**Sanzo Miyazawa**
Laboratory of Genetic Information Analysis, Center for Genetic Information Research, National Institute of Genetics, Mishima, Shizuoka 411, Japan

# DNA Data Bank of Japan: Present Status and Future Plans

Activities of the DNA Data Bank of Japan are reported; the present status of data collection, data entry and search/retrieval systems developed at the DDBJ, and future plans of the DDBJ are discussed.

## INTRODUCTION

As a center for DNA sequence databanks and related activities in Japan, the DNA Data Bank of Japan (DDBJ) was established at the National Institute of Genetics with a grant from the Japanese government in April, 1986. The government has shown its support and commitment to the databank through a permanent grant. At present, this project consists of two faculty positions, some operating funds, and computer facilities.

A primary task of the DDBJ is, of course, DNA sequence collection. However, in addition, we have a wide range of activities: (1) DNA data collection and data entry in collaboration with other databanks; (2) data distribution, including the

Computers and DNA, SFI Studies in the Sciences of Complexity, vol. VII
Eds. G. Bell and T. Marr, Addison-Wesley, 1990

secondary distribution of the GenBank[2] and EMBL[3] databases in Japan; (3) provision of on-line access to DNA and related databases; (4) development of research tools for sequence analysis; (5) regularly published newsletters to inform researchers of the activities of the DDBJ, and (6) provision of training courses for users of the DDBJ computer system. We have been developing a data entry system to manage data collection, a search/retrieval system for sequence databases, and research tools for DNA and protein information analysis. To let researchers know about these activities, we have published newsletters regularly and given training courses. The newsletters contain articles that describe the state of international collaboration among databanks and how to submit data to databanks as well as matters of specific interest to Japanese scientists such as available databases at the DDBJ, how to access the DDBJ computer system, and how to use the databases. The activities of the DDBJ are also conveyed through an on-line service of the DDBJ computer system; information about our activities may be obtained by accessing the computer system and using the "getinfo"[10] command devised for this purpose. All these activities provided by the DDBJ are open to anyone regardless of whether one works for a non-profit organization or not.

In the following paper, I will briefly report the present status of data collection, data entry and search/retrieval systems developed at the DDBJ, and future plans of the DDBJ.

## DATA COLLECTION

DNA data collection and data entry are the primary tasks of the DDBJ. Our data collection is carried out in collaboration with the GenBank[2] and the EMBL Data Library.[3] We started collecting DNA data in December, 1986. Data is currently entered in the GenBank format and fully annotated.

TABLE 1 The Number of Entries and Bases in Each Release of the DDBJ Database

| Release | Date | Entries | Bases |
|---|---|---|---|
| 1 | 07/87 | 66 | 108,970 |
| 2 | 01/88 | 142 | 199,392 |
| 3 | 07/88 | 230 | 345,850 |
| 4 | 01/89 | 302 | 535,985 |

We released the first version of the DDBJ database in July, 1987. Release 1 included only 66 entries and 108,970 bases. Since then, our database has been released every half year. The numbers of entries and bases included in each release are listed in Table 1. The DDBJ collected about 240,000 bases in one year from July 1987 to July 1988. About 8,000,000 bases were collected during the same period by the EMBL Data Library and the GenBank. In other words, the DDBJ processed about 1/30 of the total collection of DNA sequences in a year.

Each release includes a coding sequence database and a peptide sequence database that were extracted and translated from the original DNA sequence database. They are helpful for users, since the translation of base sequence to amino acid sequence is not trivial due to the exon-intron structure in a gene and the variation of genetic code. Release 2 and the later release included the journal index, accession number index, short directory, and data submission form files. Release 4 will be available in January, 1989.

## JOURNALS SCANNED

One of international collaborations in data collection is to share journals that each databank scans. The DDBJ principally has charge of journals published in Japan. The journals scanned are listed in Table 2; the MEDLINE literature databases are used to search other journals.

As expected, original DNA sequences were hardly reported in most journals published in Japan except for a few such as the *Journal of Biochemistry (Tokyo)*, *Agricultural Biological Chemistry*, and the *Japan Journal of Genetics*. Even the *Journal of Biochemistry (Tokyo)* included only about 20–25 papers per year. The total number of papers that included original DNA sequences in the scanned Japanese journals was only about 30–40 per year. We are now planning to regularly scan a few main journals and use the MEDLINE literature databases to search minor journals in which few reports of original DNA sequences were published in a year. By the way, reports from Japanese research organizations[14] numbered 148 of the 1279 papers published in 1987 according to the BIOSIS Preview database. We processed about 70 papers in the past year, nearly half of the reports from Japanese organizations. This means that the DDBJ processed about 1/20 of all papers published in a year. (BIOSYS does not always cover all of the reports, so this number may be an overestimate.)

