Personal access to sequence databases on personal computers

Philip E. Doggett and Frederick R. Blattner

DNASTAR Inc., 1801 University Ave., Madison, WI 53705, USA

Received 1 August 1985

ABSTRACT

A comprehensive package of software has been developed to access nucleic acid and protein sequence databases on stand-alone IBM personal computers. The software combines keyword search on the annotation fields of the data with pattern matching algorithms on the biological sequences. Sequences containing complex sites like promoters or kink sites can be identified as well as sequences that are similar to a query sequence. Protein sequences with particular patterns of amino acids such as hydrophobic regions can be identified as well. Considering the relatively inexpensive hard disks now available, personal computers have become a cost-effective alternative to mainframe processing for sequence databases.

SEQUENCE DATABASES ON MICROCOMPUTERS

Establishment of publicly funded resources for systematic collection and dissemination of biological sequence data such as GenBankTM (1) and Protein Identification Resource (PIR) (2) has created a broad range of research opportunities based on the analyses of these data.

The DNASTAR sequence analysis package provides access to these genetic data on a stand-alone personal computer. These programs can be operated as a stand alone package or within the context of the complete 37 program DNASTAR molecular biology system. All programs work with databases in the GenBank format. GenBank and PIR data (reformatted to agree with GenBank) are available on data cartridge tape from DNASTAR.

These programs are available as a package from DNASTAR Inc., 1801 University Ave., Madison, Wisconsin, (608) 251-9685; or in Europe from DNASTAR Ltd., 8 Walpole Gardens, Chiswick (UK), Tel. 1-994-0619. The price is \$3,000 (US). This includes both PIR and GenBank data and updates on cartridge for a period of one year.

GENEMAN

The main program for scanning the databases is called GENEMAN. In order to perform a search the user must specify: (1) What database or subdatabase to access; (2) What output to produce; (3) What search criteria to apply. Each of these choices is made from a separate menu.

DATABASE SELECTION

The program uses an index method to access complete databases or searches can be limited to subcategories specified by typing the name of a subindex file. Off-the-shelf subindex files are provided for standard phylogenetic categories. In addition, subindices can easily be created by users to fit their specific interests. Subindex files are also used to specify a second search of the data entries identified in a first search.

OUTPUT OPTIONS

The GENEMAN program can send simultaneous output to the screen, to the printer, to single-sequence data files and to a multi-sequence file in database format (termed a subdatabase). Specification of which fields will be directed to which outputs are made by by a cursor oriented selection menu. The four available outputs are listed across the top and the fields of data are arrayed down the side. The cursor is moved to the desired intersection points to tag which fields are to go to which output. In this way any subset of fields can be directed to any medium.

This flexible system allows the user to catalog and copy a database as well as to search it. In addition, the selective copy produced by the subdatabase option when indexed can also serve as an independent stand-alone database. If a database with fewer fields and/or fewer sequence entries is satisfactory, the user can save disk space by removing the original.

SEARCH CRITERIA

The GENEMAN program provides a specially designed query language that allows specification of searches for entries containing keyword matches in annotation fields, as well as

sequences that contain a "site" in the sequence field.

The GENEMAN menu prompts the user for search terms one at a time so that the query language need not be memorized. Each term of the search criteria has the general format:

what to search for / place or places to search for it

The / (slash) character separates the keyword from a list of database field names or abbreviations. Several subdivision names can be separated by slashes which in this context mean the logical or, thus

ribonuclease/def/ref

means search for the keyword RIBONUCLEASE in either the definition or reference subdivisions.

After the first term has been entered the computer prompts for another term separated from the first by "or". A blank term causes the next connective to be " and " followed by more "or"s. The series is continued until two blank terms are input. The search criteria entered become a logical function in the following canonical format:

(a OR b OR c ...) AND (d OR e OR f ...) AND ...

where a, b, c, d, e, f are TERMS and the disjunction of terms within a pair of parenthesis is called a GROUP. Each term may be preceded by the word NOT to give a fully generalized logical function.

