Volume 14 Number 1 1986 Nucleic Acids Research

A convenient and adaptable microcomputer environment for DNA and protein sequence manipulation and analysis

J.Pustell[1] and F.C.Kafatos[1,2]

[1]Department of Cellular and Developmental Biology, The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA, and [2]Institute of Molecular Biology and Biotechnology and Department of Biology, University of Crete, Heraklio, Crete, Greece

Received 16 July 1985

ABSTRACT

We describe the further development of a widely used package of DNA and protein sequence analysis programs for microcomputers (1,2,3). The package* now provides a screen oriented user interface, and an enhanced working environment with powerful formatting, disk access, and memory management tools. The new GenBank floppy disk database is supported transparently to the user and a similar version of the NBRF protein database is provided. The programs can use sequence file annotation to automatically annotate printouts and translate or extract specified regions from sequences by name. The sequence comparison programs can now perform a 5000 x 5000 bp analysis in 12 minutes on an IBM PC. A program to locate potential protein coding regions in nucleic acids, a digitizer interface, and other additions are also described.

INTRODUCTION

Since it was created in 1981 for the use of our laboratory, this system has grown from a small, user-friendly package of sequence handling utilities for 8 bit microcomputers (1) to a large, powerful package (2,3) running on a wide variety of micro, mini, and mainframe computers, under an equal variety of operating systems, in both single and multi-user configurations. Distributed at cost as FORTRAN source code, a version of the package was supplied by us to over 150 institutions and individuals world-wide, who in turn disseminated it to many other users. That version is still freely available in the public domain, although we have stopped distributing it ourselves.

An enhanced and expanded version of the package (4) has been developed for commercial distribution through an agreement between one of us (J.P.) and International Biotechnologies, Inc. (IBI). Since that version was first released in August 1984, the demand has been even higher, permitting the production of a major update within six months, and of a radically enhanced version which is described here and is being released late in 1985.

* Available to academic and commercial users for $800. See reference (4) for details.

HARDWARE COMPATIBILITY

Minimum:
- IBM PC, XT, AT, or most Compatibles
- MS-DOS 2.1 or higher
- 256 Kbytes RAM Memory
- Two 360 Kbyte Floppy Disk Drives
- Monochrome Display
- Any Printer

Optional:
- XMI additional memory to 10 megabytes
- High Capacity Floppy Disk Drives
- Hard Disk Drives (including non-IBM)
- Color Display
- Digitizer

DATA COMPATIBILITY

Symbols:
- Full IUPAC-IUB standard for nucleic and amino acids

Disk Formats Accepted:
1) Lines of text
2) Standard GenBank
3) Stanford/Intelligenetics
4) GenBank Floppy Disk Database
5) National Biomedical Research Foundation Protein Database in our format
6) Restriction Enzyme, Subsequence, and Codon Bias data files in our format

Data Utilities:
- Sequence Files:
  - Text, GenBank, and Stanford/Intelligenetics files are identified and properly read with no user intervention
- Databases:
  - The highly compressed floppy disk databases are presented to the user as if they were individual sequence files in a large directory. Any entry may be accessed from the database in seconds and used transparently in any program. Entry subsets and supersets may be made by the user. Use cross-indices as supplied with the GenBank database: by Sequence Name (Locus), Accession number, Author, or over 750 unique Keywords.
- Annotation:
  - GenBank, Stanford/Intelligenetics, GenBank database, and NBRF database files contain explanatory information which can be used to automatically annotate the sequence on display, to translate it appropriately, to extract named regions from a larger sequence, or to format specific regions of the sequence for easy identification.

NUCLEUS

User Interface, Dual Mode:
- Both Modes:
  - Help Facility- Detailed messages windowed to screen on demand.
  - Function Keys- Up to 30 single key stroke functions available simultaneously, values of keys displayed.
  - Clock Facility- All output labelled with date and time.
  - Machine Status- Always displayed on interactive screens including current output device, current free memory, current keyboard status.
- Menu Mode:
  - Screen oriented displays with multiple windows on screen and selections made by moving an arrow.
  - Values of current variables may be displayed as a list.
- Command Mode:
  - Line oriented display for rapid use by experienced user.

