ICOT Technical Memorandum: TM-1313

TM-1313

Toward Simulation-Like Representation of the Cell

by
M. Hirosawa (Kazusa DNA), R. Tanaka (IMS),
H. Tanaka, M. Akahoshi & M. Ishikawa

October, 1994

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191~5

Institute for New Generation Computer Technology

Toward Simulation-like Representation of the Cell

Makoto Hirosawa
Kazusa DNA Research Institute
1532-3 Yana-uchino, Kisarazu-shi,
Chiba 292 Japan
hirosawa@kazusa.or.jp

Reiko Tanaka
IMS
ICOT, Mita Kokusai Bldg. 21F,
1-4-28 Mita, Minato-ku Tokyo 108 Japan
ma-tanak@icot.or.jp

Hidetoshi Tanaka, Masayuki Akahoshi and Masato Ishikawa
ICOT
Mita Kokusai Bldg. 21F, 1-4-28 Mita, Minato-ku Tokyo 108 Japan
{ htanaka,akahoshi,ishikawa } @icot.or.jp

ABSTRACT

We developed a knowledge base system that describes, simulates and visualizes signal transduction in a cell. The system runs on Unix, and is mainly composed of the knowledge base module and the interface module. The interface module requires Motif. The knowledge base module in the system is described in micro-Quixote, a object-oriented database language executable on Unix.

Users can ask a question (e.g. What happens if an epidermis cell receives a stimulus such as EGF ?) through the interface module. The knowledge base module infers the result that occurred in the cell and the pathway between the stimulus and the result of the stimulus. Sometimes the pathway includes transcription of DNA and production of protein.

The interface module displays the inferred information in two levels of detail. Users can refer to related information about protein that appears in the displayed signal transduction pathway.

INTRODUCTION

Amazing advances in bio-technology now allow us to describe a range of phenomena that occurs in the human body by applying the language of genes or DNA. Genes encode the proteins that constitute the body. Proteins act not only as the building blocks of the body, but also serve to regulate it. These proteins are also coded in the genes. Consequently, a knowledge of genes and proteins is essential to understanding phenomena such as the immune system.

There is now enough accumulated information to describe and explain biological phenomena to some extent. To make the explanations possible, however, we must collect information from several sources, biological text books (Albert 1994), papers and databases(Bairoch 1992). These sources differ in their authors' interests and in their degree of abstraction of description.

Several researchers (Goto 1993) have attempted to integrate biological databases. Their works are a necessary step toward the description of biological phenomena. However, all have been interested mainly in integration of data and have paid a little attention to the representation of biological phenomena.

As a result, non-experts in biology find it difficult to integrate these information sources to adequately grasp biological phenomena. Sometimes, even experienced biologists fail to get an integrated view of biological phenomena, because many are only able to keep up with progress in their specialty.

It is important to express, in a knowledge base, the phenomena played by genes and proteins and to visualize these phenomena adequately. The use of visualization helps students of biology to understand biological phenomena in the body. Also, visualizing phenomena by referring to related databases facilitates research on genes and gives biologists further inspiration for new research.

In our previous research (Hirosawa *et al.* 1994), we studied representations of biological knowledge needed to describe biological phenomena, especially signal transduction pathways, and have developed a prototype knowledge base. In related research, (Karp 1994) studied representations of metabolic pathways using Lisp.

In our prototype knowledge base (Hirosawa *et al.* 1994) , we employed micro-Quixote, an object-oriented database language (Yokota *et al.* 1993) . From our experience with object-oriented knowledge bases (Hirosawa *et al.* 1993; Tanaka 1993), we had expected that an object-oriented knowledge base would be suitable for describing biological concepts. It proved to be a good choice.

Based on the previous study, we developed a prototype knowledge base system that displays signal transduction pathways inferred in the knowledge base, when some stimulus (eg. reception of EGF) is given to a cell. The system is executable on X11R5, and was programmed based on Motif.

In the next section, our system is presented and knowledge presentation in the system is described. Then, an example of execution with our system is presented. Finally, there is a discussion.

PROTOTYPE KNOWLEDGE BASE SYSTEM

In this section, we introduce our prototype knowledge base system. This consists of an overview followed by a description of the modules in the system. The system can predict what will happen if a specific kind of cell experiences some event (e.g. arrival of EGF).

