Ontological Foundations for Biology Knowledge Models

Carole D. Hafner

Natalya Fridman

College of Computer Science
Northeastern University
Boston MA 02115
Tel: (617) 373-2462 FAX: (617) 373-5121
{hafner, natasha}@ccs.neu.edu

Keywords: knowledge representation, ontology, frames, qualitative reasoning, information retrieval

Abstract

This paper analyzes the ontological requirements for representing biology knowledge, and identifies several areas where current knowledge representation (KR) paradigms need to be extended. We focus on the representation of experimental materials and methods, and the reasoning task of intelligent information retrieval; however, the ontological issues we raise apply to biology (and experimental sciences) in general. We have identified two important concept types in molecular biology that cause problems for standard knowledge models: 1) complex substances such as mixtures and nucleic acid sequences; 2) transformations (such as biochemical reactions) that convert one substance into another. We describe these problems, propose solutions for some of them, and give examples of the need for such knowledge representations in intelligent information retrieval.

1. Introduction

Current research aimed at the development of knowledge sharing technology [Gruber 1993, Lehman 1995] is based on the following observations:

> a. Creation of more robust intelligent systems will require domain-specific knowledge models (microtheories) to be embedded in a large substrate of "consensus" world knowledge.
>
> b. The development of a large consensus knowledge base will require research groups to share results, so that microtheories of molecular biology, for example, will build on consensus microtheories of tangible substances, events, qualities, measurements, time, space, etc.
>
> c. Knowledge sharing requires common ontological foundations to be agreed upon. In the absence of such agreements, microtheories produced by different research groups cannot be integrated.

Ontology in AI means the fundamental categories and relations that provide a framework for knowledge models. An ontology of time, for example, must include a set of formal conventions for defining points, intervals, and durations of time, and associating times with events. An ontology for molecular biology must include formal conventions for defining DNA sequences, operons, genes, and the relationships between these objects and biological processes such as protein synthesis. Although there is disagreement about the precise meaning of the term "ontology", most AI researchers agree that the ontological foundation for a knowledge model is the set of high-level categories and relations used to construct the model's more specific entities.

Most of the larger computerized resources in molecular biology, such as Genbank of the National Center for Biotechnology Information or Protein Data Bank [Bernstein 1977] use databases rather than knowledge bases, and do not define an ontological framework other than the database schema. Each object has dozens of slots that describe where it comes from, what it consists of, its properties, etc. However, there is no attempt to define general taxonomic, partonomic or role relationships among the concepts.

Other ontologies for biology knowledge are based on extensive taxonomics of concepts. Unified Medical Language System (UMLS) of the National Library of Medicine [Humphreys 1993] is an example of such a system. It has an IS-A hierarchy of more than 100 medical concepts (as of 1994) and a Semantic Network that represents relationships between categories. These include such semantic relations as *physically_related_to, spatially_related_to, functionally_related_to,*

temporally_related_to, *conceptually_related_to* and various more specific forms of these relations. The Encyclopedia of *E. coli* Genes and Metabolism (EcoCyc) [Karp 1996] employs a frame knowledge representation system. Frames are organized in a class hierarchy for various types of enzymatic reactions, metabolic pathways and chemicals used in these reactions. This system is used for retrieval and visualization.

A third example of ontologies for biology knowledge is the AI-type knowledge systems that attempt to simulate human-like reasoning (e.g. PEPTIDE by D.Weld [Weld 1986] or GENSIM by P.Karp [Karp 1993]). These systems use qualitative reasoning to predict what will happen in a biochemical system. Although they have the most "intelligence" in them, systems of this type take a "focused" approach to ontology design that includes only those concepts required to describe a narrow class of problems.

The creators of biology knowledge models have generally not considered how biology concepts would be incorporated into a general ontological framework. For example, Figures 1 and 2 show the top level concept hierarchies from [Karp 1993] and [Karp 1996]. It is not easy to see how two focused ontologies in the same domain (and by the same researcher) could be integrated with each other, let alone integrated into a general framework such as proposed by ontology researchers [Bateman 1994, Lenat 1990, Sowa 1995]. It would be very desirable to integrate ontologies for representing biology knowledge with ontologies created for other domains. Researchers in biology could then take advantage of ontologies of time and space, for example: when we say that one bacterial strain was grown for 3 hours and another one for 90 minutes, it may also be relevant that the second strain was grown half as long as the first one. When we define an experiment as a set of actions or steps, it may be important to know what it means that step A happened *during* step B. Temporal knowledge, and the conclusions that flow from it, would be defined in a general ontology and would therefore be accessible to an integrated biology/world knowledge model.

In our earlier work, aimed at developing techniques for intelligent retrieval of biology research papers [Baclawski 1993], we created an ontology for experimental materials and methods [Hafner 1994]. The top-level hierarchy (Figure 3) shows how a biology microtheory could be embedded in a more general knowledge framework – a necessity to support even limited natural language understanding. We followed what has become a standard paradigm for ontology design: frames organized into a concept taxonomy with structured inheritance; as might be expected, the most elaborated concept sub-networks in our model were substances and experimental processes. Figure 4 shows the top-level frame definition for experimental processes. Specific processes, such as insertion, have additional slots representing participants, such as inserted-object and target.