A main obstacle to increasing data entry is that it is difficult for us to employ enough annotators and reviewers. At present, we have only 0.5 full-time employees (FTEs) for annotation and 0.2 FTEs for review. It is reasonable to expect that the DDBJ could process only about 1/20 or 1/30 of the total collection if one compares the DDBJ staff with the staffs at the GenBank and the EMBL Data Library. Direct data submission from authors and even data entry by authors

TABLE 2 Journals Scanned by the DDBJ and the Number of Papers Found to Include Original DNA Sequences[1]

| | | entries | papers |
|---|---|---|---|
| JOURNALS PUBLISHED IN JAPAN | | | |
| Agricul Biol Chem | Vol. 50(01)–50(12) 1986 | 3 | 3 |
| | Vol. 51(01)–51(12) 1987 | 12 | 11 |
| | Vol. 52(01)–52(10) 1988 | 14 | 12 |
| Cell Struc Funct | Vol. 11(01)–11(04) 1986 | 0 | 0 |
| | Vol. 12(01)–12(04) 1987 | 0 | 0 |
| | Vol. 13(01)–13(05) 1988 | 0 | 0 |
| Chem Pharm Bull | Vol. 34(12)–34(12) 1986 | 0 | 0 |
| | Vol. 35(01)–35(12) 1987 | 0 | 0 |
| | Vol. 36(01)–36(10) 1988 | 0 | 0 |
| Devel Growth Diff | Vol. 28(01)–28(06) 1986 | 0 | 0 |
| | Vol. 29(01)–29(06) 1987 | 0 | 0 |
| | Vol. 30(01)–30(04) 1988 | 0 | 0 |
| J Biochem Tokyo | Vol. 99(01)–99(06) 1986 | 11 | 8 |
| | Vol.100(01)-100(06) 1986 | 27 | 14 |
| | Vol.101(01)-101(06) 1987 | 15 | 6 |
| | Vol.102(01)-102(06) 1987 | 28 | 14 |
| | Vol.103(01)-103(06) 1988 | 50 | 15 |
| | Vol.104(01)-104(05) 1988 | 12 | 6 |
| Jpn J Cancer Res | Vol. 77(01)–77(12) 1986 | 0 | 0 |
| | Vol. 78(01)–78(12) 1987 | 1 | 1 |
| | Vol. 79(01)–79(10) 1988 | 1 | 1 |
| Jpn J Genet | Vol. 61(01)–61(06) 1986 | 10 | 2 |
| | Vol. 62(01)–62(06) 1987 | 5 | 5 |
| | Vol. 63(01)–63(05) 1988 | 1 | 1 |
| Microbiol Immunol | Vol. 31(02)–31(12) 1987 | 3 | 2 |
| | Vol. 32(01)–32(10) 1988 | 1 | 1 |
| Plant Cell Physiol | Vol. 28(01)–28(08) 1987 | 2 | 2 |
| Zool Sci | Vol. 3(01)–3(06) 1986 | 0 | 0 |
| | Vol. 4(01)–4(06) 1987 | 0 | 0 |
| | Vol. 5(01)–5(04) 1988 | 0 | 0 |
| Nippon Ika Daigaku Zasshi[2] | Vol. 54 1987 | 2 | 2 |
| JOURNALS PUBLISHED OUTSIDE OF JAPAN | | | |
| J Gen Virol | Vol. 68(03)-68(12) 1987 | 38 | 27 |
| | Vol. 69(01)-69(11) 1988 | 53 | 33 |

[1] This data was collected in November 30, 1988.

[2] Not scanned.

themselves are absolutely necessary for us to further develop the DDBJ database. However, if one considers the significantly large amount of DNA sequence data that will be analyzed in near future and if one also wants to keep the quality of data annotation, obviously it is more practical and more realistic for any databank to encourage researchers to enter the data themselves rather than to increase staff.