Overview

An overview of our system is shown in Fig.1. The system is supported by any Unix machine. It is composed of four modules, the control module, the knowledge base module, the database module and the interface module. An explanation of these modules will be provided later.

Through the interface module a user can ask a question to the system (e.g. What will happen if a epidermis cell receives EGF?). The question is transferred to the knowledge base module through the control module. Then, the module infers (1) what signal transduction(including transcription of DNA and production of protein) will happen and (2) the final result of the signal transduction.

This information is transferred to the control module and the final result of signal transduction is displayed by the interface module. If information about the signal transduction pathway is necessary, it can be displayed. Furthermore, information on protein in the pathway can be displayed to users by consulting the database module.

The interface module

The interface module is coded in C, is executable on X11R5, and is programmed based on Motif. Through this module users can ask a question and the system can visualize the inferred signal transduction pathway in two levels of detail. Information on protein in the pathway can be displayed by consulting the database module. The real image of the interface will be shown in the next section, "Example of execution".

The database module

The database module has a protein dictionary. It is composed of lists of proteins and their accession number in the PIR. The dictionary is used when users request information on a protein through the interface module.

The knowledge module

The knowledge base module is written in the object oriented database language, micro-Quixote (Yokota *et al.* 1993) executable on UNIX. It is composed of an inference module and five knowledge bases, namely the intracellular process knowledge base, transcription knowledge base, cellular process knowledge base, intercellular process knowledge base and cell inheritance knowledge base. The inference module utilizes the deductive feature of micro-Quixote.

Intercellular process knowledge and cellular process knowledge are utilized to describe interaction between cells. We don't focus on interaction between cells here. For information about the two kinds of knowledge,

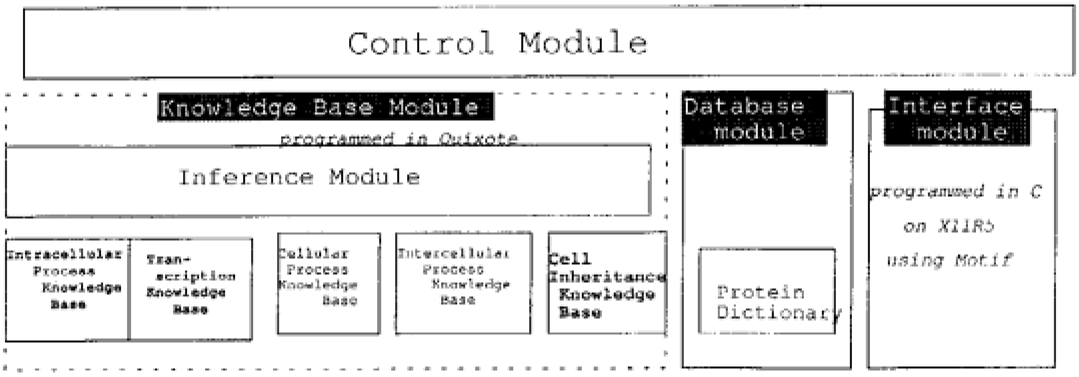

Control Module
Knowledge Base Module
programmed in Quixote
Inference Module
Intracellular Process Knowledge Base
Tran-scription Knowledge Base
Cellular Process Knowledge Base
Intercellular Process Knowledge Base
Cell Inheritance Knowledge Base
Database module
Protein Dictionary
Interface module
programmed in C on X11R5 using Motif

Fig.1

reader should refer to the previous papar (Hirosawa *et al.* 1994) .

Processes within one class of cell can be described using knowledge in the intracellular process knowledge base and transcription knowledge base. Because transcription knowldge is important information, we treat it separately from other information. And the hierarchical relationship between classes of cells is described by the cell inheritance knowledge base.

