Representations of Metabolic Knowledge

Peter D. Karp
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
pkarp@ai.sri.com
415-859-6375, fax 415-859-3735

and

Monica Riley
Marine Biological Laboratory
Woods Hole, MA 02543

mriley@hoh.mbl.edu
508-548-3705, fax 508-540-6902

February 11, 2009

Abstract

Construction of electronic repositories of metabolic information is an increasingly active area of research. Encoding detailed knowledge of a complex biological domain requires finely honed representations. We survey representations used for several metabolic databases, including EcoCyc, and reach the following conclusions. Representation of the metabolism must distinguish enzyme classes from individual enzymes, because there is not a one-to-one mapping from enzymes to the reactions they catalyze. Individual enzymes must be represented explicitly as proteins, e.g., by encoding their subunit structure. The species variation of metabolism must be represented. So must the substrate specificity of enzymes, which may be treated in several ways.

1 Introduction

The construction of electronic repositories of metabolic information is an increasingly active area of research. A crucial question in the endeavor of encoding knowledge of metabolism in electronic form is: How should that knowledge be represented? Put another way, what conceptualization of metabolism adequately captures the intricacies of this complex domain? This paper considers the representations used in a number of existing metabolic databases, and the representation under development by the authors as part of the EcoCyc project.[1]

We use the words *representation*, *conceptualization*, and *ontology* synonymously in this paper to mean the classes of entities that exist in a computer model, the relationships

[1]This paper was published in *Proceedings of First International Conference on Intelligent Systems for Molecular Biology*, Bethesda, MD, 1993, Morgan Kaufmann Publishers, pp207–215.

recorded among those entities, and the properties ascribed to them. For example, different models might employ different representations of proteins, such as: an entity that has a name; an entity that has a set of subunits, each of which is named; an entity that has a set of named subunits in specified stoichiometries, where each subunit has a property called molecular weight; an entity consisting of a sequence of amino acids, where the position of each amino acid in three-dimensional space is given.

There exists no absolute standard for comparing the goodness of two representations. We can assess the goodness of representations only relative to the problem-solving task or tasks for which they were designed. Therefore, representations that were designed for different tasks are incommensurable. Representations that serve common tasks can be compared according to a number of criteria, including computational tractability, storage requirements, and the ability to encode distinctions that are relevant to that task. This paper will be most concerned with the last criterion. That some previous researchers do not clearly state the task(s) for which their representation was designed — and for that matter, often do not clearly define the conceptualization that they employ — further complicates the evaluation of these representations.

The EcoCyc metabolic knowledge base (KB) is designed for use in a number of application tasks, including:

- Construction of an electronic encyclopedia accessible via a graphical user interface with which scientists can retrieve facts about *E. coli* metabolism and navigate through the information space of metabolism
- Answering complex queries that require biochemical reasoning, such as "what single-subunit enzymes are assisted by magnesium ions in carrying out oxidation of 3-carbon compounds?"
- Simulation of metabolic pathways, both qualitative and quantitative
- Design of novel metabolic pathways in biotechnology
- Investigating the evolution of metabolism
- Construction of a computer-aided instruction system for the metabolism
- Reference use by molecular biologists and biotechnologists who use *E. coli* as the universal tool to clone genes of other organisms, and as a host organism to produce gene products in quantity

This diverse set of activities requires a rich, detailed array of knowledge: we therefore seek *high-fidelity* representations that capture many nuances of this complex subject matter.

