KB_Bio_101: Conceptual Modeling Challenges in Creating a Biology Knowledge Base

Vinay K. Chaudhri

Acknowledgment

- This work is funded by Vulcan Inc.
  - See http://www.vulcan.com and www.projecthalo.com
- AURA/Inquire Development Team
  - Eva Banik, Peter Clark, Roger Corman, Nikhil Dinesh, Debbie Frazier, Stijn Heymans, Sue Hinojoza, David Margolies, Adam Overholtzer, Aaron Spaulding, Ethan Stone, William Webb, Michael Wessel and Neil Yorke-Smith
  - Ashutosh Pande, Naveen Sharma, Rahul Katragadda, Umangi Oza

Outline

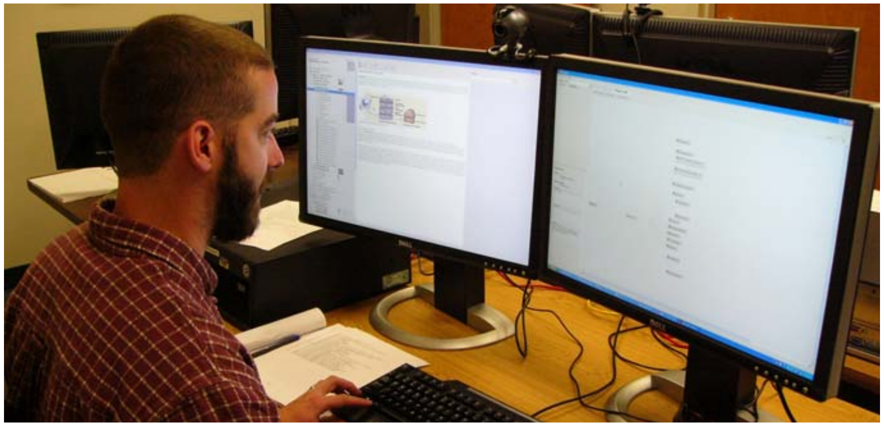
AURA

- Project Overview
- KB_Bio_101
  - Representing structure and function
    - Representation and Reasoning needs
    - Upper ontology
    - Representing structure
    - Representing function
    - Representing structure function relationship
    - Answering questions
- Some open research problems
- Summary

Project Overview

- Inquire Video
  - http://www.aaaivideos.org/2012/inquire_intelligent_textbook/

Project Overview

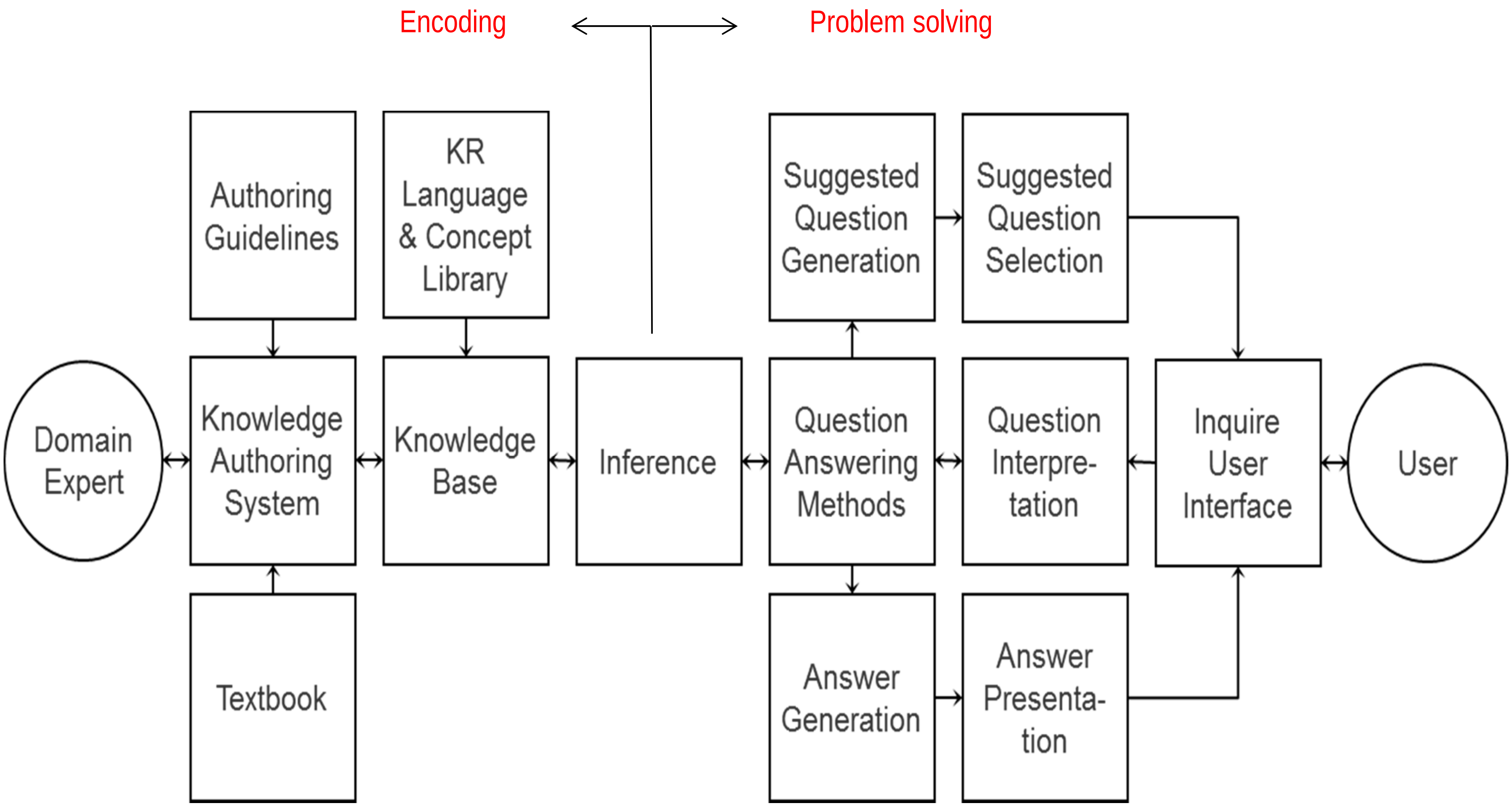


AURA
A knowledge authoring environment for domain experts

Inquire
Electronic book for biology students

System architecture

AURA



Significant effort devoted to the usability of answers by students

Knowledge Authoring in AURA

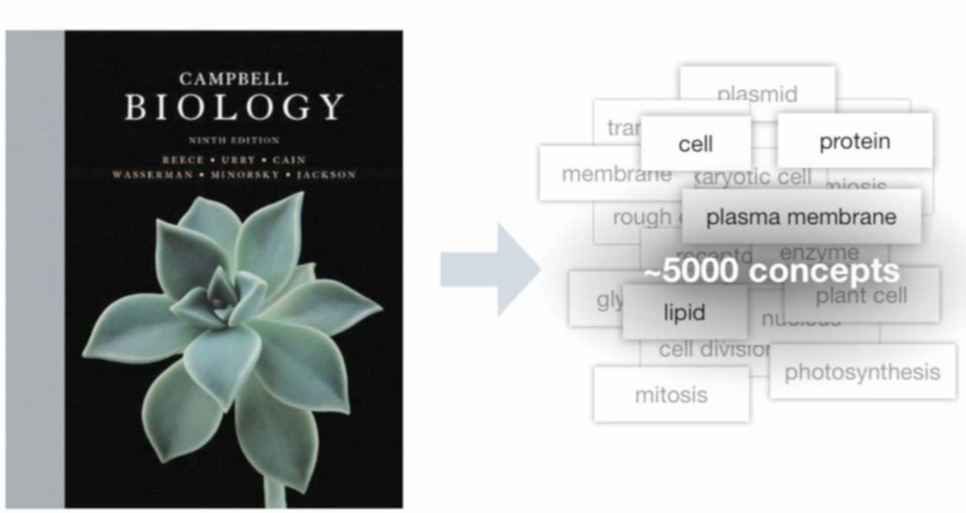
- Knowledge engineers provide a small library of domain independent representations
  - The Component Library (CLIB) contains classes representing physical actions, *e.g*., *Move, Attach, Penetrate*, and semantic relations, *e.g., agent, object, has-part (Barker, Clark, Porter, KCAP'01)*
    - See http://www.ai.sri.com/pub_list/864
- Biologists apply those representations to encode biology knowledge
  - AURA provides graphical editing
    - See http://www.ai.sri.com/pub_list/1545 and http://www.ai.sri.com/pub_list/865

Outline

- Project Overview
- KB_Bio_101
  - Representing structure and function
    - Representation and Reasoning needs
    - Upper ontology and the KR language
    - Representing structure
    - Representing function
    - Representing structure function relationship
    - Answering questions
- Some open research problems
- Summary

KB_Bio_101

AURA



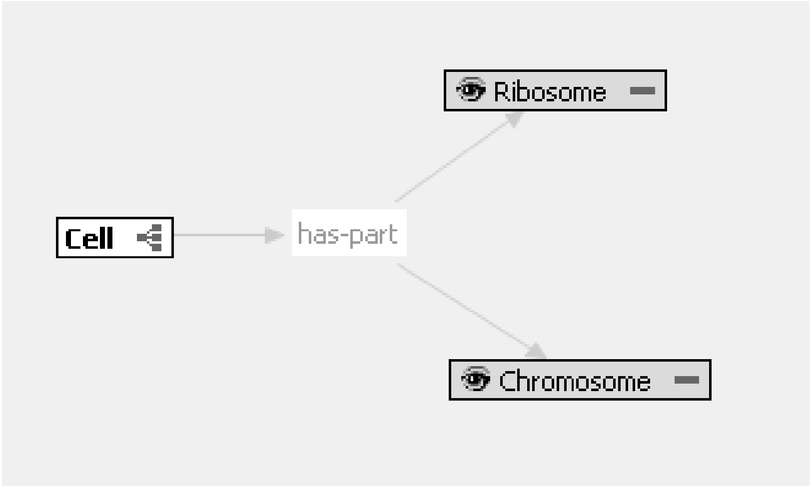
- Campbell Biology is a textbook used in an advanced placement biology course in schools and introductory biology courses in colleges
- A team of biologists used AURA to curate the KB from the textbook, using a sophisticated knowledge authoring process
- The KB is a valuable asset: it was created by an estimated 12 person years of encoding effort by biologists, and an estimated 5 person years of work on the upper ontology (CLib)
- Vulcan has released this asset free of charge for research purposes

  http://www.ai.sri.com/halo/halobook2010/exported-kb/biokb.html

KR in KB_Bio_101

- All the favorite features
  - Classes
    - Necessary and sufficient conditions
    - Disjoint-ness
    - Multiple Inheritance
  - Relations and Property Values
    - domain, range
    - inverse relations
    - transitivity
    - Relation hierarchy
    - Relation composition
    - qualified number restrictions
    - Nominals

