Sequence Analysis on Microcomputers

GORDON C. CANNON*

In recent years molecular biology has undergone a revolution that can be partially attributed to the advent of facile DNA sequencing methods. The amount of available sequence data is such that the accurate handling and analysis of these data represent a significant effort for the average molecular biologist. Since all but the simplest analyses are impossible to do manually, computer techniques have been devised for the storage, comparison, and searching of sequences. Microcomputer versions of these programs have been developed, many of which exceed the performance of their mainframe counterparts. This review will consider three packages of analysis programs that are commercially available for microcomputers with the Microsoft disk operating system (MS-DOS). These packages differ in their exact implementation of various analytic procedures and in their flexibility. DNASTAR (*1*) is a large set of programs that is essentially a nucleic acid workstation fully capable of accessing the GenBank (*2*) and Protein Information Resource (PIR) (*3*) databases, as well as performing most of the analytical functions usually associated with mainframe computers. The other two analysis packages, MicroGenie (*4*) and IBI-Pustell (*5*), are not as comprehensive as DNASTAR but offer a less expensive way to perform many of the common sequence-handling techniques.

Because the packages represent a wide range of capabilities and collectively contain more than 75 individual programs or functions, I have reviewed those functions that I consider to be most useful to the average molecular biologist and have grouped them into four classes as follows:

1) *Sequence mapping*. These functions list and number a sequence in a convenient, user-defined manner, as well as search the sequence and indicate the location of important landmarks such as restriction endonuclease recognition sites. Mapping includes functions that search for short, variable, or ill-defined subsequences of interest such as polyadenylation sites, promoters, and splice sites, and for physical arrangements such as inverted repeats.

Roche Institute of Molecular Biology, Roche Research Center, Nutley, NJ 07110.

*Present address: Department of Chemistry, University of Southern Mississippi, Southern Station, P.O. Box 5043, Hattiesburg, MS 39406-5043.

2) *Protein analysis*. The investigator may need to translate the six possible protein reading frames of a DNA sequence, to predict actual protein coding regions, and to analyze the putative proteins in terms of chemical and physical characteristics. This class of functions includes programs that predict coding regions, translate DNA or RNA sequences, and analyze such chemical and physical parameters as charge, hydrophilicity, and secondary structure of the resultant protein.

3) *Homology analysis*. These functions compare two sequences in order to identify regions of similarity and present these homologous regions graphically or numerically for evaluation.

4) *Database handling*. Such functions provide access to and easy extraction of information from sequence databases such as GenBank or PIR. Also included are functions that allow the creation of local or personal databases for the organization and analysis of raw sequencing data.

For this review I have examined how each of the program sets performs the functions described in the above four classes. Not all of the features of each package are discussed; this is particularly true for DNASTAR.

Implementation and Discussion

General operation. The specifications of the packages evaluated are listed in Table 1; some of the analytical functions they perform are shown in Table 2. Each package can be thought of as a group of individual programs connected by a user-accessible environment or "shell." The DNASTAR and IBI-Pustell systems use direct menu access, in that an analysis is chosen from a menu and the user is then presented with a submenu of selections for that individual analysis function. MicroGenie has data entry and analysis functions grouped into two modules that are selected from an initial menu. More than one analytical function can be selected for a given sequence, and more than one sequence can be listed for analysis at once. Although both the IBI-Pustell and MicroGenie packages allow use of a floppy disk system, only the hard disk versions will be discussed here. It is strongly recommended that a hard disk system, rather than a floppy disk, be used with these programs.

Before any analysis can be performed, the sequence of interest must be entered into the computer. All of the packages have facilities for direct entry of sequence data from gel digitizers; only the manual entry method was evaluated. MicroGenie and DNASTAR provide simple text editors that are tailored specifically for use with sequence data. Character-by-character editing is allowed in both systems. The IBI-Pustell editor does not allow direct cursor control for modification of existing sequences. Instead, positioning is accomplished by scrolling a number of lines entered from the console. Incorporation of a simple editor capable of direct cursor control would greatly facilitate use of this system. All three programs create unique file formats, so that the use of files created by other systems requires some editing by the user. Both DNASTAR and IBI-Pustell include automated programs for converting some foreign formats to those that are usable by the package. MicroGenie will accept and process files from other sources if they are strictly ASCII compatible and contain no characters prior to the beginning of the sequence.

All three program systems work with commonly available dot matrix printers. No alternative peripherals such as a plotter are supported. Output to letter quality printers is possible for the production of slides or publication quality tables. However, none of the packages provides an output editor that permits the user to conveniently manipulate the data, so that most presentation grade output must be generated by the user on a word processor.