Disk Interface:
- Dual Active Directories:
  - Default Path- User's directory with Read and Write permissions. May be password protected.
  - Alternate Path- A database directory with Read-Only permission.
  - Both-
    - Files displayed in scrolling window, categorized and alphabetized.
    - User sets default directory names, changes and displays directories and sub-directories, selects disk.
    - Files selected by pointing arrow at list and manipulated with function keys (Load, Rename, Delete, Browse, etc).
- Databases:
  - Are displayed as directories of sequences with the same function options as disk files.
  - May be selectively displayed by organism or gene type using "wild-cards"
  - Entries accessed as quickly as sequence files by all programs

Memory Interface:
- Programs use whatever user memory is available and display remaining free memory
- There is no limit to sequence length accepted by programs.
- Files loaded in memory are displayed, categorized and alphabetized in a scrolling window, and are selected for manipulation by moving an arrow

Printer Interface:
- Programs support up to four virtual printers (different physical printers and/or various fonts and pitches)

Screen Interface: Monochrome or Color
- User may select screen attributes to be used by displays and by automatic annotation routines: highlight, lowlight, inverse, blink, underlined, foreground color, background color
- User may select the tone to be sounded for each base, when entering sequences

Format Control:
- User may control many format parameters such as line and page length, upper or lower case, double or single stranded, DNA or RNA, numbering system, 3-letter or 1-letter AA code, printer type, automatic annotation, automatic translation, etc.
- Format specifications may be entered in sequence file annotation
- Format specifications may be changed within a sequence file. Eg. translated, untranslated, and flanking regions may be formatted differently in a single display
- Format specifications may be different for different programs
- Tokens for marking designated regions during automatic annotation may be set by user.

Some Other User Options:
- Alternate genetic codes
- Output disk file format (GenBank or Stanford/Intelligenetics)
- Digitizer port
- Audible cues on or off

DATA MANAGEMENT

Editor:
- Enter/Edit DNA or protein sequences and annotation in GenBank format
- Fully screen oriented editor: Changes made at cursor location, may involve bases or blocks from this or other files. Double entering for accuracy, option of audible tone for each base.
- Digitizer Support: Enters reading of sequencing gels into editor directly from digitizer. Handles curved or narrow lanes. Screen shows location on film and sequence entered. Can exactly locate on film last position entered, permitting interruptions. Base specific tones permit monitoring of input. Double entering for accuracy.

Restriction File Editor:
- Permits editing of more than 100 restriction enzyme entries supplied with programs

Create Peptide from Nucleic Acid:
- Peptide sequence file created by translation of nucleic acid file. Automatic notation of nucleic acid regions used.

Subsequence Editor:
- Enters or extracts from files subsequences, for use with the subsequence homology program.

Create Codon Bias File:
- Creates codon bias files from up to 50 sequences for use by protein coding region locator. Features permit rejection of deviant sequences and elimination of insignificant cells in bias table, setting baseline at the value calculated for a random sequence, determining correlation coefficients for all sequences, and calculating strand adjustment.

Database Manager:
- Merges databases, creates subsets and supersets, adds/deletes entries.

GENERAL ANALYSIS

Translation:
- Translates nucleic acid sequence into a peptide in any or all frames using standard or alternate genetic codes. Can translate just exons, as directed by annotation. Flexible output formats, 3- or 1-letter AA codes.
- Calculates codon usage table, peptide molecular weight, peptide pI, if desired

Reverse Translation:
- Predicts possible nucleic acid sequence of a peptide and adjusts for codon bias. Locates minimally redundant segments, presents as oligonucleotide probes with multiply substituted positions, and calculates probe Tm.

Restriction Analysis:
- Lists all cut sites in alphabetical order
- Prints sequence with all restriction sites shown
- Prints restriction map with single line for each enzyme
- Calculates, lists, and identifies ends of fragments produced by multi-enzyme digests.
- All restriction site functions can use various subsets of enzymes and analyze circular and linear molecules correctly, complete or in part.