Every Cell is a Living Entity and has a Ribosome and a Chromosome part



$$\forall x : Cell(x) \rightarrow \exists\, y_1, y_2 : LivingEntity(x) \wedge hasPart(x, y_1) \wedge hasPart(x, y_2) \wedge Ribosome(y_1) \wedge Chromosome(y_2)$$

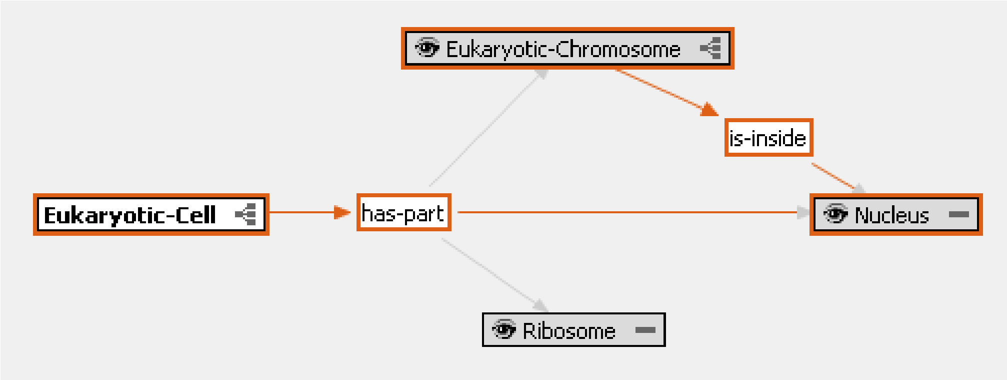
Cell ⊑ Living-Entity ⊓ (∃hasPart.Ribosome) ⊓ (∃hasPart.Chromosome)

KR in KB_Bio_101

AURA

- Unique Features
  - Graph structured descriptions

Every Eukaryotic Cell is a Cell and
has parts Eukaryotic Chromosome, Nucleus and a Ribosome such that
Eukaryotic Chromosome is inside the Nucleus



$$\forall x : EukaryoticCell(x) \rightarrow$$
$$\exists y_1, y_2, y_3 : Cell(x) \wedge$$
$$hasPart(x, y_1) \wedge hasPart(x, y_2) \wedge hasPart(x, y_3) \wedge$$
$$isInside(y_2, y_3) \wedge Ribosome(y_1) \wedge EukaryoticChromosome(y_2) \wedge$$
$$Nucleus(y_3)$$

The knowledge shown in red is not expressible in known decidable description logics such as OWL 2
This can be captured in Rule Languages

KR in KB_Bio_101

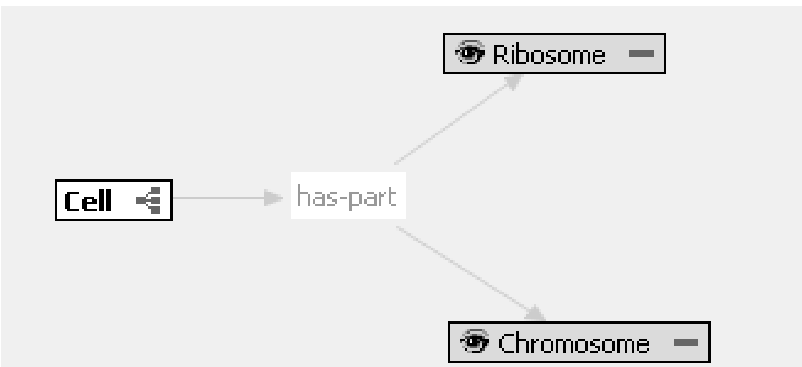
AURA

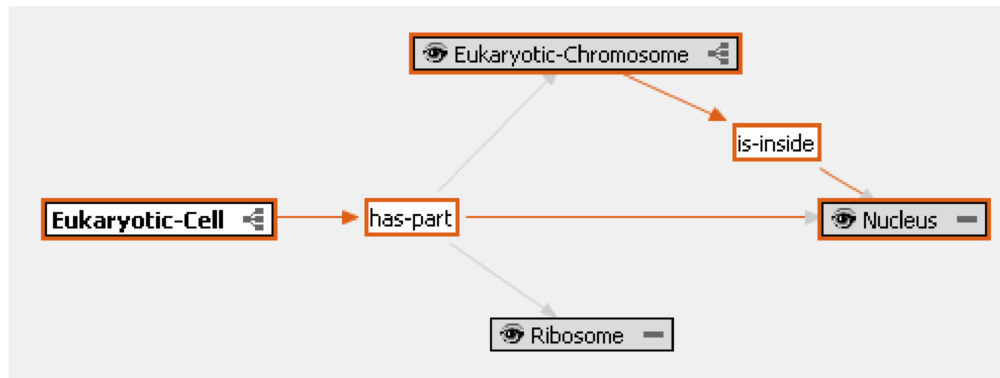
- Unique Features
  - Inherit and Specialize

In the Eukaryotic Cell

Ribosome was inherited from Cell

Chromosome was inherited from Cell and specialized to Eukaryotic Chromosome





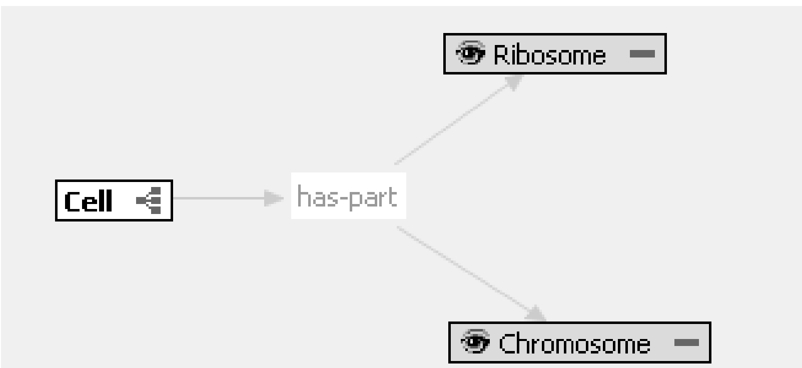
KR in KB_Bio_101

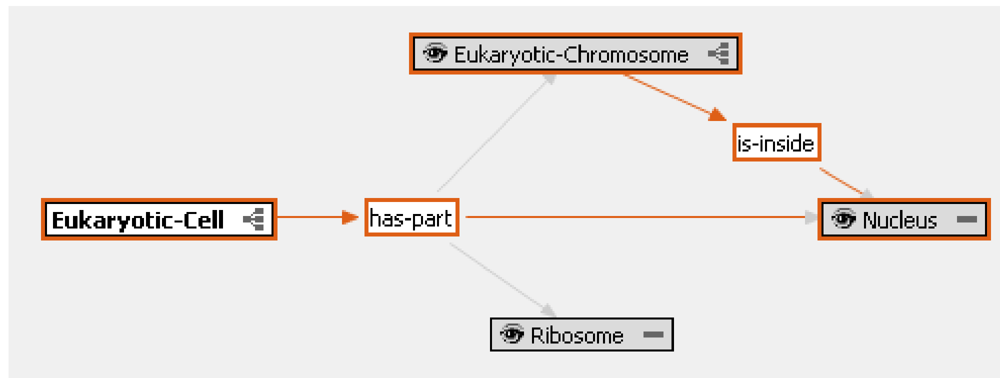
- Unique Features
  - Inherit and Specialize

In the Eukaryotic Cell

Ribosome was inherited from Cell

Chromosome was inherited from Cell and specialized to Eukaryotic Chromosome





$$\forall x : Cell(x) \rightarrow hasPart(x, f_{Cell}\#1(x)) \wedge hasPart(x, f_{Cell}\#2(x)) \wedge Ribosome(f_{Cell}\#1(x)) \wedge Chromosome(f_{Cell}\#2(x))$$

$$\forall x : ECell(x) \rightarrow hasPart(x, f_{ECell}\#1(x)) \wedge hasPart(x, f_{ECell}\#2(x)) \; hasPart(x, f_{ECell}\#3(x)) \wedge isInside(f_{ECell}\#2(x), f_{ECell}\#3(x)) \wedge Ribosome(f_{ECell}\#1(x)) \wedge Chromosome(f_{ECell}\#2(x)) \wedge Nucleus(f_{ECell}\#3(x))$$

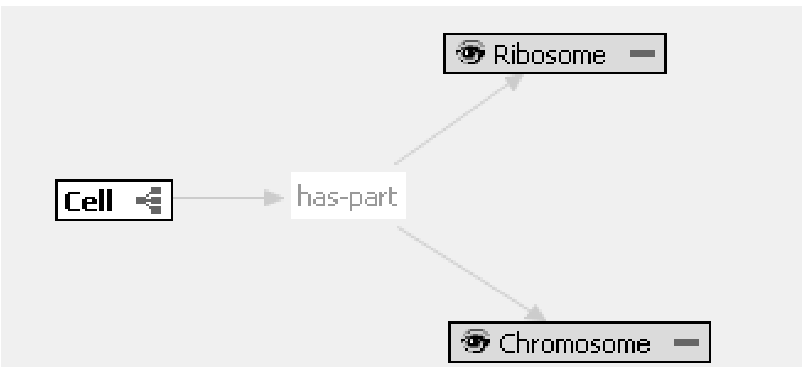
KR in KB_Bio_101

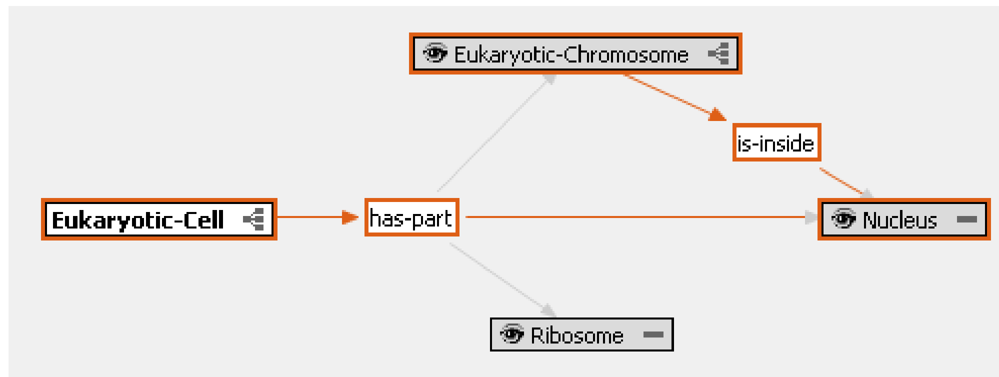
- Unique Features
  - Inherit and Specialize

