TM-1279

Application of Deductive Object-Oriented Knowledge Base to Genetic Information Processing

by
M. Hirosawa (Hitachi), R. Tanaka (IMS),
& M. Ishikawa

© Copyright 1993-08-30 ICOT, JAPAN ALL RIGHTS RESERVED

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

Application of Deductive Object-Oriented Knowledge Base to Genetic Information Processing

Makoto Hirosawa *
ICOT

Reiko Tanaka
IMS

Masato Ishikawa
ICOT

Abstract

We developed a prototype system for analyzing the genes of living organisms. The system is composed of a rule base and biological knowledge base for the analysis. The system is written in deductive object-oriented database language QUIXOTE .

We applied the system to a simple but practical problem, motif finding problem. Analysis of the results showed that the system can analyze genes by utilizing the characteristics of the deductive object-oriented database language.

1 Introduction

The analysis of the genes of livings organisms is essential to deciphering of biological phenomena. Today, with major advances in bio-technology, the number of genes that have already been identified and must be to be analyzed are increasing rapidly. To enable the automatic analysis of such massive genes and to extract biological information from them, the introduction of knowledge-based analysis is necessary, in addition to the development of high-speed computers and fast analytical algorithm. This is partly because the quality of analysis without biological knowledge is not high enough, and partly because the time required for the analysis can be considerably reduced with the introduction of biological knowledge.

Nowadays, data about genetic information is stored in databases. But these databases have not been constructed with consideration of their use for knowledge engineering. So, it is hard to create a system that enables high-level knowledge processing. To enable such a sytem, an effective representation of biological knowledge is necessary.

The representation of biological knowledge has been studied by several researchers[1][2]. They succeeded in representing of simple and small amounts of knowledge. But, the representation can not necessarily be extended to the representation of practical problems.

We selected motif discovery as a practical problem in our study of representation of biological knowledge. There were two reasons for our selection. One is that the level of complexity of biological knowledge about motif is adequate for the study of the representation of biological knowledge. Its complexity is low enough for non-specialists of biology to study its representation, thus preventing us from being overwhelmed by the biology. Nevertheless, its complexity is great enough to include essential biological concepts. We thought that the understanding gained through the study would also help us in the representation of other kinds of knowledge.

Another reason for our chice was our familiarity with the motif discovery problem. We had already developed a motif discovery system [3]. But, the representation of that system was not sophisticated one. Logic programming language, KL1, in which the system itself was written, was used to express biological knowledge. But, because KL1 is essentially a programming language rather than a knowledge representation language, its effectiveness of representation is limited.

This time we selected DOOD (Deductive Object-Oriented Database) language *QUIXOTE* [4] to represent biological knowledge . We thought that the Object-Oriented features of *QUIXOTE* would be useful for the representation. And the suitabily of the interface (deductive feature) to inference module of the system was also a major consideration.

2 What is genetic analysis

The analysis of the genes of living organisms is not only an essential technique for the analysis of living organisms but also for research of the mechanism of cancer in medicine and the development of a treatment of AIDS in virology. When a new gene is discovered, genetic analysis predicts the function that that gene plays in our body.

The genes of a living organism are stored in the nucleus of a cell, and are coded in DNA. When necessary, some part of the DNA is transcribed into mRNA, which is transformed into a chain of amino acids. When the chain takes the correct form, it can function as a protein.

Twenty kinds of amino acids are used as a unit of the chain. Conventionally a single alphabetic character is used to represent each kind of amino acids. Therefore, a protein can be represented as an alphabetic sequence. For example, "GIVE-

