BIOINFORMATICS ONTOLOGY

Vol. 16 no. 12 2000
Pages 1120–1128

A knowledge model for analysis and simulation of regulatory networks

Andrey Rzhetsky[1, 2,*], Tomohiro Koike[1, 3], Sergey Kalachikov[1], Shawn M. Gomez[1], Michael Krauthammer[2], Sabina H. Kaplan[1], Pauline Kra[2, 4], James J. Russo[1] and Carol Friedman[2, 5]

[1]Columbia Genome Center, Columbia University, [2]Department of Medical Informatics, Columbia University, [3]Hitachi Software Engineering Co. Ltd, [4]Department of French, Yeshiva University and [5]Department of Computer Science, Queens College of the City University of New York

Received on October 7, 1999; revised on April 24, 2000; accepted on July 20, 2000

Abstract

Motivation: *In order to aid in hypothesis-driven experimental gene discovery, we are designing a computer application for the automatic retrieval of signal transduction data from electronic versions of scientific publications using natural language processing (NLP) techniques, as well as for visualizing and editing representations of regulatory systems. These systems describe both signal transduction and biochemical pathways within complex multicellular organisms, yeast, and bacteria. This computer application in turn requires the development of a domain-specific ontology, or knowledge model.*

Results: *We introduce an ontological model for the representation of biological knowledge related to regulatory networks in vertebrates. We outline a taxonomy of the concepts, define their 'whole-to-part' relationships, describe the properties of major concepts, and outline a set of the most important axioms. The ontology is partially realized in a computer system designed to aid researchers in biology and medicine in visualizing and editing a representation of a signal transduction system.*

Availability: *The knowledge model can be reviewed at http://genome6.cpmc.columbia.edu/tkoike/ontology/*

Contact: *ar345@columbia.edu*

Introduction

A large body of knowledge that has become available recently through the internet consists of electronic versions of articles published in scientific journals, such as *Science* or *Cell*. This creates new opportunities for automated knowledge acquisition: if information contained in these articles can be extracted and organized, it could then be stored in a knowledge base and used in computational analyses such as data mining. While there is a prototype computer system for extracting medical knowledge from research articles in medicine (e.g. Hahn *et al.* (1996)), to our knowledge there is no equivalent system within the fields of Genomics and Molecular Biology. We aim to fill this gap by designing a system for the automatic extraction of functional information describing regulatory relationships between genes and proteins from online versions of research articles. Below, we provide the rationale for developing such a system.

The natural sciences proximal to molecular biology and medicine currently enjoy exponential growth. Individual researchers often find themselves unable to keep pace with the rate of information accumulating in multiple fields just outside their focus. Virtually every field of modern biology is likely to profit from the methodological advances that help scientists cope with information overflow. This is especially true for signal transduction-related research, where numerous investigators would benefit from systematic compilation, integration, and synthesis of thousands of disparate pieces of information scattered among individual research articles. For example, at the time that this paper was being revised, a search through the PubMed system using the keywords 'cell cycle' and 'apoptosis' produced lists of 169 293 and 29 961 articles, respectively (see http://www3.ncbi.nlm.nih.gov/Entrez/medline.html to access PubMed). It would be difficult and extremely time-consuming to navigate each of these articles manually. The biological community at large clearly needs specialized computer applications for the analysis of complex regulatory schemes; Such a program would facilitate the automatic retrieval of regulatory data from electronic versions of scientific publications using *natural language processing* (NLP) techniques (e.g. see literature on MedLEE, Friedman *et al.*, 1994), visualization and editing of representations of regulatory systems, and computer analysis and simulation of regulatory networks.

*To whom correspondence should be addressed.

© Oxford University Press 2000

Each of these tasks requires the development of a domain-specific *ontology* tailored to the specific task of analyzing complex regulatory pathways in a particular organism. In this article we introduce such an ontology and describe properties that distinguish it from the already existing ontologies.

Background

Ontology

An ontology, conceptualization or *domain model* for a specialized field of science is usually defined as a collection of concepts representing domain-specific entities, concept definitions, a set of relationships among concepts ('semantic network'), properties of each concept, the range of allowed values for each property, and, in some cases, a set of explicit axioms defined on these concepts (Gruber, 1993).

Regulatory pathways

Regulation of tissue and organ development, as well as cell cycling and differentiation, is governed by a complex series of interactions between proteins and genes, in which one protein can 'switch off' or 'switch on' another protein, which in turn may stop or start its action on other proteins or genes. When a regulatory protein A is known to *increase* expression of gene B, biologists often say that 'A upregulates gene B,' while they say that 'A downregulates gene B' if protein A causes a *decrease* in the expression of gene B. While the actual regulation process is continuous and gradual, it is convenient to represent a regulatory network as a network of logical switches that can be turned on and off by other switches within the same network. For example, a process illustrating the regulation of long-term potentiation in human neurons is shown in Figure 1. In this figure, glutamate (here defined as a small molecule) binds and activates metabotropic receptor (a protein). Metabotropic receptor, in conjunction with G-protein activates another protein, PLC/PIC, and so on. Biologists may describe this as 'glutamate acts *upstream* of PLC/PIC,' and that 'PLC/PIC acts *downstream* of metabotropic receptor.'

In graph theory, a regulatory pathway is usually represented as an oriented graph with vertices corresponding to substances and edges corresponding to interactions (*Actions* in our ontology). So long as any oriented graph is completely defined by two sets, a set of vertices and a set of edges, any complex pathway can be fully encoded with a list of substances (vertices) and a list of interactions (edges) between them.