In order to let authors submit data directly to the DDBJ, we made an agreement with some journals: a floppy diskette or a hard copy of the data submission form is sent to every author whose paper is accepted. We will try to increase the number of journals with which we have such an agreement and also to extend the relationship with all journals such that they do not accept papers without direct data submission similar to the agreement[6] between the EMBL Data Library and the Nucleic Acid Research. GenBank is developing software to help authors enter their data. At present, we plan to use their software.

## DATA MANAGEMENT

Since the DDBJ computer system became available in April, 1987, we have been developing a data management system on our UNIX[13] system. Usually a database system consists of subsystems such as (1) a data entry system, (2) a search/retrieval system, and (3) a data analysis system. It would be desirable to manage all of the three systems by using a single management system. However, building such a system would take a time. We could not afford this method, because we already had started data entry and so decided to create each system independently. Not only is such a system easy to create, users can sort out the data entry system that they do not need.

Our computer system is connected to the JUNET network, which is a UUCP[15] network for electronic mail and bulletin boards in Japan. Researchers may send the DDBJ DNA sequence data by electronic mail or any media. If the journal in which that data is supposed to appear is not one of which the DDBJ has charge, we will forward it to an appropriate databank by electronic mail; electronic mail addresses have been established for the EMBL Data Library and the GenBank so data can be forwarded to them. The EMBL Data Library and the GenBank may also communicate with submitters through the DDBJ. A special login account is available for anyone to log onto and get a restricted access to the DDBJ computer[10] (see Figure 1). A primary purpose of this special account is to provide a way for researchers to obtain a submission form and to submit data to databanks.

```
niguts
      Welcome to the NIG, UNIX System V Release 2.0

login: DDBJnews
Terminal type (pc98msdos): vt100

                DDBJ online news

      available commands

menu            # type this menu list
getinfo         # get information
man             # get the manual of commands
mailx           # send a mail; "mailx ddbj < filename" for ddbj
addresses       # list e-mail addresses
ls              # list contents of directory
cat             # concatenate or type files
pg              # pager
cp              # copy files
rm              # remove files
vi              # vi editor
kermit          # file transfer program
conv            # remove <CR> and ctrl-z
exit            # exit

DDBJnews% getinfo
```

FIGURE 1 A special login account "DDBJnews" to provide a restricted access to the DDBJ computer system.

## A DATA ENTRY SYSTEM

We built a data entry system by utilizing the Source Code Control System (SCCS)[15] available in a UNIX system. Each entry is managed as an SCCS file. SCCS has the following functions[16]: (1) version control: a record is kept with each set of changes, which includes what the changes were, why they were made, who made them, and when, and (2) file locking: only one person can modify data at a time. Both are useful in data entry where more than one person is working simultaneously.

Figure 2 shows data flow at the DDBJ and lists commands that are used at each step of the data entry. Direct data submissions both by floppy diskette and by electronic mail are managed by using mail queues. Data to be submitted is mailed to a special electronic mail address, DDBJsub. Databank staffs use the "accept" command to stamp an accession number on the mail and to forward it to a next mail address, DDBJacc. It can also send a return message of acknowledgement

