Multi-layered Representation for Cell Signaling Pathways*

Eunok Paek‡§, Jisook Park¶, and Kong-Joo Lee‖§

To understand complex signaling pathways and networks, it is necessary to develop a formal and structured representation of the available information in a format suitable for analysis by software tools. Due to the complexity and incompleteness of the current biological knowledge about cell signaling, such a device must be able to represent cellular pathways at differing levels of details, one level of information abstract enough to convey an essential signaling flow while hiding its details and another level of information detailed enough to explain the underlying mechanisms that account for the signaling flow described at a more abstract level. We have defined a formal ontology for cell-signaling events that allows us to describe these cellular pathways at various levels of abstraction. Using this formal representation, ROSPath (reactive oxygen species-mediated signaling pathway) database system has been implemented and made available on the web (rospath.ewha.ac.kr). ROSPath is a database system for reactive oxygen species (ROS)-mediated cell signaling pathways and signaling processes in molecular detail, which facilitates a comprehensive understanding of the regulatory mechanisms in signaling pathways. ROSPath includes growth factor-, stress-, and cytokine-induced signaling pathways containing about 500 unique proteins (mostly mammalian) and their related protein states, protein complexes, protein complex states, signaling interactions, signaling steps, and pathways. It is a web-based structured repository of information on the signaling pathways of interest and provides a means for managing data produced by large-scale and high-throughput techniques such as proteomics. Also, software tools are provided for querying, displaying, and analyzing pathways, thus furnishing an integrated web environment for visualizing and manipulating ROS-mediated cell-signaling events. ***Molecular & Cellular Proteomics 3:1009–1022, 2004.***

To understand complicated biological processes and disease states at a molecular level, a systematic approach is necessary to illustrate signaling pathways. Each signaling pathway in response to various external stimuli, can be regulated by changes of proteins and chemicals, and the network assembling of various signaling pathways represents physiological phenomena (1). Recent advances in large-scale and high-throughput techniques including functional genomics, proteomics, RNAi technology, and genomic-scale yeast two-hybrid assay provide a tremendous amount of information on signaling pathways. To extract the biological significance from such massively produced results, it is necessary to develop an integrated environment for a formal and structured organization of the available information, in a format suitable for analysis with bioinformatics tools.

To present a signaling pathway, a database must include the following concepts first: 1) which molecules are involved in signaling in response to each external input, 2) which direction the signal is conveyed, that is, from a certain molecule to another, and 3) how the activities and subcellular localizations of molecules are changed by protein modifications, and/or protein-protein interaction changes. Based on the first database including such information, we will further expand the database to understand what the signaling results in (proliferation, differentiation, apoptosis, etc.) and how a network can be composed of various signaling pathways in response to multiple external inputs.

It is necessary to consider two distinct characteristics of information to be stored in a database for cell signaling pathways. One is its diversity in data types. Signaling entities range from small molecules and proteins to protein states and protein complexes (2). It must be noted that these entities are not independent of one another. For instance, protein complexes are composed of proteins, and a protein state can be defined by a protein binding to small molecule. The other, more significant characteristic is its incompleteness. It is not surprising to find a lot of gaps in the current knowledge about any signaling pathway. In order to organize such diverse and incomplete information into a structured and coherent database, we find it indispensable to have a formal model. Our formal model makes it explicit that multiple levels of abstraction are necessary for signaling events as well as for signaling entities. Differing levels of abstraction are inter-related so that an essentially-the-same signaling event can be described at multiple levels of details.

As a model system that implements the formal model pro-

From the ‡Department of Mechanical and Information Engineering, University of Seoul, Seoul 130-743, Republic of Korea; ¶Division of Information and Communication Engineering, Seoul Women's University, Seoul 139-774, Republic of Korea; and ‖Center for Cell Signaling Research, Division of Molecular Life Sciences & College of Pharmacy, Ewha Womans University, Seoul 120-750, Republic of Korea

Received March 15, 2004, and in revised form July 2, 2004

Published, MCP Papers in Press, August 2, 2004, DOI 10.1074/mcp.M400039-MCP200

posed in this study, reactive oxygen species (ROS)[1]-mediated signaling has been selected. Of many known signaling pathways, involvement of ROS are intensively reported (3, 4). The transient production of ROS, including hydrogen peroxide (H_2O_2), is an important signaling event triggered by the activation of various cell-surface receptors that include growth factors (platelet-derived growth factor (PDGF), epidermal growth factor, and vascular endothelial growth factor) (5–10), cytokines (transforming growth factor-β1, tumor necrosis factor-α, and CD40) (11, 12), agonists of heterotrimeric GTP-binding protein-coupled receptors (13), and environmental stresses such as heat shock (14, 15), heavy metal, and oxidative stresses. ROS-mediated signaling pathways can be divided into three components: the ROS production system, the oxidized targets of produced ROS, and the recovery system of reduction machinery (13). The signaling pathways are turned on in different cell lines and tissues by different external inputs showing different kinetics, bringing about different results of inducing various physiological phenomena including proliferation, differentiation, apoptosis, disease, etc., indicating that these pathways include both common and distinct events (14). We believe that a better understanding of complex interactions among various participants of these pathways will be facilitated if all the available massive data on differential protein expression, protein modifications, and protein-protein interactions generated from newly developed high-throughput technologies are collated.

In the following, we first explain why a multi-layered representation is necessary for cell signaling pathways, and then present a formal ontology that supports multiple abstraction levels. Then, the ROSPath web environment is introduced by showing various ways to query the database. Especially, the ROSPath visualization software tool is provided to help researchers analyze a pathway, predict new regulatory key signaling steps, and identify key participants at a molecular level in major ROS-mediated signaling pathways. Finally, the ROSPath database and data model is compared with other pathway databases and other signaling pathway modeling efforts.

EXPERIMENTAL PROCEDURES

ROSPath data model has been fully implemented as a web-based integrated software environment and is publicly available at http://rospath.ewha.ac.kr. All the pages are manually curated by experts in cell signaling. For each of proteins and protein complexes, its name, synonyms, species, subcellular localization, tissue information, enzymatic reaction, sequence information, chromosome localization, crystal structure, domain and motif information, cellular activities in terms of Gene Ontology ID, and references to the primary literature are included. Whenever possible, such information is automatically obtained from external sources such as GenBank, Swiss-Prot, Protein Data Bank, and Gene Ontology and updated weekly.

Implementation—The ROSPath database service is made available using ORACLE 9i™ (9.2.0) and MySQL (3.23.53). A simplified form of Entity-Relationship diagram for the ROSPath database is shown in Fig. 1.

RESULTS

Structure and Content—An important challenge in constructing signaling pathway databases is to collate and integrate fragmented and incomplete pieces of evolving information, in a principled manner. ROSPath integrates and explains available information on molecules and interactions involving signaling pathways via combining the newly produced results while making links to other existing databases available.

The ROSPath database is designed to describe signaling events at differing levels of abstraction, thus enabling identification of gaps in the present knowledge. In the past, most signaling pathway databases have been focusing on directions of signal flow, often omitting detailed mechanisms of regulation, either because of lack of critical knowledge or for the sake of simplicity. Recently, more emphasis is put upon detailing how molecules coordinate with one another during cellular responses. The ROSPath data model is along the same lines as such recent change of focus and is designed to convey detailed biochemical changes that form the basis of signal flow, as well as logical information flows represented at a higher level of abstraction.

In this study, we use the PDGF receptor signaling pathway as a model system to explain the schematic representation and formal ontology of ROSPath. Fig. 2*A* is a conventional presentation of a signaling pathway depicted in a previous report by Rhee *et al.* (3). It summarizes a variety of experimental results in a compact visual representation, with hydrogen peroxide serving as a key messenger in a signaling process triggered by the activation of cell-surface PDGF receptor. Oftentimes, this type of representation does not meet the needs of a reader by omitting many details on one hand and not providing the proper level of abstraction on the other hand. For instance, there is an arrow from PDGF receptor to phosphatidylinositol 3-kinase (PI3K), indicating that PDGF receptor "sends a signal to" PI3K, but fails to indicate that this signaling event happens due to the binding between the two molecules (16). In other words, the information level of this signaling event may not be specific enough to provide us with the exact mechanism underlying the signaling event. On the other hand, the diagram may well not be considered as abstract enough to explicitly show, for instance, that the activation of NADPH oxidase by PDGF inhibits protein tyrosine phosphatase (PTP) via production of H_2O_2, thus raising the cellular protein tyrosine phosphorylation level.

