A computer program for comparative analysis of nucleic acid sequences

Paul A Blanz and Siegfried Kleindienst

Lehrstuhl Spezielle Botanik der Universität Tübingen, Auf der Morgenstelle 1, D 7400 Tübingen, FRG

Received 6 June 1985

ABSTRACT
The programs offer the possibility of comparing pairs of homologous sequences in order to find out percentage of homology, number of identical and deviating nucleotides, of transitions and transversions and, derived from these, KNUC-values according to Kimura (1) and the corresponding standard error sigmaK. The sequences can be printed in pairs underneath each other, homologies are indicated by asterisks between the identical nucleotides. Out of a set of homologous sequences stored on a disk any number of sequences can be compared in pairs in this way, and a matrix containing either the percentage of homology values, the number of deviating nucleotides or the KNUC-values together with the corresponding standard errors can be sent to screen, printer or disk. A program will be available soon which creates a dendrogram representing the similarity between the sequences by use of an average linkage clustering method deduced from this matrix. The programs are written for Apple II computers using UCSD-PASCAL and for Sirius I/Victor 9000 computers using TURBO-PASCAL.

INTRODUCTION
The use of computers has become an undisputed prerequisite for the analysis of nucleotide sequences. Many programs have been described which now handle most of the routine work. In the meantime, microcomputers have been made capable of running universal programs handling even rather long sequences. Some of these programs are written in PASCAL (2, 3). Because they are available as well-documented source files, they can be adapted for individual needs quite easily. They can also be used for handling many basic functions, such as input of data or output of the sequences in various formats. However, nearly all of these programs facilitate only the handling of single sequences.

Copies of source or object code are available to anyone requesting them. Apple users send two 5.25 inch diskettes, Sirius I/Victor 9000 users send one of the same size. The programs are not available on tape.
Here we describe a program which we originally developed for quantitative comparisons of pairs of 5S ribosomal RNA nucleotide sequences in order to calculate the numbers of identical and deviating nucleotides, of transitions and transversions, of percentage of homology and of special similarity coefficients.

PROGRAM DESCRIPTION

The package consists of the three programs: "compare", "matrix" and "dendro". They must be supplemented by a data file for each nucleotide sequence and an additional text file "ICAT.TXT", which represents the internal catalogue containing all the data file names together with comments. The data file names have to be truncated to ten or eight letters, as is allowed in UCSD-PASCAL or MS-DOS, respectively. Comments of any length can be inserted in the data files as long as a semicolon starts each line containing a comment. The same way has been used before for including comments (2, 3). Therefore, existing data files can be used. New data files can be created by use of former authors' data input procedures. Data can also be put in by using the PASCAL editor when working with UCSD-PASCAL, or by applying a text editing program under MS-DOS for TURBO-PASCAL, respectively.

The program "compare" offers two different functions. First, it compares any number of nucleotide sequences in pairs, marks identical nucleotides, counts the total number of bases, the numbers of identical and deviating nucleotides, of transitions and transversions, calculates percentages of homology and KNUC values together with their standard errors. These KNUC values and standard errors were introduced by Kimura when he described his "neutral theory" (1). Instead of only counting the number of nucleotide exchanges, Kimura also takes into account whether the type of alteration is a transition or a transversion (1). The second function of "compare" is to print many nucleotide sequences underneath each other.

Our program is interactive. First, the internal catalogue containing all names of sequence data files on disk is shown in form of a numbered list on screen. The user is then asked, how many sequences he wants to have printed underneath each other.
5S RNA COMPARISON

(1) ustihord (SEQUENCE LENGTH = 118)
(2) ustimayd (SEQUENCE LENGTH = 118)

10  20  30  40  50  60

1 AUGCUGCGGCCAUAGAAACCUCUUGCACCCGCAUCCCGGUCGCAUGCUGCGAAAGGAAGCAAG
H ******** *** ****** **************************** *** ******
2 AUGCUGCGGCCACAGAGACUUGAAAACAAACCACUCCGGCUGCAUGCUGCGCAUGUCGAAGCAAG

