



Knowledge representation of signal transduction pathways

Ken-ichiro Fukuda^{1,2} and Toshihisa Takagi¹

¹Human Genome Center, Institute of Medical Science, University of Tokyo,
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Received on December 11, 2000; revised on March 7, 2001; accepted on April 17, 2001

ABSTRACT

Motivations: Signal transduction is the common term used to define a diverse topic that encompasses a large body of knowledge about the biochemical mechanisms. Since most of the knowledge of signal transduction resides in scientific articles and is represented by texts in natural language or by diagrams, there is the need of a knowledge representation model for signal transduction pathways that can be as readily processed by a computer as it is easily understood by humans.

Results: A signal transduction pathway representation model is presented. It is based on a compound graph structure and is designed to handle the diversity and hierarchical structure of pathways. A prototype knowledge base was implemented on a deductive database and a number of biological queries are demonstrated on it.

Contact: fukuda-cbrc@aist.go.jp;
ichiro@ims.u-tokyo.ac.jp

1 INTRODUCTION

Signal transduction is the common term used to define a diverse topic that encompasses a large body of knowledge about the biochemical mechanisms (Ray, 1999). Initial studies on signal transduction were focused largely on growth factor signals that upregulate or downregulate certain genes. But the astonishingly rapid progress in this field, which covers model systems from bacteria to humans, demonstrates its importance for the understanding of various biological processes (i.e. control of cell fates, sense of direction, response to blue light, etc.).

The range of topics researchers have to cover is becoming broader and broader, but the information still resides mostly in scientific articles represented by natural language or drawings, which makes it hard to comprehensively trace the required information. This situation motivated the current blossom of Information Extraction (IE) from biological texts and other document

processing projects. But the target information is in most cases interactions or relations of proteins or genes (Proux *et al.*, 2000; Shatkay *et al.*, 2000; Thomas *et al.*, 2000; Rindfleisch *et al.*, 2000; Humphreys *et al.*, 2000).

Data on protein–protein interactions, which can be thought of as units of signals in cellular signal transduction pathways, are becoming also available through experiments and computational analyses in a comprehensive way (Uetz *et al.*, 2000; Fromont-Racine *et al.*, 1997; Flores *et al.*, 1999; Marcotte, 2000; Enright *et al.*, 1999; Marcotte *et al.*, 1999; Eisenberg *et al.*, 2000). But the problem is that these interactions derive a huge graph with a high degree of connectivity, including artifacts and false-positive predictions, which is difficult to interpret.

The cell system orchestrates biochemical interactions to form a signal transduction pathway that realize an appropriate cellular function in each tissue or environment. Hence, to decipher the mystery of the life phenomenon, it is important to understand not only the function of each interaction itself but also the pathway as a whole in which it occurs as a part.

Thus a method capable of structurally representing the signal transduction pathway in a form that can be readily processed by computers and easily understood by humans would be of great value.

The rest of the paper is organized as follows. In the next section we briefly review the features of biological pathways and then introduce our knowledge representation model. After providing examples of queries, we discuss limitations and future directions.

2 FEATURES OF SIGNAL TRANSDUCTION PATHWAYS

In the case of extensive genome knowledge bases, metabolic pathways are commonly represented by a graph-like structure in which nodes represent reactants, products or EC numbers and edges represent reactions. Owing to a matured background knowledge, KEGG has ~90 of these *reference maps* (Kanehisa and Goto, 2000) and EcoCyc has 129 metabolic *pathway objects*

²Present address: Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST) 2-41-6 Aomi, Koutou-ku, Tokyo 135-0064, Japan.

(Karp *et al.*, 1999). Each map or pathway is categorized into an existing taxonomy according to its function.

However, the signal transduction pathway has neither a widely accepted taxonomy of biological functions as does the EC classification system (Webb, 1992) nor an accepted reference pathway map, but is rather a huge archive of small fragmented pathways that concerns various species, cell-lines, and developmental stages. This makes signal transduction domain different from metabolic pathway domain. Many types of molecules are involved in signal transduction, e.g. hormones, neurotransmitters, membrane receptors, ion channels, transcription factors, metal ions, and so on. In addition, the molecules are related to each other not only by biochemical interactions (local functions; Karp, 2000) but also by direct or indirect interactions of regulations (regulatory functions, i.e. upregulation/downregulation, activation/inhibition and increase/decrease). These two functions represent different concepts. A protein that is 'regulatively' activated can be 'physically' either phosphorylated or dephosphorylated by that activation. In this sense regulatory interactions represent a protein's function in a more abstract manner with the interpretation of the authors.

