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 (PATAKA) providing an integrated, multi-user environment for visualizing and manipulating network of cellular events. PATAKA implements the basics of our ontology.

Availability: PATAKA 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.


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.
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.
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).

Two types of abstractions for representing information of incomplete nature. Transition abstraction: it is unknown whether $S_4$ activates $r_1$ or $r_2$ (left) and state abstraction: it is unknown whether $S_1$ or $S_1'$ inhibits $r_2$ (right).

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 stimuli 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).

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 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.
**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 \( V_s \times V_s \) (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. 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.

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 \( V_s^r \), \( V_s^i \), \( V_t^i \) and \( V_c^h \).

- \( E(G) \) 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 \( E_p \rightarrow V_t^i \times V_t; E_s, E_a, E_i \rightarrow V_s^i \times V_s^i \) and \( E_b \rightarrow [V_s^i]^2 \).

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

- \( E(I) \) is the union of inclusion edges \( E^c \) for defining molecular complexes and \( E^1, E^2, E^3 \) and \( E^h \) for various types of abstractions such that \( E^1 \rightarrow V_c \times V_c, E^2 \rightarrow V_a \times V_t, E^3 \rightarrow V_a \times V_a \times V_t; E^1 \rightarrow V_a \times V_t, E^2 \rightarrow V_1 \times V_t, E^3 \rightarrow V_1 \times V_1 \times V_t; E^h \rightarrow V_c^h \times V_c \).

**Fig. 7.** Tree structure used to classify transitions.
As defined below, it needs to satisfy certain invariants to represent multiple levels of detail.

- Molecular complexes cannot be nested; thus any directed path in \( I \) can contain at most one edge in \( E^1_i \). A state can be incident to a bind edge in \( E_b \), only if it has an incoming complex edge in \( E^1_i \). 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 MetaCyc (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 GenBank, 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


SPAD (2001). Signaling Pathway Database.