*Currently with System Development Laboratory, Hitachi Ltd.

```
seq1:KLGQFGEVWMGTWNGTTRVAIKTLKPGTMSPE
seq2:IGFGVYRGTLRLPSQDCKTVAIKTLKDTSPGGQWW
seq3:SQGGFGMVYEGNARDIIKGETRVAIKTVNESASLRERI
seq4:GQGAFVTVYKGLWIPEGEKIPVAIKTLREATSPKANK

seq1:KLGQFGEVWMGTWNGTTR-------VAIKTLKPGTMSPE--
seq2:-IG-FG-VYRGTLRLPSQ---DCKTVAIKTLKDTSPGGQWW
seq3:--SQGGFGMVYEGNARDIIKGET-RVAIKTVNESASLRERI
seq4:--GQGAFVTVYKGLWIPEGE-KI-PVAIKTLREATSPKANK
      ######                  VAIKT
```

Figure 1

QCCTSICSLYQLENYCN" is an amino acid sequence that signifies part of insulin, a kind of hormone. In the sequence, G stands for glysine, I stands for isoleusine and so on.

In the representation scheme, genetic analysis is equivalent to the prediction of the function of the gene from its amino acid sequence represented as an alphabetic sequence.

In the analysis of genes, motif is important information. The motif is a sequence pattern that groups of sequences have in common. In some cases, the motif is known as the common sequence pattern of a restricted class of enzyme (ex. transferase). When an amino acid sequence is discovered and it contains some motif which is attributed to enzyme A, it can be predicted that the sequence codes an enzyme that is similar to enzyme A. In another case, the motif is known as a common sequence pattern of a module of a protein that has a definite function (ex. the DNA binding function). If its subsequence corresponding to the motif is modified, the protein loses the function of the motif (It becomes unable to bind DNA).

Generally, *Multiple alignment* is used to extract motifs from group of sequences. For example, suppose we try to find motifs from the sequences in Figure 1 (top). To find motifs, patterns common through all sequences must be matched. This matching procedure, and also the result, are called multiple alignment (Figure 1 (bottom)). In the alignment, to align the same or similar amino acids in the same column, '-' is inserted in the appropriate locations. As a measure of the similarity between amino acids, the Dayhoff Matrix [5] is commonly used. In the figure, two motifs are identified. One is marked by '#', the other by 'VAIKT'. The latter is a completely common pattern, whereas the former is not. The former, however, can be regarded as a motif as well.

Multiple alignment is necessary to discover motifs. But to date, within a practical amount of time, no algorithm has been able to produce automatically high-quality multiple alignment without using knowledge. It means system which can discover high-quality motif is also hard in this line. We developed a multiple alignment and motif discovery system with knowledge engineering as part of our previous project [3]. The present research expands on this, and also in the line.

3 Motif Discovery System

The system discovers motifs from input sequences of protein. When it makes candidates of motif, it consults its biological knowledge. The consultation makes the quality of discovered motif higher and reduces the time required. The system also predicts the function of the protein. The present and previous systems are the same except in their biological knowledge.

The structure of the system is shown in Figure 2. In the followings, its structure and its method of operation are explained.

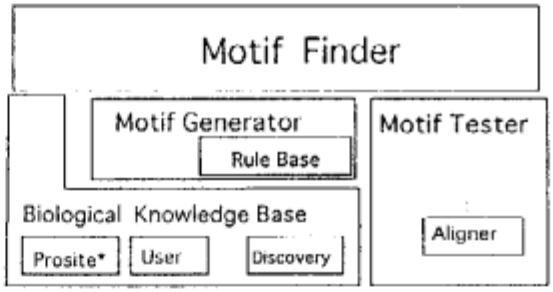

Figure 2

3.1 Overview

Aligner roughly matches similar sequences using an alignment algorithm [6] to produce multiple alignment. *Motif Finder* analyzes the multiple alignment and detects common sequence patterns. Then, it sends that information to *Motif Generator*. In the example shown in Figure 1, the system detects two common patterns. The one is marked by 'VAIKT', and the other by '#'.

Motif Generator generates candidate motifs from those offer biologically high reliability. The reliable motif candidates are those that are generated based on the biological knowledge stored in the *Prosite** and *User* modules in the *Biological*

Knowledge Base. And less reliable motif candidates are those that are generated based on purely biological statistics.

