ICOT Technical Memorandum: TM-1315

TM-1315

Simulative Representation of Biological Knowledge using Object-oriented Database Language

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

Simulative representation of biological knowledge using object-oriented database language

Makoto Hirosawa Reiko Tanaka
hirosawa@icot.or.jp ma-tanak@icot.or.jp
Hidetoshi Tanaka Masayuki Akahoshi Masato Ishikawa
htanaka@icot.or.jp akahoshi@icot.or.jp ishikawa@icot.or.jp

[1] Kazusa DNA Research Institute
ICOT, Mita Kokusai Bldg. 21F, 1-4-28 Mita, Minato-ku Tokyo 108 Japan

[2] IMS
ICOT, Mita Kokusai Bldg. 21F, 1-4-28 Mita, Minato-ku Tokyo 108 Japan

[3] ICOT
Mita Kokusai Bldg. 21F, 1-4-28 Mita, Minato-ku Tokyo 108 Japan

Abstract

Marvelous advance in bio-technology enable us to describe various phenomena played in our body with the language of genes and proteins. It is important to express these phenomena in knowledge base, and to visualize these phenomena properly. The visualization of the phenomena with reference to related databases facilitates researches on genes.

As the first step to realize a database like the one stated above, we have studied representation of biological knowledge needed to describe biological phenomena and have developed prototype knowledge base. The knowledge base is described in Quixote, object-oriented database language excutable on Unix. The knowledge base can covers the knowledge related to signal transduction within a cell and those related to transcription of genes.

In our prototype system, a sort of simulation can be done. With the arrival of signaling ligand to surface of cell, proteins along suitable pathways are activated in our simulated cell. As consequence of series of activation(a chain of inference), some biological response are deduced and shown to users.

1 Introduction

Amazing advances in bio-technology now allow us to describe a range of phenomena that occur in the human body by applying 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.

Until now, information describing biological phenomena has been accumulating to the extent that it explains 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 meshes of description. Recently, 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 not a few biologists 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 our body. Also, visualizing phenomena by referencing to related databases facilitates research on genes and gives biologists further inspiration for new research.

As a first step toward a knowledge base like that described stated above, we have studied representations of biological knowledge needed to describe biological phenomena and have developed a prototype knowledge base. From our experience with object-oriented knowledge bases[Hirosawa 1993; Tanaka 1993], we thought that an object-oriented knowledge base would be suitable for describing biological concepts. We decided to employ Quixote, an object-oriented database language [Yasukawa 1992], in this project. We expected, through the use of the object-oriented language, that concepts such as cell could be described naturally.

In the next section, processes occuring in our body are explained, prior to our explaining our knowledge base system. Both intracellular and intercellular processes are explained. Then, our prototype knowledge base, which represents the processes inside as well as outside a cell, are presented. Finally, an example of execution with our system is presented.

2 What happens in our body

Because our body is composed of cells, if cells and the interaction between cells can be described properly, it should be possible to describe the biological phenomena in our body. This section explains the intracellular and intercellular processes that occur in our body.

2.1 Intercellular process

Cells in our body communicate with each other. Cells secrete hormones such as insulin and growth factors such as EGF. When cells receive some chemical substance, be it a hormone or growth factor, they exhibit some reaction or no reaction, depending on the species of the cells. When some reaction happens, some intracelluar processes happen.

Fig. 1 is a simple example of an intercellular process. The figure shows the intercellular process between insulin secretory cell (C1), such as beta cell of pancreas, and insulin target cell (C2), such as muscle cell.

Insulin secretory cell secretes insulin(Reaction R1) when it receives glucose(Stimulus to cell S1). An insulin target cell intakes glucose (Reaction R2) when it receives insulin (Stimulus named S2). When a hormone such as insulin is secreted, it spreads thoroughout our body. So, the insulin secreted by insulin target cell travels around our body to arrive at insulin target cell. Then, the insulin target cell receives insulin to intake the glucose around it. To sum up, stimulus (reception of insulin) to insulin secretory cell (S1) results in a some reaction (intake of insulin) by the insulin target cell (R2).