Measure Fragment Sizes:
- Calculates gel fragment sizes (nucleic acid or protein) from digitizer, using spline and least squares methods.

Calculate Base Composition:
- Calculates strand assymetry, base composition, and di- and tri-nucleotide frequencies with Chi-square analyses of each cell for all frames and measures purine/pyrimidine ratio of first and third bases to predict coding regions

Plot Base Composition:
- Plots, on printer, base composition as ratio or percent for any combination of bases. Can superimpose plot on protein coding region locator plot.

Calculate Amino Acid Composition:
- Measures amino acid composition of a peptide, calculates molecular weight and estimates pI, prints a map of trypsin and CNBr cleavage sites on sequence of peptide, and lists corresponding fragments with molecular weight and pI.

Plot Amino Acid Composition:
- Plots, on printer, AA composition and protein hydropathy, from peptide files or by translating nucleic acid files. Plots illustrate protein features and suggest possibly antigenic domains.

Figure 1 - Summary of Program Package (continued on facing page)

SIMILARITY SEARCHES

Subsequence Match:
Rapid search for match with a short subsequence without insertions or deletions, with percent match set by user, and bases which may be defined by user.

Forward and Reverse Matrices:
Enhanced, high speed, "dot matrix" comparisons that locate similarities, tandem repeats, inverted repeats, potential stem loops, and palindromes
Multiple levels of noise filtration and very high speeds (5000x5000 bases on IBM PC in 12 minutes)
Accept very large sequences (up to 32,000 bases on one axis, no length limit on the other), protein or nucleic acid sequences, linear or circular. Display by printer on a graph paper-like background, using letters to indicate degree of match.

Automatic Matched Sequences:
A high-speed alignment program with the same attributes as matrices, used to examine homologies in detail at the base pair level.

Optimal Alignment
For nucleic acids and proteins respectively, the Wilbur-Lipman program (5,6) transposed into C, and the alignment pass of the Lipman-Pearson program (7) with modifications to make them compatible with the rest of the package

Global Search - Nucleic Acids:
Essentially the first pass of the Wilbur-Lipman program (5,6) with enhancements from the matrix programs above for a rapid search of the DNA sequence database.

Global Search - Proteins:
The complete Lipman-Pearson program (7) modified for compatibility with the package. It uses the database format, permitting use of protein sequences without previous extraction. The complete set of NBRF sequences fit on one floppy disk.

Manually Align Sequences:
Permits user to align sequence segments manually and produce consensus sequences by flexible criteria for slides or publication.

Locate Protein Coding Regions:
Locates potential coding regions in DNA using codon bias files created by program described above
Enhances resolution by strand adjustment, plots termination codons next to codon bias plot for all six frames (both strands) using a printer, and calculates correlation coefficients.

The current version is supplied ready to run on the IBM PC line and most compatibles. It has been designed to move into newer machine environments as they become popular, particularly into MS-DOS or UNIX based systems. The user interface is now screen and cursor oriented, large databases are supported, and a digitizer interface is provided. A number of new programs have also been added. We briefly present only some of the new aspects of the package in the text and give examples of their use.

THE NUCLEUS

We have rewritten the complete package from FORTRAN into the C programming language, which is far more powerful in terms of data structures, memory management, and flexible I/O. These characteristics are particularly evident in in the general purpose interface we call the "nucleus" (Figure 1). This nucleus has aspects of both the kernel and the shell of the popular UNIX operating system. It contains essential command and utility programs, monitors the current machine and program status, and also surrounds the user in a convenient environment tailored to the molecular biologist. Like a biological nucleus, it also contains the information to run various tasks outside itself. These tasks are the separate special function programs.

The nucleus can be run in one of two modes. Menu mode provides the most information on the screen, and will be described in more detail below. Command mode is for the rapid execution of specific tasks by the experienced user as explained in the program documentation.

In menu mode choices and information are displayed on the screen and selected by moving the cursor or an arrow to the appropriate item on the display (Figure 2). A few examples are described below. For a more complete list of capabilities refer to Figure 1.

