Proceedings of the Second Annual Conference on Advances in Cognitive Systems ACS-2013 (255-271)

Conceptual Models of Structure and Function

Vinay K. Chaudhri VINAY.CHAUDHRI@SRI.COM
Nikhil Dinesh NIKHIL.DINESH@GMAIL.COM
Artificial Intelligence Center, SRI International, Menlo Park, CA, 94025, USA

H. Craig Heller HCHELLER@STANFORD.EDU
Department of Biology, Stanford University, Stanford, CA 94305 USA

Abstract

A student studying introductory biology at the college level faces the challenges of *both* mastering a large new vocabulary *and* integrating diverse, core concepts across different levels of biological organization. This task is made even more difficult due to the metaphoric language that is typical of introductory biology textbooks. In this paper, we report our achievements in creating conceptual, multidimensional models of biological structure and function out of the linear text of an introductory biology textbook. By formalizing biological terms and the relationships between them, we created a complex biological knowledge base (KB) that is embedded into the prototype of an intelligent textbook. The resulting system enables students to explore topics across multiple levels of organization and to pose their own questions that are answered by machine reasoning. This intelligent e-book is a powerful learning tool that gives students information such as definitions and descriptions of terms, but also enables them to explore structure, function, systems, and concepts across different levels of biological organization. When scaled to the full syllabus, this approach will bring computational thinking to the teaching of biology.

1. Introduction

The introductory biology student faces the daunting challenge of learning both a huge vocabulary and an enormous volume of factual information. In general, this situation leads to memorization being the predominant learning strategy for biology. To improve biology education, and to put less emphasis on memorization and more stress on understanding, knowledge application and problem solving are necessary (Brewer & Smith, 2009). Another challenge in biology education is integrating concepts across levels of biological organization and complexity. We can address these challenges by creating conceptual models from a textbook so that they can be readily accessed and explored by students as needed and as their curiosity demands. The focus of teaching biology can then shift away from mastering the facts to discovery and knowledge application. The resulting biology course will be active, outcome oriented, and inquiry driven.

Because the facts of biology are highly interconnected, they are not amenable to a straightforward database representation. We, therefore, leverage a knowledge representation that enables representing facts that can be organized into hierarchical graph structures (Chaudhri et al., 2013b). We will show how well-known vocabularies for defining structure and function can be applied and

adapted for this problem. Once represented, these conceptual models provide the basis for machine reasoning that can be used to answer questions.

Some examples of the questions that we considered include: "What structure of plasma membrane facilities movement of ions?" "Aquaporin is to Osmosis as Stoma is to what?" We have embeded the knowledge representation of the textbook knowledge and a capability of querying it into a prototype of intelligent textbook called *Inquire* (Chaudhri et al., 2013a). In our recent experiments with students, *Inquire* proved to provide an engaging learning experience for students and also improved their test scores. A detailed description of *Inquire* and the results of the educational study are available in an AI magazine article (Chaudhri et al., 2013a). Our focus in the current paper is to describe the knowledge representation of structure and function which is one of the key aspects of biology textbook knowledge.

Although the focus of the discussion in the current paper is on representing structure and function, a similar approach can be applied to all core themes of biological knowledge (for example, continuity and change, evolution, inter-dependence in nature, and science technology and society). Several current and previous efforts apply conceptual modeling to biological knowledge for the purposes of bio-medical research (Arp & Smith, 2008b; Karp, 2001) and for the publication of research articles (Renear & Palmer, 2009). The time is now ripe to introduce these techniques at the level of introductory biology, so that biology students are well-prepared for the computational thinking (Wing, 2006) that is both so vital to practitioners in today's knowledge economy and indispensable for researchers pursuing advanced bio-medical discoveries.

We begin the discussion with a short overview of how structure and function is currently taught in an introductory biology course. We then give a short background on our conceptual modeling approach. Next, we show our conceptual models for structure and function, and then illustrate how they can be used for machine reasoning in answering questions. We conclude the paper with a discussion on future work.

2. Structure and Function

Current biology textbooks introduce and teach the concepts of structure and function informally at best. Consider one example of how this topic is introduced:

> Structure and function are correlated at all levels of biological organization. For example, a leaf has a thin and flat shape that maximizes the amount of sunlight that can be captured by its chloroplasts. Analyzing biological structure gives us clues about what it does and how it works. Conversely, knowing the function of something provides insight into its construction. For example, the wings of a bird have the function of flying which is supported by their aerodynamic shape. The wing bones have a honeycomb internal structure that is strong but lightweight (Reece et al., 2011).

A description such as the above is quite unsatisfactory from a computational perspective. For example, is the thin and flat shape of a leaf its only structural characteristic? Are its cellular and molecular constituents also not its structure? Should the constituents that do not necessarily con-

tribute to a specific function of an entity also be considered as its structure? Modern biology text do not make such descriptions rigorous and precise.

2.1 Conceptual Modeling Terminology

We say that a knowledge base (or KB) is a collection of classes, relations, properties, and rules of inference (Brachman & Levesque, 2004). Each class is associated with a set of relations and properties and their values. Classes are organized into a class hierarchy such that the relation and property values can be inherited across the class hierarchy. Biology is full of exceptions and special cases. Such special cases can be captured by associating them with an appropriately specific class. Formal modeling languages provide mechanisms that enable such exceptions to be stated.