It is common practice in biology to represent large segments of regulatory networks with non-uniquely defined 'fuzzy' names. For example, in the sentence '*activation of MKK3 kinase triggers cell death*,' '*cell death*' is a process including actions and substances that are not specified in the sentence, and '*MKK3 kinase*' is a substance (protein). Depending on the context, the term 'cell death' can represent one or a few different pathways leading to apoptosis, or it may correspond to just a fragment of the most downstream portion of a particular apoptotic pathway. Furthermore, the level of details showed in descriptions of the same regulatory pathway, usually considerably varies among journal papers. For example, the oriented graph shown in Figure 1 is commonly referred to in research articles as 'long-term potentiation' even though the number of proteins and genes included into a particular regulatory scheme can vary tenfold.

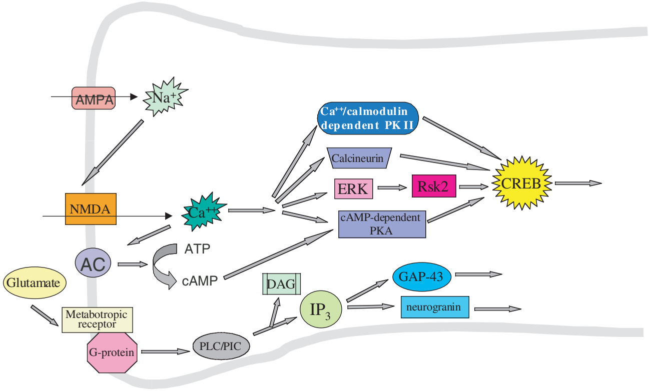

Fig. 1. Example of representation of a regulatory pathway (long-term potentiation in human neurons).

Natural language processing (NLP) techniques aimed at extracting information from electronic texts

In medicine, text reports of patient visits to their physicians are a vast source of clinical information, but such information in this 'raw' textual form is not useful for automated clinical applications (there may be numerous representations of the same piece of information) and although electronically available, this information remains locked within the text. Text is difficult to access because it is extremely diverse and meanings of words vary depending on their context. In spite of these underlying difficulties, NLP in the medical domain has begun to show promising results. For example, there are two NLP systems (Friedman *et al.*, 1994; Haug *et al.*, 1990; Hripcsak *et al.*, 1995) which are currently integrated into operational clinical information systems. Evaluating these systems clearly demonstrated that automated applications using NLP perform as well as or almost as well as medical experts in identifying abnormal conditions (Friedman *et al.*, 1994; Haug *et al.*, 1990; Hripcsak *et al.*, 1995).

Existing related systems

The molecular biology community possesses a number of online databases related to genomics research (Burks, 1999). GenBank (Benson *et al.*, 1998) contains nucleic acid sequence information, GDB (Fasman *et al.*, 1997) specializes in the location of genes on chromosomes, and PDB (Abola *et al.*, 1997) offers three-dimensional protein structures. Two databases targeted to the clinical community are OMIM (Pearson *et al.*, 1994), a database that maintains clinical phenotypes and their links to human genes, and Helix (Tarczy-Hornoch *et al.*, 1998), a database containing information about diseases, genetic testing, and laboratories in the United States that perform this testing.

Work in representing complex information has been done with a domain *independent* knowledge-based system called OWEB (Hon *et al.*, 1998), which was developed specifically for building and supporting shareable online scientific data resources. OWEB models complex information by using a knowledge base that contains meta-information about the data resources. This meta-information consists of a hierarchical taxonomy of concepts, specifications of relations between these concepts, as well as real-world objects, which are instances of concepts and relations. One application of this system resulted in the development of MHCWeb (Hon *et al.*, 1998), an online immunological knowledge base that contains information about peptide molecules and the set of major histocompatibility complex (MHC) molecules to which they bind, along with experimental and publication information.

Another complex modeling tool is the Oncology Thinking Cap (OncoTCAP) Ramakrishnan *et al.* (1998), which was developed to support learning and provide a means by which medical students, clinicians, and researchers can develop research and treatment strategies through simulation. OncoTCAP provides a simulation-modeling tool for expressing the complex concepts and explicit relationships associated with cancer research, treatment, apoptotic and mutational mechanisms, cell repair processes, treatment scheduling, and genetic characteristics.

Our research is similar to MHCWeb and OncoTCAP in that it models complex information and makes explicit relationships between informational concepts. Our work differs from the two systems in that our aim is to automatically acquire functional genomics knowledge through the extraction of relevant information from published articles using natural language processing methods.

For the construction of our ontology it was useful to include concepts from the UMLS (Unified Medical Language System of the National Library of Medicine in Washington, DC, McCray *et al.*, 1993), a comprehensive source of biomedical knowledge. The UMLS integrates biomedical terminology and organizes concepts into a hierarchy of classes. Some parts of UMLS, such as the Methathesaurus which incorporates classifications of diseases like ICD9, are immediately relevant to our ontology; other aspects of UMLS are either unrelated to our task or lack specifics essential for the computer representation and modeling of regulatory pathways. A detailed analysis of UMLS regarding the genomics domain is provided by Yu *et al.* (1999).

Other existing biological ontologies such as EcoCyc (Karp, 1991), Molecular Biology Ontology (Schulze-Kremer, 1997; Schulze-Kremer and King, 1992) and several other ontologies (Hafner *et al.*, 1994; Hafner and Fridman, 1996; Karp, 1998; Karp and Riley, 1993; Schulze-Kremer, 1998) are useful for our study as points of reference, but insufficient for our goal because they were developed for different applications.