In the Eukaryotic Cell

Ribosome was inherited from Cell

Chromosome was inherited from Cell and specialized to Eukaryotic Chromosome





$$\forall x : Cell(x) \rightarrow$$
$$hasPart(x, f_{Cell}\#1(x)) \wedge hasPart(x, f_{Cell}\#2(x))$$
$$\wedge\ Ribosome(f_{Cell}\#1(x))$$
$$\wedge\ Chromosome(f_{Cell}\#2(x))$$

$$\forall x : ECell(x) \rightarrow$$
$$hasPart(x, f_{ECell}\#1(x)) \wedge hasPart(x, f_{ECell}\#2(x))$$
$$hasPart(x, f_{ECell}\#3(x)) \wedge isInside(f_{ECell}\#2(x), f_{ECell}\#3(x))$$
$$\wedge\ Ribosome(f_{ECell}\#1(x)) \wedge Chromosome(f_{ECell}\#2(x))$$
$$\wedge\ Nucleus(f_{ECell}\#3(x))$$

$(f_{Cell}\#1(x)) = (f_{ECell}\#1(x))$ $(f_{Cell}\#2(x)) = (f_{ECell}\#2(x))$

KR in KB_Bio_101

- Computational properties
  - Reasoning with KR in KB_Bio_101 is, in general, un-decidable
  - There are, however, some decidable fragments that introduce guardedness and acyclic structure in the KB
    - Object Oriented Knowledge Bases in Logic Programming (Chaudhri, et. al., Technical Communication of the International Conference on Logic Programming, 2013)
    - Available at: http://www.ai.sri.com/pub_list/1958
  - A more thorough formal investigation is an open problem
    - Challenge for TPTP reasoners: http://www.ai.sri.com/pub_list/1937
    - Challenge for OWL reasoners: http://www.ai.sri.com/pub_list/1961

Knowledge Factory Process

- Goals
  - Avoid gaps in encoding
  - Ensure consensus
  - Catalogue conceptual modeling challenges

AURA

1) Determining Relevance and Pre-Planning

Pre-planning	Determining relevance, Diagram analysis, Pre-planning Status Labeling: Relevant, Irrelevant (closed)

2) Reaching Consensus

Universal Truth authoring, Concept chosen	QA check

3) Encoding Planning

Group common UTs, Identify KR/KE issues, Identify already encoded, Write how to encode	Planning, QA check Status Labeling: Encoding Complete, KR Issue (closed)

4) Encoding

Encode, File KR JIRA issues	QA check Status Labeling: Encoding Complete, KE Issue (closed)

5) Key Term Review

KR evaluated by modeling expert and SME, Encoder makes changes	KR evaluated by modeling expert and SME QA check

6) Question-Based Testing

Use Minimal Test Suite, File reasoning JIRA issues, Encoder fills KB gaps	QA check with screenshots of 'Passing' comparison and relationship questions

KB_Bio_101 Statistics

Regarding Class Axioms:

# Classes	# Relations	# Constants	Avg. # Skolems / Class	Avg. # Atoms / Necessary Condition	Avg. # Atoms / Sufficient Condition
6430	455	634	24	64	4

# Constant Typings	# Taxonomical Axioms	# Disjointness Axioms	# Equality Assertions	# Qualified Number Restrictions
714	6993	18616	108755	936

Regarding Relation Axioms:

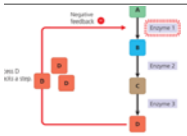
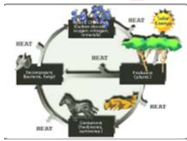
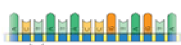
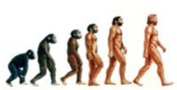
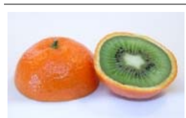
# DRAs	# RRAs	# RHAs	# QRHAs	# IRAs	# 12NAs / # N21As	# TRANS + # GTRANS
449	447	13	39	212	10 / 132	431

Regarding Other Aspects:

# Cyclical Classes	# Cycles	Avg. Cycle Length	# Skolem Functions
1008	8604	41	73815

Core Themes in Biology

AURA

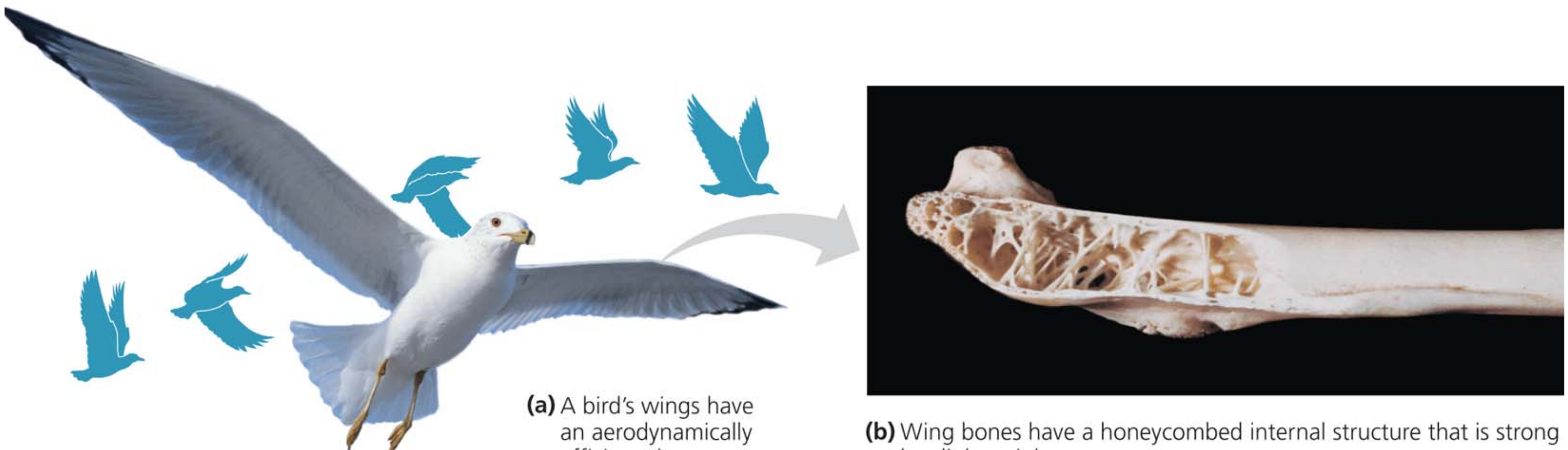
		Challenge
	Structure and Function	Relating structure and function
	Regulation	Qualitative reasoning about dynamic processes
	Energy Transfer	Representing energy production, consumption
	Continuity and Change	Representing genetic change across generations
	Evolution	Models of population dynamics
	Science as a Process	Experimentation and hypothesis testing
	Interdependence in Nature	Represent large inter-related complex systems
	Science, Technology, Society	Represent technological and social forces

Outline

- Project Overview
- KB_Bio_101
  - Representing structure and function
    - Representation and Reasoning needs
    - Upper ontology
    - Representing structure
    - Representing function
    - Representing structure function relationship
    - Answering questions
- Some open research problems
- Summary

Core Theme: Structure & Function

Structure and function are correlated at all levels of biological organization: *The form fits the function*



(a) A bird's wings have an aerodynamically efficient shape.

(b) Wing bones have a honeycombed internal structure that is strong but lightweight.

Computational Meaning of a Core Theme

- Identify the requirements in terms of a set of questions
  - Diagnostic questions
    - Help assess the basics of KR&R
  - Educationally useful questions
    - The question must be of interest to teachers and students
    - The question must be ``Google hard”
    - The question should not require solving an open-ended research problem

Diagnostic S&F Questions

- What is the structure of X?
- What is the function of X?