More concretely, Motif Generator fires rules stored in the *Motif Rule Base*, according to the priority assigned to each rule. And, it generates motif candidates. Presently, ten rules are registered in the Motif Rule Base. Among the rules, those that will be used in the section, *Example of Application*, are shown in Figure 3. Among them, the priority of Rule 1 is the highest and that of Rule 4 is the lowest. The rules consult Prosite* and/or User, if necessary. these rules are described in detail in [3].

Rule 1: **IF** Discovered motif M is similar to the motif m_i which is registered as motif of protein p in Prosite*
Then makes motif $m_j (j \neq i)$, which is also registered as motif of protein p in Prosite*, as motif candidate.

Rule 2: **IF** Discovered motif M is similar to the motif m_i which is registered as motif of protein p in Prosite*
Then makes motif $m_j (j \neq i)$, which is registered as motif of protein $p1$ being subclass of p in Prosite*, as motif candidate.

Rule 3: **IF** Discovered motif M is similar to the motif m_i which is registered as motif of protein p in Prosite*
Then makes motif $m_j (j \neq i)$, which is registered as motif of protein p in User, as motif candidate.

Rule 4: **IF** true
Then makes pattern of subsequence in multiple alignment, whose similarity is the highest, as motif candidate.

Figure 3

Motif Tester makes multiple a alignment of sequences with the constraint that matches the subsequences corresponding to the motif candidates, then checks whether it satisfies a biological statistical criterion [5]. If the criterion is satisfied, the motif candidate is confirmed as being a motif and the motif is registered to the *Discovery* module in the Biological Knowledge Base. Then, the multiple alignment is sent to the Motif Finder to initiate a new cycle of motif discovery.

4 Biological Knowledge Base

In the first subsection, the requirements for biological knowledge base are described and solutions satisfying them, using Deductive Object-Oriented Knowledge base, are briefly stated. Then, after overviewing of our Biological Knowledge Base, the description the knowledge base based on *QUIXOTE* is explained.

4.1 Requirements for a Biological Knowledge Base

There are several things that a biological knowledge base must satisfy if we want the knowledge base to be useful for for both inference machine and non-specialist of biology. The requirements, and their solutions, are listed below.

- **Requirement1** Knowledge having different reliability must co-exist in the knowledge base. But, knowledge having a different reliability must be processed differently.

  Solution1 This can be realized by using of module concept of *QUIXOTE* . We divide knowledge into modules according to the reliability of that knowledge. When we want to discover a little but reliable motifs, we should consult only modules whose reliability is high. When we want to discover many, at the sacrifice of reliability, we should also consults those modules whose reliabilities are low.

- **Requirement2** The system must manage knowledge bases belonging to public knowledge base, and those belonging to researchers, differently. Also, the system must manage knowledge bases belonging to individual differently.

  Solution2 This can also be realized by using the module concept of *QUIXOTE* . Suppose there are three modules, Public, User1 and User2. When user1 wants to consult biological knowledge, he should consult Public and User1. And, when user2 wants to consult biological knowledge, he should consult Public and User2.

- **Requirement3** Frequent situation in biology is that Person 1 regards motif C as being motif of enzyme A, but Person 2 regards that motif as being a motif having function B. At first sight, this appears to be a contradiction, but actually is not. This is because enzyme A has function B, and a portion of the protein around motif C has function B. This happens because of the different interests of researchers. In this case, Person 1 is interested in the class of the discovered protein, while and Person 2 is interested in the function of the discovered protein.

  Solution3 This kind of complexity cannot be represented by a conventional hierarchical representation. But, it can easily be represented in multiple inheritance in the Object Oriented feature of *QUIXOTE* if relation between biological concepts is untangled and properly described. To make the description, not only knowledge on DOOD but also knowledge on biology is indispensable. And we have the both.

- **Requirement4** Information described in the comments in the database is often important and the information must be readable both by machine and non-specialists of biology.