Across the top of the screen are displayed (1) the current output destination (printer, screen, or disk), which may be changed at any time with

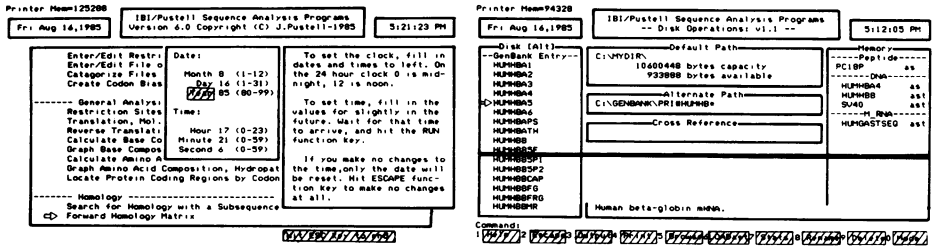

Figure 2 - Nucleus Function Screens

one keystroke, (2) the remaining free memory, (3) the current keyboard status (caps lock, insert on/off), (4) the current date and time, and (5) the current program task title. Across the bottom are the operations currently assigned to the function keys. The programs support up to 30 key assignments at any point. For efficiency, assignments change appropriately for each part of the program, or even for each choice in a list, and a special function help key permits use without straining the user's memory.

Figure 2 (left) shows the main menu from which the user may select programs by pointing with the arrow. In this case, the menu has been overlaid (center) by pressing the CLOCK function key. The date and time can now be set by moving the inverse video bar (currently on year) to the datum to change and typing in a new value. The HELP function key has also been pressed, resulting in a second overlaid box (to the right of the CLOCK data entry box), which explains use of the clock function. The function keys do not appear at the bottom of the screen in this example because the help message is displayed.

The nucleus can be configured in a great many ways in terms of parameters such as operations, file formats, printing formats, printer and screen support, etc. It also contains extensive disk and memory support accessed by the disk operations facility (Figure 2, right). The two upper central boxes in this display show the two paths supported by the program, a read/write working directory (the default path) and a read-only common directory (the alternate path). The lower central box is for use of the cross indices to the databases. Paths may be changed by simply typing in a new path in the appropriate box. The box on the right of the screen shows the files which are currently in memory. Frequently one is doing many analyses on a small group of sequences. The whole group may be loaded into memory from the disk, then manipulated quickly and easily from any program by using this memory window.

The box to the left of the screen is a scrolling window listing the disk

files in one of the two supported paths (in this case, "Alt" on the top line of the box indicates the alternate path is displayed). The programs classify disk files by type, then alphabetize the names within a type. Files are selected with the moving arrow, and single keystrokes permit the user to load the file into memory, rename, delete, get file statistics, or browse through the contents of the selected entry.

An important feature of this package is the convenient use of large databases. Databases are sets of large disk files which are not accessible by the usual operating system utilities. However, these programs present them to user as if they were large directories of normal sequence files. The GenBank primate directory has been used as the example in Figure 2 (right). In the alternate window the detailed path has been specified: disk C (C:), the GenBank directory (\genbank\), and the primate index (pri#). A specific subset of the primate index has been requested by the ambiguous name which appears after the index name. It specifies only human (hum) hemoglobin (hb) sequences, of any kind (*). Only sequence names matching all these criteria are displayed in the disk window. Database support also contains a specialized browse facility, which would normally overlay the right two-thirds of the screen. We show only the bottom half of such a display in Figure 2, reserving the top half for demonstration of the normal disk operations screen. The very convenient database browse facility presents a one line description of each entry in a parallel window to the right of the sequence names. This window scrolls as the sequence names are scrolled, and does not interfere with the usual disk functions such as loading files into memory. One may also use the cross indices provided to select a set of entries by name, by accession number, by author, or, perhaps most valuable, by one of over 750 unique keywords. Any sequence in either database can be accessed by name or by cross index in seconds.

