Volume 12 Number 1 1984 Nucleic Acids Research

A comprehensive set of sequence analysis programs for the VAX

John Devereux, Paul Haeberli* and Oliver Smithies

Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA

Received 18 August 1983

ABSTRACT

The University of Wisconsin Genetics Computer Group (UWGCG) has been organized to develop computational tools for the analysis and publication of biological sequence data. A group of programs that will interact with each other has been developed for the Digital Equipment Corporation VAX computer using the VMS operating system. The programs available and the conditions for transfer are described.

INTRODUCTION

The rapid advances in the field of molecular genetics and DNA sequencing have made it imperative for many laboratories to use computers to analyze and manage sequence data. UWGCG was founded when it became clear to several faculty members at the University of Wisconsin that the there was no set of sequence analysis programs that could be used together as a coherent system and be modified easily in response to new ideas.

With intramural support a computer group was organized to build a strong foundation of software upon which future programs in molecular genetics could be based. This initial project has been completed and the resulting programs, written in Fortran 77, are available for VAX computers using the VMS operating system. Most of the programs can be used with only a terminal, although several require a Hewlett Packard plotter.

UWGCG software has been installed for testing at eight different institutions. A simple method has been developed for transferring and maintaining this system on other VAX computers.

DESIGN PRINCIPLES

UWGCG program design is based on the "software tools" approach of Kernighan and Plauger(1). Each program performs a simple function and is easy to use. The programs can be used independently in different combinations so that complex problems are solved by the use of several programs in succession. New programming is simplified since less effort is required to bridge a gap between existing programs.

UWGCG software is designed to be maintained and modified at sites other than the University of Wisconsin. The program manual is extensive and the source codes are organized to make modification convenient. Scientists using UWGCG software are encouraged to use existing programs as a framework for developing new ones. Our copyright can be removed from any program modified by more than 25% of our original effort.

PROGRAMS AVAILABLE FROM UWGCG

The programs described below are named and defined individually in Table 1. Program names in the text are underlined.

Comparisons

Comparisons may be done with "dot plots" using the method of Maizel and Lenk(2). Optimal alignments can be generated by the methods of Needleman and Wunsch(3), of Sellers(4), and the "local homology" method of Smith and Waterman(5). The Smith and Waterman alignment algorithm is also the most sensitive method available for identifying similarities between weakly related sequences.

Mapping and Searching

Mapping is available in several formats. Graphic maps display all of the cuts for each restriction enzyme on parallel lines. This graphic map facilitates selection of enzymes for isolating any region of a sequenced DNA molecule. Sorted maps in tabular format arrange the fragments from any digestion in order of molecular weight to show which fragments are similar in size and thus likely to be confused in gels. Another frequently used mapping format, designed by Frederick Blattner(6), displays the enzyme cuts above the original DNA sequence. Both strands of the DNA and all six frames of translation are shown.

All mapping programs will search for user-specified sequences, allowing features to be marked at the appropriate position on a restriction map. The mapping and searching programs can be used to aid site-specific mutagenesis experiments by showing where mutations could generate new restriction sites. All of the positions in a sequence where a synthetic probe could pair with one or more mismatches can also be located. Sequences related to less precisely defined features such as promoters or intervening sequence splice sites, can be located with a program that uses a consensus sequence as a probe. The

Table 1

Programs Available from UWGCG

Name	Function
DotPlot+	makes a dot plot by method of Maizel and Lenk(2)
Gap	finds optimal alignment by method of Needleman and Wunsch(3)
BestFit	finds optimal alignment by method of Smith and Waterman(5)
MapPlot+	shows restriction map for each enzyme graphically
MapSort	tabulates maps sorted by fragment position and size
Map	displays restriction sites and protein translations above and below the original sequence(Blattner,6)
Consensus	creates a consensus table from pre-aligned sequences
FitConsensus	finds sequences similar to a consensus sequence using a consensus table as a probe
Find	finds sites specified interactively
Stemloop	finds all possible stems (inverted repeats) and loops
Fold*	finds an RNA secondary structure of minimum free energy by the method of Zuker(7)
CodonPreference+	plots the similarity between the codon choices in each reading frame and a codon frequency table(8)
CodonFrequency	tabulates codon frequencies
Correspond	finds similar patterns of codon choice by comparing codon frequency tables (Grantham et al,9)
TestCode+	finds possible coding regions by plotting the "TestCode" statistic of Fickett(10)
Frame+	plots rare codons and open reading frames(8)
PlotStatistics+	plots asymmetries of composition for one strand
Composition	measures composition, di and trinucleotide frequencies
Repeat	finds repeats (direct, not inverted)
Fingerprint	shows the labelled fragments expected for an RNA fingerprint
Seqed	screen oriented sequence editor for entering, editing and checking sequences
Assemble	joins sequences together
Shuffle	randomizes a sequence maintaining composition
Reverse	reverses and/or complements a sequence
Reformat	converts a sequence file from one format to another
Translate	translates a nucleotide into a peptide sequence
BackTranslate	translates a peptide into a nucleotide sequence
Spew	sends a sequence to another computer
GetSeq	accepts a sequence from another computer
Crypt	encrypts a file for access only by password
Simplify	substitutes one of six chemically similar amino acid families for each residue in a peptide sequence
Publish	arranges sequences for publication
Poster+	plots text (for labelling figures and posters)
OverPrint	prints darkened text for figures with a daisy wheel printer