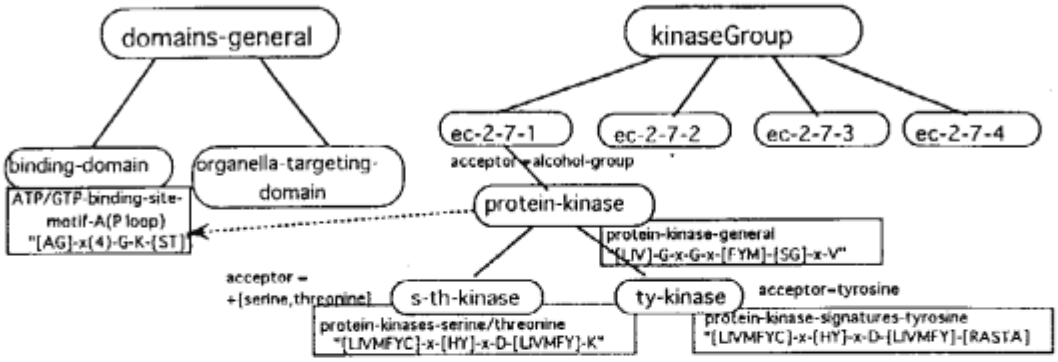

Figure 4

Solution4 This cannot be done automatically, hence must be done manually. But, this dirty and time consuming job becomes facilitated if we keep concept of Object-oriented is in mind.

4.2 Overview of Biological Knowledge Base

We constructed a Biological Knowledge Base using *QUIXOTE* . The Biological Knowledge Base contains knowledge on motif knowledge and related Knowledge. There are three modules in the knowledge base, Prosite*, User and Discovery. The system consults Prosite* and User when it tries to discover motifs, and it registers discovered motifs to Discovery.

In the system, Prosite* is more reliable than User. In Prosite*, motifs in Prosite [7] , representative database of motif, and their related knowledge are represented in hierarchical and multiple inheritance scheme. With the use of a multiple inheritance scheme, biological concepts become more properly represented (In the original Prosite, motifs are represented just in the hierachical scheme).

User is a module of the Biological Knowledge Base in which motifs collected by the user are stored. In some cases, motifs in User can be less reliable than those in Prosite*. To represent biological concepts in User and Discovery, the same scheme as that used in Prosite* is used.

To show the Biological Knowledge Base more in detail, a portion of the knowledge in Prosite* is visualized in Figure 4. In the figure, the concepts and motifs belonging to kinaseGroup and domain-general are shown. Those concepts in KinaseGroup relate to the class of protein, while those in domain-general relate to function of a protein. Because the two concepts are not independent, there is a cross-link between them.

KinaseGroup, which transfers phosphate, is classified into four groups, one of which contains protein-kinase. And, protein-kinase is futher classified into tyrosine kinase(ty-kinase) and serine/threonine kinase(s-th-kinase). The three classes, protein kinase, tyosine kinase and serine/threonine kinase didn't exist in the original Prosite. We have introduced these three classes because, although they aren't necessary for biologists to understand a knowledge base, they are necessary for us and the machine to understand it. But the introduction of the concepts is can be overridden. If a user doesn't want the introduced concepts of class (ex. tyrocine kinase), the system handles knowledge as if there were no such classes. This can be realized by the use of method of *QUIXOTE* .

For some concepts, a corresponding motif is registered. Protein kinase has a motif named protein-kinase-general, whose pattern is "[LIV]-G-x-G-x-[FYM]-[SG]-x-V". Here the expression [SG] signifies S or G; and 'x' signifies any amino acids. Tyrosine kinase has a motif named protein-kinase-signatures-tyrosine and serine/threonine kinase has a motif named protein-kinases-serine/threonine

As another motif of tyrosine kinase or serine/threonine kinase, we can refer to a motif belonging to protein kinase in addition to its own motif. This can be done by using method of *QUIXOTE* .

Also, as a motif of protein kinase, we can refer to a motif named "ATP/GTP-binding site motif A (P-loop)" , belonging to binding-domain, by the use of a method of *QUIXOTE* . In the original Prosite, the information about the cross-linking was described in natural language at comments. We read the implicit information and expressed it as explicit information to improve readability for us and the machine.