We use an upper ontology defining a small number of basic distinctions that provide the foundation of our conceptual models (Barker, Porter, & Clark, 2001). At the highest level, we distinguish between an Entity (e.g., a Cell) and a Event (e.g., transport). We use Event and Process interchangeably. An Entity is further sub-divided into TangibleEntity, Region and Spatial-Extent. A Tangible-Entity is an entity composed of material substance, having a spatial extent, and capable of independent existence. It also has properties such as mass and $\mathrm{density}$. A Region is a spatially extended entity whose existence depends on the existence of some Tangible-Entity or some group of Tangible-Entities. In related ontologies such as the Basic Foundational Ontology (BFO), a Region is also referred to as a dependent continuant (Spear, 2006). A Spatial-Extent is an entity that is independent of the existence of any material substance. It can have properties such as length, area, volume, but no mass or $\mathrm{density}$.

2.2 Modeling Structure

For creating a conceptual model, we take structure to mean an enumeration of constituents and their spatial arrangements in an entity. We will now introduce relations to capture this definition of structure.

2.2.1 Relations to represent constituents

The relations to represent constituents are often referred to as meronymic relations (Casati & Varzi, 1999). We use five such relations for representing structure: $\mathrm{has\text{-}part}$, $\mathrm{has\text{-}region}$, $\mathrm{material}$, $\mathrm{element}$ and $\mathrm{possesses}$. The following decision tree illustrates their usage.

Given an entity X and another entity Y such that Y is considered a structure of X, we say that (a) X $\mathrm{has\text{-}region}$ Y if Y is a region of space or a Spatial-Entity defined in relation to X. (b) X $\mathrm{material}$ Y if Y is an Entity and is pervasive in X. Y is usually a mass term in this case. (c) X is an $\mathrm{element}$ of Y if Y is a set of similar entities that Y is an instance of. (d) X $\mathrm{possesses}$ Y if Y is Energy, Bond or Gradient. (e) If none of the above applies, X $\mathrm{has\text{-}part}$ Y. Y must be a $\mathsf{Physical-Entity}$, and it should be a countable noun.

Let us now consider some biological examples that use these relations to describe structure. A functional group is a group of atoms that are associated with molecules (for example, a phosphate group). While modeling the structure of a molecule, we must decide which relation should be used to associate a functional group with the molecule? For example, when a functional group such as a

phosphate group is removed from a molecule, it assumes a new name — a phosphate ion. Because the functional group is only defined when attached to a molecule, considering it as a Region of that molecule is appropriate. Its constituent parts (atoms) are parts of the molecule and are located in the functional group region.

Let us consider an example of the material relation. A spider web is a structure composed of silk. Here we consider that the silk is in a material relationship to the spider web. Silk is a mass term and we cannot count the silk in the same way we would count the parts of an entity.

Next, we consider an example of the element relationship. An amino acid sequence is a series of amino acids in a particular order determined by genetic information. Here, we consider the relationship between an amino acid sequence and individual amino acids to be an element relationship. The key to this distinction is that the element relationship is not transitive.

Although our choice to use the possesses relationship with energy, bonds and gradients may at first seem arbitrary, the selection is very well thought out as we explain next. The textbook frequently uses language such as *matter possesses energy*. The matter could possess energy due to its location or structure. Thus, although, strictly speaking energy is not necessarily a structural feature of an entity, it has a strong relationship to its structure in some cases. The textbook introduces bonds as a higher level of organization of atoms, and thus, bonds are also a structural feature. In the description of various entities, bonds are considered as structures of entities (for example, *Although they may have some polar bonds associated with oxygen, lipids consist mostly of hydrocarbon regions*). But, because the bonds are neither physical entities nor spatial entities, the other relationships cannot be used, and a new relationship was needed. The textbook introduces the general concept of gradient as *A substance will diffuse from where it is more concentrated to where it is less concentrated; a substance will diffuse down its concentration gradient.* While describing the structure of entities such as membranes, the gradient features prominently. The gradient, by itself, is neither physical nor spatial. It has a magnitude and a direction. Thus, the rest of the structural relations are not applicable. Based on this analysis, we use the possesses relation to represent energy, bonds and gradients.

The has-part relationship is the most commonly used structural relationship. For example, a cell has a plasma membrane, chromosomes, ribosomes, etc. as its parts. In the decision tree for choosing a meronymic relationship, we consider has-part in the end, because biologists are not normally exposed to the other four relationships, and they tend to overuse the has-part relationship. By considering the other relationships first, we increase the likelihood of precise usage. The has-part relationship is transitive across levels of biological organization. For example, a cell has-part Plasma Membrane which in turn has-part a Lipid, and thus, by transitivity, a Cell has-part a Lipid.

Sometimes, in the textbook sentences, identifying whether an entity is a part of structure description of another entity is difficult. For example, consider the sentence from a textbook: *Recall that the plasma membrane is a phospholipid bilayer with associated membrane proteins.* From this sentence, one could also conclude a subclass-of relationship between a plasma membrane and a phospholipid bilayer, which is obviously, incorrect. Making these relationships explicit by the sort of decision tree introduced at the start of this section was instrumental in supporting biologists in making correct choices.

2.2.2 Relations to represent spatial arrangement

The structure description of biological entities is rich in spatial information and includes shapes, relative spatial locations, boundaries, portals, holes, etc.

The textbook introduces a vocabulary of more than 100 different shapes (for example, tube, helix, V-shape, etc.) We have developed the shape vocabulary necessary for the whole textbook. For the range of studied questions, we found that only referring to the shapes was necessary and that doing any deep reasoning with shapes was not required. Therefore, we will not further discuss shapes.

The key relationships needed for spatial arrangement are inside, outside, abuts and is-beween. The relations inside, outside, abuts can be reduced to well known topological relationships (Bennett, Chaudhri, & Dinesh, 2013). The choice for this small set was driven by the observation that although our ontology offered a larger set of relations, biologists most frequently and consistently used these four relations. Even with these simple relations, cases of confusion occurred, because the biological entities contain holes, and the natural language understanding of the meaning of these terms do not behave as expected. In many cases, one must use a combination of relationships to achieve the desired effect.

