systematic correspondences. But it remains to be seen whether this representation is sound and complete; indeed from the outset, there are some problematic issues with the levels of differentiation Langer has chosen.

First of all, it is necessary to ask whether any given criteria (or all criteria) are equally meaningful and appropriate; it is not clear, for example, that an increase in volume (which would be represented as an attribute in the FMA) is as important as the appearance of discrete anatomical structures such as ampulla duodeni or taeniae and haustrae (which would be FMA nodes or entities). If they should truly be equidistant, as Langer has them placed along the x- and y-axes, they should be equally important, and it is not clear that such is the case. He has also included rumination (a function) along with the structural attributes, which is clearly very different.

Additionally, he appears to have omitted important criteria. For example, as we shall see later when we examine the mouse stomach, the differentiation of the glandular part and the non-glandular part is an essential distinction, yet there is no place on this graph for it. The mouse, although an herbivore, could therefore not be accommodated under these criteria. Furthermore, he has included important criteria as a sidebar in the case of the panda and the *Hippopotamidae* surely the caecum (which plays a very important role in plant digestion) is as important as the diverticulum ventriculi (a spatial/connectivity arrangement), so it is puzzling why the latter is used as a level of differentiation when the former is not.

Finally, it is not clear that his levels of differentiation are truly differentiae in the onto-

= kangaroo and wallaby; *Neoselenodontia* = camel/llama suborder + ruminant suborder.

logical sense: for the stomach, everything which occurs below the features "taeniae, haustra, and semilunar folds" has a stomach lacking those features; they then appear for Macropodidae (kangaroos) and disappear again unsystematically for Xeoselenodontia (camels and cattle), and thus are not truly differential in the ontological sense.

Despite these problems, however, the promise of a non-quantitative distance measure is an exciting possibility, and foreshadows possible applications of our method in comparative medicine and in systematics. We will refer later to the necessary criteria for modeling attributes and relationships in an ontology for comparative anatomy, as well as the mathematical tools for manipulating the knowledge contained in the ontology.

2.2.2 Model management

Pottinger, Bernstein, and Halevy ([9], [96]) have conducted research in the area of model management to formulate an approach to mapping and merging two different models, for example, the inventory merger of a bookstore with that of a video store. Some of the issues and challenges with which they have dealt are directly relevant to developing and querying our model, so their work will be reviewed briefly here. Figure 2.13 ([96]) shows a mapping of two models that specifies that FirstName and LastName should be elements of the element Actor in the mapped model.

To implement such a mapping, they have proposed a model-matching-and-merging approach to deal with the problems of combining two or more different schemas in a database environment. Their schemas are represented as graph structures, as are ours. They allow a node in one graph to map to a node in the other graph if they are identical or "similar" entities. Using a very simple definition of similarity, they have developed a matching algorithm to find a mapping from one graph to another. The resulting match is represented as a graph structure itself, a very nice idea which we have implemented in our work.

As a result, one of the most important aspects of their work is that the mapping between two models is itself a model—*i.e.*, it is a *first-class object*, and thus can undergo the same operations as the original models. They outline a set of model management operators, of which the following will be relevant to our Structural Difference Method: 1) match, 2) apply,

3) compose, and 4) difference.

2.3 Graph theory

There is a large body of literature on the application of graphs and graph theory to the description of structural relationships. Graphs are useful mathematical structures because the *nodes* of the graph can be used to represent the anatomical structures under study, while the *edges* of the graph can be used to represent the relationships among those anatomical structures. In that way, we can formally capture what is similar and what is different in comparable structures and relationships, by constructing a graph for each anatomical structure and comparing (matching) the graphs.

Let $G_A = (A, E_A)$ be a graph with node set A and edge set E_A, and let $G_B = (B, E_B)$ be a second graph. A graph isomorphism is a one-to-one, onto mapping $f : A \longmapsto B$ such that $(a, a') \in G_A$ iff $(f(a), f(a')) \in G_B$. This means that if there is an edge between nodes a and a' in G_A, there must be an edge between the corresponding nodes $f(a)$ and $f(a')$ in G_B, and vice versa. This is called a *relational constraint.*

Let Graph A be a representation of the human heart (H), and Graph B be a representation of the mouse heart (M), as depicted in Figure 2.17. The root of each graph is `Heart`, and it has four children, connected to `Heart` by the relationship *has-part*: `Left atrium`, `Left ventricle`, `Right atrium`, and `Right ventricle`. (For simplicity of illustration, we limit the graph to `Cardiac chambers`).

In mapping the nodes of Graph A to the nodes of Graph B, mouse `Heart` matches human `Heart`, `Right atrium` matches `Right atrium`, and so forth. Similarly, the four *has-part* edges match. The mapping is therefore one-to-one and onto, and the relational constraints are satisfied, which constitutes a graph isomorphism. If a graph is isomorphic to a subgraph of another graph, the relationship between the graphs is that of a *subgraph isomorphism.*

In addition to isomorphism, which denotes an exact match between the structures under comparison, the concept of *homomorphism*, or relationship-preserving partial mapping, is useful in analyzing similar structures. Shapiro and Haralick ([108], [107]) formally define a *relational homomorphism*, in order to create a construct that will map the nodes of one graph to those of a second graph, in a way that preserves the interrelationships among the nodes. They call this homomorphism a structure-preserving function, and define it as follows:

Let A be a finite set of objects, let L be a finite set of labels, let $R \subseteq A^N \times L$ be a labeled N-ary relation, and let $h : A \longmapsto B$ be a mapping from A to a second set B. The composition of the relation R with the function h is the labeled N-ary relation $R \circ h$ defined by

$$\mathrm{R} \circ h = \{(b_1, \cdots, b_N, \ell) \in B^N \times L \mid \exists(\mathrm{a}_1, \cdots, a_N, \ell) \in R \text{ with } h(a_i) = b_i, i = 1, \cdots, N\}.$$

Suppose $R \subseteq A^N \times L$ and $R' \subseteq B^N \times L$. A relational homomorphism from R to R' is a function $h : A \longmapsto B$ such that $R \circ h \subseteq R'$.

These comparisons open up the concept of *graph distance*, or how different or similar graphs are to one another. Shapiro and Haralick utilize the concept of relational homomorphism in the development of their relational distance, which—with some differences—is an essential component of our method.

Relational distance goes one step further than relational homomorphisms; it allows for a quantitative comparison between two relational structures (graphs). In general, given a 1-1 mapping $f : A \longmapsto B$, the relational error of the mapping is defined as

$$Error_f \quad = \quad \mid E_A \circ f - E_B \mid + \mid E_B \circ f^{-1} - E_A \mid \tag{2.11}$$

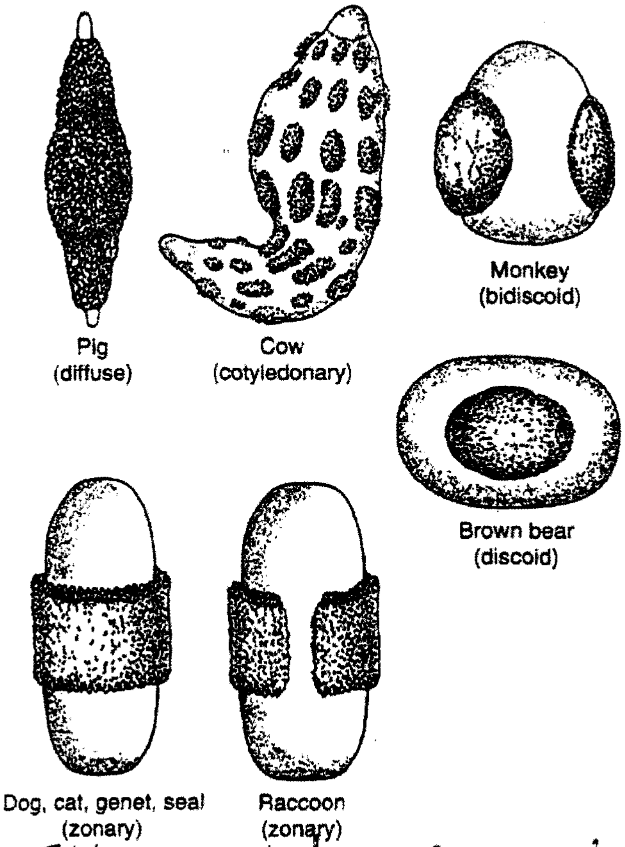
where E_A is the edge set of A, and E_B is the edge set of B.

Sanfeliu and Fu ([105]) worked on a similar problem in the context of pattern recognition. They categorized the different methods of computing a distance measure between attributed graphs, and proposed a distance measure based on cost functions. Given two graphs, a source graph and a reference graph, the cost functions were used to compute the cost of a mapping from the nodes of the source graph to those of the reference graph. Their mapping cost is a summation of the number of node insertions, node deletions, edge insertions, and edge deletions that must be performed to transform the source graph into the reference graph. The minimal mapping cost over all possible mappings (*cf.* Shapiro and Haralick's relational distance) is the distance between the graphs.

These formalisms that we have outlined are for simple graphs, but the frame-based representation of the FMA in Protégé is much more complex than a simple graph since 1) it has attributed nodes (*e.g.*, *has-mass*; *has-inherent-3D-shape*), and 2) it has multiple relationships (*e.g.*, *is-a*, *has-part*, *continuous-with*, *adjacent-to*). The edges of the complex graph structure of the FMA represent this rich mixture of structures and relationships. We have found that similarities and differences between two graphs can occur at all levels, as well as across levels, and that, as expected, there are more similarities than differences.

2.4 Summary

In this chapter, we presented the basic components of our approach in some detail. We introduced the discipline of comparative anatomy, and reviewed some of its history, which accounts for the different user groups, information needs, and anisotropic distributions of available primary data in the field. We proceeded to introduce the FMA, which we used as a template to structure the primary data that we collected, and we finished with a discussion of graph theory and existing work in the field of graph matching, which motivated the development of our model.

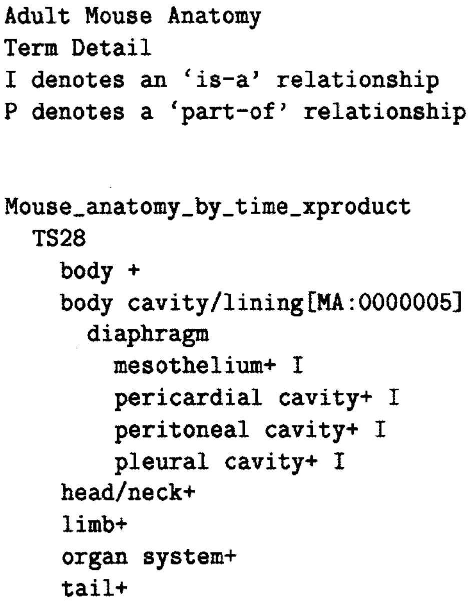


Feldhamer et al 1999 Mammalogy

Figure 9.12 **Placental villi.** The shape and distribution of placental villi vary among different groups of mammals.

I
shov
uter

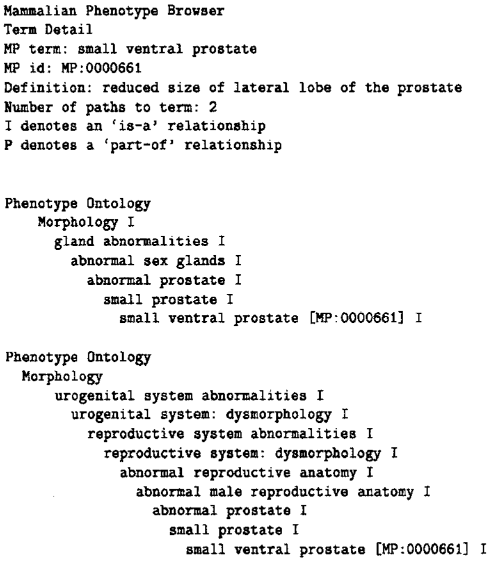
Figure 2.5: A sample of the diversity of mammalian placentae.



```
Adult Mouse Anatomy
Term Detail
I denotes an 'is-a' relationship
P denotes a 'part-of' relationship


Mouse_anatomy_by_time_xproduct
  TS28
    body +
    body cavity/lining[MA:0000005]
      diaphragm
        mesothelium+ I
        pericardial cavity+ I
        peritoneal cavity+ I
        pleural cavity+ I
    head/neck+
    limb+
    organ system+
    tail+
```

Figure 2.6: Jackson Laboratory mouse anatomy hierarchy.



```
Mammalian Phenotype Browser
Term Detail
MP term: small ventral prostate
MP id: MP:0000661
Definition: reduced size of lateral lobe of the prostate
Number of paths to term: 2
I denotes an 'is-a' relationship
P denotes a 'part-of' relationship


Phenotype Ontology
    Morphology I
      gland abnormalities I
        abnormal sex glands I
          abnormal prostate I
            small prostate I
              small ventral prostate [MP:0000661] I

Phenotype Ontology
  Morphology
      urogenital system abnormalities I
        urogenital system: dysmorphology I
          reproductive system abnormalities I
            reproductive system: dysmorphology I
              abnormal reproductive anatomy I
                abnormal male reproductive anatomy I
                  abnormal prostate I
                    small prostate I
                      small ventral prostate [MP:0000661] I
```

Figure 2.7: Jackson Laboratory Mammalian Phenotypes page.

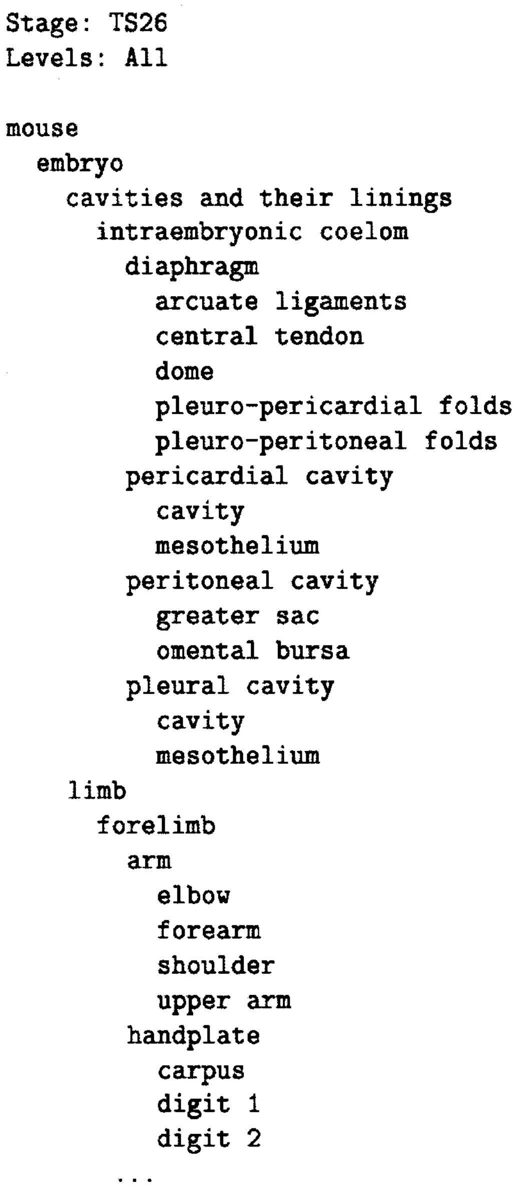
```
Adult Mouse Anatomy
Term Detail
MA term: prostate gland lobe
MA id: MA:0001738
Number of paths to term: 5
I denotes an 'is-a' relationship
P denotes a 'part-of' relationship

Mouse_anatomy_by_time_xproduct
    TS28 P
      body+ P
        body organ P
          lower body organ I
            pelvis organ I
              male reproductive gland organ I
                prostate gland I
                  prostate gland epithelium P
                  prostate gland lobe [MA:0001738] P
                    prostate gland dorsolateral lobe I
                    prostate gland ventral lobe I
                  prostate gland smooth muscle P
```

Figure 2.8: Structures other than dorsolateral and ventral lobes are missing from mouse prostate *is-a* hierarchy.

```
Phenotype Ontology
    Morphology I
      gland abnormalities I
        abnormal adrenal gland + I
        abnormal crypts of Liberkuhn + I
        abnormal lacrimal glands + I
        abnormal liver + I
        abnormal mammary glands + I
        abnormal neuroendocrine glands + I
        abnormal pancreas + I
        abnormal parathyroid glands + I
        abnormal salivary glands + I
        abnormal sebaceous glands + I
        abnormal sex glands [MP:0000653] I
          abnormal bulbourethral gland + I
          abnormal ovaries + I
          abnormal preputial glands + I
          abnormal prostate + I
          abnormal seminal gland + I
          abnormal sex gland secretion + I
          abnormal testes I
          absence of sex glands I
        abnormal sweat glands + I
        abnormal thyroid glands I
        glands: dysmorphology + I
        glands: no defect detected I
        harderian gland abnormalities I
```

Figure 2.9: Sample of phenotype ontology.



```
Stage: TS26
Levels: All

mouse
  embryo
    cavities and their linings
      intraembryonic coelom
        diaphragm
          arcuate ligaments
          central tendon
          dome
          pleuro-pericardial folds
          pleuro-peritoneal folds
        pericardial cavity
          cavity
          mesothelium
        peritoneal cavity
          greater sac
          omental bursa
        pleural cavity
          cavity
          mesothelium
    limb
      forelimb
        arm
          elbow
          forearm
          shoulder
          upper arm
        handplate
          carpus
          digit 1
          digit 2
       ...
```

Figure 2.10: Sample EMAP screen.

```
Comparative anatomy of the stomach
  Stomach (body structure)
  Parent(s):
    Abdominal viscus (body structure)
    Digestive organ (body structure)
    Hollow viscus (body structure)
  Child(ren):
    Avian stomach (body structure)
    Glandular stomach (body structure)
    Non-glandular stomach (body structure)
    Ruminant stomach (body structure)
```

Figure 2.11: Wilcke *et al*'s proposed solution to anthropocentric symbolic models.

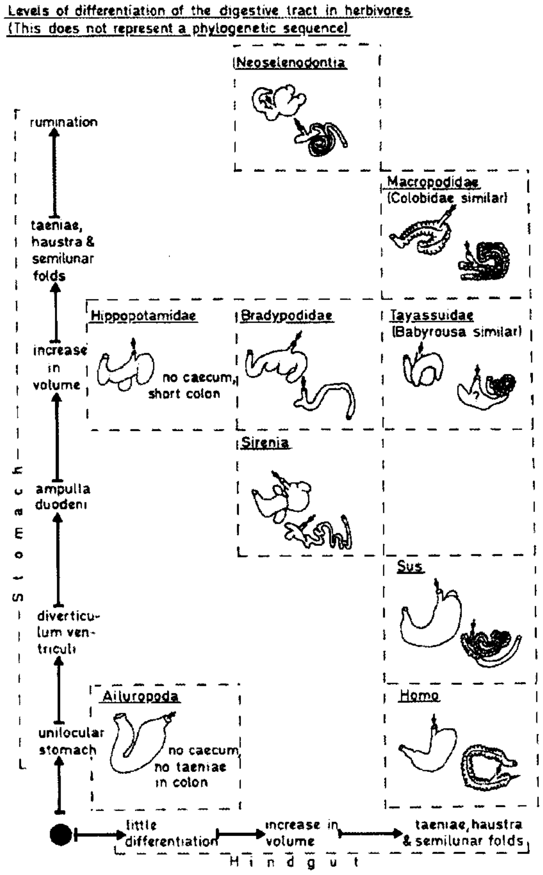


Figure 2.12: Langer's levels of differentiation for mammalian herbivore stomachs.

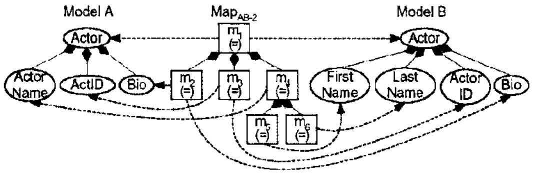


Figure 2.13: Mapping FirstName and LastName as elements of Actor in the mapped model.

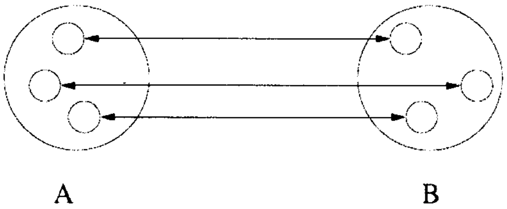


Figure 2.14: A set isomorphism for organ parts of the human (A) and mouse (B) prostates.

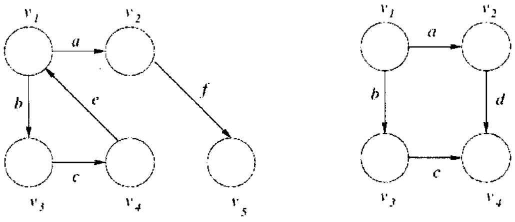
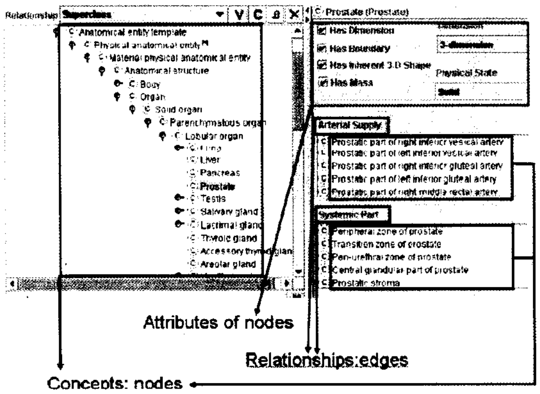


Figure 2.15: Graphs A and B for relational distance comparison.



Relationship Superclass

C Anatomical entity template
C Physical anatomical entity
C Material physical anatomical entity
C Anatomical structure
C Body
C Organ
C Solid organ
C Parenchymatous organ
C Lobular organ
C Lung
C Liver
C Pancreas
C Prostate
C Testis
C Salivary gland
C Lacrimal gland
C Thyroid gland
C Accessory thyroid gland
C Areolar gland

C Prostate (Prostate)

☑ Has Dimension
☑ Has Boundary
☑ Has Inherent 3-D Shape
☑ Has Mass

Dimension: 3-dimension
Physical State: Solid

Arterial Supply
C Prostatic part of right inferior vesical artery
C Prostatic part of left inferior vesical artery
C Prostatic part of right inferior gluteal artery
C Prostatic part of left inferior gluteal artery
C Prostatic part of right middle rectal artery

Systemic Part
C Peripheral zone of prostate
C Transition zone of prostate
C Peri-urethral zone of prostate
C Central glandular part of prostate
C Prostatic stroma

Figure 2.16: FMA entities (nodes), attributes (node attributes), and relationships (edges).

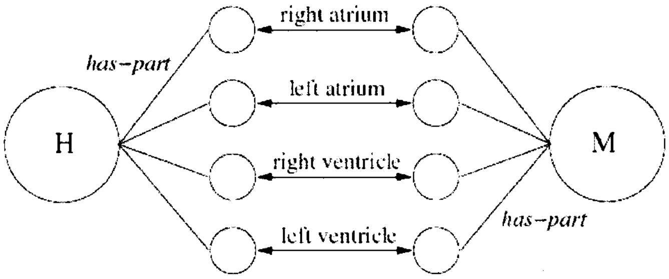


Figure 2.17: Mapping the human heart (H) to the house heart (M).

Chapter 3

COMPARATIVE ANATOMY AND THE STRUCTURAL DIFFERENCE METHOD

The previous chapter introduced the domains and methods that our approach to the problem draws upon. This chapter describes our method for symbolically modeling the similarities and differences between anatomical structures across species, and how this description meets the information needs of different types of users of the system.