In an attempt to overcome aforementioned limitations of conventional presentations of signaling pathways, we propose a new data model that allows us to present the data at multiple levels of abstraction and thus leads to improved

[1] The abbreviations used are: ROS, reactive oxygen species; PDGF, platelet-derived growth factor; PI3K, phosphatidylinositol 3-kinase; PTP, protein tyrosine phosphatase; PTEN, phosphatidylinositol 3-phosphatase; ERK-2, extracellular signal-regulated kinase 2; Trx, thioredoxin.

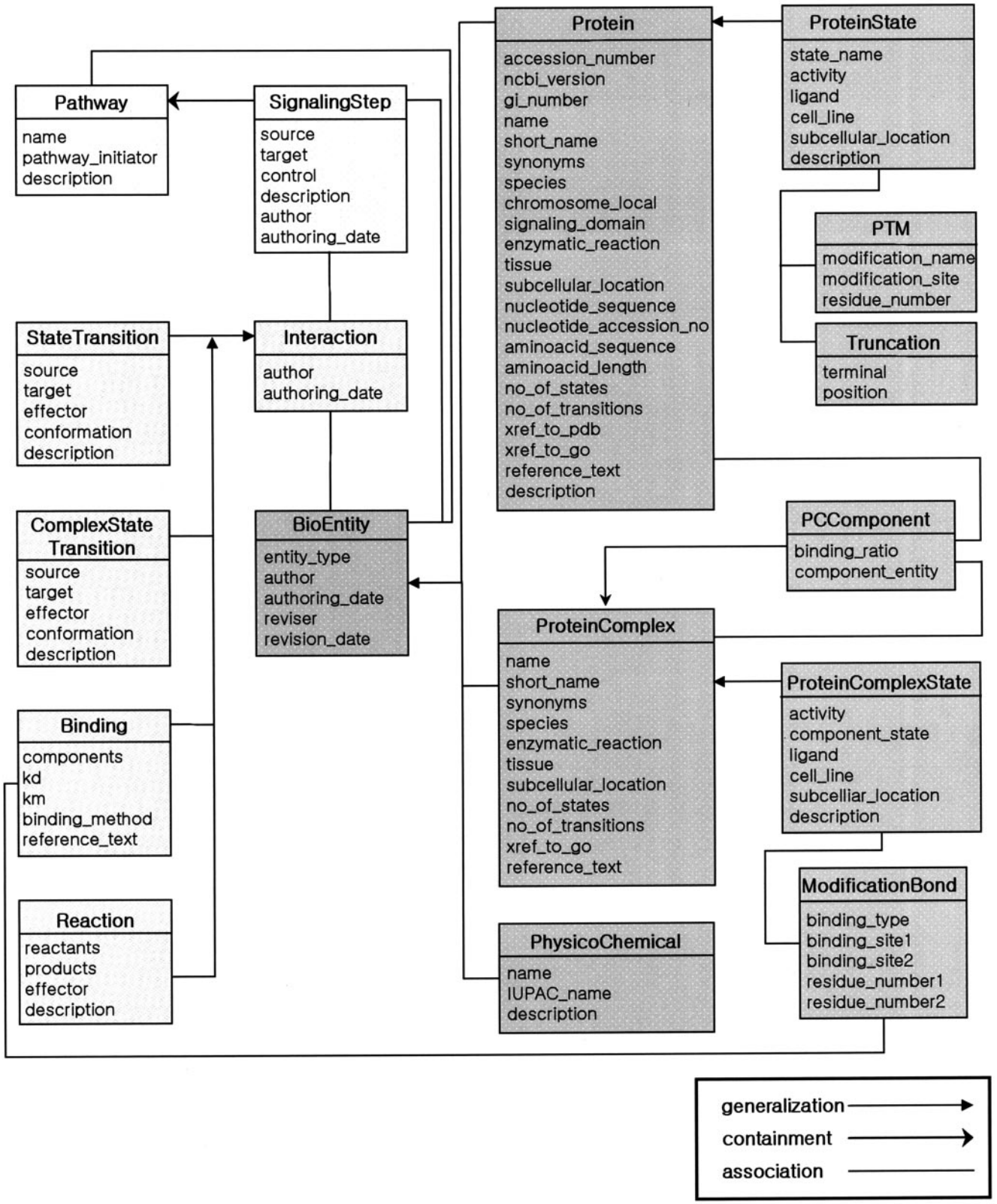


FIG. 1. **A simplified form of Entity-Relationship (ER) diagram for the ROSPath database.**

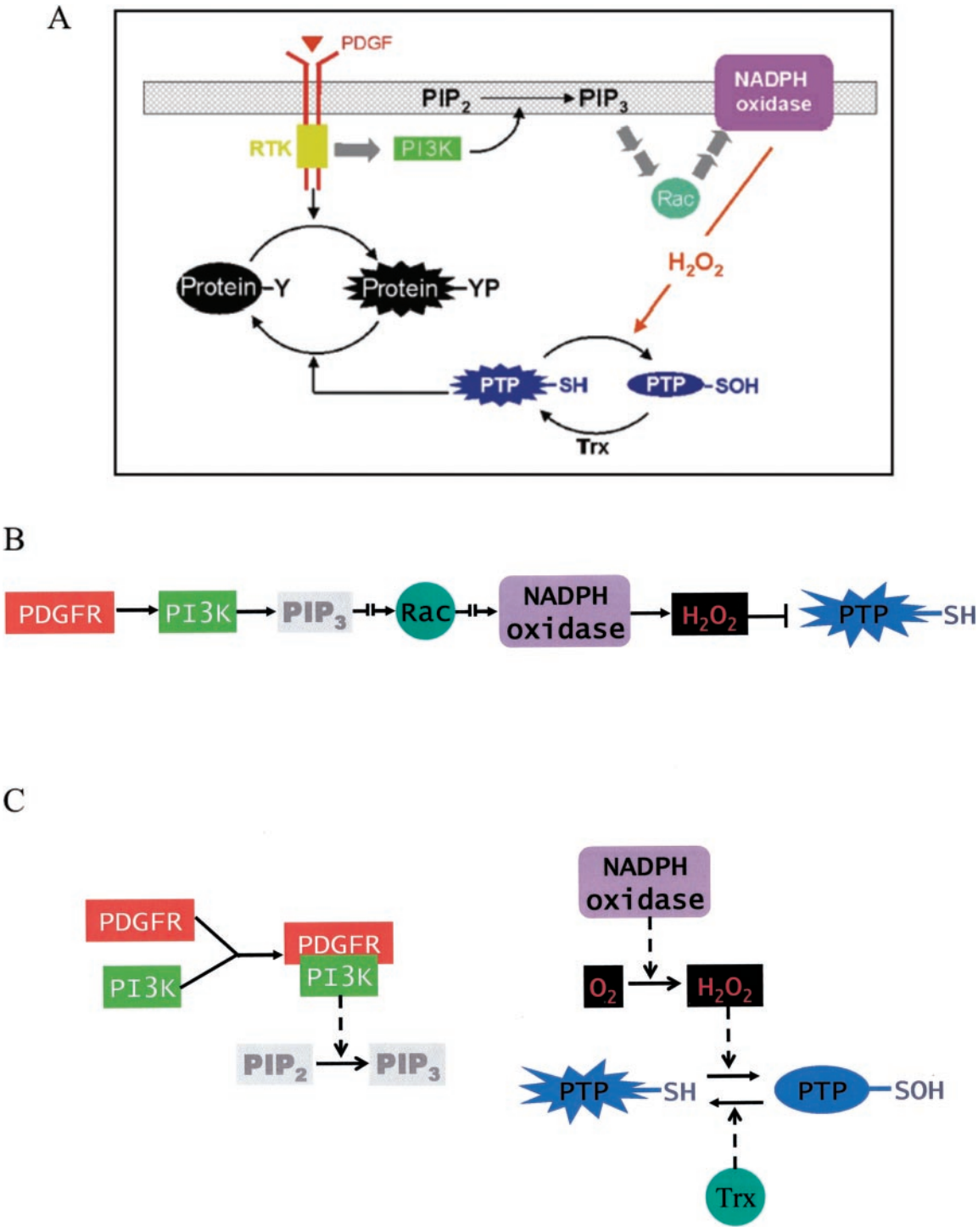


FIG. 2. **Signaling pathways in ROSPath.** *A*, conventional presentation of PDGF receptor signaling pathway: PDGF-induced production of H_2O_2 and the role of H_2O_2 as a modulator of protein tyrosine phosphorylation (3). *B*, manual reconstruction of PDGF receptor signaling pathway in terms of signaling steps as defined by the ROSPath multi-layer model. *C*, manual reconstruction of PDGF receptor signaling pathway in terms of signaling interactions, including state transitions, bindings, and reactions as defined by the ROSPath multi-layer model. In the signaling step level, regular arrows depict positive control, disconnected arrows indicate intermediate steps, not yet known, and arrows with a bar-type arrowhead depict negative control.