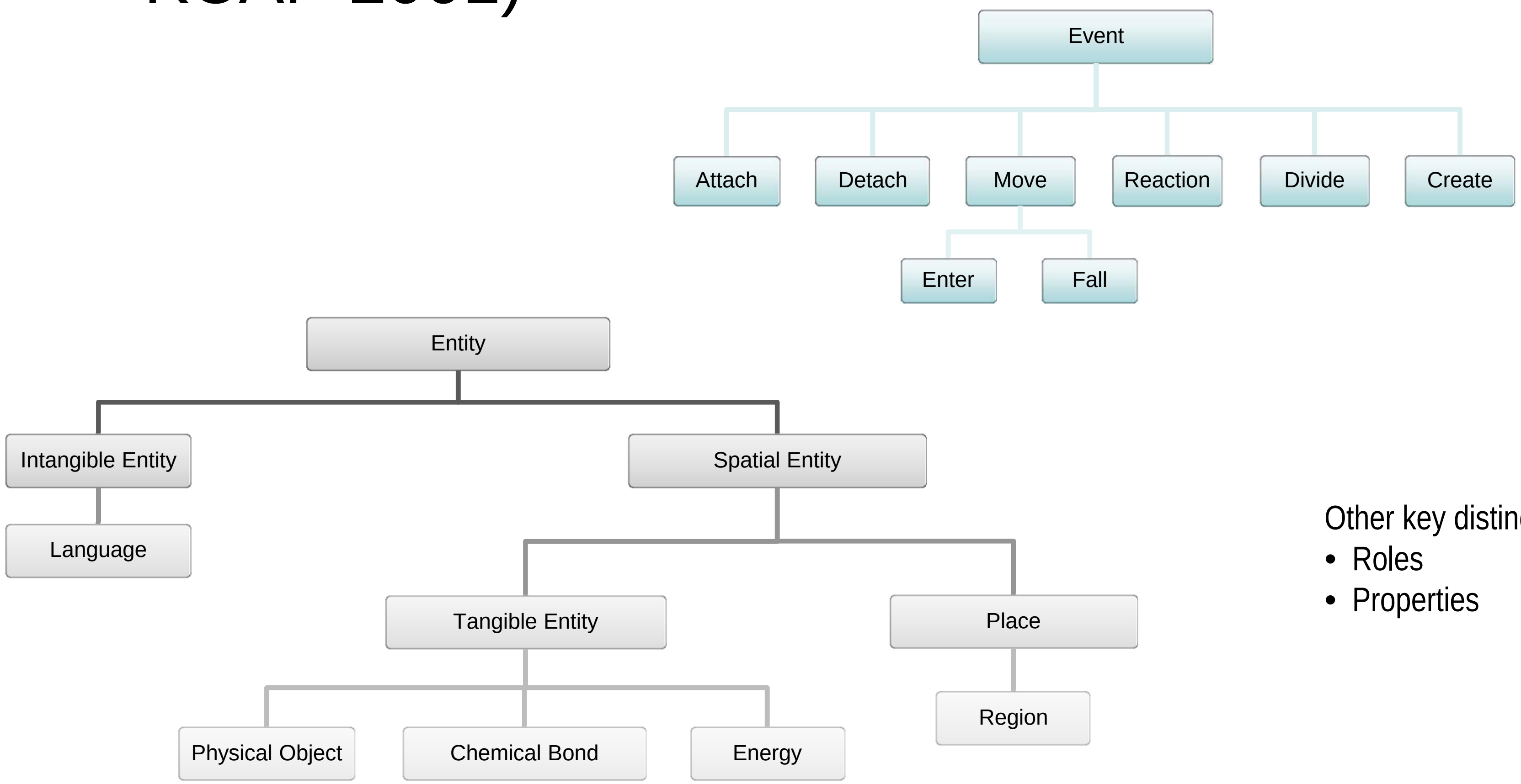
Educationally Useful Questions

- Relate Structures to Functions
  - What structure of Biomembrane facilitates a function of biomembrane, namely phagocytosis?
- Qualitative Comparisons
  - If the Loop of Henle gets longer, how will its function be impacted?
- Detailed Comparisons
  - What is the functional similarity between prions and viroids?
- Similarity Reasoning
  - Glucose is to Glycogen as ATP is to what?
- Negatively Modified Structures Impacting Functions
  - If hydrogen is removed from a saturated fatty acid, then how is its function impacted?

Starting Point - Component Library

AURA

- A simple upper ontology designed to be accessible to domain experts (Barker et. al, KCAP 2001)



Other key distinctions:

- Roles
- Properties

See http://www.ai.sri.com/halo/public/clib/20130328/clib-tree.html for more information

Component Library

AURA

- A vocabulary of relations to describe events

Event to Entity	Event to Event	Event to Value
agent	first-subevent	direction
object	next-event	distance
instrument	causes	duration
raw-material	enables	frequency
result	prevents	intensity
site	inhibits	rate
origin	by-means-of	

Representing Structure

AURA

- Structure of an entity represents its parts, their spatial arrangements and sizes

Meronymic	Spatial	Properties
has-part	is-at	length
has-region	is-inside	diameter
material	is-outside	height
possesses	abuts	area
element	is-between	depth
	is-along	volume

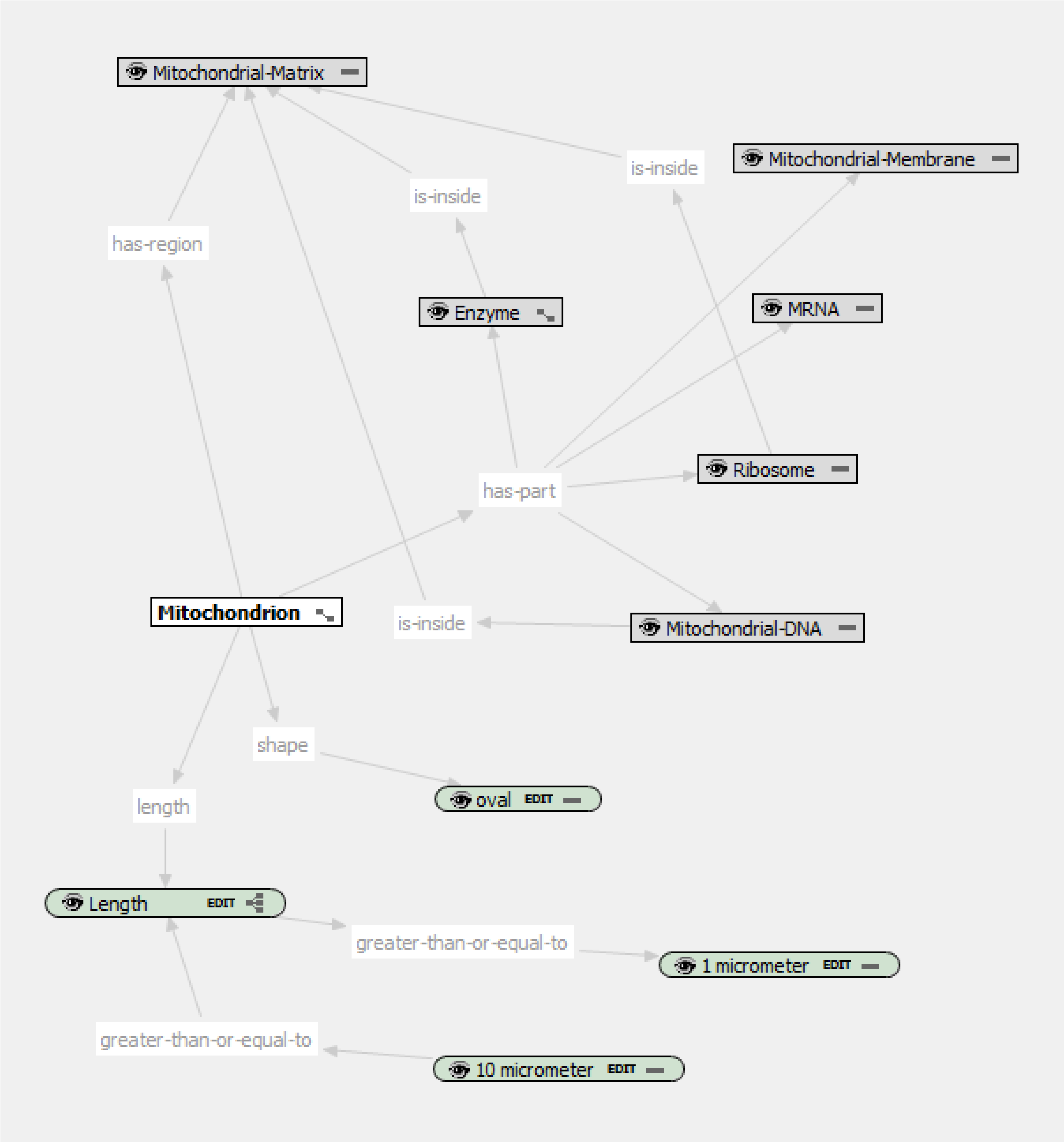
Choosing Structural Slots

AURA

- Inspired by work of Maria Keet, but simplified for use by biologists:
  - It must make sense to say ``X has Y” in English
  - X has-region Y if
    - Y is a region of space defined in relation to X
    - It does not make sense to associate Y with properties such as mass or density, but can be associated with measures such as length, area, or volume
  - X has material Y only if
    - Y is tangible and pervasive in X
  - X has element Y if
    - X is a set of entities of the same type (or sibling types) that Y is an instance of
  - X possesses Y only if
    - Y is Energy, bond or gradient
  - Otherwise X has part Y

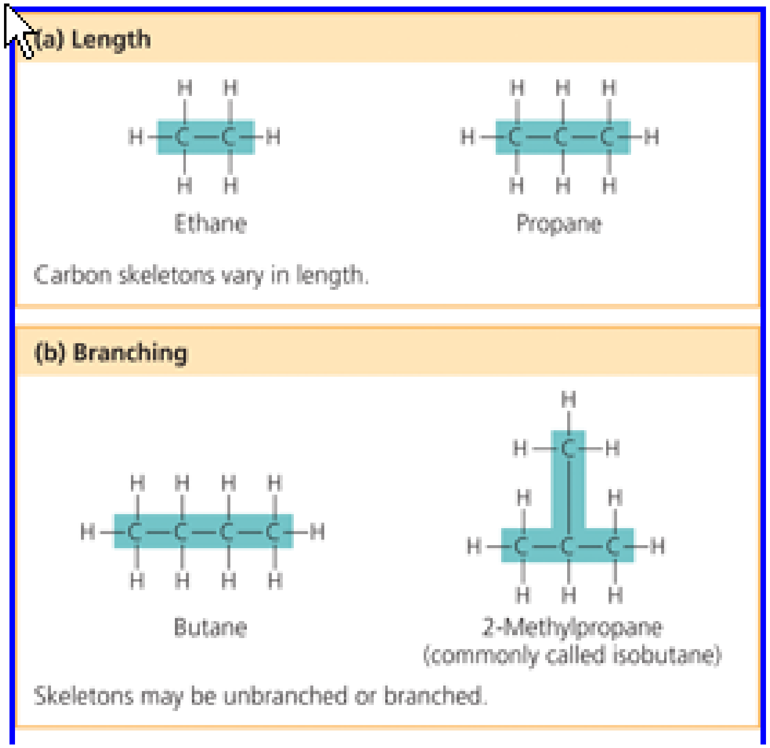
Example Structure Representation

AURA



A difficult example: Carbon Skeleton

What should be the relationship between an organic molecule and a skeleton?