EcoCyc (Karp, 1998) is a carefully designed ontology aimed at representing, modeling and visualizing *biochemical pathways in bacteria* (Karp *et al.*, 1999). Many features are relevant to our application: EcoCyc can both represent a range of complex biochemical reactions and allow for qualitative as well as quantitative modeling of each reaction. However, since this system targets only the representation of bacterial pathways, it does not reflect the multiplicity of cell types, tissues, organs, and developmental stages of multicellular organisms. Moreover, it deals mostly with linear or simple cyclic pathways typical of bacteria, rather than more complex graphs corresponding to regulatory networks in multicellular species. The design of EcoCyc was designed for input of data to be done by human experts and consequently does not consider issues arising when information is automatically extracted from literature using NLP techniques, such as conflicts between statements. Finally, EcoCyc is a proprietary commercially implemented ontology, not readily available in its complete form to the general public.

MBO, a Molecular Biology Ontology (Schulze-Kremer, 1997; Schulze-Kremer and King, 1992), is a *general* ontology for molecular biology which aims to collect ‘all relevant concepts that are required to describe biological objects, experimental procedures and computational aspects of molecular biology.’ As a general ontology, MBO contains large amounts of information that are important to their application whereas our model has a different focus. Our intention is to describe signal transduction as well as biochemical pathways in both complex and simple organisms through visualization and simulation.

Methods

In order to design the ontology, we manually collected more than 300 online journal articles from *Science, Nature, Proceedings of the National Academy of Sciences of the USA, Cell*, and *Current Biology*, which addressed regulation of ‘programmed cell death’ in animals. We then manually analyzed these articles with the aim of reconstructing a moderately complete regulatory network for this system. The selection of papers for this analysis was done through an iterative keyword search against an online reference database (MEDLINE, see http://www3.ncbi.nlm.nih.gov/Entrez/medline.html), followed by downloading relevant papers from Internet sites of the mentioned journals and manual review. The keywords used for the initial search were ‘apoptosis’ and the name of the publication, e.g. *Science*. We used the reconstructed network to iteratively and manually design a parsimonious representation capturing much of the information that a biologist might consider important when analyzing a regulatory network. While designing ontology we followed recommendations aimed at facilitating future ‘graceful evolution’ of the ontology (Cimino, 1998). Each iteration of the development process included a fine-tuning of the knowledge model followed by model verification through description of a part of the apoptotic network with the currently available tools of the ontology. Altogether four iterations of this kind were carried out.

Further, we downloaded twenty review articles from *Current Biology* and *Trends in Genetics*, all of which contained descriptions of relatively large regulatory pathways. We then manually analyzed descriptions of pathways in these articles attempting to use our knowledge model for the representation of information contained in each of these articles with the aim of verifying our ontology. Because the last verification required little fine-tuning of the model, we concluded that the model is sufficiently mature for current purposes. However, in the future it may be necessary to amend concepts and concept properties as the system moves toward novel computational problems and new aspects of the complex biological reality. Note that the current evaluation of the knowledge model is not a rigorous one because it was performed by the model developers rather than by a group of impartial experts.

Results and discussion

It is convenient to describe our ontology by considering its three main aspects: *taxonomy* of the concepts, *relations* between them, *properties* of the concepts, and axioms defined in the ontology; below we will follow this scheme.

Taxonomy of terms

The ‘IS-A’ relationships between concepts (relationships between a general concept and more concrete concepts representing instances of the general concept) are often represented as trees (taxonomies) in which each edge represents one binary relation of this kind and each node is occupied by a concept (see Figure 2). Each of the basic concepts, depicted in this figure serves as a root for a separate tree: *Action, ActionAgent, Process, Publication, Taxon, Disease, Mechanism, Result, Developmental Stage, MicroStructure, State, MacroStructure, Relation, Similarity, RelationType, and ActionTemplate*. The most important among these concepts are *ActionAgent* and *Action*. The former most frequently corresponds to a *Substance* (a protein, a gene or other molecule) and the latter to an *interaction* between two or more molecules in a regulatory