A few methodoligical comments about the motivations behind this paper are in order. *Biological Knowledge Representation* is an important subfield of ISMB. Its goal is to articulate and evaluate alternative representations of biological knowledge about topics such as protein structure, biochemical reactions, experimental techniques, genomic maps, and macromolecule sequences. Unfortunately, representations for most existing biological DBs have never been published with a detailed explanation, rationale or design history, making it difficult to build on the work of previous researchers. One might argue that since the biology literature already discusses all of the concepts that need representing, it is a

waste of time to write papers about biological knowledge representation. On the contrary, the biological literature does not consider what subset of all knowledge (or what level of abstraction) is required for a particular problem-solving task; it often lacks the precision and detail that computational tasks demand; and it does not discuss the computational tradeoffs of different representations. This paper demonstrates the difficulty of moving from biological concepts in the literature to a formal representation of that knowledge by identifying shortcomings in the representations used by several existing metabolic DBs.

2 Existing Electronic Repositories of Metabolism

This section provides a brief overview of past and present projects to encode metabolic information in computer form.

The precursor of all of these projects is the book of enzyme nomenclature developed by the International Union of Biochemistry and Molecular Biology [14]. The 1992 edition is the latest in a series of publications that began in 1961. This book contains a detailed classification of many enzyme-mediated reactions (2,800 reactions as of the 1992 edition).

Three early metabolic databases (DBs) were as follows. Seressotis and Bailey's DB (which we call S&B) described 90 reactions [12]. A database created at the 1987 Biomatrix workshop (which we call BIOMATRIX) described 120 reactions [6]. Mavrovouniotis' DB (which we call MAVRO) described 245 reactions [5].

More recently, Bairoch's ENZYME DB is an electronic version of the reactions in [14], and therefore describes 2,800 reactions [1]. Ochs and Conrow's DB (called METPROTO) was derived from ENZYME, although some reactions were omitted and some of the chemical compounds were renamed [7].

Selkov and his colleagues in Russia have had a large data-gathering operation under way for over 10 years, called DBEMP for Database on Enzymes and Metabolic Pathways [11]. This DB contains information on enzymes, pathways, reaction mechanisms, phenotypes, noncatalytic proteins, enzyme nomenclature, and mathematical models. Overbeek and his colleagues recently prepared a metabolic knowledge base, which we call OVER. No publications describes this effort, so we find it difficult characterize the DB precisely. Kazic is constructing a metabolic KB for *E. coli* [4, 3], which we call KAZ. Rouxel et al. are also constructing a metabolic DB for *E. coli* [9] called METALGEN.

Karp and Riley's EcoCyc KB currently contains *E. coli*-specific information for 50 enzymes involved in the TCA cycle, glycolysis, and amino acid biosynthesis, but its goal is to describe all of *E. coli* intermediary metabolism. It also describes 1,000 metabolic compounds [2], plus the 2,800 enzyme classes defined in [14]. EcoCyc will integrate metabolic and genetic information; the genetic data will be derived from two related compilations of *E. coli* genes and genes and gene products [8, 10].

These efforts employ a variety of techniques for encoding the information including flat files (ENZYME), keyed binary files (SELKOV), relational database management (MET-

PROTO, METALGEN), LISP S-expressions (MAVRO, BIOMATRIX, S&B), PROLOG (OVER, KAZ), and frame knowledge representation (EcoCyc). Only METPROTO and METALGEN include software for manipulating the database, although a graphical user interface is under development for EcoCyc.

3 Level of Interpretation

We begin by considering a broad conceptual decision that the DB designer must make. At what level of interpretation is the information that will be encoded? For example, does it consist of raw data from instruments, or information acquired from the primary literature, or is it a consensus view of the literature as found in a textbook or compendium? Virtually all of the DBs we consider use a consensus-level conceptualization. EcoCyc is an exception. Although its information on reactions is commonly accepted and derived from textbooks and other secondary sources, its information on the enzymes of *E. coli* is derived completely from the primary literature, with literature citations available for each finding. The most relevant and most recent articles are selected for citation. The SELKOV DB is also closely based on the primary literature.

The level of interpretation often affects the organization of the DB: consensus-level DBs are organized around the domain concepts of interest (such as enzymes and reactions) so that only one instance of a given entity exists in the DB. In contrast, observation-level DBs are often organized around the observation (such as a publication) and encode distinct observations of the same entity separately. For example, a publication-level DB such as would create two records for the enzyme fumarase if it were described in two separate publications, whereas the ENZYME DB contains only a single entry for fumarase because it presents a consensus view of the different observations about that enzyme.