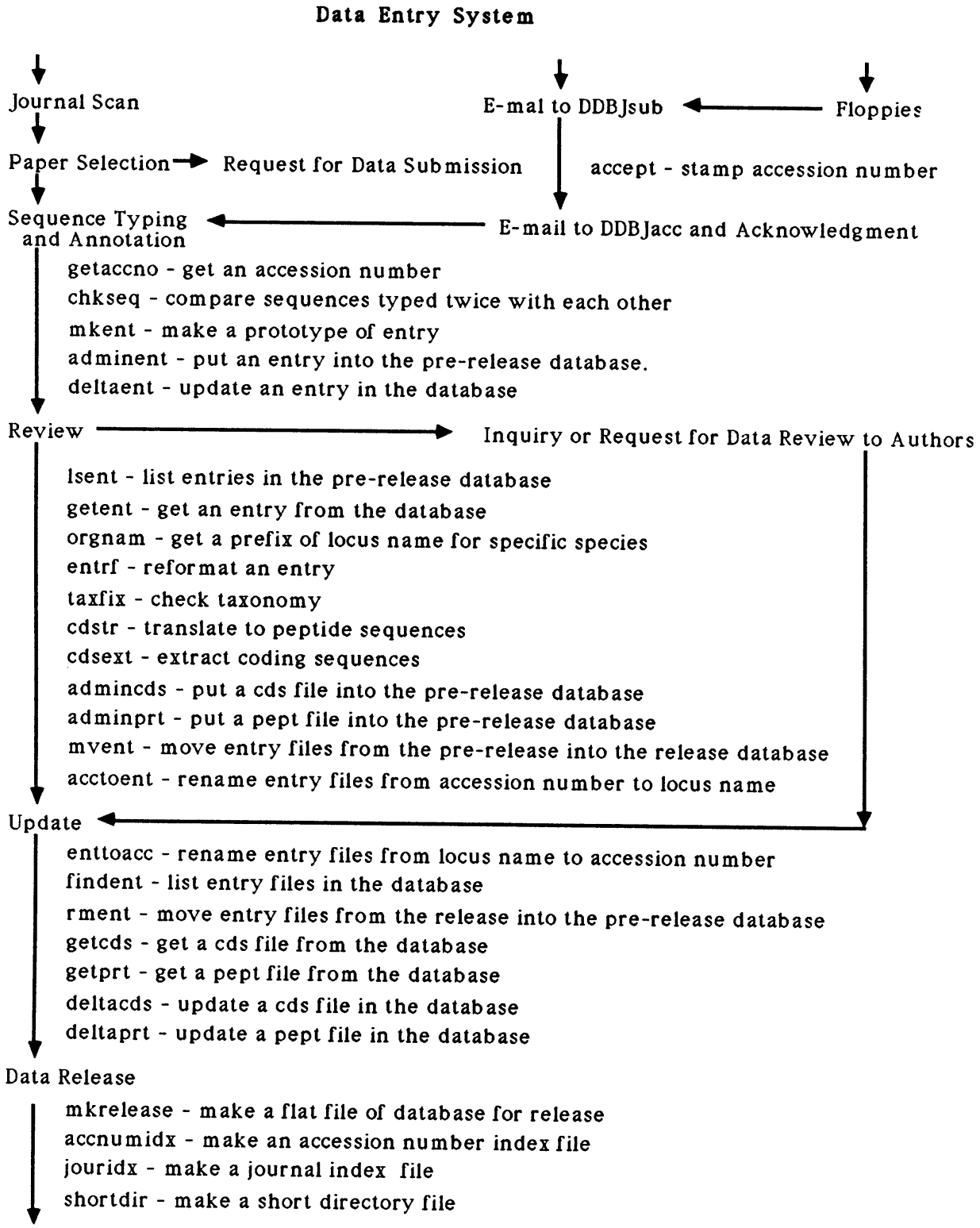


FIGURE 2 Data flow and commands that are used at each step of data entry.

to the submitter. Annotators process the mail that arrives at the DDBJacc address as well as papers that are found to include original sequence data during the journal scan. An entry is entered by using tools such as "chkseq" and "mkent," and then copied into a directory, which stores pre-release entries, by using a command "adminent." The "chkseq," which was programmed by Dr. H. Hayashida of the DDBJ, compares to each other the base sequences that are typed in twice, and the "mkent" makes a prototype of entry. The "adminent" and "deltaent" are front-ends for the "admin" and "delta" commands of SCCS, respectively; the "admin" creates an SCCS file and the "delta" is used to update it. Reviewers then get a pre-release entry by using the "getent" command, review it, and if necessary modify it. An entry may be updated by using the "deltaent" command. After data is reviewed and considered to be satisfactory, it is moved from that directory to another that stores entries ready for release.

QUALITY CONTROL. Data is checked in several ways. A DNA sequence is checked by typing each entry twice. Format, taxonomy, journal name, start and stop codons, and the codon frame in amino acid coding regions are checked by using programs such as entrf, taxfix, cdsext, and cdstr listed in Figure 2, which were programmed by Dr. J. Fickett of the GenBank. However, spelling in reference, features, and comment records is checked only by human review. We plan to use a spelling-checker for this portion. Our experience in checking only coding regions indicates that non-coding regions usually include errors. An effort to find errors in non-coding regions should be made.