Sequence mapping. In general, to map a sequence with any of the program sets, a file containing the sequence of interest is searched for the recognition sites of restriction endonucleases stored in a separate file of sites. The DNA sequence is then printed across the page, base numbers are marked at convenient intervals, and the name of the restriction enzyme is printed above its cleavage site in the sequence (Fig. 1). In each of the programs the list of enzymes used for the analysis can be modified by the user. All three program sets provide a means of representing restriction enzyme sites in tables as well as producing graphic representations of the fragments formed upon cleavage. To further map the sequence for short regions of interest that may be less well defined than restriction enzyme sites (for example, control regions, splice junctions, or promoters), each of the systems contains a subsequence search function.

DNASTAR permits searches of this type through use of a program called "Matseek." Literal searches are permitted for subse-

quences up to 150 bp in length. Ambiguity in the subsequence is allowed by inclusion of one of the IUB (International Union of Biochemists) standard codes for ambiguous bases. The search can also be set to report only matches of above a given percentage homology. Multiple sequences can be used to search simultaneously in canonical format, that is, to search for subsequence [(a or b) and (c or d)]. Up to ten such primary terms can be used with no limit on the number of secondary terms. The program automatically examines the complementary strand of the searched DNA sequence. Matseek also allows the inclusion of variable spacing between regions of the subsequence. For example, the search string GTCA 4 >>(9, 12)>> TATATA would report all instances in the searched sequence where GTCA is followed by TATATA with a maximum of 12 and a minimum of 9 nucleotides separating them. Matseek also performs matrix searches. The term matrix refers to a site matrix 4 by N (for DNA) or 22 by N (for protein) in which numerical weights are assigned to all of the possible residues at each position in the sequence. The weight is proportional to the probability of the residue appearing at that site in the searched sequence. Output from Matseek is a simple list of positions within the searched sequence that meet the match specification defined by the user.

Fig. 1. Typical output from the sequence mapping functions of (**A**) IBI-Pustell, (**B**) DNASTAR, and (**C**) MicroGenie. Files containing the restriction enzyme recognition sites can be updated by the user as well as limited or grouped to include specific enzymes sets. Note that the DNASTAR output in (B) also includes each possible reading frame in single letter amino acid code.

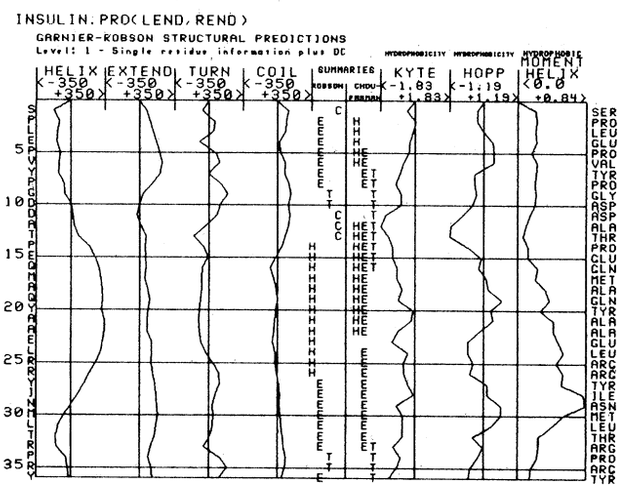
Subsequence searches with MicroGenie are accomplished with the same search function that is used to find restriction enzyme sites. Site files of up to 60 bp are created that contain the site of interest. Ambiguity can be introduced with program-defined ambiguity codes. The IBI-Pustell package performs this type of sequence mapping through an option called "subsequence homology search." Selection of this option permits the user to search a DNA sequence with a short subsequence entered from the console or with a segment of an already existing sequence file. Ambiguity is allowed by matching any lower case letter in the subsequence with any base in the sequence being searched. In addition, the user can specify a minimum percentage match below which any homology will be ignored by the program. Positions within the subsequence that must match exactly are indicated by the user. Output consists of a table of the percentage match and location of the subsequence in the searched sequence.

All three packages include a facility for calculating and displaying the frequencies of base and codon use in a given sequence as well as the dinucleotide distribution. DNASTAR and MicroGenie provide functions specifically aimed at displaying interesting sequence arrangements such as inverted and direct repeats, whereas the documentation for IBI-Pustell describes how a homology program (discussed below) can be used to locate repeats.