It is more than simply a set of entities

-Can have length and shape

Is not an entity in its own right

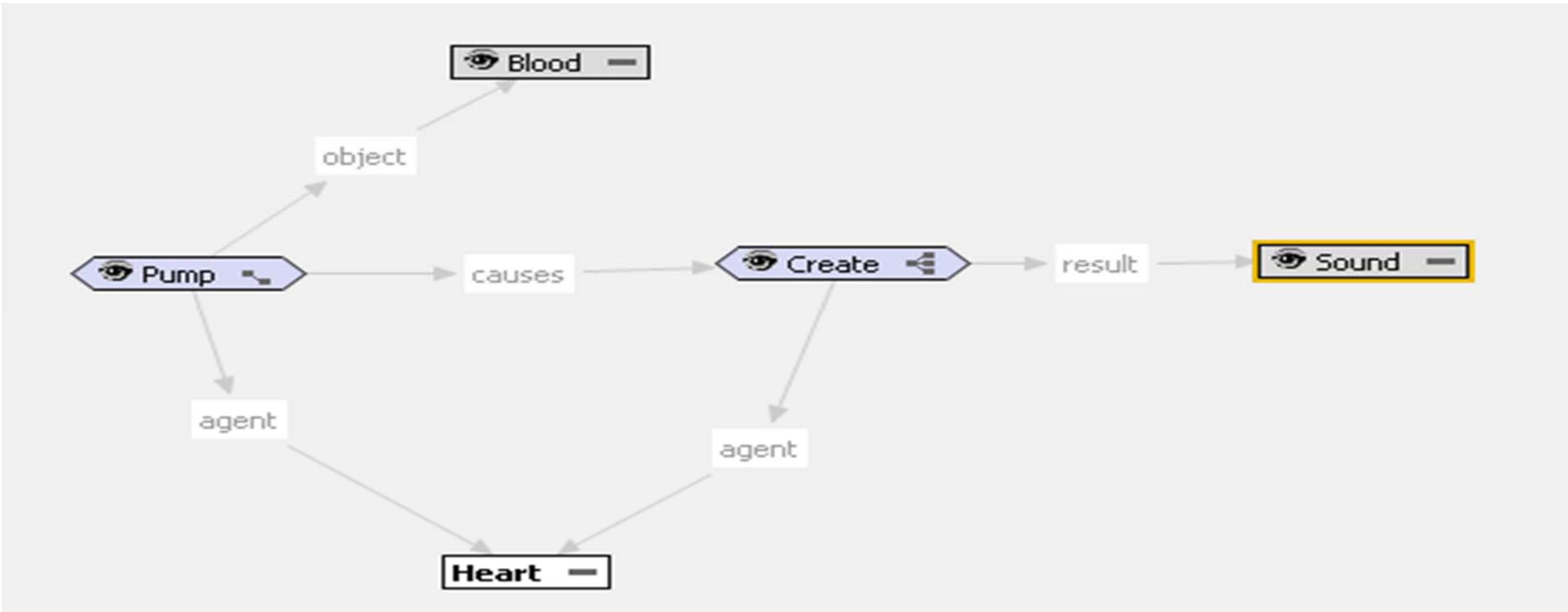
- Biologists do not associate mass with it

The remaining choice is has-region

-behaves differently than a human skeleton

Representing Functions

- Is function a primitive or a computed notion?

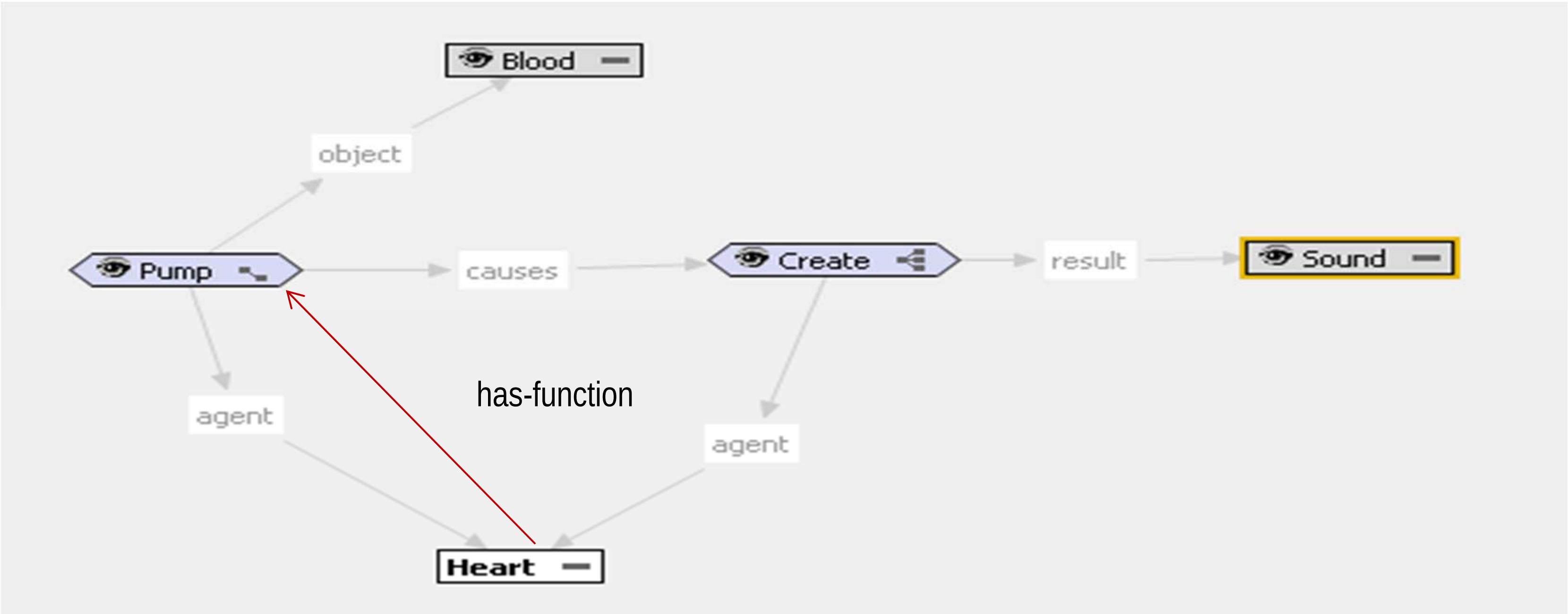


- Could function be inferred from participant relations, thus, reducing the encoding time?

Representing Functions

AURA

- Is function a primitive or a computed notion?



- It is a primitive notion and should be encoded by a biologist

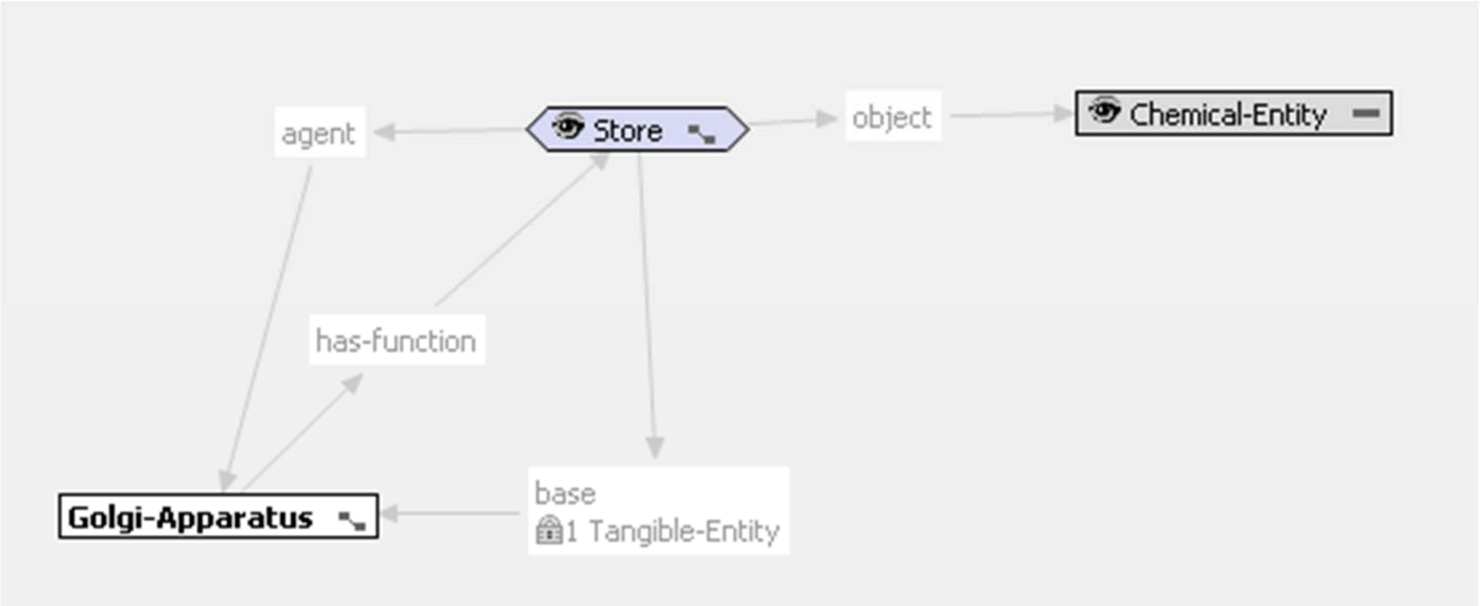
Representing Functions

- What is a function?
  - We understand functions as “special” events in which an entity participates
  - Alternatively, a function is an event which is a reason for an entity’s existence
  - The “special” nature of functions will be indicated by using a new slot called **has-function**

- Types of functions
  - Inherent functions of an entity
    - These will appear on the entity’s concept graph
  - Contextual functions of an entity
    - These will appear on *another* entity or event’s concept graph