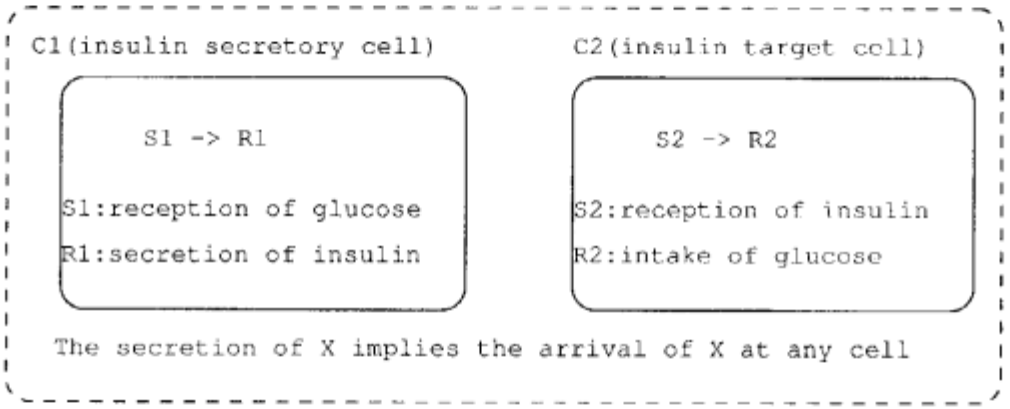

Fig. 1

2.2 Intracellular process

As mentioned in the previous subsection, an cellular process can be simply described as a stimulus-reaction pair ($S1$(reception of glucose)$- > R1$(secretion of insulin) in C1). However, many intracellular processes exist between S1 and R1. Biologists often want to determine and understand the actual intercellular process.

Fig. 2 is an example of an intracellular process. To describe the intracellular processes, two kinds of knowledge, transcription knowledge and intracellular process knowledge, are necessary. We will explain the significance of these two kinds of knowledge later.

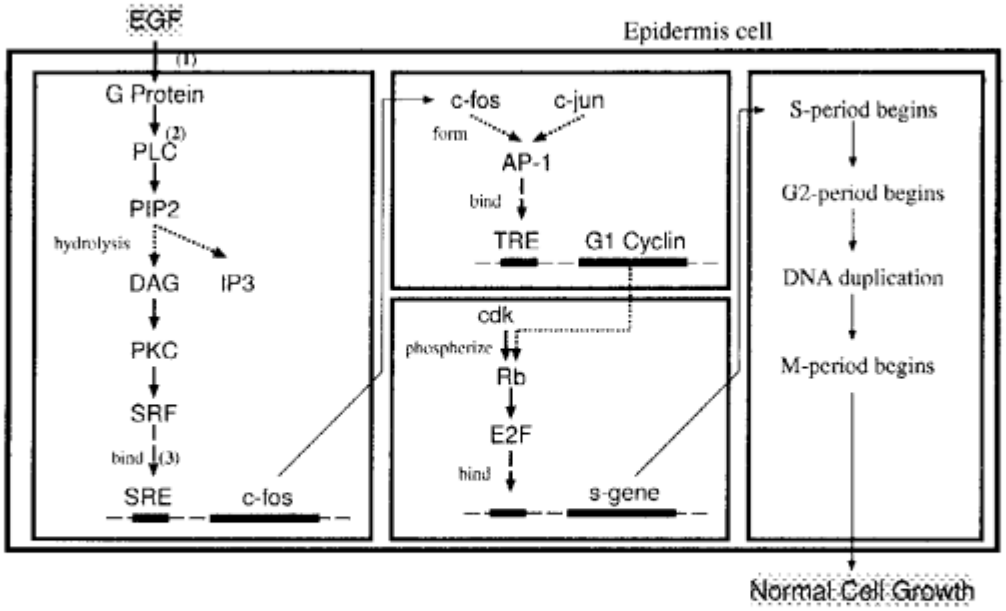

Fig. 2

In this case, the stimulus to the cell is the reception of EGF. The corresponding reaction by the cell is normal cell growth. EGF(Epidermis Growth Factor) belongs to the growth factor family. When a growth factor is received by some species of cell, the cell is duplicated as the result of intracellular process. In the case of EGF, the epidermis cell duplicates upon receiving EGF.

In the figure, a solid arrow from one entity to another indicates the activation of the latter by the former. For example, the reception of EGF activates G protein (1), then the activation of G protein activates PLC (2). A dashed arrow signifies the process indicated beside the arrow. For example, the activation of SRF enables SRF to bind to SRE(3).

When EGF arrives at the epidermis, after a series of processes, SRF binds to SRE. Then, the c-fos gene, controlled by SRE, is expressed and c-fos is produced. The produced c-fos and c-jun form AP-1 to bind to TRE. Then, the G1 Cyclin gene, controlled by TRE, is expressed and G1 Cyclin is produced (SRE and TRE are refered to as response elements) . After a series of processes, the s-gene is produced. After S-period, G2-period begins to duplicate DNA. Finally, normal cell growth occurs after M-period.