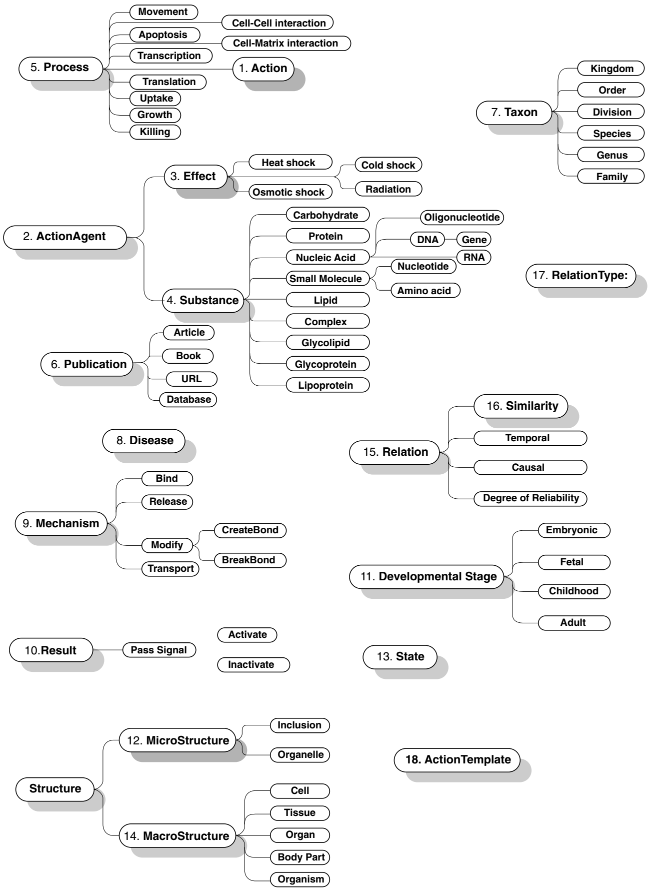

Fig. 2. Top-level taxonomy (‘IS-A’ relationships, indicated with continuous lines, and ‘is associated with’ relationships, indicated with dashed lines) of concepts. *Substances* and *Actions* are the central concepts in this ontology; *Taxon*, *Structure*, *Publication*, and *Disease* are required to specify the living organism, the structure within the organism, the source of the information and the malady associated with each individual *Action* and *Substance*. *Stimulus* and *Process* are categories associated with the external triggers of the regulatory cascades in a living organism, and the macroevents within an organism that involve multiple regulatory steps, correspondingly. Finally, *Relation* is a separate concept permitting the ordering of *Actions* and *Processes* in three different ways (cf. Figure 4).

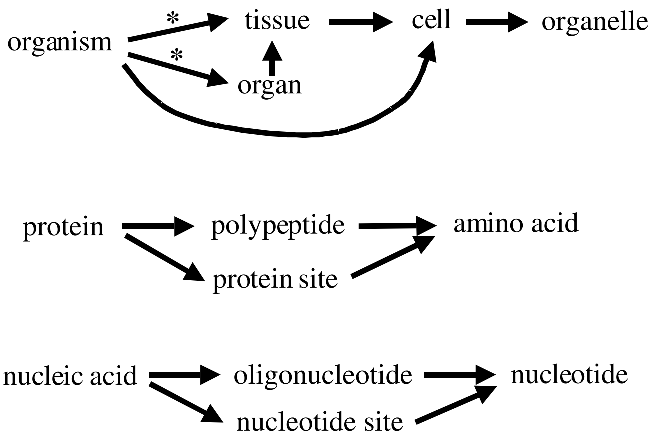

Fig. 3. Example of ‘PART-TO-WHOLE’ relationships (‘HAS-A’) between some of the categories shown in Figure 1. Each arrow corresponds to a single ‘HAS-A’ relationship between two concepts. Arrows marked with asterisks represent *optional* relations. For example, not every organism has organs and tissues because many organisms are unicellular. Note that, unlike the taxonomies of categories, these graphs are not trees.

network. For example, proteins IP3 and GAP-43 belong to the group *substances*, and activation of GAP-43 by IP3 is an *interaction* (see Figure 1). A less frequent type of *ActionAgent* is *Effect*, introduced here to reflect inputs or outcomes of an action that can not be correctly characterized within *Substances*. Among the most common examples of input *Effect* found in the biological literature are heat shock, cold shock, osmotic shock, radiation, electrical stimulation, tension, and starvation. Mechanisms of *Action* in our model are limited to their ‘normal’ repertoire in animal signal transduction systems. One can think of ‘abnormal’ actions that we are not including into our ontology at this stage.

Next in rank by importance are *Process* and *Relation*. In our ontology, *Process* represents a set of several *Substances*, at least some of which are linked by *Actions* or *Relations*. *Process* also represents a set of other *Processes* or a mixture of *Actions* and *Processes*. Finally, *Publication, Taxon, Structure, Developmental Stage, and Disease* encapsulate pieces of auxiliary information about *ActionAgents, Processes* and *Actions*. These concepts indicate the source of knowledge (e.g. a book), taxonomic position of the organism (e.g. *Homo sapiens*), anatomical structure (which can be a macro-structure, such as an organ or a tissue, or a micro-structure, such as a mitochondrion), a stage in ontogenesis (e.g. fetal), and a malady (e.g. cancer), respectively. The concept *Disease* corresponds to a large list of specific maladies and syndromes, which are taken from UMLS (data not shown).