+ requires a Hewlett Packard Series 7221 terminal plotter
* Fold is distributed by Dr. Michael Zuker not UWGCG.

mapping programs can also be used on protein sequences to identify the peptides resulting from proteolytic cleavage.

Secondary Structure

Three programs are available to examine secondary structure in nucleic acids. The program StemLoop identifies all inverted repeats. An implementation of Dr. Michael Zuker's Fold program(7) finds an RNA secondary structure of minimum free energy based on published values of stacking and loop destabilizing energies. The "dot plot" comparison (mentioned above) of a sequence compared to its opposite strand gives a graphic picture of the pattern of inverted repeats in a sequence.

Analysis of Composition and the Location of Genetic Domains

Regions of a sequence with non-random base distribution can be displayed with three graphic tools designed to identify genetic domains. The program CodonPreference(8) identifies potential coding regions by searching through each reading frame for a pattern of preferred codon choices. The CodonPreference plot predicts the level of translational expression of mRNAs and helps identify frame shifts in DNA sequence data. Patterns of codon choice can be compared with the program Correspond(9). When a strong pattern of codon preferences is not expected, the "TestCode" statistic of Fickett(10) can be plotted to show regions of compositional constraint at every third base. Another program plots asymmetries of composition by strand. Strand asymmetries have been associated with genetic domains by several authors(11)(12). A fourth program called Frame marks the positions of rare codons and open reading frames on a graph showing all six reading frames.

Several tools are available to measure content and to count dinucleotide, trinucleotide, neighbor and repeat frequencies. A program that predicts RNA fingerprint patterns and another that tabulates codon frequencies complete the group of programs that analyze composition.

Sequence Manipulation

Sequences may be entered, assembled, edited, reversed, randomized, reformatted, translated, back-translated, documented, transferred, or encrypted rapidly with a large set of sequence manipulation tools.

A screen-oriented editor is available that allows sequences to be entered and checked. After a sequence is entered, it may be reentered for proofreading. Whenever a reentered base is at variance with the original, the terminal bell rings and the position is marked. Existing sequences can be edited quickly by moving directly to a sequence position specified by either a coordinate or a sequence pattern. The program can reassign the terminal's keys to place G, A, T and C conveniently under the fingers of one hand in the same order as the lanes of a sequencing gel.

Programs are available for changing sequence file format. Sequence data from any source can be used in UWGCG programs, and sequence files maintained with UWGCG software can be converted for use in other non-UWGCG programs. For instance, the programs of Roger Staden(13) or Intelligenetics Inc.(14) could be used to assemble a sequence from the sequences of many small sub-fragments generated by DNAase I digestion. The assembled sequence could then be reformatted for use in any UWGCG program. A program is available that transfers sequences to and from other computers.

Sequence Publication

A program, Publish, will format sequences into figures. Publish has alternatives for line size, numbering, scaling, translation and comparison to other sequences. Poster is a program that will plot text on figures.

GENERAL FEATURES OF UWGCG SOFTWARE

Interactive Style

Each program is run by simply typing its name. Every parameter required by the program is obtained interactively. Questions are answered with a file name, a yes, a no, a number, or a letter from a menu. Default answers are displayed. Programs are insensitive to absurd answers and will ask the question again if, for instance, you name a file that does not exist or if you use a nonnumeric character when typing a number. Special features such as plotting features oriented to publication, are obtained by using an extra word next to the program's name when the program is run. Thus parameter queries are kept to a minimum for the normal use of each program.

