An ontology for collaborative construction and analysis of cellular pathways

E. Demir[1,2], O. Babur[1,2], U. Dogrusoz[1,2,*], A. Gursoy[2], A. Ayaz[1,2], G. Gulesir[1,2], G. Nisanci[1,2] and R. Cetin-Atalay[1,3]

[1]Center for Bioinformatics, [2]Computer Engineering Department and [3]Department of Molecular Biology and Genetics, Bilkent University, Ankara 06533, Turkey

ABSTRACT

Motivation: As the scientific curiosity in genome studies shifts toward identification of functions of the genomes in large scale, data produced about cellular processes at molecular level has been accumulating with an accelerating rate. In this regard, it is essential to be able to store, integrate, access and analyze this data effectively with the help of software tools. Clearly this requires a strong ontology that is intuitive, comprehensive and uncomplicated.

Results: We define an ontology for an intuitive, comprehensive and uncomplicated representation of cellular events. The ontology presented here enables integration of fragmented or incomplete pathway information via collaboration, and supports manipulation of the stored data. In addition, it facilitates concurrent modifications to the data while maintaining its validity and consistency. Furthermore, novel structures for representation of multiple levels of abstraction for pathways and homologies is provided. Lastly, our ontology supports efficient querying of large amounts of data.

We have also developed a software tool named pathway analysis tool for integration and knowledge acquisition (PATIKA) providing an integrated, multi-user environment for visualizing and manipulating network of cellular events. PATIKA implements the basics of our ontology.

Availability: PATIKA version 1.0 beta is available upon request at http://www.patika.org

Contact: patika@cs.bilkent.edu.tr

INTRODUCTION

Human genome is expected to create an extremely complex network of information, composed of hundred thousands of different molecules and factors (Arnone and Davidson, 1997; Miklos and Rubin, 1996). Knowing the exact map of this network is very important since it will potentially explain the mechanisms of life processes as well as disease conditions. Such knowledge will also serve as a key for further

biomedical applications such as development of new drugs and diagnostic approaches. In this regard, a cell can be considered as an inherently complex multi-body system. In order to make useful deductions about such a system, one needs to consider cellular pathways as an interconnected network rather than separate linear signal routes.

Our knowledge about cellular processes is increasing at a rapidly growing pace. Novel large scale analysis methods are already applied to yeast to provide data about the yeast proteome (Ito *et al.*, 2001; Zhu *et al.*, 2001). However, most of the time these data are in fragmented and incomplete forms. One of the most important challenges in bioinformatics is to represent and integrate this type of knowledge. Efficient construction and use of such a knowledge base depends highly on a strong *ontology* (i.e. a structured semantic encoding of knowledge). This knowledge base then can act as a blueprint for simulations and other analysis methods, enabling us to understand and predict the behavior of a cell much better.

A conventional approach for representation of cellular pathways is based on pathway drawings composed of still images (BPC, 2001; BBID, 2001, http://bbid.grc.nia.nih.gov; BioCarta, 2001, http://www.biocarta.com; SPAD, 2001, http://www.grt.kyushu-u.ac.jp/spad/index.html). Although easy to create, such drawings are often not reusable and the ontologies used are far from being uniform and consistent, highly depending on implicit conventions rather than explicit, formal rules. Clearly, this approach does not support programmatic integration regardless of the underlying ontology. Another approach is the development of interaction databases, in an attempt to cope with rapidly emerging protein–protein and protein–DNA interaction data (Xenarios *et al.*, 2001; BRITE, 2001, http://www.genome.ad.jp/brite/; Bader *et al.*, 2001). However these approaches deal with intermolecular interactions, but not with cellular processes per se, lacking desired details of information.

It appears that our knowledge about metabolic pathways are much more detailed and structured. As a result, databases mainly focusing on the metabolic parts of an organism are more extensive compared to their signaling counterparts

*To whom correspondence should be addressed.

(Ogata *et al.*, 1999; Karp *et al.*, 2002c; WIT, 2001, http://wit.mcs.anl.gov; BRITE, 2001, http://www.genome.ad.jp/brite/). In all of these databases, the enzymes are classified according to the Enzyme Commission list of enzymes (EC numbers). Although these databases have strong ontology, they are centered around enzymatic activities; thus their scope is strictly limited to metabolic pathways.