This chapter takes what may be metaphorically called a "breadth-first" approach: in presenting the classifications and results which emerged from our research, we present mappings of varied anatomical structures across a number of species. The purpose is to test the limits of our method and resulting classifications under a variety of different conditions. The more different the other species is from the human, the more variety of possibilities will be modeled, and the more the method and classifications are tested. This means that differences among the species will be emphasized in modeling structures for this chapter, and the modeling will be less granular in the interest of covering more ground where differences are likely to be found.

3.1 The Structural Difference Method (SDM)

The *structural difference method (SDM)* is a formalism for representing similarities and differences between anatomical structures across two different species, first introduced in [125], and further developed in [123] and [39]. We use graph isomorphism to illustrate anatomical correspondence and any deviation from isomorphism to represent a difference in the anatomical entities compared. In this way, we can start with an organ, construct the *part-of* hierarchy from the gross anatomical to the cellular level for each species under comparison, and determine the mappings at each level. We call this the structural difference method (SDM).

Isomorphism, or *graph identity*, indicates that there is no difference at a given level of organization; in other words, the mapping between the entities across species is one-to-one and onto. Examples include the `Heart chambers`, the `Lungs` (in mammals), and the mouse and human `Stomachs` at the `Organ` and `Organ part` levels. If two structures are isomorphic at some level of abstraction and resolution, they are identical at that level. But if they are not isomorphic, how do we gauge the difference between two corresponding structures?

Based on our preliminary studies and the relational distance work of Shapiro and Haralick ([108], [107]), we propose the following types of differences for our approach: *node (structure) differences* and *edge (relationship) differences.* Node mappings may be one-to-one and onto (isomorphism), one-to-one but not onto (subgraph isomorphism), one-to-nothing (null mapping), one-to-many, many-to-one, or many-to-many. Furthermore, the edges provide relationship constraints that may or may not be satisfied (edge differences). We illustrate each type of symbolic difference with examples, treating the node differences first, and then proceeding to edge differences.

Node set differences are differences between the number of entities in the source species and the corresponding entities in the target species—in other words, a structure that exists in one species but does not exist in the other species, or it does exist but the correspondences are distributed among a different number of entities than in the source species. Node set differences are illustrated in Figure 3.1.

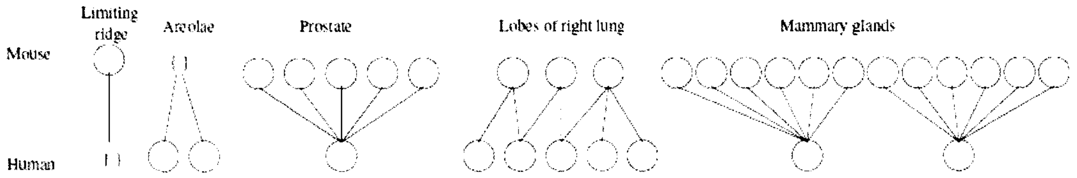


Figure 3.1: Node set differences for various structures in the human and the mouse.

Examples of such mapping differences include null mappings, which may be one-to-zero (one mouse limiting ridge to none in the human, discussed below) or many-to-zero (two areolae of breast in the human to none in the mouse mammary glands). Null mappings for structures in the human breast and the mouse mammary gland are illustrated in Figure 3.2.

Additionally, there are mappings that may be one-to-n (one human prostate organ to five mouse organs), or n-to-m (three lobes of the human right lung to five lobes of the mouse right lung; two mammary glands in the human to twelve in the mouse). The 1:5 mapping between the human prostate and the mouse prostate organs are illustrated in Figure 3.3.

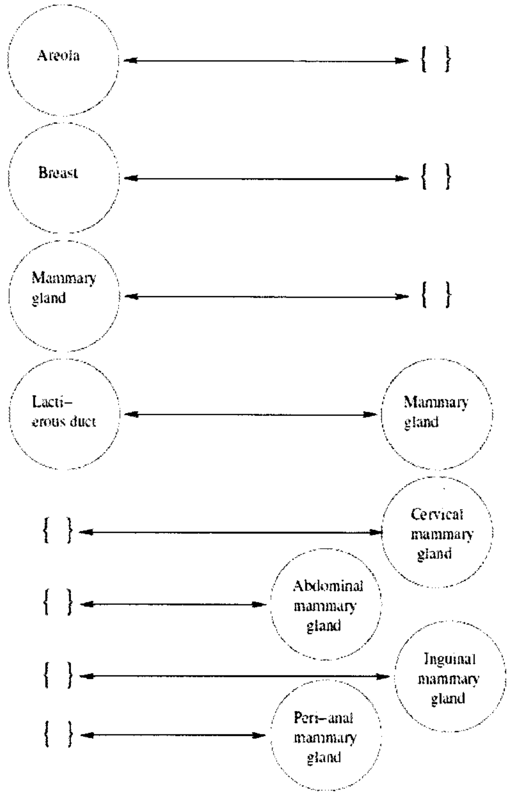


Figure 3.2: Null mappings in gross anatomical mammary structures found in humans and mice.

Node attribute differences (Figure 3.4) are differences in the existence of an attribute between two corresponding structures in the source and target species—in other words, the structure exists in each species, but it occupies a different place in the AT, and thus, the slots required for a *sound* and *complete* description of the structure differ across species. For example, *has-member* (which is a specialization of the *partonomic* relationship constrained

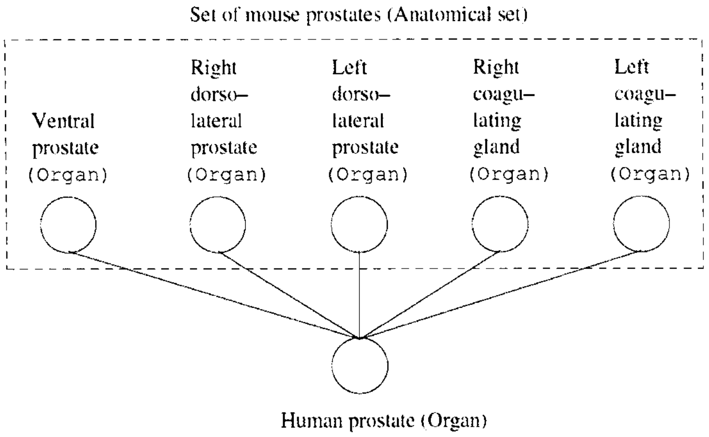

Set of mouse prostates (Anatomical set)
Ventral prostate (Organ)
Right dorso-lateral prostate (Organ)
Left dorso-lateral prostate (Organ)
Right coagu-lating gland (Organ)
Left coagu-lating gland (Organ)
Human prostate (Organ)

Figure 3.3: The 1:5 correspondence between the human and the mouse `Prostates` at the `Organ` level.

in the FMA to `Anatomical sets`) is an attribute of the node `Set of mouse prostates`. In this partonomic scheme, `Anatomical set` is made up of member `Organs`. In the human, the prostate is a single organ. The class `Organ`, however, lacks the attribute *has-member*, and therefore a node attribute difference exists between the `Prostates` of the two species. This category of differences is necessary, because it is the only explicit way of acknowledging the difference in roles of the different structures in the AT. In accordance with Stevens' principle that the parameters of a measurement system be exhaustive and mutually exclusive ([115]), these attributes are necessary to fully describe the structure and its anatomical role. To correspond to another kind of structure in the AT is to lose those specific attributes of its role in the other species, as well as to gain other attributes, and this category of differences accounts for that shift in anatomical role across species.

In [122], we proposed *vestigial* as an attribute of an anatomical structure, rather than as a separate class in the AT. Our reasoning at the time was that since vestigial structures are brought about by the same epigenetic and genetic processes as their retained homologues, that to move them to a separate class, as proposed by Rosse [271], would artificially magnify the differences between them. For example, Hildebrand asserts that a 19 m whale has a 4 cm vestigial femur ([49]). Despite the fact that the femur exists (although minimally), the phe-

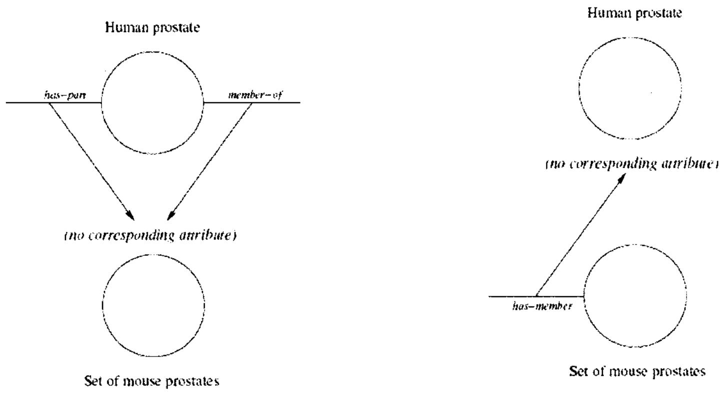


Figure 3.4: Node attribute value differences.

notype of the whale is legless. We asserted that the graph representation of the comparison of the human and the whale femur should both show the existence of a `Femur` in relation to the `Pelvis` (isomorphism at the level of existence of `Femur`); the specific differences should emerge in the missing `Shaft`, `Distal head`, and other cetacean femoral structures. We argued that to move the whale femur to another class, as entailed by Rosse's call for a `Vestigial anatomical structure class`, would artificially add graph distance to the CAIS representation, and so we proposed that *vestigial* should be considered an attribute of a structure, rather than an entirely different class of structure. In light of the information gathered by the domain experts in the course of this dissertation, we now regard that proposal as hasty. Our current understanding is that the decision of the correct way to classify vestigial structures should be informed by the modeling of the evolutionary transformation processes involved. Since that modeling is a part of recommended future work, at this point we make no recommendation on the appropriate representation of vestigial structures.

Node attribute value differences are differences in values of corresponding attributes shared between corresponding nodes of two species—in other words, the structure exists in both species, and (to some extent) shares an anatomical role, but there is some difference in the values of its attributes from one species to the other. For example, an isomorphism exists between the mouse (or rat) and human Stomachs at the levels of whole Organ and Organ part: the mapping is one-to-one and onto for {Fundus of stomach, Body of stomach, Pyloric antrum}. The isomorphism propagates to the next level, namely, the Stomach wall, the parts of which are: Mucosa (GM), Submucosa (SM), Muscularis (M) and Serosa (S). The difference between mouse and human begins to emerge in the attribute values for the node Mucosa. Unlike the body of the human stomach (HS), which is lined throughout by the Glandular mucosa (GM), the Mucosa of the Body of the (mouse) stomach (MS) is divided into two structurally-different regions: Glandular mucosa (GM) and Non-glandular mucosa (NM). GM and NM are demarcated from one another by the Limiting ridge (LR), which has no corresponding node in the human ([99]).

Figure 3.5 depicts both node attribute value differences and node set differences. The mappings involving the Serosa, Submucosa, and Muscularis are isomorphisms, indicated by the two-headed arrows. The Mucosa, however, is not isomorphic across species: in the human its attribute value is "glandular", whereas in the mouse the values are "glandular" and "non-glandular"[1]. The dashed line represents a mapping between nodes with different values for the same attributes. Additionally, there is no corresponding structure for the Limiting ridge in the human: the difference in node mapping is represented by the dotted line. This is an example of a null mapping, and the non-existent structure is represented by the empty set notation {}.

Edge set differences are differences in the existence of relationships (edges) between structures across species. For example, the *dorsolateral prostates* of the mouse are adjacent to the coagulating glands, which do not exist as organs in the human. Another example is the inguinal mammary glands of the mouse, which are adjacent to the inguinal ligament, whereas the human mammary glands are adjacent only to the pectoralis major muscle. Because they are located in different places in the body in different species, the spatial relationships (such as *continuous-with* or *adjacent-to*) among the anatomical entities are

[1] Here we gloss over the issue of whether Mucosa is the appropriate term for the non-glandular region of the rodent stomach; the rodent literature is approximately evenly divided among authors who use the term Nonglandular mucosa and those who use Nonglandular region or Nonglandular part. For the sake of simplicity in comparison, we use the term Nonglandular mucosa as it is widely used in the literature, while stipulating that the term is indeed problematic.

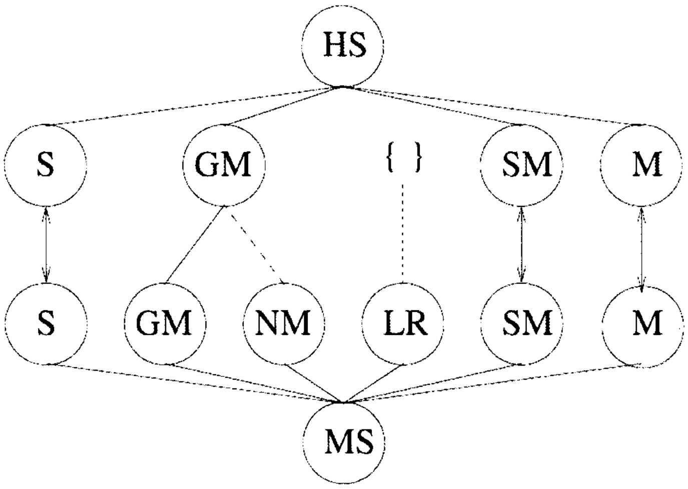


Figure 3.5: Node set and node attribute value differences between the human and rodent stomachs.

changed, and this change is reflected in the relationship differences across species.

Edge attribute value differences are differences in the attributes of existing relationships between structures across species. In the same way that nodes can have attributes, edges can as well, and the differences between those attributes can also be expressed symbolically.

There is an asymmetry between the number of node differences and the number of edge differences, due to the lack of *edge attribute differences*, which would correspond to node attribute differences. This category of edge difference does not exist, because there is no hierarchy of spatial relationships to correspond to the structural hierarchy in the AT.

3.1.1 Other vertebrates

Because of their longer evolutionary history earlier (more basal) vertebrates are a potentially very rich source of anatomical difference for testing the SDM. Intuitively, it would seem that the longer the evolutionary distance between species, the more time they have had to evolve significant differences from each other. Although this is not an absolute rule, the `Pituitary gland`, viewed from the earliest vertebrates through to mammals, bears out that intuition, and provides a useful test case for the SDM.

Figure 3.6 is a phylogenetic tree of the structures and spatial relations (ASA) of the Pituitary from cyclostomes (hagfish and lamprey) through sharks, bony fishes, lungfishes, amphibians, reptiles, birds, and mammals. The different parts of the Pituitary (Anterior pituitary, Posterior pituitary) and the relevant parts of the lower Brain (Median eminence, Third ventricle) are represented in the differently-shaded sections. Additionally, for the first time, we explore the application of the SDM to the ASA, specifically to attributed relationships, in order to determine whether the method is robust enough to adequately represent those relationships. The source for Figure 3.6 is *The Encyclopedia of Endocrinology* after Gorbman's illustrations.

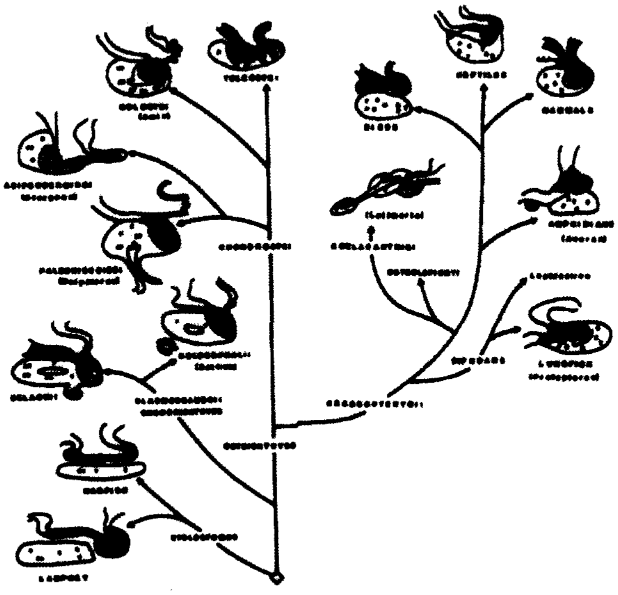


Figure 3.6: Variations in spatial relations among the parts of the vertebrate pituitary.

Note that the Anterior pituitary (white rectangle) is totally separated from the Median eminence (hatched) and the Posterior pituitary (black) in the lamprey and the hagfish. In the sharks and bony fishes, we see them begin to come into contact (the *continuous-with* attribute of the *adjacency* relationship in the FMA, written from this point as *adjacency:contiguous-with*), and then begin to interdigitate and penetrate each other, leading to richer vascularization as we move "up" the phylogenetic tree. By the time we reach birds, they are distinct from each other, yet communicating through vascularization (*supplies:arterial-supply*, *supplies:venous-supply*, *supplies:capillary-supply*) in the FMA. Yet all of these different conditions fall into the edge-differences we have delineated—edge-set differences, and differences in attributed edges as well. By contrast, in mammals, the `Anterior pituitary` and `Posterior pituitary` are fused (one node/organ, rather than two, as in birds: a 1:2 Node-set mapping), and so our node-set differences apply. Again, the SDM we have proposed is sufficient to handle variation as significant as that demonstrated by the `Pituitary` across 400 million years of vertebrate evolution [59].

The hagfish and the lamprey are the earliest extant vertebrates: representations of their relevant structures appear in Figure 3.7 and Figure 3.8. A couple of details about the diagrams need to be noted: first, the structures are not limited to only the `Pituitary` itself, but include the lower part of the `Brain` (the `Hypothalamus`), as well as the vasculature and innervation between those structures. So this is more accurately described as a model of the `Hypothalamo-pituitary complex` across vertebrates.

Second, the discipline of endocrinology has a couple of centuries of history of experimentation, and has had time to develop multiple terms for the same entity, depending on the structure and the species. In the literature, and in our examples, `Adenohypophysis` and `Anterior pituitary` are synonyms for each other, as are `Neurohypophysis` and `Posterior pituitary`. As we mentioned earlier in the discussion of the mouse prostate, the FMA, and thus our method, is capable of handling these synonyms, because we model entities, rather than only terms.

Figure 3.7 is a representation of the hagfish `Pituitary` and related structures and relationships. Although later in the phylogenetic tree, as we have seen, they fuse into a single `Lobular organ` (as for humans in the FMA), in the hagfish the `Adenohypophysis` and the `Neurohypophysis` are separate `Organs`, and in fact do not even touch each other (*adjacency:adjacent-to*, as opposed to *adjacency:contiguous-with*). The `Neurohypophysis` is, however, *adjacency:contiguous-with* the `Third ventricle` of the `Brain`, a relationship

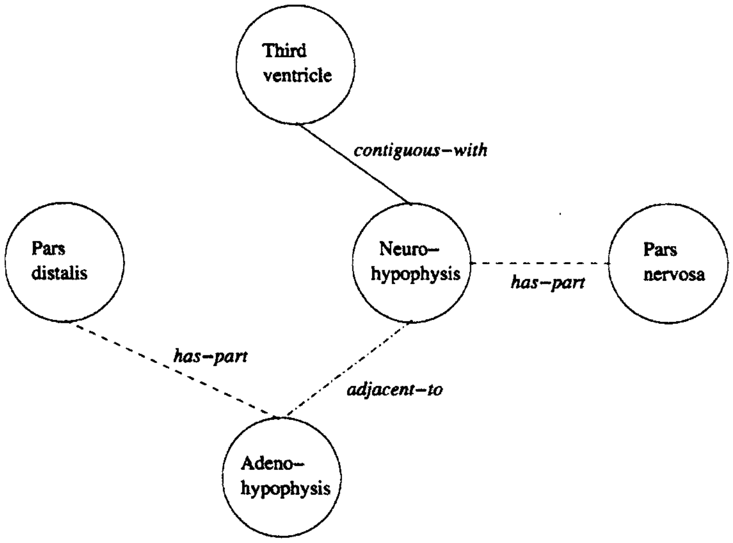


Figure 3.7: Hypothalamo-pituitary structures and relationships in the hagfish.

which will remain constant as we proceed through the tree. The `Neurohypophysis` has only one part in this species, the `Pars nervosa`, and the `Adenohypophysis` has only the `Pars distalis`. Although having only one part would seem to indicate synonymy between the two entities in question, in later species we will encounter more differentiation of these parts, along with the associated node and edge set differences; maintaining the *part-of* relationship, even for only one part, keeps the integrity of the ontology for adding later structures and relationships to it for later species.

Although along with the hagfish, the lamprey is one of the earliest and most basal vertebrates (*Agnathans*, or jawless vertebrates), we can see that these structures in the lamprey already exhibit more complexity than in the hagfish.

The `Third ventricle`, `Neurohypophysis`, and `Pars nervosa` remain essentially the same as in the hagfish, but both the `Adenohypophysis` and the `Pars distalis` exhibit more differentiation, the `Adenohypophysis` acquiring the part `Pars intermedia`, and the `Pars distalis` differentiating into two histologically-distinct zones. (We are modeling only to the level of cellular granularity for this example: modeling hormone products would produce

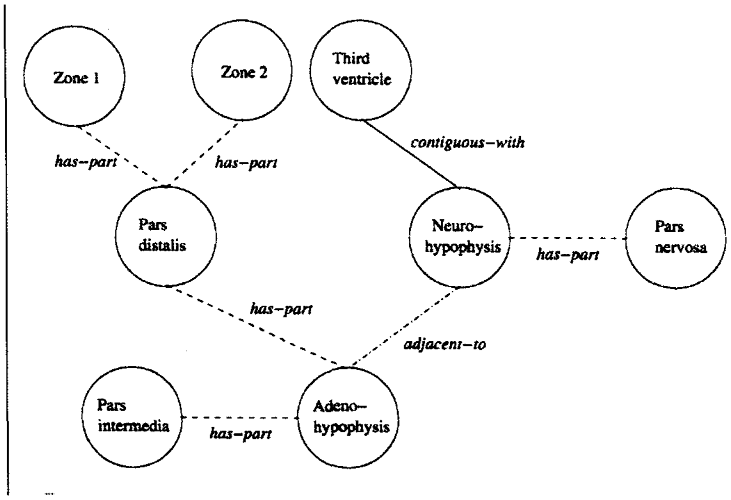


Figure 3.8: Hypothalamo-pituitary structures and relationships in the lamprey.

even more node and edge-set differences.) Thus we have node and edge-set differences already, even in a comparison of two of the earliest living vertebrates.

Moving to the *holocephalans* (represented by the chimaera), a type of cartilaginous fish, even more differences become apparent, as in Figure 3.9. The `Third ventricle-Neurohypophysis` relationship remains constant, but already we see a great deal of change in the structures and relationships of the `Neurohypophysis`. It has acquired a second part (another node), the `Median eminence`, which *adjacency:surrounds* the `Portal blood vessels` which *supply:capillary-supply* the `Neurohypophysis` (edge-set differences). Additionally, it has acquired another part, the `Saccus vasculosus`, with the associated node and edge differences.