In trying to represent actual experiments reported in the literature, we discovered important areas of mismatch, where the ontological conventions used in other AI applications appeared inadequate for our purposes. For example, the fundamental notion that every "real" object is defined as an instance of a category seems incompatible with a universe where objects can change their category. In addition, the traditional representation of tangible objects as either objects whose parts are an unordered set or totally unstructured "stuff", does not fit well with the complex mixtures and topological structures described in biochemistry experiments. We argue that molecular biology, and experimental science in general, impose requirements that should be, but have not yet been, taken into account in the effort to create a common set of ontological conventions.

2. Complex substances

Most general ontological models include a high-level division between discrete objects (those with distinct parts or components) and quantities of "stuff". Discrete objects such as cars have discrete components such as wheels and engines, organized into a "parts" hierarchy (sometimes called the "partonomy"). Stuff (such as water, sand, air, and so forth) does not have parts and every sub-chunk of stuff is still the same stuff. The basic inference rules (axioms) that distinguish these categories are:

> Axiom 1. If X is a discrete object such as a car, then there exist identifiable subparts of X such as wheels, an engine and so forth. The parts of a car are not themselves cars.
>
> Axiom 2. If Y is a chunk of stuff such as water, then every sub-chunk of Y is also a chunk of water.

(Lenat and Guha in [Lenat 1990] note that Axiom 2 only applies down to a minimal granularity.)

In creating a representation for biology knowledge, there are many complex substances that do not easily fit within this taxonomy. We discuss two of them below: 1) populations and mixtures, and 2) parts and sequences.

2.1. Populations and Mixtures

Practically all biology experiments deal with molecules or cells: bacteria, protein, DNA. When any of these is part of an experiment, it is usually not a single object, but a collection of such cells or molecules (e.g. *E. coli* strain

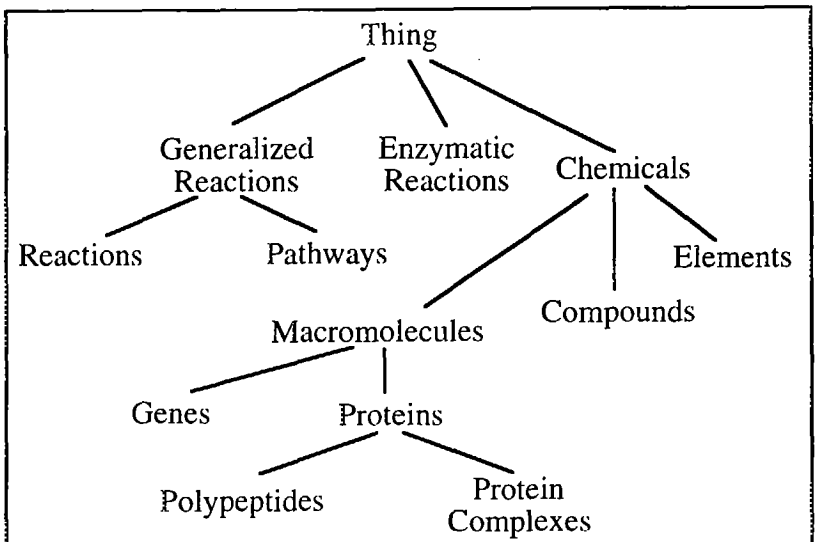

Figure 1. Top-level Ontology for EcoCyc [Karp 1996]

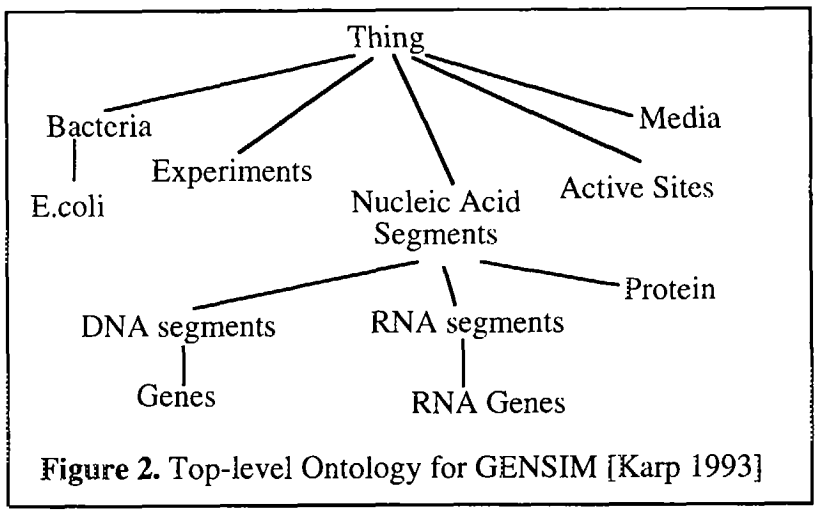

Figure 2. Top-level Ontology for GENSIM [Karp 1993]