In general, signaling pathway databases focus on the direction of signal flow, showing activation and inhibition relations among signaling molecules (Wingender and Chen, 2001; Takai-Igarashi and Kaminuma, 1999). In these systems one can follow the transduction of a signal. However, the mechanisms of regulation is often omitted in favor of simplicity, leading to ambiguities in the model, and hindering any possible functional computations. Considering a molecule to be only in active and inactive states is clearly an oversimplification since a molecule often times has more than one active state, each performing a different activity.

Karp *et al.* deserves a special attention due to their extensive work on an ontology for cellular processes (Karp, 2000; Karp *et al.*, 2002a). They define different types of molecules, each with its own class, and consider different states of a molecule as different actors. In addition, reactions are defined to be independent entities, and molecules are linked to the reactions by distinct relations, which they call slots. Each molecule may optionally be tagged with a cellular compartment. Their ontology also makes use of the 'pathway' concept to define summary abstractions, which may be used for defining data at varying levels of detail.

Collaborative construction of cellular pathways poses additional constraints on an ontology. Our current knowledge on signaling pathways is often incomplete due to the nature of data and the way it is collected. For instance, most experimental data captures indirect regulations among molecules, leaving out intermediate steps. An ontology should also be able to represent available information even when it is in incomplete form, so as to support incremental construction of pathways.

Another issue to consider in constructing a strong ontology for cellular pathways is dealing with concurrent modifications to the data. It is reasonable to assume that a user will view only a limited portion of the complex network of available cellular pathways at a time. Hence a modification to the existing data in this small window may affect the integrity of this entire network. On account of this, an ontology should also state the integrity rules of the pathway data, enabling us to construct a rigid model. Only with the help of such rules, automated integration of data into the existing knowledge base is possible.

Because of the aforementioned reasons, efforts for developing common, standard ontologies are gaining increasing support in the scientific community. There are efforts in multiple levels (Hucka *et al.*, 2003; Ashburner *et al.*, 2000). We believe that coercion between these different levels are important for the integration of biological data at different levels. For example, sequence, yeast two hybrid, microarray and metabolic simulation data have different perspective and level of detail, although they describe the same system. An ontology which could integrate and store data from such different sources and present them seamlessly in different perspectives, isolating a user from such heterogeneities, is critical to modeling of such a complex system.

Even though the ultimate goal in analysis of pathway data is support for functional computations and simulations on the model created, a simpler yet very effective form of analysis is possible through visualization. First of all, an effective visualization is only possible through an ontology that permits drawings of pathways with intuitive images (i.e. graphical user interfaces). Another necessary tool for effective visualization is automated layout, with which aesthetically pleasing, comprehensible drawings of pathways can be produced. It is also crucial to have proper complexity management tools for analysis of complex pathways. Such techniques are necessary at both the visualization level and at the level of knowledge base, which is free of geometrical information for pathways. Thus the ontology should suggest various ways to reduce the complexity of the information that the user deals with at one time. Another way of dealing with complexity is by supplying powerful querying mechanisms. Such mechanisms enable researchers to find their ways around in the 'jungle' of paths, again requiring a rigid ontology.

In what follows, we describe our ontology to model networks of cellular processes through integration of information on individual pathways. Our ontology is suitable for modeling incomplete information and abstractions of varying levels for complexity management. Furthermore, it facilitates concurrent modifications and extensions to existing data while maintaining its validity and consistency. Then we present a partial implementation of our ontology, and end up with a discussion and concluding remarks.

SYSTEM AND METHODS

An ontology for cellular processes

States and bioentities In every second, a cell makes hundreds of decisions, based on its internal status and its inputs. The underlying decision-making mechanism is a complex network of molecular level interactions.

Actors of this network are macromolecules (e.g. DNAs, RNAs and proteins), small molecules (e.g. ions, ATP and lipids), or physical events (e.g. heat, radiation and mechanical stress).

More than often these actors, especially macromolecules, have a common path of synthesis and/or are chemically very similar. For example, p53 protein has many states like its native, phosphorylated and MDM2-bound forms. In pathway drawings, it is common to represent these molecules as a

single biological entity. This is an oversimplification as different *states* can have very different and sometimes conflicting effects. It is therefore preferable to represent states individually while maintaining their biological or chemical groupings under a common *bioentity*.

Our ontology has similarities with (hierarchical) Petri-net approach of Choo (1982) in the way that Petri-net modeling has places (states) and transitions as nodes in the interaction graph and certain concepts such as abstractions may be defined recursively as will be discussed later on.