## A SEARCH/RETRIEVAL SYSTEM FOR SEQUENCE DATABASES: FLAT

We have been developing a search and retrieval system for flat file databases in order to provide simple tools to use DNA and protein sequence databases. This system called FLAT[11] consists of primitives, most of which perform a single operation and work as a filter in a UNIX system; a filter[8] program reads a line from standard input, processes it, and then writes some output onto standard output. Some basic commands available in the FLAT are listed in Table 3. They perform basic single operations such as (1) extraction of specified types of records from database files, (2) search of strings in each entry of database, and, if found, output of those entry names, (3) performance of "and," "or," and "xor" in respect of entry names, and (4) extraction of specified entries from database files. These filters may be combined with a UNIX pipe[8] to perform a complicated task. One may search and retrieve entries from databases by key words such as author name, journal name, title, organism name, source name, and any combination of such items. An example of such a search and retrieval request is given in Figure 3. This is a typical approach for designing programs in a UNIX system.

Strings for these programs are specified in the regular expression,[8,15] so that one can search and retrieve entries in databases by fuzzy key words and entry names. The "seqgrep" program also allows operators to use the regular expression

to specify sequence patterns to be searched for in databases. Some of these filters were programmed in the Bourne shell and use UNIX tools such as sed, egrep, sort, and awk,[15] so that they are flexible enough to support many formats of databases and to easily keep up with the format changes that often occur. At present, the GenBank, EMBL, PIR[7] (Protein Identification Resource) and PRF[12] (Protein Research Foundation) data formats are supported. However, this approach tends to trade computational speed for flexibility. So applications whose processing speed is critical are written in C language. A program "getgb," which extracts specified entries from databases, uses a "pseud" index file to quickly find the location of the entries in a flat database file.

This FLAT[11] search/retrieval system for sequence databases is designed to be portable among UNIX systems that are available for a wide range of computers from super- to microcomputers.

TABLE 3 Some Basic Commands Available in Flat Software

**{and | or | xor}** *file1 file2 [file3...]*
- and/or/xor entry names in *files*

**{dirgb | dirembl | dirpir | dirprf}** *[database-file...]*
- make short directory of *database-files*

**{fromgb | fromembl | frompir | fromprf}** *[file...]*
- convert *files* from the GenBank/EMBL/PIR/PRF format into the BIONET-like format

**{getgb | getembl | getpir | getprf} [-1] [-0]** *"database-files" [entry...]*
- get *entries* from *database-files*

**{rcdgb | rcdembl | rcdpir | rcdprf} [-f** *"database-files"]* *record-type...*
- get specific *record-types* from *database-files*

**rsites** *reg.-expr.-file [file...]*
- search sequence patterns specified in *reg.-expr.-file* in *files;* appropriate for search of restriction enzyme sites

**seqgrep [-l** *max-pattern-length]* *reg.-expr. [file...]*
- search sequence patterns of *full regular expression* in *files*

**{srchgb | srcdembl | srcdpir | srcdprf}** *[options-for-egrep] reg.-express. [database-file...]*
- search patterns of *full regular expression* in the text portion of *database-files*

```
niguts% flat
flat% set embl=annseq.dat
flat% rcdembl -f $embl OC | srchembl Vertebrata > vrt
flat% wc -l vrt
  8110 vrt
flat% rcdembl -f $embl DE KW RT | srchembl -i oncogene > onco
flat% wc -l onco
   537 onco
flat% and onco vrt > onco+vrt
flat% xor onco onco+vrt > onco-vrt
flat% wc -l onco+vrt
   386 onco+vrt
flat% wc -l onco-vrt
   151 onco-vrt
flat% getembl $embl < onco+vrt > onco+vrt.seq
flat% exit
niguts%
```

FIGURE 3 An example of search and retrieval by using FLAT software.

## A GET-INFORMATION COMMAND: GETINFO

An online help program called "getinfo"[10] has been devised to provide databank staffs and users an easy way to access necessary information. One may use the "getinfo" to learn how to submit DNA data and to which databank it should be submitted; the user can even get a data submission form. An example of the "getinfo" command is shown in Figure 4.