Although EcoCyc contains a single frame for each enzyme,[2] its properties such as subunit structure, prosthetic group, and sensitivities to activators and inhibitors, are all annotated with specific literature citations to the experimental work on each property. When different laboratories report conflicting information about a property, we have the capability to record those conflicts explicitly, with each labeled with a citation.

4 Separating Enzymes from Reactions

A reaction is a transformation in which chemical bonds are formed, or broken, or both. An enzyme is a protein catalyst that lowers the activation energy of a reaction. The official convention is to assign each enzyme a name and a unique identifying number called the EC (Enzyme Commission) number, based on the reaction it catalyzes [14]. However, it is crucial to recognize that enzymes and reactions are different concepts, and

[2]The next section explores a different reason for breaking the description of an enzyme into multiple frames, namely that multiple proteins (isozymes) can exhibit the same enzymatic activity.

that a one-to-one correspondence does not always exist between the two. That is, one protein sometimes catalyzes more than one reaction, and the same reaction is sometimes catalyzed by more than one protein. Multiple proteins that catalyze the same reaction may or may not show sequence homology.

For example, *E. coli* contains three different enzymes that catalyze the hydration of fumarase to malate. Two of the enzymes are homodimers; the third is a homotetramer. The polypeptide subunits of these enzymes are encoded by three distinct genes. Other examples in *E. coli* are the reaction glyceraldehyde 3-phosphate-dehydrogenation, which is catalyzed by the products of the genes *gapA* and *gapB*; and the reaction pyruvate phosphorylation, which is catalyzed by the products of both *pykA* and *pykF*. Conversely, the polypeptide encoded by the *E. coli* gene *eda* catalyzes two different reactions: 4-hydroxy-2-oxoglutarate aldolase and 2-dehydro-3-deoxyphosphogluconate aldolase. Other polypeptides that catalyze multiple reactions in *E. coli* are encoded by the genes *hisB*, *hisI*, *fadB*, *trpC*, and *trpD*.

No metabolic DB except for EcoCyc draws this distinction. Most other DBs (such as ENZYME, MAVRO, and METPROTO) are organized as a list of what their developers call enzymes, but what in fact are enzyme *classes* [14]. An enzyme class is defined by a set of enzymes (from one species or from multiple species) that all carry out the same reaction. There is little more to say about an enzyme class than to describe the associated reaction. Although a one-to-one mapping exists between an enzyme class and a reaction, that relationship does not hold between enzyme classes and enzymes. Therefore, it is less appropriate to informally call an enzyme class an enzyme — as have most previous researchers [1] — than it is to call an enzyme class a reaction.

The failure to draw this distinction between enzymes and reactions becomes problematic when we attempt to encode more detailed information about enzymes and reactions, such as molecular weight, activators, inhibitors, subunit composition, and equilibrium constant. Some of this information is specific to the enzyme, and some is specific to the reaction. If we blur together a reaction with two enzymes that catalyze it, we cannot record the fact that the two enzymes have different molecular weights, different inhibitors, and different genes; the equilibrium constant, in contrast, is specific to the reaction and is independent of the enzyme(s) that catalyze the reaction.

The EcoCyc representation employs three KB classes for this conceptualization: `reaction, protein`, and `enzymatic-reaction.` Instances of the latter class link an enzyme with the reaction(s) it catalyzes, and encode information about a reaction as it is catalyzed by a particular enzyme.

Figure 1 illustrates this representation for the three enzymes that hydrate fumarate to form malate. The reaction is described as an instance of class `reaction`. The three enzymes are instances of class `protein`. Three separate instances of class `enzymatic-reaction` describe the catalysis of fumarate hydration by each of the three isozymes.