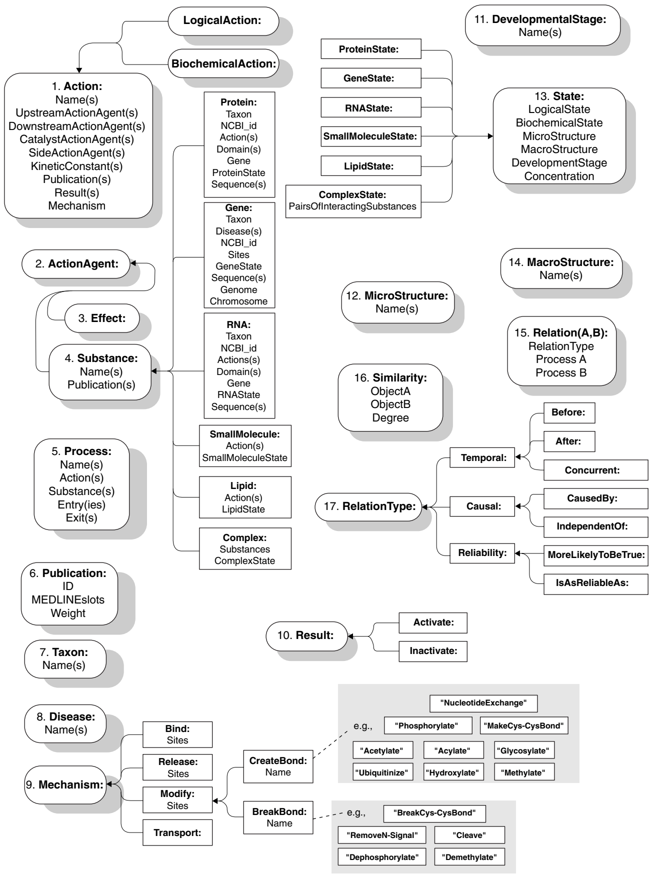

Fig. 4. Properties of major concepts. The frame *Publication* contains a slot called ‘MEDLINE slots’ referring to types of information (‘fields’) used in the MEDLINE reference database (http://www3.ncbi.nlm.nih.gov/Entrez/). These fields include: title of publication, names of the authors, year, journal volume, journal number, page numbers, abstract, keywords and some others. The frame *Actions* has a slot called *KineticConstants,* which refers to constants characterizing kinetics of the interaction between two or more molecules. The slots *MicroStructure* and *MacroStructure* in the frame *State* must refer to the location of the corresponding substance within the organism at several structural levels: organ, tissue, cell type, and organelle (see Figure 7).

Note, that although the current ‘IS-A’ graph is a strict taxonomy (tree), the future evolution of the knowledge model may require addition of edges representing relation of multiple inheritance between concepts. In this case the oriented graph will cease to be a tree but it will always remain acyclic.

Relations between concepts

To bring our ontology closer to the language used by biologists, we defined a network of ‘HAS-A’ relations (= ‘part-whole’ relations between concepts). A part of this network is shown in Figure 3. This network explicitly specifies ‘part–whole’ relations between concepts and the possibility that terms may have several different meanings, common in the language of experimental biology and medicine, where the same word, e.g. *cell*, may refer to distinct entities with very different properties (e.g. cells in Eukaryotes, Eubacteria and Archaea), and can correspond to either a part of an organism, or the whole organism.

Properties of concepts, slots and admissible values

The properties of the above concepts can be conveniently described using *frame* representation (Gruber, 1993).

In this representation, each concept is described with a 'frame' having a *name* (e.g. *Substance*) unique for the given domain and *slots* which can be filled by values of a specified type and range (see Figure 4). For example, following an expert's advise (Cimino, 1998) that a well-designed ontology needs to 'have the unique identifiers for the concepts which are free of hierarchical or other implicit meaning (i.e. nonsemantic concept identifiers),' we provided each concept with a slot 'ConceptID,' which should hold a unique numeric value for each concept. The slot *Name(s)* may contain a single string or a set of strings representing alternative names of the same substance, and the slot *Publication(s)* must refer to a concept object defined with the frame named *Publication*. Figure 4 depicts the current knowledge model, which utilizes these features and illustrates the concepts first introduced in Figure 2.

Similarities and differences with EcoCyc

We adopted a number of important features from the EcoCyc (Karp, 1998) ontology. Most importantly, the structure of the concept *Action* in our ontology is based on the concepts *reaction* and *enzymatic reaction* in EcoCyc. Following EcoCyc, we explicitly specify *sideActionAgents* and *mainActionAgents*, a *CatalystActionAgent*. Likewise, our concepts *UpstreamActionAgents and DownstreamActionAgents* are not unlike the EcoCyc concepts reaction left side and reaction right side, respectively. In the Biochemical Representation A phosphorylates B in Figure 6, The *sideActionAgent* is ATP, the *mainActionAgent* is B, and the *CatalystActionAgent* is A. Meanwhile, the logical representation of the same action, the *UpstreamActionAgent* is A and the *DownstreamActionAgent* is B.

Also analogous to EcoCyc's *isozyme-sequence-similarity* is our concept *Similarity*. Our definition describes both primary sequences and three-dimensional protein structures, where similarity at the sequence level is not required (see Murzin and Bateman, 1997).

The concept *ActionAgent* is somewhat parallel to the *compound* concept of EcoCyc. In addition, it includes *Effect*, which is not a substance. For example, in the statement 'ultraviolet light can induce expression of the p53 gene,' 'ultraviolet light' is an *effect* and p53 gene is a *substance* (gene) activated via an unspecified mechanism.

EcoCyc (Karp, 1998) is very well designed for representing metabolic pathways of bacteria, and therefore it includes details of cell morphology of prokaryotic species (e.g. *cytoplasm, membrane, inner-membrane, outer-membrane, membrane-spanning, periplasm*). Because our model seeks to describe signal transduction and biochemical pathways in eukaryotes as well as bacterial, it includes the additional concepts of tissues, body parts, organs, developmental stages, and cellular organelles that distinguish multicellular eukaryotes from prokaryotes (see Figure 4).

Signal transduction pathways are currently described as a mixture of two representations, logical and biochemical

Analyzing the language of current biological literature reveals a curious mixture of two different representations of regulatory pathways, which we will denote here as *logical* and *biochemical* (see Figure 6). The 'logical' representation handles changes in 'logical states' of the proteins and genes, while the 'biochemical' representation defines the chemical or physical mechanism leading to a change in the logical state of a substance. A single logical description 'A activates B' can correspond to a multiplicity of biochemical descriptions. Regulatory diagrams in present-day research articles often blend both logical and biochemical descriptions in the same figure. Moreover, situations where only the logical description of an action may be inferred from experimental data are common. The availability of only a biochemical action description is a dominant feature in the portrayal of *metabolic pathways*, such as synthesis of fatty acids; in contrast, actions with only well-defined logical descriptions are foremost in the portrayal of *signal transduction pathways*, such as cell cycle regulation. Similar to the described duality of *Action*, we define two properties of *State* for each substance: *LogicalStat*e and *BiochemicalState*. Without contextual information, a property cannot be directly inferred from the other. For example, a protein that is phosphorylated (*BiochemicalState*) can be in either active or inactive *LogicalState*.