The `Adenohypophysis` at the `Organ` level is isomorphic to the lamprey `Adenohypophysis`, but at the `Organ subdivision` level, it continues to undergo differentiation, generating additional symbolic differences. It retains the histological zones acquired by the lamprey, and additionally has developed the regional parts `Rostral pars distalis` and `Proximal pars distalis`. It now *adjacency:surrounds* a `Cavity`, and there is a unique structure referred to in the literature by its German name, `Rachendachhypophyse` (pharyngeal roof pituitary), *adjacency:exterior-to* `Cranium`, *adjacency:inferior-to* `Pharyngeal mucosa`, and

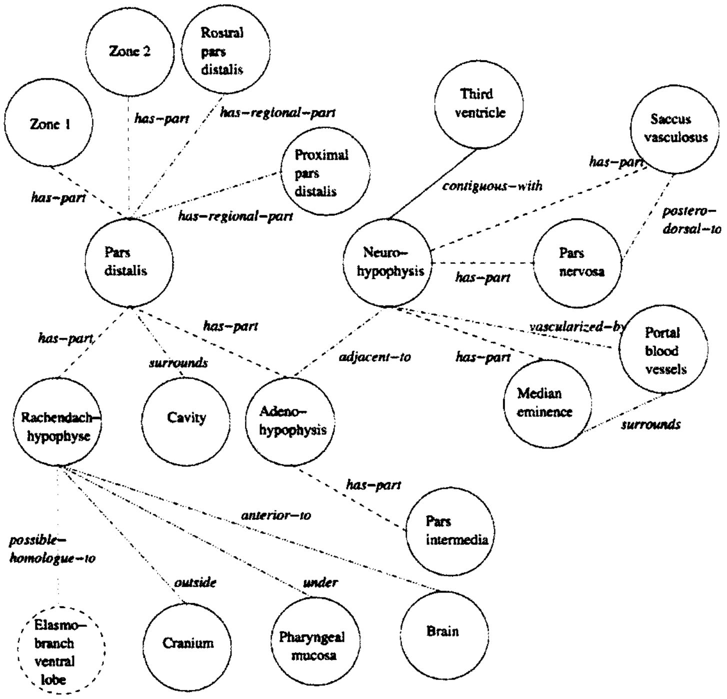


Figure 3.9: Hypothalamo-pituitary structures and relationships in the chimaera.

adjacency:anterior-to `Brain`.

So far, we clearly have multiple node-set and edge-set differences associated with the transition from jawless vertebrates to early jawed fishes. There is one node and one edge outlined with dashes and a dotted line to indicate an ambiguity—when we examine the *elasmobranchs*, we will find that there is a unique structure called the `Ventral lobe of the pars distalis`. Until the homology of the `Ventral lobe of the pars distalis` and the `Rachendachhypophyse` is either definitively established or ruled out, there is a risk of being off by one node, as well as the associated edges. If we count non-homologues as purported homologues, the cardinality of our node-set differences is one less than the true cardinality for each such difference, and the cardinality of our edge-set differences is off negatively by the number of associated edges. Similarly, if we count homologous structures as non-homologues, the cardinality of our node-set differences is one more than the true cardinality for each such difference, and the cardinality of our edge-set differences is again off—this time positively—by the number of associated edges[2].

The hypothalamo-pituitary axis of the coelacanth is modeled in Figure 3.10; it continues to exhibit the same sort of node and edge set differences that we have seen so far. Again, the structure `Rostral pars distalis` is a possible homologue of the holocephalan `Rachendachhypophyse`; we have indicated it as ambiguous by means of a dashed outline.

As we progress through the phylogenetic tree, the diagrams become more and more complex; for the sake of space, we have stopped including the diagrams at this point. To summarize our findings, the SDM proved sufficient for modeling every interesting trend in the evolution of the `Hypothalamo-pituitary complex` all the way through the phylogenetic tree: the replacement of *innervation* of the `Neurohypophysis` by *vascularization* as animals moved from the sea and being surrounded by water containing hormones to become more complex land animals who no longer could get these hormones from the air, and needed a corresponding delivery system; the gain and loss of certain structures (*e.g.*, birds don't have a `Pars intermedia` and neither do adult humans, while the phylogenetically closer-to-humans adult cats so); and the fusion of the two organs `Adenohypophysis` and `Neurohypophysis` (in everyone before mammals) into one *Lobular organ* characteristic of mammals. Additionally, the SDM proved robust enough to handle the ambiguity of structures which are presumed homologous, but whose ultimate disposition has not yet been definitively established.

The totality of these descriptions of differences constitutes the Structural Difference Method. Using the SDM, we have already carried out mappings between the human and the mouse for the mammary gland, prostate, ovary, cervix, and lung. Additionally, we applied

[2] The importance of this is twofold: first, the purpose of the SDM is to describe soundly and completely the difference between the anatomies of two species. This discrepancy in the Node-set differences is a threat to the integrity of that description. Second, although the implications of this fact are outside the scope of this dissertation, this issue once again highlights the problem of missing and conflicting information regarding taxic homology. The scope of this problem has implications for inheritance, rendering strictly monotonic inheritance unfeasible—there are too many gaps and conflicting authorities to safely assume monotonic inheritance of properties from one structure to its descendants. As will be mentioned briefly later, birds and adult humans have no Pars intermedia, while adult cats do. So in this regard, humans are more like birds than the phylogenetically much closer mammal (cats). Clearly, this is not a simple case of monotonic inheritance of development of the structure, and in this regard, like the discrepancies in the number of mammary glands or prostates between closely-related mammals, indicates a potentially very interesting unsolved question in comparative anatomy.

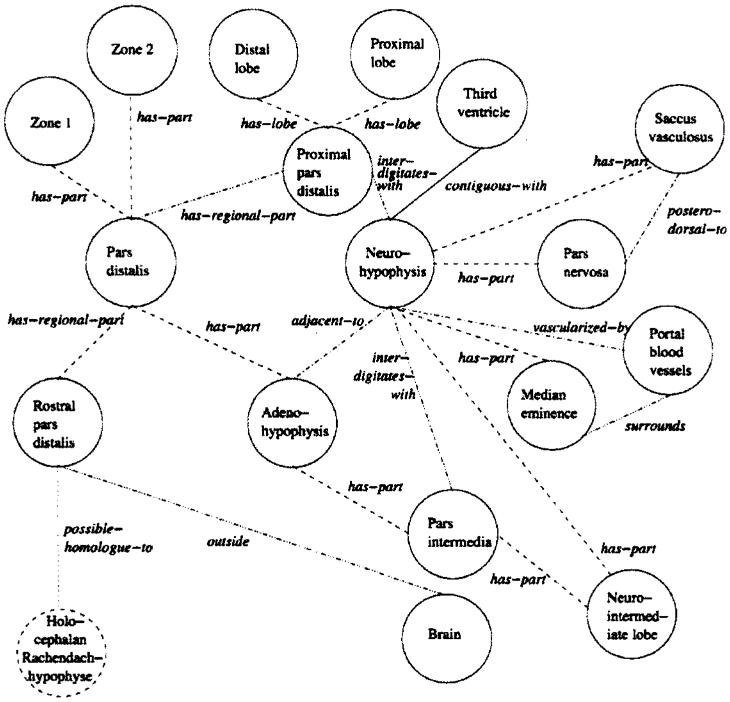


Figure 3.10: Hypothalamo-pituitary structures and relationships in the coelacanth.

the model to a problem in conservation biology, and were able to clarify the classification of sun bear vaginal epithelial cells (in the "formalization improves conceptualization" approach mentioned in Chapter 2), and to clarify the information space the SDM operates in (the intersection of the ranges of "Normal" for each species, and the corresponding need for an ordinal measurement capacity to determine that range).

3.2 Summary

In this chapter, we provided examples of modeling anatomical structures and spatial and anatomic relations among those structure, based on the FMA. We then applied the SDM to determine whether it was sufficient to handle the range of cross-species anatomic variation provided by the examples. The examples were drawn from our previous work in modeling the `Prostate` and `Mammary gland` in humans and mice, and from the `Hypothalamo-pituitary` `complex` from hagfish to mammals. The number and types of differences in the species compared grew greater as we moved from intra-mammalian comparision for `Prostate` and `Mammary gland` to pan-vertebrate comparisons for the `Hypothalamo-pituitary complex`. On a preliminary basis, we have carried out comparisons between different taxa, at varying levels of detail, and from the point of view of differing medical disciplines, such as endocrinology. While more thorough validation is necessary for our model, it is nevertheless encouraging that no significant obstacles were encountered as we surveyed such a wide range of topics. More work needs to be done in evaluating the SDM, but it seems on a preliminary basis to be well-equipped to deal with the kinds of differences and similarities this range of examples provided.

In carrying out the mouse mappings, we had to first resolve the problems we encountered in the literature. The non-equivalence of entities was a major problem that we encountered in the first symbolic models based on the mouse. Because the FMA was developed based on the human, and because the human is such an exception from other mammals in so many attributes, there is a great deal of terminology that is not part of the human FMA, but is needed for the appropriate representation of structures in other species. This fact made it necessary to develop new regional terms to extend existing FMA terms for our murine symbolic models.

In order to address the issues we encountered to develop the models of the mouse organs, we performed the following steps:

- developed a standard of preferred existing terms as validated by mouse anatomists and pathologists;

- established consistent and systematic regional terms for organs with no human matching term;

- incorporated these terms and definitions into ontologies for mouse mammary gland and mouse prostate in a separate database from the human FMA;

- identified gaps in the literature on spatial relationships among mouse structures;

- carried out comparisons of different structures across species at varying levels of granularity.

We developed the following categories for classifying anatomical difference across species according to the SDM:

- Node differences
  - Node set differences
  - Node attribute differences
  - Node attribute value differences
- Edge differences
- Edge set differences
- Edge attribute value differences

We applied the SDM to a real and current problem in conservation biology, and established that:

- the SDM can inform the research by illuminating the gaps and inconsistencies in current biological knowledge which act as an obstacle to principled modeling;
- the research can inform the SDM by providing real-life examples of where constraints and conditions need to be added to the original specification.

Chapter 4

DESIGN OF THE COMPARATIVE ANATOMY INFORMATION SYSTEM (CAIS)

The previous chapter introduced the structural difference method and its classification of differences in anatomical structure. This chapter provides a description of the design of anatomical mappings and other design considerations, as well as of implementation of the knowledge base in Protégé-2000.

This chapter takes what may be metaphorically called a "depth-first" approach: in presenting the classifications and results which emerged from our research, we present mappings of selected human anatomical structures in mice and rats. The purpose is to create a proof-of-concept implementation of a knowledge base that can be useful in a real-world research situation. The research that underlies this knowledge base relies upon the relevance of animal models to human disease. This means that similarities among the species—while neither emphasized nor deprecated in modeling structures—will nevertheless be better represented in this chapter as a consequence of the choice of species and their appropriateness as animal models. The modeling will be more granular in the interest of providing a prototype knowledge base that is populated enough to be useful.

4.1 Introduction

In previous work [319], we proposed an approach to correlating the anatomy of *Homo sapiens* with selected species, using the Foundational Model of Anatomy (FMA) as a framework, and graph matching as a method, for determining similarities and differences in the nodes and relationships (edges) defined by the attributed graph of the FMA. In addition, we hypothesized that the frame-based ontology of the FMA furnishes a comprehensive set of concepts and relationships for correlating human anatomy, at all levels of structural organization, with the anatomy of any mammalian or vertebrate species. In this way, our method can serve as a basis for navigating the rapidly emerging databases and knowledge bases that are evolving as reusable resources in bioinformatics. This chapter describes a comparative anatomy information system between *Homo sapiens*, *Rattus norvegicus*, and *Mus musculus*, which will serve as a pilot project for cross-species anatomical information collection, storage, and retrieval. The underlying data structure of a mapping, and the syntax and semantics of the system's query language are presented.

4.2 Components of the Proposed Information System

Our comparative anatomy information system (CAIS) accepts queries posed by the user about similarities and differences in human and mouse anatomy. The implementation of this version of the comparative anatomy system is a single database of mappings, from which the query engine accesses and returns a result set. Automatic and dynamic generation of mappings from separate databases by species is a possible future goal of this research, but is specifically outside the scope of this version of the project.

The anatomical mapping data structure and the syntax and semantics of the system's query language are particularly significant, and will be discussed in more detail below.

4.3 Anatomical Mapping

Mappings are the data structure at the heart of the proposed information system. As developed in [319], there are two main kinds of mapping classess: `Node mappings` and `Edge mappings`, corresponding to the components of the directed graph described by the FMA. The structures which are mapped across species are selected on the basis of homology (evolutionary relatedness); homoplasy (similarity of appearance) and analogy (similarity of function) are not considered in creating mappings. `Node mappings` are further divided into `Node set mappings`, `Node attribute mappings`, and `Node attribute value mappings`, and `Edge mappings` are further divided into `Edge set mappings` and `Edge attribute value mappings`.

At a conceptual level, a `Mapping` across `Species` between `Anatomical structures` can be represented as in Figure 4.1, which shows `Mappings` between the human and mouse `Prostates` at the `Organ` level. The edges of the graph in green represent isomorphisms, or anatomical identity: one-to-one, onto, and structure-preserving. For example, the anatomical abstraction Lobular organ in the mouse is isomorphic, or identical, to the Lobular organ in the human. The edges of the graph in blue represent non-isomorphic similar matches. For example, there is a 5:1 mapping between the different mouse prostate organs and the single human prostate. The edges in red represent null mappings. For example, there is no corresponding Set of human prostates to map to the Set of mouse prostates, so that constitutes a null mapping.

The underlying Mapping data structure contains pointers in both directions between species: *i.e.*, the human can be either the source or the target species, as can the mouse or rat. Both directions are necessary for a complete answer to queries on similarities and differences between species, as, from the user's point of view, the answer returned to the query "what is the difference between the human and mouse (or rat) prostates?" should be the same as the answer returned to the query "what is the difference between the mouse (or rat) and human prostates?". This data structure provides that consistency of response, yet at the same time allows a more refined query to return a more granular answer, depending on the level of detail the user wishes to specify. Although the usual query will be bidirectional, there will be users who want information in one direction only. For example, a user may want to know what prostatic zone in the human is homologous to the murine dorsal prostate. This structure is able to accommodate those queries as well.

The examples for each type of Mapping are taken from [317]. As a class, Mappings are *first-class* objects (*cf.* Pottinger and Bernstein), and can thus undergo the same operations as the models from which they are derived. Mappings are thus objects comprised of two species-specific anatomical structures and the Mapping relationship between them. They correspond to, for example, a mouse node, a human node, and the edge between those nodes in Figure 4.1, or to one rectangle in Figure 4.2.

Mappings are implemented in Protégé in the following manner: the Protégé template slots for Mapping are the two Species being compared, and the two corresponding Anatomical structures. Most of the time, due to our appreciation of real similarities consequent upon the vertebrate *Bauplan*, the structures will have the same name across species (Prostate (mouse) and Prostate (human)), but not always (*cf.* Oviduct (shark) and Fallopian tube (human)). Species names are required to always be single; Anatomical

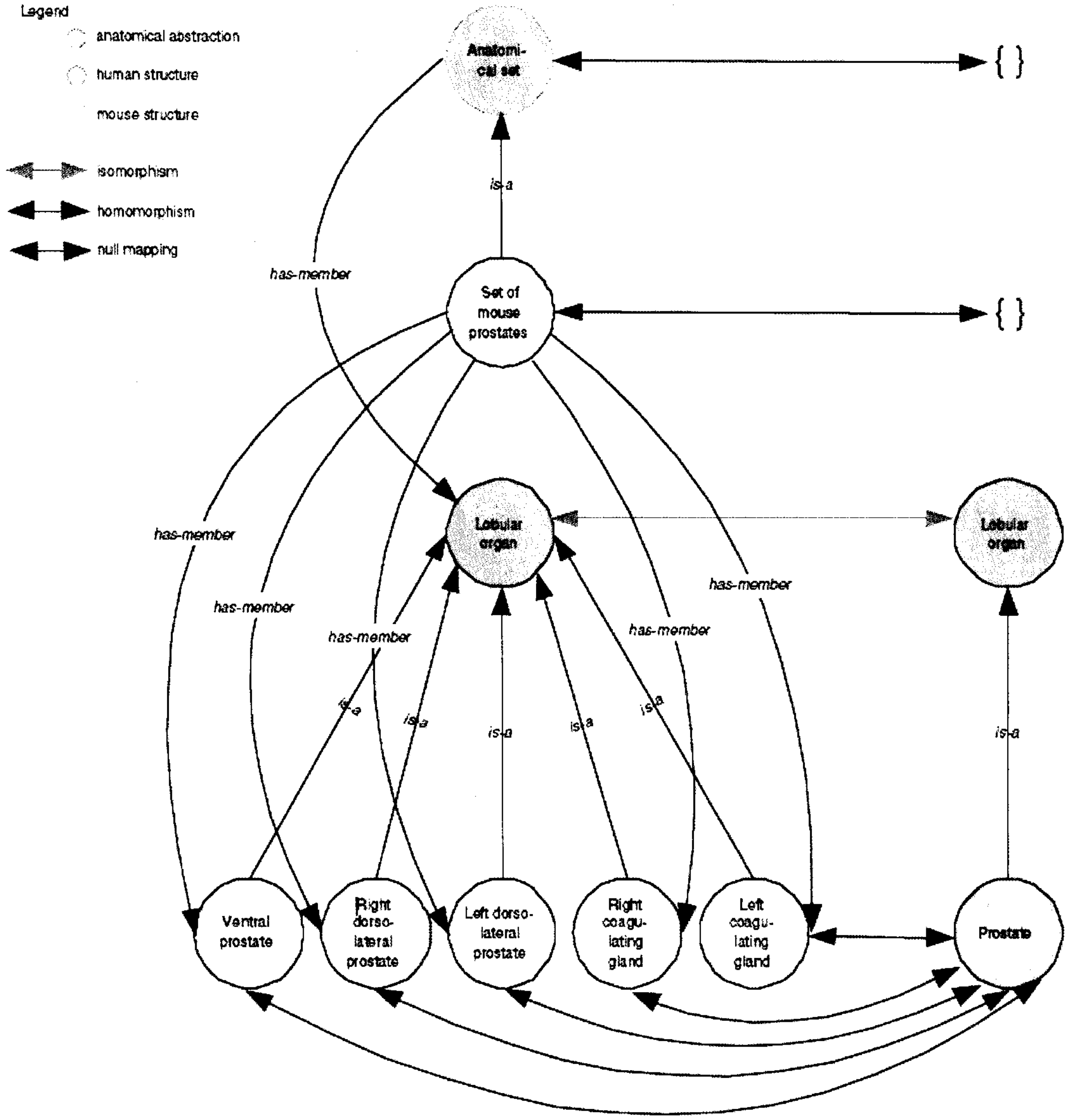


Figure 4.1: Conceptual mapping between the human and mouse prostates.

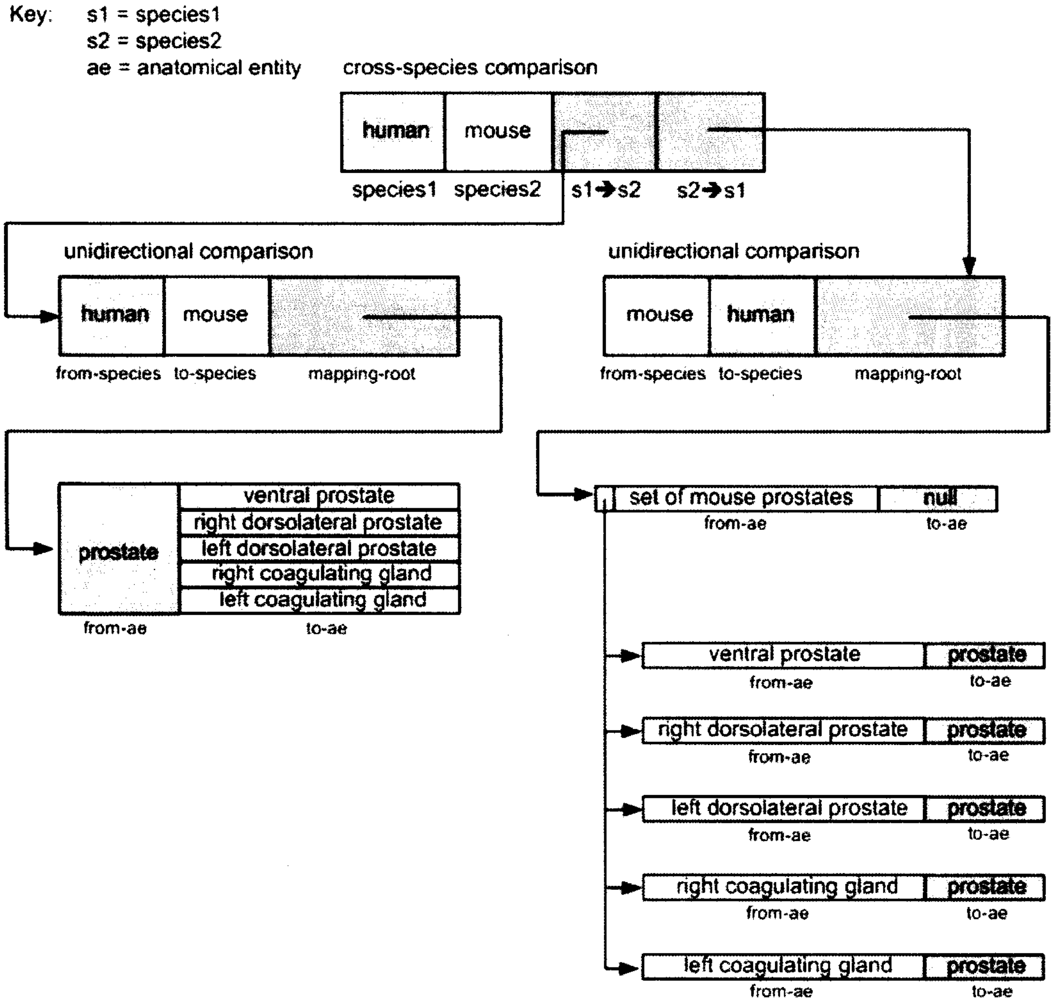


Figure 4.2: Abstraction of the data structure to be used to represent a cross-species comparison for the human and mouse prostates.

`structures` can be 1 or more in a particular `Species`. *Cardinality* specifies whether the correspondence is 1:null, null:1, 1:1, 1:many, many:1, many:many, many:null, or null:many.

The slots for `Node mapping` are inherited verbatim from the `Mapping` class, and `Edge mappings` have the additional slot *Relationship* to describe which FMA ASA slot is being compared across `Species` for the given `Anatomical structures`. These examples demonstrate the definitions for the different kinds of `Mappings`. A `Cross-species comparison` is made up of all the `Mappings` of the `Anatomical structure` at the level under comparison. We use these structures to return answers to anatomical queries about similarities and differences between these structures—the `Mapping` contains the information about similarity and differences of particular discrete structures, and the `Cross-species comparison` provides the context (hierarchy) for those structures in relation to other anatomical structures.

4.4 Syntax and semantics of the query language

For the purpose of defining this comparative anatomy information system, it is useful to draw a distinction between different kinds of queries, based on how many models the system handles at a time. These classifications will specify what types of queries our system handles, and what is outside its scope. We define the classification of a query as follows:

Single-species queries hold for species models taken one at a time. For example, in the human, the `Heart` is *inside* the `Thoracic cavity`, so the query "what is the relationship between `Heart` and `Thoracic cavity` [*implied: in the human*]?" is a single-species query. Note that a single-species query can be simple or compound; the classification of the query refers to the number of species models participating in the query, NOT to the complexity of the query. Single-species queries are the basis of queries in the FMA using *Emily* [93], and involve existence, location, connectivity, and similar features of anatomical structures. Single-species queries are not implemented in our current CAIS system.

Two-species queries hold for species models taken two at a time, and are the basis of what is unique about our CAIS system. They involve comparisons between anatomical structures across two different species and are the main difference between the proposed system and *Emily*. For example, the query "how is the human prostate different from the mouse prostate?" is a two-species query. Two-species queries involve similarity, difference, homology, identity, and synonymy of anatomical structures in two different species, as described below. While the concepts of homology, identity, and synonymy overlap to some degree in natural language, the syntax below suffices to deal with them at the level of the users' needs. Higher-degree queries (as well as sex and stage of development [3]) represent future work, and will explicitly be omitted from this specification. We developed the syntax for two-species queries, as follows.