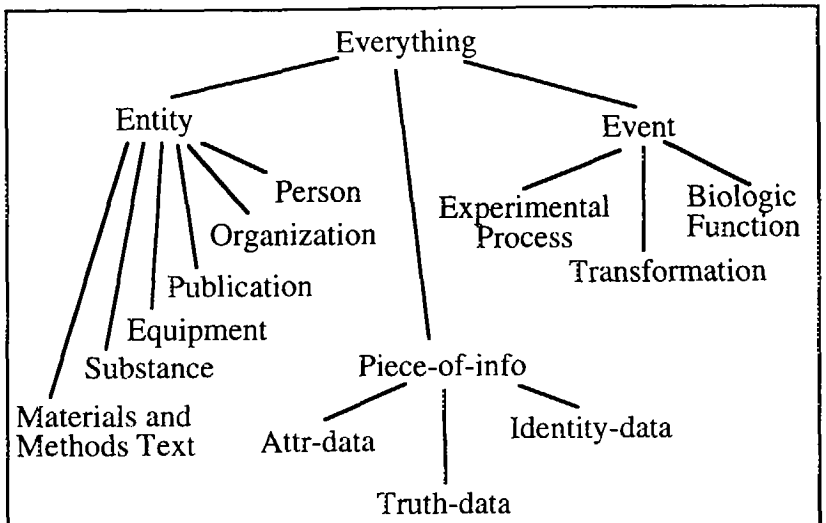

Figure 3. Top Level Ontology from [Hafner 1994]

EXPERIMENTAL-PROCESS frame

Superclass:	**EVENT**
Species:	`<string>`
Result:	`<substance> or <transformation>`
Parent:	`<process>`
End-test:	`<property-expression>`
Substeps:	`<process>...`
Equipment:	`<string>` or `<equipment-use>`
Manner:	`<property-expression>...`

Figure 4. Experimental process frame from [Hafner 1994]

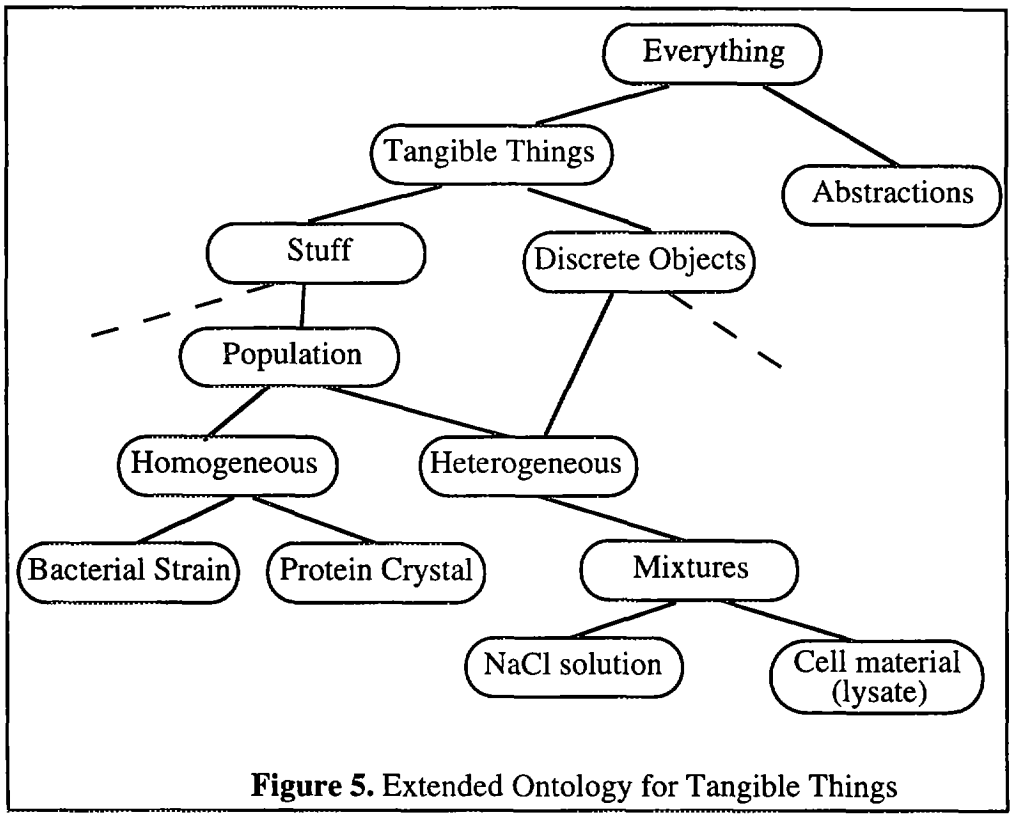

Figure 5. Extended Ontology for Tangible Things

in the example in Section 4) called a *population*. A population can be homogeneous or heterogeneous (See Figure 5). A crystal of a particular protein is an example of a homogeneous population, and a mixture such as a buffer is a heterogeneous population.