Protein coding regions and analysis. Prediction of protein coding regions within nucleic acid sequences and their subsequent characterization have proven useful for identify-

Fig. 2. Output from the protein analysis program "Protein" of DNASTAR. (**A**) Graphic representation of calculated structural predictions for the sequence of insulin. The Garnier-Robson predictions (the four leftmost graphs) can be predicted on three levels: single-residue information, single-residue information plus a weighting factor derived from amino acid composition data, and sum-of-neighboring-residue information. The center two columns contain summaries of structural predictions and denote in which protein conformation (H, helix; E, extended; T, turn; and C, coil) each amino acid residue is likely to be found. Two hydropathy plots are produced according to the calculation method of Kyte and Doolittle (*9*) as well as Hopp and Woods (*10*). The rightmost plot is the hydrophobic moment for either the α-helix (example shown) or β-pleated sheet. The user can modify the *p*H values used in the calculations. (**B**) A summary of the data is presented as well as the amino acid distribution, molecular weight, and isoelectric point.

A



B

```
                    SUMMARY DATA ON B:\INSULIN.PRO

RESIDUES ANALYZED:      36                  WINDOW SIZE:  HOPP     7
                                                          KYTE     9
FIRST PASS MADE:  YES                       pH:   7.000

ROBSON                                      CHOU-FASMAN
CONFORMATION:   HELIX       36%             CONFORMATION:  HELIX       44%
                EXTENDED    42%                            EXTENDED    69%
                TURN        11%                            TURN        39%
                COIL        11%

Robson Decision Constants:
HELIX  EXTEND  TURN  COIL
 -75     -88     0     0

ANTIGENIC SITES:  COORDINATE  HOPP INDEX
                      13          1.19
                      12          1.16
                      11          0.73
                      14          0.57
                      23          0.56

MOLECULAR WT:  TOTAL     4212 g/mol     AVE HYDROPHOBICITY:  HOPP (-100X)   -13
               AVERAGE    117 g/mol                          KYTE (100X)    -85

ISOELECTRIC PT:  pH= 4.83

AMINO ACID DISTRIBUTION:

RESIDUE  NUMBER PERCENT      RESIDUE  NUMBER PERCENT
Ala (A)      4    11%        Phe (F)      0     0%
Arg (R)      4    11%        Pro (P)      5    14%
Asn (N)      1     3%        Ser (S)      1     3%
Asp (D)      2     6%        Thr (T)      2     6%
Cys (C)      0     0%        Trp (W)      0     0%
Glu (E)      3     8%        Tyr (Y)      4    11%
Gln (Q)      2     6%        Val (V)      1     3%
Gly (G)      1     3%        Asx (B)      0     0%
His (H)      0     0%        Glx (Z)      0     0%
Ile (I)      1     3%        Ter (.)      0     0%
Leu (L)      3     8%        Xxx (X)      0     0%
Lys (K)      0     0%        ??? (?)      0     0%
Met (M)      2     6%
```

Table 1. Specifications of the software packages reviewed. All programs were evaluated on an IBM PC AT that was equipped with 640 kilobyte (kybte) of random access memory (RAM), an internal 20-megabyte (Mbyte) hard drive, an external 35-Mbyte hard drive–60-Mbyte tape backup unit (Tall Grass Technologies, model TG-6135), and an Epson FX-85 printer. All of the packages have on-line help.

	IBI-Pustell	MicroGenie	DNASTAR
Version or release date	Ver. 1.2	Ver. 4.0	Dec. 86
Language*	Fortran or C	Pascal	C
Host computer	IBM PC-XT or AT†‡	IBM PC-XT or AT†	IBM PC-XT or AT†
Hardware§ required	20-Mbyte hard disk; printer	20-Mybte hard disk; printer	30-Mbyte hard disk with tape backup or CD ROM reader; printer
RAM required (kbyte)	256	640	512
Copy protection	Key disk	Hardware (requires one-half slot)	Key disk
Database or program updates per year	Databases not supplied	2	4
Cost	$800	$1995	$1250–$6000

*Programs are compiled; source code is not available to the customer. †Also some compatibles. ‡Apple II version is available but does not contain all capabilities. §Floppy disk versions are available but are not recommended. ‖Price depends on programs selected in package. Total cost of all programs reviewed was $5,900; the complete DNASTAR package costs $12,000.

ing cryptic sequences as well as for determining the potential antigenic sites of a predicted peptide. Short peptides representing such sites can be synthesized and used for the production of antibodies against the genuine protein. Molecular weight, hydropathy, or charge data are useful to the investigator attempting to purify a protein that is identified only by its coding sequence. The three packages can translate a DNA sequence into all its possible reading frames. Since only the analysis of proteins from true coding regions yields useful information, one must decide which frame and starting point to use for the "in scripto" translation.

