# XI. DATABASE

## Constructing a DNA Database

Sanzo MIYAZAWA

A data entry system and data search/retrieval system for DNA sequence data were developed on a unix operating system. We built a data entry system by utilizing the SCCS (Source Code Control System) available in the unix operating system. The SCCS is a source code management system with such features as version control and exclusion control. Version control means that a record is kept with each set of changes; of what the changes are, why they were made, and who made them and when. Ecclusion control means that only one person can modify data at a time. Both are critical in data entry with more than one user. Quality control of data is also important in any database. We installed programs made by Dr. J. Fickett in the GenBank for error checking of data. These programs check whether the following features are illegal or not; record format, journal name, start and stop codons, and codon frame. A codon table which depends on an organism is maintained as a database. Taxonomy records in the data are automatically generated by using a taxonomy database which is also maintained by GenBank. More extensive automatic error-checking should be developed on the basis of studies of the telationship between DNA sequence pattern and function. A simple search and retrieval system which uses flat files and therefore called "flat" has also been developed. Some basic tools are available as filters in the unix system to output specified types of records, to search strings expressed in regular expression over database and output names of entries which include such strings, to take "and", "or" and "xor" over sets of entry names, and to cut out specified entries from the database. By joining such filters together with "pipe", one can search and retrieve entries from the database by keywords such as author name, journal name, title, organism name, gene name, and any combination of such items. This "flat" system is designed to be portable and easy to maintain at the cost of speed; it is portable among unix systems which are available for a wide range of computers from super to personal computer.