When GenBank solicited help on its floppy disk database release from sequence software suppliers, we elected to donate all the work we had done in that direction for their unrestricted use, since we much prefer to help the existing database efforts than to duplicate them. We are pleased that much of our work was incorporated into their format, and have changed our software to accomodate the ways in which they differ from our original conception. For the protein database we used a variable number of bits per amino acid, with the most common amino acids having the shortest codes. This permits us to compress the complete set of proteins onto a single 360 kbyte IBM floppy disk. Thus one may easily do a global protein search without needing a hard disk.

The annotation is also compressed onto additional disks.

The databases can be used as supplied on floppy disks, but a database manager provided with the package greatly extends their usefulness. With it, one may create subsets or supersets of the database, or add one's own sequences to it.

Large amounts of sequence data are relatively useless without annotation. For some types of information, such as general description or references, it is sufficient to provide functions to display or print it. For locating specific regions of sequence, such as coding regions, promoters, inserted viruses, etc, it is far more useful and accurate for the machine to read the annotation and use it. Our programs separate the sites and features tables, which contain such information, from the rest of the annotation. The tables may be printed out, if desired. They may also be used to translate specified regions, to extract specified regions into another file, to annotate the sequence itself at the specified location much like a restriction map, or some combination. For example, if the user requests "Large T antigen", the programs can locate SV40 among the thousands of sequences in the database, find the location of large T from the annotation, read the two exons into memory, splice them, reverse complement them (large T is on the minus strand of SV40), locate the entries in the features table which pertain to this region, renumber them to fit the new sequence, reorder them in reverse order, and store the new sequence in memory under the name LARGE T in less than 10 seconds. In addition to the information supplied by GenBank, the user may add information to format the printout in different ways for different regions, or to note features of specialized interest.

THE PROGRAMS

Unlike the nucleus, which is an environment within which one works, the programs are specific specialized tasks done by command from the nucleus. They are summarized in Figure 1 and only selected new features will be described here.

Editor

Since we are no longer bound to line oriented displays for transportability, the sequence editor has been endowed with full screen and cursor functions. It features a split screen which enables the user to view both annotation and sequence simultaneously and to move easily from one to the other. During cutting and pasting of sequence fragments, the sites/features table information associated with a particular fragment is moved with it, and

renumbered appropriately. Similarly, numbered annotation entries are renumbered automatically during editing within a sequence. The editor has a double-entry capability in which a sequence may be entered twice for checking accuracy. The second entry is typed over the first as it is entered. When the second does not agree with the first, the user is stopped with an audible tone and corrections can be made immediately.

The editor also has a digitizer interface for direct entry of data from autoradiograms of sequencing gels. It can correctly interpret curved lanes and/or very narrow lanes. It displays both the sequence entered and a map of the gel showing how much of each lane has been entered. Should a data entry session be interrupted, it has a facility to locate the last position entered on the gel before resuming data acquisition. The interface supports many editor features on the digitizer itself such as double-entry, ambiguity codes, insert, delete, etc. The digitizer can also be used to estimate size from mobility on a gel by both a least squares (9) and spline (smooth curve) method.

At this time we are preparing a "shotgun" sequencing program to automatically overlap gel readings during a sequencing project, of the type originally pioneered by Roger Staden (8). It will be available by the time this paper is published and is a free update to users. It is designed to interface with the editor, digitizer routines, and disk file facilities described above.

Matrices and Automatic Matched Sequences

We have increased the speed of the forward and reverse matrices and of the automatic matched sequence programs up to several hundred-fold by adding a hashing algorithm. Such algorithms create look-up tables from the sequences to find short matches directly, rather than by searching for them. They are used extensively in the very fast global database search programs (5,6,7). In our implementation, the programs will look only at places with at least a minimal number of matching positions, thus performing the more extensive and time-consuming analyses at only a subset of all possible positions. Using these methods one can compare all of polyoma to all of SV40 (5000 x 5000 bp) in 12 minutes on an IBM PC and in less than 5 minutes on a 16 bit minicomputer (similar to a PDP-11).