In signal transduction literature it is possible to observe both logical and biochemical representations of an action combined within a single sentence of a research article. For example, in the following sentence taken from an actual research article (Boussiotis *et al.*, 1997)

'Activated Raf-1 phosphorylates and activates MEK-1...'

one can clearly distinguish the logical action between proteins Raf-1 and MEK-1 ('activate') and the biochemical mechanism of the action ('phosphorylate').

Because signal transduction pathways use a rather limited repertoire of biochemical reactions, such as *phospho-*

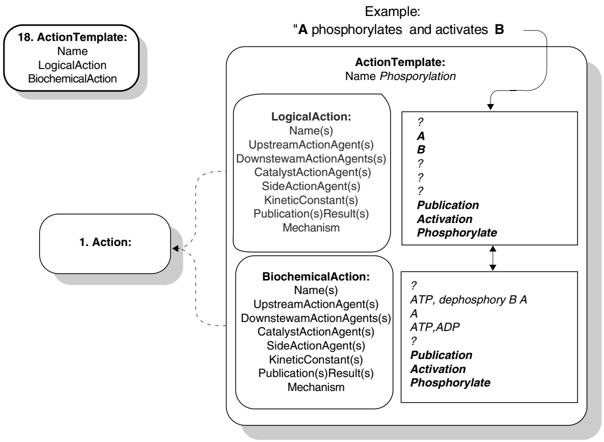

Fig. 5. An additional concept of *ActionTemplate* is designed to store rules of conversion of *LogicalAction* to corresponding *BiochemicalAction* and *vice versa.*

rylation, dephosphorylation, cleavage, etc. (see Figure 6), a logical description can often be converted to its corresponding biochemical description (or vice versa) automatically. For this purpose we introduce the *ActionTemplate* concept, which defines rules of converting of a logical action representation into the corresponding biochemical representation, or the other way around, assuming that the required information is present (see Figure 5).

Consensus-level conceptualization vs observation-level conceptualization

Most of the existing knowledge models, including EcoCyc (Karp, 1998), are oriented toward *consensus-level* conceptualization. This means that data inconsistencies are removed or resolved by domain experts before the data is represented with a knowledge model. This is impractical in the case of automatic extraction of data from original research articles: the ‘raw’ data obtained in this way are bound to contain inconsistencies and even mutually exclusive statements. Therefore, in our application we adopt an *observation-level* conceptualization rather than a consensus-level conceptualization. Each statement in our knowledge base will be given a ‘weight,’ a real number measuring the ‘credibility’ of the statement and computed as described below.

We reason that as controversies and occasional errors are unavoidable realities of a productive research process, Natural Language Processing analysis is destined to produce some contradictory and incorrect statements. Observation-level conceptualization would potentially allow the retention of original natural language passages containing primary information, thus enabling an individual researcher accessing our system to make independent decisions on the quality of different pieces of information. In order to determine the ‘credibility’ of a statement, we introduce a set of weights that take into account the reliability of an individual journal, the publication year, particular authors, and the section within an article (e.g. ‘Discussion’) in which particular statements appeared. (Similarly, if a statement in a research article is a citation of the previous work, it may be given lower weight than that of an original experimental observation.) This should provide a flexible mechanism for re-defining the regulatory model along with an accumulation of knowledge about previous errors. It would be desirable, for example, to exclude from analysis a statement that appeared in a retracted article, in an article with an erratum published later, or derived from an experiment that was not successfully reproduced. (A hypothetical example: one may decide to exclude all papers published before 1990 because of an unreliable, older experimental technique that was replaced in 1990.)

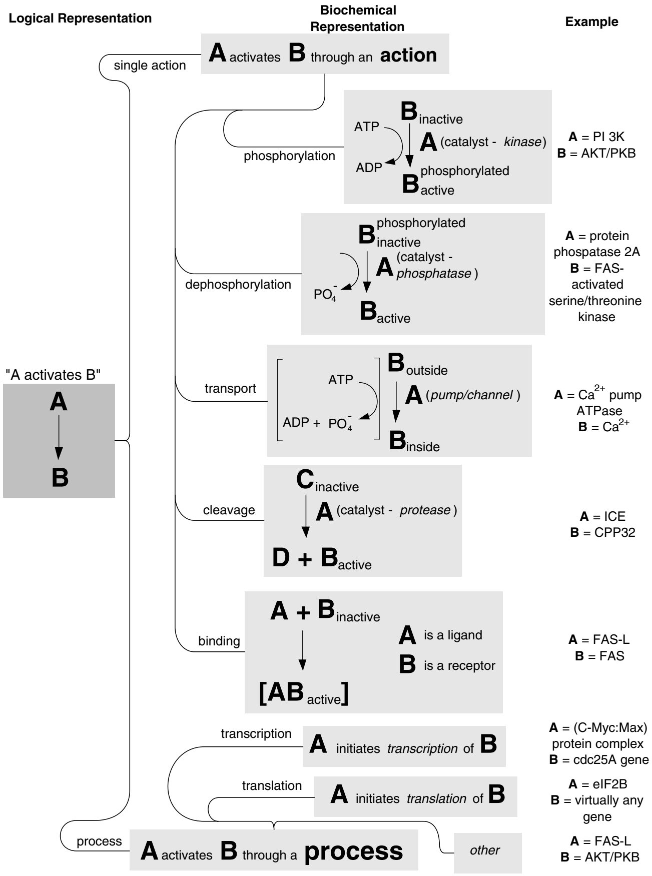

Fig. 6. Multiple biochemical representations corresponding to a single logical representation. A similar figure corresponds to the logical representation ‘A inhibits B.’