Let us now consider a few examples to illustrate some of the issues. A plasma membrane is-inside the extra-cellular fluid, and the cytoplasm is-inside the plasma-membrane. But, we cannot say that the plasma membrane is-outside the cytoplasm. That is because the spatial extent of the plasma membrane overlaps the cytoplasm. A much better natural language reading of is-outside is to view it as is-external-to.

Consider the spatial relationship between a peripheral protein to a plasma membrane. As discussed above, because the peripheral protein is part of the plasma membrane, saying that the peripheral protein is-outside the plasma membrane is inappropriate. The spatial extent of an entity includes the spatial extent of all its parts, and so the peripheral protein has to be is-inside the plasma membrane. If we wish to express that the protein is on the extracellular side, we add that the protein is-outside the cytoplasm. Thus, in some cases, we have to use a combination of spatial relationships to get an accurate representation of the spatial structure.

2.2.3 A sample conceptual model of bio-membrane structure

As an illustration of the application of the relationships for representing constituents and spatial arrangements of entities, we show a partial conceptual model of a bio membrane. In this model, the nodes of the graph represent the entities, and the labeled edges represent the relationships. The node labeled Bio-membrane is distinguished in the sense that this is a model of bio-membrane and represents a prototypical bio-membrane. We call this model prototypical because often exceptions exist to the most common case, and we only capture the most common case in this model. We further emphasize that this is only a partial model — a complete structural model of a bio-membrane is much more complex and too large for the purposes of this paper.

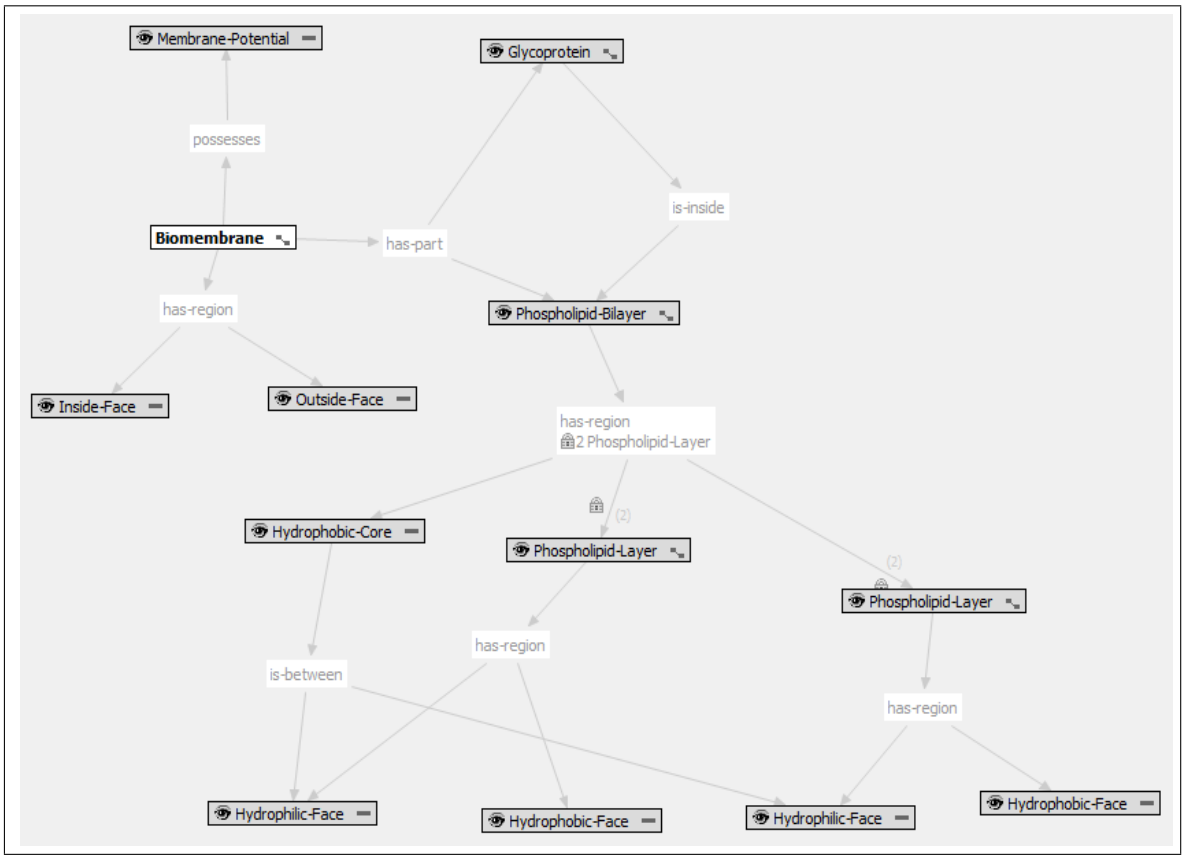

Figure 1. A Simplified Conceptual Model of Biomembrane Structure

2.2.4 Related work on representing structure

The relations to represent meronymic information and the spatial arrangement of entities have been known in the literature for a long time. An extensive vocabulary of such relations is available in the Cyc KB (Lenat, 1995). The distinguishing aspects of our work are in its simplicity and the set of guidelines and definitions that are appropriate for introductory biology and that can be applied by biologists, and understood by students. Our work on guidelines is inspired by the recent work on representing part-whole relationships (Keet & Artale, 2008).