Figures 2 and 3 show the `reaction` and `enzymatic-reaction` frames in more detail. In a step that may initially be counterintuitive, the EC number is a property of reactions,

Figure 1: Three different proteins catalyze the hydration of fumarate to malate. In this semantic-network depiction, boxes represent frames whose class is shown in italics and whose name is shown in bold face. Arrows designate relations between frames. This figure omits additional information that is stored at each frame.

not of enzymes, since there is a one-to-one correspondence between reactions and EC numbers. Other properties of the reaction include the compounds on the two sides of the reaction and their coefficients in the reaction, and the change in Gibbs free energy of the reaction. The reaction also lists the enzymatic-reactions for enzymes that catalyze the reaction. As for several of the other classes, slots are defined for comments, a general list of citations to the reaction, and synonyms for its name.

The enzymatic-reaction frame records information that is specific to the catalysis of this reaction by this enzyme. It lists different types of activators, inhibitors, and cofactor molecules. Although the ENZYME database lists that information on a per-reaction basis, the different enzymes that catalyze a reaction may respond to different regulatory molecules, and an enzyme that catalyzes multiple reactions may interact with different regulatory molecules for each. In the `reaction` class, We call the two sides of the reaction `left` and `right` to avoid the directionality connotations of the words "reactants" and "products," since in a purely chemical sense, almost all reactions are reversible to some degree. The slot `reaction-direction` allows us to record the directionality of those few reactions that are essentially irreversible. It is tempting to try to record the "normal" physiological direction of a reaction in the cell, but it is common for a reaction to be driven in varying directions due to the changes in substrate concentrations caused by different growth media.

5 Enzymes Are Proteins

Although enzymes are proteins, few past DB architects have conceptualized enzymes as proteins. Three exceptions are that ENZYME lists the accession numbers of SwissProt entries that contain sequence data for members of each enzyme class, and OVER lists the

```
(define-instance FUMHYDR_RXN
  :template (REACTION)
  :slots
  (COMMON-NAME "fumarate hydration"
         :max-cardinality 1 :value-class STRING)
  (SYNONYMS "malate dehydration"
         :value-class STRING)
  (EC-NUMBER "4.2.1.2" :max-cardinality 1
         :value-class STRING)
  (CATALYZED-BY FUMARA_ENZRXN
         :value-class ENZYMATIC-REACTION)
  (LEFT (fumarate H2O) :value-class CHEMICAL)
  (LEFT-COEFFICIENTS (1 1) :value-class integer)
  (RIGHT (malate) :value-class CHEMICAL)
  (RIGHT-COEFFICIENTS (1) :value-class integer)
  (DELTAGO' 0 :max-cardinality 1
         :value-class NUMBER)
  (COMMENT)
  (CITATIONS)
  )
```

Figure 2: An instance frame of types `reaction`. This definition gives the name of the frame, its template (defining) classes, and the slots it contains. The definitions of individual slots list the slot names, their values (if present), and constraints on the allowable values of the slots, such as the maximum number of values, and the type (value-class) of individual values.

```
(define-instance FUMARA_ENZRXN
  :template (ENZYMATIC-REACTION)
  :slots
  (ENZYME FUMARASE_A :max-cardinality 1
          :value-class PROTEIN)
  (REACTION FUMHYDR_RXN :max-cardinality 1
          :value-class REACTION)
  (REACTION-DIRECTION REVERSIBLE :max-cardinality 1)
  (COFACTORS/PROSTHETIC-GROUPS :value-class CHEMICAL)
  (ACTIVATORS-ALLOSTERIC :value-class CHEMICAL)
  (ACTIVATORS-NONALLOSTERIC :value-class CHEMICAL)
  (ACTIVATORS-MECHNOTSTATED :value-class CHEMICAL)
  (INHIBITORS-COMPETITIVE :value-class CHEMICAL)
  (INHIBITORS-ALLOSTERIC :value-class CHEMICAL)
  (INHIBITORS-NEITHER :value-class CHEMICAL)
  (INHIBITORS-MECHNOTSTATED :value-class CHEMICAL)
  (ALTERNATE-SUBSTRATES :value-class CHEMICAL)
  (ALTERNATE-COFACTORS :value-class CHEMICAL)
  (CITATIONS)
  (COMMENT "Three isozymes in E. coli:
            [88268900], [88193096]")
  )
```