In the DNASTAR package the same program used to map the sequence (Mapseq) prints the six translational reading frames alongside their respective DNA strands. By following the length of the peptide produced before the appearance of a stop codon, a guess can be made as to which regions of the sequence are actually encoding protein. Three additional programs, "ORF," "Geneplot," and "Findcode," aid in locating genuine protein coding regions. ORF locates regions of the sequence that begin with initiator codons and that are followed by any uninterrupted string of coding triplets of a length defined by the user. Findcode searches a sequence for regions of potential coding capacity based on Hunkapillar's modification of Fickett's algorithm (*6*). The program selects potential coding regions based on statistical observations of triplet order in a large number of genuine coding sequences. Geneplot locates potential protein-coding regions based on species-specific codon usage tables. A simple method for producing codon usage tables from Genbank is described, and a codon usage table for *Escherichia coli* is provided. IBI-Pustell also contains a function designed to find protein coding areas based on codon usage within the sequence with a codon usage table. However, unlike DNASTAR, no simple means of producing codon usage tables is provided, and only a drosophila table is included with the package. Without an easy means for producing usage tables for other organisms, this function is of little use.

Once a likely coding region of the sequence is chosen for translation, all three analysis systems contain programs to create protein files from DNA or RNA files, which can then be analyzed directly. The DNASTAR system performs the bulk of this task through a program called "Protein." Output from this program includes a linear map of the sequence printed vertically down the left side of the page (Fig. 2). Robson structural predictions (*7*) of the likelihood of the residues being in helix or coil conformation, in

addition to Chou-Fasman (*8*) predictions, are presented graphically in the five columns to the immediate right of the sequence. Predictions of hydropathy based on the chemical nature of the residue and its neighbors are graphically displayed with the Kyte-Doolittle method (*9*), as are the alternative means of calculation described by Hopp and Woods (*10*).

The analysis module of MicroGenie contains two procedures (numbers 14 and 15) that accept either a nucleic acid or protein file as input and display a structural plot of the protein as output. The first plots hydropathy according to the Hopp and Woods algorithm (*10*), whereas the second procedure lists the sequence and under each amino acid residue prints a letter indicating in which secondary conformation that residue is likely to be found with the prediction formula of Garnier *et al.* (*11*). The information provided by the two procedures taken as a whole then allows the use to make estimates of good antigenic sites within the putative peptide. The IBI-Pustell system produces a Kyte-Doolittle plot of an input protein file and also prints the sequence with cyanogen bromide and trypsin cleavage sites marked. The program also predicts peptide fragment sizes produced by such cleavages. When data obtained from protein sequences predicted from DNA or RNA sequences are used, it must be remembered that these are only predictions. Post-transcriptional or post-translational modifications that may occur cannot yet be evaluated by the programs.

Homology searches. Once a sequence has been characterized, it can be compared with

Table 2. Programs, procedure numbers, or functions with descriptions for the IBI-Pustell, MicroGenie, and DNASTAR software. The list is not exhaustive, especially for DNASTAR.

IBI-Pustell		MicroGenie		DNASTAR	
Program or function	Description	Procedure number or function	Description	Program or function	Description
		Sequence mapping			
Restriction site	Prints sequence with restriction sites	1, 2	Prints sequence with restriction sites	Mapseq	Displays sequence with sites, translation, and annotation
Subsequence search	Locates user-defined subsequence	6, 7	Produces table or graph of fragments	Sitelist, Sitelook	Displays restriction map graphically or in lists
Calculate base composition	Residue frequencies	8	Locates user-defined subsequence	Mapper	Best-fit analysis of restriction fragment data
		2, 5	Residue frequencies	Basedis	Nucleotide frequencies
		9, 10, 11, 14, 15	Searches for repeated sequences	Matseek	Searches for a site described by a matrix
				Sizegel	Measures fragments
				Loop, Bend, Model	DNA structure analysis
		Protein analysis			
Seq translation	Translates protein from DNA	15, 16	Translates all reading frames	Protein	Plots hydropathy, charge, and helix-turn potential
Plot AA composition	Plots either amino acid composition or hydropathy	12	Predicts protein structure by Garnier method	Revtrans	Reverse-translates protein to DNA
Coding-region locator	Finds coding regions based on codon usage tables	13	Predicts protein hydrophobicity by Kyte-Doolittle	Titrate	Titration curve of protein
				Probe	Designs DNA probes from protein sequence
Reverse translation	Reverse translates protein to DNA			Trans	Protein sequence, codon usage, molecular weight, and charge
Peptide analyzer	Amino acid composition, molecular weight; displays CNBr and trypsin sites			Findpro	Finds region of known protein
				Findcode	Coding region by Fickett method (*6*)
				ORF	Finds open reading frames
				Geneplot	Finds coding regions based on codon usage tables.
		Homology			
Forward matrix-reverse matrix	"Dot plot" homology searches	Homology comparison	Compares sequences for areas of homology	Compare, Aaccomp	Compares two sequences by the sliding-window method
Automatic alignment	Lists homologous regions in aligned format	Matrix comparison	"Dotplot" homology comparison	Align, Aalign	Align sequences by the Wilbur-Lipman method
		Alignment of two DNA sequences	Alignment by Korn and Queen method	Gap	Aligns sequences interactively
		Protein alignment	Dayhoff method	Seqcmp	Aligns sequences by method of Martinez, Needleman, and Wunsch
				Nucscan, Proscan	Scans public databases for homology
		Database management			
Cyborg	Reading and retrieval of GenBank entries	List	Retrieval of PIR and GenBank entries	Geneman	Searches PIR or GenBank databases
		Merge sequences	Shotgun sequencing	Seqman	Personal database for raw sequencing information; shotgun sequencing