The diversity of concepts involved in signal transduction together with the rapid and uninterrupted progress of research in this area results in the lack of a well-defined language to write down the knowledge in articles. And this lack makes it difficult to model a signal transduction pathway diagram as a *graph* as described below.

2.1 Heterogeneous knowledge granularity

A signal transduction pathway diagram, which represents a mechanism of a certain cell function, consists of a set of orchestrated local functions. Thus each part of a pathway can be decomposed into fragments of pathways at an arbitrary level of detail, depending on the author's interest. This results in a diagram that consists of biological entities of irregular knowledge granularity.

Figure 1 shows examples that describe the same TGF- β /Smad signaling pathway. Figure 1a is from a review (Ulloa *et al.*, 1999). It shows that TGF- β assembles a ligand-induced receptor complex that phosphorylates a member of the R-Smad class, Smad3, which enables Smad3 to associate with Smad4 to accumulate in the nucleus. The receptor complex consists of TGF- β and two receptor serine/threonine kinases, type I and II. Type II receptor activates type I receptor through phosphorylation. Smads in the nucleus associate with DNA binding cofactors and activators or repressors. Figure 1b is taken from a study that reported an interplay between TGF- β signaling and IFN- γ signaling. Accordingly, this figure does not go into detail, but represents the pathway in a more abstract way. Both examples represent a pathway by means of a graph-based structure with circles and arrows. While in the former ex-

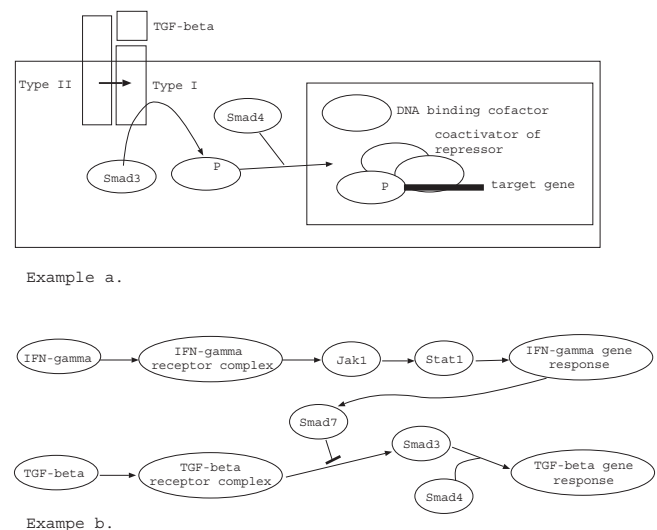


Fig. 1. Examples of TGF β signaling diagram.

ample a circle represents a protein, in the latter a circle represents one of three different concepts: a protein, a receptor complex consisting of several proteins, and a cell response to outer stimuli which itself can be represented as a pathway.

Therefore to represent these diagrams as a graph, one needs to decide what the primitive components are in each pathway, and to comprehensively determine how each node can be decomposed into these primitives. However, considering the number of available articles and the diversity of the signal transduction domain, a bottleneck is likely to arise in the data-retrieval phase when constructing a knowledge base.

Another drawback of a graph representation is that it omits the author's sectioning of a pathway (i.e. receptor complexes, downstream cascades, etc.), which is important to understand the pathway's structure. For example, a detailed graph with primitives as its nodes connected by edges constitutes an overly complex model and is not informative to a user who only wants to know the substrate of a TGF- β receptor. This loss of information caused by omitting the sectioning can be crucial, as we do not have a widely accepted taxonomy of signal transduction pathways.

2.2 Incomplete knowledge description

Studies on signal transduction pathways often do not provide comprehensive information about every relation in the pathway (Takai-Igarashi and Kaminuma, 1999). For example, the author may omit a well-known portion of the pathway under the presumption that the reader is already familiar with it. A portion of the pathway not clearly understood yet may also be omitted.

Because of this incompleteness, even a well-known pathway can be fragmented in a knowledge base. An example is the description of a complex such as ‘*AtCUL is a component of an SCF(SKP-cullin-F-box) ubiquitin ligase complex that includes TIR1, the F-box protein, and ASK1, a homolog of yeast SKP1*’ (McCarty and Chory, 2000). This description defines the members of a complex, but not the direct relations between them. Consider a pathway of proteins A , B and C , where C is a complex with subcomponents C_1 , C_2 , and C_3 , and where A interacts with C_1 and B interacts with C_3 , but direct interactions between (C_i, C_j) are not provided. A graph will represent A , B , C_1 , C_2 , and C_3 as nodes with edges (A, C_1) , (C_3, B) . In this representation the pathway from A to B through C is fragmented because a path from C_1 to C_3 is not defined. A way to avoid this fragmentation is to add virtual edges between all (C_i, C_j) . But this does not allow representation of subcomponents of a subcomponent. In addition the graph will have become overly intricate, and to find a complex one needs to find a clique[†].

