Volume 14 Number 1 1986 Nucleic Acids Research

The protein identification resource (PIR)

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

National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, NW, Washington, DC 20007, USA

Received 29 August 1985

ABSTRACT

The Protein Identification Resource, which provides the scientific community with an efficient on-line computer system designed for the identification and analysis of protein sequences and their corresponding coding sequences, has been established. The resource consists of an integrated computer system composed of a number of protein and nucleic acid sequence databases and the software necessary to analyze this information effectively.

INTRODUCTION

Under the sponsorship of the Division of Research Resources of the National Institutes of Health (NIH), the National Biomedical Research Foundation (NBRF) has established the Protein Identification Resource (PIR), a resource designed with the necessary capabilities to aid investigators in the identification and interpretation of protein sequence 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 the corresponding genetic sequences. The resource is the direct outgrowth of an on-line protein and nucleic acid database system that has been operational since 1980 (1,2,3). It consists of several protein and nucleic acid databases and a set of interactive computer programs integrated into a sophisticated computer system that is completely documented with on-line help messages; written documentation is also provided. The databases include the NBRF Protein Sequence and Nucleic Acid Databases, the GenBank(TM) Nucleic Acid Database, and the European Molecular Biology Laboratory's (EMBL) Nucleotide Data Library, as well as several other special purpose databases. Additional databases may be added as they become available. The PIR on-line system is

Information about procedures and charges for ordering tapes of the programs or databases, or for obtaining on-line access, can be obtained by writing to Ms. K. Sidman, Protein Identification Resource, at the above address.

accessible by direct telephone connection or through the TYMNET communications network, which also permits international access. All NBRF programs are written in VAX Fortran-11, a superset of Fortran-77. The system also includes FASTP, a protein database searching program, developed by William Pearson of the University of Virginia and David Lipman at the NIH (4); this program is written in the C computer language. All NBRF programs and databases of the PIR system are in the public domain and are available on magnetic tape. They are distributed either on a VAX-labeled tape or on a nonlabeled tape for non-VAX computer systems. The software, however, is designed specifically for VAX/VMS systems and requires considerable modification to run on other computer systems.

THE PIR ON-LINE SYSTEM

The Protein Sequence Query (PSQ) Program and the Nucleic Acid Query (NAQ) Program, the main retrieval programs of the NBRF Protein and Nucleic Acid Sequence Databases, formed the core programs for the development of the PIR on-line system. These programs were initially developed to aid our scientists in the maintenance of the databases; they provide very rapid access to the database information and, with the addition of several sequence manipulation routines, serve as extremely powerful research tools. Entries may be located and retrieved by entry title, biological source, taxonomic classification, keyword phrases and other text material, author name, journal or other reference citation, and by sequence length, molecular weight, and amino acid composition. The subset of the database thus located can be saved, and all of the display and sequence manipulation routines can be made to operate on this subset or 'current list.' These programs also contain routines to search for open reading frames, to translate and back-translate nucleotide and protein sequences, to splice and edit sequences, to compute codon usage, to compute amino acid and nucleotide frequency tables, and to search for restriction enzyme cut sites in nucleotide and back-translated protein sequences. The sequence searching routines of these programs are described below. The PIR programs are designed so that information can conveniently be exchanged between them. The programs can directly read sequences from any of the databases and also access sequences in user-owned files. Facilities are provided that allow the user to create and/or modify sequence files easily and that allow files to be transferred to and from the PIR computer. The identification codes of entries on the 'current list' of the PSQ and NAQ programs can be stored in files that can be read by other programs and the action of these programs can thus be restricted to a selected subset of the database.

DATABASE SEARCHING USING THE PIR SYSTEM

We strongly recommend the searching of protein sequence data rather than nucleic acid sequence data whenever possible. Because the protein databases contain information obtained by the translation of published nucleotide sequences, in most cases searching the protein database for translated sequences will identify any sequences that can be identified by a direct nucleic acid database search. Furthermore, protein database searching is not only faster by a factor of at least three but is generally considered to be much more sensitive. Nucleic acid database searches can identify only very similar sequences, whereas techniques developed for protein database searching allow very distant protein relationships to be discovered. Although several routines are provided that directly search the nucleic acid sequence database, the PIR principally supports protein database searching.

The PIR on-line system provides four methods for searching protein databases: the SCAN and MATCH commands of PSQ, the SEARCH program, and the FASTP program. The particular method used to search the database depends upon the amount of sequence information the user has concerning a particular sequence and the type of relationship the user wishes to detect. (See refs. 5, 6, 7, and 8 for recent reviews of sequence searching and comparison methods.) The simplest searching method is to look for sequences that are identical with the test piece. The most effective way to do this is to select a subsequence of five to seven residues from the test sequence and search the database for any exactly matching subsequences. Only about 10% of all possible pentapeptides and less than 1% of all possible hexapeptides are present in the current NBRF protein database, making the occurrence of any identically matching hexapeptide a rare event. A string of six consecutive amino acids is usually sufficient to identify a closely related protein. The SCAN command utilizes a precompiled tripeptide catalog and a simple lookup algorithm to locate all occurrences in the database of amino acid strings from 3 to 30 residues in length. The response time is instantaneous from the point of view of the user. If the protein is in the database, it can be found immediately by using this routine.

