Volume 10 Number 1 1982 Nucleic Acids Research

Computer programs to analyze DNA and amino acid sequence data

Katsumi Isono

Max-Planck-Institut für Molekulare Genetik, Abteilung Wittmann, Ihnestrasse 63-73, 1000 Berlin 33, GFR

Received 4 September 1981

ABSTRACT

Extensive modifications have been incorporated into many of the computer programs written by Staden (1-4) to make them easier to use. In addition, six programs have newly been created to cope with DNA and amino acid sequence data. These programs can be easily used by persons with minimal knowledge of computers.

INTRODUCTION

Analysis of DNA and amino acid sequence data by computer is a very powerful technique. In recent years several papers have appeared describing computer programs suitable for this purpose (1-7). In particular, the programs published by Staden (1-4) are very useful for analyzing DNA and amino acid sequence data. With them one can easily store, improve and analyze the sequence data. When these programs are combined with the M13 phage-cloning method for DNA-sequencing (8), one can easily handle a vast amount of sequence data in a short time.

In practical use of these computer programs (1-4), however, I found that they could be improved to become much more useful and easier to use. Therefore, I started rewriting many of the programs which are most frequently used. Furthermore, six programs have newly been written to cope with sequence data, which I found very practical for our purpose. All these programs are written in FORTRAN for the DEC-10 (Digital) system. The programs as well as a detailed user's manual are available upon request.

DESCRIPTION OF THE PROGRAMS

Table 1 lists the computer programs written by Staden (1-4) which we use most frequently. Almost all of them have been considerably modified to meet our demand and to make them easier to use. One of the most common modifications I have incorporated is to avoid use of a subroutine termed ASDAT. Use of this subroutine is very convenient to write a program, but it often creates a situation in which the user gets confused, because it is not clear when he is asked "PLEASE TYPE FILE 1", or, "PLEASE TYPE FILE 2", which file he should enter now. Another type of common modification is the incorporation of options at the end of each program which enable the user to continue working with the program for the same or other data files without unloading the program from the computer. Furthermore, some of the programs listed in Table 1 have been modified and renamed. AASERK is derived from AATNUK and the original AASERK combined, BASPAR is derived from BPFTBK, GELFIT is from OVRLAP, REVCOM from SQRVCM, and SEARCH from SRCHMU.

The program DBFIX is the one that has been most extensively modified. This is one of the key programs for handling DNA sequence data obtained by the M13 phage-cloning method. One of the modifications I have incorporated here is to allow withdrawal of any gel readings from the database, and another is to make it possible to alter individual gel readings by editing them separately, but by using the common relative position numbers within a "contig" (for the definition of the term contig, see ref. 4). Furthermore, during the course of sequencing DNA using the M13 cloning method, several times we encountered a situation in which two fragments of DNA had been tandemly ligated into one M13 phage clone. By using other restriction enzymes for cloning and sequencing, we found that these two fragments were derived from two different regions of the original DNA. Therefore, to cope with such a situation, an additional modification has been incorporated into DBFIX, with which the operator can split the gel reading (sequence) in question into two halves. Either the right or the left half can be maintained in the database. Program takes all possibilities into account and recalculates the paramters for the remaining gel within

Table 1: List of computer programs for sequence analysis

Program	Function
AACOMP*	Amino acid sequence comparison from new gel readings.
AAEDIT*	Storing and improving amino acid sequence data.
AAFIT	Amino acid sequence comparison.
AALIST*	Amino acid sequence display.
AASERK†	Search for probable restriction enzyme sites within a 'DNA' sequence created from an amino acid sequence.
ARCLST*	Display of sequences stored in an archive data file.
BASPAR†	Search for base-pairs in a DNA sequence
BATIN	Storing DNA and amino acid sequencing data.
CONLST*	Display of sequences stored in a consensus sequence file.
CONSEN	Writing consensus sequences.
CODTOT	Calculation of codon usage in a given gene.
DBCOMP	Comparison of new gel readings vs. consensus sequences.
DBEXAM	Statistical analysis of a database.
DBFIX	Solving various problems in a database.
DBUTIL	Alignment of gel readings to create a database.
DIRREP	Search for direct repeats.
FILINS	Creation of a new sequence file by extracting regions of existing files.
GELFIT†	Search for homology among gel readings.
HAIRPN	Search for palindromes within a DNA sequence.
MWCALC	Calculating amino acid composition and MW of a protein either from a DNA or amino acid sequence.
REVCOM†	Conversion of DNA sequence into its complementary one.
SEARCH†	Search for various restriction enzyme sites.
SEQEDT	Storing and improving DNA sequence data.
SEQFIT	DNA sequence comparison.
SEQLST	DNA sequence display with various format options.
TRANSQ	DNA to amino acid translation with various format options.
TRANDK	Storing amino acid sequence data on disk.
TRANDS*	Modified version of TRANDK.

*These are new programs (see text).
†These programs are modified and renamed (see text).

the contig.

Programs BATIN, DBCOMP and CONSEN are altered so that the file of file names is read by computer in random access mode. This makes it easier for one to store the file names of new gel readings, because one has to use only one and the same file name upon repeated execution of BATIN. To compare the new gel readings with the consensus sequences of the previous gel readings, DBCOMP asks the operator from which gel to which gel he wants to compare. A new option to avoid repetitive comparison among the new gel readings is also incorporated into DBCOMP, since the necessity for such an option has often been encountered. To make it easier to have the stored sequence data of gel readings displayed or printed, a new program termed ARCLST has been created. This program enables the user to get any gel readings printed by simply supplying the program with the gel numbers. Similarly, sequences stored in a consensus sequence file can be printed by the new program CONLST.

By incorporating more gel readings into a database using the program DBUTIL, some of the contigs become modified, and therefore, the consensus sequences for them must be updated. The newly incorporated option in the program CONSEN exists to cope with such a situation. With this option, one can delete the old consensus and create a new one.

TRANDS is a modified version of TRANDK, which automatically writes translation of DNA sequence in three phases into a disk file. AAEDIT and AALIST are modified versions of SEQEDT and SEQLST, respectively, to deal with amino acid sequence data more specifically. AALIST displays amino acid sequences from either a DNA or one-letter amino acid sequence file in either a one-letter or three-letter format.

A new program AACOMP has been created to deal with such a situation in which one wants to compare the DNA sequence data with any genes of which one only knows the amino acid sequence. This program uses the same algorithm used in DBCOMP for sequence comparison.

ACKNOWLEDGEMENT

I am very grateful to Dr. R. Staden for providing all his

computer programs and to Dr. P. Wills for assistance in the early stage of this work.

REFERENCES

1. Staden, R. (1977) Nucleic Acids Res. 4, 4037-4051.
2. Staden, R. (1978) Nucleic Acids Res. 5, 1013-1015.
3. Staden, R. (1979) Nucleic Acids Res. 6, 2601-2610.
4. Staden, R. (1980) Nucleic Acids Res. 8, 3673-3694.
5. Queen, C.L., Korn, L.J. (1980) Methods in Enzymology 65, 595-609.
6. Gingeras, T.K., Roberts, R.J. (1980) Science 209, 1322-1325.
7. Sege, R., Söll, D., Ruddle, F.H., Queen, C. (1981) Nucleic Acids Res. 9, 437-444.
8. Messing, J., Crea, R., Seeburg, P.H. (1981) Nucleic Acids Res. 9, 309-321.