Development of New DDBJ DNA Sequence Database with Data Annotation Tool Yamato II

T. Koike [3]
tkoike@genes.nig.ac.jp

T. Okayama [3]
tokayama@genes.nig.ac.jp

J. Ishii [3]
jishii@genes.nig.ac.jp

T. Mizunuma [3]
tmizunum@genes.nig.ac.jp

T. Tamura [2]
tatamura@genes.nig.ac.jp

Y. Tateno [1]
ytateno@genes.nig.ac.jp

H. Sugawara [1]
hsugawar@genes.nig.ac.jp

K. Nishikawa [1]
knishika@genes.nig.ac.jp

T. Imanishi [1]
timanish@genes.nig.ac.jp

K. Fukami-Kobayashi [1]
kfukami@genes.nig.ac.jp

K. Ikeo [1]
kikeo@genes.nig.ac.jp

T. Gojobori [1]
tgojobor@genes.nig.ac.jp

[1] Center for Information Biology, National Institute of Genetics
1111 Yata, Mishima, 411 Japan

[2] Association for Propagation of the Knowledge of Genetics
1171-195 Sakuragaoka, Yata, Mishima, 411 Japan

[3] Hitachi Software Engineering Co., Ltd.
5-79 Onoe-cho, Naka-ku, Yokohama, 231 Japan

Abstract

As the molecular biology has made a rapid progress these years, there has been a great number of changes required of the methodology for maintaining and utilizing DNA sequence data. For example, annotation to sequences has become complex and extensive. DDBJ which recognized the impending requirements decided to develop a new DNA sequence database system in 1995. To tolerate with frequent changes of the data structures and significant increment of the data in terms of quality and quantity, we designed a completely new database schema. In the new system, physical changes of the data structure do not affect such applications as a tool for annotation. We also designed a new annotation tool with object oriented concept that allows us to handle DNA sequence data in computers as intuitively as in the real world. The annotation tool is named as YAMATO II. We also take care of needs from DDBJ itself in the new system. Data traffics and security in the database access are especially analyzed and reviewers of data for DDBJ who are distant from DDBJ are now able to process the data safely and comfortably in the new system. The new system also realized more robust and effective data exchange with partners in the international nucleotide sequence banks, EMBL and GenBank.

1 Introduction

DDBJ (DNA Data Bank of Japan) organizes an international nucleotide sequence data bank with EBI (European Bioinformatics Institute) and NCBI (National Center for Biotechnology Information) and they cooperate in developing the international DNA sequence database DDBJ/EMBL/GenBank [1].

DDBJ itself accumulates data directly from scientists and disseminates the international database. The bank used to utilize a relational database management system as database engine and Annotator's Work Bench (AWB) [2] to annotate sequence data. This conventional system was developed by Los Alamos National Laboratory in U.S. more than 5 years ago. Meantime, data items [3] included in the database have frequently changed due to the rapid development of molecular biology and the wide spread of sequencing technology, and it has been almost impossible for us to tolerate with the changes by the conventional system.

In order to catch up with the change induced by the science and technology, we were urged to promptly develop our own new DNA database system which we could enhance it by ourselves.

In addition, users expect us to introduce the state-of-the-art technology for data processing such as the internet and World Wide Web (WWW). Therefore we defined a new schema, prepared object oriented libraries for the programming, created an annotation system for DDBJ staffs, and developed a new data submission system by use of WWW [4].

In this paper, the annotation system, which guarantees the credibility of the DDBJ database, is described in details after a short summary of the new schema.

2 System and Methods

We used Sun SPARC server 2000/E with Solaris 2.4, 1GB memory, 603GB hard disk, and SYBASE SQL server 10 for the database engine.

The database accessing library was written in C++ language with the size of 17.5k lines. In order to access the database server through the SYBASE-DB-client library, the database accessing library can be compiled and executed on both Apple Macintosh and Sun SPARC server by using SYBASE-DB-client library for each platform.

YAMATO II, the database annotation tool, is a program which runs on Apple Macintosh. It was written in C++ language and the size was 64.5k lines. It is connected to the database server with TCP/IP protocol and uses the previously described database accessing library for database access. The user interface was developed by using Neuron Data Open Interface which is a platform independent graphical user interface (PIGUI).