Transcription Knowledge Base

The entries in the transcription knowledge base describe which response element controls what gene. Two examples are shown below. The entries can be read as follows: SRE controls c-fos (1) and TRE controls G1 Cyclin (2).

```
gene[name="SRE",coded="c-fos"];;      (1)
gene[name="TRE",coded="G1cyclin"];;  (2)
```

Intracellular Process Knowledge Base

Each entry in the knowledge base represents individual processes inside a cell. For the sake of simplicity, only simple knowledge is shown.

```
receive[name="EGF"]/[result=increase[name="DG"]];;  (3)
increase[name="DG"]/[result=active[name="PKC"]];;   (4)
active[name="PKC"]/[result=active[name="SRF"]];;    (5)
```

"EGF", "DG", "PKC" and "SRF" are proteins or a sort of protein. The entries can be read as follows. If "EGF" is received, "DG" is increased (3). If "DG" is increased, "PKC" becomes active (4). If "PKC" is active, "SRF" becomes active (5). The description is done in the form "A/[result = B];;". It signifies that if A is satisfied, then B becomes true.

To understand the collective result of each process, let us assert knowledge " receive[name="EGF"]" to the knowledge base. If we ask whether "active[name="SRF"]" is true, the knowledge base answers "yes". To prove "active[name="SRF"]" the above three entries are used. An important point that must be noted here is that only individual processes are described in the knowledge base and that the proof of "active[name="SRF"]" is the result of a series of inferences. Here, the deductive feature of micro-Quixote is used.

Cell Inheritance Knowledge Base

The cell Inheritance knowledge base describes the hierarchical relationship between classes of cell. By means of the knowledge base, intracellular processes described in some class of cell are inherited by its lower class.

Two examples are shown in Fig.2. Knowledge (6) means that liver, muscle and fat cells are subclasses of the insulin_target_cell. Muscle has striated_muscle, smoothe_muscle and heart_muscle as its subclasses (Knowledge (7)). In this case, if intracellular processes (eg. intracellular knowldege (3)(4)(5)) collectively represented by *Stimulus*$-$ $>$ *Reaction* are assigned as intracellular processes of insulin_target_cell by the use of the module concept of micro-Quixote(Yokota *et al.* 1993) , these intracellular processes are inherited by their three subclasses. As a result we don't have to describe the intracelluar processes in the three subclasses thanks to the inheritance of object-oriented database language (in this case the module concept of micro-Quixote).

```
insulin_target_cell >=
                {liver,muscle,fat_cell};; (6)
muscle >=
{striated_muscle,smooth_muscle,heart_muscle};;(7)
```

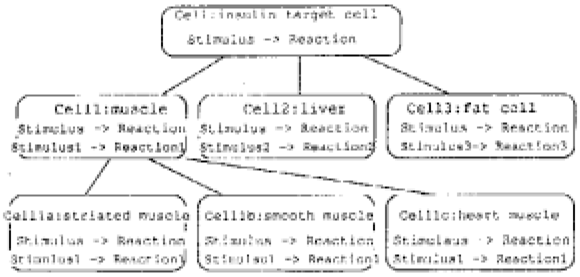


Fig.2

EXAMPLE OF EXECUTION

In this section, we will show an example execution of the knowledge base system(Fig.5). First, the system is invoked by the command named 'pathway'. Then, users specify a file name of the knowledge base. In this example, test2.qxt. is selected. In test2.ql, knowledge related to cell division cycle is stored.

Users can ask question by specifying it in question box. In this example, the question is what will happen if an epidermis cell receives EGF. Then, the knowledge base module infers processes that successively occur in the cell and the final result. The system shows a result in the question box, 'normal cell growth' in this case.

When users want to know a pathway between the reception of EGF and the normal cell growth, a rough pathway can be shown in the left side of the display by clicking 'Display' in the question box. In this example, users can see six big events, namely expression of c_fos, G1 cyclin and s-gene, then occurrence of s-period, G2-period and M-period.

If users want to know more detailed information between any two big events, users can obtain the information by specifying the two events using mouse. In this case, 'start'(= arrival of EGF) and 'express c-fos' are specified. Then, detailed pathway between the two big events is displayed in the right side of the rough pathway. The detailed events are (1)activation of G protein, (2)activation of PLC, (3)hydrolysis of PIP2, (4)production of DAG, (5)activation of PKC, (6)activation of SRF, (7)binding of SRF to SRE and (8)expression of c-fos. Here, activation of some entity is symbolized by a solid arrow. Other processes such as hydrolysis and production are symbolized by dashed arrows.

