XI. DATABASE

Activities of the DNA Data Bank of Japan

Sanzo MIYAZAWA and Hidenori HAYASHIDA

A primary task of the DDBJ is, of course, DNA sequence collection. However, in addition to that, we have a wide range of activities; 1) DNA data collection and data entry in collaboration with other databanks, 2) data distribution, including the secondary distribution of the GenBank and EMBL databases in Japan, 3) development of research tools for sequence analysis, and 4) the regular publication of newsletters to inform people of activities of the DDBJ. The DDBJ computer system is open to researchers for data submission and for providing online access to DNA and related databases. In the following, I will briefly report this year's activities of the DDBJ. For details, see "DDBJ News Letter No. 8, 1989" and "No. 9, 1990".

1) Data Entry and Management

Sanzo MIYAZAWA and Hidenori HAYASHIDA

Our data collection began in December, 1986 and is carried out in collaboration with the GenBank and the EMBL Data Library. The collaboration includes projects on designing a new feature table and rebuilding a DNA database. DDBJ has charge of mainly scanning journals published in Japan, and before, data directly submitted by authors had been forwarded to either GenBank or EMBL according to journal split. However, DDBJ started locally processing all direct submissions in October 1989; Direct data submissions which the EMBL Data Library is supposed to process are still forwarded to the EMBL at this stage.

Since we released the first version of the DDBJ database in July, 1987, our database has been released every half year; version 5, which included 679,378 bases in 395 entries, was released in July, 1989, and version 6 including 841,236 bases in 496 entries was released in January, 1990. The DDBJ collected about 305 kb in the year from January 1989 to January 1990. This amount of data is almost equal to 1.25 times the data collected last year, but only 1/40 of the total DNA sequences collected in the world.

However, this may be reasonable, if the number of staff of the DDBJ is compared with those of GenBank and the EMBL Data Library. Each release included a coding sequence database and a peptide sequence database that were extracted and translated from the original DNA sequence database, and files of journal index, accession number index, short directory, and data submission form.

2) FLAT Database and Sequence Analysis System for DNA and Proteins: Release 1.2.

Sanzo MIYAZAWA

We have been developing a search and retrieval system for flat file databases in order to provide simple tools for using DNA and protein sequence databases. This system called FLAT consists of primitives most of which perform a single operation and work as filters in the UNIX system. Its first version, 1.0β, was released in 1988. The most recent version, 1.2, was released in October, 1989. New features added to this release include functions of (1) electronic mail database server and (2) automatic update of databases according to new data sent by e-mail. One can search and retrieve DNA and protein sequence databases by sending search/retrieval commands by electronic mail to the specific address through Junet, Usenet, Internet, and Bitnet. Available databases include PIR, SwissProt and PRF as well as DDBJ, EMBL, and GenBank databases. The EMBL and GenBank databases are updated twice a day by adding new entries which are daily sent by e-mail from those databanks. The DDBJ database is updated as well. Functions provided by the server include keyword search on definition lines, search by author name, journal name and accession number, and entry retrieval by specifying entry names or accession numbers. This FLAT search/retrieval system for sequence databases is designed to be portable among UNIX systems which are available for a wide range of computers from super to micro computers.

For details, see "DNA Data Bank of Japan: present status and future plans" by Sanzo Miyazawa in "*Computers and DNA, SFI Studies in the Sciences of Complexity*, Eds. G. Bell and T. Marr (Reading MA: Addison-Wesley), vol. VII, pp. 47–61, 1989".

Basigin, a New, Broadly Distributed Member of the Immunoglobulin Superfamily, Has Strong Homology with Both the Immunoglobulin V Domain and the β-Chain of Major Histocompatibility Complex Class II Antigen

Teruo MIYAUCHI*, Takuro KANEKURA*, Akihiro YAMAOKA*,
Masayuki OZAWA*, Sanzo MIYAZAWA
and Takashi MURAMATSU*

Lotus tetragonolobus agglutinin (LTA) binds preferentially to early embryonic cells in the mouse. The affinity-purified antibody raised against LTA receptors from embryonal carcinoma cells was used to screen a λgtll expression library of F9 embryonal carcinoma cells, resulting in detection of a cDNA clone specifying a new glycoprotein termed "basigin". The glycoprotein has been suggested as being a transmembrane one, and was found to be a new member of the immunoglobulin (Ig) superfamily. The molecular weight of basigin was largely in the range between 43,000 and 66,000, while that of the peptide portion with a putative signal sequence was inferred to be about 30,000. Significant levels of basigin mRNA were detected not only in embryonal carcinoma cells, but also in mouse embryos at 9–15 days of gestation and in various organs of the adult mouse. The Ig-like domain of basigin is unique, since it has strong homology to both the β-chain of the major histocompatibility class II antigen and the Ig V domain. The number of amino acids between the two conserved cysteine residues is intermediate between those of the Ig V and C domains. Therefore, basigin is an interesting protein in connection with the molecular evolution of the superfamily.

For details, see "*J. Biochem.* **107**, 316–323, 1990".

* Faculty of Medicine, Kagoshima Unversity.