Data

Both the NIH-GenBank(15) and the EMBL(16) nucleotide sequence data libraries are available "on-line" to any UWGCG program. A Search utility will locate sequences in the libraries by key word. A Find utility will locate library entries containing any specified sequence. A program is available that installs the new data sent periodically from GenBank and EMBL to update their data libraries.

All of the data in the system are stored in text files that can be read and modified easily. Every data file has an English heading describing the contents. The data files may be copied by each user for analysis or modification. Programs recognize and read user-modified input data automatically. Data files can be modified with any text editor.

Sequence File Structure

Sequences are maintained in files that allow documentation and numbering both above and within the sequence. This file format is compatible with both of the nucleic acid sequence libraries and has been adopted as the standard sequence file format by the data base project at the European Molecular Biology Lab. Because genetic manipulations commonly involve linking several molecules of known sequence, UWGCG sequence files are designed to support concatenation by allowing comments to appear within the sequences at any location. Coding sequences or the boundaries between cloning vector and insert, for instance, can be marked within the sequence itself for immediate identification.

Sequence Symbols

All possible nucleotide ambiguities and all standard one-letter amino acid codes are part of the UWGCG symbol set that includes all alphabetic characters plus five additional characters. The proposed IUB-IUPAC standard nucleotide ambiguity symbols(17) are used for the mapping, searching and comparison programs. Lower case characters are used in sequences to indicate uncertainty as distinct from ambiguity. This allows the entire lexicon of symbols to be reused with same meaning, but with the prefix "maybe-." This reuse of the symbol set in lower case makes the uncertainty symbols more complete, understandable and visible.

Symbol Comparison

Sequence analysis programs generally make comparisons between sequence symbols (bases or amino acids) in order to find enzyme sites, create alignments, locate inverted repeats etc. These symbol comparisons are handled in several ways.

Symbol comparisons for alignment, comparison and secondary structure analysis are made by looking up a value in a symbol comparison table for the quality of the match. The table might contain 1's for matches and 0's for mismatches. If amino acids are being compared, however, a real number could be assigned at each position based on some previously assigned chemical similarity of the pair of residues or on the mutational distance between their codons. Standard symbol tables are provided by UWGCG, but the system is designed to allow each user to specify his own values.

Symbols comparisons for mapping and searching operations in nucleic acids are made by converting the IUB-IUPAC symbols into a binary code. The bits of this code represent G, A, T and C with ambiguity symbols causing more than one bit to be set. A group of library functions identify overlap between the bits for each IUB-IUPAC symbol.

Documentation

Documentation is available both in printed form and on the terminal screen. A 350 page manual describes the operatiọn of each program in detail, gives practical considerations and shows what will appear on the screen during a session with the program. Output files and plots are shown for the session. The data for the session shown in the documentation are included with the system so that the each program's operation can be checked. The "on-line" documentation is the same as the manual, but can be changed immediately when a program is modified.

All programs write output to files that are completely documented and sensibly organized for input to other programs. The input data, the program and the parameters used are clearly identified in every output file.

Procedure Library

UWGCG programs are written largely as calls to a library of 250 procedures designed to manipulate biological sequences. These procedures use data and file structures which have been designed to simplify program modification. For instance, standard operations such as reading sequences from files are always handled by a single library procedure. Thus a change in sequence file format requires only one subroutine to be modified for the new format to be acceptable to all of the programs in the system. Command procedures are available to help modify the library. The procedure library can be used by programs written in any language.

DISTRIBUTION OF UWGCG SOFTWARE

Intent

The intent of UWGCG is to make its software available at the lowest possible cost to as many scientists as possible.

Fees

A fee of $2,000 for non-profit institutions or $4,000 for industries is being charged for a tape and documentation for each computer on which UWGCG software is installed. While no continuing fee is required, UWGCG software, like the field it supports, is changing very rapidly. A consortium of industries and academic laboratories is planned to support the project in the future. The consortium will entitle its members to periodic updates and to influence the direction of new programming undertaken by UWGCG in return for a pledge of continuing financial support.

Copyrights

UWGCG retains the copyrights to all of its software and UWGCG must be contacted before all or any part of the its software package is copied or transferred to any machine. UWGCG is, however, mandated to provide research tools to help scientists working in the area of molecular genetics and we are glad to see our source codes become the basis of further programming efforts by other scientists. Copyright can be removed for any program modified by more than 25% of its original effort.