4.3 Description by DOOD

The description of knowledge (visualized in Figure 4) by *QUIXOTE* is shown in Figure 5 (class structure) and in Figure 6 (entries of motif). In Figure 5, $A >= B$ means that B is sub-concept of A.

There are four motif entries in Figure 6. Among them, the upper three motifs belong to Prosite* and the lowest motif belongs to User. The specification of a module is done by the specification preceding the double semi-colon. We can specify Prosite* or User or Discovery as a module.

```
Prosite*::protein-kinase[name = "Protein kinases general"]/
   [pattern = "[LIV]-G-x-G-x-[FYM]-[SG]-x-V",
    comment  = "involved in  ATP binding ",
    otherMotif = binding_domain
       [name = "ATP/GTP-binding site motif A (P-loop)"]
   ];;
Prosite*::ty_kinase[name = "Protein kinases signatures tyrosine"]/
   [pattern = "[LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTA]-x(2)-N-[LIVMFYC](3)",
    acceptor = tyrosine,
   ];;
Prosite*::s_th_kinase[name = "Protein kinases serine/threonine"]/
   [pattern = "[LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYC](3)",
    acceptor +={serine,threonine},
   ];;
User::protein-kinase[name = "Protein kinases ATP-bind"]/
   [pattern = "A-x-K",
   ];;
```

Figure 6

```
kinaseGroup >= ec-2-7-1 ;;
kinaseGroup >= ec-2-7-2 ;;
kinaseGroup >= ec-2-7-3 ;;
kinaseGroup >= ec-2-7-4 ;;
    ec-2-7-1 >= protein-kinase ;;
         protein-kinase >= s-th-kinase;;
         protein-kinase >= ty-kinase;;
```

Figure 5

In the top entry(protein kinase), there is an attribute, otherMotif. The attribute enables the system to refer a motif belonging to aother class (binding-domain). The procedure for the reference is also written in a method of *QUIXOTE* . This multiple inheritance is one of the features of DOOD.

Protein kinase has as attribute value of acceptor alcohol-group, inherited from its upper class ec-2-7-1. There are two sub-groups of protein kinase, tyrocine kinase and serine/threonine kinase, each of which has tyrosine and serine or threonine (expressed by acceptor $+=\{serine, threonine\}$) as the attribute value respectively. In this case, the inherited attribute value is overwritten. From biological point of view, the attribute value becomes more precise because tyrocine, serine and threonine are the members of alcohol group. This is another feature of DOOD.

Using this information, the acceptor of protein can be determined more precisely if we find motif which belongs to higher hierarchy. The identification as tyrosine kinase or serine/threonine kinase brings us more information than identification as protein kinase, the upper concept of the two.

5 Example Application

In this section, we present an example where the system discovers a motif using biological knowledge.

Preparatory Analysis The results of alignment (similarity matching) of seven sequences by using Aligner equipped in Motif Generator are shown in Figure 7 (top). Only the part that is relevant to this explanation is shown. From the result, Motif Generator detects a common pattern 'G-x-G-x-F-G'.

Identification of class of protein Motif Generator searches for a pattern in Prosite* and identifies it as a motif named "protein-kinase-general" which belongs to the protein kinase. The system searches for other motifs belonging to the protein kinase (Application of Rule 1) and identifies "ATP/GTP-binding site motif A (P-loop)" in the alignment (not shown in the Figure). The reference of "ATP/GTP-binding site motif A (P-loop)" is realized by the multiple inheritance feature of *QUIXOTE* . As a result the proteins is found to belong, at least, to protein kinase which can bind to ATP/GTP.

The system tries further identification of classes of protein. It checks if the alignment contains motif of either tyrosine kinase or serine threonine kinase, both of which are sub-class of protein kinase (Application of Rule 2). But, the system finds neither motif. It can be concluded that the protein belongs to the general class of protein kinase and its acceptor is alcohol-group (Overwriting information is consulted but not used in this case)

Identification of known motif The system tries to identify motifs belonging to User in the multiple alignment (Application of Rule 3). The identification of known motifs helps the system to enhance the quality of multiple alignment high. Consequently, this enhances the quality of motif discovery.

