CABIOS Vol.9 no.4 1993
Page 474

A simple method for accessing and transferring nucleotide database files

Gregory Smutzer

Introduction

Many software programs are available to the molecular biologist for analysis of nucleic acid sequence information. Many researchers utilize sequence information from comprehensive databases in conjunction with these programs. Direct access to CD-ROM data releases (ENTREZ: Sequences, National Center for Biotechnology Information; or GenBank® by IntelliGenetics, Inc.) with a personal computer allows the researcher to quickly import nucleotide sequences without a mainframe computer. GenBank releases from IntelliGenetics, Inc. contain limited utilities for data retrieval (e.g. GBMENU) while the replacement CD-ROM from GenBank, ENTREZ: Sequences, includes text retrieval software linked to MEDLINE citations which can be used to extract files interactively or as a linkable library of function calls. GenBank files, however, contain descriptive annotations which must be partially or fully removed before sequence files can be accessed with many software programs. Furthermore, programs such as DNASIS® (Hitachi America, Ltd) and PC Gene® (IntelliGenetics, Inc.) are designed to use their own CD-ROM source data files that are in a special format. A method is presented for rapidly accessing data files from GenBank CD-ROM releases for use with a variety of software programs.

Methods

Primer Detective® software (v. 1.01, CLONTECH Laboratories) is utilized to import DNA sequences from CD-ROM releases of GenBank (Release 67.0, floppy disk format) for subsequent use with a variety of software programs. This method should also be applicable with ENTREZ: Sequences CD-ROM releases. Briefly, the GenBank Utilities Menu of Primer Detective is used to select sequence files and save them in the Primer Detective sequence directory. Because the sequence files contain descriptive annotations that are not recognized by many software programs, WordPerfect® (v. 5.1, WordPerfect Corporation) is used to remove these annotations from the beginning of GenBank files by blocking and deleting non-sequence information. (Other word processing programs may also be used for this purpose.) The other directory function of WordPerfect must be used to open files in the Primer Detective sequence directory. WordPerfect sequence files are then saved as DOS text files by pressing the Control-F5, 1, and 1 keys. The modified sequence file can be directly saved to the appropriate sequence directory (i.e. DNADATA/GeneY.seq) with the word processing program or can be moved to the appropriate directory with the copy commands in WordPerfect or DOS. The sequence file must possess the three letter extension .seq for recognition by many software programs. Nucleotide sequence files can also be accessed from GenBank releases since this databank contains retrieval software (e.g. GBMENU, GenBank release notes, 1991) and descriptive annotations can be removed with WordPerfect as described above. WordPerfect can also be used to remove non-software annotations from files obtained through an electronic mail server.

Some software programs have an upper limit to the number of nucleotides that can be analyzed in a single file (Lowe, 1990; Hillier and Green, 1991). DNASIS, for example, can only analyze the first 63 000 nucleotides in a sequence, but Primer Detective has no apparent full-size limitations for the transfer of genetic information to word processing programs (up to 229 400 bases have been transferred). Large sequence files accessed with either Primer Detective or GenBank retrieval software can be broken up into smaller files with WordPerfect so that all potential sequence information from the original file can be transferred to the appropriate software program for analysis. In addition to its intended use in primer design, Primer Detective has been successfully utilized for the transfer of sequence information to a variety of software programs for subsequent data analysis. Utilization of this program with a word processor may also be useful for handling extremely large nucleotide sequence files.

Acknowledgements

GenBank® is a registered trademark of the US Department of Health and Human Services. This work was supported by NSF grant BNS 91-14153 and the Johnson & Johnson Focused Giving Program.

References

GenBank® Release Notes, release 67.0, Floppy Diskette Distribution CD-ROM Version, March 1991. IntelliGenetics Inc. Mountain View, CA.

Hillier,L. and Green,P. (1991) OSP: a computer program for choosing PCR and DNA sequencing primers. *PCR Methods Applic.*, 1, 124–128.

Lowe,T. (1990) *Primer Detective® User Manual.* CLONTECH Laboratories, Inc. Palo Alto, CA.

Monell Chemical Senses Center, 3500 Market Street, Philadelphia, PA 19104-3308, USA