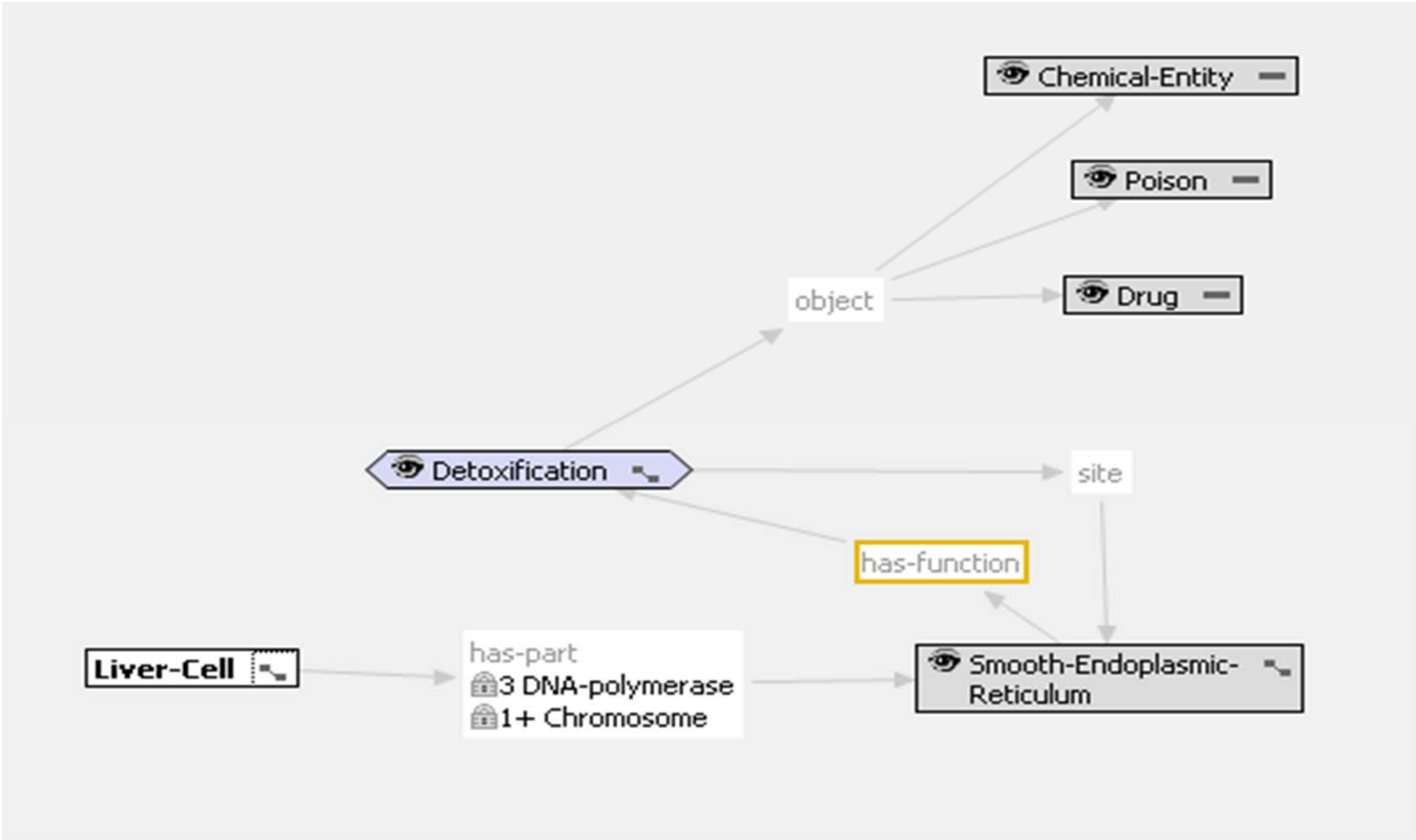
Example of an Inherent Function

- An inherent function of a Golgi Apparatus is to store chemicals
  - This is true regardless of which specific type of cell it is a part of
  - Inherent functions are placed on the Entity graph, using the has-function slot



Example of Function in an Environment

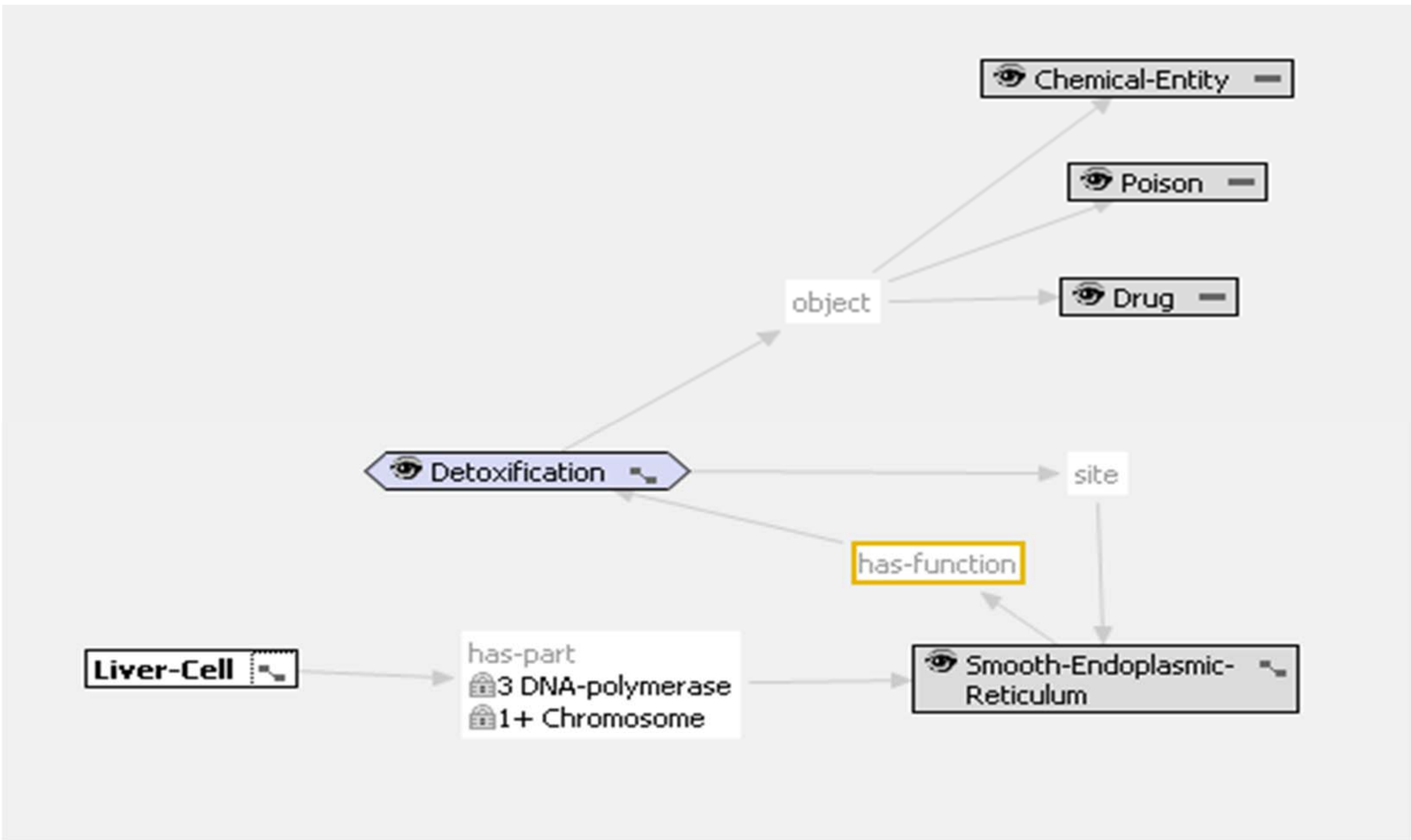
- Not every smooth ER detoxifies drugs
- However, drug detoxification is the function of a smooth ER in a liver cell



Structure Function Relationship

AURA

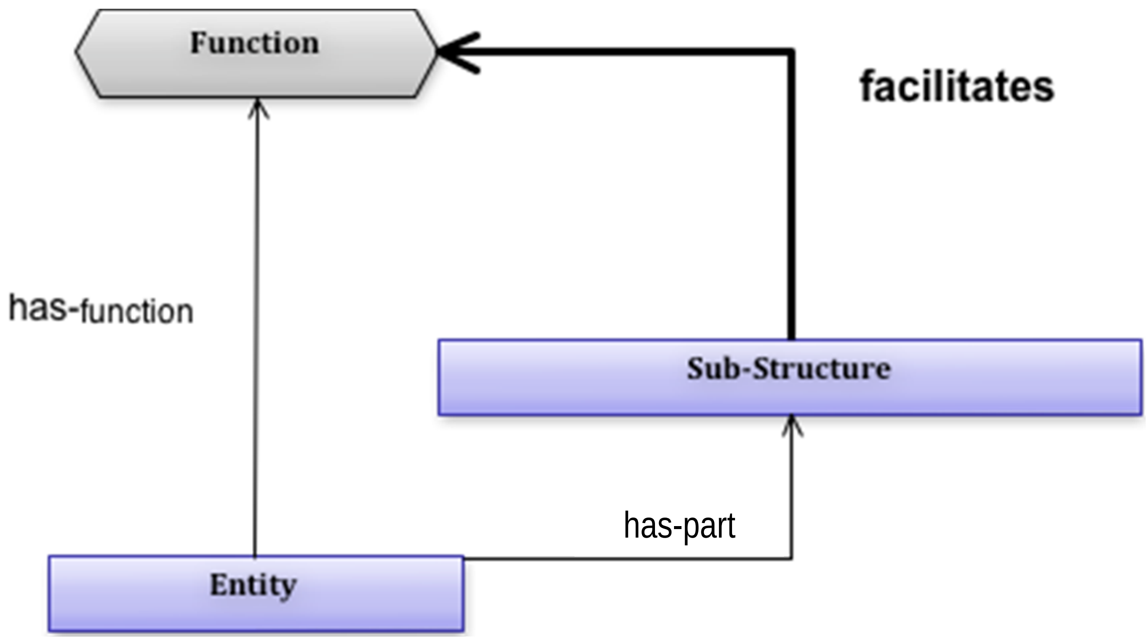
- We know how an entity participates in a function