Since there is a certain loss of sensitivity in the hashing process we made the degree of hashing user adjustable. We also introduced a variable we call a jump into the hashing process. Using a jump of three, for example, enables the user to create a hash table in which the bases considered are

always three positions apart. For the coding regions, in two out of three possible registrations a jump of three will produce a hash table containing only conserved first and second bases, and the matching regions will be readily located even if every third base is substituted. Thus the programs are very flexible in terms of trading off speed versus sensitivity and in adjusting to compare coding versus non-coding regions. The same additional variables have been added to the global nucleic acid search which has otherwise been previously described (5,6).

Locate Protein Coding Regions

This program locates regions of DNA whose potential codon bias suggests they may code for proteins, a method first suggested by others (10,11,12). The method involves a heuristic measure which we call the "C-statistic". It involves measuring codon usage over a short interval of sequence and comparing it to that for an appropriate set of known coding regions (eg. a number of genes from the same species). Following (12), we use a ratio of the number of actual occurrences of each codon within a synonymous group versus the number of occurences expected if there were no bias within that group. Potential codon usage is then measured over a region of an unknown sequence, and a value calculated for overall codon usage in that region in each frame, as a measure of the likelihood that it is a bona fide coding region translated in that frame. We have added several important improvements in the construction of the reference codon bias table and in the analysis itself.

A bias table is constructed by a program which accepts a number of sequences and measures their usage of codons within synonymous groups. A bias table is constructed for each individual sequence and an aggregate table is constructed using all sequences together. The program measures the significance of bias in each synonymous group, compares each sequence in the table to every other, and sets a baseline for a random sequence of the same base composition as the real sequences in the table. Thus one may readily evaluate the internal consistency of the table and have a standard of comparison to values for known genes.

Previous programs of this type only analyze the three possible reading frames on a DNA strand. Our programs analyze six frames at once, three on each DNA strand. This involves special difficulties due to the way in which codon usage seems to be biased. As part of a larger study of codon bias, to be published elsewhere (13), we have noted that the pattern of codon bias tends to produce long open reading frames and strong positive bias on both strands of a known coding region. To deal effectively with this problem, the