TABLE I
Physicochemical factors currently used by ROSPath

Factor
PIP_2
PIP_3
IP_3
DAG
H_2O_2
ROS
O_2
NADPH
Calcium
UV
Heat shock
Small molecules

understanding of cell signaling pathways. We believe that it is essential to have a formal model for cell signaling events, so that a comprehensive yet coherent database structure shall be realized as well as the model support functional computations and simulations. We will first give an intuitive explanation for types of entities and relations that our ontology provides for, explaining how they may be used to describe various cell signaling events. Then our formal model will be presented.

Signaling is mediated through interactions of participating molecular entities. Proteins and their states, protein complexes and their states, and physicochemical factors are the major signaling entities of ROSPath. The concepts of protein state and state transition were introduced to describe the protein-protein interaction in the Live DIP database (2). For the purposes of ROSPath, we define "protein" as a single polypeptide gene product; "protein complex" is an entity comprising of more than two protein units regardless of the nature of their interaction or the stability of the complex; "physicochemical factor" refers to external stimuli including oxidative and heat stress, as well as chemical entities and small molecules involved in signaling. Table I shows the complete list of physicochemical factors currently used by ROSPath. "Protein states" are defined so that we can differentiate proteins existing in various forms in post-translational modifications, activity profiles, conformational changes, subcellular localizations, ligand binding, etc. Table II lists the attributes used to define protein states in ROSPath, together with possible descriptions for each attribute. "Protein complex states" are defined as structural changes that occur in the process of protein complex formation. The protein complex state is critical to many signaling processes because protein complex formation can determine a signaling event. A protein complex exemplifies the use of multiple levels of abstraction in defining ROSPath entities. Protein complexes can be defined among various component proteins or protein complexes. Structured representation of protein complex is made possible because ROSPath allows the protein complex itself to be a component of another protein complex. When such structural information is not available, one can always define protein complex as an aggregation of proteins only.

Dynamic interactions among signaling entities control a course of signaling events. State transitions and binding interactions between two signaling entities, possibly together with their effectors, and any accompanying chemical and structural changes constitute the details of signaling events. "State transitions" are defined for a pair of protein states or protein complex states. For example, two state transitions of PTP are defined in Fig. 2*A*: oxidation with H_2O_2 as an effector and reduction with thioredoxin (Trx) as an effector. Signaling events are often mediated by physical interactions of more than two signaling entities. Some proteins carry out their functions exclusively when complexed with other protein(s), while other proteins interact with each other only when the signaling events occur. The physical interaction of two protein entities is called "binding." In Fig. 2*A*, a signaling event from PDGF receptor to PI3K is shown without explaining the detailed mechanism. The ROSPath model shown in Fig. 2, *B* and *C*, allows us to specify that this signaling event occurs due to the binding of specifically phosphorylated PDGF receptor to PI3K and provides information on the features of the binding phenomenon such as whether the binding molecules must be in a specific state; which subcomponent of a binding molecule (PI3K, in this example) actually participates in the binding, and what the binding site is. Signaling events that do not occur as binding and/or state transition are described as "reactions." Reactions are general interactions involving any signaling entities including protein, protein complexes, protein/protein complex states, and physicochemical factors. Reactions describe chemical transitions among proteins and chemicals, and association and dissociation of more than two protein/protein complexes. For example, a chemical transition from phosphatidylinositol 4,5-bisphosphate (PIP_2) to phosphatidylinositol-3,4,5-trisphosphate (PIP_3) catalyzed by PI3K in Fig. 2*A* is defined as a reaction in ROSPath.

"Signaling events" are composed of chemical and biological transformations among signaling entities, but these individual changes do not necessarily convey the information flow involved in signaling pathways. A "signaling step" is the basic logical unit that corresponds to a signal flow within a single signaling event. The mechanism underlying the signal flow represented by a signaling step is represented as a set of "signaling interactions." A signaling interaction can be a state transition, a binding interaction, or a chemical reaction. Signaling steps and signaling pathways are designed to convey signal flow in a complex biological system, at a higher level of abstraction than the abstraction level of signaling interactions. In ROSPath, protein state transition from PTP-SH to PTP-SOH with H_2O_2 as an effector can also be modeled at the same time as a signaling step from H_2O_2 to PTP-SH, where the type of signaling step is "inhibition" as shown in Fig. 2*B*. A "signaling pathway" is defined as a set of signaling steps that

TABLE II
Description of protein states

Attributes		Description
Post-translational modification	Carbohydrate modifications	Glycation/deglycation *N*-Glycosylation, *O*-glycosylation/deglycosylation
	Lipid modifications	Farnesylation/defarnesylation Geranylation/degeranylation Myristoylation/demyristoylation Palmitoylation/depalmitoylation Prenylation/deprenylation Stearoylation/destearoylation
	Oxidation/reduction of cysteine	Cysteine-cysteine disulfide/disulfide-cysteine Cysteine-cysteine sulfenic acid/cysteine sulfenic acid-cysteine Cysteine-cysteine sulfinic acid/cysteine sulfinic acid-cysteine Cysteine-cysteine sulfonic acid/cysteine sulfonic acide-cysteine Cysteine-cysteine sulphenyl-amide/cysteine sulphenyl-amide-cysteine
	Oxidation/reduction of selenocysteine	Selenocysteine-selenocysteine diselenide/selenocysteine diselenide-selenocysteine Selenocysteine-selenocysteine selenenate/selenocysteine selenenate-selenocysteine Selenocysteine-selenocysteine selenenylsulfide/selenocysteine selenenylsulfide-selenocysteine Selenocysteine-selenocysteine seleninate/selenocysteine seleninate-selenocysteine
	Other oxidation or reduction	Carbonylation/decarbonylation *S*-Glutathiolation/deglutathiolation Hydroxylation/dehydroxylation Iron-sulfur oxidation/iron-sulfur reduction Methionine sulfoxide/Methionine sulfoxide-methionine *S*-Nitrosylation
	Phosphorylation or dephosphorylation	Phosphorylation/dephosphorylation
	Proteolytic modifications	Neddylation/deneddylation Sumoylation/desumoylation Ubiquitination/deubiquitination
	Other modifications	5′-Adenosylation Acetylation/deacetylation ADP-ribosylation Amidation/deamidation Biotinylation Carboxylation/decarboxylation Crosslinking Formylation/deformylation Lipoylation Methylation/demethylation Nitration/denitration Sulfation/desulfation
Truncation	N-terminal, C-terminal, cleavage site	
Ligand	ADP, ATP, Ca^{2+}, cAMP, cGMP, DAG, GDP, glucose, GTP, H_2O_2, heme, NADP, NADPH, NO, O_2^-, PIP, PIP_2, PIP_3	
Cell line	B cell, endothelial cell, epithelial cell, smooth muscle cell	
Subcellular localization	Cytosol, endoplasmic reticulum, endosome, extracellular region, Golgi apparatus, membrane, mitochondrion, nucleus, outer mitochondrial membrane, perinuclear region, peroxisome, plasma membrane, ribosome	
Activity	Active, inactive, unknown	

conveys the information flow in signaling events. The extent of molecules used to define a signaling pathway is variable: it may apply to biological events in a relatively closed experimental system or may apply to the entire cellular system of the signaling network. Fig. 2*B* presents ROSPath reconstruction of the pathway shown in Fig. 2*A* as a cascaded set of signaling steps, with regular arrows depicting positive control, disconnected arrows indicating that intermediate steps, not yet known, are likely to exist, and bar-type arrowheads depicting negative control.

Each signaling step in a pathway can have another description at a more detailed level, that is, an interaction level, in terms of state transitions, bindings, and reactions. When each step presented in Fig. 2*B* is described at a more specific level of abstraction, the diagram shown in Fig. 2*C* results.