When users want to know information on any protein in the pathway, users can obtain the information by clicking the interested protein. In this version of the system, accession numbers of the protein in PIR are displayed. We plan to extend the system so that it can display PIR entries corresponding to the accession numbers

DISCUSSION

1. In our system, possible events within a cell(eg. if "EGF" is received, "DG" is increased) are stored in the knowledge base. If we want to know the event that occurs if some stimulus comes from outside the cell, and if we ask so, inducible events are successively calculated in the system. The series of events that occur after the stimulus are then displayed. This can be regarded as being a simulation of the phenomena that occur in the body. Through the simulation and visualized presentation, users can experience the biological processes happening in the body. Also, they can understand what proteins play a role in the phenomena. *The program of the system will be obtainable through ftp by the end of next March.*

   An important point that must be noted is that only individual processes are described in the knowledge base and that the result of normal cell growth is derived as a result of a series of inferences. Here, the deductive feature of micro-Quixote is used.

2. In this knowledge base, we utilize inheritance of the object-oriented database language (namely, the module concept of micro-Quixote) to effectively describe intracellular processes that happen when a cell receives some stimulus from outside. Because the intracellular processes of some class of cell are automatically inherited by its lower classes, there is no need to describe the relationship in those lower classes. This reduces the amount of knowledge necessary to describe biological phenomena.

3. We can describe the reaction of a cell to a stimulus from outside in two levels of abstraction. In one way, we can describe it by describing every intracellular process involved in producing the reaction from the stimulus to a cell. In the other way, we can describe it by describing only the stimulus to a cell and its reaction. So, we can see biological phenomena in different levels of abstraction. In this paper, we concentrated on the first type of description. However, the latter descriptive approach(higher abstraction) is also possible in our system with the use of the cellular process knowledge base. It is useful when users focus on intercellular interaction.

   Often, biologists want to understand biological phenomena in greater detail. This, for example, happens when they want to compare two biological processes. In such a case, the former approach is preferable.

   However, it happens that not every intracellular processes describing the describing the reaction to a stimulus is possible. In this case, biological phenomena can be described only in the latter way.

ACKNOLWLEDGEMENT

This research was done using mainly the facilities of ICOT when M.Hirosawa stayed at there. We thank Dr. Uchida of ICOT and Dr. Takanami of the Kazusa DNA Research Institute for establishing the research collaboration between ICOT and the Kazusa DNA Research Institute.

REFERENCE

Albert,B.*et al.* 1994." Molecular biology of cell." Garland Publishing, INC.

Bairoch,A. 1991. "Prosite : A dictionary of Protein site and pattern : User manual Release 7.00." May 1991.

Goto,S. *et al.* 1993. "A deductive language in object-oriented database for genome analysis." *Proceedings of International Symposium on Next Genaration Database Systems and Their Application* , pp123-129.

Hirosawa,M. *et al.* 1993. "Application of deductive object-oriented knowledge base to genetic unformation processing." *Proceedings of International Symposium on Next Generation Database Systems and Their Application* , pp116-122.

Hirosawa,M. *et al.* 1994. "Simulative representation of biological knowledge using object oriented language." *Proc. Genome Informatic Workshp, ,Yokohama, Japan, 1994.*

Tanaka,H. 1993. "A Private Knowledge Base for Molecular Biological Research." *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences,***Vol.1** pp803-812,1993.