From these observations we recognized that a signal transduction pathway diagram should be modeled as a graph whose nodes can be decomposed into another graph for arbitrary depth.

3 METHODS

3.1 Modeling pathways

To model a signal transduction pathway we use a *compound graph*. The method is designed to capture directly the structure of pathways that biologists bear in mind or that are described in articles. The model does not assume the existence of a well-defined semantics of the domain.

A compound graph is an extension of a graph definition in which each node can contain a graph inside itself. First, we give a general definition of a compound graph.

3.1.1 Compound graph definition. A graph $G = (V, E)$ is defined by a finite set V of *nodes* and a finite set E of *edges*, which are ordered pairs (a, b, l) of nodes a, b with a label l . a is called the *source node* and b is called the *target node* of the edge, respectively. A *path* of length m between a node a_1 and a node a_m is a sequence of nodes a_1, a_2, \dots, a_m such that $(a_i, a_{i+1}, l_i) \in E$, for $i = 1, \dots, m - 1$.

A *rooted tree* $T = (V, E, r)$ is a graph such that for every node $a \in V$, except for the node r called the *root* of the tree, there is a unique path from r to a . For every node $a \in V$, except r , a unique node b such that $(b, a, l) \in E$ exists and is called the *parent* of a . For a node $a \in V$, the nodes in $\{b | (a, b, l) \in E\}$ are called the *children* of a . If

[†] A clique of a graph is a maximal complete subgraph. A complete graph has every pair of its nodes adjacent.

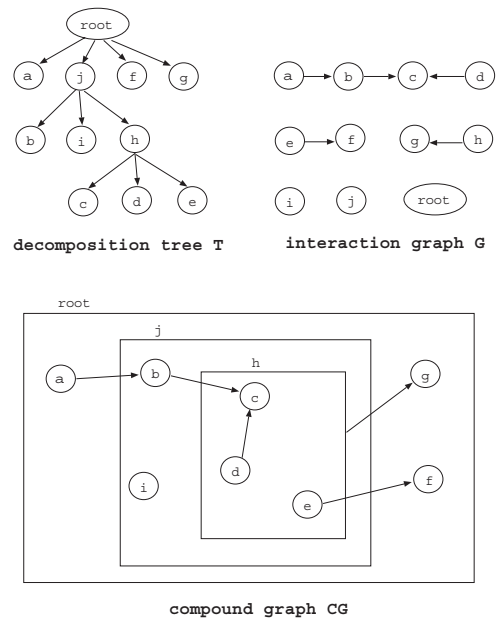


Fig. 2. An example of a compound graph.

node $a \in V$ does not have a child, a is called a *leaf* of the tree, and otherwise a is an *internal node*. The *ancestors* of a are nodes that are on the path from the root node r to a , with the exception of a itself.

A *compound graph* $CG = (G, T)$ is defined as the pairing of a graph $G = (V, E^G)$ and a rooted tree $T = (V, E^T, r)$ that share the same set of nodes. In this paper, we refer to the graph G of CG as an *interaction graph* and the tree T of CG as a *decomposition tree*. An edge $e_i^G \in E^G$ is called an *interaction edge* and an edge $e_i^T \in E^T$ is called a *decomposition edge*. A *fragment* $Frag(a)$ of CG is defined as a compound subgraph derived from the nodes of the subtree T' of T , rooted by the internal node a of T . Figure 2 is an example of a compound graph.

3.1.2 Signal transduction pathway model. A compound graph CG represents a signal transduction pathway SP. Then, a knowledge base of signal transduction pathways KB is defined by (S, R) , where S is a set of CGs and R is a set of rules to manipulate the data. Every biological entity that appears in the diagram is an individual node of CG . For example, a phosphorylated Smad3 and an unphosphorylated Smad3 are two distinct nodes in CG of the TGF- β signal transduction pathway. Smad3 in the cytosol and Smad3 accumulated in the nucleus are also represented as two different nodes. A local interaction with the name l in SP is an interaction edge of CG with the label l . The labels of interaction edges used throughout this paper are shown in Table 1. *phospho-*

Table 1. The type of edges

Type of edges	
Interaction	Decomposition
bind	has_component
release	has_state
modify	has_a
attach_modifier	default
phosphorylate	
detach_modifier	
dephosphorylate	
statechange	
translocate	
default	