The knowledge necessary to describe an intracellular process like the above can be classified into two classes. One class relates to which response element controls which gene (SRE controls c-fos, TRE controls G1 Cyclin, and RE(s-gene) controls s-gene). We named such knowledge "transcription knowledge". The other class named "intracellular process knowledge" is the knowledge other than transcription knowledge. For example, knowledge (1) (reception of EGF activates G protein), knowledge(2) and knowledge(3) are intracellular process knowledge.

3 Prototype knowledge base

In this section, we explain our prototype knowledge base. An overview of our system is shown in Fig.3. The system is supported by any Unix machine.

The system is divided into two modules, one being the knowledge base module and the other the display module. Currently, only the knowledge base module has been finished. The display module is under developement. Once the display module has been completed, the result of inference performed by the knowledge base module will be transferred to the display module and the result, like that shown in Fig.2, will be displayed using the GUI. The display module is coded in C, is executable on X11R5, and is programmed based on Motif. The knowledge base moduleis written in Object oriented database language, Quixote[Yasukawa 1992] excutable 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 interitance knowledge base. The inference module utilizes the deductive feature of Quixote.

These knowledge bases are explained below.

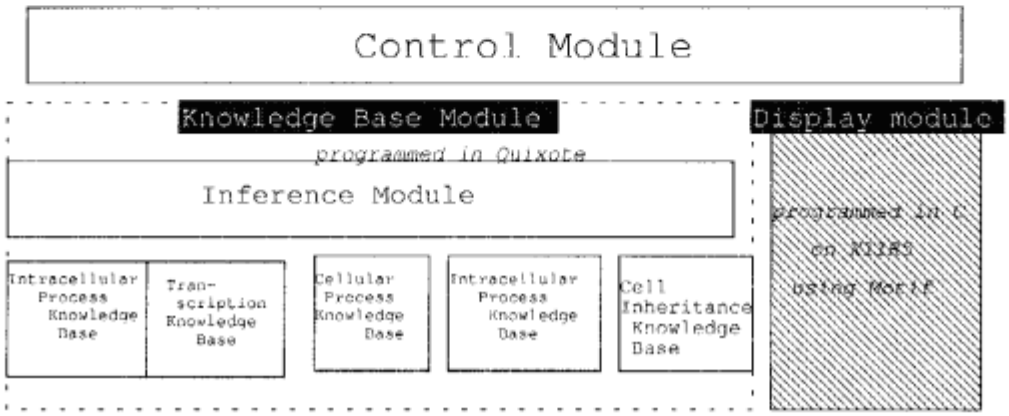

Fig.3

Intracellular Process Knowledge Base

The following is a portion of the intracellular process knowledge base. Processes within a cell can be described using knowledge in the intracellular process knowledge base and transcription knowledge base. Each entry in the intracellular 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"]];;   (1)
increase[name="DG"]/[result=active[name="PKC"]];;    (2)
active[name="PKC"]/[result=active[name="SRF"]];;     (3)
```

"EGF", "DG", "PKC" and "SRF" are proteins or a sort of protein. The knowledge can be read as follows. If "EGF" is received, "DG" is increased (1). If "DG" is increased, "PKC" becomes active (2). If "PKC" is active, "SRF" becomes active (3). The description is done in the form "A/[result = B];;". It signifies that if A is satisfied, then B becomes true. In this way, individual processes within a cell can be described.

In the three knowledge (1)(2)(3), 'receive', 'increase' and 'active' are objects in Quixote. However, we could choose "EGF","DG","PKC" and "SRF", which correspond to entities in

the entity-relationship model. Database researchers might regard the latter as being the most suitable choice. But, when we describe intracellular processes, it is essential that we describe possible events. We selected 'receive', 'increase' and 'active' as objects because possible events can be described using these words.

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 knowledge 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 Quixote is used.

Transcription Knowledge Base

The entries in the transcription knowledge base describe which response element controls what gene. Two examples are shown below. The knowledge can be read as follows. SRE controls c-fos (4) and TRE controls G1 Cyclin (5).