Structure Function Relationship

AURA

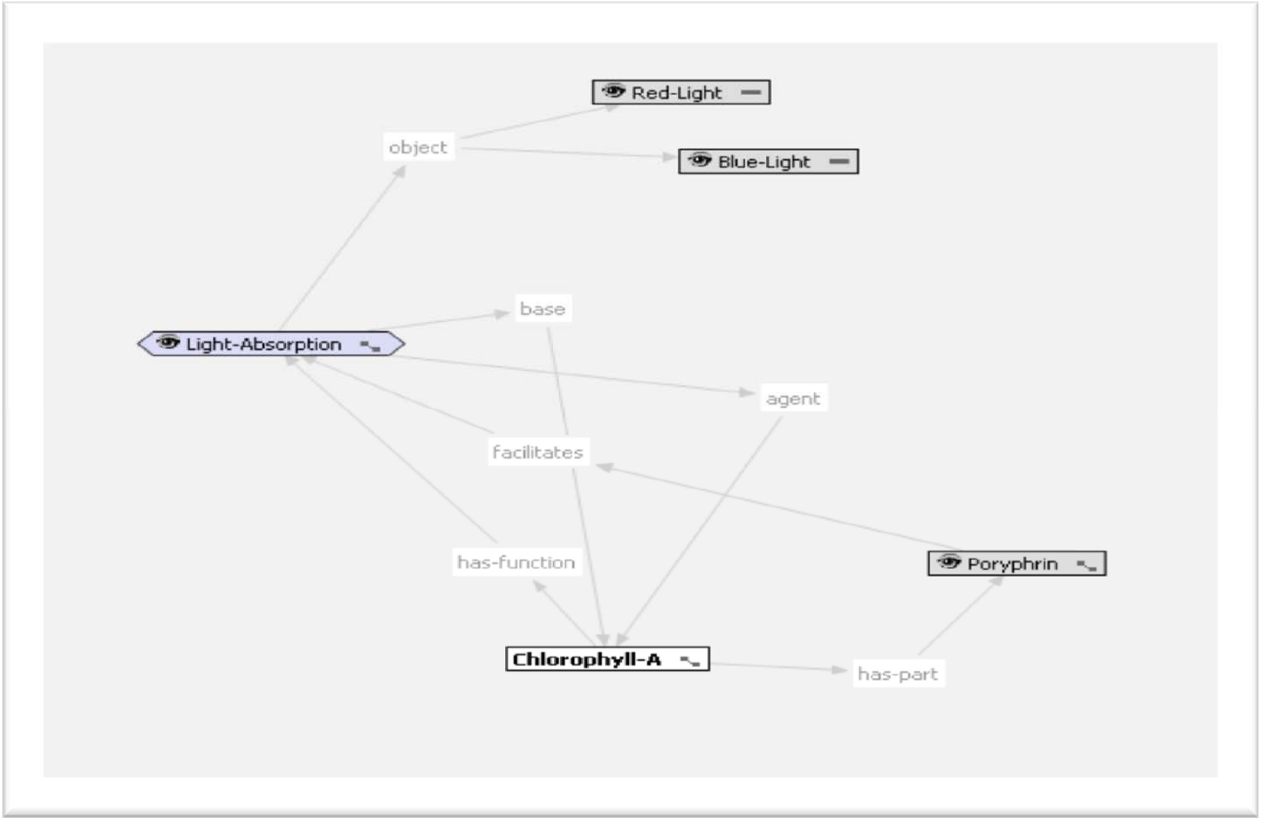
- We do not know how an entity participates in a function



For example, Chlorophyll-A contains Poryphrin. The textbook says that Poryphrin facilitates Chlorophyll-A's function of absorbing violet-blue light, but does not say how.

Structure Function Relationship

- We do not know how an entity participates in a function



Answering Questions

AURA

- Create an ABOX
  - Instantiate every concept in the knowledge base and compute the individuals it is related to up to depth three
- Conjunctive query answering
  - Reduce questions to conjunctive queries on an ABOX
- Path finding
  - Find all possible paths between two individuals
- Comparisons
  - Computer intersection and difference between two sets of triples

what is the structure of a biomembrane?

AURA

Biomembrane is a type of: membrane.

Structure of a biomembrane

Encloses

- carbon skeleton

Is between

- aqueous solution
- aqueous solution

Has region

- surface
- inside face
- outside face ▾

  An outside face is inside an aqueous solution

Possesses

- membrane potential

Has part

- membrane protein
- phospholipid bilayer
- glycoprotein ▾

  A glycoprotein is inside a phospholipid bilayer

- glycolipid
- saturated fatty acid

what are the functions of a biomembrane?

Inquire found this answer to your question about biomembrane:

Functions of a biomembrane

- A chemical is moved through a biomembrane
- A chemical is blocked by a biomembrane

Some functions of different types of biomembranes

- Plasma membrane 4 FUNCTIONS
- Outer membrane 1 FUNCTION

AURA

what are the functions of a biomembrane?

Inquire found this answer to your question about biomembrane:

Functions of a biomembrane

- A chemical is moved through a biomembrane
- A chemical is blocked by a biomembrane

Some functions of different types of biomembranes

- Plasma membrane HIDE
  - Transport of an organic molecule in a cell to an extracellular side with a transport vesicle from cytoplasm through a plasma membrane using another organic molecule. This process requires chemical energy and free-energy.
  - A cell is recognized by another cell at a plasma membrane
  - In receptor mediated endocytosis, directed motion — a chemical is moved in a plasma membrane through a coated pit, a cytoplasmic side, an extracellular side, a hydrophobic core and an intermembrane space
  - In coated pit, transport of a chemical by means of a cell surface receptor in a eukaryotic cell to cytoplasm inside a coated vesicle from an extracellular side using an organic molecule. This process requires chemical energy and free-energy.
- Outer membrane HIDE
  - In cell communication with mitochondrial signal leading to apoptosis, a semiautonomous organelle is confined by an outer membrane

what are the functions of a biomembrane?

AURA

Inquire found this answer to your question about biomembrane:

Functions of a biomembrane

- A chemical is moved through a biomembrane
- A chemical is blocked by a biomembrane

Some functions of different types of biomembranes

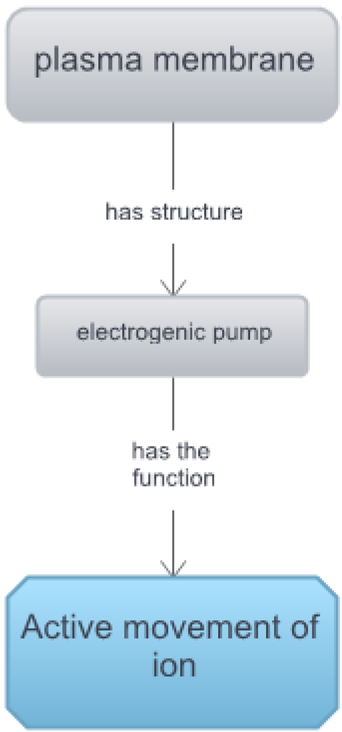
- Plasma membrane HIDE
  - Transport of an organic molecule in a cell to an extracellular side with a transport vesicle from cytoplasm through a plasma membrane using another organic molecule. This process requires chemical energy and free-energy.
  - A cell is recognized by another cell at a plasma membrane
  - In receptor mediated endocytosis, directed motion — a chemical is moved in a plasma membrane through a coated pit, a cytoplasmic side, an extracellular side, a hydrophobic core and an intermembrane space
  - In coated pit, transport of a chemical by means of a cell surface receptor in a eukaryotic cell to cytoplasm inside a coated vesicle from an extracellular side using an organic molecule. This process requires chemical energy and free-energy.
- Outer membrane HIDE
  - In cell communication with mitochondrial signal leading to apoptosis, a semiautonomous organelle is confined by an outer membrane

Generating good sentences is a research problem in natural language generation

See http://kbgen.org

Relate Structure to Function

What structures of a plasma membrane facilitate a function of the plasma membrane, namely active movement of ions?



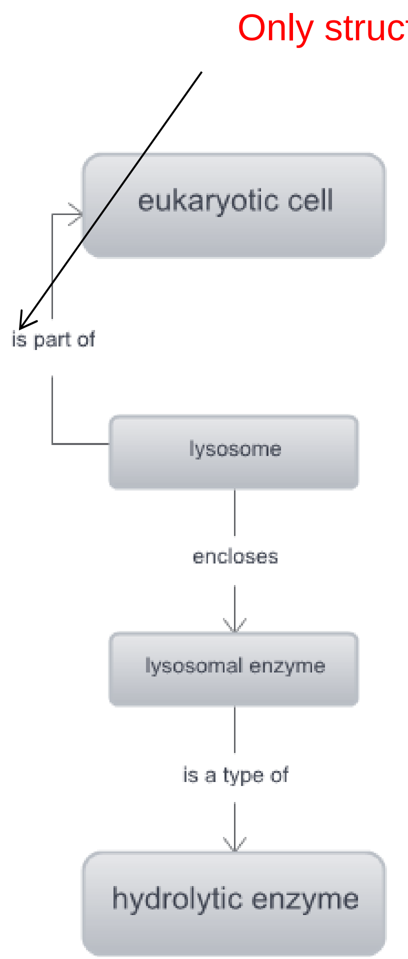
The membrane at the boundary of every cell that acts as a selective barrier, regulating the cell's chemical composition.

An ion transport protein that generates voltage across a membrane.

The proteins mediate the movement of ions to move across a membrane.

Structural and Functional Relationships

What is the structural relationship between hydrolytic enzymes and eukaryotic cells?



A type of cell with a membrane-enclosed nucleus and membrane-enclosed organelles. Organisms with eukaryotic cells (protists, plants, fungi, and animals) are called eukaryotes.

A membrane-enclosed sac of hydrolytic enzymes found in the cytoplasm of animal cells and some protists.