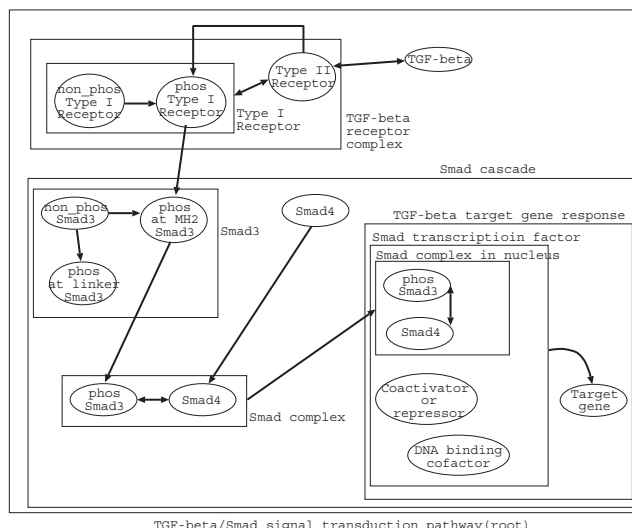
rylate is an *attach_modifier* edge and *attach_modifier* is a *modify* edge. *dephosphorylate* is a *detach_modifier* edge and *detach_modifier* is a *modify* edge. An edge $(a, b, statechange)$ means that state a has changed to b , i.e. activated or inactivated by phosphorylation, dephosphorylation, etc. For more detailed descriptions, one can use other available ontologies (bioontology, 2000). A regulatory function is an interaction edge of CG with the label *activation* or *inhibition*.

If a set of biological entities $ent_i (i = 1, \dots, n)$ can be integrated into a biological concept b (e.g. a protein complex, a downstream cascade) in SP, then b is a node in CG which has ent_i as its children in the decomposition tree T . An ontology for this *integration/decomposition relation* is shown in Table 1. A protein complex is decomposed into subunits with the *has_component* relation. If a protein has several modified forms, i.e. a non-phosphorylated form, a form phosphorylated at the C-terminal, a form phosphorylated at the linker region, etc., the protein is decomposed into these modified forms with the *has_state* relation. An integration/decomposition relation in SP with a name L is a decomposition edge in CG with the corresponding label. We refer to $Frag(b)$ as a *pathway fragment* b of SP. Thus a fragment can contain other fragments and its hierarchical structure is defined by the decomposition tree T . Note that in our model this hierarchical relation is not necessarily a hierarchy of 'is_a' relations that is in a concept tree.

Figure 3 shows the TGF- β /Smad pathway represented by a compound graph. Pathways represented by compound graphs can be easily integrated by adding a new root node as shown in Figure 4.

3.2 Implementing pathways

3.2.1 HiLog. A prototype knowledgebase is implemented on XSB ver.2.2 (xsb, 2000) that can process

**Fig. 3.** A compound graph of the TGF- β /Smad pathway.

a predicate logic called HiLog (Chen *et al.*, 1993)[‡]. A distinguishing feature of HiLog is that even though it has a higher-order syntax its semantics is first-order. HiLog allows arbitrary terms to be relation names so that terms can appear in places where predicates, functions, and atomic formulas occur in predicate calculus.

EXAMPLE.[§] A Prolog predicate calculating a transitive closure of edge can be written as follows:

```
path(X,Y) :- edge(X,Y).
path(X,Y) :- edge(X,Z), path(Z,Y).
```

With HiLog syntax, a more *generic* transitive closure predicate can be defined:

```
closure(R)(X,Y) :- R(X,Y).
closure(R)(X,Y) :-
    R(X,Z), closure(R)(Z,Y).
```

Here, *closure* is syntactically a second-order function which, given any relation R , returns its transitive closure $closure(R)$.

3.2.2 Example: the TGF- β /Smad pathway. A part of the pathway in Figure 3 (the TGF- β receptor complex) is shown in Figure 5 as a HiLog code. For simplicity, a node has only two attributes, i.e. its object ID and its name. Each interaction edge and decomposition edge is implemented as a predicate with the name of the corresponding edge label. For example, an interaction edge $(nid1, nid3, bind)$ is a predicate $bind(eid1, nid1, nid3)$ in a HiLog

[‡] Note that a user does not need to understand the HiLog code or any prolog syntax. Queries will be submitted through a Graphical User Interface (GUI) that presents directly the structure of the stored knowledge as shown in Figure 3 (although the GUI is not implemented yet).

[§] The example is borrowed from Chen *et al.* (1993).