A weight for each binary action (statement) can be

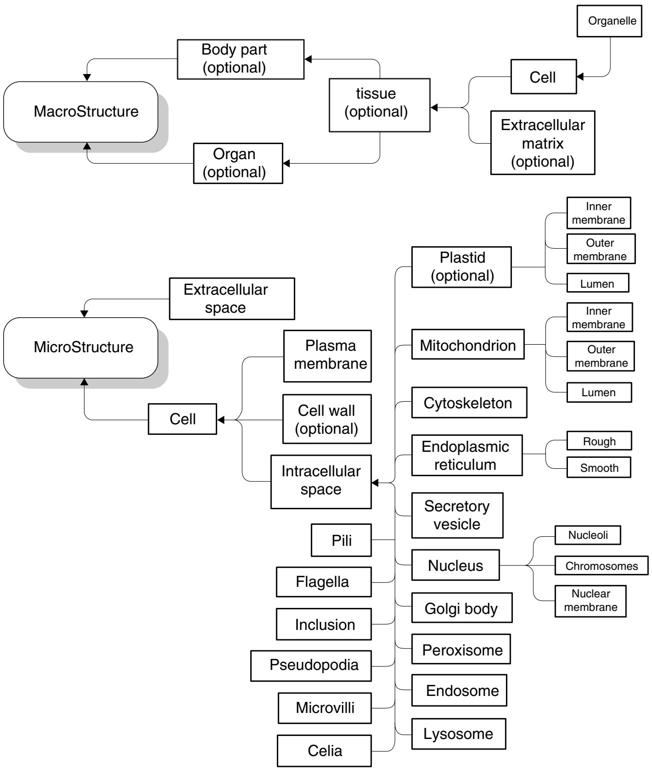

Fig. 7. Models corresponding to concepts *MicroStructure* and *MacroStructure.*

computed according to the following 'equations' where *W(x)* is defined as 'the weight of *x*.'

$$\boldsymbol{W(\text{a statement}) = \text{A weighted sum of } W(\text{all sentences related to this statement})}$$

$$\boldsymbol{W(\text{a sentence related to the statement}) = W(\text{journal, year, authors, article, etc.}).}$$

The sum of the first 'formula' is weighted in order to allow for downweighting multiple identical statements from the same paper and from different papers of the same authors. In the simplest scheme, the weight of a sentence would be defined as a product of the individual weights of journal, year, author, etc. Once all the weights are explicitly defined, it is trivial to compute a weight for each statement and for each statement's negation. For each set of conflicting statements, only the statement with the largest weight would be selected for visualization and simulation. The goal of this is to create a scenario in which a researcher can easily trace individual statements and corresponding sentences, remove erroneous data, and downweight equivocal statements. We intend to implement an 'iterative editing' regime, where a researcher can begin to visualize or simulate a pathway, then 'descend' to statements and sentences (each sentence highlighted within the text of the corresponding article or shown independently) related to a particular part of the pathway, redefine weights, and again 'ascend' to a modified visual representation and dynamic model of the same pathway, repeating this cycle as often as desired. This weight scheme may potentially accommodate fluctuations in the dominant scientific paradigm: different paradigms would correspond to different sets of weights for the same collection of statements.

Axioms

The area of molecular biology describing signal transduction incorporates a set of axioms that are implicit for biologists, but should be specified explicitly within a knowledge model. The fundamental axiom of molecular biology can be formulated in the following way; 'For every protein there is a unique mRNA, and for every mRNA there is a unique gene or a unique set of genes.' Among other important axioms are the following three:

1. If *ActionAgent* A is upstream to *ActionAgent* B in an action, and the action *Result* is defined, the *ActionAgent* A is active.

2. If *ActionAgent* A is situated downstream in an action and the action *Result* is *Activate*, the *ActionAgent* A is active.

3. If *ActionAgent* A is situated downstream in an action and the action result is *Inactivate*, the ActionAgent A is *inactive*.

A few other axioms used in logical and biochemical representation of actions are shown in Figure 8.

Conclusion

With the goal of developing computational tools for the analysis of signal transduction pathways, we have introduced ontology suitable for the description and modeling of regulatory pathways in multicellular and unicellular species. This ontology has a number of common features with EcoCyc and several other existing ontologies, but differs significantly from them in a number of features, most important of which is that it allows for modeling both Boolean ('logical') and biochemical pathways, multiple cells, tissues, and organ types, as well as representing partial and/or conflicting information.

Acknowledgements

This study was partially supported by grants from the Center for Advanced Technology, New York State, and from Hitachi Software Engineering Co. Ltd.

Logical result-related axioms

1. **Transitivity**. If A activates B and B activates C, it is valid to state that A activates C.

2. **Additivity.** A logical result of a pathway can be calculated as a sum of results of separate actions along the pathway.