other known sequences to find areas of similarity. Domains within the sequence can thus be equated with known functional regions of other molecules, and evolutionary relations between sequences can be studied. Homology searches through the large public databases have also often been used to identify cryptic nucleic acid sequences. The problem of determining similarity between two large sequences (greater than 200 bp in length) differs in two ways from searching large sequences for small, well-defined subsequences. First, the computational task is not trivial. To directly compare a sequence of 200 amino acids, for example, with a protein database of 500,000 residues requires 10^8 comparisons (*12*). This problem is partially solved with search algorithms such as that of Lipman and Pearson (*12*) that obviate the need for direct comparison. Second, and more difficult to overcome, is the fact that "homology" as a parameter is a subjective term. In some cases 50 percent homology over a segment of 1000 bp is more significant than 75 percent over 25 bp. What is or is not homology is best defined by the individual investigator within the context of his or her study. A program for comparing two sequences needs to be flexible to be of general use.

DNASTAR contains a group of programs, each of which are designed for a particular approach to sequence comparisons. The program called "Seqcmp" compares and aligns two nucleic acids for similarity with an algorithm described by Martinez (*13*). The program first locates regions of perfect match and then optimizes the region between homologous segments with the Needleman-Wunsch method (*14*). Two large sequences can be compared in a reasonable length of time (comparison of two 8000-bp sequences requires less than 1 hour). The output consists of the two compared sequences displayed in the calculated best-fit alignment. Bases that are matched have the consensus bases printed between them, and gaps introduced by the alignment process are represented by dashes. For comparisons in which the user is less interested in local areas of exact homology but would rather examine two sequences for long regions of imperfect homology, DNASTAR contains a program named "Compare." This program uses a sliding window method of comparison in which the two sequences are "slid" past one another one base at a time, and after each shift the degree of homology is evaluated. The user is prompted to select the window size for the comparison as well as to set percentage and length thresholds for acceptable matches. Output consists of a list of all homologies, which are then sorted in order of decreasing similarity and written to a file for subsequent use with another program called "gap."

"Gap" accepts output from Compare and introduces gaps in either sequence, thereby maximizing the overall alignment of the two sequences. Gap is extremely user interactive and affords the operator a chance to direct the optimization process to areas of interest within the sequence. The output from Gap can be presented as a table of matches, a matrix plot, or a plot of cumulative similarity index (X) versus register shift. These two programs in combination offer the user a flexible means of studying homology between two sequences. However, this flexibility requires close attention by the operator for effective use. The alignment algorithm of Wilber and Lipman (*15*) is implemented by DNASTAR with the program "Align" (for nucleic acids) and "Aalign" (for proteins). In these two programs sequences are compared and the alignment with the highest similarity score is printed out. The score is defined as the sum of short areas of perfect match (k-tuples) minus the sum of gap penalties. The gap penalty can be adjusted by the user, and by varying these and the length of the k-tuple the user can control the "stringency" of the comparison. The output is the calculated best alignment of the two sequences.