A Formal Definition of an Ontology—To model a signaling pathway, we use a hierarchical hyper graph. A hierarchical hyper graph (17) is an extension of a graph definition in which each edge can contain a graph inside itself and each node can have an internal hierarchical structure.

A "signaling pathway graph" $SP = (I, S)$ is defined as the pairing of an "interaction graph" I and a "signaling graph" S. An interaction graph is a model to describe a pathway at the level of signaling interactions, and each hyper edge corresponds to a signaling interaction. A signaling graph is a model for a given signal pathway at the level of information flow, and each edge in the signaling graph corresponds to a signaling step.

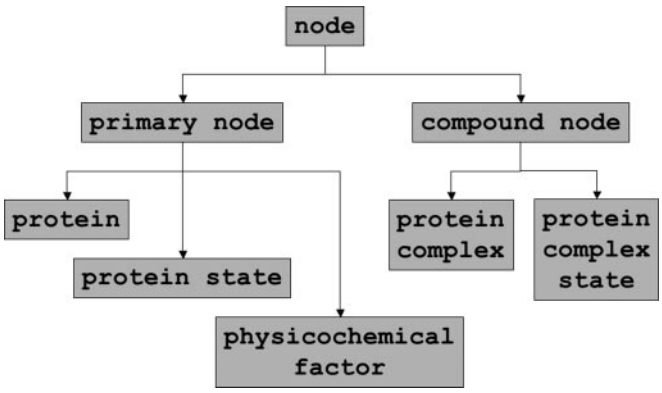


FIG. 3. **Node hierarchy in the ROSPath signaling pathway.**

Signaling entities participating in signaling pathways are modeled as a set V of nodes and are classified into two groups: primary nodes and compound nodes as explained in Fig. 3.

Primary nodes represent a single signaling unit such as proteins, protein states, and physicochemical factors, denoted by V_p, V_{ps}, and $V_{c,}$ respectively. Compound nodes represent protein complexes and protein complex states, denoted by V_{pc} and V_{pcs}, respectively. Compound nodes V_{pc} and V_{pcs} are recursively defined by a set of primary nodes and compound nodes as follows.

$$V_{pc} = \{\{v_1, v_2, \ldots, v_n\} | v_i \in V_p \cup V_{pc}, 1 \le i \le n\}$$

$$V_{pcs} = \{\{v_1, v_2, \ldots, v_m\} | v_i \in V_{ps} \cup V_{pcs}, 1 \le i \le m\}$$

An interaction graph $I = (V^I, E^I, T^{II})$ is a hyper graph defined by a finite set of nodes V^I ($\subseteq V$), a finite set of hyper edges E^I, which are ordered pairs ($<a_1, a_2, \ldots, a_i>$, $<b_1, b_2, \ldots, b_j>$, I) of two sequences of nodes with a label I, and a mapping table T^{II}. The start terminal of a hyper edge, $<a_1, a_2, \ldots, a_i>$ represents a list of source nodes, and the end terminal $<b_1, b_2, \ldots, b_j>$ represents a list of target nodes of the edge. A label of a hyper edge describes an edge type such as binding, state transition, and reaction. The semantics for each edge type is summarized in Table III. A mapping table T^{II} contains ordered pairs (e^I_i, $\{e^I_1, e^I_2, \ldots, e^I_k\}$) of an hyper edge $e^I_i \in E^I$ and a set of hyper edges $\{e^I_1, e^I_2, \ldots, e^I_k\}$, where $e^I_j \in E^I$ ($1 \le j \le k$). Each pair defines the abstraction relation between interaction graphs. An example mapping from a hyper edge to another hyper edge is illustrated in Table IV. This example describes that a state transition from Trx_{ox} to Trx_{red} with TrxR as an effector can also be modeled as a reaction among Trx_{ox}, Trx_{red}, $TrxR_{red}$, and $TrxR_{ox}$.

A signaling graph $S = (V^S, E^S, T^{SI})$ is defined by a finite set of nodes V^S($\subseteq V$), a finite set of edges E^S, and a mapping table T^{SI}. Each edge $e^S_i \in E^S$, called a "signaling step," is defined by an ordered pair (a, b, I), where a is a source node, b is a target

TABLE III
Edge type ($\langle a_1, a_2, \ldots, a_i\rangle$, $\langle b_1, b_2, \ldots, b_j\rangle$) in interaction graph

Type of edges	Semantics	Restriction
Binding	Proteins (or protein complexes) a_1 and a_2 bind together to form b_1.	$i = 2, j = 1$
State transition	Protein (or protein complex) in state a_1 is transformed into another state b_1, possibly with an effector a_2. Effector may be any signaling entity, primary or compound.	$i \le 2, j = 1$
Reaction	A set of signaling entities is involved in any type of reaction, that cannot be described as binding or state transition.	None

TABLE IV
An example mapping table T^{II}

Hyper edge (from)	Hyper edge (to)
($\langle Trx_{ox}$, TrxR$\rangle$, Trx_{red}, state transition)	($\langle Trx_{ox}$, $TrxR_{red}\rangle$, $\langle TrxR_{ox}$, $Trx_{red}\rangle$, reaction)

node, that has a label *l*. Edge labels describe the type of edges as belonging to one of activation, inhibition, suspected positive control, suspected negative control, positive control with intermediate steps omitted, and negative control with intermediate steps omitted. Edge type is summarized in Table V together with a graphical notation for each type. A mapping table T^{SI} contains ordered pairs (e^S_i, {e^I_1, e^I_2, . . . , e^I_k}) of a signaling step $e^S_i \in E^S$ and a set of hyper edges {e^I_1, e^I_2, . . ., e^I_k}, where $e^I_j \in E^I$ ($1 \leq j \leq k$). Each pair describes the abstraction relationship in which one or more hyper edges are involved in a signaling step. An example of a mapping table T^{SI} can be found in Table VI. Here, a signaling edge from H_2O_2 to PTEN representing that H_2O_2 inhibits PTEN involves the following two interaction edges: a state transition from Trx to Trx1 with H_2O_2 as an effector and a state transition from PTEN = $(S)_2$ to PTEN = $(SH)_2$ with Trx as an effector.

One can see that the recursive definition of protein complexes and their states enables structured representation for signaling entities because protein complex as well as protein can play as signaling entities. Because signaling pathway can be presented with signaling steps and their detail interactions, the mappings between signaling graphs and interaction graphs as well as those between two interaction graphs make it possible to represent signaling events in a structured format.

Data Queries—Database users can easily browse through various entities defined in ROSPath, such as proteins, protein complexes, and their individual states. Alternately, one can query these entities with a name, species, tissue source, or other characteristics that define each protein state, such as post-translational modification and/or ligand binding. Fig. 4*A* shows a sample query page together with its search result from which one can follow various links to find out more about protein information, protein states, etc.

Fig. 4*B* shows a sample protein information page. One can follow various links from the protein page, if more detailed information about proteins or their states are needed. An example is shown in Fig. 4*C*, which reveals that the modification is phosphorylation at 308 serine, 473 threonine, and 474 tyrosine, and that this information has been obtained from human umbilical vein endothelial cell (HUVEC) line.

One can search for signaling interactions that simply involve certain signaling entities. For example, when a search request is to find all the binding interactions that involve PDGF receptor, the results shown in Fig. 4*D* are obtained. One learns that PDGFR-β forms noncovalent binding with Shc, and that phosphotyrosine at amino acid 740 of PDGFR can interact with SH2 domain of Shc.

Visualization and Graphical Queries—ROSPath provides prototype pathway visualization software that allows automatic visual inspection of stored signaling pathway information. A detailed description of its layout algorithm will be published elsewhere. It is possible to dynamically generate a graphic view of signaling pathway data stored in the ROSPath database by communicating with the ROSPath web and database servers. Different types of signaling entities and signaling interactions have distinct graphical representation for quick and easy visual discrimination. It can also process the graph-theoretic queries and present the search results as a highlighted subgraph within the context of the target pathway. Fig. 5 is an example snapshot of the ROSPath visualization software. A search for all the signaling paths from PDGFR-β to extracellular signal-regulated kinase 2 (ERK-2) will retrieve six different paths within a PDGF signaling pathway. Simply by clicking on any path from the search results, one can highlight it within the pathway diagram.