Structural information has also been represented in numerous biological ontologies. For example, the Foundational Model of Anatomy (FMA) focuses on structural decomposition of entities (Rosse & Mejino Jr, 2003). FMA supports several meronymic relations such as has-part, has-region, has-member, and several spatial relationships such as continuous-with, attached-to, adjacent-to, surrounds, etc. The SNOMED-CT ontology is an ontology that contains a collection of clinical terms (Wang, Sable, & Spackman, 2002). It has only two relationships isa and has-part, and thus, a very limited representation of structure.

Obviously, the vocabulary considered here is inadequate to capture all the intricacies of biological structure. The most notable exceptions are representations for surfaces, cavities, and holes. Example sentences requiring the representation of these concepts are as follows. *Active site is a pocket or groove on the surface of the enzyme where catalysis occurs. In hydras, jellies, and other cnidarians, a central gastrovascular cavity functions in the distribution of substances throughout the body and in digestion.* We have taken first steps toward designing those representations which are described in more detail elsewhere (Bennett, Chaudhri, & Dinesh, 2013).

2.3 Modeling Function

In philosophy literature, the function of an entity is viewed as its reason for existence (Arp & Smith, 2008a). A direct application of this definition to biology can lead to unnecessary debate. For the purposes of our discussion, we adopt an operational definition: a process is a function of an entity if a biologist considers it to be a function of that entity within the scope of the knowledge in the textbook. Such a consensus must exist because for a student answering a question on an exam, there has to exist a correct answer to the question: "What is the function of X?"

We distinguish between two kinds functions: inherent functions and contextual functions. Inherent functions of entities are always true regardless of where that entity is found or which process it participates in. For example, storing chemicals is an inherent function of a Golgi apparatus. The contextual functions of an entity are realized only when that entity is part of some other entity or only in the context of a specific process. For example, smooth endoplasmic reticulum has the function of drug detoxification in a liver cell. But, associating drug detoxification as a universal function of smooth endoplasmic reticulum is incorrect. We refer to drug detoxification as a contextual function of smooth endoplasmic reticulum that can only be stated in the context of a liver cell.

Some processes have a natural one or two word biological name (for example, photosynthesis or cellular respiration). But, some processes have a much longer name composed of several words (for example, drug detoxification in a liver cell or aerobic cellular respiration in eukaryotes). From a conceptual modeling perspective, these longer names should be represented in more detail than by just creating a process with that name. For example, for aerobic cellular respiration in eukaryotes, indicating that it is an aerobic process and that it occurs in eukaryotes provides a more complete representation. Even for the processes with one word names, a detailed representation is useful for modeling in more detail how functions are realized, for stating contextual functions, and for explaining how various entities facilitate different steps of a function. Hence a vocabulary to describe process participants is essential for describing functions.

To specify the conceptual model for representing functions, we first give relations for describing process participants, then introduce our representation of functions, and then discuss the modeling of structure function relationships.

2.3.1 Representing process participants

Our relations for describing process participants are inspired by a comprehensive study of case roles in linguistics (Barker et al., 1997). These relations include agent, object, instrument, raw-material,

result, source, destination, and site. We developed both syntactic and semantic definitions for these relations which are available elsewhere (Chaudhri, Dinesh, & Inclezan, 2013). As an example, we consider the definition of raw-material. The semantic definition of raw-material is that it is any entity that is consumed as an input to a process. The syntactic definition of raw-material is that it is either the grammatical object of verbs such as*to use*, *to consume*, or it is preceded by: *using*.

As an example illustration of the use of these participant relations, consider the examples discussed earlier. For drug detoxification in a liver cell, we could represent a detoxification process which has an agent of smooth endoplasmic reticulum, an object of drug and a site of liver cell. Similarly, we could represent eukaryotic cellular respiration as a cellular respiration process that has base a eukaryotic cell.

2.3.2 Associating function with entities

We associate a function with an entity using the has-function relationship. For example, a biomembrane has-function movement of chemicals and blocking of chemicals. While associating a function with an entity using the has-function relationship is straightforward, identifying these functions from the language used in the textbook is not always easy. We illustrate this using a few examples.

Consider the sentence: *Channel proteins function by regulating movement of molecules across a membrane.* In this sentence, even though the word function is used, one must guess that a channel protein indeed has-function the process of regulating the movement of molecules across a membrane. In one possible interpretation of this sentence, it could have been understood as a description of the operation of channel proteins.

In the sentence: *Hydrophobic core impedes direct passage of ions*, the textbook does not use the word function, but here the reader must interpret to recognize that hydrophobic core has-function the process of impeding the direct passage of ions.

We considered the possibility in which has-function is defined as a super relation of all the participant relations. That approach, however, leads to many spurious inferences. For example, a heart is an agent of two different processes: pumping blood and making sound. If has-function is made a super relation of agent, we will conclude that making sound is also a function of heart, which is obviously, incorrect. Because of this, we needed to make has-function as an additional primitive relationship in the ontology that needs to be curated by biologists.

2.3.3 Representing structure function relationship

Biology educators are especially interested in students learning about how particular aspects of structure enable a certain function. (for example, learning that the aerodynamic shape of the wings of a bird enables the function of flying). In most situations, the structure function relationship is evident through the participant relationship between an entity and the process in which it functions. For example, a smooth endoplasmic reticulum has-function drug detoxification, and drug detoxification site smooth endoplasmic reticulum (ER). Thus, we know that the specific relationship between the smooth ER and its function as a site of the process of drug detoxification. However, many ex-

amples exist in which a substructure of an entity contributes to its function but the specific way in which it participates in that function is either unknown, or outside the textbook's scope. For example, chlorophyll A has-part poryphrin which contributes to its function of violet light absorption in an unknown way. We handle such under-specification by using a relation called facilitates. In the example considered here poryphrin facilitates violet light absorption by chlorophyll A.