Figure 3: An instance frame of type `enzymatic-reaction`.

names of genes that encode the enzymes that carry out a listed reaction, when known. SELKOV lists physical properties of enzymes, and describes enzyme-purification procedures.

As shown in Figure 1, EcoCyc explicitly separates the description of a protein that is an enzyme from the reaction that it catalyzes. The EcoCyc representation also distinguishes enzyme complexes that contain subunits from enzymes that consist of a single polypeptide chain. The class `protein` is covered by two subclasses: `protein-complex` and `polypeptide`. A protein complex is built of one or more smaller components — which in turn can be protein complexes or polypeptides (Figure 4). The stoichiometries of the subunits are also encoded. In contrast, a polypeptide has no subunits. Our representation includes various physical properties of polypeptides and protein complexes, such as their molecular weight as determined from sequence, their pI, and their accession number in the Niedhardt compilation of electrophoretic properties of *E. coli* proteins [13]. We expect that information about the physical properties of proteins may be particularly useful to biotechnologists. A single polypeptide is encoded by a single gene, therefore the slot `gene` in the `polypeptide` class identifies the gene associated with a polypeptide. Slot `locations` allows us to record the cellular location of a protein; allowable values for *E. coli* are `cytoplasm`, `membrane` (when exact membrane location is unknown), `inner-membrane`, `outer-membrane`, `membrane-spanning` (i.e., both inner and outer membrane), and `periplasm`. The slot `isozyme-sequence-similarity` allows us to record whether significant known sequence similarity exists between a given polypeptide and known isozymes.

Information on individual genes is obtained from the Ecogene database [10]. For each of 1,500 *E. coli* genes, Ecogene lists the gene name, the physical map position of the gene on the *E. coli* chromosome in centisome units (1 centisome = 1% of a chromosome length), and the direction of transcription (+ or − strand). Each gene is encoded as a member of the `gene` class (Figure 6).

6 Species Variation of Metabolism

Different species have different complements of enzymes. For example, *E. coli* is able to synthesize all 20 amino acids, whereas humans have lost several of the amino-acid biosynthetic pathways; yet humans have many secondary metabolic pathways that bacteria lack, such as those that synthesize steroids.

Because enzymes are proteins, and because protein sequences vary among different species, it follows that the precise function and physical properties of enzymes will vary among different organisms and sometimes among different tissues of the same organism. In some cases, enzymes from different sources that carry out the same reaction are very similar to one another, regardless of differences between their hosts. In other cases, the enzymes for a given reaction from different sources are entirely different, and it is necessary to specify the host and to take care not to combine data on the same enzyme from different sources.

Figure 4: The enzyme succinate dehydrogenase has four subunits. Two have catalytic functions and two anchor the protein to the cell membrane.

```
(define-instance FUMA_MONOMER
  :template (POLYPEPTIDE)
  :slots
  (COMMON-NAME "fumarase A monomer"
          :max-cardinality 1 :valueclass STRING)
  (SYNONYMS :valueclass STRING)
  (COMMENT "Similar to fumB, not fum C"
          :valueclass STRING)
  (CITATIONS)
  (GENE fumA :max-cardinality 1  :valueclass GENE)
  (MOLECULAR-WEIGHT-SEQ 60.16 :max-cardinality 1
          :valueclass NUMBER)
  (LOCATIONS)
  (NEIDHARDT-SPOT-NUMBER)
  (PROSTHETIC-GROUP :max-cardinality 1
          :valueclass CHEMICAL)
  (pI :max-cardinality 1  :valueclass NUMBER)
  (ISOZYME-SEQUENCE-SIMILARITY)
  )
```