3 Design and Implementation

3.1 Overview

The skeleton of the new database system is presented in Fig.1. DNA Data are submitted with mainly two different ways: a) with electronic mail or floppy disk and b) with the SAKURA through WWW. And Yamato II, the database annotation tool running on Macintosh, is used to input data into SYBASE database server on the UNIX machine and to modify the data in the database. The data which is stored in the database can be converted into the text file format for data distribution. That is used for disseminating data on-line at DDBJ and for exchanging data off-line with satellite sites of DDBJ and other data banks.

3.2 Schema

The new system schema was designed with these factors in mind which are:

1. Consistency
2. Efficiency
3. Sufficiency
4. Expandability
5. Interoperability

We adopted an object orient design to model the real world objects as naturally as possible since the structure that is developed from this strategy is much easier to manipulate and maintain.

The design of the new database schema basically follows the "ANSI/SPARC three level schema architecture" [5]. It classifies the abstract levels of database schema into three levels. First of all, "Conceptual schema" directly models real world objects or phenomena onto the surrogates. Second, "Internal schema" describes an implementation of the physical structure on the database details. And "External schema" is used for each user's or application's view of data. Although an interpretation of this concept varies from person to person, the classification of layers in the design makes it possible for us to focus on the characteristic problems of each layer, and insures independency of a layer from the underlying structure.

Fig.2 shows an overview of the new DDBJ schema methodology. In summary, the conceptual schema was designed first by using an object-oriented functional model (AIS) [6] with visual diagrams and it was implemented on a commercial RDB (relational database) system. Here the term "relational" is defined as internal or physical modeling because we define the functional model as being located and implemented by the relational model. All physical information is represented in the relational layer. For the user/application interface, another modeling was adopted for external schema to manipulate data via a tree structure. To support these operations, an object oriented database access library was created by using C++ language. This interface enables access to abstract objects by way of tree/set operation, that features "set at a time" basis with the hierarchical description.

More detailed information about this section will be reported elsewhere. (We are now preparing the report to publish[7]).

3.3 Yamato II

During the development of Yamato II, we adhered to the following design policies:

a) using a graphical user interface (GUI) on Macintosh GUI is easier for users to understand and operate than non-graphical one. And the most terminals in DDBJ are Macintosh. Almost all software for running on Macintoshes is designed with GUI.

b) layered structure We separated the structure of the Yamato II into three different layers (Fig. 3) which are consisted of the commercial database client layer, the database accessing layer and the application layer. Consequently the layered structure makes Yamato II easier to modify and increases the portability.

c) platform independent To give our Yamato II a good inter-platform portability, we decided to use a PIGUI product. With the product, the user interface of Yamato II is independent of platforms (Macintosh, Microsoft Windows series, OS/2 and UNIX (X windows)). And we developed other parts of the application layer and the database accessing layer with ANSI C++ language because of the same reason. In the case of commercial database client layer, a product which runs on many platforms (Macintosh, Windows NT, UNIX, etc.) was chosen. As a result, Yamato II expected to be executable on various platforms by only recompiling with appropriate commercial libraries on the platform.

4 Discussion

Here we discuss the features and the new database system by comparing with the traditional system.

4.1 Easy-to-modify

First of all, the new system is designed to be flexible enough to accept any modifications occurred in the future. Since the conventional system was given from the outside of DDBJ, the modification was almost impossible. We had experienced so many problems which were caused by the difficulty of the modification. Therefore, we were more concerned about maintaining better availability of modification through exercising the system design. (i.e. more simple schema of the database, strongly moduled and layered the software structure.)

4.2 Easy-to-operate

The Yamato II is a part of the user interface of the new database system for annotators and reviewers in DDBJ. Yamato II runs on Macintosh and is equipped with a natural graphical user interface. The application software named AWB running on UNIX in the conventional system corresponds to Yamato II in the new system. AWB was designed to be available on character-based terminals. The screen is written with ASCII characters and the input is only feasible from keyboard. Thus, the operators were required to acquire the skills in manipulating AWB. In contrast, Yamato II requires significantly less trainings for the operation. It was designed that almost all operations are done with choosing items from the menus which are prepared for each data item on the screen (Fig.4). As a result, it contributes not only to shorten the operation period but also to decrease the operational mistakes.