Some forms of structure function relationship exist that we have not yet captured. One notable example is when a certain property of a structure enables a function (for example, the shape of a red blood cell is round which enables its function of movement through the blood veins). A more comprehensive analysis of such relationships is open for future work.

2.3.4 A sample conceptual model of bio-membrane function

As an illustration of the application of the representation introduced so far, we show in Figure 2 two of the functions of a bio-membrane. One of the functions of the bio-membrane is allowing the movement of chemicals to which it is otherwise impermeable. In Figure 2, this movement function has a path of Hydrophobic-Core which is a region of a phospholipid bilayer, which is one of its parts that we have shown earlier. A second function of a bio-membrane is blocking hydrophilic compounds. In Figure 2, this blocking action is done by the Hydrophobic-Core using a Fatty-Acid-Tail which is a region of Phospholipid which in turn is an element of the Phospholipid-Layer. This model brings together both the structure and function representation in one place.

2.3.5 Related work on functions

Structure, behavior and function (or SBF) is a modeling language to capture the structure and function of engineering artifacts. Our work is very similar in spirit to the SBF modeling in that it provides a vocabulary for defining structure and in connecting structure to function. In an SBF model, the functions are defined using a set of pre-conditions and post-conditions. In our approach, we do not explicitly model pre-conditions and post-conditions. The SBF model refers to the unfolding of the process as a behavior. Our process modeling primitives (for example, next-event, subevent, etc. enable the specification of process steps, but we do not have a special name for those primitives. In the SBF modeling language, if the function of a substructure is a step of the function of the whole (Goel, Rugaber, & Vattam, 2009), then a way to state that the substructure "owns" the function of the whole (Umeda & Tomiyama, 1997) is provided which is very similar to the idea of using the facilitates relation. Just as in SBF, the functions can be decomposed hierarchically.

In the literature on upper ontologies, a precedent exists for using has-function as a primitive relationship (Burek et al., 2006). The novelty of our representation is that it combines an ontological representation of functions (with has-function) with a linguistic representation (using relations like agent, object, etc.). This approach is crucial for our particular application. In various biomedical ontologies, relating entities to processes via a single relation such as has-participant which does not provide enough detail for the knowledge in the textbook. The linguistic relations offer us good coverage for a wide variety of events. Finally, to the best of our knowledge no prior work has distinguished between inherent and contextual functions, which is quite critical in biology.

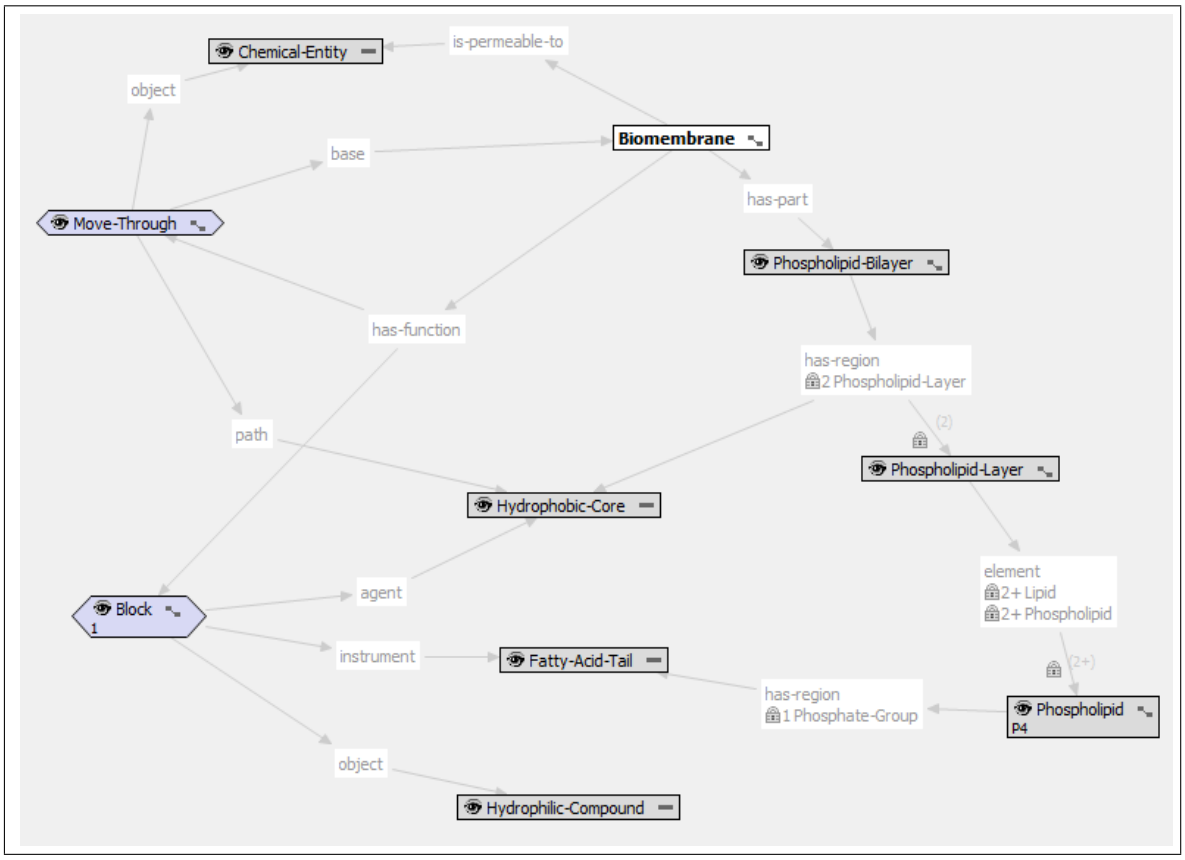

Figure 2. A Simplified Conceptual Model of Biomembrane Function

3. Answering Questions by Machine Reasoning

We begin this section by first describing our question development process, then describing different categories of questions handled by the system, and finally discussing implementation and example outputs.