Tape Format

The UWGCG package is usually distributed in VAX/VMS "backup" format on a 9 track magnetic tape recorded at 1600 bits/inch. The system consists of about 1000 files using about 20,000 blocks at 512 bytes/block. The current versions of the GenBank and EMBL nucleotide sequence data bases are normally included which add another 3,000 files and require another 20,000 blocks.

Upon request UWGCG will make a card image tape of all of the Fortran 77 programs and procedures for reading on computers other than the VAX. The card image tape is usually provided at 1600 bits/inch with 80 characters/record and 10 records/block. Adaptation of UWGCG software to systems other than VAX/VMS may take considerable effort.

Equipment Required

UWGCG programs and command procedures will run on a Digital Equipment Corporation (DEC) VAX computer that is using version 3.0 or greater of the DEC VMS operating system. A tape drive is necessary; a floating point accelerator and a DEC Fortran compiler are helpful, but not required. All programs can be run from a DEC VT52 or VT100 terminal. Seven programs, as noted in table 1, require a Hewlett Packard 7221 terminal plotter wired in series with the terminal. Several utilities support a daisy wheel compatible printer attached to the terminal's pass-through port, however, all programs write output files suitable for printing on any standard device.

Inquiries

Inquiries may be sent to John Devereux at the Laboratory of Genetics, University of Wisconsin, Madison, WI, USA 53706, (608) 263-8970. UWGCG is not licensed to distribute Fold(7), but the UWGCG implementation is available from Michael Zuker, Division of Biological Sciences, National Research Council of Canada, 100 Sussex Drive, Ottawa, Canada, K1A OR6 (613) 992-4182.

ACKNOWLEDGEMENTS

UWGCG was started with software written for Oliver Smithies' laboratory with NIH support from grants GM 20069 and AM 20120. UWGCG is directed by John Devereux and is operated as a part of the Laboratory of Genetics with the advice of a steering committee consisting of Richard Burgess, James Dahlberg, Walter Fitch, Oliver Smithies and Millard Susman. UWGCG is currently supported with intramural funds and with fees paid by the faculty and industries using the facility in Madison. This article is paper number 2684 from the Laboratory of Genetics, University of Wisconsin.

*Current address: Silicon Graphics Inc., 630 Clyde Court, Mountain View, CA 94043, USA

REFERENCES

1. Kernighan, B.W. and Plauger, P.J. (1976) Software Tools, Addison-Wesley Publishing Company, Reading, Massachusetts.
2. Maizel, J.V. and Lenk, R.P. (1981) Proceedings of the National Academy of Sciences USA 78, 7665-7669.
3. Needleman, S.B. and Wunsch, C.D. (1970) Journal of Molecular Biology 48, 443-453.
4. Sellers, P.H. (1974) SIAM Journal on Applied Mathematics 26, 787-793.
5. Smith, T.F. and Waterman, M.S. (1981) Advances in Applied Mathematics 2, 482-489.
6. Schroeder, J.L. and Blattner, F.R. (1982) Nucleic Acids Research 10, 69-84, Figure 1.
7. Zuker, M. and Stiegler, P. (1981) Nucleic Acids Research 9, 133-148.
8. Gribskov, M., Devereux, J. and Burgess, R.R. "The Codon Preference Plot: Graphic Analysis of Protein Coding Sequences and Gene Expression," submitted to Nucleic Acids Research.
9. Grantham, R. Gautier, C. Guoy, M. Jacobzone, M. and Mercier R. (1981) Nucleic Acids Research 9(1), r43-r74.
10. Fickett, J.W. (1982) Nucleic Acids Research 10, 5303-5318
11. Smithies, O., Engels, W.R., Devereux, J.R., Slightom, J.L., and S. Shen, (1981) Cell 26, 345-353.
12. Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) Nucleic Acids Research 11, 2205-2220.
13. Staden, R. (1980) Nucleic Acids Research 8, 3673-3694.
14. Clayton, J. and Kedes, L. (1982) Nucleic Acids Research 10, 305-321.
15. The GenBank(TM) Genetic Sequence Data Bank is available from Wayne Rindone, Bolt Beranek and Newman Inc., 10 Moulton Street, Cambridge, Massachusetts 02238, USA.
16. The EMBL Nucleotide Sequence Data Library is available from Greg Hamm, European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 6900 Heidelberg, West Germany.
17. Personal communication from Dr. Richard Lathe, Transgene SA, 11 Rue Humann, 67000 Strasbourg, France.