Transitions   A cell is not a static entity, neither are its actors. Molecules in a cell are synthesized, modified, transported and degraded constantly to respond to changes in the environment, or to accomplish a task. One can model such changes as quantitative chemical reactions. However this would reduce the coverage of the model, as currently both molecular concentrations and rate constants for most of these reactions are unknown. It is often preferred to represent these changes qualitatively since this better suits current experimental data.

A *transition* occurs only when all of its substrates are present and activation conditions are satisfied; therefore it is a function of the presence (or absence) of certain other actors. Under different conditions, different subsets of transitions may occur, leading to different cellular responses. A state may go through a certain transition, may be produced by a transition, or may effect a transition without getting changed. When a transition occurs, all of its products are generated.

Under certain circumstances, multiple transitions having the same state as a substrate may affect each other through depleting this common substrate. This happens when the equilibrium constant of a transition is relatively much higher than the others. If such a difference occurs among the equilibrium constants of transitions, we call the transition with the highest equilibrium constant *exhaustive* over other transitions for the common substrate. Transitions having the same order of equilibrium constant, on the other hand, are said to be *cooperative*.

Compartments   A significant number of transitions transport molecules between cellular compartments. The set of transitions that a state can participate in is strictly related to its compartment; thus a change in the compartment means a change in the state's information context. So we choose to incorporate the state's compartment in the model.

As the compartments and their adjacencies are cell type dependent, compartmental structure should be modeled as part of the ontology. Membranes pose an additional problem since not only a molecule may be located completely inside the membrane but also it may span one or both of its neighboring compartments.

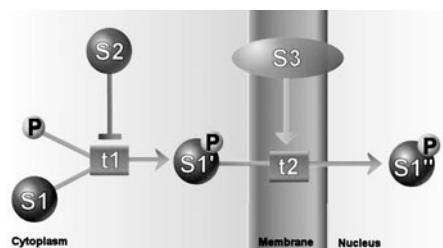Figure 1 illustrates the basics of our ontology with an example.

Fig. 1. An example illustrating the basics of our ontology, where states, transitions and interactions are represented with circles, rectangles and lines, respectively.
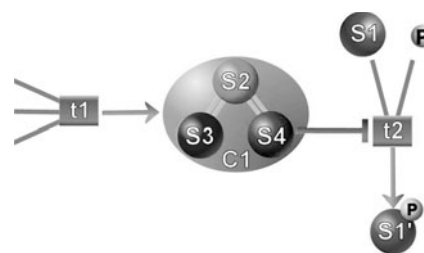

Fig. 2. A portion of a pathway containing a molecular complex of three states.

Molecular complexes   In biological systems molecules often form clusters for performing proper tasks, behaving like a single state. We consider each member of a *molecular complex* as a new state of its biological entity.

The function of a molecular complex is affected by the specific binding relations within itself. Therefore these binding relations must be represented in the model as well. Moreover, members of a molecular complex may independently participate in different transitions; thus one should be able to address each member individually (Fig. 2). In addition, a molecular complex may contain members from multiple neighboring compartments.

Abstractions   Network of molecular interactions derived from current biological data is incomplete and complicated. Complete network of cellular events is clearly beyond human perception. So different levels of *abstractions* are necessary to make effective analysis of cellular processes and dealing with complexity better.

Representing a cellular pathway as a single process or grouping related processes under a certain cellular mechanism would enhance the comprehensibility of the network of events (Fig. 3).