70  80  90  100

1 GAUUGCUCGAUGUCAGUACUGCGGUGGACCACCGGGAAAUCCUA-GGCGCGCAGGCUU
H ******** **************** ********* ********
2 UCGUCGCUCAGCCAGYACUGCGGUGGGGACCAGCGCGGAAUCCUA-GGUGCGGAGGCUU

HOMOLOGY BETWEEN THE TWO SEQUENCES: 89.83 %
NUMBER OF HOMOLOGOUS NUCLEOTIDES: 106
NUMBER OF DEVIATING NUCLEOTIDES: 12
NUMBER OF TRANSITIONS: 9
NUMBER OF TRANSVERSIONS: 3
Knuc-value: 0.111
sigmaK-value: 0.033

Figure 1: Comparison of the 5S ribosomal RNA nucleotide sequences from the smut fungi Ustilago hordei and Ustilago maydis (5) by use of the "compare" program. The asterisks mark identical nucleotides. For further explanations see text.

If the answer is "2", the user will be asked if he likes to have homology indicated. If yes, all the calculations mentioned above will be executed and, as a matter of choice, sent to screen, be printed on a line printer or stored on disk. The user is asked to state from which nucleotide sequence on and how far he wants the comparison to run. He can enter many combinations at a time. Then the starting point may be determined differently for each nucleotide sequence or at nucleotide number 1 for all sequences. After a heading line has been entered, the computer starts with the comparisons and prints them on screen or printer, or stores them on disk. This process may take hours depending on the number of nucleotide sequences to be compared, on disk access time and on the speed of the printer. Fig. 1 shows a printout of a sequence comparison indicating identical nucleotides.

If no indication of homology is wanted, the chosen number of sequences will only be printed underneath each other. The user will be asked which order the nucleotide sequences shall
be printed in. The output shows first a numbered list of sequence names and the length of the individual sequences. Then the nucleotide sequences will be printed underneath each other, 60 nucleotides per row, every tenth position numbered. This type of printout is especially helpful for the manual alignment of the nucleotide sequences.

The program "matrix" arranges the data of the sequence comparisons stored on disk in a matrix pattern. The user can chose whether he wants the percentages of homology, the number of deviating nucleotides or the KNUC values arranged in a matrix. The matrix will be stored on disk and can be printed on a line printer.

At present, we are about to finish the program "dendro", which will draw dendrograms from the matrix created by the program "matrix". "Dendro" is written in FORTRAN. It offers the choice between single, complete and average linkage clustering of the data in the matrix. It is completely menue driven. Dendrograms can be drawn by any 80 column line printer. Because this simplification causes a loss in precision, a list of all intersections of the dendrogram can be printed in order to enable users to draw a more precise dendrogram by hand and to use a different arrangement of the clusters if wanted. "Dendro" will be available on request together with the programs "compare" and "matrix" by the time this article will be published.

IMPLEMENTATION DETAILS

These programs can be run either on an Apple II plus with 64 kB RAM, an 80 column card and two Apple disk drives using UCSD-PASCAL, or on a Sirius I/Victor 9000 computer using MS-DOS and TURBO-PASCAL. The latter version can be easily transferred to an IBM PC or other computers using MS-DOS. Because no high resolution graphic was used, any 80 column line printer will be appropriate. The Apple PASCAL and the TURBO-PASCAL system must be purchased from Apple or from Borland International, respectively.

CONCLUSIONS

Our program package has already been extensively used for comparing about 50 different 5S rRNA nucleotide sequences at
once, calculating the similarity between them and, derived from this, for presenting graphic presentation of relationships between different sequences (4, 5). These programs proved to be a helpful tool for studying phylogenetic relationships between different organisms by comparison of corresponding nucleotide sequences. However, the applicability for this purpose first depends to a great extent on the quality of the nucleotide sequence used as a phylogenetic marker, and secondly it is limited by the fact that only a single characteristic is used. It is advisable, however, to draw phylogenetic conclusions from more than one single characteristic, e.g. by comparison of nucleotide sequences of different genes or by considering other markers. Our programs are capable of providing an impression of phylogenetic relationships between organisms.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Axel Brennicke and Winfried Hellrung for helpful discussions and the Deutsche Forschungsgemeinschaft for financial support by grant Bl 190/3-1 to P.B.

REFERENCES