Figure 5: An instance of class `polypeptide`.

```
(define-instance TRPA-GENE
  :template (GENE)
  :slots
  (TRANSCRIPTION-DIRECTION "-" :max-cardinality 1
                :valueclass STRING)
  (COMMON-NAME "trpA" :max-cardinality 1
                :valueclass STRING)
  (MAP-POSITION 28.342 :max-cardinality 1
                :valueclass STRING)
  )
```

Figure 6: An instance of class `gene`.

We expect that most scientists who query a metabolic database will wish to know if the answer pertains to their organism of interest. For example, data on the properties of the lactate dehydrogenase of mammalian muscle are not applicable to the mammalian heart lactate dehydrogenase, nor to either the d or l-lactate dehydrogenases of *E. coli*. Each of these enzymes has unique physical and chemical properties, and unique sensitivities to regulatory molecules that have been studied by scientists on an enzyme-by-enzyme basis. But because most metabolic databases do not encode the validity of their data on a per-species basis, the relevance of their data to a particular species is impossible to ascertain. For example, ENZYME is a compendium of many different reactions from many species.

METPROTO is an exception because each reaction contains a field that in the future will indicate whether that reaction occurs in each of several species. In addition, the METPROTO representation encodes what metabolic space (i.e., cellular compartment) the reactants and products of a reaction reside in, thus allowing transport events to be encoded alongside biochemical reactions. EcoCyc, METALGEN, and KAZ are also exceptions because each encodes information for *E. coli* only. The conceptualization for EcoCyc described in the previous sections generalizes nicely to the task of representing information about multiple species, since enzymes from additional species are simply additional proteins that can be described as instances of the `protein` class. These enzymes might carry out reactions that are already encoded in the KB, or previously unknown reactions that must be added to the KB. In either case, the linkages between enzymes and reactions are expressed with new `enzymatic-reaction` frames that allow us to encode the fact that enzymes from different species may catalyze different reactions in different ways, e.g., subject to the control of different regulatory molecules. EcoCyc will not only define those reactions present in *E. coli* but will provide specific information on each *E. coli* enzyme and enzymatic reaction.

7 Substrate Specificity

Different enzymes show varying degrees of specificity for their substrates. Although a single reaction equation is typically written to describe the activity of an enzyme, the enzyme might actually catalyze a family of related reactions involving similar substrates. Biochemical knowledge of the degree of substrate specificity for different enzymes is incomplete, but it is desirable to be able to encode as much of the existing information as possible.[3]

Substrate specificity is represented in two different ways in the ENZYME database, as shown in Figure 7. The entry for alcohol dehydrogenase encodes the reaction in a generalized fashion, listing compound classes such as "alcohol" rather than specific compounds such as methyl alcohol. The entry for glucose 1-dehydrogenase illustrates the second approach, in which a comment field describes alternative substrates. We refer to these approaches as (1) and (2), respectively, and we describe some other possible approaches.

[3]Many of the problems and solutions discussed in this section also pertain to representing the variability of the cofactors, activators, and inhibitors that modulate the activity of an enzyme.

(a)

```
DE   ALCOHOL DEHYDROGENASE.
CA   ALCOHOL + NAD(+) = ALDEHYDE OR KETONE + NADH.
CC   -!- ACTS ON PRIMARY OR SECONDARY ALCOHOLS OR HEMIACETALS.
CC   -!- THE ANIMAL, BUT NOT THE YEAST, ENZYME ACTS ALSO ON CYCLIC
CC       SECONDARY ALCOHOLS.
```

(b)

```
DE   GLUCOSE 1-DEHYDROGENASE (NADP+).
CA   D-GLUCOSE + NADP(+) = D-GLUCONO-1,5-LACTONE + NADPH.
CC   -!- ALSO OXIDIZES D-MANNOSE, 2-DEOXY-D-GLUCOSE AND 2-AMINO-2-DEOXY-D-
CC       MANNOSE.
```

Figure 7: Two entries from the ENZYME database that demonstrate different representations of substrate specificity: (a) naming substrates as classes, and (b) encoding substrate variation within comments. Lines marked DE name the enzyme class, CA give the reaction equation, and lines marked CC are comments.