4.4.1 Syntax

The following BNF syntax represents a textual abstraction of our allowable cross-species queries. In the introductory paragraph, we described our proposed system as correlating phenotypes across species at varying levels of granularity and detail; these parameters will control that refinement of the query. For example, the parameters (development of which is included in our future work) will determine whether similarity or difference is being assessed at the `Organ` level, the `Organ part` level, the `Tissue` or `Cell` level, or at some other level of resolution. Relevant levels for our prostate example would include `Organ`, `Lobe`, `Zone`, and `Tissue` levels, for example.

$< query >::=< concept1 >< relationship >< concept2 > (< parameters >)$

$< concept1 >::=< species1 >< anat.ent1 > \ | unknown | \ < result - set >$

$< concept2 >::=< species2 >< anat.ent2 > \ | unknown | \ < result - set >$

$< species1 >::=< name - of - species >$

$< species2 >::=< name - of - species >$

$< anat.ent1 >::=< name - of - anatomical - entity >$

$< anat.ent2 >::=< name - of - anatomical - entity >$

Species1 and *species2* can both be either human or mouse or rat; *anat.ent1* and *anat.ent2* can be any of the anatomical structures specified earlier or any of their parts.

Including the FMA relationships as allowable queries makes future work possible in extending the system to higher-order combinations of models ($n > 2$, where $n =$ the number of species being compared) and metamodels (*e.g.*, `Mammal`, `Rodent`, `Vertebrate`), as well as to compound and complex queries. By incorporating lower-order relationships in each succeeding type of comparison, backwards compatibility is preserved, and emerging patterns in the relationships are not prematurely ruled out by disallowing earlier relationships.

At the same time, it is necessary to point out that the type of query posed here—simple, two-species queries—may be considered as a *degenerate* case of the higher-order queries which remain in our plans for future work. So because we do not rule out any FMA relationships in our system, the possibility of queries such as "Is the heart of the mouse adjacent to the liver of the human?" remains. While such a query is semantically absurd on its face, it is syntactically well-formed, and the answer is "no". More important, by permitting such seemingly nonsense queries at this level, the door remains open for more complex queries, such as "Is (***the structure which ultimately becomes the Head kidney in the Flounder***) adjacent to (***the structure which ultimately becomes the left adrenal gland in the Mammal***) in the `proto-Vertebrate`?", in our future work. This, in turn, ensures that the usefulness of our system is not limited to humans, mice, and rats, but in fact can be used to compare the anatomy of any species to that of any other species.

We use this syntax as the basis for queries and responses about anatomical similarities and differences between the human and the mouse. This notation represents an abstraction of the basis for the queries and responses; there is a low-level syntax that is used by the system for accessing and returning information, as well as a higher-level graphical user interface for the users of the system.

4.4.2 Semantics

Queries are of two major types, set queries and Boolean queries. Boolean queries return `T` or `F` when the user queries whether structures in two different species map to each other. Set queries return result sets, such as the set of shared mappings between two species for a structure at a given level of granularity. The semantics of the proposed operators are as follows.

Set queries

The set query operators are *differs-from*, *similar-to*, *shared*, *not-shared*, and *union*.

- `species1.anatomical-entity1` *differs-from* `species2.anatomical-entity2` returns the difference between anatomical-entity1 in species1 and anatomical-entity2 in species2 as computed by the structural difference method (SDM). If anatomical-entity1 and anatomical-entity2 are isomorphic, it will return null.

- species1.anatomical-entity1 *similar-to* species2.anatomical-entity2 returns the complement of the set returned by (species1.anatomical-entity1 *differs-from* species2.anatomical-entity2), which is all of the similarities between species1. anatomical-entity1 and species2.anatomical-entity2 as computed by the SDM.
- species1 *shared* species2 returns the set of non-null mappings between anatomical entities of species1 and those of species2.
- species1 *not-shared* species2 returns the set of null mappings between anatomical entities of species1 and those of species2. In other words, it is the inverse operation of *shared.*
- species1 *union* species2 returns the set of all (null as well as non-null) mappings between anatomical entities of species1 and those of species2.

Boolean queries

The Boolean query operators are *is-different?* and *is-homologous?*.

- species1.anatomical-entity1 *is-different?* species2.anatomical-entity2 returns T if species1. anatomical-entity1 does not map to species2.anatomical-entity2, and F if the two anatomical entities do map to each other.
- species1.anatomical-entity1 *is-homologous?* species2.anatomical-entity2 returns F if species1.anatomical-entity1 does not map to species2.anatomical-entity2, and T if the two anatomical entities do map to each other. In other words, it is the inverse operation of *is-different?*.

These Boolean and set query operators suffice to deal with the questions of similarity and difference that a user would ask the system about the comparisons between mouse and human anatomy, and this aim serves to provide the structure (syntactic and semantic) for those operators.

4.5 Summary

In this chapter, we described the design of a pilot comparative anatomy information system that can answer queries regarding cross-species similarities and differences in structural phenotypes and serves the dual purpose of addressing important scientific questions in both medical informatics and comparative anatomy. In informatics, the inherent complexities of comparing such different anatomical data at so many levels of complexity for so many species carries the promise of developing techniques and tools that can be applied to genomic ontology alignment problems, taken as another level of anatomical complexity. In comparative anatomy, the structure and organization of massive amounts of anatomical data in one resource will serve multiple purposes of making information accessible and visualizable in different views for different users with different information needs, as well as for identifying gaps and inconsistencies in the scientific literature for future research. We hypothesize that our system will prove to be an initial step toward meeting these needs.

Chapter 5

INTERFACE AND SAMPLE QUERIES

This chapter presents a description of the system's interface and a detailed discussion of the components and their significance to the user, with examples of queries that can be executed in the system.

5.1 Introduction

In previous work, we described the development of the Structural Difference Method (SDM) formalism for representing the similarities and differences between homologous structures across different species [125]. Additionally, we proposed the design of a comparative anatomy information system (CAIS), based on the SDM, to support queries about those similarities and differences [124]. This chapter reports on the development and implementation of a graphical user interface for that system, as well as on our experiments with the use of CAIS, including scenarios from rodent-human research that show how the system can be used for realistic studies.

5.2 The CAIS System

As described in Chapter 4, the CAIS system [124] was designed to allow a user to study the similarities and differences between anatomical entities in two species. Similar to the *Emily* query interface to the FMA [29], queries to the CAIS system have the basic form:

`<anat. entity1>` <*query relation*> `<anat. entity2>`

where <`anat. entity 1`> is an anatomical entity from the first species, <`anat. entity 2`> is an anatomical entity from the second species, and the query relation is one of the following operators: *similar-to*, *different*, *shared*, *not shared*, *union*, *is-homologous?*, and *is-different?*. Either <`anat. entity1`> or <`anat. entity2`> can be `Unknown`, in which case the system returns a mapping for the specified anatomical entity if one exists in the database. If there are two anatomical entities specified, one in each species, or if the `Unknown` reference has been resolved, the system returns the information as requested by the operator. The operators, which are based on the design described in Chapter 4 with some improvements that came about during the programming phase, can be summarized as follows.

5.2.1 Result set operators

The result set operators consist of the following:

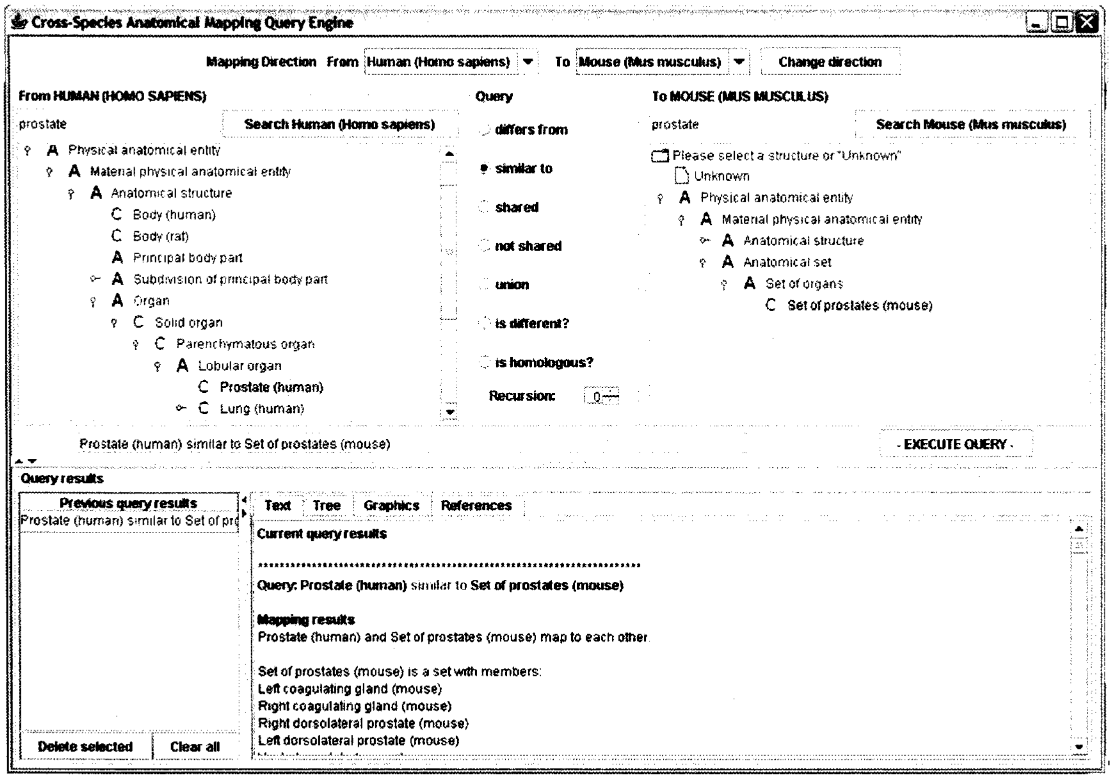
similar-to: returns an anatomical isomorphism (1-to-1 and onto correspondence) between the two homologous structures across species at the level of granularity (*e.g.*, `Organ`, `Organ part`, `Cell`) of the query if there is one, and returns `False` otherwise. For example, the `Left` and `Right atria` and `Left` and `Right ventricles` of the `Heart` are similar between the mouse and the human.

different: returns a non-null correspondence other than anatomical isomorphism (*e.g.*, a one-to-many relationship) between two homologous structures across species at the level of granularity of the query if there is one, and `False` if there is no mapping in the database. For example, the `Right lobes` of the mouse and human `Lungs` are different because they are in a 4:3 relationship.

shared: returns all the parts of the structure which occur in both species to the level of granularity specified. For example, the human and mouse brains both contain an `Amygdala`, so `Amygdala` would be one of the structures returned on a *shared* query on human and mouse `Brain`.

not shared: returns all the parts of the structure which occur in one species or the other, but not both, to the level of granularity specified; this is the set complement of the structures returned by *shared*. For example, the human brain *includes* `Gyri` and `Sulci` that mouse brains do not, so the *not shared* relation between human and mouse brains would contain those `Gyri` and `Sulci` (among other structures).

union: returns all the parts of the structure that occur either in one species or the other, or in both, to the level of granularity specified: in other words, the set union of the



prostate

- A Physical anatomical entity
  - A Material physical anatomical entity
    - A Anatomical structure
      - C Body (human)
      - C Body (rat)
      - A Principal body part
      - A Subdivision of principal body part
      - A Organ
        - C Solid organ
          - C Parenchymatous organ
            - A Lobular organ
              - C Prostate (human)
            - C Lung (human)

○ differs from

● similar to

○ shared

○ not shared

○ union

○ is different?

○ is homologous?

Recursion: 0

prostate

□ Please select a structure or "Unknown"

- □ Unknown
- A Physical anatomical entity
  - A Material physical anatomical entity
    - A Anatomical structure
    - A Anatomical set
      - A Set of organs
        - C Set of prostates (mouse)

Prostate (human) similar to Set of prostates (mouse)

Prostate (human) similar to Set of pr

Current query results

```
*******************************************************************
Query: Prostate (human) similar to Set of prostates (mouse)

Mapping results
Prostate (human) and Set of prostates (mouse) map to each other.

Set of prostates (mouse) is a set with members:
Left coagulating gland (mouse)
Right coagulating gland (mouse)
Right dorsolateral prostate (mouse)
Left dorsolateral prostate (mouse)
```

Figure 5.1: Results of a query to the knowledge base in text mode.

structures returned by the CAIS relationships *shared* and *not shared.*

5.2.2 Boolean operators

The Boolean operators consist of the following:

is-homologous? returns `True` if the two entities selected for the query are homologous, and `False` if they are not.

is-different? is the opposite of *is-homologous?*—it returns `False` if the two entries selected for the query are homologous, and `True` if they are not.

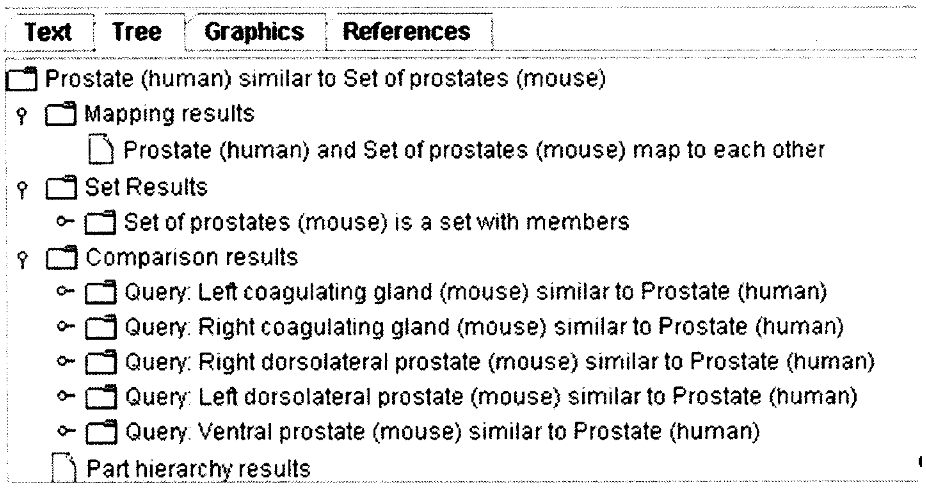
Figure 5.1 illustrates a screen shot of the CAIS graphical user interface that shows the results of a query to the knowledge base in text mode.

5.3 The CAIS Interface

To make the CAIS query functionality available to users, we have designed and implemented a graphical user interface. The CAIS interface is written in Java, and uses the Java API to access the Protégé-2000 database, in which rat, mouse, and human anatomical structures comprise a single hierarchy ([124], [126]). The CAIS interface provides the following functionalities.

1. choose the pair of species to compare from all species in the database,
2. select an anatomical entity from a hierarchy or search for one that the user has entered and give him/her a choice if the entry is ambiguous,
3. inform the user if selected entities cannot be directly compared and indicate reasonable alternatives if they exist,
4. select the query operator from a list of choices,
5. show the user query in a string form as the user constructs it from the GUI,
6. compare the selected structures at multiple levels of the parts hierarchy as selected by the user (default is 1 level)
7. keep track of results from prior queries so the user can return to them, and
8. show the output in multiple forms including text, tree, graphics, and references.

Figure 5.1 shows a screen shot of the full user interface. The user has selected the species human on the left and mouse on the right. She has typed in "prostate" in the search area on the left, and the system has found the human prostate in the hierarchy and displayed it. She has also typed in "prostate" in the search area on the right, and the system has responded with a message, "Select from search results," and displayed four possibilities from which the user has selected `Set of prostates (mouse)`. She has then selected the query operator *similar* and clicked on the Execute Query button. The query has been executed, and the results displayed in text mode, since the text tab is the default display tab. As the text



Prostate (human) similar to Set of prostates (mouse)
- Mapping results
  - Prostate (human) and Set of prostates (mouse) map to each other
- Set Results
  - Set of prostates (mouse) is a set with members
- Comparison results
  - Query: Left coagulating gland (mouse) similar to Prostate (human)
  - Query: Right coagulating gland (mouse) similar to Prostate (human)
  - Query: Right dorsolateral prostate (mouse) similar to Prostate (human)
  - Query: Left dorsolateral prostate (mouse) similar to Prostate (human)
  - Query: Ventral prostate (mouse) similar to Prostate (human)
- Part hierarchy results

Figure 5.2: Tree display mode.

mode is very verbose, the user may wish next to look at the results in tree mode (Figure 5.2) or graphics format (Figure 5.3). Tree results are returned as a structured hierarchy, down as many levels of the tree as was specified in the selected recursion level. In the graphics results a representative graphic is included at each level of the hierarchy.

5.4 Scenarios

In order to illustrate the potential use of the CAIS system, we give several research scenarios from the literature. We motivate the need for such a tool in each scenario and give examples of CAIS queries (in simplified string form) that can be used by the researchers in these studies.

5.4.1 Scenario 1: Correlating prostatic lobes/organs

Dr. A is a pharmacological scientist who is studying the effect of candidate compounds for new prostate cancer drugs. Because different regions of the human prostate are subject to different diseases, those regions that develop benign prostatic hyperplasia do not develop cancer, and vice-versa. Dr. A wants to determine the rat-human homologies for the dor-

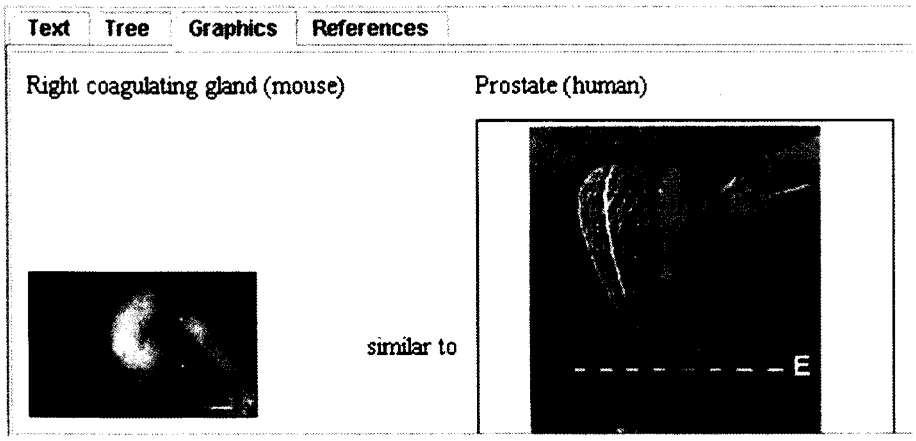


Figure 5.3: Graphics display mode.

solateral and ventral regions of the prostate, so that she can correlate the observed effects of the compounds in rat tissue with predictions for the effects in humans expected to be observed in later clinical trials. Specifically, her questions are: do the dorsolateral prostates of the rat correspond to the dorsolateral regions of the human prostate, and does the rat ventral prostate correspond to the ventral region of the human prostate (called anterior lobe in humans)?

The CAIS operator *similar-to* provides information for the researcher on what structures are homologous across species, what evidence exists that they are homologous (*e.g.*, traditional embryological studies, genetic assays), and the provenance or source of that information. Dr. A's queries will be:

<`Dorsolateral prostate (rat)`> <*is-homologous?*> <`Dorsal lobe of prostate (human)`>

<`Ventral prostate(rat)`> <*is-homologous?*> <`Anterior lobe of prostate (human)`>.

The attributed relationship returned by CAIS answers the researcher's query: documented in Dorothy Price's embryological work on "Comparative Aspects of Development and Structure in the Prostate" in the *National Cancer Institute Monograph* 1963 Oct. 12:1-

27 [97], the rat dorsolateral prostates are homologous to the dorsolateral lobes of the human, while the rat ventral prostate is *not* homologous to the anterior lobe of the human prostate. Based on this information, Dr. A. adjusts her expected correlations of the compound's effect accordingly.

5.4.2 Scenario 2: Correlating prostatic zones/organs

Dr. B is a pathologist who is formally developing new mouse models of human prostate cancer. Part of his evaluation is the application of analyses of previous results in mouse modeling of human prostate cancer and the determination of what those analyses imply for a mouse model that more soundly mirrors the initial development and the subsequent progression of prostatic tumors.

He has a candidate model in mind, pending confirmation of certain homologies. Given that the human peripheral zone is the region in which most prostate carcinomas originate, his question to establish the validity of that candidate rests on the results of the following correspondence: what is the mouse prostate region corresponding to the human peripheral zone of the prostate? His CAIS query based on that question will be in the form

`<Unknown (mouse)>` <*similar-to*> `<Peripheral zone of prostate (human)>`.

In this case, CAIS can be used to return not only the result set for the query, but also the references that back up the result, including, for example, the information that on the basis of an epidemiological study, [132] reports that the mouse dorsolateral prostate corresponds to the peripheral zone of the human prostate, and that [104] concurs on a preliminary basis, but cautions that Xue's assertion is based on descriptive data, and that the molecular studies that would confirm the correspondence remain to be carried out. Based on this information, Dr. B. determines that his mouse model is as yet insufficiently validated, and incorporates certain molecular assays on the dorsolateral prostate as part of the validation process for this model.

5.4.3 Scenario 3: Shared similarities and differences in prostate for tumor microenvironment

Dr. C wants to determine the best mouse tumor model for determining clinically relevant information on the response of tumors to a particular treatment effect. Bearing in mind the significant role the tumor host microenvironment (in this case, vasculature among other variables) can play in establishment of the tumor and its response to treatment, Dr. C. requires information on what aspects of the prostatic epithelium-the tumor microenvironment-are similar between the mouse and the human, and what aspects are different.

The queries *shared*, *not-shared*, and *union* provide information about the documented evolutionary possibilities for a given anatomical structure. For example, to confirm that the basic cellular structure of the mouse and human prostates are similar enough to support generalizing from the mouse tumor microenvironment to the human (a subset of Dr. C's eventual result set) the researcher may wish to verify that the prostates in both species consist of the same types of cells. This researcher's query would take the form

<`Prostatic epithelium (mouse)`> <*shared*> <`Prostatic epithelium (human)`>.

CAIS would return the result set that the prostatic epithelium in both species share the following cell types: `Secretory epithelial cell`, `Basal epithelial cell`, `Neuroendocrine epithelial cells`, citing [74] and [44], among others, as sources for this information, and verifying for this researcher that the species are histologically similar enough to validate a particular proposed study. The results of previous queries are accessible for use in building the compound query, which will return the totality of the shared features of the tumor microenvironment.

5.4.4 Scenario 4: Union of all normal stem cells as basis of a breast cancer tumor cell taxonomy

Dr. D is a tumor biologist who uses genome-wide expression analysis on normal luminal epithelial and myoepithelial/basal lineages of tumor cells for molecular classification of breast cancer, to the end of developing therapies that are less toxic than traditional radiation or chemotherapy treatment. As a first step in this research, he is working on a cross-species stem cell hierarchy, which he expects to reveal important aspects of the histogenesis of breast cancer evolution.

The CAIS operator *union* gives the range of all normal possibilities of these structures in the species under examination. CAIS will return all of the similarities and differences at all levels of granularity in the knowledge base in response to a *union* query. In order for Dr. D to obtain the desired information for his hierarchy, a detailed compound query on the relevant anatomical sites is necessary. One representative component of this compound query is

`<Epithelial cell of mammary gland (mouse)>` <*union*> `<Epithelial cell of lactiferous duct tree (human)>`.

The researcher builds up the query from components like this, and submits the query *in toto* to CAIS. Based on the information returned, Dr. D now has a result set from which he can derive his cell hierarchy, which will underlie his examination of breast cancer histogenesis.

5.5 Summary

Drawing on the Structural Difference Method (SDM), developed in our previous work, we developed and implemented an application that extracts cross-species anatomical information from a Protégé-2000 database file and allows the users to query the application about correspondences and differences in those anatomical structures. We implemented features to make the application more user-friendly, such as allowing the user to build a query by clicking, rather than being forced to remember the syntax, and by allowing the user to view and change the query as it is being constructed.