Since the data on cellular processes is not complete, different levels of information may be available for certain events. In cases where it is not identified which state among a set of states constitutes the substrate, product or effector of a transition, or where target transition of an effector is obscure, we may need to abstract these states (transitions) as a single state
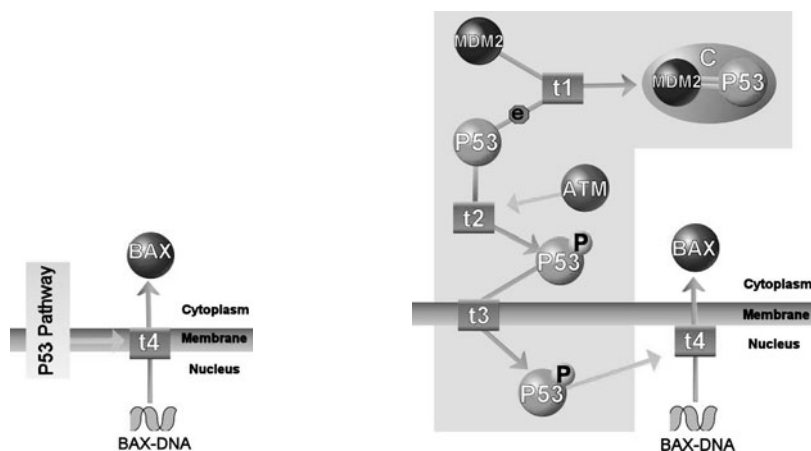
Fig. 3. Abstractions help better handling of complex information. For instance, part of a pathway graph may be 'collapsed' (left) to simplify a relatively more complex pathway graph (right).
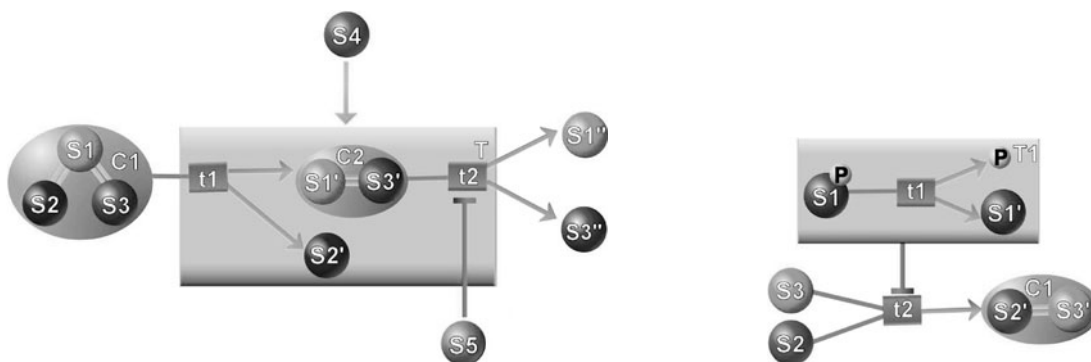

Fig. 4. Two types of abstractions for representing information of incomplete nature. *Transition abstraction:* it is unknown whether $S4$ activates $t1$ or $t2$ (left) and *state abstraction:* it is unknown whether $S1$ or $S1'$ inhibits $t2$ (right).

(transition) to represent the available information despite its incomplete nature (Fig. 4).

In biological systems, a gene is often duplicated throughout its evolution serving a different function. A special case occurs when this differentiation serves as a specialization of a generic mechanism. For example, when referring to the *wnt* gene, we actually mean 19 various similar genes in human (Miller, 2001). These genes are all activated by different stimulus at different tissues and can lead to different responses even though the signal processing mechanism is similar. Bhalla also describes common process motifs in signaling pathways, which are even more elementary operations that are reused through the entire network (Bhalla, 2002). Our ontology supports representation of such homologies using abstractions (Fig. 5).
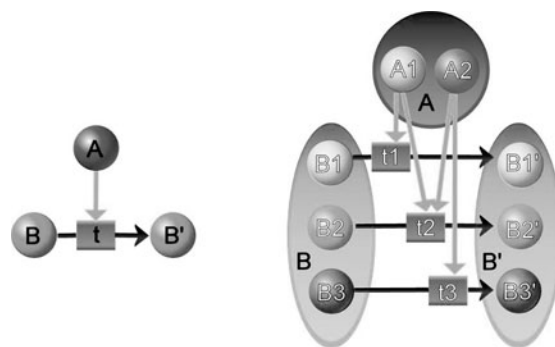
Fields, tissues, phases Contents of a complete network of pathways may be classified according to varying fields of studies such as apoptosis, lipid metabolism, cell cycle, etc. Similar classification may be performed based on tissue or phase specific processes. Looking at such an entire, complex

Fig. 5. In our ontology, homologies form another type of an abstraction.

network from the point of specific interest fields, tissues, or phases of cellular processes would simplify the understanding of the network by filtering out the undesired parts.

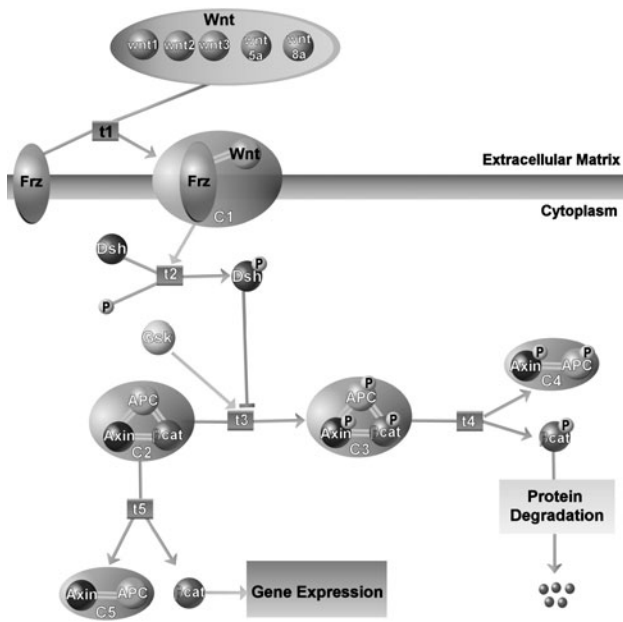Figure 6 shows a sample pathway described using our ontology.