Keyword searches in general are for separated words and the case of letters is ignored. Various wild cards and exception symbols are provided that allow searches for imbedded keywords and other less usual cases. You may place search specifications into a file, which will be read by GENEMAN when the search term "=filename" is encountered.

RECOGNITION SITES

Many biological recognition sites in DNA consist of consensus sequence blocks separated from one another by plus or minus a distance. A classic example is the *E. Coli* promoter which consists of minus 35 and minus 10 hexamers usually separated by a 17 ± 1 base interval (6). The DNASTAR site search capability is designed to locate sequences in the database that contain such sites. Implementation of this scheme occurs whenever the sequence field is specified in a search term. For example, a search for a

possible promoter sequence could be defined:

```
TTGACA >> (22,24) TATAAT /Seq
```

which pulls out of the database any sequence which includes TTGACA (the minus 35 consensus) upstream of TATAAT (the minus 10 consensus) with a spacing of 22 to 24 residues measured between the 5' ends of the two sequence blocks. The default search covers both direct and reverse complement strands; a + or ~ preceeding the term would be used to confine the search to only one strand. You may allow a certain percentage of mismatches using the following format.

```
TTGACA %83 >> (22,24) TATAAT %50 /Seq
```

In order to extend the consensus sequence to include the most likely variations at specific positions within the consensus, ambiguous residues can be specified in the standard degeneracy codes.

```
TTGACATNTW %83 >> (17,19) TNTGNTATAAT %50 /Seq
```

In this example W refers to A or T and N can be any residue.

A more precise method for capturing the pattern of variation in a site is through a matrix that gives an individual weight to each residue at each position in the sequence. Since it is impractical to key in a matrix from the keyboard, this is read in from a file (m35.mat and m10.mat in this example).

```
m35.mat %83 >> (17,19) m10.mat %50 /Seq
```

The following example matrix describes a donor RNA splice site.

	G	A	C	T
c	-41	-13	-12	-13
c	-33	-12	-3	-31
g	48	13	13	-153
.g	30	*	*	*
t	*	*	*	30
a	-8	9	-22	-21
a	24	32	-59	-75
g	52	22	10	19
c	-22	-20	-14	-23
t	-20	-20	-22	-17
c	-25	-16	-8	-30
c	-15	-22	-8	-28
c	-29	-26	-9	-15
c	-16	-10	-7	-13
c	-13	-18	-7	-8

The numerical entries in each row are the weights assigned to each residue at that position. The * means a residue is never allowed. In the case of a matrix, percent matches are scaled so

that 100% corresponds to the consensus sequence (the most likely residue at each position) and zero % corresponds to the anti-consensus (the least likely residue at each position). When the word NOT precedes a sequence search term, it means to select only those entries which do not include the matching sequence. The orientation of site components can be varied: A>>B means A precedes B; A << B means A follows B; A <> B and A>< B mean that order doesn't matter. Orientation is always interpreted in the 5' to 3' direction on the strand being searched. You may leave the separation specification out of the relation to signify that the any separation is allowed.

As another example, the following relation can be used to locate DNA sequences that are "kinky" (7):

AAAA >> (9,12) AAAA >> (9,12) AAAA >> (9,12) AAAA /Q

It finds sequences that have regularly spaced areas of AAAA at separations corresponding to the pitch of the DNA helix.

GAATT(>) (200,210) GGATCC>> (1000,1100) (GAATTC)

could be used to find sequences with a restriction fragment pattern. Sequence and keyword searches can be mixed in any combination.

SEQUENCE SIMILARITY SEARCH

NUCSCAN provides the ability to search the GenBank for sequences that are similar to a query sequence up to 4096 residues including the introduction of gaps to optimize the alignment. NUCSCAN imposes no limit on the sizes of sequence in the database that can be scanned. PROSCAN performs the same function on the PIR. These programs utilize the approaches and some segments of code developed by Wilbur and Lipman (3) and Pearson and Lipman (4). Both programs operate within the GENEMAN data management environment discussed above. Database or subindex selection are basically the same for all three programs, although their outputs are different.