4.3 Quick Preparation of Database Releases

DDBJ issues releases of the whole DNA database four times a year for off-line dissemination. With the conventional system, it took us about a week to convert the data stored in the database into a format for the release. The format is called 'flat file format'.

With the new system, the 'flat file format' is stored in the database and updated on the fly. For example, when the data is modified, all other data related with the modified ones are flagged as 'need updating' to prevent to be read without corrections. The update process goes to the background in idle time to be ready for the service. Currently we just read the 'flat file' on demand. It now takes a few hours to prepare the release.

4.4 Data Confirmation and Error Checking

Yamato II has an window on which 'flat file format' is displayed and the window can be opened while an window for review is opened (Fig. 5). When you are converting data into 'flat file format', Yamato II checks insufficient data and inconsistency and lists the warnings on the 'flat file' window. Therefore, the operator can confirm the data modification in the 'flat file format' upon his or her convenience as well as seeing the errors associated with the modified data at the same time. Consequently it decreases the errors in the data construction and keeps the quality of DDBJ's database high.

4.5 Auto-decision for Specific Data

Yamato II chooses a candidate and constructs data automatically for the specific items of data. For example, it constructs peptide sequence data when CDS (Coding sequence) exists in the DNA sequence, and it chooses a table of genetic code which is rendered to translate DNA sequence into peptide sequence. This kind of automatic decision by the system is remarkably effective to save the man power for annotation. It also contributes to purge the inconsistency and to standardize each data into database.

4.6 Connecting Server with Client

Yamato II is a client software which connects with a server through TCP/IP protocol. By using the methods like PPP, it can be used from a remote site even with a telephone line.

Since the database client library which is included in Yamato II manages direct connection with a server, an operator of YAMATO II need not be qualified as a user of the server machine, but the accessibility only to the database is enough for his or her task. This fact is good for the security reason associated with the server machine which includes other resources.

At DDBJ, the main database server now runs on Sun SPARC server 2000/E. However, since our database engine SYBASE frequently releases many different database server products on various hardware, various choice of hardware is available as the database server machine. For example, a personal computer with Microsoft Windows NT operating system could be a database server machine for a small scaled database. In fact, our test bed for the database system development is a Windows NT machine. Besides this, Yamato II is platform independent (on Macintosh, Microsoft Windows series, OS/2 and UNIX). Therefore we will be able to switch to the most efficient combination of server / client machines anytime.

5 Conclusion

We had defined the new database schema and implemented the new DNA sequence database system which replaced the old one in January 1996. Because of this, we could realize a high productivity in the implementation and quickly start a fairly stable operation. The advantages mentioned in the previous section are actually proved in daily use.

As you might know, DDBJ daily exchanges data for the update of the database with EBI and NCBI [8]. Each of them identifies the requirement to the specifications of the database from scientific communities and they keep discussing about the relevant upgrade. Thanks to the new system, we now be able to modify the database system in a relatively and reasonably short period when the policy of the three banks are changed.

However, we already have had some plans for the improvement of the new system. The major ones are: a) to share the taxonomy database with EMBL/GenBank, b) to rewrite the programs with Java, c) to share some program modules with the SAKURA, the new data submission system on WWW, and d) to distribute the database system to the external satellite sites of DDBJ.

References

[1] Gojobori, T. and Tamura, T. "Current State of Scientific Information Systems in Biology," *Japanese Scientific Monthly*, Vol. 49, pp. 687-692, 1996.

[2] GenBank at Los Alamos *User Manual Training Guide and Reference Manual For the ASCII AWB*, 1993.

[3] DNA Data Bank of Japan, EMBL Nucleotide Sequence Database, GenBank(NCBI) *The DDBJ/EMBL/GenBank Feature Table Definition Version 1.08 Dec 1, 1995*

[4] Yamamoto, H., et al. "SAKURA: A new data submission system in DDBJ for the age of mass production for DNA sequence data.," *The Seventh Workshop on Genome Informatics at YEBISU Garden Place* on December 2 and 3, 1996.

[5] Tsichritzis, D. C. et Klug, A. "The ANSI/X3/SPARC DBMS frameworks: report of the study group on data base management systems," *Information Systems 3*, 1978.