The most significant feature of the application is its ability to provide two mutually related graphical representations of a signaling pathway, one as a signaling graph and the other as an interaction graph. When there is a mapping between signaling and interaction graphs specified in the ROSPath database, one could easily browse the correspondence between the two graphs by clicking on the signaling step arrow. According to the formal ontology, it is also possible to define a mapping between interaction graphs to an arbitrary depth for finding the new connectivity.

Another feature of our application is its ability to draw

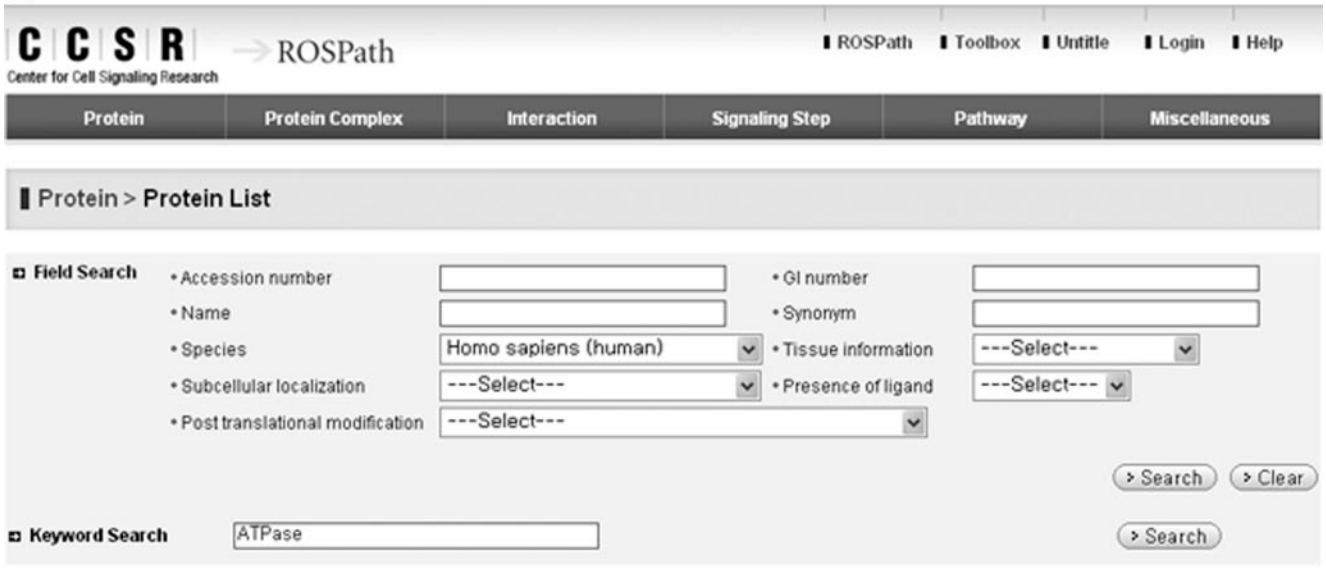
TABLE V
The type of signaling edges

Edge Type	Graphical Notation
Activation	——▶
Inhibition	——⊣
Suspected (positive control)	- - - -▶
Suspected (negative control)	- - - -⊣
Intermediate steps omitted (positive control)	—⊣ ⊢▶
Intermediate steps omitted (negative control)	—⊣ ⊢⊣

TABLE VI
An example of a mapping table T^{SI}

	Signaling step	Hyper edges in interaction graph
1	(H_2O_2, PTEN, inhibition)	(⟨Trx, H_2O_2⟩, Trx1, state transition) (⟨PTEN = $(S)_2$, Trx⟩, PTEN = $(SH)_2$, state transition)
2	(PI3K, PIP_3, activation)	(⟨PIP_2, PIP_3⟩, PI3K, reaction)

A



Sort Protein ID Total: 11 New Protein

Protein ID	Protein Name	GI No.	Link to NCBI	Synonym	Species	No of State	No of Transition
PR000336	Similar to proteasome (prosome, macropain) 26S subunit, ATPase, 1	16741033	AAH16368	Similar to proteasome (prosome, mac ...	Homo sapiens (human)	0	0
PR000329	AHA1, activator of heat shock 90kDa protein ATPase homolog 1	6912280	NP_036243	AHA1, activator of heat shock 90kDa ...	Homo sapiens (human)	0	0
PR000314	proteasome 26S ATPase subunit 2	4506209	NP_002794	proteasome 26S ATPase subunit 2,pro ...	Homo sapiens (human)	0	0
PR000310	proteasome 26S ATPase subunit 4	5729991	NP_006494	proteasome 26S ATPase subunit 4,pro ...	Homo sapiens (human)	0	0
PR000292	Transitional endoplasmic reticulum ATPase	6094447	P55072	Transitional endoplasmic reticulum ...	Homo sapiens (human)	0	0
PR000279	Na,K-ATPase alpha-4 subunit	17149816	AAK72396	Na,K-ATPase alpha-4 subunit,Na,K-AT ...	Homo sapiens (human)	0	0
PR000212	26S proteasome ATPase subunit	2791680	AAC26843	26S proteasome ATPase subunit	Homo sapiens (human)	1	0
PR000164	proteasome 26S subunit, ATPase 2	12803525	AAH02589	proteasome (prosome, macropain) 26S ...	Homo sapiens (human)	1	2
PR000158	transitional ER ATPase	2984586	AAC07984	TERA_HUMAN,VCP,valosin containing p ...	Homo sapiens (human)	2	1
PR000109	proteasome 26S subunit, ATPase 2	13529470	AAH05462	proteasome (prosome, macropain) 26S ...	Mus musculus (house mouse)	1	0

FIG. 4. **User interface for the ROSPath web pages.** *A*, database search for signaling entities. *B*, display of protein information. *C*, display of protein states. *D*, database search for signaling interactions.

hyper graphs. Signaling graphs are general directed graphs, while interaction graphs are hyper graphs. Most graph generation tools avoid the problem of drawing hyper graphs by indicating signaling interactions as nodes, rather than as edges, and by connecting related entities with so called "interaction nodes" (18–20). Hyper graphs can thus be transformed into regular directed graphs. This visual representation, however, is very different from the intuitive graph representations that biologists have long been using to represent cellular signaling pathways. While these systems automatically generate graphs from a database of signaling interactions, there are efforts to provide software tools that allow users to manually draw a signaling pathway diagram at will (21). Related work can be found for drawing interaction graphs (22–24), but interaction network diagrams are intrinsically different from signaling graphs in that edges are not directed, and they are not hyper graphs. In our system, we have successfully generated hyper graphs, thereby facilitating visual inspection and use of the computer-generated pathway diagram.

When the application draws each pathway as a graph, it tries to minimize the edge crossings. The current implemen-