The "getinfo" command apparently mimics the help utility of the VAX/VMS system.[16] However, unlike the VMS help utility, each item of information is stored as a flat file and organized into a tree-like structure, if necessary, by using symbolic links[15] or "pseud" symbolic links; the "pseud" symbolic link was devised because the symbolic link is not available in the System V UNIX. The "getinfo" displays a specified item and, if available, a list of help items at the next level and prompts the user to choose one of them. A "pager" program, "pg" or "less," available in a UNIX system, is used to print files on terminals so that one may read a help item page by page and may save it into another file, if necessary.

```
DDBJnews% getinfo
Type the name of item in which you are interested. The item will be displayed.
If you type
      <CR>,   "getinfo" will back up to more recent topic.
      ctrl-c, "getinfo" will quit at that point.
      '?',    "getinfo" will output an item list again.
Meta characters for file names in "csh" may be used to specify an item.
      ex. "ddbj*", "*LAN"

Pager "jpg" is used to output files; to get help, type ": h" or "(page .): h".

DDBJ_news->       Info_as_file      Learn_unix        Learn_vms
Welcome_msg       background_job    bugs/             bulletin_board/
emacs/            file_transfer/graphics_lib/         imsl_stat_math/
ingres->          inquiries         junet/            line_printer
local_commands/   mails/            manuals->         nig_system/
printing_man      tex               troff             tty_emulator/
work_directory

Item, <CR>, ctrl-c or ? DDBJ

                          DDBJ news

Application/      data_submit/      db_catalog        db_growth/
db_manuals/       db_version        dir_of_files      iris_softwares
manuals/          newsletters/      softwares/        vms_softwares

Item, <CR>, ctrl-c or ? data_submit
```

**FIGURE 4 An example of using the "getinfo" command.**

# FUTURE PLANS

## DATA COLLECTION

The DDBJ aims to process at least all of the DNA sequence data analyzed in Japan. However, at present, three DNA databanks—the EMBL Data Library, the GenBank, and the DDBJ—collaborate by sharing journals for scanning rather than sharing data entry on the basis of the geographic location of submitters. This method of cooperation is used because of a technical problem concerning data collection of regularly scanned journals. The system of data collection will change, as direct submissions from authors increase. Until then, the DDBJ will collect sequence data analyzed in Japan, but will forward them to the databank in charge without processing. We believe that this is a step toward the DDBJ taking charge of the collection of data analyzed in Japan.

DNA sequence data that we collect is sent to the EMBL Data Library and the GenBank when the DDBJ database is released every half year. They incorporate it into their databases. So the delay between data submission and its appearance in the EMBL and GenBank databases may be significant. In order to improve this situation, we plan to send the data as soon as it is ready for release. This new system will begin early in 1989.

## COMPUTER NETWORK

The JUNET network to which the DDBJ computer system is connected is a UUCP[15] network that is connected by public telephone line and so it is a flexible but slow means of communication. Communication over telephone lines via modem is not stable. A current project of building and maintaining identical databases at three sites—the GenBank, the EMBL Data Library, and the DDBJ—needs high-speed communication among their computers. The DDBJ has a plan to connect the DDBJ computer system to the Internet in the U.S.A. to make high-speed communication feasible. Also, the DDBJ is planning to network computers of related organizations in Japan with the DDBJ computer system by using a X.25 packet communication line. This network would be useful in data collection and in DDBJ's provision of database access.

## E. COLI SEQUENCING PROJECT

Because of significant breakthroughs, the large amount of DNA sequences of various genes in a wide range of organisms from prokaryote to human have been analyzed. Now, sequencing the entire genome is not just a dream but a feasible project.

In 1987, Kohara et al.[9] constructed the physical map of the whole *Escherichia coli* (*E. coli*) genome by isolating 3,400 lambda phage clones that contain segments of *E. coli* chromosome and by constructing a restriction map for eight 6-base-recognizing enzymes. Those clones, which may be used for the isolation of any

desired *E. coli* genes, are maintained at the Laboratory of Gene Library in the National Institute of Genetics in Japan, and distributed to anyone who wants to use them. Kohara et al.'s work is the first case of the complete physical mapping of the whole genome. A project of sequencing the entire *E. coli* genome in Japan will use those *E. coli* clones. Directed by Dr. T. Yura and Dr. K. Isono, this project will start in April, 1989, under the direction of Dr. T. Yura and Dr. K. Isono. At present, about 15 laboratories are involved and more laboratories will join later. The DDBJ will join this project in managing sequence data.