[6] Arisawa, H. et al. "Representation of complex objects in semantic data model "AIS" and implementation of set operators," *IEICE Trans.*, Vol. E74, pp. 191-203, 1991.

[7] Okayama, T., et al. "Formal design and implementation of improved DDBJ DNA database with a new schema and object oriented library," *in preparation.*

[8] Tamura, T. and Tateno, Y. "Data Bank Activity and System Development in DNA Data Bank of Japan," Presented at *the 1995 Meeting of CODATA Task Group at George Mason University* on June 16, 1995.

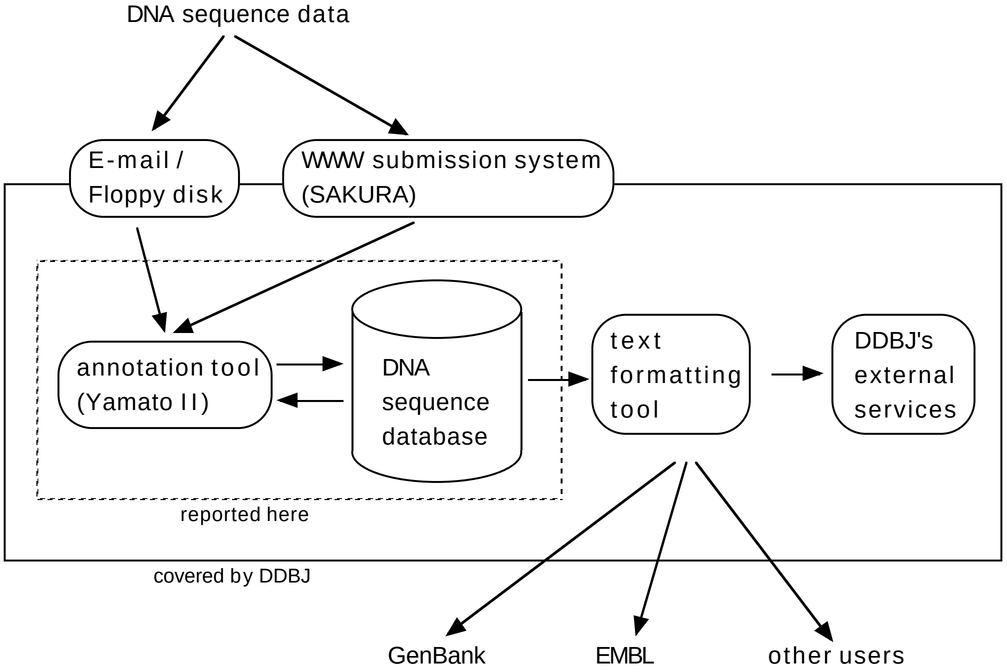


Fig.1 overview of DDBJ DNA sequence database system

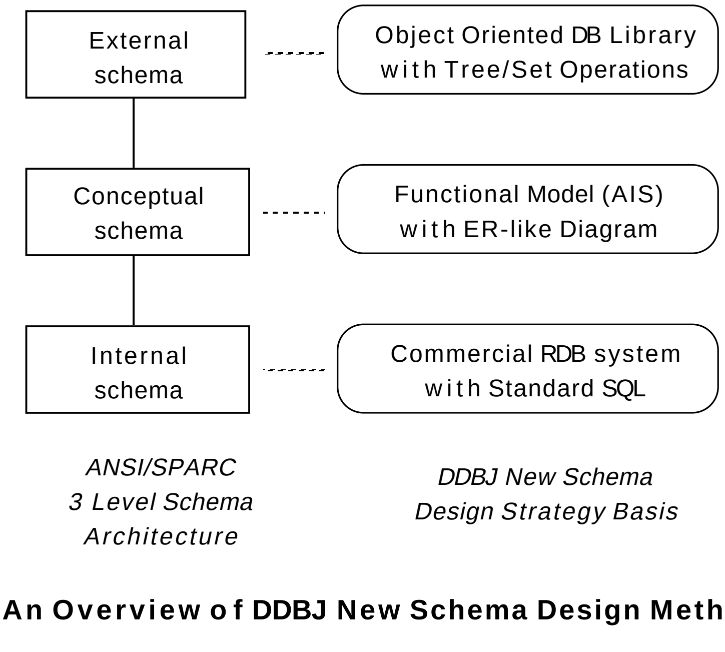

External schema
Conceptual schema
Internal schema
Object Oriented DB Library with Tree/Set Operations
Functional Model (AIS) with ER-like Diagram
Commercial RDB system with Standard SQL
ANSI/SPARC 3 Level Schema Architecture
DDBJ New Schema Design Strategy Basis