B

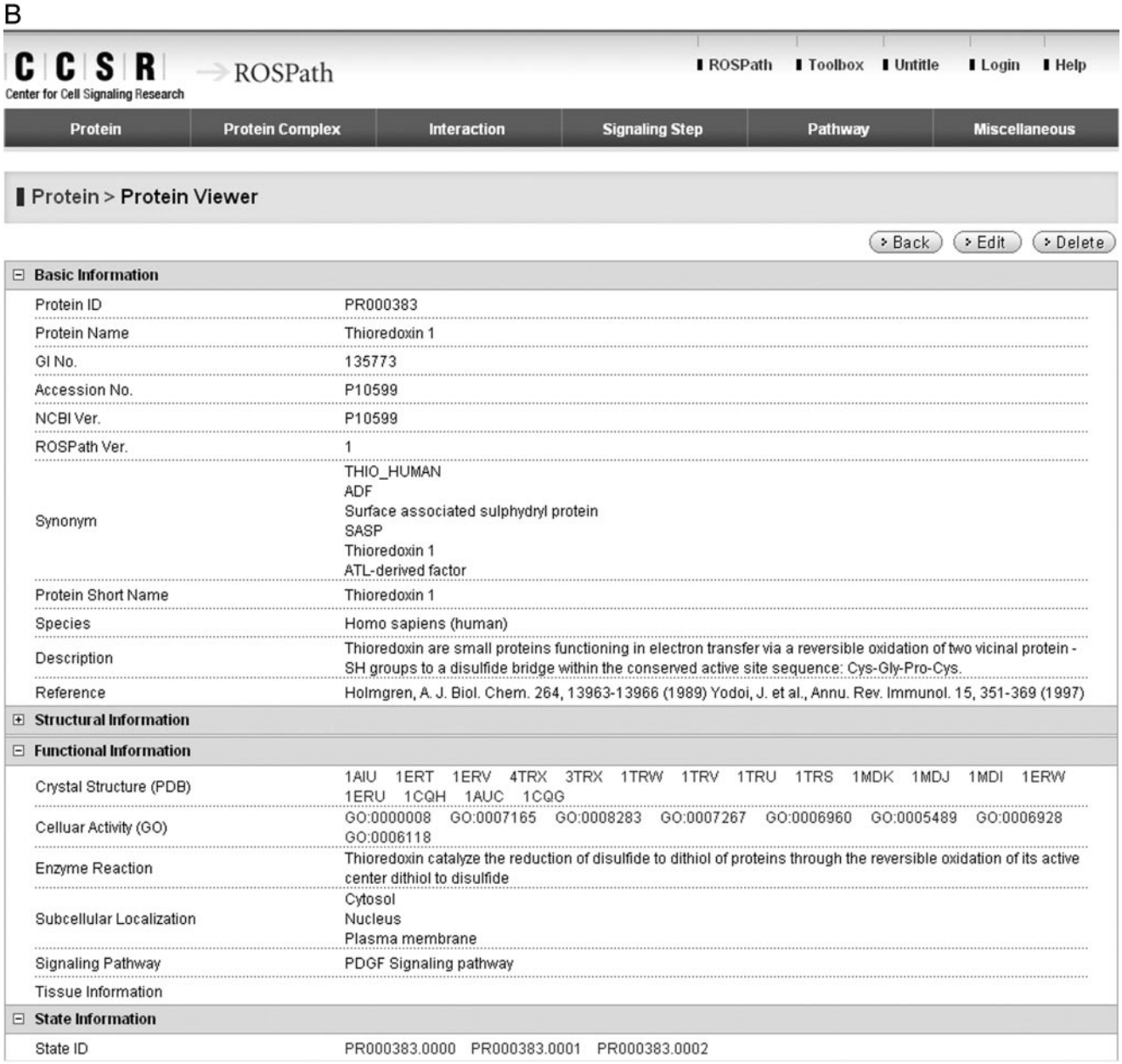


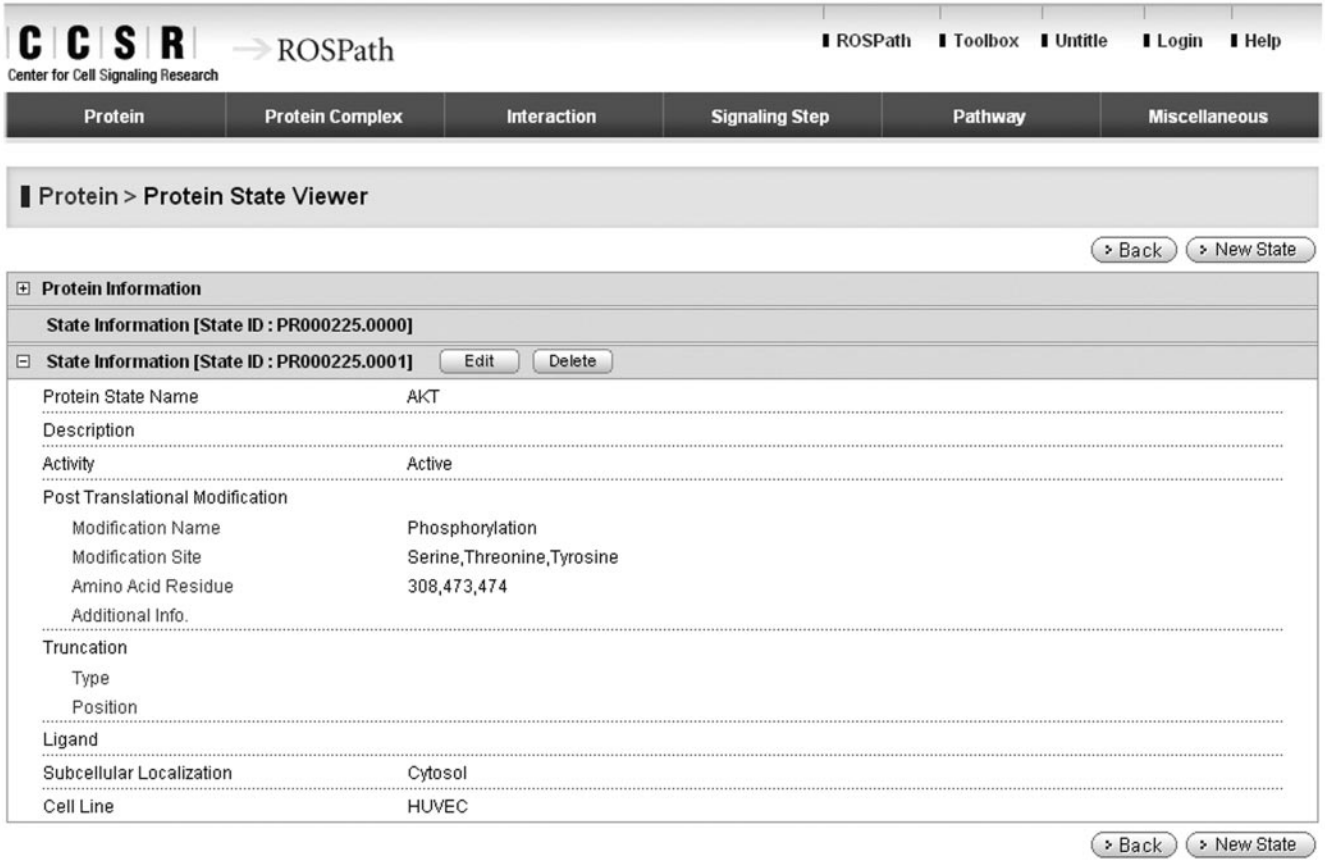
FIG. 4—*continued*

tation adopts various layout algorithms including Mincut and Sugiyama algorithms (25), depending on one's preference.

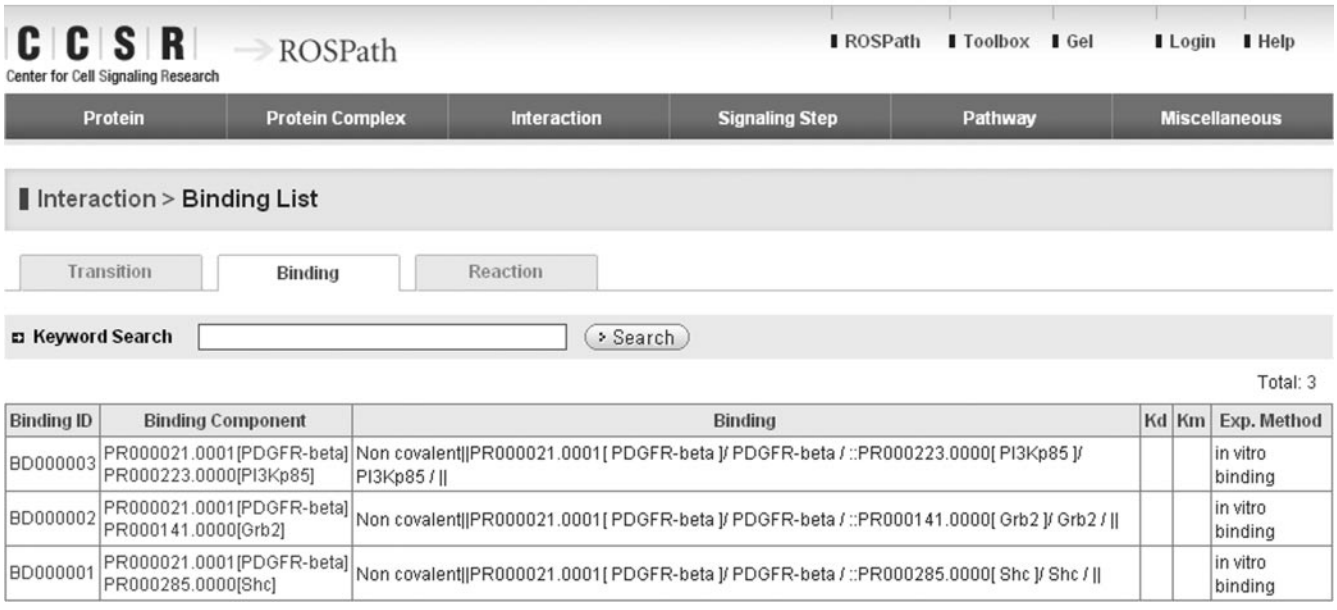
DISCUSSION

There have been various efforts to develop a formal model for signaling pathways that supports multiple abstraction levels and representation of incomplete knowledge. Demir *et al.*, in their work on PATIKA (20), defined an "abstraction node" that represents a complex molecule or abstracts a partial pathway. When modeling a signaling pathway as a graph, biochemical entities are generally represented as nodes and interactions such as state transitions or chemical reactions are represented as edges. In PATIKA, on the other hand, a node of a graph may represent an interaction or a partial pathway, as well as a biochemical entity. Such an abstraction simplifies the model because signaling pathways can always be represented as a directed acyclic graph. The model, however, is limited in that it does not support multi-level abstraction when defining complex molecules, and that it does not support multi-layered abstraction of signaling events. In PATIKA, signaling events are described only in terms of interactions, and cannot explicitly represent how the signal flow should be understood. Fukuda *et al.* (26) introduced a "decomposition tree" that supports multi-level abstraction of