Fig. 6. Canonical wnt pathway represented by our ontology: there are 19 *wnt* genes and eight frizzled genes identified (Dale, 1998), both represented as homology abstractions (drawn with black labels). Wnt homology abstraction also has an expanded view with 5 of the 19 *wnt* genes, which activate the wnt pathway. In addition, there are regular abstractions, 'protein degradation' and 'gene expression', represented as solid rectangles with black labels. Examples of complex molecule structure include APC:Axin:beta-Catenin and APC:Axin complexes.

A formal definition for the ontology

A pathway is an abstraction of a certain biological phenomena and is the uppermost abstraction in our ontology. Its context can change from a single molecule–molecule interaction to a complete set of all the interactions in a cell. In our ontology a pathway is represented by a *pathway graph*, which is a compound graph (Fukuda and Takagi, 2001). For the sake of simplicity, we will first describe a simple pathway graph and extend our definition to a more complete, complex compound graph.

A *simple pathway graph* is defined by an interaction graph $G = (V, E)$ along with a number of constraints on the topology as discussed below. V is the union of a finite set of *states* V_s and a finite set of *transitions* V_t. E is a union of *interactions* of four sets: *substrate* edges E_s, *product* edges E_p, *activator* edges E_a and *inhibitor* edges E_i, each directed edge belonging to either $V_t \times V_s$ (for product edges) or to $V_s \times V_t$ (for remaining interaction edge types).

Every state has a *type*: DNA, RNA, protein, small molecule or physical factor. It is also associated with a specific *compartment*. Chemically identical molecules in different compartments are considered as separate states. States of the same biological origin and/or similar chemical structure are grouped under a *biological entity* or simply *bioentity*.
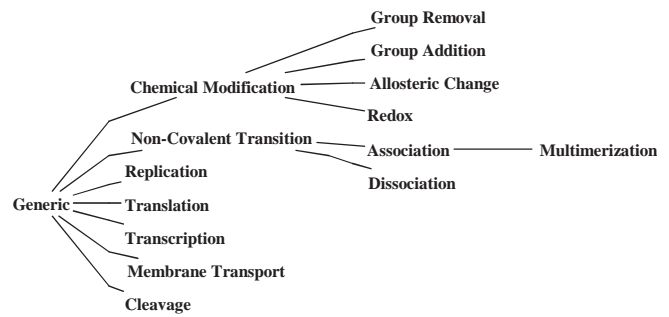

Fig. 7. Tree structure used to classify transitions.

Every transition must be incident with at least one substrate and one product edge. It may have an arbitrary number of effectors, a combination of which define the exact activation condition for the transition. Transitions are classified as a tree, according to the chemical nature of the transition (Fig. 7). A transition is not associated with a compartment; instead, its compartment is implied by its adjacent (interacting) states. A substrate edge can be labeled as *exhaustive* indicating the exhaustive effect of the associated transition for the incident substrate.

Every pathway graph has an associated cell model, which defines compartments, sub-cellular locations (e.g. axon) and their adjacencies. Cell models do not necessarily represent a single cell type. For users who want to model and analyze at a more generic level, a generic model comprised of compartments common to all cells of that organism may be used.

A more comprehensive ontology addressing molecular complexes as well as various types of abstractions can be defined with the notion of a compound graph. A *compound pathway graph* $CG = (G, I)$ is a 2-tuple of a *pathway graph* G and a directed acyclic *inclusion graph* I where:

- $V(G) = V$ is the union of states V_s, transitions V_t, molecular complexes V_c, and abstractions of four distinct types: regular, incomplete state, incomplete transition and homology, respectively denoted by V_a^r, V_a^s, V_a^t and V_a^h.