Among the drawbacks of (1) are the following. A name may not always exist for the compound class that precisely describes the activity of the enzyme. The statement that alcohols are a substrate of alcohol dehydrogenase may be overly general if in fact the enzyme will not accept certain alcohols as a substrate, and no name may exist that encompasses the exact subset of alcohols that are acceptable substrates. Another approach (3) is to describe substrates as generalized compound structures such as **CH–OH, where the asterisks denote wildcard positions in the structure. This approach is also flawed in that the constraints on structures will usually be three-dimensional in nature, thus requiring a richer description language. But since precise three-dimensional constraints on substrate structures are usually not known in great detail, this approach will not usually be worth the effort.

Another approach (4) is to explicitly list the equations for every known reaction that the enzyme catalyzes. Although prolix, this approach has the advantage of precisely stating those (and only those) reactions that have been observed experimentally. Its precision becomes a disadvantage, however, when we do not know every substrate in the equation; that is, we may know that an enzyme accepts a range of reactants, but we may not know or be able to easily deduce the corresponding products.

The next approach (5) is less precise; it merely states that compound X will substitute for compound Y in a reaction, without identifying the other reactants of the corresponding reaction. However, (5) is more precise than (2); the comments used in (2) do not always make clear what compounds substitute for what others. The other drawback of (2) is that employing English comments prevents the representation from being machine parsable.

None of these approaches provide information about the relative rates of the reactions involving alternative substrates. Another shortcoming common to all of the approaches is that none distinguish physiologically significant substrates of an enzyme from substrates

that have only been demonstrated in *vitro*.

Although we have chosen (5) for use in EcoCyc, none of these representations are perfect, and a combination of representations will probably be required to represent the range of information that exists on substrate specificity. As the project proceeds we will better ascertain the frequency of the different types of information in the literature.

8 Summary

We survey representations used for several metabolic databases, and reach the following conclusions. The representations used in other metabolic databases have overlooked important aspects of metabolism. Representation of the metabolism must distinguish enzyme classes from individual enzymes, because there is not a one-to-one mapping from enzymes to the reactions they catalyze. Individual enzymes must be represented explicitly as proteins, e.g., by encoding their subunit structure. The species variation of metabolism must be represented. So must the substrate specificity of enzymes, which may be treated in a variety of ways to capture the different types of partial knowledge that may exist about substrate specificity. Our future work will explore additional representational issues not explored here, such as the representation of reaction mechanisms.

Acknowledgments

Many of the ideas in this paper, especially that of separating enzymes from enzyme classes, benefited from early discussions with G. Christian Overton. This work was supported by Grant 1-R01-RR-07861-01 from the National Center for Research Resources.

References

[1] A. Bairoch. Enzyme database, centre medical universitaire, geneva. 1992.

[2] P.D. Karp. A knowledge base of the chemical compounds of intermediary metabolism. *Computer Applications in the Biosciences*, 8(4):347–357, 1992.

[3] T. Kazic. Biochemical databases: Challenges and opportunities. In *Thirteenth International CODATA Conference*, 1993. In press.

[4] T. Kazic. Representation, reasoning, and the intermediary metabolism of *Escherichia coli*. In *Proc 26th Annual Hawaii International Conference on System Sciences*, volume I, pages 853–862. IEEE Computer Society Press, 1993.

[5] M. Mavrovouniotis. Group contributions for estimating standard Gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnology and Bioengineering*, 36:1070–1082, 1990.