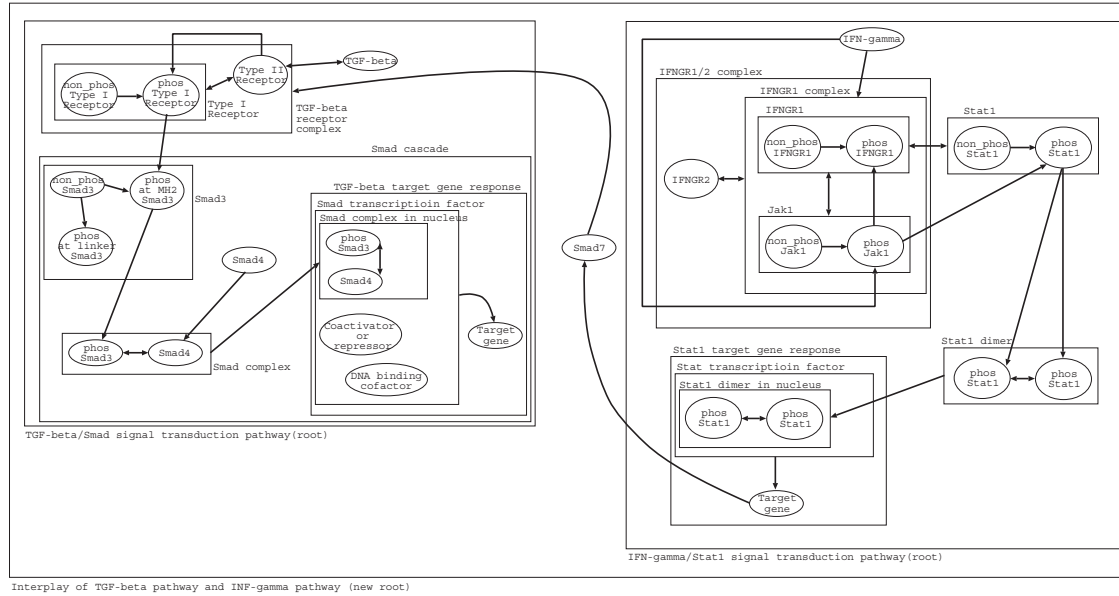


Fig. 4. Interplay of the TGF-beta/Smad pathway and IFN-gamma/Stat1 pathway.

program. The hierarchy of the edge ontology is defined as in Figure 6. `edge(bind)` means that `bind` is an edge relation. `modify(R) :- attach_modifier(R)` means that relation `R` is an `attach_modifier` relation and is a `modify` relation.

4 COMPUTATION

In this section, examples of queries are presented.

4.1 Path search

As mentioned before, pathways may be fragmented in a graph representation due to the lack of precise information of relations between every node.

In Figure 3 the path from ‘Smad complex’ to ‘Target gene’ is fragmented in a conventional representation because there are no edges between ‘Smad complex in nucleus’ and ‘Target genes’. But with the background knowledge that ‘Smad complex in nucleus’ is a member of ‘Smad transcription factor’ and that the transcription factor binds to ‘Target genes’, one can infer a path from ‘Smad complex’ to ‘Target gene’.

The compound graph model captures this background knowledge by its decomposition tree.

Thus one can infer the path by traversing the decomposition edges. Figure 7 is a path inference program. It checks if the target node b_1 and the source node a_2 of two interaction edges (a_1, b_1, l_1) and (a_2, b_2, l_2) are related by the decomposition tree. For instance, if a_2 is a child of b_1 it infers that there is a path `inf_path(a1, b2)` from a_1 to b_2 . By the relation `int_edge` and `dec_edge` we can

control the interaction edges and the decomposition edges that can be used for the inference. For example, by lines 6 and 7 only the connectivity between members of a node decomposed by `has_component` or `has_state` relation is inferred.

4.2 Search pathway functions

Suppose one wants to know the cellular function of a protein. An ontology or a gene catalogue with controlled vocabularies of cellular functions is able to answer the question with function names but will not provide the underlying molecular mechanism of the function. A graph of protein interaction will answer with its connected component that has the protein as its node, which is a transitive closure of possible biochemical interactions and does not necessarily represent a cellular function.

The compound graph model will answer the question with its fragment $Frag(x)$, where x is the parent of the protein in the decomposition tree (Figure 8). If the database curator gave the appropriate name to each node[‡], the pathway fragment will provide both the name of a function and the molecular mechanism of the function by a simple program:

```
cell_function(X) :- parent(Y,X),
                    list_fragment(Y) ..
```

`list_fragment(Y)` does a depth first search on the subtree T_Y of T rooted by Y and lists the nodes V in T_Y and