- $E(G)$ is the union of directed interaction edges of four distinct types: substrate, product, activation and inhibition, respectively denoted by E_s, E_p, E_a and E_i, and undirected bind edges E_b, used to form molecular complexes such that $E_p \rightarrow V_t \times V_s$; $E_s, E_a, E_i \rightarrow V_s \times V_t$; and $E_b \rightarrow [V_s]^2$.

- $V(I) = V(G)$.

- $E(I)$ is the union of inclusion edges E_i^c for defining molecular complexes and E_i^r, E_i^s, E_i^t and E_i^h for various types of abstractions such that $E_i^c \rightarrow V_c \times V_s$; $E_i^r \rightarrow V_a^r \times V$; $E_i^s \rightarrow V_a^s \times V$; $E_i^t \rightarrow V_a^t \times V$; and $E_i^h \rightarrow V_a^h \times V_s$.
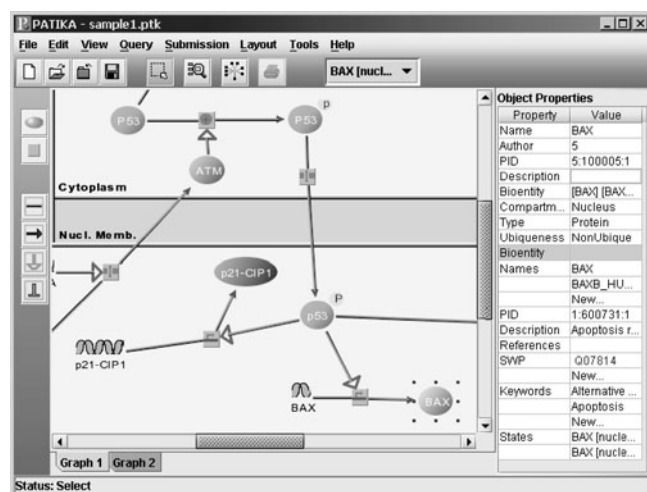
Fig. 8. A snapshot look at a signaling pathway in the PATIKA client.

In order for a compound pathway graph $CG = (G, I)$ to comply with our ontology, it needs to satisfy certain invariants as defined below:

- Molecular complexes cannot be nested; thus any directed path in I can contain at most one edge in E_i^c. A state can be incident to a bind edge in E_b, only if it has an incoming complex edge in E_i^c. Complexes are not allowed to overlap, a state can have at most one incoming complex edge. A complex state has no associated bioentity, although its children in I have their own bioentities.

- Regular abstractions represent pure grouping; thus they are not allowed to have incident edges in $E(G)$. However, they may be nested for representing multiple levels of detail.

- Homology abstractions are not allowed to be nested; therefore, any directed path in I can contain only one homology abstraction edge.

- A vertex in V is allowed to have any number of incoming abstraction edges in $E(I)$ since abstractions may overlap. Two overlapping abstractions do not necessarily define two vertex sets where one is a proper subset of the other.

IMPLEMENTATION

The basics of our ontology has been implemented within a software tool named PATIKA (Demir *et al.*, 2002).

Different types of molecules (e.g. protein, DNA and RNA) have distinct user interfaces for easier visual discrimination in PATIKA. Compartmental information is also modeled (Fig. 8).

In addition, advanced, graph theoretic querying facilities on the existing knowledge base is facilitated in PATIKA. The results are presented as a PATIKA (pathway) graph.

PATIKA also implements collaborative construction and concurrent modification issues addressed by our ontology.

PATIKA maintains version numbers as part of the ID of each graph object. Thus it is possible that while a user is working on a PATIKA graph locally, others might change the topology and/or properties of states and transitions in the PATIKA database. In that case, some of the local graph objects will have version numbers smaller than the ones in the database making the user's local PATIKA graph (partially) out-of-date. In other words, the user will have an out-dated view of the PATIKA database.

Whether a PATIKA graph is up-to-date can be checked by the client. As a result the user's graph objects are colored to indicate their status. For instance, green means this graph object exists in the database but its properties are locally modified.

If a user has any out-of-date graph objects, they may update their view of the database. For all graph objects that are out-of-date, the system will perform a check to see if the local copy can be updated automatically. If not, the user is asked to resolve any conflicts. When the user completes the update process, each object in the current PATIKA graph is brought up-to-date, and should have the same version as the ones in the database.

There are certain invariants that each PATIKA graph must satisfy, so that it is sensible from a biological point of view. For instance, each transition must have at least one substrate and at least one product. Once your graph is up-to-date and satisfies invariants imposed by the underlying ontology for validity, you may submit it to be integrated into the PATIKA database. Notice that validity of your data locally does not guarantee that its integration to the database will not create invalid situations in the 'big picture'. Upon submission, PATIKA checks for such global inconsistencies and notifies the user of any such integration problems.