Fig.2 An Overview of DDBJ New Schema Design Methodology

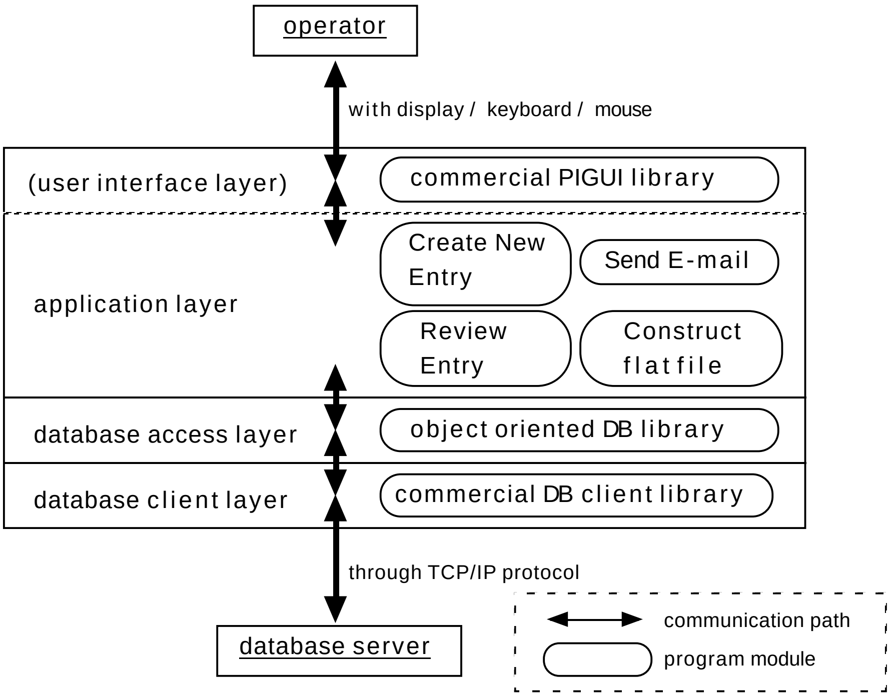
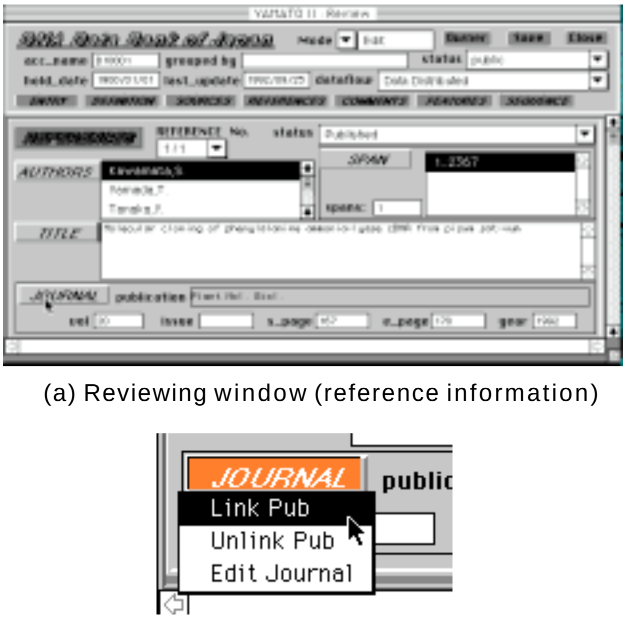