In the text of papers and in the representation, when it is stated that process A was applied to substance B, what is actually meant (in the cases where B is a population) is that process A is applied to B and has an effect on the *members* of the population (i.e. on the cells or molecules of the population). If B is homogeneous, than all the members of B are affected. If B is a heterogeneous population, process A may only affect some of the members of the population. For example if we say that "cells were broken by a press", the following is the exact interpretation of this: the press was applied to the population of cells and each cell in the population was broken. Alternatively, if a salt solution evaporates, the evaporation process only involves the water, not the salt.

Mixtures, which are extremely common in biology experiments, can be viewed as a sub-type of stuff since they satisfy Axiom 2 above, or as a subtype of discrete object, since mixtures also satisfy Axiom 1 above (i.e., they have identifiable components). The following features distinguish mixtures from heterogeneous populations in general:

1. components of mixtures are meaningful entities by themselves; they exist not only in and for this mixture. This is different from, say, parts of an engine that are normally found only within an engine.
2. in biology experiments, mixtures are often created by the experimenter for a purpose (to effect a transformation of one or more ingredients, or as environment for a certain cellular or chemical component).

As a first approximation, we can represent ingredients of a mixture using a role "ingredients" similar to the "parts" role in structured objects. Depending on how the substance is being used, there can be named component-roles (such as *medium* in the mixture of cells and a growth medium) that indicate what function the particular component is performing. Ingredients of a mixture cannot be heterogeneous, since they would mingle with other elements of the mixture. A problem arises if we want to be able to represent the concentration of ingredients within a mixture. The straightforward "ingredients role" approach described above will not suffice, and standard alternatives are not appealing. Another approach is to define a relational frame "substance-in-concentration" with two slots (the substance and the concentration), and fill the "ingredients" slot with instances of this kind.

Further discussion of mixtures appears in Section 3, where we explore the transformations that combine ingredients into new mixtures, and extract ingredients from the mixtures and that result in a category change for the ingredients or the mixtures.

2.2 Parts and Sequences

Standard ontologies of objects and substances are founded on two hierarchical structures: taxonomic ("isa") structures and partonomic ("has-part") structures. However, in building the knowledge model of [Hafner 1994], problems arose in applying the "has-part" relation to the complex objects found in molecular biology, such as DNA and proteins. The standard partonomic inference rule is:

```
If X has-part Y, then every object of
type X contains an object of type Y.
```

This rule needs to be supplemented with a variation describing the component relationships in mixtures, as follows:

```
If X has-ingredient Y, then every
population X contains a population Y.
```

Next, we observe that characterizing proteins as having parts or ingredients which are amino acids is correct but insufficient. Clearly, it is necessary to include in our ontological foundation another partonomic relationship, "made of", which would permit the following inferences:

```
If X is-made-of Y, then every X
contains a population Y.
```

```
If X is-made-of-Y, then for every X
there is a population Y which is co-
extensional* with X
```

Finally, an ontology for biochemistry knowledge must be able to represent the fact that proteins (and other chemicals) are not merely collections or mixtures of components, but are formed in particular structures -- in the case of proteins, a chain or sequence of amino acids. We can consider creating additional partonomic relations, "is-structure-of" and "is-sequence-of"; an example of the kind of reasoning these ontological structures would support is:

```
If X is-sequence-of Y, then for each
homogeneous subclass Xj of X there
exists a unique sequence
S = [Y1 . . . Yn] such that:
if S contains Yi k times, then:
    1.every molecule of Xj contains k
    instances of Yi and
    2.every population of Xj is k/n
    percent made of Yi
```

Although the structures proposed here would undoubtedly be useful in representing biology knowledge, it is unclear how to integrate them with other knowledge models, including those of biological processes such as DNA transcription. In creating a consensus ontological base, the AI community should develop a general mechanism for characterizing the relationships between objects and their physical parts that can encompass the variety of structures found in molecular biology.

3. Processes and Transformations

Current ontological models make a primary distinction between tangible objects, processes, and abstractions such as numbers. The structure of processes in most AI systems follows a well known pattern: a list of participant objects (the "parameters"), the preconditions for the process to occur, and the effects or changes engendered. This approach to representing processes has been used for planning [Barr 1981], natural language understanding [Cohen 1979], and simulation [Forbus 1984] with good results.

A model of experimental biology must include a large number of different processes – natural processes such as growth, DNA transcription, and chemical reactions; and experimental processes such as mixing, removing and inserting. For guidance, we compiled a list of verbs in research articles, for example: add, assay, centrifuge, combine, dialyze, disrupt, elute, extract, harvest, incubate, inoculate, label, measure, precipitate, purify, rinse, suspend, tether, wash. Our goal was to define a useful taxonomy of experimental processes, and create a model of their preconditions and effects. This in turn led to a re-examination of the ontological foundations of our knowledge representation.

3.1 Transformations and Identity

In examining the verbs above, it is clear that many experimental processes are characterized in terms of their end result, which is a transformation of the substances involved in the experiment. But most AI systems model the effects of a process on an object only as changes in its property values or relations with other objects. For example, if a heating process is applied to an object, its temperature rises. If a block in a blocks-world model (or an object in a manufacturing plant) is moved, a change occurs in its location property and its support relations with other objects.