[‡] In this example we assume that a node name represents a function, but usually this information will be carried by an attribute of the predicate `node`, i.e. `node(nid, name, function)`.

```

% part of the interaction graph
node(nid1,tf_beta).
node(nid2,tf_beta_receptor_complex).
node(nid3,type2_receptor).
node(nid4,phos_type1_receptor).
node(nid5,nonphos_type1_receptor).
node(nid6,type1_receptor).
node(nid12,smad_cascade).

% part of the decomposition tree
default(cid1, nid0, nid2).
default(cid2, nid0, nid12).
default(cid3, nid0, nid1).
has_component(chid4, nid2, nid3).
has_component(chid5, nid2, nid6).
has_state(chid6, nid6, nid5).
has_state(chid7, nid6, nid4).

% part of the interaction edges
bind(eid1,nid1,nid3).
phosphorylate(eid2,nid3,nid4).
statechange(eid3,nid5,nid4).
bind(eid4,nid3,nid6).

```

Fig. 5. Pathway fragment of TGF- β receptor complex.

all interaction edges (a, b, l) , where $a, b \in V$. Figures 9 and 10 are the result for `?- cell_function(nid3)`.

4.3 Search conserved pathway structures among species

Searching conserved modules of signal transduction pathways would be of great importance. For example, it is suggested that a scaffold structure in MAP-kinase signaling of *Saccharomyces cerevisiae* is conserved in mammals (Whitmarsh and Davis, 1998).

With a graph representation, the problem can be modeled as a subgraph isomorphism determination, an NP-complete task (Garey and Johnson, 1979). However, compared to intermediary metabolism, signal transduction pathways are more divergent and may not be conserved ‘exactly’ as a graph structure. Hence, defining a submodule by a graph is a too-strict constraint in some cases.

Since our model represents biologically related objects not only by interaction edges but also by a decomposition tree, the constraints can be relaxed to the problem ‘find a pathway fragment that shares the same biological entities with the input structure’.

```

edge(bind).
edge(release).
edge(modify).
edge(R) :- modify(R).
         modify(attach_modifier).
         modify(R) :- attach_modifier(R).
         attach_modifier(phosphorylate).
         modify(breakbond).
         modify(R) :- detach_modifier(R).
         detach_modifier(dephosphorylate).
edge(statechange).
edge(translocate).
edge(default).

```

Fig. 6. Interaction edge ontology.

```

int_edge(statch).
int_edge(bind).
int_edge(mod).
int_edge(move).
int_edge(trans).

dec_edge(has_state).
dec_edge(has_component).

inf_path(X,Y) :- int_edge(R),R(_,X,Y).
inf_path(X,Y) :- dec_edge(R),R(_,X,Y).
inf_path(X,Y) :- int_edge(R),R(_,X,Z),
                 inf_path(Z,Y).
inf_path(X,Y) :- int_edge(R),R(_,X,Z),
                 dec_edge(S),S(_,W,Z),
                 inf_path(W,Y).
inf_path(X,Y) :- int_edge(R),R(_,X,Z),
                 dec_edge(S),S(_,Z,W),
                 inf_path(W,Y).
inf_path(X,Y) :- dec_edge(S),S(_,X,Z),
                 inf_path(Z,Y).
inf_path(X,Y) :- dec_edge(S),S(_,Z,X),
                 inf_path(Z,Y).
inf_path(X,Y) :- int_edge(R),R(_,X,Z),
                 dec_edge(S),
                 S(_,V,W),S(_,V,Z),
                 inf_path(W,Y).

```

Fig. 7. Path finding program.

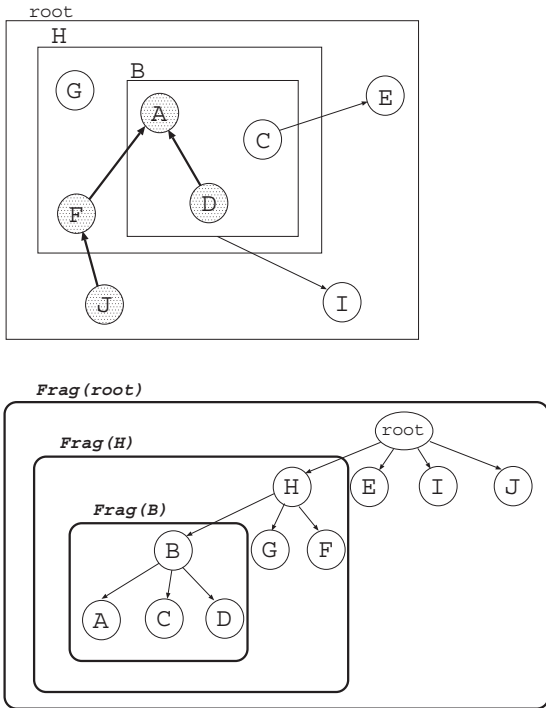


Fig. 8. Search pathway functions of protein A: in a graph of protein interactions, the answer will be the connected component (the shaded nodes), while in a compound graph the answer is the pathway fragments *Frag(B)*, *Frag(H)*, *Frag(root)*.