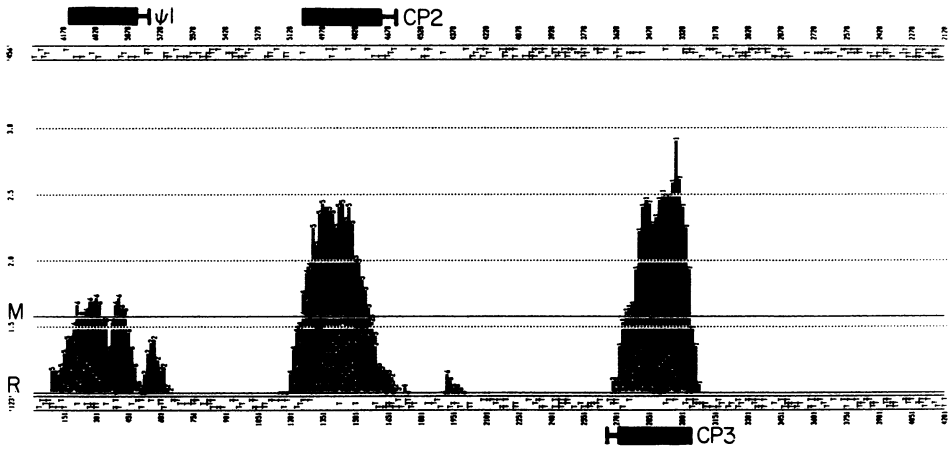

Figure 3 - Codon bias plot of the first 4200 base pairs of the Drosophila cuticle protein locus (DROCTCL2, accession number J01081). The region contains two cuticle protein genes (CP2 on the minus strand in frame 5, and CP3 on the plus strand in frame 1) and a pseudogene (ψ1) on the plus strand. Wide bars denote coding regions, narrow bars denote introns. The first exons are only four amino acids long and are too short to be detected by this method. The lower column shows a "T" in one of three horizontal lanes for each termination codon found in frames 1,2, and 3, respectively, on the plus strand. The upper column shows the termination codons in frames 4,5, and 6 on the minus strand. The numbering on the bottom is 5'-3' for the plus strand, and on the top is 5'-3' on the minus strand, counting from the end of the complete 6314 bp sequence. The black bars extending up from the bottom of the plot show the C-statistic at that point. The numbers from 1-6 within the plot give the values for each frame, with the correct frame at that point appearing at the top of the bar. The baseline of the plot has the C-statistic value of 1.0 for a random sequence, and this is noted by the "R" on the left. Between the C-statistic values of 1.5 and 2.0 is a line running through the plot marked by "M". This the minimum value, the lowest C-statistic value for any of the known sequences which were used to make the particular codon bias table. Empirically, anything above the M line is probably significant, anything below the R line is probably not significant (and is, in fact, not shown on the plot).

program creates a second C-statistic called a "strand adjustment", by making a codon bias table using only those codons which are positively biased on one strand and negatively biased on the opposite strand. This value is calculated independently of the more usual C-statistic described above and then used to adjust the C-statistic up or down. The results of this process substantially increase the discrimination between strands.

The analytical program which uses the codon bias tables generated above can produce a plot of the C-statistic (Figure 3) with or without the strand adjustment. It can also calculate the statistic by either multiplying bias

values over the window and taking the root, or by summing the values over the window and taking the average. We find the product method generally more discriminating, but the average method is helpful when one has a poor bias table (one with too few sequences). Once a potential coding region has been located with the plot, this program can also calculate the correllation coefficient between the codon bias table for the selected region versus the codon bias table previously established for the set of standard genes for a more accurate measure of the significance of the comparison.

ACKNOWLEDGEMENTS

We would like to thank D.George at NBRF and W.Rindone and D.Swindell at GenBank for their willing advice on the databases, W.Pearson for his help with the global search, M.Kreitman for sharing his experience with digitizers, and P.Cherbas for helpful discussions on the codon bias work. The initial basic research on codon bias was done in collaboration with Michael Rosbash and partially supported by grant GM33205 to him. Additional basic reasearch was partially supported by ACS and NIH grants to F.C.Kafatos. Development of the commercial package was funded by IBI and by program sales.

REFERENCES

1. Pustell, J. and Kafatos, F.C. (1982) Nuc. Acids Res. 10:51-59.
2. Pustell, J. and Kafatos, F.C. (1982) Nuc. Acids Res. 10:4765-4782.
3. Pustell, J. and Kafatos, F.C. (1984) Nuc. Acids Res. 12:643-655.
4. The programs are available for IBM PC, XT, AT and compatibles for $800 (commercial or academic) on 360 kbyte floppy disks. Compacted NBRF protein database and support for GenBank nucleic acid database provided. Nucleic acid database available directly from GenBank. Support provided for IBI digitizer (digitizer itself available separately). Price includes telephone support and free updates for one year. Updates after that time are available at low cost to registered users. Contact International Biotechnologies, Inc, P.O.Box 1565, New Haven, Connecticut 06506 or telephone 800-243-2555.
5. Wilbur, W.J. and Lipman, D. (1983) Proc. Natl. Acad. Sci., USA 80:726-730.
6. Wilbur, W.J. and Lipman, D. (1984) S.I.A.M. J. Appl. Math. 44:557-567.
7. Lipman, D.J. and Pearson, W.R. (1985) Science 227:1435-1441.
8. Staden, R. (1982) Nuc. Acids Res. 10:4731-4751.
9. Schaffer, H.E. and Sederoff, R.R. (1981) Anal. Biochem. 115:113-122.
10. Smith, M.M. and Andresson, O.F. (1983) J. Mol. Biol. 169:663-690.
11. Staden, R. (1984) Nuc. Acids Res. 12:551-567.
12. Gribskov, M., Devereux, J., and Burgess, R.R. (1984) Nuc. Acids Res. 12:539-549.
13. Pustell, J. et. al., manuscript in preparation.
14. Snyder, M., Hunkapillar, M., Yuen, D., Silvert, D., Fristrom, J., and Davidson, N. (1982) Cell 29:1027-1040.