We provide a search feature, and control which classes can be searched and selected. All of the set operators developed for the SDM have been implemented, and they permit different aspects of anatomical correlation to be queried. The tabs provide different views for users to choose among, including unstructured set results, a structured hierarchy of results, graphics for comparison, and attributed slots that describe the basis (embryological or genetic) of the anatomical correlation, and the provenance of the information it was based upon.

Based on correspondence with domain experts, we are continuing to incorporate information on five different rodent organs (mammary gland, prostate, lung, ovary, and cervix), and preliminary feedback from users indicate a very welcome reception. In fact, the need for communicating these anatomical correspondences is becoming greater as the research into animal models of disease becomes more interdisciplinary and as researchers come from other backgrounds than traditional comparative anatomy. Our scenarios reflect the real need expressed by users for valid comparative anatomy information available in a user-friendly manner.

Chapter 6
DATA AND RESULTS

This chapter presents a review of the selection of the data and the methods used to acquire it, as well as the results of a set of queries representative of real-world comparative anatomy problems.

6.1 Motivation: the need for biological research data

In their evaluation of ontology exchange languages for bioinformatics, McEntire *et al* list reasons for the importance of ontology development. Among others, they mention:

1. the necessity for modeling at the appropriate level of granularity, in order to capture sufficient data elements for some problem-solving task;

2. the necessity for correctly forming the semantics of an ontology, in order to preserve the integrity of the information it contains;

3. the value of sharing the knowledge constituted by the ontologies themselves;

4. the ability of the process of ontological representation of biological principles to shed insight on the underlying biology it represents [78].

Biological research data from scientists actively working on real-life problems is an important way to validate our model in light of the reasons listed by McEntire. While in previous chapters, we presented a wider range of more high-level examples in order to test the limitations of the SDM, for the application itself we present deeper, more detailed examples, supplied by the users themselves, as well as by reference sources.

Obtaining biological research data from working scientists has multiple purposes. By verifying that our method can model and return accurate results at the level of granularity used by researchers, we verify that its data structures suffice for the problem-solving carried out by researchers in the lab (item 1, above). By translating the information provided to us in natural-language format into our syntax, executing the relevant queries, and comparing the semantics of the result set against the original, we can verify that our method preserves the integrity of the information it accepts, stores, and retrieves (item 2, above).

Since the information provided to us has value to the researchers who provided it, it is likely that our ontology based on that information will in turn be of value to other researchers (item 3, above), and so we plan to make our "core data set of information" [20] publicly and freely available to access. Finally, in the process of modeling the information, we clarify what information used by the researchers is explicit and what information is implicit [101]. By recognizing these gaps in explicit information, we shine a light on areas where anatomical information needs to be collected and tested, providing biological insights (item 4, above), and possibly leading to hypothesis generation.

6.2 Getting the data: domain expert input

6.2.1 Domain expert selection

Domain experts in mouse anatomy were selected by personal referral, as well as by a PubMed search. The selection of authors from PubMed was carried out in the following way:

1. For each one of our designated subset of the anatomical sites of interest identified by the MMHCC—*i.e.*, ovary, lung, cervix, prostate, and mammary gland—a PubMed search was carried out:

    - "Ovary"[MeSH] AND "Rodentia"[MeSH]
    - "Lung"[MeSH] AND "Rodentia"[MeSH]
    - "Cervix Uteri"[MeSH] AND "Rodentia"[MeSH]
    - "Prostate"[MeSH] AND "Rodentia"[MeSH]
    - ("Breast"[MeSH] OR "Mammary Glands, Human"[MeSH] OR "Mammary Glands, Animal"[MeSH]) AND "Rodentia"[MeSH]

Each one of these searches returned a MEDLINE-tagged corpus of PubMed literature for the particular anatomical site of interest as of January 2006, sorted by date from most recent to least recent.

2. Starting with the most recent item in each corpus, the affiliation of the first author of the article was identified by the item tagged with AD, which is one of the data tags returned when the MEDLINE display format is selected.

3. For each corpus, the list of affiliations of first authors identified by the AD tag was searched for the first 10 authors who provided email contact identification.

In all, 52 researchers were contacted from the information from PubMed articles and 7 through personal referrals for a total of 59 contacts, from which 6 agreed to provide a detailed description of the kind of comparative anatomical information that would be useful in their work.

6.2.2 Questionnaire

Researchers who agreed in the initial contact to participate in our information-gathering were sent a questionnaire that we developed. The questionnaire, and the rationale behind the questions, are described below.

The questionnaire sent to researchers was intended to elicit specific comparative anatomical information of the type dealt with every day in their work, and that they would consider essential in any comparative anatomy information system. It consisted of 4 questions, and the associated instructions, which were discussed and refined in our laboratory meetings before being sent out to researchers. The questionnaire itself is included in Appendix A; the questions and their purposes are summarized here.

The first question asked of the researchers was what their research was about. The purpose of this question was to define the scope of anatomical knowledge relevant to the researcher—what anatomical site is of interest, and in what level of granularity (*e.g.*, gross anatomy, microscopic anatomy, etc.) the researcher is most interested. Sample responses included:

- “investigate the immune response in mice after vaccination or infection with live influenza virus”;
- “veterinary pathologist primarily involved in the histopathologic evaluation of...wild type mouse mammary gland tissues”.

The second question modeled examples of each kind of query (similar, different, intersection, complement, and union). It asked the researcher to describe in free text the anatomical structures and their spatial and other relationships that the researcher would consider important to include. The purpose of this question was to gather relevant data with real-world application for the knowledge base. Sample responses included:

- “The anatomic location and the number of mammary glands varies between the rodent and human. Although primarily located along the ventral abdomen in the mouse, mammary gland tissue can be found in several other subcutaneous locations, including along the lateral or dorsal surfaces as evidenced by the occasional formation of mammary tumors in these locations.”
- “Mice usually have 5 pairs of mammary glands numbered 1 to 5 from anterior to posterior. Three pairs are in the cervicothoracic region and two are in the inguinoabdominal region. Males usually only have four pairs and do not have nipples.”
- “There is a difference in the antibody classes between man and mouse (different between rodents also!). In mouse serum IgG (IgG1, IgG2a, IgG2b, IgG3), IgM, IgA. In man IgG1, IgG2, IgG3 and IgG4 (no correspondence between mouse and man antibody subclass) and IgM, IgA (IgA1 and IgA2).”

The third question asked what content the researcher would consider essential for the knowledge base to contain. The purpose of this question was to establish what knowledge is fundamental to include in such an information system, to the degree that its omission would be considered a serious or fatal flaw in the content. In other words, the researchers were asked to evaluate what information, at a minimum, should be included for the knowledge base to be considered adequate. The researchers were also asked to describe briefly why this content was so fundamentally significant. Space was provided for the researchers to answer this question in as much length as they desired. Sample responses included:

- "Duct (intralobular and interlobular)", "Ductule", "Alveoli", "Lobule", "Stroma", "Nipple" [no reasons given];
- "Spleen, because part of the immune system, similar in function", "Blood, because distribution of bioactive substances (cytokines etc.) in rodents *contra* man";
- "BALT[1], because differs greatly between species, "NALT[2], because not [found] in humans".

The fourth question was optional, and was included in order to evaluate the relative importance of the entities and the relationships in the model. Researchers were asked which they would use more: queries about anatomical entities or about the relationships among those entities. They uniformly responded that the relationships among the entities were more useful than the entities alone.

Selected responses to the questionnaire are summarized in Appendix B. We obtained about 125 total entities (including parts of structures) that we were able to use in our model, a number representing less than 40% of the narrative description in the responses. Most of what was unable to be used consisted of descriptions of physiology or pathology, which are explicitly outside of CAIS' scope. In addition, we expanded our original set of entities to include others mentioned as relevant to the domain experts' work, even when they were not in the original MMHCC site. For example, one lung expert who responded works in an immunological capacity with lymphoid tissues, so we added `Tonsil (human)` and similar structures to our list of entities.

6.2.3 Use of PubMed and other references

In order to provide more detailed anatomical information, we extracted information from PubMed abstracts and other references, in addition to the data provided by the domain

[1] Bronchus-associated (or bronchial-associated) lymphatic tissue

[2] Nasal-associated (or nasopharynx-associated) lymphatic tissue

experts in their responses to the questionnaire. Relevant PubMed abstracts (mouse and rat ovary, lung, cervix, prostate, and mammary gland) were downloaded, and Perl scripts run on them to perform the following analysis:

- create a list of unique (non-duplicated) terms in the corpus;
- remove stop words and other non-functional terms from the corpus;
- tag anatomical entities for collection in a cumulative list;
- tag anatomical relations for collection in a cumulative list;
- return the cumulative lists for entry in the ontology.

The Perl scripts represent a very rudimentary approach to mining a PubMed corpus for anatomical entities and relationships. The scripts' basic method is that the first time they are run on a corpus of PubMed abstracts in XML format, they compile an alphabetized list of every word in the corpus, removing all duplications, for review. Review consists of manually examining the list, and marking every term as either an entity (for incorporation into CAIS), a relationship (for incorporation into CAIS), a stop word (to be ignored/excluded in subsequent iterations when the scripts are re-run), or context n (this word needs clarification; include n words on both sides of it when the scripts are re-run and a new list is generated). Subsequent runs of the scripts are cumulative—they add changes on to the original list generated in the first run. In this way, the corpus can be reviewed and marked up as many times as necessary to extract entities and relationships to populate the knowledge base.

6.3 The data

This section shows examples of how the free-text anatomical descriptions obtained from the domain experts and the literature were converted into our syntax and modeled in our knowledge base.

6.3.1 Prostate model and queries

Dorothy Price's embryological work on "Comparative Aspects of Development and Structure in the Prostate" in the National Cancer Institute Monograph 1963 Oct. 12:1-27 [97] states that the rat dorsolateral prostates are homologous to the dorsolateral lobes of the human, while the rat ventral prostate is not homologous to the anterior lobe of the human prostate. (Source: PubMed corpus)

We model that information in the following way:

- `Dorsolateral prostate (rat)` *is-a* `Lobular organ`
- `Dorsolateral prostate (rat)` *maps-to: embryologically* `Dorsal lobe of prostate (human)`
- `Dorsal lobe of prostate (human)` *maps-to: embryologically* `Dorsolateral prostate (rat)`
- `Ventral prostate (rat)` *is-a* `Lobular organ`
- `Ventral prostate (rat)` *maps-to: embryologically* `TBD-not null (human)`
- `Anterior lobe of prostate (human)` *maps-to: embryologically* `TBD-not null (rat)`

and support, among others, the following queries:

- Natural-language query: What structure in the rat corresponds to the dorsal lobe of the human prostate?
  - Corresponding CAIS query: `Unknown (rat)` *similar-to* `Dorsal lobe of prostate (human)`
- Natural-language query: Is the anterior lobe of the human prostate homologous to the ventral prostate in the rat?
  - Corresponding CAIS query: `Anterior lobe of prostate (human)` *is-homologous?* `Ventral prostate (rat)`

As mentioned in our example in the previous chapter, on the basis of an epidemiological study, [132] reports that the mouse dorsolateral prostate corresponds to the peripheral zone of the human prostate. [104] concurs on a preliminary basis, but cautions that the assertion in [132] is based on descriptive data, and that the molecular studies that would confirm the correspondence remain to be carried out. (Source: PubMed corpus)

We model that information in the following way:

- `Dorsolateral prostate (mouse)` *is-a* `Lobular organ`
- `Dorsolateral prostate (mouse)` *maps-to: embryologically* `Peripheral zone of prostate (human)`
- `Peripheral zone of prostate (human)` *maps-to: embryologically* `Dorsolateral prostate (mouse)`

and support, among others, the following queries:

- Natural-language query: What structure in the mouse corresponds to the peripheral zone of the human prostate?
  - Corresponding CAIS query: `Unknown (mouse)` *similar-to* `Peripheral zone of prostate (human)`

6.3.2 Mammary gland queries

Mice usually have 5 pairs of mammary glands numbered 1 to 5 from anterior to posterior. Three pairs are in the cervicothoracic region and two are in the inguinoabdominal region. (Source: domain expert's response to questionnaire)

We began by modeling that information in the following way:

- `Mammary gland (mouse)` *is-a* `Lobular organ`
- `Cervical mammary gland` *is-a* `Mammary gland (mouse)`
- `Thoracic mammary gland` *is-a* `Mammary gland (mouse)`

- Abdominal mammary gland *is-a* Mammary gland (mouse)
- Inguinal mammary gland *is-a* Mammary gland (mouse)
- Peri-anal mammary gland *is-a* Mammary gland (mouse)
- Cervical mammary gland *is-a* Mammary gland (mouse)

Although modeling the subsumption hierarchy was simple, it was insufficient—as mentioned previously [126], the *part-of* hierarchy is more useful for biological researchers, and reflects more closely the entities and their relationships that are most biologically relevant. Although the human mammary gland comprises multiple lactiferous duct trees (LDTs) communicating to a single nipple, the mouse mammary gland consists of a single LDT communicating to a nipple. (Source: domain expert's response to questionnaire)

We model that information in the following way:

- Mammary gland (mouse) *is-a* Lobular organ
- Mammary gland (human) *is-a* Anatomical set
- Breast (human) *maps-to: embryologically* Null (mouse)

  Right mammary gland 1 (mouse) *part-of* Cervicothoracic region (mouse)

  Left mammary gland 1 (mouse) *part-of* Cervicothoracic region (mouse)

  Right mammary gland 2 (mouse) *part-of* Thoracic region (mouse)

  Left mammary gland 2 (mouse) *part-of* Thoracic region (mouse)

  Right mammary gland 3 (mouse) *part-of* Abdominal region (mouse)

  Left mammary gland 3 (mouse) *part-of* Abdominal region (mouse)

  Right mammary gland 4 (mouse) *part-of* Inguinoabdominal region (mouse)

  Left mammary gland 4 (mouse) *part-of* Inguinoabdominal region (mouse)

  Right mammary gland 5 (mouse) *part-of* Peri-anal region (mouse)

  Left mammary gland 5 (mouse) *part-of* Peri-anal region (mouse) and support, among others, the following query:

- Query: Has a transformation in the structure of the mammary gland occurred between mice and humans?
- Answer (via the composite query, below):

  1. Query$_1$: `Unknown (mouse)` *similar-to* `Mammary gland (human)`

     - Answer$_1$: `Set of mammary glands (mouse)`

  2. Query$_2$: Is the superclass (parent) of Answer$_1$ identical to the superclass of `Mammary gland (human)`? (Note that this question must currently be answered by looking up one level of the hierarchy for each entity in the results returned for Query$_1$. The next version of the application will provide a quick and simple way to query on superclasses (parents) in order to automate this process.)

     - Answer$_2$: The superclass of Answer$_1$ = `Anatomical set`, while the superclass of `Mammary gland (human)` = `Lobular organ`.

  The fact that the superclasses differ indicate that there is an *edge-set difference* between the two entities, according to the SDM. An edge-set difference indicates that between two comparable entities, a transformation sufficient to change the class has occurred, and since this comparison is between mice and humans, the transformation is therefore a phylogenetic one.

  Therefore, the answer to the original query is:

  - True—a transformation between `Anatomical set` and `Lobular organ` has occurred in the structure of the mammary gland between mice and humans.

Note that while our information system does not provide an explanation for this transformation, it indicates a point at which potentially fruitful hypotheses can be generated as possible explanations for this transformation. Additionally, since we are dealing only with two species at a time at this point, we can indicate that there is a difference between mice and humans, but we have insufficient information to characterize that difference in terms of evolutionary change.

To describe evolutionary change, we require a phylogenetic tree, and a phylogenetic tree requires at a minimum a parent node and a child node. For example, mice and humans are both *chordates* (members of the phylum *Chordata*; for our purposes here, effectively a superset of *vertebrates*). One of the distinctive characteristics of chordates is a post-anal tail. Relative to the ancestral condition of possessing a tail, the mouse retains the *basal* condition (retains the tail of its chordate parent), while the human has the *derived* condition of a vestigial tail (losing the tail of its chordate parent) via some type or types of evolutionary transformation after their divergence.

By contrast, when we compare the mammary gland in the mouse and in the human, we are comparing leaf nodes, and so—without a parent node for reference—we can only qualitatively describe the differences (`Lobular organ` as opposed to `Anatomical set`). Without a parent node against which to reference *basal* vs. *derived*, we cannot put those differences in the leaf nodes into the larger context of evolutionary change. For the scope of this dissertation, we only compare leaf nodes; modeling phylogenetic trees and supporting queries regarding evolutionary change (as opposed to modeling and querying on simple difference) is an area of future research.

6.3.3 Lung queries

There are 5 lobes in the right mouse lung, but unlike the human the mouse has only a single left lobe. (Source: domain expert's response to questionnaire)

We model the top level (lung) in the following way:

- `Right lung (mouse)` *maps-to* `Right lung (human)`
- `Left lung (mouse)` *maps-to* `Left lung (human)`
- `Right lung (human)` *maps-to* `Right lung (mouse)`

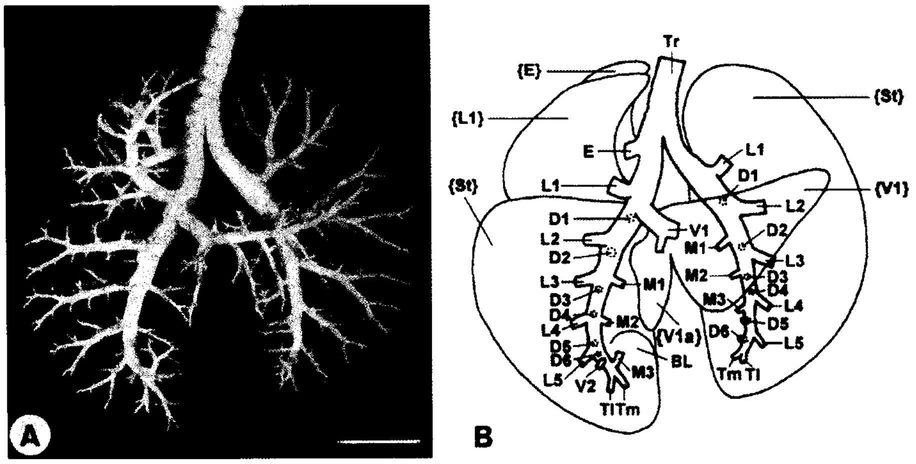


Figure 6.1: The relative symmetry of both sides of the mouse/rat tracheobronchial tree stands in contrast to the pronounced asymmetry of the lobes, with the attendant modeling implications (Image source: [129]).

- `Left lung (human)` *maps-to* `Left lung (mouse)`

The natural next step is to model the lobes of the mouse lung, but that step raises some problematic modeling issues.

First, the fact that the mouse lung is viewed by biologists as a single lobe is conceptually inconsistent, but the implicit knowledge behind that terminology makes it workable in daily practice. However, for our ontology, these inconsistencies must be dealt with. We resolve this in the following way:

- `Null (mouse)` *maps-to* `Upper lobe of left lung (human)`

- `Null (mouse)` *maps-to* `Lower lobe of left lung (human)`

- `Tracheobronchial tree (mouse)` maps to `Tracheobronchial tree (human)` as expected (see Figure 6.1 for reference). This demonstrates the principle previously mentioned that mappings can be more or less symmetrical at varying levels of organization, while skipping (null) layers of organization in between.

Mapping the lobes of the mouse right lung to the lobes of the human right lung does not pose the same logical problem, but rather a practical one. We know that a determinable many-to-many mapping exists between the lobes across species ([17], [90], [98]), but we do not yet know exactly what that mapping is ([53], [81], [92], [119], [129], [130], [134], [133]). The embryological information necessary to determine those mappings has not been adequately documented in the literature. At present, pending further clarification of the appropriate mappings, we model that information in the following way:

- `TBD`[3] `(mouse)` *maps-to* `Upper lobe of right lung (human)`
- `TBD (mouse)` *maps-to* `Middle lobe of right lung (human)`
- `TBD (mouse)` *maps-to* `Lower lobe of right lung (human)`
- `Lobe 1 of right lung (mouse)` *maps-to* `TBD (human)`
- `Lobe 2 of right lung (mouse)` *maps-to* `TBD (human)`
- `Lobe 3 of right lung (mouse)` *maps-to* `TBD (human)`
- `Lobe 4 of right lung (mouse)` *maps-to* `TBD (human)`
- `Lobe 5 of right lung (mouse)` *maps-to* `TBD (human)`

This example demonstrates two features of the CAIS system:

1. the process of determining cross-species mapping can illuminate gaps in the existing literature, where necessary knowledge is missing (*cf.* Rosse's "formalization improves conceptualization");
2. the system supports the entry of tentative or incomplete knowledge, that maintains the integrity of the knowledge base, and that can be updated later as the necessary knowledge is generated or discovered.

[3] TBD: to be determined. This is a convention in our knowledge base to distinguish between null mappings (to an entity which does not exist in the target species) and between unknown mappings. TBD means that the mapping has not yet been done (no information at all), and TBD-not null means that we cannot yet definitively put an entity in the slot, but we know that one exists—*i.e.*, is **not** null.

Instead of tonsils, mice have NALT (nasal-associated lymphatic tissue). (Source: domain expert's response to questionnaire)

We model that information in the following way:

- `Tonsil (human)` *maps-to: embryologically* `NALT (mouse)`
- `NALT (mouse)` *maps-to: embryologically* `Tonsil (human)`

and support, among others, the following queries:

- Natural-language query: What structure in the mouse corresponds to the tonsils in humans?
  - Corresponding CAIS query: `Tonsil (human)` *similar-to* `Unknown (mouse)`.

6.3.4 Ovary queries

The ovary's surface is composed of surface epithelium. The next layer is the tunica albuginea ovarii, which is composed of dense connective tissue. In the human and in rodents, as in most species, the cortex of the ovary surrounds the medulla of the ovary. Ovarian follicles, which are made up of follicular cells containing developing oocytes, interstitial gland cells, and stromal elements make up the cortex of the ovary. The medulla, by contrast, is composed of loose fibrous connective tissue, and large blood vessels, nerves and lymphatic vessels, which communicate with the rest of the body through the hilus of the ovary. (Source: domain expert's response to questionnaire)

We model that information in the following way:

- `Ovary (mouse)` *has-part* `Epithelium (mouse)`
- `Ovary (mouse)` *has-part* `Tunica albuginea ovarii (mouse)`
- `Ovary (mouse)` *has-part* `Ovarian cortex (mouse)`
- `Ovary (mouse)` *has-part* `Ovarian medulla (mouse)`

- Ovarian cortex (mouse) *has-part* Ovarian follicle (mouse)
- Ovarian cortex (mouse) *has-part* Oocyte (mouse)
- Ovarian cortex (mouse) *has-part* Interstitial gland cell (mouse)
- Ovarian cortex (mouse) *has-part* Ovarian stroma (mouse)
- Ovarian medulla (mouse) *has-part* Connective tissue (mouse)
- Ovary (mouse) *has-part* Hilum of ovary (mouse) ...

Because of the anatomical isomorphisms at multiple levels of human and rodent ovaries, the queries on this organ are extremely straightforward, and present no particular modeling issues.