Processes in molecular biology, unlike the above, often involve fundamental changes in the structure (and even the

* Physically the same although conceptually different - see the stuff-of function in Sec. 3.3.

category identity) of the substances in the model. To consider a blocks world analogy: suppose enough heat is applied to melt the plastic blocks in a blocks world model into liquid. There are no longer any blocks in existence! However, it still may be important to represent that the goo which now exists used to be a particular block or set of blocks. In molecular biology one strain of bacteria is turned into a different strain by the introduction of a plasmid; bacterial cells are turned into lysate by pressing or sonication; a mixture is turned into a pellet and a supernatant by centrifugation, etc.

We define a *transformation* as a process in which at least one participant changes its category identity during the process. From the standpoint of a computer database, we may say that the transformed participant ceases to exist, but this is not an accurate reflection of the way people think about the situation. Thus, a straightforward process model that simply represents inputs (participants) and outputs (objects that come into existence) is inadequate for modeling transformations, because a) it does not represent the fact that the inputs no longer exist and b) it does not represent the relationship between the outputs and the original inputs, one of the most important relationships being the fact that the stuff the inputs were made from is now the stuff the outputs are made from.

For example, Figure 6 shows the definition of "boiling" in qualitative process theory [Forbus 1984], a transformation of some liquid to some gas. The input to the process is contained-liquid *w* (an individual view). The object that comes into existence is *g*, a population which is a gas. The process description says that *w* and *g* are the same substance (meaning they are both made of the same liquid, since they cannot be co-extensional) and the amount of *g* increases, while the amount of *w* decreases at the same rate. It does not express the fact that the "stuff" of *g* is derived from the "stuff" of *w*. Therefore, the history of a boiling process would not make this connection. (It is also interesting to consider the situation where *w* is a mixture such as salt solution. In that case, the assertion that substance(*g*) = substance(*w*) is incorrect, and it is not clear how the correct relationship would be expressed in this framework.)

Note that, according to our definition of transformations, different underlying representation of objects may lead to different classification of processes. For example, suppose a model of physical substances has a slot called "SLG" with values solid, liquid, and gas. In that case, there would not be two different substances *w* and *g*, there would only be one substance whose SLG property gradually changed. However, this representation gives rise to problems also -- a model of the boiling process would require a representation for populations of objects where the proportion of individuals with a particular value for a property would be the "quantity" subject to a qualitative influence. (This is probably a better representation of

Process Boiling
Individuals:
w is a contained-liquid
hf is a heat flow process instance where
dst(hf) = w
QuantityConditions:
Status(hf, Active)
A(temperature(w)) < A(t-boil(w))
Relations:
There is g ∈ piece-of-stuff
gas(g)
substance(g) = substance(w)
temperature(w) = temperature(g)
Let generation-rate be a quantity
A(generation-rate) > ZERO
generation-rate Q+ flow-rate(hf)
Influences:
I - heat(w), A(flow-rate(hf)
I - amount-of(w), A(generation-rate)
I + amount-of(g), A(generation-rate)
I - heat(w), A(generation-rate)
I + heat(g), A(generation-rate)

Figure 6. Boiling Process Description from [Forbus 1994]

boiling, but not all processes are amenable to this treatment.)

3.2 Types of Transformations

We have made a preliminary classification of the transformations occurring in experimental biology:

a. **Topological transformations** involve inserting objects into other objects, creating new mixtures by combining substances, and separating components (including ingredients) from objects (for example, by centrifugation or precipitation) but without chemical reactions. Complex topological transformations occur, for example when two substances are mutually transformed by the transfer of a component from one substance to the other, as in dialysis; or in the case of "rinsing", when one substance is removed from another by adding a rinse agent which combines with the ingredient to be removed, and then separating the (now dirty) rinse agent from the mixture. The rightmost part of Figure 7 shows graphically the effect of another complex topological transformation, where cells in a buffer are broken, creating a new mixture containing the buffer and the ingredients of the cells. In topological transformations the low-level ingredients are preserved, but their arrangement in the objects changes, which may lead to the changes in the identity of the higher-level objects.