DISCUSSION

Several aspects have been especially kept in mind when designing our ontology. *Coverage* refers to the amount of data an ontology is able to model, compared to the entire biological knowledge corpus. *Content* describes an unambiguous and regular structure in the information to be modeled. Finally, *clarity* refers to the intuitiveness and comprehensibility of the model itself. These principles often conflict with each other, and a compromise must be made, considering the nature of the data at hand.

One conflict arises due to the heterogeneous nature of biological knowledge. There are fields, as in metabolic pathways, where our understanding is deeper, with a nearly complete map of reactions, their reaction constants and even typical concentrations. On the other hand, data on most signaling pathways are still vague at best, with indirect relations, ambiguous mechanisms and unknown reaction constants. A detailed model would dismiss a lot of signaling data, where a lax model would poorly model metabolic pathways. Ability to represent multiple levels of detail gets more important,

when we consider collaborative construction, as desired modeling detail level of one user can be drastically different from another. A user may not be able to integrate their knowledge if the existing level of detail in the database does not match theirs. We address this problem by allowing multiple levels of detail. A user can represent a metabolic pathway in a very detailed form, and can include an abstract level signaling pathway regulation in the same graph using incomplete abstractions, even though the exact knowledge of mechanism is unknown at the metabolic level.

Another important tradeoff is between clarity and content. A more vigorous model, for most of the time, means a more complex representation, which in turn leads to models cluttered with states and interactions that are possibly of no interest to certain users. It is therefore desirable to manage complexity, such that the part of the model that a user currently focuses on is represented in full detail, where other portions are hidden or represented at a more abstract level. Our regular and homology abstractions are an attempt to reduce complexity through capturing groupings and similarities, and hiding their details when desired. Molecular complexes provide yet another way to hide unnecessary details.

Specification of inhibitors and activators of a transition does not necessarily establish an exact activation condition. Similarly, exhaustive substrate edges are an oversimplification of depletion of a substrate. This is a choice made to increase coverage since vigorously modeling activation condition and substrate concentrations require a linear (and possibly stochastic) set of equations (Tomita *et al.*, 1999; Schaff and Loew, 1999; McAdams and Arkin, 1997; Regev *et al.*, 2001), which are unknown for most signaling pathways. Our primary aim is to build a framework, albeit not precise, with the available biological data. However it would still be possible to add simulation support, at the software level by using a pluggable interface to a simulation engine. Our ontology would then serve to intuitively represent and investigate a model, where the simulation engine would be used for functional computations.

Our transition tree is far from being complete even though we believe it provides a fair amount of coverage without disturbing clarity. It can be expanded both vertically and horizontally. However, a more elaborate classification could be harder to represent, at least visually.

Up to now we have implicitly assumed that our model is built for a single organism. We believe that representing multiple organisms would overcomplicate the model and is not necessary since for most purposes, cellular networks of different organisms do not interact. Still a hybrid database such as Meta-Cyc (Karp *et al.*, 2002b), with the ability to encompass more than one organism, would be useful for experimental studies where two molecules from different organisms are allowed to interact (e.g. yeast two hybrid) or for modeling homologies between organisms. As our ontology distinguishes states of different bioentities and provides facilities for representing

homologies, it can be readily extended to encompass a hybrid model.

Finally, bioentities such as small molecules and macromolecules are not always grouped using the same criteria. For example, based on their path of synthesis, gene, mRNA and protein of p53 are all associated with the same bioentity even though they are chemically very different. On the other hand, cytosolic and extracellular Ca^{++} ions are associated with the same bioentity purely based on their chemical structure. Such choices are for practical reasons since one of the main use of the bioentity concept is linking a molecule to external databases such as GeneBank, SWISS-PROT and Ligand (Karsch-Mizrachi *et al.*, 2000; Bairoch and Apweiler, 2000; Goto *et al.*, 2002).

CONCLUSION

We have described an ontology for collaborative construction and analysis of cellular pathways. Based on this ontology, we have also developed a software tool named PATIKA providing an integrated, multi-user environment for visualizing and manipulating network of cellular events. PATIKA promises quite important benefits for many research fields in life sciences, including but not limited to, rapid knowledge acquisition, microarray data analysis and drug development.