6.3.5 Cervix queries

The female mouse has a duplex uterus with uterine horns communicating just prior to entering the single cervix. (Source: PubMed corpus)

We model that information in the following way:

- Uterus (mouse) *has-part* Right uterine horn (mouse)
- Uterus (mouse) *has-part* Left uterine horn (mouse)
- Right uterine horn (mouse) *maps-to: embryologically* Uterine cavity (human)
- Left uterine horn (mouse) *maps-to: embryologically* Uterine cavity (human)
- Uterine cavity (human) *maps-to: embryologically* Right uterine horn (mouse)
- Uterine cavity (human) *maps-to: embryologically* Left uterine horn (mouse)
- Cervix (mouse) *maps-to: embryologically* Cervix (human)
- Cervix (human) *maps-to: embryologically* Cervix (mouse)

and support, among others, the following queries:

- How do the mouse uterus and cervix differ from their human homologues?: `Unknown (mouse)` *similar-to Cervix (human)*; `Unknown (mouse)` *similar-to* `Uterus (human)`

6.4 Evaluation of results

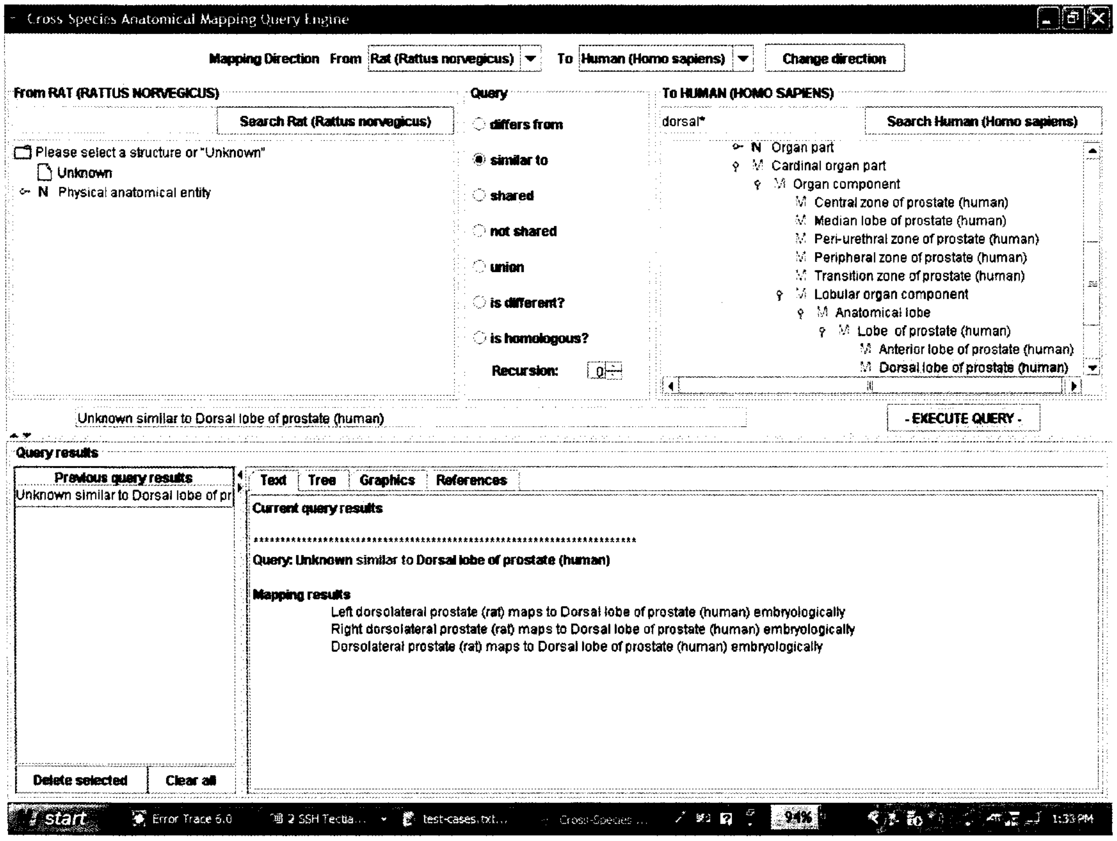
Because we do not determine what the content of the knowledge base is, but rather, we model expert consensus [125], that determines how we evaluate the application in regard to the correctness of content. Results, therefore, are correct if they match the results provided by the expert or reference. That means that they have to "survive" 1) the process of normalization, according to our syntax and semantics, and 2) entry into Protégé in such a way that the result set based on that information corresponds to what the resource originally said in natural language.

6.4.1 Testing the results of the process

The testing process for the application consisted of developing and carrying out a suite of test cases based on the scenarios and associated queries. The test cases were all associated with an underlying query, and consisted of the query and the expected results, to be verified against the results obtained when the query was actually run. Below is a set of representative test cases. Figure 6.2 shows an example of testing a prostate query.

Test prostate queries

- *Query:* `Dorsolateral prostate (rat)` *similar-to* `Unknown`
  - *Expected response:* `Dorsal lobe of prostate (human)`
  - *Obtained expected response:* Yes
- *Query:* `Right dorsolateral prostate (rat)` *similar-to* `Unknown`
  - *Expected response:* `Dorsal lobe of prostate (human)`
  - *Obtained expected response:* Yes


Cross Species Anatomical Mapping Query Engine
Mapping Direction From Rat (Rattus norvegicus) To Human (Homo sapiens) Change direction
From RAT (RATTUS NORVEGICUS)
Search Rat (Rattus norvegicus)
Please select a structure or "Unknown"
Unknown
N Physical anatomical entity
Query
○ differs from
● similar to
○ shared
○ not shared
○ union
○ is different?
○ is homologous?
Recursion: 0
To HUMAN (HOMO SAPIENS)
dorsal*
Search Human (Homo sapiens)
N Organ part
M Cardinal organ part
M Organ component
M Central zone of prostate (human)
M Median lobe of prostate (human)
M Peri-urethral zone of prostate (human)
M Peripheral zone of prostate (human)
M Transition zone of prostate (human)
M Lobular organ component
M Anatomical lobe
M Lobe of prostate (human)
M Anterior lobe of prostate (human)
M Dorsal lobe of prostate (human)
Unknown similar to Dorsal lobe of prostate (human)
- EXECUTE QUERY -
Query results
Previous query results
Unknown similar to Dorsal lobe of pr
Delete selected
Clear all
Text
Tree
Graphics
References
start
Error Trace 6.0
2 SSH Tecia...
test-cases.txt...
Cross-Species ...
1:33 PM

Current query results

```
********************************************************************
Query: Unknown similar to Dorsal lobe of prostate (human)

Mapping results
            Left dorsolateral prostate (rat) maps to Dorsal lobe of prostate (human) embryologically
            Right dorsolateral prostate (rat) maps to Dorsal lobe of prostate (human) embryologically
            Dorsolateral prostate (rat) maps to Dorsal lobe of prostate (human) embryologically
```

Figure 6.2: Test of a representative prostate query.

- *Query:* `Left dorsolateral prostate (rat)` *similar-to* `Unknown`
  - *Expected response:* `Dorsal lobe of prostate (human)`
  - *Obtained expected response:* Yes
- *Query:* `Dorsolateral prostate (rat)` ***is-homologous?*** `Dorsal lobe of prostate (human)`
  - *Expected response:* `T`
  - *Obtained expected response:* Yes
- *Query:* `Dorsolateral prostate (rat)` ***is-different?*** `Dorsal lobe of prostate (human)`
  - *Expected response:* `F`
  - *Obtained expected response:* Yes
- *Query:* `Right dorsolateral prostate (rat)` ***is-homologous?*** `Dorsal lobe of prostate (human)`
  - *Expected response:* `T`
  - *Obtained expected response:* Yes
- *Query:* `Right dorsolateral prostate (rat)` ***is-different?*** `Dorsal lobe of prostate (human)`
  - *Expected response:* `F`

  - *Obtained expected response:* Yes

- *Query:* `Left dorsolateral prostate (rat)` ***is-homologous?*** `Dorsal lobe of prostate (human)`
  - *Expected response:* `T`
  - *Obtained expected response:* Yes

- *Query:* `Left dorsolateral prostate (rat)` ***is-different?*** `Dorsal lobe of prostate (human)`
  - *Expected response:* `F`
  - *Obtained expected response:* Yes

- *Query:* Ventral prostate (rat) **is-homologous?** Anterior lobe of prostate (human)
  - *Expected response:* **F**
  - *Obtained expected response:* Yes

- *Query:* `Ventral prostate (rat)` ***is-different?*** `Anterior lobe of prostate (human)`
  - *Expected response:* `T`
  - *Obtained expected response:* Yes

- *Query:* `Unknown` *similar-to* `Ventral prostate (rat)`
  - *Expected response:* `TBD-not null (human)`

  - *Obtained expected response:* Yes

- *Query:* `Anterior lobe of prostate (human)` *similar-to* `Unknown`
  - *Expected response:* `TBD-not null (rat)`
  - *Obtained expected response:* Yes

- *Query:* `Dorsolateral prostate (mouse)` *similar-to* `Unknown`
  - *Expected response:* `Peripheral zone of prostate (human)`
  - *Obtained expected response:* Yes

- *Query:* `Right dorsolateral prostate (mouse)` *similar-to* `Unknown`
  - *Expected response:* `Peripheral zone of prostate (human)`
  - *Obtained expected response:* Yes

- *Query:* `Left dorsolateral prostate (mouse)` *similar-to* `Unknown`
  - *Expected response:* `Peripheral zone of prostate (human)`
  - *Obtained expected response:* Yes

- *Query:* `Dorsolateral prostate (mouse)` ***is-homologous?*** `Peripheral zone of prostate (human)`
  - *Expected response:* `T`
  - *Obtained expected response:* Yes

- *Query:* Right dorsolateral prostate (mouse) ***is-homologous?*** Peripheral zone of prostate (human)
    - *Expected response:* T
    - *Obtained expected response:* Yes
- *Query:* Left dorsolateral prostate (mouse) ***is-homologous?*** Peripheral zone of prostate (human)
    - *Expected response:* T
    - *Obtained expected response:* Yes
- *Query:* Dorsolateral prostate (mouse) ***is-different?*** Peripheral zone of prostate (human)
    - *Expected response:* F
    - *Obtained expected response:* Yes
- *Query:* Right dorsolateral prostate (mouse) ***is-different?*** Peripheral zone of prostate (human)
    - *Expected response:* F
    - *Obtained expected response:* Yes
- *Query:* Left dorsolateral prostate (mouse) ***is-different?*** Peripheral zone of prostate (human)
    - *Expected response:* F

- *Obtained expected response:* Yes

The motivation behind the next example is the search for an appropriate mouse model for the correlation between tissue cultures of human tumors and the clinical course of cancer (clinical significance) ([12], [54]). This example is based on the fact that different breast tumors exhibit different degrees of aggressiveness—some cancers are relatively *indolent* (slow to spread), while other cancers metastasize rapidly. Different *histogenesis* (tissue origin) is correlated with different rates of tumor growth, and understanding the origins of the tissues involved in the tumor will help in refining the correlation ([40], [91], [112], [58,]. Bratthauer investigated the incidence of invasive lobular and ductal cancers, and concluded that—while the reasons "remain...unclear and...unexplored", it is possible for the characteristics of the neoplastic cells to retain their stem cell characteristics, accounting for the possibility of developing into an invasive phenotype [14]. This may accord with and reinforce Stingl's model in which "the commitment to the luminal versus the myoepithelial lineage may play a determining role in the generation of alveoli and ducts ([116])". Al-Hajj goes so far as to identify this consideration as "challeng[ing] our current paradigms of experimentation" [4], [5], [32]. The potential importance of this model is the basis for our choosing it as an example of what CAIS can handle in the way of compound queries.

Spanakis *et al* studied fibroblasts and myofibroblasts from different types of breast tissue and reported that fibroblasts from malignant tumors were phenotypically more distant from normal cells compared with other pathological types. They propose that stromal and epithelial tissues interact with each other during the development of breast tumors, in what they term "co-adaptive transformation", and further, they propose that different types of fibroblasts give rise to different types of myofibroblasts [114]. This correlation of different pathological phenotypes, and their qualitative description of phenotypical distance, indicates that it may be useful to classify normal mammary cells involved in cancerous tumors in each species, and to use the SDM to determine what similarities and differences exist between the two hierarchies. This possibility is additionally reinforced by Stingl's and Villadsen's observation of the developmental nature of the human mammary gland—the hierarchy of progenitor cells [116] "holds promise for the existence of a stem cell hierarchy, the understanding of which may prove to be instrumental in further dissecting the histogenesis of breast cancer evolution" [128] Such a hierarchy lends itself well to symbolic modeling in an FMA/CAIS template, and these assumptions underlie the following example.

In our example, we propose a scenario in which a researcher wishes to compare the hierarchy of mouse and human mammary stem cells. In preparation for this scenario, Fridriksdottir's two breast epithelial stem cell lineages have been modeled:

- `Luminal epithelial cell of lactiferous duct (human)` *is-a* `Endo-epithelial cell`
- `Luminal epithelial cell of lactiferous duct (mouse)` *is-a* `Endo-epithelial cell`
- `Myoepithelial cell of lactiferous duct (human)` *is-a* `Meso-epithelial cell`
- `Myoepithelial cell of lactiferous duct (mouse)` *is-a* `Meso-epithelial cell`
- `Stem cell of lumen of lactiferous duct (human)` *is-a* `Stem cell`
- `Stem cell of lumen of lactiferous duct (mouse)` *is-a* `Stem cell`
- `Stem cell of myoepithelium of lactiferous duct (human)` *is-a* `Stem cell`
- `Stem cell of myoepithelium of lactiferous duct (mouse)` *is-a* `Stem cell`

and so on, in order to populate the two stem cell lineages (for our purposes, subsumption hierarchies) for each species in our model.

- *Query:* `MESC lineage (mouse)` *different* `MESC lineage (human)` [recurse 2 levels]
  - *Result*$_1$*:* Mouse has 3 progenitor cell populations *is-a*: ductal-restricted, lobular-restricted, and bipotent; human has one multipotent population. According to Fridriksdottir, "more elaborate characterization is warranted" [40]; pending that characterization, this is as far as CAIS is able to describe the difference.
- *Query:* `MESC lineage (mouse)` *different* `MESC lineage (human)` [recurse 4 levels]

- *Result*$_2$: Result 1 + Mouse cell *has-part* `CALLA`, `MUC1`; human cell does not.

The examples above illustrate two CAIS features: first, the user's ability to select how deep to search the tree (recurse n levels), and second, the composition of compound queries or results by concatenation, as in *Emily* [29].

6.4.2 Evaluation and feedback by domain experts

Our evaluation process consists of two parts: evaluation of the interface, and evaluation of the content. We have begun evaluating the interface by instructing users in the operation of the application, then giving them some sample queries to try, and recording their usage of the application (using the commercial application Camtasia for video screen capture) for analysis for difficulties (based on Kim's criteria for evaluating anatomical software and Web pages [60], [61]), as well as having them fill out a survey.

Feedback to date, while yet sparse, does show some distinct tendencies, although with so few responses, it remains to be seen whether these are real trends in the evaluation, or simply artifacts due to as-yet inadequate power of the small sample size. Nevertheless, the tendencies are plausible, and they reflect design issues we have encountered, so if they continue to be borne out, this will not be a surprising development.

The interface is fairly universally agreed to be easy to use and to navigate, acting as experienced Windows users would expect a well-behaved application to perform. The real-time feedback on the query under construction has been especially well-received. However, in order to use the application, non-trivial knowledge of the content plays a role, which has caused difficulty for reviewers without specific anatomical knowledge. Additionally, our use of the specialized terms *map*, *similar*, and *homologous* has created some confusion. This fact indicates two issues, one in designing the system, and one in implementing it.

The system design issue is that our concept of homology, as implemented, is not granular enough—in other words, more work remains to be done in the recognition and representation of complex and partial homologies. Dividing the *maps-to* relationship into the attributed *maps-to: genetically* and *maps-to: embryologically* is a beginning, but more examination of complex homologies, incorporating their appropriate knowledge representation, remains to be done. In addition to the biological reality, the appropriate representation remains to be developed more finely—our level of detail in mapping does not account for purported homologies based on epidemiological studies, and the corresponding levels of evidence for the mapping, for example.

Alongside this system design issue, it is also clear that some kind of user help, whether a help file, a tutorial, pop-ups, or some combination is an appropriate addition in the next version of our application. Even after the concepts are explained to the user, the names of the relationships available to choose from can be uninformative, and some kind of assistance for the user would be very helpful at this point. In addition, the "Recursion" feature seems to be particularly unintuitive for the user; whether a help feature would be sufficient to remedy this issue, or whether revisiting the presentation of the entire concept is called for remains to be seen in our next version.

Despite an initial steep learning curve, however, users are excited about the potential of the system, and are already asking for more species (most notably, zebrafish) to be included in the knowledge base. We anticipate that future versions, which will be more user-friendly, will be even more enthusiastically received.

Additionally, we are lining up domain experts for the second part of the CAIS evaluation—the content. The domain experts will be instructed in the operation of the application, and (as in the interface evaluation described above), their usage of the application will be recorded and analyzed for content gaps. In addition, they will be surveyed for the completeness and correctness of the content of CAIS.

The write-up of the application has been peer-reviewed for the AMIA 2006 conference, and one of the reviewers' comments has been incorporated as evaluation/request for features in future versions of CAIS. The references to sources of the anatomical structures in the scientific literature is well-received, but as the reviewer points out, we have not yet made an attempt to deal with sources that conflict with each other or with our model. That issue will be an important component of our future research.

6.5 Summary

In this chapter, we reviewed the selection of the data and the methods used to acquire it, as well as the results of a set of queries representative of real-world comparative anatomy problems. The evaluation of our application was also briefly treated.

Chapter 7

PUTTING THE BIOLOGY IN BIOINFORMATICS: CONCLUSIONS AND FUTURE WORK

This chapter provides a summary of our completed work and its contributions, as well as a preview of future work.

7.1 Our work and its contributions

In this dissertation, we have described our work in developing a comparative anatomy information system (CAIS) by:

1. proposing an approach to creating a symbolic model of cross-species anatomical comparisons—the structural difference method (SDM);
2. implementing the SDM by symbolically modeling the similarities and differences in selected anatomical structures between humans and other species;
3. gathering domain knowledge to populate a knowledge base for rodent structures of selected site cancers of interest to the Mouse Models of Human Cancer Consortium (MMHCC);
4. developing a query interface to retrieve information from the knowledge base;
5. validating our approach by domain expert evaluation.

An introduction to the domains upon which our approach draws—including comparative anatomy, knowledge representation and modeling, and graph theory—was provided and discussed in the context of their implications for our approach, and representative examples of modeling and mappings that we carried out were presented.

Our work for this dissertation provided a prototype of a working comparative anatomy information system (CAIS) to support researchers' queries about differences between rodent and human anatomical models for five sites of interest to human cancer researchers—prostate, mammary gland, lung, ovary, and cervix. In order to design the knowledge base underlying this application, we developed and refined a theoretical model to represent anatomical similarities and differences, and in order to populate that knowledge base, we surveyed domain experts, representing their natural-language responses in our syntax and semantics.

Future work will include the following:

- developing interface and feature enhancements for the CAIS application itself;
- expanding the mappings in the content of the knowledge base to include more of the anatomical structures involved in the MMHCC site cancer working groups;
- extending the theoretical basis behind the application through the development of models, metamodels, and an anatomical algebra for dealing with them;
- determining appropriate and more rigorous methods of validation for our approach.

As a first step to translating information from these animal models into effective clinical treatments, much more work is needed to determine which of these cross-species anatomical differences are medically significant. Clarifying the meaning of these differences and representing them appropriately in our knowledge base will constitute an important part of validation of our method.

7.2 Future work

> "[B]ioinformatics is going to be critical to the evo-devo research program, which to date has emphasized the 'devo' part with much work on model systems, but is going to put increasing demands on comparative molecular information from genomics and bioinformatics to fulfill the promise of the 'evo' part." Paul Z. Myers (developmental biologist) [80]

As demonstrated in the literature review in Chapter 2, there are a myriad of species of medical interest. From the genome sequence in yeast and insects, to the pathogenicity of molds and bacteria and the toxicology of arthropods, to the beginnings of the immune system in ocean invertebrates, to the basic mammalian similarity that results in a useful animal model of disease—while some species are of more immediate usefulness than others, there is literally no species that does not have *any* comparative anatomical feature of interest.

This number and diversity of species and their significance means that our approach of modeling and comparing two closely-related species at a time—while useful for a particular set of biomedical informatics queries, and valid for establishing a proof of concept of the information system—needs to lead to an efficient and valid method of extending knowledge capture, modeling, and query support to multiple species. Our system provides the first very preliminary steps toward those efforts, yet much remains to be done, as will be described below.

The FMA's role as a reference ontology [103], coupled with the SDM as a way of symbolically describing cross-species similarities and differences [125], together provide a strong informatics foundation for such a system. The potential of the FMA as a reference for organizing knowledge about the human body has been well-documented ([102], [103]). Combining the FMA's representation of knowledge with the capacity of the SDM to compare different species representations for similarity and difference affords the opportunity to transcend the usual level of experimental and observational anatomical detail at the leaf node, and to create meaningful and useful abstractions about those comparisons which represent theoretical principles, support reasoning about those structures, and indicate fruitful areas for hypothesis generation.

For any one researcher or research group, the problem of capturing and modeling that much data would, for all practical purposes, be intractable. Therefore, knowledge-capture and data-mining techniques, supporting direct domain expert knowledge entry, and the attendant curation and resource issues, as well as knowledge representation and modeling and—most important—collaborative work issues—must all be addressed in future work.

7.2.1 Interface and feature enhancements

In order to support the usage of an expanded CAIS system by the comparative anatomical community, the system needs to be extremely user-friendly. Currently the only users are the development team, so working directly in Protégé has been feasible. However, Protégé is not designed as an end-user application, and even the development team has encountered interface issues that hinder usage. Examples of such issues include lack of support for quickly and easily populating inverse slots automatically, and inconsistent inheritance of species slots. A separate, user-friendly, visually-oriented knowledge-capture application is called for, if this knowledge base is to benefit from wide-scale knowledge sharing and usage in the comparative anatomical community.

Because the current CAIS system is a prototype, storing the small and selected knowledge base in a native Protégé .pont file has been sufficient to date. However, as the experience of the FMA shows, once the knowledge base becomes sufficiently rich, a database structure becomes necessary to accommodate the volume of information involved. The FMA, with its more than 100,000 entities and more than 1.5 million relationships among those entities just to describe the human body [103], already encounters storage and performance degradation issues; a project of the scope of multiple species models and metamodels will have to deal with those issues to an even greater degree. We do not propose a specific database-management system at this point; we merely indicate that this is an issue which will need to be addressed.

The issues above have financial, temporal, and resource implications for development and testing of the system which need to be delineated and planned for at the outset of the next stage of the research.

7.2.2 Expanding the mappings–knowledge capture and representation

> "It might be even more useful for someone like me (who has worked with mice) to find homologies in other, more distantly related, organisms—such as zebrafish (which are also being used as a model for group A strep pathogenesis). Can [the CAIS system] be extended to that, or is this more mammal-specific?"—Tara C. Smith (infectious-diseases epidemiologist), personal communication [109]

The CAIS can absolutely be extended to other, more distantly-related organisms, as well as to sex and developmental stage [3], increasing its utility to different researchers working with different species, and with the intent to translate the findings from that research into different branches of clinical treatment. We have already modeled the rat to the same degree as the mouse, and other more different species more sparsely, in order to determine whether any anatomical structures were so different from the human as to present an insuperable modeling challenge, and thus indicate a limitation in the FMA's capacity to model anatomical structure. So far, despite the diversity of species and structures we have modeled, the FMA + SDM has been sufficiently extensible to accommodate it. We hypothesize that the class structure of the FMA suffices to model anatomical structure at least as far down the phylogenetic tree as chordates.

On the other hand, that distance represents an incredible amount of data. In order to populate a useful comparative anatomy information system, the collaboration of the comparative anatomy community in providing the data is essential. One way of getting at the content of the shared intellectual capital of this community is through automated tools for natural-language processing of the research literature.