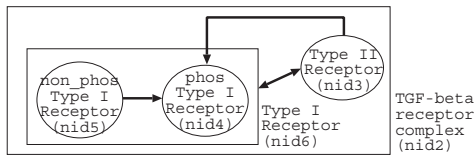


Fig. 9. TGF-*beta* receptor complex.

Figure 11 is an example. By executing the query `?- submod(['list of entities'])`, it searches the pathway fragment that includes the entities in 'list of entities'.

5 DISCUSSION AND FUTURE WORKS

In this paper we proposed a knowledge representation model for signal transduction pathway diagrams. By modeling a pathway as a compound graph, the method (1) prevents fragmentation of pathways caused by incomplete knowledge descriptions, and (2) handles the irregular knowledge granularity problem in diagrams; and, at the same time, (3) succeeds in not omitting the information about sectioning of pathways intended by the authors.

By implementing a deductive database with the pro-

```
% xsb
[xsb_configuration loaded]
[sysinitrc loaded]
[packaging loaded]

XSB Version 2.2 (Tsingtao) of April 20, 2000
[x86-pc-windows; mode: optimal;
engine: chat; scheduling: batched]
| ?- [pathways]
[pathways loaded]

yes
| ?- cell_function(nid3).
>>> root of fragment: nid2
      with name: tgf_beta_receptor_complex
>> decomposition_tree:
has_component(chid4,nid2,nid3),
has_component(chid5,nid2,nid6),
has_state(chid6,nid6,nid5),
has_state(chid7,nid6,nid4),
<<< end_of_the_tree

>> interaction_graph:
bind(eid4,nid3,nid6)
phosphorylate(eid2,nid3,nid4)
statechange(eid3,nid5,nid4)
<<< end_of_the_graph

yes
| ?-
```

Fig. 10. Result of `?- cell_function(nid3)`.

posed method, we demonstrated that several queries of biological importance could be realized with a set of simple rules.

Another good feature of a compound graph is its suitability for a Graphical User Interface (GUI) not only for pathway editing tools for data registration but also to visualize the results of pathway queries in a knowledge base. Owing to its ability to represent complex structures, several computer-aided creativity systems use the compound graph in their GUI systems (Misue, 1990), and several automatic compound graph drawing algorithms have been developed (Battista *et al.*, 1994).

Takai-Igarashi and Kaminuma (1999) show a signal transduction database and represent interaction data as

```

all_dec_edge(has_state).
all_dec_edge(has_component).
all_dec_edge(default).

anc(X,Y) :- all_dec_edge(R),R(_,X,Y).
anc(X,Z) :- all_dec_edge(R),R(_,X,Y), anc(Y,Z)

in_tree([C|CL],X) :- anc(X,C), in_tree(CL,X).
in_tree([],_).

submod([C|CL]) :-
anc(X,C), inctree(CL,X), node(X,XN),
print('>>> root of fragment: '), print(X),
print(' with name: '), print(XN), nl,
list_fragment(X).
submod([]).

```

Fig. 11. A submodule finding program.

binary relations. Thus its inference engine calculates only a transitive closure of the relations. Eilbeck *et al.* (1999) illustrate a database of protein–protein interactions that displays the interaction map as a graph with a straight-line drawing algorithm (Basalaj and Eilbeck, 1999). As discussed before, each node in signal transduction diagrams does not necessarily represent a single protein. Thus, at least for the signal transduction domain, pathways should be visualized according to their underlying structures. Our model is suitable for that purpose.

In this paper we have not mentioned causality or time representations. Information of timing is as important as that of localizations when inferring whether two biological entities interact in a certain biological process. However, it is not clear whether one can use a standard time scale: i.e. can we compare two events such as ‘The increase of A was maximal 15 min after addition of B’ and ‘The increase of C was maximal 20 min after addition of B’ in two different cells? As a consequence, the granularity of time representation will not be fine, as in, for example, ‘short term response’ and ‘long term response’. But even this limited information will help to control the inference process in a knowledge base.

Although our model defines the structure for a diagram, it does not provide a unique representation. For the same pathway diagram, several decomposition trees may exist. Thus some guidelines for data registration will be needed.

The naming of sections and submodules (the naming of pathway fragments) in a pathway depends on the authors, and no naming convention or ontology for such sectioning exists. The primary goal of the proposed method is to make the knowledge buried in articles accessible and has not assumed a well-defined language to describe the

domain. Therefore, knowledge base development will be an interactive process involving the construction of signal transduction pathway taxonomies.