MicroGenie's search functions are invoked from the analysis portion of the program and include "alignment of two sequences," "homology comparison," "matrix comparison," and "protein alignment by similarity." "Homology comparison" scans the two sequences for short areas of high homology and then optimizes the matches using the algorithm described by Korn and Queen (*16*). The output consists of a list of matched regions and the ratio of correct matches to total length. This program can be used for long sequences (sum of the lengths of the sequences must be less than 60,000 bp) but the documentation warns that long sequences require a long time and indiscriminant selection of threshold parameters can result in large amounts of output. An interesting and useful feature of this program is its ability to compare protein files with nucleic acid files directly. For long sequences it is recommended by the authors of MicroGenie that a matrix comparison be used. "Matrix comparison" writes one sequence vertically along the y-axis of the display and the other along the x-axis. A mark is placed at the junction of a column and row with the same base, thus creating a diagonal line of marks in areas where there is homology between the two sequences. This simple form of comparison is somewhat limited because of the high background usually caused by random matches throughout the sequences. This problem is overcome by only placing marks at a match if that match is part of a segment containing other matches. The degree of matching and length of the segment required to produce marks are determined by the user with the proper parameter values. The other significant limitation of matrix or dot plots is that most homology searches will involve sequences longer than the display width of the computer, but provisions are made for compressing the sequence to fit on the screen. The compressed mode is used to preliminarily establish areas of probable homology (Fig. 3). Subsequent searches are restricted to these areas for a more detailed look at the similar regions. "Alignment of two sequences" outputs the optimal alignment of two sequences in much the same manner as DNASTAR's program Align. Two sequences are compared for imperfect homology by introducing gaps in either sequence to produce optimal alignment. The alignment algorithm maximizes the similarity index S. (S equals the number of matches minus the number of gaps minus the number of residues in the gap.) Sequences as long as 30,000 bp or 4,000 amino acids can be used. "Protein alignment by similarity" is a program module based on the Dayhoff algorithm (*17*). Two proteins are aligned optimally with gaps in either peptide to maximize homology. Conservative amino acid changes are counted as partial matches according to the rules as outlined by Dayhoff.

The IBI-Pustell package addresses homology analysis by including three programs, all of which were designed around the concept of matrix comparison of two sequences similar to MicroGenie's "matrix comparison." However, a user-defined window of comparison is evaluated through the matrix over the length of the two sequences. Background caused by random matching is filtered with rather long windows (20 to 30 residues). The smoothing effect usually associated with such an approach is lessened by numerically weighting the central residues of matching clusters to allow more distant residues to contribute to the scoring. The program speed is accelerated by hashing—a technique common to most homology algorithms that reduces the number of initial comparisons required. The output is hindered in that neither the display nor printer allow very long sequences to be used in the horizontal position. As with MicroGenie, this is overcome by compression of the sequence so that a preliminary evaluation is required followed by a closer examination of areas demonstrating promise. Forward and reverse homology searches can be performed, and data obtained from a matrix

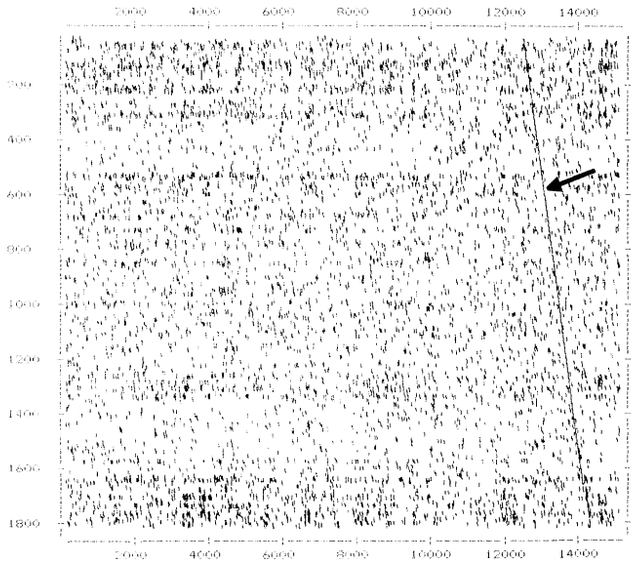


Fig. 3. A dot plot homology comparison generated by MicroGenie. The DNA sequence of the Rubisco large subunit from maize (retrieved from GenBank) is compared to a 15,000-bp region of tobacco chloroplast genome (*x*-axis). The diagonal line (indicated by the arrow) signifies a high degree of homology between the maize gene and a segment of the tobacco chloroplast DNA (approximately 12,750 to 14,250 bp).

search can be displayed in a tabular format in which each matching region is printed under its matching area in the other sequence.

Database management. The GenBank and PIR databases contain systematically collected and catalogued sequence data including pertinent annotations. All three packages contain facilities for accessing the GenBank. In addition, DNASTAR and MicroGenie provide versions of the GenBank and of the PIR sequence bank compatible with their programs. The IBI-Pustell package can use the floppy disk version of the GenBank that is provided by BBN, the distributors of the database.