Biochemical mechanism-related axioms (in the form **mechanism**[subject substance, object substance])

1. **Transcribe**[gene]

2. **Translate**[mRNA]

3. **Phosphorylate**[protein (kinase), protein]

4. **Dephosphorylate**[protein (phosphatase), protein]

5. **Methylate**[protein (methylase), gene OR protein OR lipid]

6. **Demethylate**[protein (demethylase), protein OR lipid]

7. **Cleave**[protein, gene OR protein OR lipid OR RNA OR small molecule OR carbohydrate]
Cleave[RNA, RNA]
Cleave[small molecule, protein OR gene OR lipid OR RNA OR small molecule OR carbohydrate]

8. **Bind**[protein, protein OR RNA OR gene]
Bind[RNA, RNA OR gene]
Bind[small molecule, protein OR gene OR lipid OR RNA OR small molecule]

9. **Transport**[protein, protein OR gene OR lipid OR RNA OR small molecule]
Transport[RNA, small molecule]

Fig. 8. Major axioms of the knowledge model described in this paper. A more complete set of axioms will be made available through the Internet in near future.

References

Abola,E.E., Sussman,J.L., Prilusky,J. and Manning,N.O. (1997) Protein Data Bank archives of three-dimensional macromolecular structures. *Meth. Enzymol.*, **277**, 556–571.

Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) GenBank. *Nucleic Acids Res.*, **26**, 1–7.

Boussiotis,V.A., Freeman,G.J., Berezovskaya,A., Barber,D.L. and Nadler,L.M. (1997) Maintenance of human T cell anergy: blocking of IL-2 gene transcription by activated Rap1. *Science*, **278**, 124–128.

Burks,C. (1999) Molecular biology database list. *Nucleic Acids Res.*, **27**, 1–9.

Cimino,J.J. (1998) Desiderata for controlled medical vocabularies in the twenty-first century. *Meth. Infect. Med.*, **37**, 394–403.

Fasman,K.H., Letovsky,S.I., Li,P., Cottingham,R.W. and Kingsbury,D.T. (1997) The GDB Human Genome Database Anno 1997. *Nucleic Acids Res.*, **25**, 72–81.

Friedman,C., Alderson,P.O., Austin,J.H., Cimino,J.J. and Johnson,S.B. (1994) A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.*, **1**, 161–174.

Gruber,T.R. (1993) *Towards Principles for the Design of Ontologies Used for Knowledge Sharing: Knowledge Systems Laboratory*. Stanford University.

Hafner,C.D., Baclawski,K., Futrelle,R.P., Fridman,N. and Sampath,S. (1994) Creating a knowledge base of biological research papers. *Ismb*, **2**, 147–155.

Hafner,C.D. and Fridman,N. (1996) Ontological foundations for biology knowledge models. *Ismb*, **4**, 78–87.

Haug,P.J., Ranum,D.L. and Frederick,P.R. (1990) Computerized extraction of coded findings from free-text radiologic reports. Work in progress. *Radiology*, **174**, 543–548.

Hon,L., Abernethy,N.F., Brusic,V., Chai,J. and Altman,R.B. (1998) MHCWeb: converting a WWW database into a knowledge-based collaborative environment. *Proc. AMIA Symp.*, 947–951.

Hripcsak,G., Friedman,C., Alderson,P.O., DuMouchel,W., Johnson,S.B. and Clayton,P.D. (1995) Unlocking clinical data from narrative reports: a study of natural language processing. *Ann. Int. Med.*, **122**, 681–688.

Karp,P.D. (1991) Artificial intelligence methods for theory representation and hypothesis formation. *Comput. Appl. Biosci.*, **7**, 301–308.

Karp,P.D. (1998) Metabolic databases. *Trends Biochem. Sci.*, **23**, 114–116.

Karp,P.D. and Riley,M. (1993) Representations of metabolic knowledge. *Ismb*, **1**, 207–215.

Karp,P.D., Riley,M., Paley,S.M., Pellegrini-Toole,A. and Krummenacker,M. (1999) Eco Cyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **27**, 55–58.

McCray,A.T., Aronson,A.R., Browne,A.C., Rindflesch,T.C., Razi,A. and Srinivasan,S. (1993) UMLS knowledge for biomedical language processing. *Bull. Med. Libr. Assoc.*, **81**, 184–194.

Murzin,A.G. and Bateman,A. (1997) Distant homology recognition using structural classification of proteins. *Proteins*, (Suppl 1), 105–112.

Pearson,P., Francomano,C., Foster,P., Bocchini,C., Li,P. and McKusick,V. (1994) The status of online Mendelian inheritance in man (OMIM) medio 1994. *Nucleic Acids Res.*, **22**, 3470–3473.

Schulze-Kremer,S. (1997) Adding semantics to genome databases: towards an ontology for molecular biology. *Ismb*, **5**, 272–275.

Schulze-Kremer,S. (1998) Ontologies for molecular biology. *Pacific Symp. Biocomput.*, 695–706.

Schulze-Kremer,S. and King,R.D. (1992) IPSA-inductive protein structure analysis. *Protein Eng*, **5**, 377–390.

Tarczy-Hornoch,P., Covington,M.L., Edwards,J., Shannon,P., Fuller,S. and Pagon,R.A. (1998) Creation and maintenance of helix, a web based database of medical genetics laboratories, to serve the needs of the genetics community. *Proc. AMIA Symp.*, 341–345.

Yu,H., Friedman,C., Rzhetsky,A. and Kra,P. (1999) Representing genomic knowledge in the UMLS semantic network. *AMIA 1999 Fall Annual Symposium*.