We expect that the encoding of every pathway of every cell-line reported in articles into a knowledge base will provide a case-base for the molecular mechanism of each cellular function, which will compensate for the coarseness of cellular function name ontologies.

ACKNOWLEDGEMENTS

This work was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) ‘Genome Information Science’ from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), by Special Coordination Funds for Promoting Science and Technology, and by JSPS (the Japan Society for the Promotion of Science) Research Fellowships for Young Scientists.

REFERENCES

- Basalaj,W. and Eilbeck,K. (1999) Straight-line drawings of protein interactions—system demonstration. In Kratochvil,J. (ed.), *Graph Drawing, 7th International Symposium,GD'99*. Springer, pp. 259–266.
- Battista,G.D., Eades,P., Tamassia,R. and Tollis,I.G. (1994) Algorithms for drawing graphs: an annotated bibliography. *Comput. Geom.: Theor. Appl.*, **4**, 197–204.
- Bio-Ontology, (2000) The third annual Bio-Ontologies workshop (Bio-Ontologies 2000) <http://img.cs.man.ac.uk/stevens/workshop/>.
- Chen,W., Kifer,M. and Warren,D.S. (1993) HiLog: a foundation for higher-order logic programming. *J. Logic Program.*, **15**, 187–230.
- Eilbeck,K., Brass,A., Paton,N. and Hodgman,C. (1999) INTERACT: an object oriented protein–protein interaction database. In *Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 87–94.
- Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
- Enright,A.J., Iliopoulos,I., Kyripides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
- Flores,A., Briand,J.-F., Gadal,O., Andrau,J.-C., Rubbi,L., Mullem,V.V., Boschiero,C., Goussot,M., Marck,C., Carles,C., Thuriaux,P., Sentenac,A. and Werner,M. (1999) A protein–protein interaction map of yeast RNA polymerase III. *Proc. Natl Acad. Sci. USA*, **96**, 7815–7820.
- Fromont-Racine,M., Rain,J.-C. and Legrain,P. (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genet.*, **16**, 277–282.
- Garey,M.R. and Johnson,D.S. (1979) *Computers and Intractability*. Freeman, New York.
- Humphreys,K., Demetriou,G. and Gaizauskas,R. (2000) Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. In *Pacific*

- Symposium on Biocomputing 5*. pp. 502–513.
- Kanehisa, M. and Goto, S. (2000) Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
- Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. and Krummenacker, M. (1999) EcoCyc: electronic encyclopedia of *e.coli* genes and metabolism. *Nucleic Acids Res.*, **27**, 55–58.
- Karp, P.D. (2000) An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**, 269–285.
- Marcotte, E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, **10**, 359–365.
- Marcotte, E.M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
- McCarty, D.R. and Chory, J. (2000) Conservation and innovation in plant signaling pathways. *Cell*, 201–209.
- Misue, K. (1990) On abridgment of figures and its application to abduction support. *Inf. Process. Soc. Japan SIG-HI, HI-31*, **31**(31-1).
- Proux, D., Rechenmann, F. and Julliard, L. (2000) A practical information extraction strategy for gathering data on genetic interactions. In *Eighth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 183–189.
- Ray, L.B. (1999) The science of signal transduction. *Science*, **284**, 755–756.
- Rindfleisch, T.C., Tanabe, L., Weinstein, J.N. and Hunter, L. (2000) EDGAR: Extraction of Drugs, Genes and Relations from the biomedical literature. In *Pacific Symposium on Biocomputing 5*. World Scientific Publishing Co. Pte. Ltd., pp. 514–525.
- Shatkay, H., Edwards, S., Wilbur, W.J. and Boguski, M. (2000) Genes, themes and microarrays. In *Eighth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 317–328.
- Takai-Igarashi, T. and Kaminuma, T. (1999) A pathway finding system for the cell signaling networks database. *Silico Biol.*, **1**, 129–146.
- Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M. (2000) Automatic extraction of protein interactions from scientific abstracts. In *Pacific Symposium on Biocomputing 5*. pp. 538–549.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–631.
- Ulloa, L., Doody, J. and Massague, J. (1999) Inhibition of transforming growth factor- β /*smad* signalling by the interferon- γ /stat pathway. *Nature*, 710–713.
- Webb, E. (1992) *Enzyme Nomenclature, 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, London.
- Whitmarsh, A.J. and Davis, R.J. (1998) Structural organization of map-kinase signaling modules by scaffold proteins in yeast and mammals. *TIBS*, 481–485.
- xsb (April 21, 2000) xsb Version 2.2 <http://xsb.sourceforge.net/>.