[6] H.J. Morowitz and T. Smith. Report of the matrix of biological knowledge workshop. Technical report, Santa Fe Institute, Santa Fe, NM, 1987.

[7] R.S. Ochs and K. Conrow. A computerized metabolic map. *J. Chem. Inf. Comput. Sci.*, 31:132–137, 1991.

[8] M. Riley. Functions of the gene products of *Escherichia coli*. *Microbiological Reviews*, 57:862–952, 1993.

[9] T. Rouxel, A. Danchin, and A. Henaut. METALGEN.DB: Metabolism linked to the genome of *Escherichia coli*, graphics oriented database. *Computer Applications in the Biosciences*, 9(3):315–324, 1993.

[10] K. E. Rudd, G. Bouffard, and W. Miller. Computer analysis of *E. coli* restriction maps. In K. E. Davies and S. M. Tilghman, editors, *Genome Analysis, Vol.4: Strategies for Physical Mapping*, pages 1–38. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992.

[11] E.E. Selkov, I.I. Goryanin, N.P. Kaimatchnikov, E.L. Shevelev, and I.A. Yunus. Factographic data bank on enzymes and metabolic pathways. *Studia Biophysica*, 129(2–3):155–164, 1989.

[12] A. Seressiotis and J.E. Bailey. MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways. *Biotechnology and Bioengineering*, 31:587–602, 1988.

[13] R.A. Van Bogelen, M.E. Hutton, and F.C. Niedhart. Gene-protein database of *Escherichia coli* K-12: Edition 3. *Electrophoresis*, 11(11), 1990.

[14] Edwin C. Webb. *Enzyme Nomenclature, 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, 1992.

Introduction to Knowledge Representation. Note. Slides are at the bottom of the page! until I correct itâ€”thereâ€™s a typo on slide 46, the conjunction and disjunction symbols are backwards (true: ^ = and ^ v = or). Latex should have caught this error :) This section is largely based on Sowaâ€™s Conceptual Structures; Sussman and Abelsonsâ€™s SICP, Suppeâ€™s Introduction to Logic, Partee et alâ€™s Mathematical Methods in Linguistics, Cruseâ€™s Lexical Semantics and Thinking about Android Epistemology. The big questions.Â We can load the knowledge into the computer, or we can build a learning system that builds its own representation from examples. Using the knowledge. Assuming weâ€™ve got the knowledge, how do we use it to solve problems? (And what problems are worth solving?)

Knowledge Representation in AI describes the representation of knowledge. Basically, it is a study of how the beliefs, intentions, and judgments of an intelligent agent can be expressed suitably for automated reasoning. One of the primary purposes of Knowledge Representation includes modeling intelligent behavior for an agent. Knowledge Representation and Reasoning (KR, KRR) represents information from the real world for a computer to understand and then utilize this knowledge to solve complex real-life problems like communicating with human beings in natural language. Our representation of metabolic pathways -- such as the TCAcycle, glycolysis, and tryptophan biosynthesis -- facilitates both knowledgeacquisition of pathways, and automatic pathway drawing. The chief contributions of this paper are a minimized representation for biochemical pathwayscalled the predecessor list, and inference procedures for converting the predecessor list into a pathway-graph representation that can serve as input to a pathway-drawing algorithm. The predecessor list has several advantages over the pathway graph, including its compactness and its lack of redundancy.Â The usefulness of the predecessor list for knowledge acquisition should make this representation a staple of any metabolic database project. î€Imperial College Press. Knowledge-based generalization of metabolic. Networks: a practical study. Anna zhukova.Â To support the graphical representation of "cancer" pathways, we have adapted our Pathway Browser to display disease variants and events in a way that allows comparison with the wild type pathway, and shows connections between perturbations in cancer and other biological pathways. The curation of pathways associated with cancer, coupled with our efforts to create other disease-specific pathways, will interoperate with our existing pathway and network analysis tools.