Because a prohibitive amount of time would be required for a rigorous calculation of similarity between a query sequence and every sequence in the database, the programs perform two passes through the data. The first pass employs the hash method of Wilbur and Lipman (3) to look for runs of identical matching nucleotides

or amino acids of a fixed length K (k-tuples). A preliminary match score is computed by locating the area of highest concentration of k matches and counting them. Since it is quite difficult to determine what is a significant match score given that biological as well as statistical insight is required, the program retains a list of the 400 sequences receiving best preliminary scores. A full alignment score is then computed for the top sequences on the list. This is done with the Wilbur and Lipman algorithm in the case of NUCSCAN and the Lipman and Pearson algorithm for PROSCAN. In the latter case an amino acid similarity matrix is used to increase the sensitivity of the search. This matrix allows amino acid sequences that don't match exactly to receive a score depending on their degree of relatedness. The standard matrix used is the PAM score matrix (4). Other matrices can be used that exploit the similarity of codons or the similarity of physical characteristics of the amino acids. The number of full alignments to be performed can be selected by the user.

SCAN RESULTS

After the first and second passes through the data, a ranked list of sequence names is presented on the screen in the order of match score. By use of the cursor keys the entire list may be inspected. Individual sequences or groups of sequences can be selected for further processing and inspection of the alignment. Sequences may be aligned, database entries displayed, individual sequences extracted from the database, and subindex files may be constructed from this option.

This output screen provides user interaction is most important for effective use of the database program which is very difficult to predict in advance the number of sequences and the range of scores that will emerge from a given search. It is equally impractical to set an *a priori* acceptance criterion for significant match. Sometimes the best match found by the program will be completely insignificant. In other cases, less significant match scores will have great significance biologically. This can really be determined only by the investigator.

EXAMPLES

Two examples of the NUCSCAN and PROSCAN outputs are

```

          10v      20v      30v      40v      50v      60v
is5 TAAAGATGGTAATTCCTGGTTATACTCGTGTGTTGGGAACGCAGAACTAGGGAAG-----GTGCG
    TAAAGATGGTAATTCCTGGTTATACTCGTGTGTTGGGAACGCAGAACTAGG  A      GTGC
h.h.TAAAGATGGTAATTCCTGGTTATACTCGTGTGTTGGGAACGCAGAACTAGGACAACTCTTATCAAGAGTGC
          850^      860^      870^      880^      890^      900^      910^

```

PROSCAN OUTPUT (partial)

Transcription termination factor rho from Escherichia coli aligned with
ATPase (EC 3.6.1.3) alpha chain from Tobacco chloroplast

```

          130v      140v      150v      160v      170v      180v      190v
rho FDKPENARNKILFENLTPLHANSRLRMERGNSTEDLTARVLDLAS--PIGRGQRGLIVAPPKAGKTMLLQN
    ::KP::R::I :: :: :: ::R:::E:L:: ::::S PIGRGQR:LI:: ::GKT:::
ATP LAKPIDGRGEISASEFRLIESAAPGIISR-RSVYEPLQTGLIAIDSMIPIGRGQRELIIGDROTGTAVATD
ase  120^      130^      140^      150^      160^      170^      180^
          200v      210v
rho IAQSIAYNHPCVLMVLLID
    : : : : CV : :
ATP TILNQOQONVICVYVAIGQK
ase  190^      200^

```

Fig. 1 NUCSCAN OUTPUT (partial) Insertion element is5 from *E. coli*, complete sequence aligned with *h. haemolyticus* hhaili modification (m) and restriction (r) genes

presented. In the first we scanned the GenBank for a match to Insertion Sequence IS5 (8,9). Surprisingly, this turned up a very strong match to the *h. haemolyticus* restriction and modification enzyme sequence (10). The match is in a non-coding region and may indicate transgenic exchange of insertion elements between distantly related species, or this may be an artifact since the *h. haemolyticus* sequence was determined from an *E. coli* plasmid.