```
gene[name="SRE",coded="c-fos"];;     (4)
gene[name="TRE",coded="G1cyclin"];; (5)
```

Cellular Process Knowledge Base

Processes within a cell can be described using knowledge in the intracellular process knowledge base and transcription knowledge base. Then, the reaction of a cell to a specified stimulus can be described theoretically with the two knowledge bases. However, it is often the case that some intracellular processes, necessary to deduce some reaction to a specified stimulus, have not yet been discovered. In such a case, case, knowledge that directly relates stimulus to reaction is neccesary. Cellular process knowledge base describes the stimulus-reaction relationship.

Two examples are shown below. Knowledge can be read as follows. When IL-4 and IL-5 arrive at B cell, B cell receives IL-4 and IL-5 (6). When B cell receives IL-4 and IL-5, it secretes IgE (7). *If it did not have knowledge (6), B cell would not be able to secrete IgE, even when IL-4 and IL-5 arrive at B cell.*

```
arrive[name1="IL4",name2="IL5",cell="B cell"]/
  [ result= receive[name1="IL4",name2="IL5",cell="B cell"]];; (6)
receive[name1="IL4",name2="IL5",cell="B cell"]
   /[result=secrete[name="IgE",cell="B cell"] ];; (7)
```

Intercellular Process Knowledge Base

The intracellular process knowledge base describes the interaction between cells. There are many species of interaction. However, the prototype knowldege base has only one entry for the intercellular process knowledge base. The knowldege (8) shown below is very important and it can be applied to a large portion of intracellular processes. This means that chemical substance P, secreted by some cell named C1, can arrive to any cell named C2.

```
secrete[name=P, cell=C1]/
     [ result= arrive[name=P,cell=C2]];; (8)
```

Cell Inheritance Knowledge Base

The cell Inheritance knowledge base describes the hierarchical relationship between classes of cell. By means of the knowledge base, the stimulus-reaction relationship described in some class of cell is inherited to its lower class.

Two examples are shown in Fig.4. Knowledge (9) means that liver, muscle and fat cells are subclasses of the insulin_target_cell. Muscle has muscle1, 2 and 3 as its subclasses. In this case, stimulus-reaction relationship $S2- > R2$, described in insulin_target_cell, is inherited by its three subsclass. As a result we don't have to describe the relationship in the three subclasses thanks to the inheritance of the object-oriented database language.

```
insulin_target_cell >= {liver,muscle,fat_cell};; (9)
muscle >= {muscle1,muscle2,muscle3};;            (10)
```

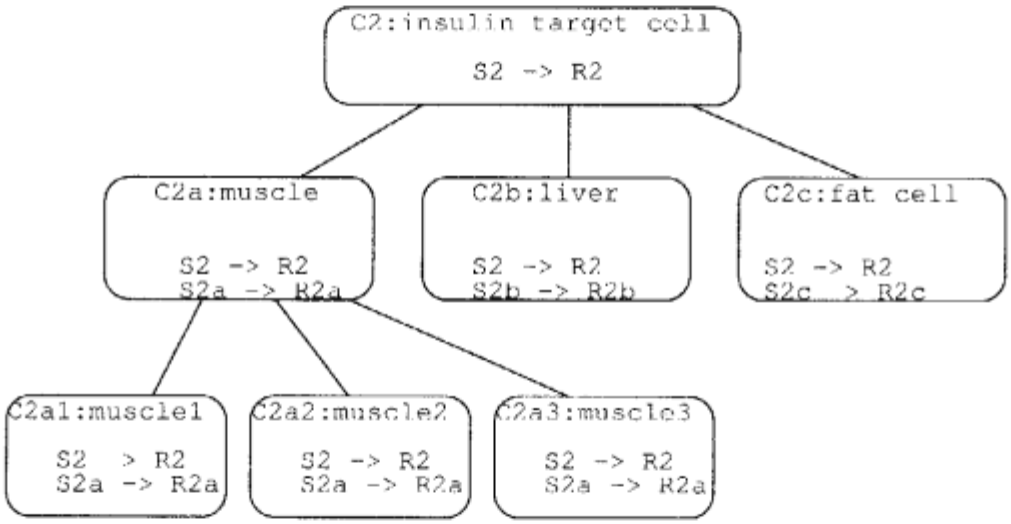

Fig.4

4 Example of execution

In this section, we give an example of execution of our knowledge base. In this case, "what happens if EGF arrives at a epidermis cell" is asked (a0). The knowledge base answers that normal cell growth occurs (a1). The knowledge base also indicates which intercellular processes occur between the reception of EGF (a2), and normal cell growth (a28). The intracellular process is almost the same as that explained as an example of an intracellular process using Fig.2.

Among a series of processes(a2 $-$ > a28), there are three processes corresponding to the gene expression for producing protein(a11,a15,a19). The three processes are expression of c-fos,

G1 cyclin and s-gene. The three processes are derived from the transcription knowledge base. Other processes are derived from the intracellular process.

After three gene expressions, the cell enters the s-period (a20) then the G2-period (a23). After cell duplication (a24), the cell enters the G2-period (a27) to complete normal cell growth (a28).