Most of the data retrieval functions in the DNASTAR package are executed by the program "Geneman." Short sequences (150 bases and less) or matrix definition files can be entered in canonical format for searches throughout either GenBank or PIR. In addition, keywords, author's names, or any literal character phrase that might appear in the annotation file associated with a sequence can be used as search criteria, as can any pattern definition as used by Matseek. For example, a search through GenBank specifying the search terms [(chloroplast or plastid) and (nicotiana)] resulted in the retrieval of six sequence entries in which the word chloroplast or plastid occurred with the word nicotiana in the annotation. The mode of output can be selected by the user and includes options of printing all or part of an entry or directing the output to a file on any available disk drive. Retrieved entries can also be used to create a user-defined subdatabase that can then be accessed just as if it were a conventional GenBank division. Global database homology searches with large sequences (<8000 bp) are performed by the programs "Nucscan" and "Proscan." Either of these programs compares sequences in all or any number of database divisions to a query sequence and scores for the best fit alignment with a variation of the Lipman-Pearson search algorithm (*12*). The sequences in the bank are listed by decreasing similarity score (the difference between the number of matched bases and the number of unmatched bases divided by the gap tolerance) and the matches are displayed with their name, database location, and similarity index. The user can then peruse the list and display any of the listed sequences as an alignment with the query sequence. The time required to complete a homology search depends upon the length of the query sequence and fraction of the database surveyed, as well as how search parameters are defined. Sensitivity of the search can be balanced against the time required by modifying the window size and gap tolerance. A quicker version of the nucleic acid program is provided (Snucscan) that limits query sequences to less than 1000 bases. A typical search for a 541-bp sequence against the entire GenBank with default search parameters took 54 minutes.

The MicroGenie database search facilities treat both GenBank and PIR as if they were individual user storage sections assigned a password or group name. The group name is the same as the divisional name in the database, such as "organelle" or "bacteria". To reduce the storage space required, the annotations are removed from each sequence entry; instead, a descriptive title of 60 characters or less is assigned to each sequence. Each entry also contains a comment line with its database name and the literature citation reporting the sequence. To locate a sequence in the database the "list" function is used for the division of the base that contains the entry. Standard MS-DOS wild cards can be used to list all of the entities within a division that contain the search word or phrase in its description. For example, the command **List organelle: chloroplast** produces the names of 86 entries that contain sequences from chloroplast DNA. Due to the lack of annotation, searching for an entry by author or literature reference is impossible, and by keywords is difficult. The "make search" function is used to conduct a global homology search. The current limit for a query sequence is 2200 residues. The user is allowed to alter the minimum length of perfect homology to qualify as a match; however, the stringency is not under the control of the user and in general the program will find and report matches of 75 percent or greater homology with nucleic acid searches and 40 percent or greater for proteins. With the same 541-bp query sequence reported above, a complete search of the GenBank with MicroGenie required 66 minutes. Output consists of the name of the sequence found in the database to match the query sequence.

The IBI-Pustell package includes a program for accessing the GenBank database. The program Cyborg creates an environment separate from the normal analysis program menu that permits the user to browse through the GenBank entries and extract in a number of formats any information contained there, including full annotations and comments. Finding a particular sequence depends on having a prior idea of its name; the search then proceeds in a fashion analogous to looking through a large file cabinet for a folder. Apparently, the Cyborg environment will eventually contain all the analysis functions of the present IBI-Pustell system (*18*).

In addition to public databases, the user may want to create a personal database, especially users of the shotgun sequencing method of Sanger (*19*). With this method an investigator produces random sequence segments from a master sequence. As these random bits are sequenced, each must be compared with all the others to determine overlaps. This process is continued until a consensus sequence representing the master sequence is completed. DNASTAR contains a sequencing database manager called "Seqman" that allows the user to enter individual sequence segments (typically the results of one electrophoresis gel). As the sequence is entered the program compares it to all others in the database or project. If homology or overlap is detected, the program adds that sequence to the homologous segment and creates a "contig" and the members of the contig are aligned to create a consensus

sequence. As more and more entries are made, contigs are bridged and joined until the resulting consensus sequence represents the master sequence. Data produced from Maxam-Gilbert sequencing, in which relative locations of individual segments are most often known, are also accommodated by Seqman and can be properly placed before any bridging sequences are known. Seqman produces output that allows the user to examine the project at any stage of completion. Furthermore, the individual entries are always maintained just as they were added and thus can be examined firsthand to resolve any conflicts. Graphic representations of the segments and their orientation respective to the contig can be printed as well.

MicroGenie also provides a facility for sorting and arranging shotgun sequence segments. As each segment is produced it is entered into a storage area that is provided with an individual name or password by the user. When a sufficient number of segments has been added, the "merge" function is then invoked. Merge compares every segment in the storage area and produces contigs where overlaps occur. This process continues until all the segments have been assigned a position or until too many conflicts exist within the contigs. Output consists of two types: a consensus contig that contains all of the overlapping segments entered up to that point, and a listing of the contigs with each segment (exactly as it was entered) aligned upon the contig according to the overlap. Both the DNASTAR and MicroGenie programs allow the user to modify the comparison criteria.