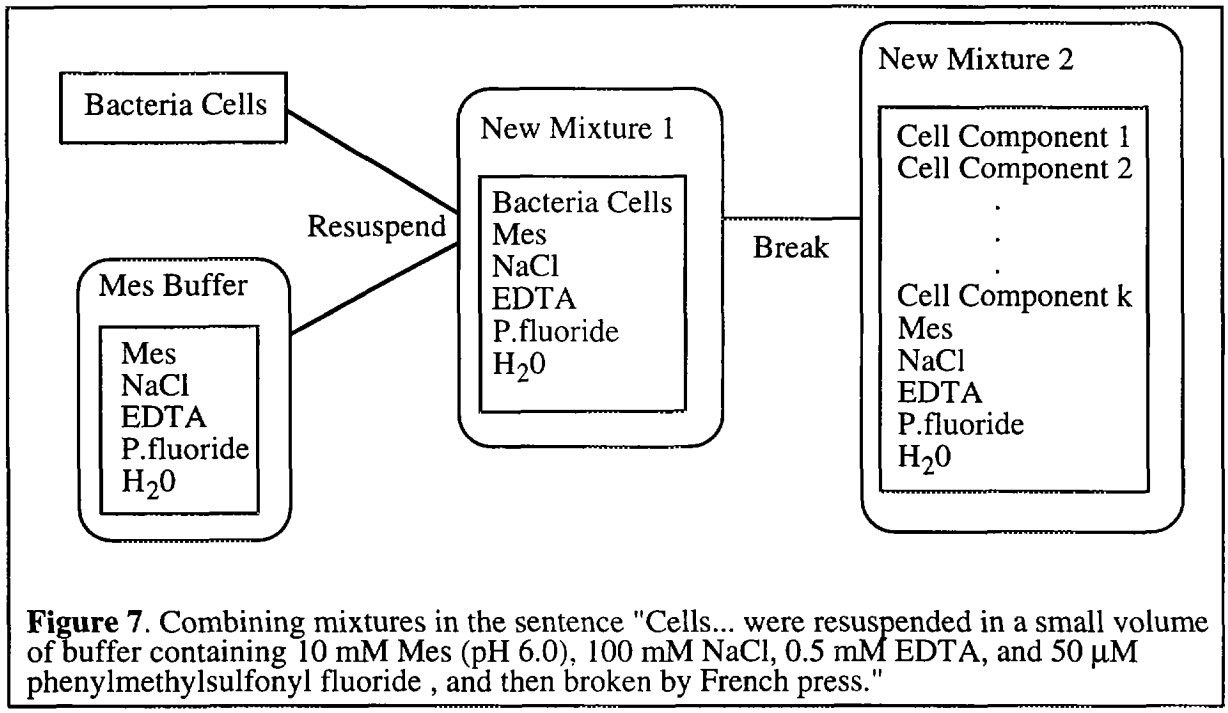

Figure 7. Combining mixtures in the sentence "Cells... were resuspended in a small volume of buffer containing 10 mM Mes (pH 6.0), 100 mM NaCl, 0.5 mM EDTA, and 50 μM phenylmethylsulfonyl fluoride , and then broken by French press."

b. **Reaction transformations** involve substances that combine at the molecular level to produce different substances (normally through covalent bonds). The classic model of a chemical reaction occurs when two chemicals are mixed, and they react to form a new chemical (or two different chemicals). In biology, many reactions require enzymes to be present; in that case, three chemicals are mixed, and two of them are transformed to form one or two new chemicals.

3.3 Formalizing Transformations

In molecular biology, as in other experimental sciences, participants in transformations are normally populations (e.g. population of cells, molecules, etc.). Since populations are chunks of stuff, we can assume a dual representation: for every population object P, there is another object stuff-of (P) representing all the stuff that comprises the population. Let us model a transformation of one object to another by creating two object descriptions, and linking some of the stuff in one object to some of the stuff in the other.

Transformations can be instantaneous (at a certain time granularity) or gradual. To model an instantaneous transformation of *A* into *B*, we can define a predicate Tr(stuff-of(A), stuff-of(B)). To model the gradual transformation of one substance to another, let us also define the concept of a "transfer path" analogous to a heat path in qualitative physics, with a Quantity transfer-rate. There is a predicate Tr-Connects (p, w, g) which states that population *w* is being transferred into population *g* via a transfer path *p*. This means that the stuff-of (*w*) gradually becomes stuff-of (*g*) In the boiling process description, we can replace the statement that *g* and *w* are made of the same substance by a statement that the stuff of *g* is coming from the stuff of *w*, as shown below.

Process Boiling
Individuals:
w is a contained-liquid
hf is a heat flow process instance where dst(hf) = w
QuantityConditions:
Status(hf, Active)
A(temperature(w)) < A(t-boil(w))
Relations:
There is g ∈ population
gas(g)
temperature(w) = temperature(g)
There is p ∈ Transfer-Path
Tr-Connects(p, w, g)
transfer-rate(p) Q+ flow-rate(hf)
Influences:
I - heat(w), A(flow-rate(hf)
I - amount-of(w), A(transfer-rate(p))
I + amount-of(g), A(transfer-rate(p))
I - heat(w), A(transfer-rate(p))
I + heat(g), A(transfer-rate(p))

4. Knowledge-based Information Retrieval

In this section we describe how the ontological structures discussed above exhibit in the biology literature and how

In this section we describe how the ontological structures discussed above exhibit themselves in the biology literature and how they can be used for intelligent information retrieval. The text in Figure 8 is an excerpt from a molecular biology paper [Gegner 1991]. This **Growth of Cells and Protein Purification.** The *cheW* and *cheA* plasmids were expressed in *E. coli* mutant strain RP3098 (a Δ*flhA-flhD* mutant), which was provided by J.S. Parkinson (University of Utah). Cells were grown at 30°C in L broth....