C



D



Binding ID	Binding Component	Binding	Kd	Km	Exp. Method
BD000003	PR000021.0001[PDGFR-beta] PR000223.0000[PI3Kp85]	Non covalent||PR000021.0001[PDGFR-beta]/ PDGFR-beta / ::PR000223.0000[PI3Kp85]/ PI3Kp85 / ||			in vitro binding
BD000002	PR000021.0001[PDGFR-beta] PR000141.0000[Grb2]	Non covalent||PR000021.0001[PDGFR-beta]/ PDGFR-beta / ::PR000141.0000[Grb2]/ Grb2 / ||			in vitro binding
BD000001	PR000021.0001[PDGFR-beta] PR000285.0000[Shc]	Non covalent||PR000021.0001[PDGFR-beta]/ PDGFR-beta / ::PR000285.0000[Shc]/ Shc / ||			in vitro binding

FIG. 4—*continued*

complex molecules and an "interaction graph" that represents interactions between arbitrary nodes in a given decomposition tree. Using these two graph structures, it is possible to define the interrelationship between various levels of abstraction. This representation describes signaling as interactions and decomposition, but does not explicitly represent the in-

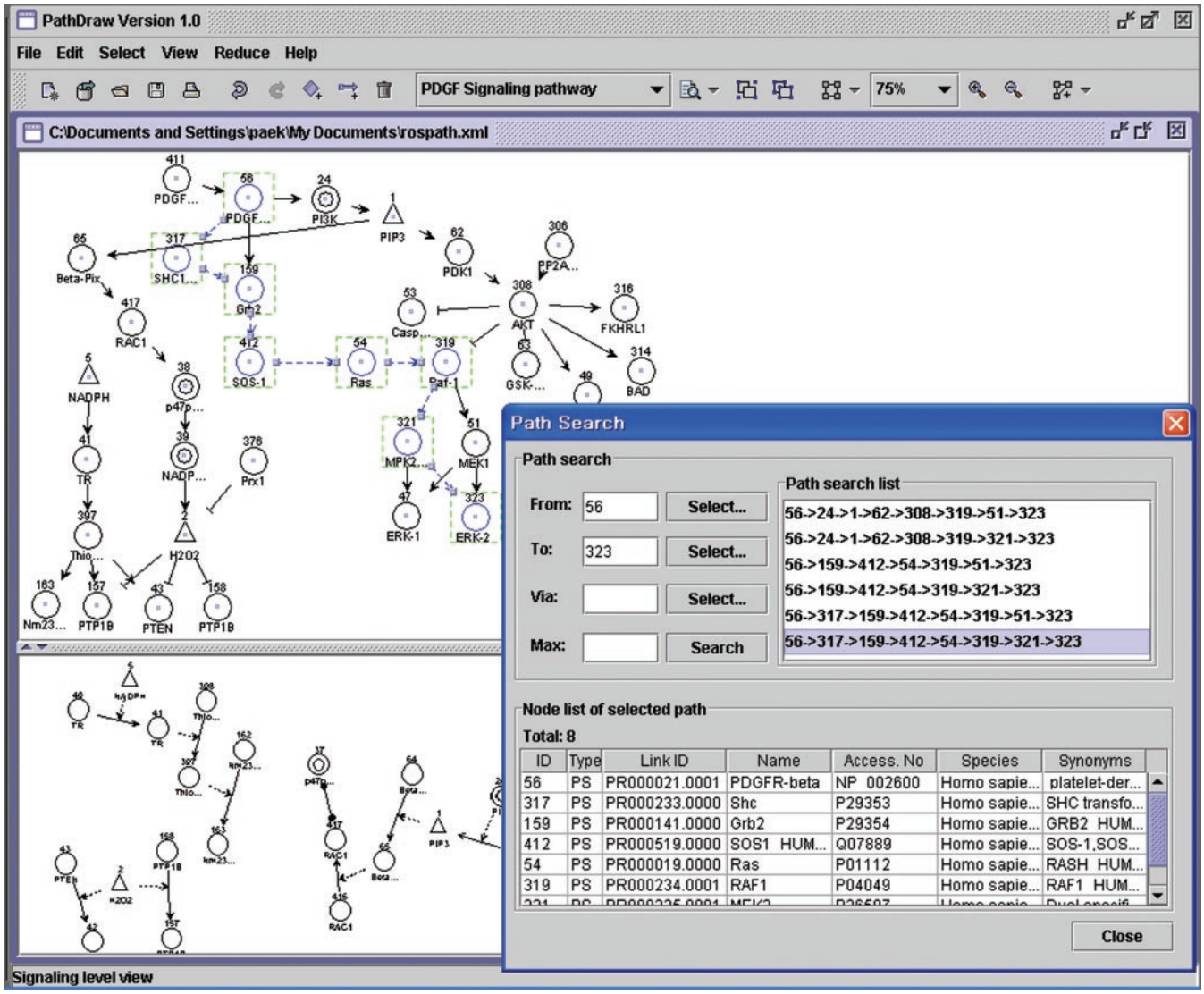


FIG. 5. **Snap shot of pathway visualization software.** In each signaling pathway, signaling entities are represented as nodes and signaling interactions as edges; the upper window displays signaling level and the bottom window interaction level; each edge at the signaling level may correspond to a hyper graph in the bottom window. A search request to find all the paths from PDGFR-β to ERK-2 yields six different paths within a PDGF signaling pathway as shown in a pop-up window. A path selected from the search result is highlighted within the pathway diagram.

formation flow within a signaling pathway.

Database systems such as KEGG (27) and EcoCyc (28) have been developed that computerize knowledge about information pathways of genes and gene products. KEGG's pathway database contains graphical representations of cellular processes such as metabolism, membrane transport, and signal transduction. The generalized protein interaction network in KEGG's pathway database is drawn manually as a graphical pathway diagram. It is also stored as a set of binary relations such as enzyme-enzyme relations, protein-protein interactions, and gene expression relations, which enables to reproduce the known pathways. EcoCyc is an organism-specific pathway and genome database that describes the metabolic and signal-transduction pathways of *Escherichia coli*. The EcoCyc database contains enzymes, transport proteins, and mechanisms of transcriptional control of gene expression of *E. coli*. The pathway data in the EcoCyc database is stored as a predecessor list that encodes the ordering of reactions within the pathway, and the graphical representation of the pathway is derived from this predecessor list. Although EcoCyc can provide pathways at multiple levels of detail, ranging from a skeletal view of a pathway to a detailed view showing full structures for each chemical compound, it dose not focus on describing multiple levels of abstraction between signaling events and corresponding reactions as in the ROSPath data model.

On the other hand, there have been efforts such as SBML (29), CellML (30), and BioPAX (www.biopax.org) that aim to

standardize the format of biochemical and/or biophysical data. SBML proposes a computer-readable format for representing biochemical reaction network encompassing metabolic, cell-signaling, and genomic regulatory networks so that different software tools can be used without rewriting models for each tool. CellML focuses on providing mathematical models for cellular function. BioPAX's focus is on providing a unified framework for sharing biological pathway information, encompassing metabolic pathways, signal transduction pathways, gene regulatory pathways, genetic interaction pathways, pathways through word relationships found in paper using text mining, computationally predicted pathways, and pathways of cellular interactions. We are working with the BioPAX group so that ROSPath can be compatible with BioPAX specification.

