XI. DATABASE

Activities of the DNA Data Bank of Japan

Sanzo MIYAZAWA and Hidenori HAYASHIDA

The 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) to provide on-line access to DNA and related databases, 4) to develop research tools for sequence analysis, 5) to regularly publish newsletters to inform people of the activities of the DDBJ, and 6) to provide 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 people know of such activities of the DDBJ, we published newsletter No. 7, and had a training course in 1988. Newsletters contained articles that described the state of international collaboration among databanks and how to submit data to databanks as well as matters of interest specific for 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 the on-line service of the DDBJ computer system. All these activities provided by the DDBJ are open to anybody irrespective of whether one works for a non-profit organization or not. In the following, I will briefly report this year's activities of the DDBJ.

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 of designing a new feature table and rebuilding a DNA database. Data is currently entered in the GenBank format and fully annotated. Since we released the first version of the DDBJ database in July, 1987, our database has been released every half year; version 3, which included 345,850 bases in 230 entries, was released in July, 1988, and version

4 including 535,985 bases in 302 entries was released in January, 1989. The DDBJ collected about 240,000 bases in the 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; this number 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. Release 2 and later releases included the 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 1.0 β. 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; a filter program reads a line from standard input, processes it and then writes some output into standard output. Some basic commands available in FLAT perform single operations such as 1) extracting specified types of records from database files, 2) searching for strings in each entry of database and, if found, outputing those entry names, 3) performing "and", "or", and "xor" in respect of entry names, and 4) extracting specified entries from database files. These filters may be combined with the UNIX pipe to perform complicated jobs; 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. This is a typical approach for designing programs in the UNIX system. Strings for these programs are specified in the regular expression, so that one can search and retrieve entries in databases by fuzzy key words and entry names. The "seqgrep" program also allows users 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, so that they are flexible enough to support many formats of databases and to easily keep up with format changes which often occur. At present, the Genbank, EMBL, PIR (Protein Identification Resource) and PRF (Protein Research Foundation) data formats are sup-

ported. However, this approach tends to trade computational speed for flexibility, and so applications whose processing speed is critical are written in the C language. A program "getgb", which extracts specified entries from databases, uses a pseud index file to quickly find the location of entries in a flat database file. 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. Its β version of release 1.0 was released in 1988.

3) The Qanalys Sequence Analysis System for Molecular Evolution: Version 1.0 β Hidenori HAYASHIDA

Programs for sequence analysis were developed to study molecular evolution, including programs for the calculations of base or amino acid composition and similarity between sequences, a sequence alignment program, and a display program for postprocessing. The β version of release 1.0 was released in 1988.

4) A Guide to the DDBJ Computer System: the "getinfo" Command Sanzo MIYAZAWA

An online help program called "getinfo" was made to provide databank staff and users an easy way to get necessary information. One may use the "getinfo" to learn how to submit DNA data and to which databank the data should be submitted, and even to get a data submission form. The "getinfo" apparently mimics the help utility of the VAX/VMS system. However, unlike the VMS help utility, each help information is stored as a flat file and organized into a tree-like structure, if necessary, by using symbolic links 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 specific topic and, if available, a list of help items at the next level and prompts users to choose one of them. A pager program, pg or less, which is available in the UNIX system is used to print files on a terminal, so that one may read a file page by page, and may save it in a file, if necessary. A hard-copy version of the guide to the DDBJ computer system was published in 1988.

For details, see "DNA Data Bank of Japan: present status and future plans" by Sanzo Miyazawa in "*The interface between Computational Science and Nucleic Acid Sequencing, Santa Fe Institute Studies in the Sciences of Complexity*, Eds. G. Bell and T. Marr (Reading, MA: Addison-Wesley), vol. VIII, 1989".

Estimation of the Average Energy Increments of Protein Native Structures due to Amino Acid Exchanges and Its Use in Evaluating a Substitution Probability Matrix for Homology Search

S. MIYAZAWA and R. L. JERNIGAN*

The energy change due to an amino acid exchange in protein structures was estimated on average, and used to evaluate a substitution probability matirix, which was then used to measure the similarity between protein sequences. In a statistical sense, each type of amino acid residue is found at a particular location in the three dimensional structure of proteins; non-polar residues are more often found in the non-polar environment of the protein core and polar residues on the protein surface. Residues surrounding an amino acid in protein structures are specific to the type of amino acid. We consider a typical or average protein which satisfies statistical features observed in a large set of protein structures, and the average energy increment due to an amino acid exchange in such an average protein. In a previous study (Miyazawa & Jernigan, 1985), we estimated the effective inter-residue contact energy of each type of amino acid pair for proteins in solution from 18,192 residue-residue contacts observed in 42 globular proteins, and also compiled the types of residues and their average numbers in contact with each type of amino acid. By using their contact energies and statistical data of surrounding residues, we evaluated the average energy increment due to an amino acid exchange in the average protein for each pair of the 20 kinds of amino acid residues. These estimates show reasonable characteristics of physico-chemical similarities of amino acids. The average energy increment due to an amino acid exchange would be a good measure of structural instability caused by an amino acid exchange. We assume that the fitness of an amino acid exchange in the evolutionaly process is proportional to the Boltzmann factor of the average energy increment due to the amino acid exchange. The substitution process of codons consists of two steps, mutation and selection at the DNA level, and selection at the protein level. The process of codon substitution is assumed to be in equilibrium and for simplicity its rate at the DNA level is assumed to be proportional to the equilibrium frequency of the codon. Thus a transition matrix of amino acid substitution for a long time interval which corresponds to 250 PAM (accepted point mutations per 100 residues) is generated from

* National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA.

that of codon substitution for a short time interval, and the log of each element of the 250 PAM matrix divided by the amino acid frequency is calculated as a scoring matrix in the same way as Dayhoff et al. (1978). The correlation coefficient between this scoring matrix and Dayhoff's scoring matrix (MDM_{78}), which corresponds to the 250 PAM substitution probability matrix calculated from amino acid substitutions observed in closely related proteins, is about 0.55 for all off-diagonals but 0.82 when infrequent amino acids of met, trp, cys and tyr are excluded. The poor correlation for all off-diagonals may result from statistical errors due to small numbers of substitutions. This scoring matrix has demonstrated the same degree of detection power of sequence homology as the Dayhoff's scoring matrix. These results indicate that the average energy increments due to an amino acid exchange as estimated here reflect the structural instability caused by the amino acid exchange. This manuscript is in preparation.

Publications from Genetic Resources Section, Genetic Stocks Reseach Center

Shin-ya IYAMA

The following publications were released from the Genetic Resources Section this year.

1. "Silkworm Strains in Japan". (Compiled by S. Iyama, H. Doira and A. Murakami) 181 pages (in Japanese). This was completed in cooperation with the Silkworm Genetic Resources Subcommittee, the Japanese Society of Sericulture. The catalogue listed 943 entries from 18 locations in Japan, with special emphasis on experimental strains. Each stock was described with its strain name, location kept, genotype, historical record of establishment and acquisition, genetic characteristics, and information on maintainance and distribution. A list of gene symbols with the description of their characteristics was added for the convenience of research workers in using materials.

2. "Rice Genetics Newsletter Vol. 5, 1988". 162 pages. (in English). This volume contained the following:

(1) Report of the Committee on Gene Symbolization, Nomenclature and Linkage Groups, describing 8 newly registered genes, 13 newly adopted gene symbols and 16 marker genes newly confirmed for their linkage relations.