CheW purification is based on the procedure described by Stock *et al.* (14) with the following modifications. **Cells** were harvested by centrifugation at 5000rpm (Beckman JA 10 rotor) for 5 min, resuspended in a small volume of **buffer** containing 10 mM Mes (pH 6.0), 100 mM NaCl, 0.5 mM EDTA, and 50 μM phenylmethylsulfonyl fluoride, and then broken by French press. The **lysate** was ultracentrifuged at 50,000 rpm (Beckman Ti 60 rotor) for 1 hr to remove **cellular debris. Protein** was precipitated from the **supernatant** by adding $(NH_4)_2SO_4$ to 40% saturation and pelleted by centrifugation. The **pellet** was resuspended, dialyzed against the Mes **buffer** and loaded onto a Whitman DE-52 column.... **CheW** was >99% pure as determined by Coomassie Blue staining.

Figure 8: An excerpt from a biology research paper [Gegner 1991] that describes the process of protein purification

they can be used for intelligent information retrieval. The text in Figure 8 is an excerpt from a molecular biology paper [Gegner 1991]. This excerpt describes a process of purifying CheW protein from a certain strain of *E. coli* bacteria.

The sequence starts out with the strain of *E. coli* bacteria which is grown to get the necessary amount of cells. The grown cells contain CheW protein which now needs to be purified. The purification process consists of first breaking the cells and then achieving higher and higher concentration of CheW in the mixture that remains.

Along the way, various substances (buffers, chemicals) are added to the mixture and then removed, carrying some of the unwanted stuff away with them. In the end, what is left is a mixture 99% of which is CheW protein. The following Sections illustrate several information retrieval problems using this paragraph.

4.1 Mixtures

In Section 2.2 we talked about representing mixtures in biology experiments. In this paragraph we have:

> "Cells... were resuspended in a small volume of **buffer** containing 10 mM Mes (pH 6.0), 100 mM NaCl, 0.5 mM EDTA, and 50 μM phenylmethylsulfonyl fluoride"

There are two mixtures described here. First is the mixture of the cells and the buffer, which is the top level output of the combining process described in the sentence. "Buffer" is the role of the second ingredient. The substance which functions as a buffer is also a mixture of four ingredients at certain concentrations, plus water. In the buffer, ingredients don't have named roles. Figure 7 shows a representation of the process described. The basic inference rules for combining mixtures is:

> Rule 1. If X and Y are mixtures combined into Z, then Z is also a mixture, and the ingredients of Z
>
> are the union of the ingredients of X and Y.

Consider the following query:

> "What were the chemicals used in resuspension of the cells?"

The user specifies "chemicals" as the class of the substance or substances used in resuspension. There is no mention of buffer in the query. Buffer itself is not a chemical but a mixture and the paragraph in Figure 8 would not be brought up as an answer without additional reasoning. We know, however, that cells were mixed with the buffer in the resuspension process and the buffer in turn had particular chemicals as its ingredients. We can then infer what chemicals were used in resuspension.

4.2 Indirect match of transformants

One of the issues discussed in the previous section was tracking substances and their properties through processes, mixtures and transformations (like lysing) in particular. This knowledge can be utilized to give a more complete answer to user queries.

> Rule 2. If substance X is transformed into substance Y and process A is applied to Y, then process A is indirectly applied to X.

One of the substeps of purification in the experiment described above is breaking cells by French press. As a result of the process, cells are replaced by a lysate, which is a mixture of all the cells' ingredients but without

separation by cell walls. That is, all the stuff is now in one unstructured mixture, as shown at the right of Figure 7. This raises two issues. One is representing this change formally. This was discussed in Section 3. Second, representing the fact that cells don't exist any more and all the subsequent processes are applied to the lysate (or purified parts of it). Being able to represent this will give us the knowledge that, first, all the stuff that was in the cells, is still present, and, second, it is not referred to as "cells" anymore. So, if someone asks:

> "Were the cells ever ultracentrifuged in the process of purifying CheW by the procedure described in Stock *et al.* ?"

the desirable answer from investigating this paper would be "ultracentrifugation was not applied to the cells directly, but to the lysate derived from the cells". This type of query-answer interface is similar to *cooperative responses* technique used in natural language [Kaplan 1982].

In general, if a query is made about a process applied to a substance and we know that this process was applied to a transformant or a precursor of this substance, we would like to bring it up to the user as a possible answer. In particular, if, as in this example, the transformant has exactly the same stuff in it as the sought substance, but is structured differently.

4.3 Coherence Inferences

An intelligent retrieval system's goal is, in large part, to match a user's query to the appropriate text strings which can then be displayed. Domain and world knowledge is used in a variety of ways to match the query to elements of the text's meaning that are not explicitly present. *Coherence* is the assumption that consecutive sentences in a text are related, by cause and effect, goal and means, or some other relationship at the meaning level.

The second paragraph above starts with a sentence that states the goal: "purification of CheW". The assumption that the text is coherent tells us that the following sentences describe the means of achieving this goal. It also tells us that the order of steps described represents their temporal sequence (since there is no mention of the contrary). This allows us to, first, answer queries about the paragraph as a whole and, second, relate the output of each step as the input to the next.

> Rule 3. If G is goal of an event sequence S and process P is part of sequence S, then G is a goal of process P.