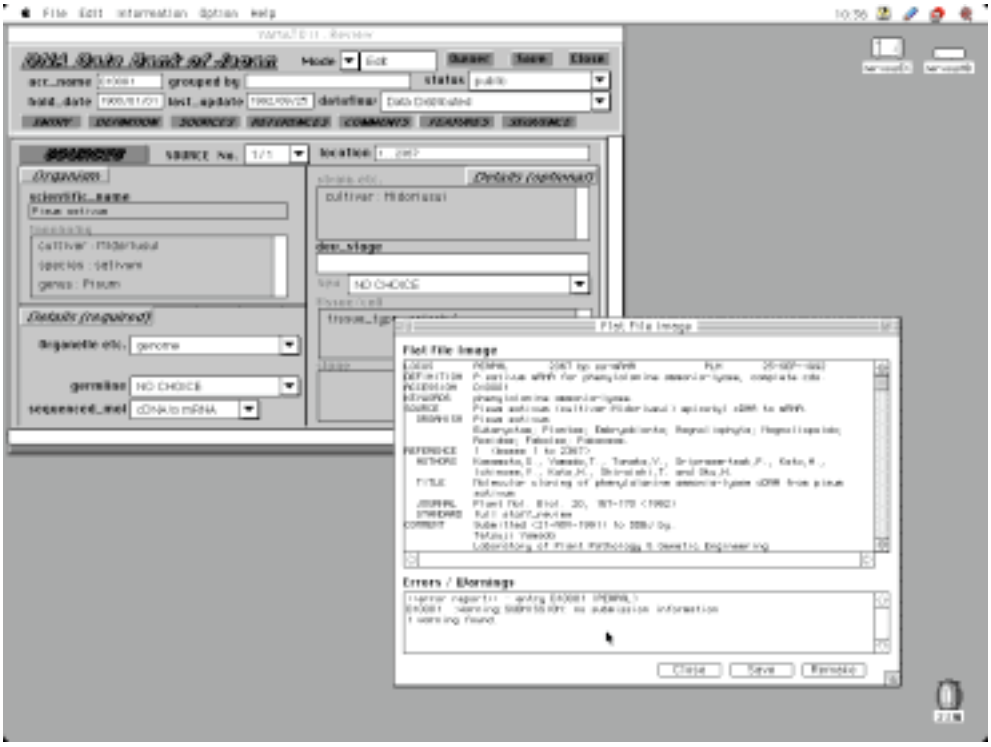
Fig.3 Internal structure of Yamato II



(a) Reviewing window (reference information)

(b) menu for 'Journal' information

Fig.4 Reviewing window of Yamato II



Mode: Edit
acc_name: D10001
grouped by:
status: public
hold_date: 1995/01/01
last_update: 1992/09/25
dataflow: Data Distributed
SOURCE No. 1/1
location: 1..2367
scientific_name: Pisum sativum
taxonomy: cultivar : Midoriusui; species : sativum; genus : Pisum
strain etc.: cultivar : Midoriusui
dev_stage:
sex: NO CHOICE
tissue/cell: tissue_type: epicotyl
organelle etc.: genome
germline: NO CHOICE
sequenced_mol: cDNA to mRNA

Flat file Image

```
LOCUS       PEAPAL       2367 bp ss-mRNA        PLN       25-SEP-1992
DEFINITION  P.sativum mRNA for phenylalanine ammonia-lyase, complete cds.
ACCESSION   D10001
KEYWORDS    phenylalanine ammonia-lyase.
SOURCE      Pisum sativum (cultivar Midoriusui) epicotyl cDNA to mRNA.
  ORGANISM  Pisum sativum
            Eukaryotae; Plantae; Embryobionta; Magnoliophyta; Magnoliopsida;
            Rosidae; Fabales; Fabaceae.
REFERENCE   1  (bases 1 to 2367)
  AUTHORS   Kawamata,S., Yamada,T., Tanaka,Y., Sriprasertsak,P., Kato,H.,
            Ichinose,Y., Kato,H., Shiraishi,T. and Oku,H.
  TITLE     Molecular cloning of phenylalanine ammonia-lyase cDNA from pisum
            sativum
  JOURNAL   Plant Mol. Biol. 20, 167-170 (1992)
  STANDARD  full staff_review
COMMENT     Submitted (21-NOV-1991) to DDBJ by:
            Tetsuji Yamada
            Laboratory of Plant Pathology & Genetic Engineering
```

Errors / Warnings

```
!!error report!! - entry D10001 (PEAPAL)
D10001 :warning SUBMISSION: no submission information
1 warning found.
```

Fig.5 Flat file window and Reviewing window