3.1 Question Development

Because the goal of the current phase of work was to embed the resulting capability into an electronic textbook, we needed to identify a set of educationally useful questions. To determine the questions that will be useful and interesting to answer, we first convened a focus group of teachers and students who generated a list of questions that they considered educationally useful and of the kinds of the questions that students typically have as they study for the course. Next, we analyzed these questions to determine a broad set of categories. These categories include: relating structure to function, qualitative comparisons, similarity reasoning, and the effect of modifying the structure of an entity on its function. We specified each question category by using one or more question

templates which could then be instantiated in a variety of ways. We will now enumerate these five categories, some example questions in each category and the question templates.

Question category: Relating structure to function
Example question: What structures of chlorophyll-A enable it to absorb violate light?
Question template: What structure of <X> facilitates its function <Y>?

Question category: Qualitative comparisons
Example question: Which is easier to break down through mechanical digestion: saturated fatty acid or unsaturated fatty acid?
Question template: What is the relationship between property <P1> of concept <C1> to property <P2> of concept <C2>?

Question category: Detailed comparisons
Example question: What is the difference between the hydrogen bonds of water as ice, and the hydrogen bonds of water as a liquid?
Question template: What is the difference between <structure-1> that has <relationship-1> to <concept-1> and <structure-2> that has relationship <relationship-2> to <concept-2>

Question category: Effect of modifying a structure
Example question: If hydrogen is removed from a saturated fatty acid, then how is its function impacted?
Question template: If entity <A> is negatively impacted, what events <B> will be impacted?
If entity <A> no longer has relationship <R> with entity <B>, what events will be impacted?

Question category: Similarity reasoning
Example question: Leaf and photosynthesis have a certain relationship. This same relationship applies to a bio-membrane and what?
Question template: <A> is to <B> as <C> is to what?

Needless to say, we resort to significant simplification of the original questions while developing these templates. For example, the question *Which is easier to break down through mechanical digestion: saturated fatty acid or unsaturated fatty acid?* could be supported by instantiating the template as *What is the relationship between the digestibility of saturated fatty acid and the digestibility of unsaturated fatty acid?* Our end-user application addresses the issues of the usability of these templates by mapping input questions into these templates and automatically suggesting the instantiated versions. Such an approach goes a long ways towards keeping the focus on reasoning with the structured representation (Chaudhri et al., 2013a).

3.2 Approach to Reasoning

We have implemented reasoning methods for all the question templates considered in the previous section. The system supports some core reasoning tasks such as taxonomic queries, relation value

computation, and path computation, which are adapted to support all of the question templates (Chaudhri et al., 2013c). The taxonomic queries involve computing superclasses and subclasses of a class. The relation value queries compute the values of all applicable relations to an individual. The path computation queries usually begin from an individual, and recursively compute the relation values applicable to that individual. Such queries are computationally very expensive. Next, we describe how these reasoning operations are adapted to answer each of the five question types.

We answer the questions asking for structure function relationship by path reasoning. Thus, for the question template: What structure of <X> facilitates its function <Y>, we look for <Z> such that the path from <X> to <Z> only contains structural relationship, and there is a facilitates relationship between <Z> and <Y>.

To answer a question involving qualitative comparison, we look for those paths between <X> and <Y> that are labeled by qualitative relationships, where <X> represents the property <P1> of concept <C1> and <Y> represents the property <P2> of <C2>. The qualitative relationships are: positive-influence, negative-influence, directly-proportional, and inversely-proportional (Forbus, 1984).

To answer a comparison question, we first compute a description of the two entities, and then take a set difference between their descriptions. The description of an entity involves taxonomic information (ie, its superclasses and subclasses) and its relation values (ie, values of all relations applicable to it.) When the question asks for a specific kind of difference, for example, strutural differences, the system will only display the difference between the structural relations. The structual relations are precisely the relations considered in section 2.2.1.

To answer questions involving the effect of modifying structure, we first compute paths from the entity in question to various events in the KB with special preference given to those paths that contain the has-function relationship. The intuition behind this query is that if an entity is removed from the structure, its functions as well as other events it participates in will be affected. We refer to such reasoning as process interruption reasoning which we have described in a previous paper (Chaudhri, Heymans, & Yorke-Smith, 2012).

Finally, let us consider the similarity reasoning question template: "<A> is to <B> as <C> is to what?" This is a powerful question as it can be instantiated to ask for structural relationships (e.g., "Cell is to a cytoplasm as bio-membrane is to what?") and functional relationships (e.g., "Aquaporin is to stoma as osmosis is to what?"). It can also be instantiated within a single level of biological organization (e.g., "Nucleus is to a chromosome as cytoplasm is to what?") or across different levels (e.g., "Heart is to a human body as electrogenic pump is to what?") A huge space of such analogical reasoning questions is possible. Reasoning to answer questions of this form involves first computing a path connecting A and B, and then searching the whole KB for the same path between C and some D. Because arbitrary search can be computationally expensive, we prioritize the search process by using the following heuristics: (1) Look for a taxonomic path between A and B, and if found, look for a similar taxonomic path between C and some D; (2) look for a path between A and B in the concept definition of A, and then look for the same path in the concept definition for C; (3) look for a path between A and B in any single concept definition in the KB, and then look for the same path between C and some entity D in any concept definition in the KB; (4) search the whole KB for a path between A and B and then look for the same path between C and some D. Obviously,

such reasoning will return multiple answers, and the system will need to sort and rank them based on some criteria. More advanced forms of similarity reasoning could relax the requirement that the paths between A and B must be identical to the path between C and D.