```
?- arrive[name="EGF",cell=epidermis]/[result=Result,path=Path].          (a0)
?- yes.
Result = normal_cell_growth,                                            (a1)
Path = state[now= receive[name="EGF",cell=epidermis],                   (a2)

  next=state[now= active[name="G protein",cell=epidermis],              (a3)
  next=state[now= active[name="PLC",cell=epidermis],                    (a4)
  next=state[now= hydrolysis[name="PIP2",cell=epidermis],               (a5)
  next=state[now= produce[name=list[element1="DAG",element2="IP3"],
                                                  cell=epidermis],(a6)
  next=state[now= increase[name="DAG",cell=epidermis],                  (a7)
  next=state[now= active[name="PKC",cell=epidermis],                    (a8)
  next=state[now= active[name="SRF",cell=epidermis],                    (a9)
  next=state[now= bind_element[name="SRE",cell=epidermis],              (a10)
  next=state[now= express[name="c-fos",cell=epidermis],                 (a11)

  next=state[now= form[name="c-jun",with="c-fos",cell=epidermis],       (a12)
  next=state[now= produce[name="AP-1",cell=epidermis],                  (a13)
  next=state[now= bind_element[name="TRE",cell=epidermis],              (a14)
  next=state[now= express[name="G1cyclin",cell=epidermis],              (a15)

  next=state[now= active[name=cdk,with="G1cyclin",cell=epidermis],      (a16)
  next=state[now= phosphorize[name="Rb",cell=epidermis],                (a17)
  next=state[now= active[name="E2F",cell=epidermis],                    (a18)
  next=state[now= expresse[name="s-gene",cell=epidermis],               (a19)

  next=state[now= begin[name="s-period",cell=epidermis],                (a20)

  next=state[now= phosphorize[name="Cdc2 kinase",cell=epidermis],       (a21)
  next=state[now= bind[name="Cdc2 kinase",to=cyclinB,cell=epidermis],   (a22)

  next=state[now= begin[name="G2-period",cell=epidermis],               (a23)
  next=state[now= duplicate[cell=epidermis,object="DNA"],,              (a24)
  next=state[now= active[name="Cdc25 phosphatase",cell=epidermis],,     (a25)
  next=state[now= active[name="Cdc2 kinase",cell=epidermis],,           (a26)

  next=state[now= begin[name="M-period",cell=epidermis],,               (a27)

  next=normal_cell_growth]]]]]]]]]]]]]]]]]]]]]]]]]]]]                  (a28)
```

5 Discussion

1. So far, a little attention has been paid to the representation and visualization of biological phenomena. 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, inducable events are successively calculated in the system. The series of events that occur after the stimulus are then shown to the user. This can be regarded as being a simulation of the phenomena that occur in the body. Through the simulation, users can experience the biological processes happening in the body. Also, they can understand what proteins play a role in the phenomena.

   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 Quixote is used.

2. In this knowledge base, we utilize inheritance of the object-oriented database language to effectively describe reaction of a cell to a stimulus received from outside. Because the reaction to stimulus relationship 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 ways. In one way, we can describe it by describing every intracellular process involved in producing the reaction from the stimulus. In the other way, we can describe it by describing only the stimulus reaction relationship. So, we can see a biological phenomena in different levels of detail.

   Often, biologists want to understand biological phenomena in greater detail. This, for example, happens when they want to compare two biological processes. In this case, the former way of description is preferable.

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

4. To make the system more powerful, we believe that knowledge to explain protein, eg. EGF, is neccessay. We are now investigating how to construct such a knowledge base and how to connect our system to existing databases.

5. In the present stage of our research, the simulation result is output as language. With display module in Fig.3, under development, The system can visualize the simulation result. With the graphic interface, users can view simulation results in several levels of detail.

   Visualizating the simulation result will enhances our understanding of biological phenomena in the body, with the comparison of similar processes giving biologists a deeper insight

into biologial phenomena. We believe that our system will contribute to the progress of biology and medicine.

Acknowledgement

This research has been done using facility of ICOT when M.Hirosawa stays at ICOT. We thank Dr. S.Uchida of ICOT and Dr. Takanami of the Kazusa DNA Research Institute for establishing research collaboration between ICOT and the Kazusa DNA Research Institute.

Reference

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

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

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

[**IITakaka 1993**] Tanaka, H. 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.

[**Yasukawa** *et al.* **1992**] Yasukawa, H *et al.* Objects, Properties, and Modules in Quixote. *Proc. of FGCS92, pp89-112, 1992*