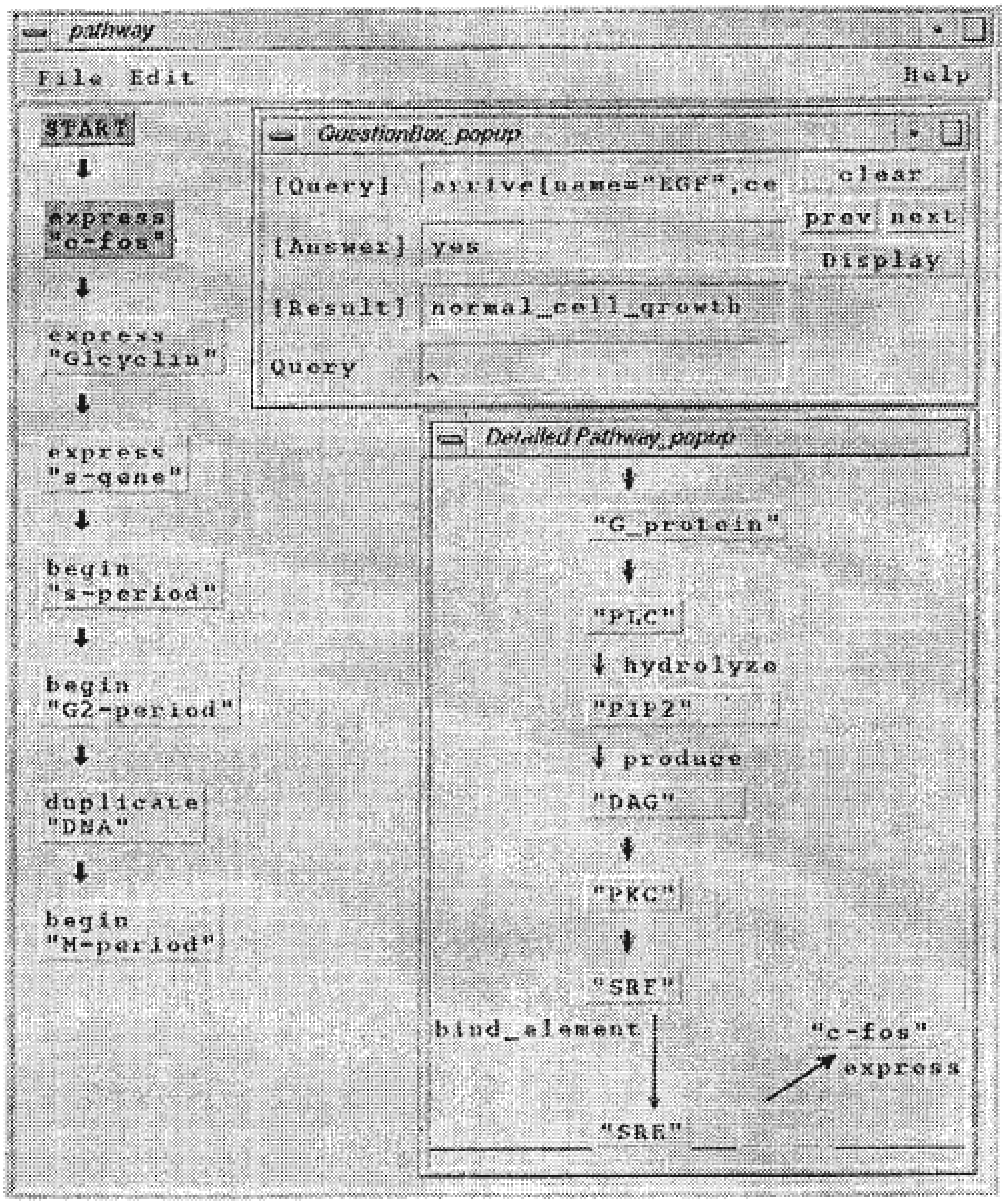
Karp,P. 1994. "Representations of Metabolic Knowledge : Pathway." *Proceedings of the Second International Conference on Intelligent System for Molecular Biology* pp203-211,1994.

Yokota,K *et al.* 1993. "Specific Features of a Deductive Object-Oriented Database Language Quixote." *Proc. ACM SIGMOD Workshop on Combining Declarative and Object-Oriented Databases, Washington DC, USA, May 29, 1993.*

BIOGRAPHY

Makoto HIROSAWA is a researcher in Department of Genome Informatics at KDRI(Kazusa DNA Research Institute) in Japan.

He was graduated from Science Univ. of Tokyo and recieved master degree of physics from Tokyo Univ. After research a period at Hitachi Ltd., He had studied knowledge engineering, then the application of knowledge enginnering to biology at ICOT(Research center of japanese 5th generation computer project). He returned to Hitachi Ltd. and studied communication network in a short period. However, he came to realize that bio-informatics was his Promised Land, and he joined KDRI.



[Query] arrive[name="EGF",ce

[Answer] yes

[Result] normal_cell_growth

Query ^

Fig.5