The protein identification resource (PIR)

Kathryn E.Sidman, David G.George, Winona C.Barker and Lois T.Hunt

National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, N.W.
Washington, DC 20007, USA
Submitted August 17, 1987

ABSTRACT

The Protein Identification Resource consists of an integrated computer system composed of a number of protein and nucleic acid sequence databases and software designed for the identification and analysis of protein sequences and their corresponding coding sequences. The PIR serves the scientific community through on-line access, distributing magnetic tapes, and performing off-line sequence identification services for researchers.

INTRODUCTION

The Protein Identification Resource is sponsored by the National Biomedical Research Foundation (NBRF) and supported by the Division of Research Resources of NIH. This Resource includes protein and nucleic acid sequence databases and a powerful computer system for protein identification and for retrieval of associated information. Although the resource is primarily concerned with problems related to the analysis of protein sequence information, facilities are also available for the analysis of genetic sequences.

Central to the resource are the sequence databases. The NBRF-PIR Protein Sequence Database includes the following: all substantially sequenced proteins, including sequences translated from nucleic acid sequences; bibliographic citations for amino acid sequences, nucleic acid sequences, X-ray crystallography, active site determination, etc.; annotations identifying posttranslational modifications, active sites, signal sequences, activation cleavages, disulfide bonds, intron locations, etc. An auxiliary file includes sequences in preparation and fragmentary and "hypothetical" sequences. The NBRF Nucleic Acid Sequence Database contains entries annotated to show the locations of protein coding regions.

The PIR performs three major types of activities that are designed to serve the biological community by making the technology of the Resource available to scientists worldwide at minimal cost: the on-line PIR computer system, the distribution of PIR databases and software on magnetic tapes, and the off-line sequence identification service.

On-line Access

In addition to its own protein and nucleic acid sequence databases, the PIR makes recent releases of several other databases available to its on-line users. Among these are the GenBank(R) Genetic Sequence Databank and the EMBL Nucleotide Sequence Data Library.

The sequence databases are accessed by the PSQ and NAQ programs. Sequence entries can be retrieved by sequence, by species, by author, by superfamily classification, by taxonomic classification, and by text terms. Sequences can be back-translated and restric-

tion enzyme cut points located. Amino acid compositions or codon usage tables can be compiled. Entries can be copied from the database to a user-owned file for modification. Confidential user-created sequence files can be used with the retrieval systems as well as with other programs. The SCAN command of PSQ instantaneously finds short identical segments. The MATCH command can be used to look for well-defined sequence patterns as the user can specify the number of mismatches allowed as well as which regions of the test string must match exactly. Some other database searching programs available include ISEARCH, FASTP, and FASTN. The ISEARCH program is useful for searching ambiguous segments in which two or more amino acids may be found at some positions. The FASTP and FASTN programs of William Pearson (1) can use an entire sequence to rapidly search the database, producing near-optimal alignments of segments identified in the search. The ALIGN program can produce pairwise optimal alignments of sequences up to 1200 residues in length. Program PRPLOT plots hydrophobicity or other characteristics.

Newly available sequence analysis programs (Table 1) include programs from Minoru Kanehisa's IDEAS package (2), most of which were adapted for PIR's system by Carolyn Bucholtz of the CSIRO in Australia.

The databases and programs of the PIR are accessible via TYMNET. For 1987, the annual fee is \$350 for qualified researchers at nonprofit institutions. Others pay a \$50 set-up fee and \$50 per hour connected.

Magnetic Tapes

The NBRF-PIR Protein Sequence Database and the NBRF Nucleic Acid Sequence Database are distributed on magnetic tape in VAX/VMS and ASCII card image formats on a quarterly basis. The PSQ (Protein Sequence Query) and NAQ (Nucleic Acid Query) programs for browsing and information retrieval are distributed with the respective databases in VAX/VMS format. The VAX/VMS version of the nucleic acid tape includes programs

Table 1. New Programs

SEQFP (IDEAS): Kanehisa's adaptation of the Wilbur-Lipman diagonal method for the comparison of two sequences.
SEQHP (IDEAS): Kanehisa-Goad local homology method.
MTXANL (CSIRO): Reisner-Bucholtz local homology method (3).
HPLLOT (IDEAS): Hydrophobicity plot including 5-point polynomial smoothing.
ALOM (IDEAS): Prediction of membrane associated regions.
CHOFAS (IDEAS): Chou-Fasman method of secondary structure prediction.
DELPHI (IDEAS): Garnier method of secondary structure prediction.
ALNED: A screen-oriented interactive alignment editor that incorporates a pairwise sequence alignment algorithm but allows the resultant gaps (insertions/deletions) to be interactively adjusted.
ALNPLT: Plots % similarity or Wu-Kabat variability.
HIGHCOMP: Searches for compositional similarities.

to reformat GenBank(R) and EMBL nucleotide databases to be accessible by the NAQ program.

The NBRF-PIR Sequence Analysis Software is designed to run on a VAX-11/780 operating under VAX/VMS. All programs are written in VAX-11 Fortran (a superset of ANSI Fortran 77), with the exception of the Lipman-Pearson programs, which are written in VAX-11 C. Release 3.0 contains the following programs: FASTP, FASTN, RDF, SEARCH, ALIGN, RELATE, DOTMATRIX, PRPLOT, ISEARCH, IALIGN, CHOFAS.

The databases and software of the PIR are in the public domain and may be freely copied and redistributed provided the Protein Identification Resource is acknowledged as the source. There is a \$200 fee per tape release.

Off-line Services

The staff of the PIR performs off-line sequence identification services on a cost-recovery basis. The fee is generally \$200-\$300 depending on computer and staff time required. The programs used are described above. Sequences submitted for analysis are kept strictly confidential. Services requested vary from simple database retrieval and searches to complex alignments and evolutionary topologies.

ACKNOWLEDGMENTS

The Protein Identification Resource is supported by grant RR01821 from the Division of Research Resources of the National Institutes of Health.

Inquiries regarding the services of the PIR should be directed to K.E. Sidman, Protein Identification Resource, at the above address. Prices subject to change. BITNET address: PIRMAIL@GUNBRF.

REFERENCES

1. Lipman, D.J. and Pearson, W.R., *Science* 227, 1435-1551, 1985.
2. Kanehisa, M., IDEAS: Integrated Database and Extended Analysis System for Nucleic Acids and Proteins (User Manual), National Institutes of Health, Bethesda, MD, 1985.
3. Reisner, A.H. and Bucholtz, C.A., *Nucl. Acids Res.* 14(1), 233-238, 1986.