The second example involves a scan of the protein database vs the *E. coli* transcription termination factor *Rho* (12). In this case the result showed strong homology with several ATPases including the tobacco chloroplast sequence presented (11). This is quite interesting since the rho factor also has an ATPase activity. The area of match may show where this activity maps on the peptide chain. In this output the ":" symbol denotes PAM related amino acids.

IMPLEMENTATION

The DNASTAR Genetic Database System is designed to run on IBM PC, XT, AT, or compatible microcomputers, provided sufficient harddisk storage is available to hold the database. A cartridge tape unit is required for acquisition and backup of the databases

from DNASTAR. At least 256 Kilobytes of random access memory are needed.

The programs are written in the C programming language. Since microcomputers are limited both by disk access time and by memory size we have implemented a design that conserves both resources. Disk files are accessed nonsequentially so that only the needed fields of needed sequence entries are ever read into memory. Buffered memory is used for all scanning requirements so that no practical limit is imposed by the program on the number or lengths of sequences or annotation fields of the database. The maximum query sequence length in the present implementation is 4096 residues. Another strategy for optimization of database access is the judicious selection by the program of what field to search first in order to minimize the number of searches to be performed. Similarly the program searches matrix terms and matrix elements in an optimal order. Any combination of fields can be searched in a single pass through the data. Database fields can be added or altered by the collecting resource without necessitating a change in the program.

PERFORMANCE

To give an idea of the processing speeds of these programs, a GENEMAN keyword search for all the papers of a single author on the July 1985 GenBank database (5102 sequences) took 14 minutes. GENEMAN matrix search for donor splice sites took 42 minutes. A search of the entire GenBank by NUCSCAN for homologies to a 654 base sequence took 61 minutes. A search by PROSCAN for homologies to a 400 amino acid sequence on the July 1985 PIR database (3384 sequences) took 37 minutes. These would be proportionally reduced had smaller indexes been accessed. All times were determined on the IBM personal computer model AT with a Tallgrass Technologies model 6135 35 Mb hard disk with 60 Mb tape cartridge.

ACKNOWLEDGEMENTS

We thank Allison Koenke who in 1983 wrote the first version of Geneman for a 64 K microcomputer, proving that it could be done. We also thank David Lipman, Wm.R Pearson and W. J. Wilbur for providing source code of their programs (References 1 and 2).

Walter Goad of the Los Alamos National Labs provided much help in getting the GenBank data onto our personal computers. Terry Platt suggested that we search the PIR vs Rho factor. This work was supported in part by a grant from the NIH SBIR #GM33107.

REFERENCES

1. Burks, Christian, Fickett, James W., Goad, Walter B., Kannhesa, Minoru, Lewitter, Frances T., Rindone, Wayne R., and Bilofsky, Howard S. This issue
2. George, D. G., Barker, W. C., and L. T. Hunt, Protein Identification Resource (PIR), this issue
3. Wilbur, W. J. and Lipman, D. (1983) Proc. Natl. Acad. Sci., USA 80:726
4. Lipman, David J. and Pearson, William. R. (1985) Science 227:4693
5. Dayhoff, M. Atlas of Protein Sequence and Structure (1978) Vol. 5, Supl 3.
6. Hawley, Diane K., and McClure, William R. (1983) Nuc. Acids Res. 11:2237-2255.
7. Levene, S. and Crothers, D. M. (1983) J. Biomol. Str. and Dyn. 1:429-435.
8. Engler, J. A., and Van Bree, M. P. (1981) Gene 14:155-163.
9. Schoner, B., and Kahn, M. (1981) Gene 14:165-174.
10. Schoner, B., Kelly, S., and Smith, H. O. (1983) Gene 24:227-236.
11. Deno, H., Shinozaki, K., and Sugiura, M. (1983) Nuc. Acids Res. 11: 2185-2191.
12. Pinkham, J. L., and Platt, T. (1983) Nuc. Acids. Res. 11:3531-3545.