If no identically matching subsequences are detected, the next step is to search the database for subsequences, allowing for mismatches between the test string and the database subsequences. In most cases this will be accomplished using the FASTP or SEARCH programs described below. However, the MATCH command of PSQ and NAQ provides a useful facility not available through the other searching methods. The MATCH command allows the user to search the database for test strings of up to 30 residues and to specify an allowable number of mismatches. It also allows the user to select certain regions of the test segment where the database sequences must match exactly. This facility makes the MATCH command very effective at locating entries containing well-defined sequence patterns.

The NBRF SEARCH program (9) has been used as the standard database searching method for the last 10 years. Rather than scoring sequence comparisons based on the number of mismatches or matches, this program uses a similarity scoring matrix to evaluate the contribution of aligned pairs of amino acids to the total alignment score. The SEARCH program compares a segment of a specified length with all other segments of the same length in the database and computes a score based on the specified scoring matrix; the highest scoring segment comparisons are selected.

A more sophisticated searching algorithm has been described by Wilbur and Lipman (10,11). Lipman and Pearson (4) implemented a modified version of this method, designed specifically for searching protein databases, and have donated a version of this program, FASTP, to the PIR. Although it only approximates a full-scale search of the database allowing for gaps, it is extremely fast and very sensitive for detecting sequence similarities. The program runs faster than our SEARCH program and, in most cases, is the method of choice for searching the protein database. There are instances when one would want to use the SEARCH program rather than FASTP, however. FASTP is designed to locate the best contiguous regions of similarity between the test sequence and any database sequence. If there is more than one such region within a protein, only the best is chosen. There are a number of biological mechanisms for producing noncontiguous regions of sequence similarity within proteins, such as the insertion or deletion of large sequence segments, internal gene duplication, or gene splicing or transposition. The SEARCH program can more readily identify multiple regions of similarity within sequences than can FASTP.

SEQUENCE COMPARISON METHODS

Database searching programs are most effectively used as methods of selecting sequences for further comparison with the test sequence; these methods yield a number of possible candidates that may eventually be demonstrated to be related to the test sequence. Once the selections have been made the next step is to produce an alignment of the test sequence or portions of the test sequence with portions of each of the selected database sequences. The PIR on-line system provides two methods for aligning sequences, the ALIGN program and the COMPARE program, which have been described in more detail elsewhere (2,3,9). Both methods are based on the algorithm of Needleman and Wunsch (12). The ALIGN program also generates statistics based on the comparison of random permutations of the two sequences. Because of the method of implementation, the ALIGN program cannot be used to compare extremely long sequences (greater than 1,200 residues). The COMPARE program is an approximate method that allows the alignment of long, closely related sequences. Both the ALIGN and COMPARE programs will operate on the nucleic acid sequence data as well as on the protein data.

The present on-line PIR system provides many of the necessary facilities for the analysis of sequence data. To provide a resource that effectively supports scientific research in a field changing as rapidly as molecular biology, the PIR must adapt as quickly as the biological understanding grows. This can be accomplished only by continually upgrading the presently available programs and by adding additional capabilities as necessary. In the future, both NBRF-developed software and that contributed by outside sources will be made available on the PIR on-line system.

ACKNOWLEDGMENTS

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

REFERENCES

1. Dayhoff, M.O., Schwartz, R.M., Chen, H.R., Barker, W.C., Hunt, L.T. and Orcutt, B.C. (1981) DNA 1, 51-58.
2. Orcutt, B.C., George, D.G., Fredrickson, J.A. and Dayhoff, M.O. (1982) Nucl. Acids Res. 10, 157-174.
3. Orcutt, B.C., George, D.G. and Dayhoff, M.O. (1983) Annu. Rev. Biophys. Bioeng. 12, 419-441.
4. Lipman, D.J. and Pearson, W.R. (1985) Science 227, 1435-1441.
5. Orcutt, B.C. and Barker, W.C. (1984) Bull. Math. Biol. 46, 545-552.
6. Waterman, M.S. (1984) Bull. Math. Biol. 46, 473-500.
7. Kruskal, J.B. (1983) SIAM Rev. 25, 201-237.
8. Sankoff, D. and Kruskal, J.B. (1983) Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, Addison-Wesley, Reading.
9. Dayhoff, M.O., Barker, W.C. and Hunt, L.T. (1983) Methods Enzymol. 91, 524-545.
10. Wilbur, W.J. and Lipman, D.J. (1983) Proc. Natl. Acad. Sci. USA 80,726-730.
11. Wilbur, W.J. and Lipman, D.J. (1984) SIAM J. Appl. Math. 44, 557-567.
12. Needleman, S.B. and Wunsch, C.D. (1970) J. Mol. Biol. 48, 443-453.