3.3 Example Output of Reasoning

To test the reasoning templates, one requires a well-curated KB. We have created such a KB called KB_Bio_101 which encodes significant portions of an introductory biology textbook (Reece et al., 2011). The representation of structure and function in KB_Bio_101 is based on this paper. The KB, of course, contains representations broader than just structure and function, for example, representation of processes, roles, states, etc. The KB is available for research purposes at http://www.ai.sri.com/~halo/public/exported-kb/biokb.html. In Table 1, we show a few representative question and answer pairs for each question type considered in the previous section.

For the questions asking for structure function relationship, if the question does not specify a function (e.g., Q1), we first compute the function of the entity in the question, and then for each of its functions, determine which structure facilitates it. In the case of Aquaporin, one of its functions is Facilitated-Diffusion-of-Water, which is facilitated by Hydrophilic-Channel. If the question specifies a function (e.g., Q3), the system finds all the structures that facilitate that particular function which, in this case, includes Stroma and Protein-Enzyme.

The questions asking for qualitative relationships are answered by searching for a path between the two items mentioned in the questions such that the path contains the qualitative relationships. Some of the relationships can be direct (e.g., Q4), or could involve tracing through a series of relationships (e.g., Q6). The sentences shown in the answer are automatically synthesized by the natural language generation facility in the system (Banik, Kow, & Chaudhri, 2013).

For the detailed comparison questions, the system presents a well-organized table in which the specific differences are shown. For example, for an answer such as "A saturated fatty acid has single bonds and a linear structure while an unsaturated fatty acid has double bonds and a kinked structure", the system will display the values of the possesses relation to have respective values of singlebond and doublebond, and the respective values of the shape relation as linear and kinked.

For the question asking for the effect of modifying structure, the system will return the list of affected events (as in Q10 and Q11), and a confirmation if the event of interest will indeed be affected (as in Q12).

For the similarity reasoning questions, the system will return one similar entity (as in Q15) or multiple similar entities (as in Q13, Q14) that have an analogous relationship. In the actual presentation of the answer the details of the analogous relationship are shown. For example, for Q13, the system will display that Photosynthesis is the function of Chloroplast, and analogously, Cellular-Respiration is a function of Mitochondrion.

In the process of creating KB_Bio_101 we created test question sets for the first 11 chapters such that each question template had a few instantiations. The questions shown in have been drawn from that test suite. Our full test suite contained over 2000 questions from which 207 questions pertained to structure and function from which 153 questions were answered by the system to the user's satisfaction. Some of the failing questions were due to software engineering problems.

Table 1. Range of Questions Answered

Question type	QID	Question	Answer
Relate structure to function	Q1	What structure of Aquaporin facilitates its function?	Hydrophylic channel facilitates diffusion of water
	Q2	What structures of thylakoid facilitate light reaction?	Thylakoid membrane
	Q3	What structures of chloroplast facilitate photosynthesis?	Stroma, Protein enzyme
Qualitative relationships	Q4	What is the qualitative relationship between quantity of channel protein and the function of channel protein?	Quantity of the channel protein is directly proportional to rate of facilitated diffusion using channel protein.
	Q5	What is the qualitative relationship between quantity of mitochondria and function of mitochondrion?	Quantity of the mitochondrion is directly proportional to rate of energy transformation by organism.
	Q6	What is the qualitative relationship between Hydrogen Ion gradient and free energy?	Quantity of the hydrogen ion gradient is directly proportional to quantity of the proton-motive force which is directly proportional to free-energy.
Detailed Comparisons	Q7	What is the structural difference between a monomer and a diamer?	A diamer has exactly two monomers
	Q8	What are the functional differences between protein phosphatase and protein kinase?	The function of protein phosphatase is dephosphorylation and the function of protein kinase is phosphorylation
	Q9	What are the structural differences between a saturated fatty acid and an unsaturated fatty acid?	A saturated fatty acid has single bonds and a linear structure while an unsaturated fatty acid has double bonds and a kinked structure
Effect of modifying structure	Q10	If smooth endoplasmic reticulum is removed from plant cell, what events will be affected?	Synthesis of lipids (as it is the function of smooth ER)
	Q11	If ATP synthase is removed from thylakoid membrane, what events will be affected?	Synthesis of ATP, Facilitated diffusion, generation of hydrogen ion graident, holding together phospholipids
	Q12	If lysosome is removed from eukaryotic cell, will autophagy be affected?	Yes because autophagy is the function of lysosome
Similarity reasoning	Q13	chloroplast is to photosynthesis as mitochondrion is to what	Cellular Respiration, Chemiosmosis
	Q14	Calvin cycle is to CO2 as citric acid cycle is to what?	NAD Plus, oxaloacytate, acetyl COA
	Q15	ATP Synthase is to Chemiosmosis as Electron transport chain is to what?	Electron transport chain reaction

4. Summary and Conclusions

In this paper, we have shown how we can define a conceptual representation to model biological structure and function. Although much of the vocabulary we used has already been known, our contributions include developing guidelines that can be used by biologists and applying those relations to a new domain. Our vocabulary has a few novelties: the use of the possesses and facilitates relationship, combining a linguistic representation of processes with functions, and the distinction between inherent and contextual functions. Through a user study, we determined the range of questions that would be educationally interesting and used that data to develop a set of question templates. We illustrated how these questions templates could be implemented and presented the results of reasoning performed by the system. None of the previous work on representing structure and function has considered the range of questions that we have considered in our work.