We made a very tentative initial foray into natural-language processing (NLP) tools in this research—in order to supplement the information provided by the domain experts, we developed a suite of Perl tools to extract entities and relationships from the appropriate PubMed corpora. However, the need for NLP tools, well-documented in the literature [6], is even more pressing when the volume of literature on the anatomical structures of all the species of medical interest is considered. A great deal of further research remains to be done in this domain—how do we develop and utilize NLP tools to gather and organize comparative anatomy data for our ontology?

7.2.3 Models, metamodels, and an anatomical algebra

Extending our approach to more and different species brings up many modeling issues, similar to the one we dealt with in modeling the mouse left lung and its relationship (if any) to lobar structure. Just to name one of many examples, should `Vestigial anatomic structure` be a class, as proposed by [103], or an attribute of existing classes? In light of the biological reality which such a change represents, which representation provides the proper graph difference, as opposed to a false over- or underestimation of its significance in evolutionary transformation? How do we integrate the knowledge emerging from the non-isomorphic mappings into a sound and complete description of complex and partial homologies? These are just a few examples of many modeling issues that will be confronted in moving the research into other, more different, species of animal model.

Additionally, in order to move from describing species at the leaf node (as we have done to date) to describing animal models at the level of abstraction used by comparative anatomists (*e.g.*, "insects", "vertebrates", "mammals"), and to create a true phylogenetic tree on that basis, we need to push the research in the direction of metamodels.

7.3 All Anatomy Is Comparative Anatomy: The Pan-Vertebrate Foundational Model of Anatomy (PVFMA)

At AMIA 2003, the vision of a Pan-Vertebrate Foundational Model of Anatomy (PVFMA) was introduced by Cornelius Rosse. This section of the dissertation outlines a roadmap toward such a pan-vertebrate model, and reviews the state of existing work toward that goal. Multi-species modeling and metamodeling are reviewed, and representative comparative anatomy queries are examined for commonalities that permit abstraction and representation of their underlying structure. This underlying structure is then used as a basis for classification of queries. In addition, by classifying the types of queries posed by different researchers using comparative anatomy, it provides preliminary desiderata for further work in integrating anatomy and informatics.

The term "pan-vertebrate", coined by Rosse, indicates that this model leverages off of the vertebrate *Bauplan*, or common structural similarities shared by animals with backbones. While not every single vertebrate has every single structure of the *Bauplan*, most vertebrates have some variation of most of them, including paired pectoral and pelvic limbs (the arms and legs of the human), five digits on the *pes* (human foot, animal hindpaw) and *manus* (human hand, animal forepaw), regional vertebrae (cervical, thoracic, lumbar, etc.), and a tripartite brain.

The fact that so many animals have variations on these same structures indicate that there is an advantage in leveraging off of those similarities, to ensure consistency among models, to promote efficiency of implementation of models by avoiding the duplication of modeling effort, and preventing the introduction of error in repeating the same work for different models. In other words, the phylogenetic shared similarities offer an opportunity for the development of anatomical metamodels to address the problem of computational complexity. Indeed, as mentioned earlier, the problem is practically intractable if modeled one species at a time.

So what makes the PVFMA special? As alluded to above, it is the first major step in modeling anatomical structures of species more further removed from the human. Many species of vertebrates have medically-important applications as animal models or in genome sequencing, so it is a productive place to begin development of models with practical applications. Perhaps most importantly, the FMA has been modeled in such an extensible way that much of the groundwork for the PVFMA has already been laid—the model has already successfully been extended to selected rat and mouse organs. The commonalities of the *Bauplan*, combined with the extensibility of the FMA, means that much of the groundwork for other vertebrate models has already been laid. The term PVFMA, therefore, is to be construed as a first stage of development, not as any kind of constraint on the ultimate scope of the possible modeling. A schematic of these stages of development and implementation is presented in Figure 7.1.

The first step in communicating the scope of such a potential comparative anatomy information system, and how the PVFMA model supports that scope, is to understand what kinds of questions and answers the users of a comparative anatomy information system

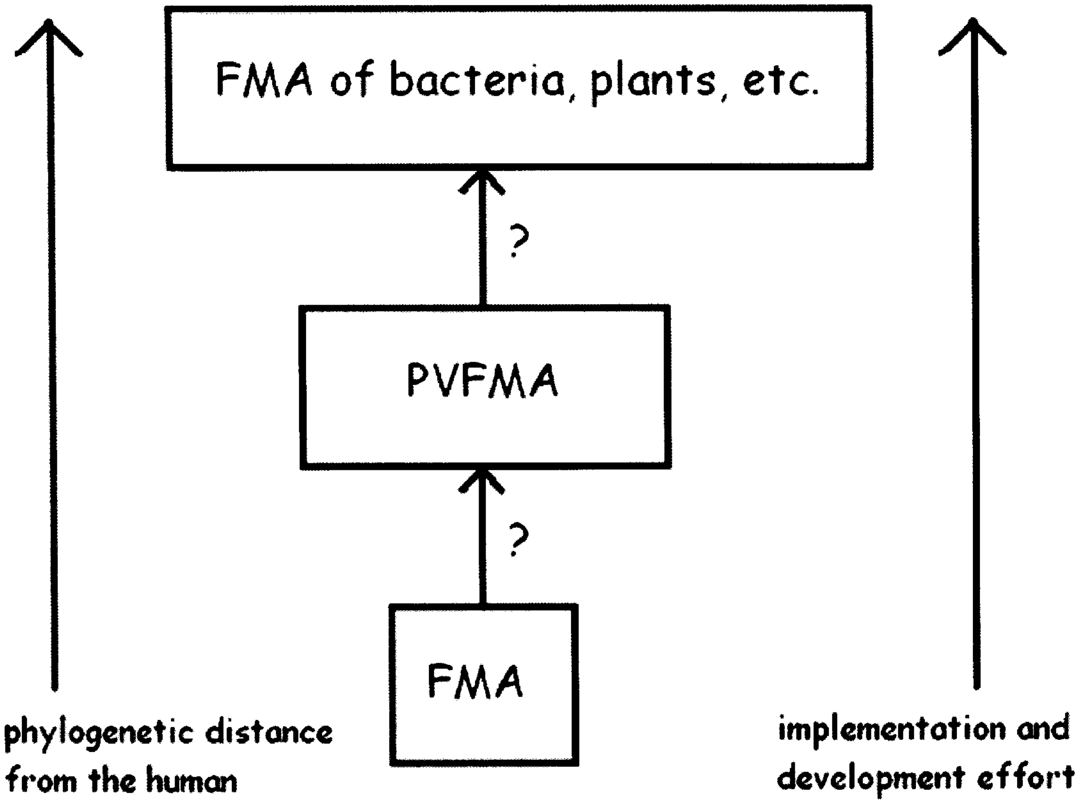

FMA of bacteria, plants, etc.
?
PVFMA
?
FMA
phylogenetic distance from the human
implementation and development effort

Figure 7.1: Stages in further modeling via the FMA.

want to know or query on. We present four sample questions, all of which are taken from senior-level comparative anatomy university exams.

1. Do crocodile epidermal scales **contain** β-keratin?

2. What structure in humans **is homologous to** the abomasum chamber of a cow's stomach?

3. How do human and dolphin/seal testes **differ**?

4. What are the unique reproductive structural traits **shared by** monotremes and marsupials?

The representative operations for this set of queries are then: "contain", "homologous to", "differ", and "shared by".

We collected hundreds of sample questions from comparative anatomy sources on the Web and in print. We are currently engaged in developing a classification for them, and while it would be premature to pronounce this classification as definitive, it is nevertheless the case that questions in the following categories recur very frequently, potentially indicating an ultimately useful categorization of comparative anatomical queries:

- Descriptive queries (*i.e.*, queries which return information explicitly stored in the knowledge base):
  - Description queries
    - Association queries: Is the gene Laminin associated with the mammary gland in the mouse? T
    - Component queries: Do bird feathers contain β-keratin? F
    - Distance queries: How do the pectoral girdle bones differ between *Aves* (birds) and therian mammals? AT: birds have Interclavicle; ASA: anterior Coracoid and mammals have posterior Coracoid
    - Evolutionary history queries: Name the muscle in the cat with the same evolutionary origin as the intermandibularis in the shark: Mylohyoid
    - Existence queries: Do male cats have a prostate gland as an organ? T
    - Gradient queries: In order from least to most, which type of uterus has the most fusion of its uterine horns? Simplex > Bicornuate > Bipartite > Duplex
    - Homology queries: What structure in humans is homologous to the abomasum chamber of a cow's stomach? Fundic & Pyloric regions of human Stomach
    - Simulation queries: What are the parts of a mammalian hepatic portal system in the order of blood flow after blood exits the celiac artery? Left

gastric artery, Gastrosplenic vein, Hepatic portal vein, Hepatic vein, Vena cava

  - Evaluation queries: What data is missing from the mouse ovary model?

  - Model development queries: Given the mouse model and the rat model, develop a tentative rodent metamodel for evaluation and verification.

- Inferential queries (*i.e.*, queries which use information explicitly stored in the knowledge base as a basis for reasoning in order to derive knowledge): In which taxa does the Archinephric duct transport urine in adults? *Chondrichthyes*, *Actinopterygii*, *Lissamphibia*

While elaboration on these categories is outside of the scope of this dissertation, they do tie into the roadmap for future work in the following way: obtaining the answers to particular types of queries in this classification scheme lends itself to the association of particular operations with particular types of queries, as indicated in bold above (*e.g.*, `shared by`, which necessarily presumes $n > 2$ species). In turn, some of those operations are inherently more closely associated with paired models, others with multiple ($n > 2$) models, and yet others with metamodels. In this way, our classification of queries is a first step to the specifications of what will be required to develop multiple, merged, and metamodels, such as the PVFMA—a component of what we refer to as an "anatomical algebra".

7.3.1 Validation

Validation issues about such an ambitious system are relatively easy to state, but will require a massive effort to implement. Perhaps the most interesting from an informatics point of view is what the implications of a non-monotonic knowledge system are for validation—in other words, when the experts do not agree and—pending more and better knowledge—the status quo is unclear and conflicting—what constitutes an appropriate validation of the relevant knowledge, and how is it to be carried out?

7.4 Summary

> "To understand the puzzles of the diversity of animal forms and development, Minelli points out that we need not only molecular developmental genetics, but also the theoretical tools of updated comparative morphology. As a comparative developmental morphologist, I could not agree more."—Paula M. Mabee [70]

In this dissertation, we have described the theoretical work we have carried out in comparative anatomy informatics, and the development and implementation of our system based on this theoretical foundation. Our system, CAIS, is a first step in the direction of the "theoretical tools of comparative morphology" that Mabee calls for, above. It currently compares anatomical structures across species two at a time, and—even more significantly—contains deliberate design decisions that will permit it, in conjunction with more theoretical work, to be extended to meet the needs of evolutionary developmental biologists to compare more species more different from each other across more phylogenetic space and time. Not just of abstract or aesthetic interest, understanding these similarities and differences better than we currently do is crucial to understanding the biomedical implications of animal models in health and disease.

As informaticists, we have a crucial role to fill in providing biologists with the tools for this task, because without automated tools to capture, organize, manage, visualize, and mine this vast amount of data, the task is overwhelming. CAIS is a very preliminary attempt to address this need, and because of decisions deliberately made in its design, it contains the capacity to nimbly and flexibly be extended in the different directions outlined in the desiderata for evo-devo and bioinformatics collaboration as outlined by Mabee [71]—in other words, CAIS has the capability to evolve to meet the biologists' information needs as we work together to establish, refine, and implement them.

BIBLIOGRAPHY

[1] F. I. Achike and C. W. Ogle. Information overload in the teaching of pharmacology. *Journal of Clinical Pharmacology*, 40(2):177–183, Feb 2000.

[2] Nadav Ahituv, Edward M. Rubin, and Marcelo A. Nobrega. Exploiting human–fish genome comparisons for deciphering gene regulation. *Human Molecular Genetics*, 13 Spec No-2:261–266, Oct 2004.

[3] Stuart Aitken. Formalizing concepts of species, sex and developmental stage in anatomical ontologies. *Bioinformatics*, 21(11):2773–2779, Jun 2005.

[4] Muhammad Al-Hajj and Michael F. Clarke. Self-renewal and solid tumor stem cells. *Oncogene*, 23(43):7274–7282, Sep 2004.

[5] Muhammad Al-Hajj, Max S. Wicha, Adalberto Benito-Hernandez, Sean J. Morrison, and Michael F. Clarke. Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):3983–3988, Apr 2003.

[6] Marc Aubry, Annabelle Monnier, Celine Chicault, Marie de Tayrac, Marie-Dominique Galibert, Anita Burgun, and Jean Mosser. Combining evidence, biomedical literature and statistical dependence: new insights for functional annotation of gene sets. *BMC Bioinformatics*, 7:241, 2006.

[7] Pedro Beltrao and Luis Serrano. Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. *PLoS Computational Biology*, 1(3):e26, Aug 2005.

[8] Richard N. Bergman. Pathogenesis and prediction of diabetes mellitus: lessons from integrative physiology. *Mount Sinai Journal of Medicine*, 69(5):280–290, Oct 2002.

[9] P.A. Bernstein, A.Y. Levy, and R.A. Pottinger. A Vision for Management of Complex Models. Check title: Model management: managing complex information structures. *Microsoft Research Technical Report MSR-TR-2000-53*, June 2000.

[10] F. Biering-Sorensen. Evidence-based medicine in treatment and rehabilitation of spinal cord injured. *Spinal Cord*, 43(10):587–592, Oct 2005.

[11] Leslie G. Biesecker. Phenotype matters. *Nature Genetics*, 36(4):323–324, Apr 2004. Comment.

[12] Blase Billack and Alvaro N. A. Monteiro. Methods to classify BRCA1 variants of uncertain clinical significance: the more the merrier. *Cancer Biology Therapy*, 3(5):458–459, May 2004. Comment.

[13] Olga O. Blumenfeld. Mutation databases and other online sites as a resource for transfusion medicine: history and attributes. *Transfusion Medicine Reviews*, 16(2):103–114, Apr 2002. Historical Article.

[14] Gary L. Bratthauer and Fattaneh A. Tavassoli. Lobular intraepithelial neoplasia: previously unexplored aspects assessed in 775 cases and their clinical implications. *Virchows Archiv: An International Journal of Pathology*, 440(2):134–138, Feb 2002.

[15] Jarle Breivik. The evolutionary origin of genetic instability in cancer development. *Seminars in Cancer Biology*, 15(1):51–60, Feb 2005.

[16] I. Brigandt. Conceptual role semantics, the theory theory, and conceptual change. In *First Joint Conference of the Society for Philosophy and Psychology and the European Society for Philosophy and Psychology.*, 2004.

[17] V. P. Cabral, F. S. Oliveira, M. R. Machado, A. A. Ribeiro, and A. M. Orsi. Study of lobation and vascularization of the lungs of wild boar (*Sus scrofa*). *Anatomia, Histologia, Embryologia*, 30(4):205–209, Aug 2001.

[18] W. Ceusters, B. Smith, and M. van Mol. Using ontology in query answering systems: scenarios requirements and challenges. In *Proceedings of the 2nd CoLogNET-ElsNet Symposium*, pages 5–15, 2003.

[19] Werner Ceusters, Barry Smith, Anand Kumar, and Christoffel Dhaen. Ontology-based error detection in SNOMED-CT. *Medinfo*, 11(Pt 1):482–486, 2004.

[20] I. R. Chambers, J. Barnes, I. Piper, G. Citerio, P. Enblad, T. Howells, K. Kiening, J. Matterns, P. Nilsson, A. Ragauskas, J. Sahuquillo, and Y. H. Yau. BrainIT: a transnational head injury monitoring research network. *Acta Neurochirurgica Supplement*, 96:7–10, 2006.

[21] Lifeng Chen and Carol Friedman. Extracting phenotypic information from the literature via natural language processing. *Medinfo*, 11(Pt 2):758–762, 2004. Evaluation Studies.

[22] Mayo Clinic. http://www.mayoclinic.org/breast-cancer/.

[23] Apelon Corporation. http://mmr.afs.apelon.com/contents.html#hierarchies. Accessed 25 June 2006.

[24] F. Coulier, C. Popovici, R. Villet, and D. Birnbaum. MetaHox gene clusters. *Journal of Experimental Zoology*, 288(4):345–351, Dec 2000.

[25] Bernard Crespi and Kyle Summers. Evolutionary biology of cancer. *Trends in Ecology Evolution*, 20(10):545–552, Oct 2005.

[26] L. F. da Costa. Return of de-differentiation: why cancer is a developmental disease. *Current Opinion in Oncology*, 13(1):58–62, Jan 2001.

[27] Jamie A. Davies. Do different branching epithelia use a conserved developmental mechanism? *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 24(10):937–948, Oct 2002.

[28] Landon T. Detwiler and James F. Brinkley. Custom views of reference ontologies. In *Proceedings, American Medical Informatics Association Fall Symposium, Bethesda, MD.*, 2006.

[29] Landon T. Detwiler, Emily Chung, Ann Li, Jose L. V. Jr Mejino, Augusto Agoncillo, James Brinkley, Cornelius Rosse, and Linda Shapiro. A relation-centric query engine for the Foundational Model of Anatomy. *Medinfo*, 11(Pt 1):341–345, 2004. Evaluation Studies.

[30] C. J. DiGiorgio, C. A. Richert, E. Klatt, and M. J. Becich. E-mail, the Internet, and information access technology in pathology. *Seminars in Diagnostic Pathology*, 11(4):294–304, Nov 1994.

[31] R. Doelz. Hierarchical Access System for Sequence Libraries in Europe (HASSLE): a tool to access sequence databases remotely. *Computer Applications in the Biosciences*, 10(1):31–34, Feb 1994.

[32] Gabriela Dontu, Muhammad Al-Hajj, Wissam M. Abdallah, Michael F. Clarke, and Max S. Wicha. Stem cells in normal breast development and breast cancer. *Cell Proliferation*, 36 Suppl 1:59–72, Oct 2003.

[33] W. Dooley. Surgery in breast cancer. *Current Opinion in Oncology*, 11(6):447–462, Nov 1999.

[34] K. M. Downs and T. Davies. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development*, 118(4):1255–1266, Aug 1993.

[35] B. A. Eckman, A. S. Kosky, and L. A. Jr Laroco. Extending traditional query-based integration approaches for functional characterization of post-genomic data. *Bioinformatics*, 17(7):587–601, Jul 2001.

[36] Edinburgh Mouse Atlas Project (EMAP). http://genex.hgu.mrc.ac.uk/. Accessed: 14 December 2004.

[37] G. Feldhamer, L. Drickamer, S. Vessey, and J. Merritt. *Mammalogy: Adaptation, Diversity, and Ecology.* Boston: McGraw-Hill, 1999.

[38] Foundational Model Explorer (FME). http://fme.biostr.washington.edu. Accessed 14 December 2004.

[39] Cheryl Frederick, Florence W. Patten, and Ravensara S. Travillian. Bearly Different? An Application of the Structural Difference Method to an Ursine Reproductive Conservation Initiative. *Unpublished*, 2006.

[40] Agla Jael Rubner Fridriksdottir, Rene Villadsen, Thorarinn Gudjonsson, and Ole William Petersen. Maintenance of cell type diversification in the human breast. *Journal of Mammary Gland Biology and Neoplasia*, 10(1):61–74, Jan 2005.

[41] G. Fusco. How many processes are responsible for phenotypic evolution? *Evolution Development*, 3(4):279–286, Jul 2001.

[42] F. Galis. On the homology of structures and Hox genes: the vertebral column. *Novartis Foundation Symposium*, 222:80–91, 1999.

[43] F. Galis. Why do almost all mammals have seven cervical vertebrae? Developmental constraints, Hox genes, and cancer. *Journal of Experimental Zoology*, 285(1):19–26, Apr 1999.

[44] E. M. Garabedian, P. A. Humphrey, and J. I. Gordon. A transgenic mouse model of metastatic prostate cancer originating from neuroendocrine cells. *Proceedings of the National Academy of Sciences of the United States of America*, 95(26):15382–15387, Dec 1998.

[45] Jordi Garcia-Fernandez. The genesis and evolution of homeobox gene clusters. *Nature Reviews Genetics*, 6(12):881–892, Dec 2005.

[46] Stanford Protege Group. http://protege.stanford.edu, accessed 10 June 2006.

[47] U. Hahn and S. Schulz. Towards a broad-coverage biomedical ontology based on description logics. *Pacific Symposium on Biocomputing*, pages 577–588, 2003.

[48] G. Halder, P. Callaerts, and W. J. Gehring. New perspectives on eye evolution. *Current Opinion in Genetics Development*, 5(5):602–609, Oct 1995.

[49] Milton Hildebrand. *Analysis of vertebrate structure (3rd ed.)*. New York: Wiley, 1988.

[50] Masanao Honda, Hidetoshi Ota, Showichi Sengoku, Hoi-Sen Yong, and Tsutomu Hikida. Molecular evaluation of phylogenetic significances in the highly divergent karyotypes of the genus Gonocephalus (Reptilia: Agamidae) from tropical Asia. *Zoological Science*, 19(1):129–133, Jan 2002.

[51] Masanao Honda, Yuichirou Yasukawa, Ren Hirayama, and Hidetoshi Ota. Phylogenetic relationships of the Asian box turtles of the genus Cuora sensu lato (Reptilia: Bataguridae) inferred from mitochondrial DNA sequences. *Zoological Science*, 19(11):1305–1312, Nov 2002.

[52] O. Hook. Scientific communications. History, electronic journals and impact factors. *Scandinavian Journal of Rehabilitation Medicine*, 31(1):3–7, Mar 1999. Historical Article.

[53] M. Ishaq. A morphological study of the lungs and bronchial tree of the dog: with a suggested system of nomenclature for bronchi. *Journal of Anatomy*, 131(Pt 4):589–610, Dec 1980.

[54] F. C. Izsak, T. Gotlieb-Stematsky, E. Eylan, and A. Gazith. Search for correlation between tissue cultures of human tumors and the clinical course of cancer in man. *European Journal of Cancer*, 4(4):375–381, Aug 1968.

[55] The Jackson Laboratory (JAX). http://www.jax.org. Accessed: 14 December 2004.

[56] J. David Johnson, Donald O. Case, James E. Andrews, and Suzanne L. Allard. Genomics–the perfect information-seeking research problem. *Journal of Health Communication*, 10(4):323–329, Jun 2005.

[57] Craig E. Jones, Ute Baumann, and Alfred L. Brown. Automated methods of predicting the function of biological sequences using GO and BLAST. *BMC Bioinformatics*, 6:272, 2005. Evaluation Studies.

[58] C. Y. Kao, K. Nomata, C. S. Oakley, C. W. Welsch, and C. C. Chang. Two types of normal human breast epithelial cells derived from reduction mammoplasty: phenotypic characterization and response to SV40 transfection. *Carcinogenesis*, 16(3):531–538, Mar 1995.

[59] Kenneth Kardong. *Vertebrates: Comparative Anatomy, Function, Evolution.* McGraw-Hill, 2001.

[60] S. Kim, J. F. Brinkley, and C. Rosse. Design features of on-line anatomy information resources: a comparison with the Digital Anatomist. *Proceedings: American Medical Informatics Association Annual Symposium*, pages 560–564, 1999.

[61] S. Kim, J. F. Brinkley, and C. Rosse. Profile of on-line anatomy information resources: design and instructional implications. *Clinical Anatomy*, 16(1):55–71, Jan 2003.

[62] Asako Koike and Toshihisa Takagi. PRIME: automatically extracted PRotein Interactions and Molecular Information databasE. *In Silico Biology*, 5(1):9–20, 2005.

[63] A. S. Kondrashov. Comparative genomics and evolutionary biology. *Current Opinion in Genetics Development*, 9(6):624–629, Dec 1999.