*E. coli* is one of those organisms whose genetics has been best studied at the molecular level. Its genome size is about 4.7 million bases and is presumed to consist of about 3000 to 4000 genes. About 15-20% of its base sequence and about 1000 genes[1] are known at present. The genome size of *E. coli*, 4.7 Mb, is not very long but only about one fifth of the total bases that have been collected by databanks. Even so, the physical mapping and sequencing of its entire genome and the management of the sequence data are not trivial tasks. (1) New types of information such as absolute and relative map positions and overlaps between sequence segments must be managed by a database management system (DBMS). (2) Researchers would certainly like to perform more diversified and complicated retrieval of sequence segments than what is currently available, such as retrieval of sequence segments by gene name, product name, and map position. DBMS must be flexible enough to satisfy these needs. (3) In addition,, the project involves many researchers in more than 10 laboratories who are located at geographically distant places. Even though the genome size is not long, the DDBJ cannot afford to enter all the data by itself, and so sequence data including annotation must be processed by each researcher. That is, distributed data entry must be solved. Data entry by authors is not a problem peculiar to this project but a general task which databanks must promote, as already mentioned. In order to make distributed data entry feasible, a user-friendly program for author entry is needed.

The problems that I listed above would be common in any large-scale sequencing project. Of course, the difficulty of a project would depend on the characteristics of the genome that one wants to analyze. The longer the genome size is, the more efficient the experimental and computer systems are required to be. Also, the existence of highly repetitive sequences in eukaryotes would make the physical mapping of their genomes difficult. The genome of *E. coli* would be easier to analyze and so it may be a good exercise for genome sequencing.

A new feature table definition[5] has been just completed by the EMBL and the GenBank with the assistance of the DDBJ in September, 1988, in order to represent in proper form a wide range of new information on DNA sequences that have been discovered in the recent development of molecular biology. A relational database whose schema[4] has been designed by the GenBank may be flexible enough to manage a large quantity of sequence data with the proper representation of information required for large-scale sequencing. Such a database would also make it feasible to search and retrieve sequence segments in various ways. The DDBJ will prepare for this new trend in DNA database reconstruction.

## ACKNOWLEDGMENTS

I would like to thank the GenBank, especially Dr. James W. Fickett, for kindly providing us useful programs and data for managing the DDBJ database. I would like to acknowledge Dr. Hidenori Hayashida, for his efforts in managing data review, and also the staff of the DDBJ.

## REFERENCES

1. Bachmann, Barbara J. "Linkage Map of *Esherichia coli* K-12, Edition 7." *Microbiol. Rev.* **47** (1983):180–230.
2. Burks, Christian. This volume.
3. Cameron, Graham N. "The EMBL Data Library." *Nucl. Acids Res.* **16** (1988):1865–1867.
4. Cinkosky, Michael J., Debra Nelson, and Thomas G. Marr. *A Technical Overview of the GenBank/HGIR Database.* Los Alamos, NM: GenBank/HGIR.
5. DDBJ, EMBL Data Library, and GenBank. "The DDBJ/EMBL/GENBANK Feature Table: Definition, version 1." Mishima, Japan: DDBJ; Heidelberg, FRG: EMBL Data Library; Mountain View, CA: GenBank, IntelliGenetics, 1988.
6. EMBL and GenBank staffs. "A New System for Direct Submission of Data to the Nucleotide Sequence Databases." *Nucl. Acids Res.* **15(18)** (1987).
7. George, David. Personal communication, 1987.
8. Kernigan, Brian W., and Rob Pike. *The UNIX Programming Environment.* New Jersey: Prentice-Hall Inc., 1984.
9. Kohara, Yuji, Kiyotaka Akiyama, and Isono Katsumi. "The Physical Map of the Whole *E. coli* Chromosome: Application of a New Strategy for Rapid Analysis and Sorting of a Large Genomic Library." *Cell* **50** (1987):495–508.
10. Miyazawa, Sanzo. *A Guide to the DDBJ Computer System.* Mishima, Japan: DDBJ, National Institute of Genetics, 1987.
11. Miyazawa, Sanzo. *The Manual of the FLAT Database and Sequence Analysis System for DNA and Proteins.* Mishima, Japan: DDBJ, National Institute of Genetics, 1988.
12. PRF Peptide Sequence Database. Maintained by the Peptide Institute, Protein Research Foundation, 4-1-2 Ina, Minoh, Osaka, Japan.
13. Ritchie, D., and K. Thompson. "The UNIX Time Sharing System." *CACM* **17** (1974):365–375.
14. Uchida, Hisao. Personal communication, 1988.
15. *UNIX User's Manual and Programmer's Manual.* Berkeley, CA: Computer Science Division, Univ. of California, 1984.

16. *VAX/VMS Command Manual.* Massachusetts: Digital Equipment Corporation, 1987.