The work presented is only a starting point and can be expanded in a variety of ways. First, many aspects of structure and function exist that we do not yet cover and that require design of new vocabulary. The most notable omissions from the current vocabulary are ways to represent boundaries, cavities and surfaces (Bennett, Chaudhri, & Dinesh, 2013). Second, a similar level of analysis could be applied to other forms of biological knowledge such as process regulation, energy transfer, continuity and change, evolution, etc. Finally, a variety of pedagogical tools could be built that leverage these representations. For example, machine reasoning could use these representations in a student assessment program that evaluates a student by engaging in a dialog, thus, checking for depth of knowledge. An online tutor could use such representations as a precise model of the relationships that a student must master.

To conclude, we have put forth a concrete proposal to incorporate rigor in describing biological knowledge in an introductory textbook. Every scientific discipline must go through such a process in its evolution toward maturity. We hope that when pursued on a large scale such rigor will fundamentally alter not only biology education, but the way that we understand and practice biology.

Acknowledgements

This work has been funded by Vulcan Inc. and SRI International. We thank Bryan Wiltgen and Debbie Frazier for their contribution to this work.

References

Arp, R., & Smith, B. (2008a). Function, Role, and Disposition in Basic Formal Ontology. *Nature Preceedings*, 1–4.

Arp, R., & Smith, B. (2008b). Ontologies of Cellular Networks. *Science Signaling*, *1*, mr2.

Banik, E., Kow, E., & Chaudhri, V. K. (2013). User-Controlled, Robust Natural Language Generation from an Evolving Knowledge Base. *ENLG 2013 : 14th European Workshop on Natural Language Generation*.

Barker, K., Copeck, T., Delisle, S., & Szpakowicz, S. (1997). Systematic Construction of a Versatile Case System. *Journal of Natural Language Engineering*, *3*, 279–315.

Barker, K., Porter, B., & Clark, P. (2001). A Library of Generic Concepts for Composing Knowledge Bases. *First International Conference on Knowledge Capture*.

Bennett, B., Chaudhri, V. K., & Dinesh, N. (2013). A Vocabulary of Topological and Containment Relations for a Practical Biology Ontology. *Conference on Spatial Information Theory*.

Brachman, R. J., & Levesque, H. J. (2004). *Knowledge Representation and Reasoning*. Morgan Kaufmann.

Brewer, C., & Smith, D. (2009). *Vision and Change in Undergraduate Biology Education: A Call to Action* (Technical Report). Final Report of a national conference organized by AAAS.

Burek, P., Hoehndorf, R., Loebe, F., Visagie, J., Herre, H., & Kelso, J. (2006). A Top-Level Ontology of Functions and its Application in the Open Biomedical Ontologies. *Bioinformatics*, *22*, e66–e73.

Casati, R., & Varzi, A. C. (1999). *Parts and Places: the Structures of Spatial Representations*. Bradford Books.

Chaudhri, V. K., Cheng, B., Overholtzer, A., Roschelle, J., Spaulding, A., Clark, P., Greaves, M., & Gunning, D. (2013a). Inquire Biology: A Textbook that Answers Questions. *AI Magazine*, *34*.

Chaudhri, V. K., Dinesh, N., & Inclezan, D. (2013). Three Lessons in Creating a Knowledge Base to Enable Explanation, Reasoning and Dialog. *Second Annual Conference on Advances in Cognitive Systems*.

Chaudhri, V. K., Heymans, S., Wessel, M., & Tran, S. C. (2013b). Object-Oriented Knowledge Bases in Logic Programming. *Technical Communication of International Conference in Logic Programming*.

Chaudhri, V. K., Heymans, S., Wessel, M., & Tran, S. C. (2013c). Query Answering in Object Oriented Knowledge Bases in Logic Programming. *Workshop on ASP and Other Computing Paradigms*.

Chaudhri, V. K., Heymans, S., & Yorke-Smith, N. (2012). Process Interruption Reasoning. *Proceedings of the 2nd Deep Knowledge Representation and Reasoning Challenge Workshop*. Playa Vista, CA.

Forbus, K. D. (1984). Qualitative process theory. *Artificial intelligence*, *24*, 85–168.

Goel, A. K., Rugaber, S., & Vattam, S. (2009). Structure, Behavior, and Function of Complex Systems: the Structure, Behavior, and Function Modeling Language. *AI EDAM*, *23*, 23–35.

Karp, P. D. (2001). Pathway Databases: a Case Study in Computational Symbolic Theories. *Science*, *293*, 2040–2044.

Keet, C. M., & Artale, A. (2008). Representing and Reasoning over a Taxonomy of Part–Whole Relations. *Applied Ontology*, *3*, 91–110.

Lenat, D. B. (1995). CYC: A Large Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, *38*, 33–38.

Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., & Jackson, R. B. (2011). *Campbell Biology*. Boston: Benjamin Cummings imprint of Pearson.

Renear, A. H., & Palmer, C. L. (2009). Strategic Reading, Ontologies, and the Future of Scientific Publishing. *Science*, *325*, 828–832.

Rosse, C., & Mejino Jr, J. L. (2003). A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy. *Journal of biomedical informatics*, *36*, 478–500.

Spear, A. D. (2006). Ontology for the Twenty First Century: An Introduction with Recommendations. `http://www.ifomis.org/bfo/documents/manual.pdf`.

Umeda, Y., & Tomiyama, T. (1997). Functional reasoning in design. *IEEE expert*, *12*, 42–48.

Wang, A. Y., Sable, J. H., & Spackman, K. A. (2002). The SNOMED clinical terms development process: refinement and analysis of content. *Proc AMIA Symp*, 845–9.

Wing, J. M. (2006). Computational Thinking. *Communications of the ACM*, *49*, 33–35.