The system notes that pattern 'A-x-K' is hidden in the alignment. Aligner makes an alignment with the constraint that aligns 'A-x-K' in the same columns (Figure7 (bottom). The system evaluates whether the alignment is acceptable according to biological criteria [5]. It identifies it as a known motif.

Discovery and Registration of motif Finally, the system investigates whether there are any new motifs in the

Figure 7

alignment (Application of Rule 4). As a result, it identifies pattern '[LIM]-x-E-[ARK]' and registers, to Discovery, the pattern as new motif of protein kinase.

In this way, the system predicts the function of proteins coded in genes and it discovers new motifs of the proteins.

6 Discussion

Knowledge Base

The Biological Knowledge Base in the system satisfies the four conditions stated in Section 4.1. Requirement1, the management of multiple level reliability, is satisfied by the module feature of *QUIXOTE* . Requirement2, the management of multiple knowledge bases of different owners, doesn't correspond to the system. But, it can also be satisfied by the module feature of *QUIXOTE* .

Requirement3, the proper representation of biological concepts, is satisfied by the multiple inheritance, the overwriting of inheritance and the usage of method, which are not special features of *QUIXOTE* but standard features of a Deductive Object-Oriented Database. Requirement4 is satisfied by our endeavor to biology.

Motif Discovery

The motif discovered by the system was also discovered by the previous system. The difference is the time required to make the discovery. The previous system couldn't notice the motif as a motif of the protein kinase, ATP/GTP-binding site motif A (P-loop), belonging to binding domain, whereas the present system can notice the motif with the help of multiple inheritance. The motif information makes the consumed time of the present system lower than that of the previous system. This proves that proper representation of knowledge can contribute to the speed-up of the genetic analysis

Others

The usage of the module feature of *QUIXOTE* which used to satisfy Requirement 1 and 2, is not indispensable to our system. It can be substituted by other methods of description without the feature. But we believe the module feature will be useful when the volume of the biological knowledge base is increased.

7 Conclusion

A biological knowledge base for motif was constructed by the use of DOOD (Deductive Object-Oriented Database) language *QUIXOTE* . A motif discovery system with a biological knowledge base was developed. The system discovered a new motif of the protein kinase. The time required for the discovery was reduced in the present system because the system could access the motif information with the help of the multiple inheritance feature of DOOD (precisely OO feature).

Also the D (Deductive) feature of DOOD is also suitable for to interfacing with the inference module of the system. Consequently, DOOD is suitable as a way of representing a biological knowledge base that attached to genetic analysis system.

8 Acknowledgement

The authors thank Prof. Fuchi, ex-director of ICOT, for introducing us to knowledge engineering and for giving us with a research environment. We also thank Dr. Uchida, director of ICOT, for accessing us to genetic information processing.

We also thank to *QUIXOTE* groups at ICOT, especially H. Tanaka, for their assistance.

Finally, the first author gives special thanks to N. Mizuhara and M. Kunimoto at the System development Lab. of Hitachi Ltd. for providing the computer environment used to write this paper.

References

[1] **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**.

[2] **Kazic,T.** *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*,**Vol.1** pp803-812, **1993**.

[3] **Hirosawa,M.** *et al.* Protein Multiple Sequence Alignment using Knowledge. *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*,**Vol.1** pp803-812, **1993**.

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

[5] **Dayhoff,M.O.,** Schwatz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in protein. In Dayhoff,M.O.(ed), *Atlas of Protein Sequence and Structure Vol.5,--Suppl.3,* Nat. Biomed. Res. Found., Washington, D. C., 363-373.

[6] **M.Hirosawa,** M.Hoshida, M.Ishikawa. Iterative Multiple Alignment with Similarity yConsideration. *Proceedings of the Thirty-Sixth Convention of Information Processing Society of Japan*,**Vol.2** pp289-290, **1993**.

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