Lysosomal enzymes are synthesized in the cytosol and the endoplasmic reticulum. Some important lysosomal enzymes include lipase, carbohydrase, proteases, nucleases and phosphoric acid

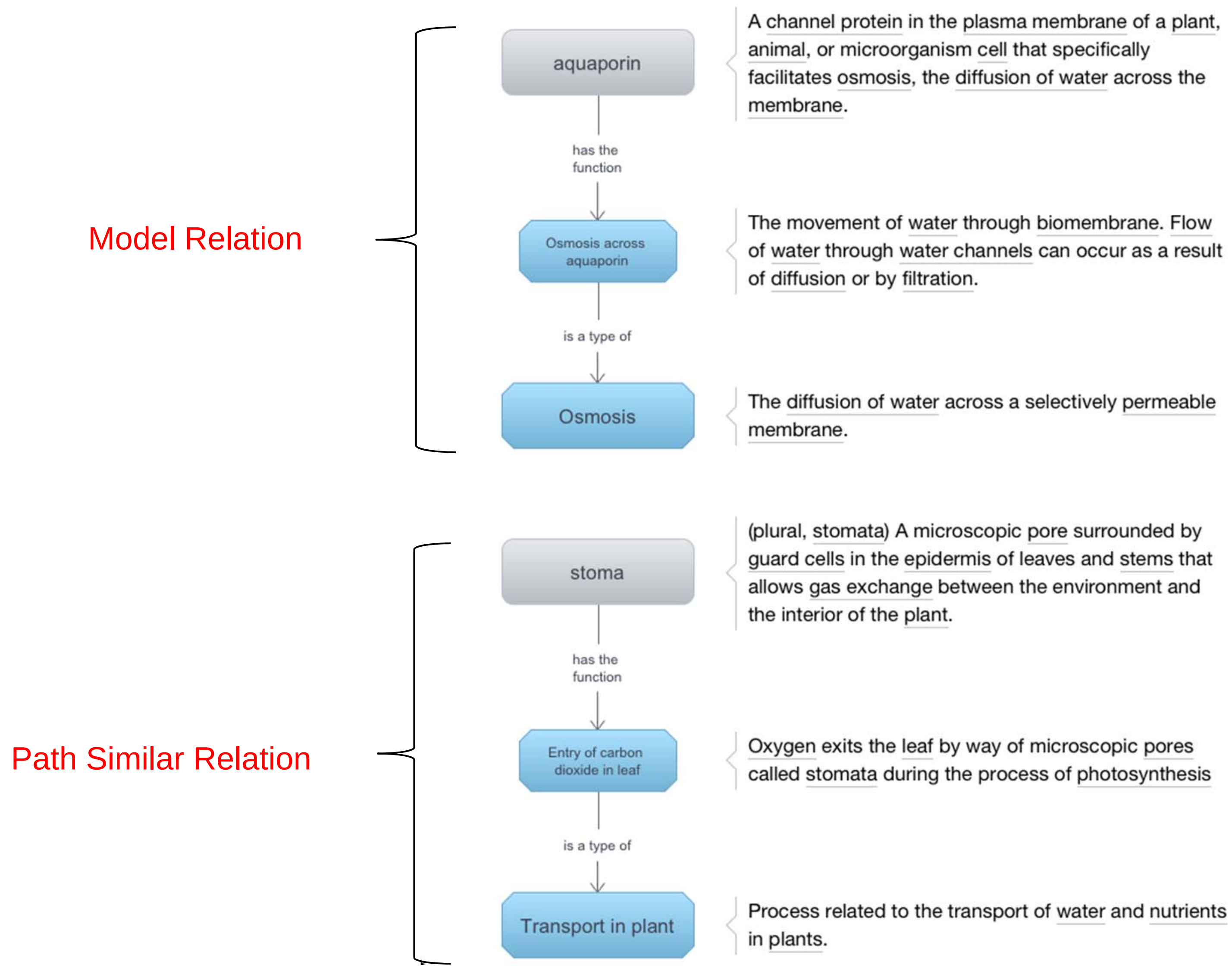
Hydrolytic enzymes are enzymes which are involved in break-down of biological molecules

Path-Based Similarity Reasoning

AURA

Aquaporin is to osmosis as stoma is to what?

Aquaporin, which is a hydrophobic substance is to Osmosis as stoma is to Transport in plant. Here are the similar relationships:



Open Research Problems

- What are some longer-term research problems?
  - Defining spatial slots for the whole book
    - Specifically, boundaries, regions and cavities
    - Preliminary work done by Bennett et. al. published at the 2013 Conference on Spatial Information Theory
      - Available at: http://www.ai.sri.com/pub_list/1959
  - Specifying the structure at multiple levels of detail and from multiple perspectives

Prior Work

AURA

- Structure, Behavior & Function (Chandrasekhran, 2000)
- Basic Foundational Ontology (Arp & Smith, 2008)
- General Formal Ontology (Herre, et. al., 2006)
- DOLCE (Borgo et. al. 2010)

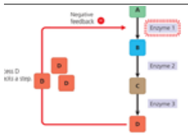
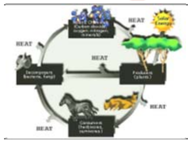
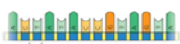
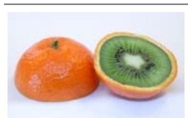
Outline

AURA

- Project Overview
- KB_Bio_101
  - Representing structure and function
    - Representation and Reasoning needs
    - Upper ontology
    - Representing structure
    - Representing function
    - Representing structure function relationship
    - Answering questions
- Some open research problems
- Summary

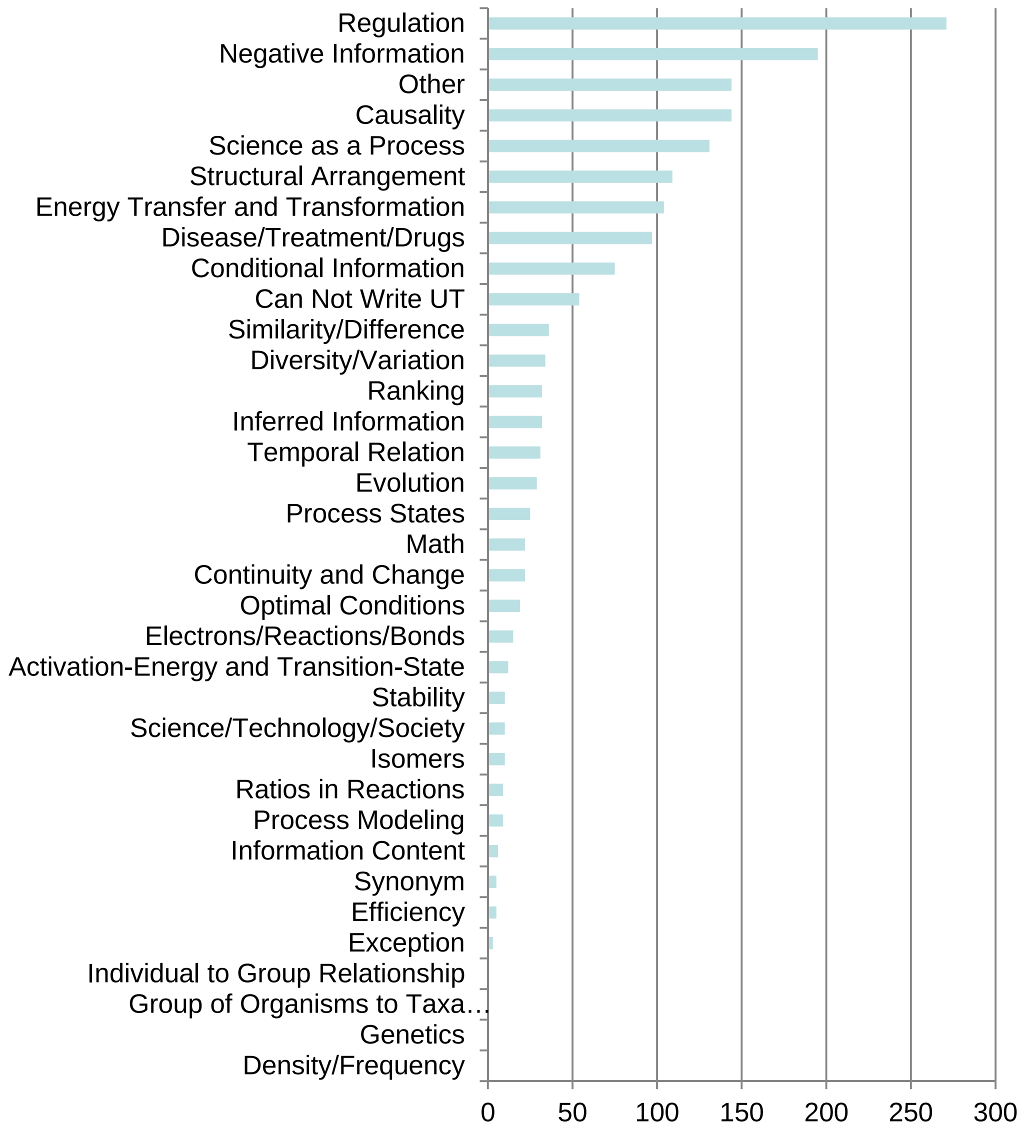
Conceptual Modeling Challenges

AURA

						Challenge
	Structure and Function					Relating structure and function
	Regulation					Qualitative reasoning about dynamic processes
	Energy Transfer					Representing energy production, consumption
	Continuity and Change					Representing genetic change across generations
	Evolution					Models of population dynamics
	Science as a Process					Experimentation and hypothesis testing
	Interdependence in Nature					Represent large inter-related complex systems
	Science, Technology, Society					Represent technological and social forces

The design and experimentation methodology that we showed for structure and function needs to be applied to all core themes

Conceptual Modeling Challenges



Number of Sentences from chapters 2-12

Frequently, the challenge lies in taking a piece of biological information and reducing it to a known modeling approach

Deep KR Challenge Workshop
https://sites.google.com/site/dkrckcap2011/

Ontology Management Challenges

AURA

- Ontology evaluation
  - Which parts of KB_Bio_101 are high quality? Which can be improved?
- Ontology modularization
  - Can we extract meaningful modules from KB_Bio_101?
- Ontology design patterns
  - What are repeating patterns of axioms in KB_Bio_101?
- Ontology testing and debugging
  - Is KB_Bio_101 consistent? What are maximally consistent subsets?

Summary

- KB_Bio_101
  - is a core component of an intelligent biology textbook
  - was authored using a state of the art modeling tool
  - a valuable resource for AI research on
    - efficient reasoning with hierarchical graphs
      - pushing the boundaries of logic programming and description logics
    - conceptual modeling techniques
    - ontology management methods
    - natural language generation
    - automated knowledge acquisition

Thank You!