The ROSPath data model focuses on cell signaling pathways and provides a refined model for protein and protein complexes, and their states, while enabling multi-layered representation based on distinctive conceptualizations at a different layer.

The most distinguishing characteristic of ROSPath is its foundation on a formal representation of signaling and interaction among signaling entities, while maintaining the connection between different levels of representations. Pathway visualization software is associated with the database so that visual and dynamic inspection of stored signaling pathway information is possible. The most significant feature of the application is its ability to provide two related graphical representations of a signaling pathway, one as a signaling graph and the other as an interaction graph. Another feature is its ability to draw hyper graphs, thereby facilitating visual inspection and use of the computer-generated pathway diagram.

The current contents of the ROSPath database is concentrated only on cell signaling events that allows description of ROS-mediated signaling cellular pathways at various levels of abstraction. However, the proposed data model is applicable to various other cell signaling pathways as well.

Future Development—Plans are underway 1) to include in the database, details of entities and interactions in each signaling pathway and additional pathway relevant information from literature search, 2) to refine the analysis, simulation, and visualization tools for presenting the dynamic signaling data, and 3) to extract and predict the signaling pathway information from the database.

Acknowledgments—We thank K. Choi of Seoul National University for graphic viewer and S. R. Lee, K. S. Yang, and S. H. Lee of Ewha Womans University for the discussion on signaling pathways.

* This work was supported by IMT2000 project (IMT2000-C5–2) for IT-BT and by KOSEF, through the Center for Cell Signaling Research (CCSR) at Ewha Womans University. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed: Eunok Paek, Department of Mechanical and Information Engineering, University of Seoul, Seoul 130-743, Republic of Korea. Tel.: 82-2-2210-2680; Fax: 82-2-2248-5110; E-mail: paek@uos.ac.kr. Kong-Joo Lee, Center for Cell Signaling Research, Division of Molecular Life Sciences & College of Pharmacy, Ewha Womans University, Seoul 120-750, Republic of Korea. Tel.: 82-2-3277-3038; Fax: 82-2-3277-3760; E-mail: kjl@ewha.ac.kr.

REFERENCES

1. Hunter, T. (2000) Signaling-2000 and beyond. *Cell* **100,** 113–127
2. Duan, X. J., Xenarios, I., and Eisenberg, D. (2002) Describing biological protein interactions in terms of protein states and state transitions: The LiveDIP database. *Mol. Cell. Proteomics* **1,** 104–116
3. Rhee, S. G., Bae, Y. S., Lee, S.-R., and Kwon, J. (2000) Hydrogen peroxide: A key messenger that modulates protein phosphorylation through cysteine oxidation. *Sci. STKE* **53,** 1–6
4. Torres, M., and Forman, H. J. (2003) Redox signaling and the MAP kinase pathways. *Biofactors* **17,** 287–296
5. Park, H. S., Lee, S. H., Park, D., Lee, J. S., Ryu, S. H., Lee, W. J., Rhee, S. G., and Bae, Y. S. (2004) Sequential activation of phosphatidylinositol 3-kinase, βPix, Rac1, and Nox1 in growth factor-induced production of H_2O_2. *Mol. Cell. Biol.* **24,** 4384–4394
6. Kreuzer, J., Viedt, C., Brandes, R. P., Seeger, F., Rosenkranz, A. S., Sauer, H., Babich, A., Nurnberg, B., Kather, H., and Krieger-Brauer, H. I. (2002) Platelet-derived growth factor activates production of reactive oxygen species by NAD(P)H oxidase in smooth muscle cells through Gi1,2. *FASEB J.* **17,** 38–40
7. Meng, T. C., Fukada, T., and Tonks, N. K. (2002) Reversible oxidation and inactivation of protein tyrosine phosphatases *in vivo*. *Mol. Cell* **9,** 387–399
8. Maulik, N., and Das, D. K. (2002) Redox signaling in vascular angiogenesis. *Free Radic. Biol. Med.* **33,** 1047–1060
9. Abid, M. R., Tsai, J. C., Spokes, K. C., Deshpande, S. S., Irani, K., and Aird, W. C. (2001) Vascular endothelial growth factor induces manganese-superoxide dismutase expression in endothelial cells by a Rac1-regulated NADPH oxidase-dependent mechanism. *FASEB J.* **15,** 2548–2550
10. Ushio-Fukai, M., Tang, Y., Fukai, T., Dikalov, S. I., Ma, Y., Fujimoto, M., Quinn, M. T., Pagano, P. J., Johnson, C., and Alexander, R. W. (2002) Novel role of gp91(phox)-containing NAD(P)H oxidase in vascular endothelial growth factor-induced signaling and angiogenesis. *Circ. Res.* **91,** 1160–1167
11. Thannickal, V. J., Day, R. M., Klinz, S. G., Bastien, M. C., Larios, J. M., and Fanburg, B. L. (2000) Ras-dependent and -independent regulation of reactive oxygen species by mitogenic growth factors and TGF-β1. *FASEB J.* **14,** 1741–1748
12. Ha , Y. J., and Lee, J. R. (2004) Role of TNF receptor-associated factor 3 in the CD40 signaling by production of reactive oxygen species through association with p40phox, a cytosolic subunit of nicotinamide adenine dinucleotide phosphate oxidase. *J. Immunol.* **172,** 231–239
13. Hirotani, S., Otsu, K., Nishida, K., Higuchi, Y., Morita, T., Nakayama, H., Yamaguchi, O., Mano, T., Matsumura, Y., Ueno, H., Tada, M., and Hori, M. (2002) Involvement of nuclear factor-κB and apoptosis signal-regulating kinase 1 in G-protein-coupled receptor agonist-induced cardiomyocyte hypertrophy. *Circulation* **105,** 509–515
14. Kim, H.-J., Song, E. J., and Lee, K.-J. (2002) Proteomic analysis of protein phosphorylations in heat shock response and thermotolerance. *J. Biol. Chem.* **277,** 23193–23207
15. Kim, Y. M., Kim, H. J., Song, E. J., and Lee, K.-J. (2004) Glucuronic acid is a novel inducer of heat shock responses. *Mol. Cell. Biochem.* **259,** 23–33
16. Klinghoffer, R. A., Duckworth, B., Valius, M., Cantley, L., and Kazlauskas, A. (1996) Platelet-derived growth factor-dependent activation of phosphatidylinositol 3-kinase is regulated by receptor binding of SH2-domain-containing proteins which influence Ras activity. *Mol. Cell. Biol.* **16,** 5905–14
17. Berge, C. (1973) *Graphs and Hypergraphs.* Elsevier, New York.
18. Schacherer, F., Choi, C., Gotze, U., Krull, M., Pistor, S., and Wingender, E. (2001) The TRANSPATH signal transduction database: A knowledge base on signal transduction networks. *Bioinformatics* **17,** 1053–1057
19. Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A., and Wingender, E.

(2003) TRANSPATH: An integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res.* **31,** 97–100
20. Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetain-Atalay, R., and Ozturk, M. (2002) PATIKA: An integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* **18,** 996–1003
21. Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* **31,** 19–20
22. Shannon, P. Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504
23. Han, K., and Ju, B. (2003) A fast layout algorithm for protein interaction networks. *Bioinformatics* **19,** 1882–1888
24. Bader, G. D., Betel, D., and Hogue, C. W. V. (2003) BIND: The biomolecular interaction network database. *Nucleic Acids Res.* **31,** 248–250
25. Sugiyama, K., Tagawa, S., and Toda, M. (1981) Methods for visual understanding of hierarchical system structures. *IEEE Trans. Sys. Man. Cybernetics* **11,** 109–125
26. Fukuda, K., and Takagi, T. (2001) Knowledge representation of signaling transduction pathways. *Bioinformatics* **17,** 829–837
27. Kanehisa, M., and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.* **28,** 27–30
28. Karp, P. D. (2001) Pathway databases: A case study in computational symbolic theories. *Science* **293,** 2040–2044
29. Finny, A., and Hucka, M. (2003) Systems Biology Markup Language: Level 2 and beyond. *Biochem. Soc. Trans.* **31,** 1472–1473
30. Lloyd, C., Halstead, M., and Nielsen, P. (2004) CellML: Its future, present and past. *Progr. Biophys. Mol. Biol.* **85,** 433–450