[64] Shigeru Kuratani. Craniofacial development and the evolution of the vertebrates: the old problems on a new background. *Zoological Science*, 22(1):1–19, Jan 2005.

[65] M. D. Landry and W. J. Sibbald. From data to evidence: evaluative methods in evidence-based medicine. *Respiratory Care*, 46(11):1226–1235, Nov 2001.

[66] P. Langer. *The mammalian herbivore stomach: Comparative anatomy, function, evolution.* Stuttgart, New York: G. Fischer, 1988.

[67] S. Letovsky. Beyond the information maze. *Journal of Computational Biology*, 2(4):539–546, Winter 1995.

[68] Tangliang Li, Patricia C. M. O'Brien, Larisa Biltueva, Beiyuan Fu, Jinhuan Wang, Wenhui Nie, Malcolm A. Ferguson-Smith, Alexander S. Graphodatsky, and Fengtang Yang. Evolution of genome organizations of squirrels (Sciuridae) revealed by cross-species chromosome painting. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, 12(4):317–335, 2004.

[69] Gurutz Linazasoro. Recent failures of new potential symptomatic treatments for Parkinson's disease: causes and solutions. *Movement Disorders*, 19(7):743–754, Jul 2004.

[70] Paula M. Mabee. Removing finalism from developmental biology: Review of Minelli's The Development of Animal Form: Ontogeny, Morphology, and Evolution. *Bioscience*, 54(9):868–870, 2004.

[71] Paula M. Mabee. Integrating evolution and development: the need for bioinformatics in evo-devo. *Bioscience*, 56(4):301–309, 2006.

[72] Paula M Mabee and Michael Noordsy. Development of the paired fins in the paddlefish, Polyodon spathula. *J Morphol*, 261(3):334–344, Sep 2004.

[73] Paul C. Marker, Rajvir Dahiya, and Gerald R. Cunha. Spontaneous mutation in mice provides new insight into the genetic mechanisms that pattern the seminal vesicles and prostate gland. *Developmental Dynamics*, 226(4):643–653, Apr 2003.

[74] Paul C. Marker, Annemarie A. Donjacour, Rajvir Dahiya, and Gerald R. Cunha. Hormonal, cellular, and molecular control of prostatic development. *Developmental Biology*, 253(2):165–174, Jan 2003.

[75] Toshiyuki Matsuoka, Per E. Ahlberg, Nicoletta Kessaris, Palma Iannarelli, Ulla Dennehy, William D. Richardson, Andrew P. McMahon, and Georgy Koentges. Neural crest origins of the neck and shoulder. *Nature*, 436(7049):347–355, Jul 2005.

[76] J. Mayer and L. Piterman. The attitudes of Australian GPs to evidence-based medicine: a focus group study. *Family Practice*, 16(6):627–632, Dec 1999.

[77] E. Mayr. Uncertainty in science: is the giant panda a bear or a raccoon? *Nature*, 323(6091):769–771, Oct 1986.

[78] R. McEntire, P. Karp, N. Abernethy, D. Benton, G. Helt, M. DeJongh, R. Kent, A. Kosky, S. Lewis, D. Hodnett, E. Neumann, F. Olken, D. Pathak, P. Tarczy-Hornoch, L. Toldo, and T. Topaloglou. An evaluation of ontology exchange languages for bioinformatics. *Proceedings: International Conference on Intelligent Systems for Molecular Biology*, 8:239–250, 2000.

[79] M. O. Mosse, P. Linder, J. Lazowska, and P. P. Slonimski. A comprehensive compilation of 1001 nucleotide sequences coding for proteins from the yeast Saccharomyces cerevisiae (= ListA2). *Current Genetics*, 23(1):66–91, Jan 1993.

[80] P.Z. Myers. Modules and the promise of the evo-devo research program. Available at http://scienceblogs.com/pharyngula/2006/06/modules_and_the_promise_of_the.php. Accessed: 15 August 2006.

[81] S. Nakakuki. The bronchial tree and lobular division of the dog lung. *Journal of Veterinary Medical Science*, 56(3):455–458, Jun 1994.

[82] A. Narath. *Der Bronchialbaum der Saeugetiere und des Menschen. Eine vergleichend anatomische und entwicklungsgeschichtliche Studie.* Stuttgart: Bibliotheca Med., Abth. A, Anatomie, H. 3, S. 1-380. Taf. I-VII. Erwin Naegele, 1901.

[83] Bhagavathi A. Narayanan, Narayanan K. Narayanan, Brian Pittman, and Bandaru S. Reddy. Regression of mouse prostatic intraepithelial neoplasia by nonsteroidal anti-inflammatory drugs in the transgenic adenocarcinoma mouse prostate model. *Clinical Cancer Research*, 10(22):7727–7737, Nov 2004.

[84] Wenhui Nie, Jinhuan Wang, Patricia C. M. O'Brien, Beiyuan Fu, Tian Ying, Malcolm A. Ferguson-Smith, and Fengtang Yang. The genome phylogeny of domestic cat, red panda and five mustelid species revealed by comparative chromosome painting and G-banding. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, 10(3):209–222, 2002.

[85] Mark Noble and Joerg Dietrich. The complex identity of brain tumors: emerging concerns regarding origin, diversity and plasticity. *Trends in Neurosciences*, 27(3):148–154, Mar 2004.

[86] Mouse Models of Human Cancer Consortium. http://emice.nci.nih.gov/, accessed 10 June 2006.

[87] C.K. Ogden and I.A. Richards. *The meaning of meaning; a study of the influence of language upon thought and of the science of symbolism.* New York, Harcourt, Brace company, inc., 1925.

[88] International Committee on Veterinary Gross Anatomical Nomenclature. *Nomina Anatomica Veterinaria.* Ithaca NY: Distributed by Dept. of Veterinary Anatomy, Cornell University, 2004.

[89] Roberta A. Pagon, Peter Tarczy-Hornoch, Patricia K. Baskin, Joseph E. Edwards, Maxine L. Covington, Miriam Espeseth, Christine Beahler, Thomas D. Bird, Bradley Popovich, Charli Nesbitt, Cynthia Dolan, Kathi Marymee, Nancy B. Hanson, Whitney Neufeld-Kaiser, Gina McCullough Grohs, Tracy Kicklighter, Cynthia Abair, Audin Malmin, Matthew Barclay, and Rajasri Dharani Palepu. GeneTests-GeneClinics: genetic testing information for a growing audience. *Human Mutation*, 19(5):501–509, May 2002.

[90] DM. Palmer. Early developmental stages of the human lung. *The Ohio Journal of Science*, 36(2):69–79, March 1936.

[91] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Borresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, Aug 2000.

[92] Petros Petrou, Evangelos Pavlakis, Yannis Dalezios, Vassilis K. Galanopoulos, and Georges Chalepakis. Basement membrane distortions impair lung lobation and capillary organization in the mouse model for Fraser syndrome. *Journal of Biological Chemistry*, 280(11):10350–10356, Mar 2005.

[93] Stephan Philippi. Light-weight integration of molecular biological databases. *Bioinformatics*, 20(1):51–57, Jan 2004. Evaluation Studies.

[94] P. Popesko. *A colour atlas of the anatomy of small laboratory animals.* London: Wolfe Publishing, 1992.

[95] C. Popovici, M. Leveugle, D. Birnbaum, and F. Coulier. Homeobox gene clusters and the human paralogy map. *FEBS Letters*, 491(3):237–242, Mar 2001.

[96] R.A. Pottinger and P.A. Bernstein. Merging models based on given correspondences. *University of Washington Technical Report UW-CSE-03-02-03*, 2003.

[97] D. Price. Comparative aspects of development and structure in the prostate. *National Cancer Institute Monograph*, 12:1–27, Oct 1963.

[98] B. Q. Qi and S. W. Beasley. Stages of normal tracheo-bronchial development in rat embryos: resolution of a controversy. *Development, Growth Differentiation*, 42(2):145–153, Apr 2000.

[99] A. Robert. Proposed terminology for the anatomy of the rat stomach. *Gastroenterology*, 60(2):344–345, Feb 1971.

[100] Rosario Rodriguez, Jose M. Pozuelo, Rocio Martin, Nuno Henriques-Gil, Maria Haro, Riansares Arriazu, and Luis Santamaria. Presence of neuroendocrine cells during postnatal development in rat prostate: Immunohistochemical, molecular, and quantitative study. *Prostate*, 57(2):176–185, Oct 2003.

[101] C Rosse, J L Mejino, B R Modayur, R Jakobovits, K P Hinshaw, and J F Brinkley. Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base. *J Am Med Inform Assoc*, 5(1):17–40, Jan 1998.

[102] Cornelius Rosse, Anand Kumar, Jose L V Jr Mejino, Daniel L Cook, Landon T Detwiler, and Barry Smith. A strategy for improving and integrating biomedical ontologies. *AMIA Annu Symp Proc*, 2005.

[103] Cornelius Rosse and Jose L. V. Jr Mejino. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6):478–500, Dec 2003. Evaluation Studies.

[104] P. Roy-Burman, H. Wu, W. C. Powell, J. Hagenkord, and M. B. Cohen. Genetically defined mouse models that mimic natural aspects of human prostate cancer development. *Endocrine-Related Cancer*, 11(2):225–254, Jun 2004.

[105] A Sanfeliu and K.S. Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3):353–362, 1983.

[106] S. Schulz, M. Romacker, and U. Hahn. Part-whole reasoning in medical ontologies revisited–introducing SEP triplets into classification-based description logics. *Proceedings: American Medical Informatics Association Annual Symposium*, pages 830–834, 1998.

[107] L.G. Shapiro. Relational matching. *Handbook of pattern recognition and image processing (vol. 2): computer visions, Orlando: Academic Press, Inc.*, pages 475–496, 1994.

[108] L.G. Shapiro and Haralick R.M. A Metric for Comparing Relational Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-7(1):90–94, 1985.

[109] T. Smith. Personal communication, 7 February 2006.

[110] American Cancer Society. http://www.cancer.org.

[111] J. R. Sommer. Comparative anatomy: in praise of a powerful approach to elucidate mechanisms translating cardiac excitation into purposeful contraction. *Journal of Molecular and Cellular Cardiology*, 27(1):19–35, Jan 1995.

[112] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–10874, Sep 2001.

[113] K. A. Spackman and K. E. Campbell. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. *Proceedings: American Medical Informatics Association Annual Symposium*, pages 740–744, 1998.

[114] E. Spanakis and D. Brouty-Boye. Discrimination of fibroblast subtypes by multivariate analysis of gene expression. *International Journal of Cancer*, 71(3):402–409, May 1997.

[115] S.S. Stevens. On the theory of scales of measurement. *Science, New Series*, 103(2684):677–680, Jun 1946.

[116] John Stingl, Afshin Raouf, Joanne T. Emerman, and Connie J. Eaves. Epithelial progenitors in the normal human mammary gland. *Journal of Mammary Gland Biology and Neoplasia*, 10(1):49–59, Jan 2005.

[117] T. Suwa, A. Nyska, J. C. Peckham, J. R. Hailey, J. F. Mahler, J. K. Haseman, and R. R. Maronpot. A retrospective analysis of background lesions and tissue accountability for male accessory sex organs in Fischer-344 rats. *Toxicologic Pathology*, 29(4):467–478, Jul 2001.

[118] Takahiko Suwa, Abraham Nyska, Joseph K. Haseman, Joel F. Mahler, and Robert R. Maronpot. Spontaneous lesions in control B6C3F1 mice and recommended sectioning of male accessory sex organs. *Toxicologic Pathology*, 30(2):228–234, Mar 2002.

[119] Takao Suzuki and Tatsuo Kasai. Morphological and embryological characteristics of bronchial arteries in the rat. *Anatomy and Embryology (Berlin)*, 207(2):95–99, Sep 2003.

[120] P. Tarczy-Hornoch, M. L. Covington, J. Edwards, P. Shannon, S. Fuller, and R. A. Pagon. Creation and maintenance of helix, a Web based database of medical genetics laboratories, to serve the needs of the genetics community. *Proceedings: American Medical Informatics Association Annual Symposium*, pages 341–345, 1998.

[121] Ying Tian, Wenhui Nie, Jinhuan Wang, Malcolm A. Ferguson-Smith, and Fengtang Yang. Chromosome evolution in bears: reconstructing phylogenetic relationships by cross-species chromosome painting. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, 12(1):55–63, 2004.

[122] Ravensara S. Travillian. All Anatomy is Comparative Anatomy: Issues on the Path Toward A Pan-Vertebrate FMA *or* What's In My Dissertation, What's Not, and How To Tell the Difference. *Presentation to the Structural Informatics Group, University of Washington*, Oct. 2004.

[123] Ravensara S. Travillian. From homology to ontology : comparing anatomy across species with the structural difference method. *University of Washington thesis*, 2004.

[124] Ravensara S. Travillian, John H. Gennari, and Linda G. Shapiro. Of mice and men: design of a comparative anatomy information system. *AMIA Annual Symposium Proceedings*, pages 734–738, 2005.

[125] Ravensara S. Travillian, Cornelius Rosse, and Linda G. Shapiro. An approach to the anatomical correlation of species through the Foundational Model of Anatomy. *AMIA Annual Symposium Proceedings*, pages 669–673, 2003.

[126] RS. Travillian, K. Diatchka, TJ. Judge, K. Wilamowska, and LG. Shapiro. A Graphical User Interface for a Comparative Anatomy Information System: Design, Implementation and Uses. *AMIA Annual Symposium Proceedings*, 2006.

[127] Sacha A. F. T. van Hijum, Aldert L. Zomer, Oscar P. Kuipers, and Jan Kok. Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic Acids Research*, 33(Web Server issue):560–566, Jul 2005. Evaluation Studies.

[128] Rene Villadsen. In search of a stem cell hierarchy in the human breast and its relevance to breast cancer evolution. *APMIS: ACTA PATHOLOGICA, MICROBIOLOGICA, ET IMMUNOLOGICA SCANDINAVICA*, 113(11-12):903–921, Nov 2005.

[129] B. R. Wallau, A. Schmitz, and S. F. Perry. Lung morphology in rodents (Mammalia, Rodentia) and its implications for systematics. *Journal of Morphology*, 246(3):228–248, Dec 2000.

[130] David Warburton, Saverio Bellusci, Pierre-Marie Del Moral, Vesa Kaartinen, Matt Lee, Denise Tefft, and Wei Shi. Growth factor signaling in lung morphogenetic centers: automaticity, stereotypy and symmetry. *Respiratory Research*, 4:5, 2003.

[131] J.R. Wilcke, P. Livesay, and L. Freeman. http://snomed.vetmed.vt.edu/presentations.

[132] L. Xue, K. Yang, H. Newmark, and M. Lipkin. Induced hyperproliferation in epithelial cells of mouse prostate by a Western-style diet. *Carcinogenesis*, 18(5):995–999, May 1997.

[133] Takaho Yamada, Eiichi Suzuki, Fumitake Gejyo, and Tatsuo Ushiki. Developmental changes in the structure of the rat fetal lung, with special reference to the airway smooth muscle and vasculature. *Archives of Histology and Cytology*, 65(1):55–69, Mar 2002.

[134] Hongwei Yu, Andy Wessels, Jianliang Chen, Aimee L. Phelps, John Oatis, G. Stephen Tint, and Shailendra B. Patel. Late gestational lung hypoplasia in a mouse model of the Smith-Lemli-Opitz syndrome. *BMC Developmental Biology*, 4:1, Feb 2004.

Appendix A
DOMAIN EXPERT QUESTIONNAIRE

Comparative Anatomy Information System Content Questionnaire

(*Thank you very much for taking the time to fill out this questionnaire. The information you provide will help me to make my information system more useful to the comparative medicine research community, by ensuring that I include content that researchers in the field consider essential for such a system. Questions for you to answer are in normal type, questionnaire instructions or additional information are in italics.*)

1. What is your research about? (*The answer to this question can help define what comparative anatomy content is relevant to the knowledge base we are building.*)

2. Given an anatomical structure of the mouse and of the human, the system I am building is able to answers queries such as:

- How are they similar?
- How are they different?
- What parts are common between mouse and human?
- What parts occur in one species but not in the other?
- What are all the structures that occur in either or both species?

Here are some examples in natural-language form:

- "How do the human and mouse prostates differ at the organ part level"? answer—the human prostate consists of 4 lobes: (Anterior lobe of prostate, Left dorsal lobe of prostate, Right dorsal lobe of prostate, Posterior lobe of prostate); the mouse prostate consists of 5 lobular organs: (Ventral prostate, Right dorsolateral prostate, Left dorsolateral prostate, Right coagulating gland, Left coagulating gland).
- "How do the mouse and human hearts compare at the organ level?" answer—the mouse and human hearts are made up of the identical configuration of chambers (left and right atrium; left and right ventricle)
- "Is the mouse left lung similar to the human left lung?" answer—no; the human left lung has 2 lobes; the mouse left lung has 1 lobe

Our system will translate them into a different structure at the underlying level, but you do not need to do the translation yourself. Please write as concisely yet completely as you can descriptive statements about anatomical structures and their spatial relationships that you would consider important to include. For example, you might write: "The coagulating glands are caudal to the ventral and dorsolateral prostates in the mouse", or "The HER2 receptors are embedded in the cell membrane of the epithelial cells of the mouse mammary ducts", depending on the level of anatomical structure (systemic, gross anatomical, microscopic, submicroscopic) that is most important to your work.

3. What content would you consider essential for the knowledge base to contain? (*For example, when someone is evaluating a dictionary, they usually have a list of words that they check to make sure the dictionary has a good definition for. If those words are missing, they don't even bother looking at the rest of the dictionary, because they already know it is inadequate. In an analogous way, what anatomical information on the mouse would you check up front to make sure the knowledge base has, in order to evaluate its content?*)

It would also be helpful if you could include a short description of ***why*** *you consider this knowledge important. For example, "this is important because the different spatial relationships between mouse* ***structure*** *and human* ***structure*** *lead to very different patterns of metastatic spread".*

(*Please feel free to copy and paste as many additional cells to this table as you need to in order to answer completely. Any anatomical information that you consider important enough to include here will be included in my system.*)

What mouse anatomical knowledge to include	**Why it is important**

4. *Optional*: Would you use queries about anatomical entities or about the relationships among those entities more? For example, is the query:
"What structure in the human corresponds to the murine left coagulating gland?"
or
"What is the difference in the arterial supply between the human prostate and the murine set of prostates?" more representative of a query useful to you?

Thank you very much for taking the time to fill out this questionnaire,

Ravensara S. Travillian

Appendix B

SUMMARY OF RESPONSES TO QUESTIONNAIRE

1. What is your research about?

Responses to question 1

- I am a veterinary pathologist primarily involved in the histopathologic evaluation of rodent tissues for research institutes, government agencies, and pharmaceutical companies. My previous research focused on the evaluation of wild type mouse mammary gland with comparison to genetically modified mice.
- My primary focus has been to investigate the immune response in mice after vaccination or infection with live influenza virus. During my experiments we have vaccinated i.m. or infected the mice i.n. and harvested organs like lungs, spleen, and bone marrow as well as blood.

2. Please write as concisely yet completely as you can descriptive statements about anatomical structures and their spatial relationships that you would consider important to include.

Responses to question 2

- Although I have limited experience in the microscopic anatomy of the human mammary gland, I have noticed that the periglandular stroma in the rodent mammary gland tends to have a much higher percentage of adipose tissue (including both white and brown fat) as opposed to collagenous tissue, which seems to be more prominent in humans. However, I would presume this can vary depending on age.
- The histologic appearance of the male and female rat mammary gland are significantly different. Females have more tubuloalveolar structures with frequent ducts, whereas

males exhibit more of a lobuloalveolar pattern. This sexual dimorphism is not present in the other species I've examined, such as the mouse and dog, and I do not believe it is present in humans.

- Another important issue with reproductive tissues in general is that they are very dynamic tissues that change with time (*e.g.*, developmental stage) and with the stages of the reproductive cycle. Appearance of the mammary gland also varies with pregnancy and with overall parity. There are also significant anatomical and functional differences between a lactating and non-lactating mammary gland. All of these issues should be considered when describing the comparative anatomy of the mammary gland, particularly considering the significant differences between the reproductive cycles of humans and most other animals.

- Of course, the anatomic location and the number of mammary glands varies between the rodent and human. Although primarily located along the ventral abdomen in the mouse, mammary gland tissue can be found in several other subcutaneous locations, including along the lateral or dorsal surfaces as evidenced by the occasional formation of mammary tumors in these locations.

- Mice usually have 5 pairs of mammary glands numbered 1 to 5 from anterior to posterior. Three pairs are in the cervicothoracic region and two are in the inguinoabdominal region. Males usually only have four pairs and do not have nipples.

- My work has revolved around the differences of the immune system, not the anatomical structures in itself. If you are interested in these differences as well, I would direct you to a very good review by P.J Haly in toxicology (2003).

- One difference that comes to mind, is the NALT, nasal-associated lymphatic tissue, which is often described as a highly-organised mucosal tissue involved in mounting a mucosal antibody response in the rodent's nose. There are no (detected) NALT in humans. The same goes for BALT, bronchial-associated lymphatic tissue.

- There is a difference in the antibody classes between man and mouse (different between rodents also!). In mouse serum IgG (IgG1, IgG2a, IgG2b, IgG3), IgM, IgA. In man IgG1, IgG2, IgG3 and IgG4 (no correspondence between mouse and man antibody subclass) and IgM, IgA (IgA1 and IgA2).

- The anatomy of the rodent nose is not similar to man. The mouse can for instance not breathe through the mouth as humans do.

3. What content would you consider essential for the knowledge base to contain?

Responses to question 3

- Duct (intralobular and interlobular)
- Ductule
- Alveoli
- Lobule
- Stroma
- Nipple
- Terminal end buds - if including developing mammary gland
- **Two major cell types:** Epithelial and Mesenchymal
- **Epithelial:** 3 types: cells lining ducts and alveoli and myoepithelial cells.
- **Mesenchymal:** the stroma consisting of adipose and fibrous connective tissue. Permeated by blood vessels and nerves.
- One of the primary concerns with comparison between rodent and human anatomy in my field is in determining common sites of neoplasia development. Are the sites comparable between human and mouse? Can the mouse be used as an adequate animal model for human disorders of the mammary gland? The anatomic similarities or differences at each stage of development may be important for answering this question.

- spleen: part of the immune system, similar in function
- lungs: among other things, part of the immune system
- BALT: differs greatly between species
- NALT: not in humans
- Bone marrow: part of the immune system, similar function in harbouring memory cells of the immune system
- Blood: distribution of bioactive substances (cytokines etc.) in rodents *contra* man
- Serum antibody: different subclasses, different in function
- Nose: not similar in function

4. *Optional*: Would you use queries about anatomical entities or about the relationships among those entities more?

Responses to question 4

- Possibly helpful for determining the potential significance of pathology in the rodent as it relates to human health.
- I would prefer to read about the similarities and differences. If possible I would like to see a lot of figures illustrating positions of veins, arteries and so on (photographs are not as good, in my opinion—too many details).

VITA

Ravensara Siobhán Travillian received B.A. degrees in French and German from the University of Alabama at Birmingham in 1980, an M.A. degree in Southeast Asian Studies from the University of Michigan in 1987, and an M.S. degree in Biomedical and Health Informatics from the University of Washington in 2004. She was a National Library of Medicine Informatics Research Fellow from 2002 to 2005, and is currently a predoctoral student at the Department of Medical Education and Biomedical Informatics at the University of Washington in Seattle. Her research interests include symbolic modeling of comparative anatomical information; information systems for comparative anatomy; applications of comparative anatomy informatics in cancer research, reproductive biology, and conservation biology; and knowledge representation in systematics and taxonomy. She expects to receive her Ph.D. degree in August 2006.