Summary

Overall, each of the program packages performed their tasks satisfactorily. For analyses where there was a well-defined answer, such as a search for a restriction site, there were few significant differences between the program sets. However, for tasks in which a degree of flexibility is desirable, such as homology or similarity determinations and database searches, DNASTAR consistently afforded the user more options in conducting the required analysis than did the other two packages. However, for laboratories where sequence analysis is not a major effort and the expense of a full sequence analysis workstation cannot be justified, MicroGenie and IBI-Pustell offer a satisfactory alternative. MicroGenie is a polished program system. Many may find that its user interface is more "user friendly" than the standard menu-driven interfaces. Its system of filing sequences under individual passwords facilitates use by more than one person. MicroGenie uses a hardware device for software protection that occupies a card slot in the computer on which it is used. Although I am sympathetic to the problem of software piracy, I feel that a less drastic solution is in order for a program likely to be sharing limited computer space with other software packages. The IBI-Pustell package performs the required analysis functions as accurately and quickly as MicroGenie but it lacks the clearness and ease of use. The menu system seems disjointed, and new or infrequent users often find themselves at apparent "dead-end menus" where the only clear alternative is to restart the entire program package. It is suggested from published accounts (*18*) that the user interface is going to be upgraded and perhaps when that version is available, use of the system will be improved. The documentation accompanying each package was relatively clear as to how to run the programs, but all three packages assumed that the user was familiar with the computational techniques employed. MicroGenie and IBI-Pustell further complicated their documentation by mixing instructions for the version based on floppy disk operation with that for the hard disk version. Considering the currently low prices for hard disks, elimination of floppy-disk-based versions would seem to be the most reasonable solution for this problem. Each of the packages had some degree of on-line help that in most cases consisted of abbreviated pertinent sections of the user manual.

All of the program packages reviewed are constantly being revised (*20*). A number of newer program packages are available, including some for the Apple-MacIntosh computers, that were not evaluated here but show promise as valuable tools for sequence analysis. As microcomputers progress in power and sophistication, no doubt so will sequence analysis programs be available for them.

REFERENCES AND NOTES

1. DNASTAR, Inc., 1801 University Avenue, Madison, WI 53705.
2. C. Burke *et al.*, *Comput. Appl. Biosci. (CABIOS)* **1**, 225 (1985).
3. D. G. George *et al.*, *Nucleic Acids Res.* **14**, 11 (1986).
4. MicroGenie, Beckman Instruments, 1050 Page Mill Road, Palo Alto, CA 94304.
5. IBI Pustell, International Biotechnologies, 25 Science Park, New Haven, CT 06511.
6. J. W. Fickett, *Nucleic Acids Res.* **10**, 5303 (1982).
7. E. Suzuki and B. Robson, *J. Mol. Biol.* **107**, 357 (1976); B. Robson and E. Suzuki, *ibid.*, p. 327; B. Robson, *Biochem. J.* **141**, 853 (1974).
8. P. Y. Chou and G. D. Fassman, *Adv. Enzymol.* **47**, 45 (1978).
9. J. Kyte and R. R. Doolittle, *J. Mol. Biol.* **157**, 105 (1982).
10. T. P. Hopp and K. R. Woods, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 3824 (1981).
11. J. Garnier *et al.*, *J. Mol. Biol.* **120**, 97 (1978).
12. D. J. Lipman and W. R. Pearson, *Science* **227**, 1435 (1985).
13. H. M. Martinez, *Nucleic Acid Res.* **11**, 4629 (1983).
14. S. Needleman and C. Wunsch, *J. Mol. Biol.* **48**, 443 (1970).
15. W. J. Wilbur and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 726 (1983).
16. L. J. Korn *et al.*, *ibid.* **74**, 4401 (1977).
17. M. O. Dayhoff, in *Atlas of Protein Sequence and Structure*, M. Dayhoff, Ed. (National Biomedical Research Foundation, Silver Spring, MD, 1979), vol. 7, suppl. 3, p. 1.
18. J. Pustell and K. C. Kafatos, *Nucleic Acids Res.* **14**, 479 (1986).
19. F. Sanger *et al.*, *J. Mol. Biol.* **143**, 161 (1980).
20. The programs were reviewed during January and February of 1987. Since that time DNASTAR and IBI-Pustell have released updated versions that have resulted in improved performance in some of the functions discussed.