The ultimate goal is to build a model for a cell as a whole with mechanistic details and to be able to perform functional computations and simulations over this model. Although tools such as PATIKA are far from fulfilling such an expectation, their concepts and ontology may serve helpful for future efforts in this direction.

REFERENCES

Arnone,M. and Davidson,E. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, **124**, 1851–1864.

Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinski,K., Dwight,S., Eppig,J., *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bader,G., Donaldson,I., Wolting,C., Ouellette,B., Pawson,T. and Hogue,C. (2001) BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **29**, 242–245.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucleic Acids Res.*, **28**, 45–48.

BBID (2001) Biological biochemical image database.

Bhalla,U. (2002) The chemical organization of signaling interactions. *Bioinformatics*, **18**, 855–863.

BioCarta (2001) Charting pathways of life.

BPC (2001) Biochemical pathways chart.

BRITE (2001) Biomolecular Relations in Information Transmission and Expression.

Choo,Y. (1982) Hierarchical nets: a structured petri net approach to concurrency. Technical Report 5044-tr-82 Computer Science California Institute of Technology.

Dale,T. (1998) Signal transduction by the Wnt family of ligands. *Biochem. J.*, **329**, 209–223.

Demir,E., Babur,O., Dogrusoz,U., Gursoy,A., Nisanci,G., Cetin-Atalay,R. and Ozturk,M. (2002) PATIKA: An integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics*, **18**, 996–1003.

Fukuda,K. and Takagi,T. (2001) Knowledge representation of signal transduction pathways. *Bioinformatics*, **17**, 829–837.

Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.

Hucka,M., Finney,A., Sauro,H., Bolouri,H., Doyle,J., Kitano,H., Arkin,A., Bornstein,B., Bray,D., Cornish-Bowden,A. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Karp,P. (2000) An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**, 269–285.

Karp,P., Paley,S. and Romero,P. (2002a) The pathway tools software. *Bioinformatics*, **18** (Suppl. 1), S225–232.

Karp,P., Riley,M., Paley,S. and Pellegrini-Toole,A. (2002b) The MetaCyc database. *Nucleic Acids Res.*, **30**, 59–61.

Karp,P., Riley,M., Saier,M., Paulsen,I., Collado-Vides,J., Paley,S., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002*c*) The EcoCyc database. *Nucleic Acids Res.*, **30**, 56–58.

Karsch-Mizrachi,I., Lipman,D., Ostell,J., Rapp,B. and Wheeler,D. (2000) GeneBank. *Nucleic Acids Res.*, **28**, 15–18.

McAdams,H.H. and Arkin,A. (1997) Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA*, **94**, 814–819.

Miklos,G. and Rubin,G. (1996) The role of the genome project in determining gene function: insights from model organisms. *Cell*, **86**, 521–529.

Miller,J. (2001) The Wnts. *Genome Biol.*, **3**, reviews 3001.1–3001.15.

Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleid Acids Res.*, **27**, 29–34.

Regev,A., Silverman,W. and Shapiro,E. (2001) Representation and simulation of biochemical processes using the pi-calculus process algebra. In *Pacific Symposium on Biocomputing* Vol. 6, pp. 459–470.

Schaff,J.C. and Loew,L.M. (1999) The virtual cell. In Altman,R.B., Dunker,A.K., Hunter,L. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing*, World Scientific Press, Singapore, Vol. 4, pp. 228–239.

SPAD (2001). Signaling PAthway Database.

Takai-Igarashi,T. and Kaminuma,T. (1999) A pathway finding system for the cell signaling networks database. *In Silico Biol.*, **1**, 129–146.

Tomita,M., Hashimoto,K., Takahashi,K., Shimizu,T.S., Matsuzaki,Y., Miyoshi,F., Saito,K., Tanida,S., Yugi,K., Venter,J.C. and Hutchison,III,C.A. (1999) ECELL: software environment for wholecell simulation. *Bioinformatics*, **15**, 72–84.

Wingender,E. and Chen,X. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.

WIT (2001). What Is There? Interactive Metabolic Reconstruction on the Web.

Xenarios,I., Fernandez,E., Salwinski,L., Duan,X., Thompson,M., Marcotte,E. and Eisenberg,D. (2001) Dip: the database of interacting proteins: 2001 update. *Nucleic Acids Res.*, **29**, 239–241.

Zhu,H., Bilgin,M., Bangham,R., Hall,D., Casamayor,A., Bertone,P., Lan,N., Jansen,R., Bidlingmaier,S., Houfek,T. *et al.* (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.