> Rule 4. If substance X is used in process P and G is the goal of P then X was used to achieve G.

For example, as we mentioned, all the processes in the sequence described in the second paragraph in Figure 8 are serving one particular goal: purification of CheW protein. But this goal, after being stated in the first sentence of the paragraph, is hardly mentioned in the subsequent processes' descriptions. However, a query may often pertain to the general goal of the paragraph. For example, consider the following query:

> "Has anyone attempted to purify CheW from *E. coli* strain RP3098?"

It is easy to see that the entire excerpt should be retrieved as an answer to the query. The only process that *E. coli* strain RP3098 explicitly participates in is Grow.

Assuming that the paragraph is coherent, we can infer that CheW was, in fact, purified from this strain, unless stated otherwise.

5. Conclusion

In this paper we have presented the challenges to the current ontological paradigms that arise in the formalization of biological materials and methods. We claim that these problems also manifest themselves in other experimental sciences. We discuss the possible solutions to the outlined problems.

Many ontologies divide Tangible objects into Stuff and Discrete objects. Since this is not adequate to categorize our domain, we introduce a new ontological category, *mixture*, which exhibits the properties of both discrete objects and stuff. Mixture is an important class of heterogeneous population. We introduce made-of and sequence relations to represent extensions of the traditional part-of hierarhcy.

Transformations are one of the most common biological and experimental processes. Transformations not only change certain properties of their participants but participants can change their category and, therefore, their identity as a result of a transformation. It is important, however, to represent the fact that the stuff the inputs to a transformation process were made from is the same stuff the outputs are made from. This information can be then used in intelligent information retrieval. We suggest complementing the representation used in Qualitative Process Theory with predicates and concepts to track the changes in the stuff that occur during transformations.

This is research in progress and our goal is overall ontology design for biological materials and methods.

Acknowledgements

This research is supported in part by the National Science Foundation under Grant No. IRI-9117030. Our thanks to Robert P. Futrelle, Arthur W. Miller and the anonymous reviewers for helpful comments on an earlier draft of this paper.

References

Baclawski, K., Futrelle, R., Fridman, N. and Pescitelli, M. (1993). Data/knowledge bases for biological papers and techniques. In *Proceedings of the Symposium on Advanced Data Management for the Scientist and Engineer*, 23-28: AAAS.

Barr, A. and Feigenbaum, E.A. (1981). STRIPS. In *The Handbook of Artificial Intelligence*, , 128-134. Stanford, CA: HeurisTech Press.

Bateman, J.A., Magnini, B. and Rinaldi, F. (1994). The Generalized Italian, German, English Upper Model. In *ECAI'94 Workshop: Comparison of Implemented Ontologies*. Amsterdam.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Mayer, E.F.J., Bryce, M.D., Rodgers, J.R., Kennard, O., Simanouchi, T. and Tasumi, M. (1977). The protein data bank. *Journal of Molecular Biology* 112: 535-542.

Cohen, P.R. and Perrault, C.R. (1979). Elements of plan-based theory of speech acts. *Cognitive Science* 3: 177-212.

Forbus, K.D. (1984). Qualitative Process Theory. *Artificial Intelligence* 24: 85-168.

Gegner, J.A. and Dahlquist, F.W. (1991). Signal transduction in bacteria: CheW forms a reversible complex with the protein kinase CheA. *Proceedings National Academy Sciences* 88: 750-754.

Gruber, T.R. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. KSL 93-04. Knowledge Systems Laboratory, Stanford University.

Hafner, C., Baclawski, K., Futrelle, R., Fridman, N. and Sampath, S. (1994). Creating a Knowledge Base of Biological Research Papers. In *2nd Inter'l Conf. on Intelligent Systems for Molecular Biology*. Stanford, CA: AAAI Press.

Humphreys, B.L. and Lindberg, D.A.B. (1993). The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association* 81(2): 170.

Kaplan, S.J. (1982). Cooperative responses from a portable natural language query system. *Artificial Intelligence* 19(2): 165-187.

Karp, P.D. (1993). A Qualitative Biochemistry and Its Application to the Regulation of the Trytophan Operon. In *Artificial Intelligence and Molecular Biology*, ed. L. Hunter, 289-325. AAAI Press/ The MIT Press.

Karp, P.D., Riley, M., Paley, S.M. and Pelligrini-Toole, A. (1996). EcoCyc: Encyclopedia of *E. coli* Genes and Metabolism. *Nucleic Acids Research* 24(1): 32-40.

Lehman, F. (1995). Combining Ontologies, Thesauri, and Standards. In *IJCAI Workshop on Basic Issues in Knowledge Sharing*, 84-94. Montreal, Canada.

Lenat, D.B. and Guha, R.V. (1990). *Building large knowledge-based systems: representation and inference in the Cyc project*. Reading, Mass: Addison-Wesley Pub. Co.

Sowa, J.F. (1995). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Boston, MA: PWS Publishing Company.

Weld, D.S. (1986). The Use of Aggregation in Causal Simulation. *Artificial Intelligence* 30: 1-34.