Update for users of the Cornell sequence analysis package

John S. De Banzie, Evan W. Steeg and John T. Lis

Section of Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, NY 14853, USA

Received 17 August 1983

ABSTRACT

This paper reports some additions and improvements to the Cornell sequence analysis package for microcomputers (Fristensky, Lis, and Wu, 1982, Nucleic Acids Res. 10, 6451-6463).

INTRODUCTION

Following the publication of the sequence analysis package of Fristensky, Lis, and Wu (1), we have received many requests for copies of the programs. Here we report five additional programs that extend the capabilities of the original package and some improvements which can be made to those programs already published. Note that these are all written in Apple Pascal and will require modification for use with other systems. The programs were written and tested on an Apple IIe microcomputer.

NEW PROGRAMS

Five new programs are available. These are described below. All are extremely resistant to user errors such as failure to insert a diskette, typing mistakes, or entry of data which lies outside the acceptable range.

ALIGNSEQ: This program allows the detection and detailed examination of homologies between two sequences. It will take two sequences and, using a short segment of one, align them to give the greatest possible number of correct matches within that short segment. Once a satisfactory alignment has been found it is displayed (Figure 1a) and can be manipulated using the editing functions of the program (Figure 1b). The search segment can be selected by the program, in which case the user must define the expected region of overlap between the two sequences (i.e., is it at the 5' end, the 3' end, or the middle of the first sequence?), or it can be determined by the user. Repeated searching is possible and can be used to detect loop-outs. The editing functions allow deletions, insertions, or replacements of bases to

A) Output from ALIGNSEQ before editing.

```

1
ACCGTTTGAA CCGACTGGTC GATCGGATTC CGATTATGCG CTAGCTGAGC
          *****
          GGATTC CGATTATGCG CTAGCTGCAG

```

```

51
TAGCTTGATA GCTCCCGGCT AGTCGATTGC
          ** * *****
CTAGCTGAT AGCTCCCGGC TGTCGATTGC GACTGTGTCTG TGCTG

```

Upper strand is master sequence written 5' to 3'.

Lower strand is new sequence written 5' to 3'.

Search segment was 20 bases long, starting at 1.

B) Output from ALIGNSEQ after adding loop-out to master sequence.

```

1
ACCGTTTGAA CCGACTGGTC GATCGGATTC CGATTATGCG CTAGCTG-AG
          ***** ***** **
          GGATTC CGATTATGCG CTAGCTGCAG

```

```

51
CTAGCTTGAT AGCTCCCGGC TAGTCGATTG C
***** ***** * *
CTAGCTGAT AGCTCCCGGC TGTCGATTGC GACTGTGTCTG TGCTG

```

Upper strand is master sequence written 5' to 3'.

Lower strand is new sequence written 5' to 3'.

Search segment was 20 bases long, starting at 1.

C) Output from ALIGNSEQ after adding loop-out to new sequence.

```

1
ACCGTTTGAA CCGACTGGTC GATCGGATTC CGATTATGCG CTAGCTG-AG
          ***** ***** **
          GGATTC CGATTATGCG CTAGCTGCAG

```

```

51
CTAGCTTGAT AGCTCCCGGC TAGTCGATTG C
***** ***** * ***** *
CTAGCTGAT AGCTCCCGGC T-GTCGATTG CGACTGTGTCTG TGCTG

```

Upper strand is master sequence written 5' to 3'.

Lower strand is new sequence written 5' to 3'.

Search segment was 20 bases long, starting at 1.

Figure 1

be made in the final alignment. They also permit the user to align the two sequences manually if desired. The format used for these editing functions is very similar to that used by the Apple Pascal Editor, and they are therefore easy to learn and use. The program was designed specifically for those involved in DNA sequencing. It facilitates the alignment of overlapping sequences from different restriction sites and the generation of a cumulative master sequence. The editing functions can be used, in conjunction with the original autoradiographs, to resolve any differences or ambiguities in the original alignment. The program thus bears some resemblance to that of Staden (2). It can also be used for more general sequence comparisons, such as examining homologies detected by FASHOM or GRAHOM in more detail.

MARKSITE: This program takes the output from the previously published restriction site search programs INTREST and BACHREST (1) and prints out the sequence which was searched with the restriction sites marked on it (Figure 2). We have found this format to be much more convenient for many applications, such as designing cloning strategies or looking for restriction sites to use in high resolution S1 mapping or in vitro mutagenesis. The program deals automatically with isoschizomers and other enzymes which share recognition sites in accordance with certain preset rules. For example, BamHI and Sau3AI cut at the same site--the Sau3AI site will be suppressed and the BamHI site printed; FnuD2 and HhaI often cut at the same site (in the sequence GCGCG)--the fused name HhaFnu will be printed. These preset rules are contained in a separate data file and can be printed if desired. The program will, if requested, print a list of the enzymes which do not cut. Provision

Figure 1: Output from ALIGNSEQ. (A) The program was instructed to carry out an automatic search and informed that the new sequence (lower strand) extended beyond the 3' end of the master sequence (upper strand) and was from the same strand. The resulting output is shown. Asterisks indicate matches. The master sequence (which, for DNA sequencers, is intended to be the cumulative sequence) can be up to 10,000 bases long. The new sequence (which, for DNA sequencers, is intended to be the reading from a new gel) can be up to 1,000 bases long. The new sequence can be from either the same strand as the master sequence or from the complementary strand: the program will generate the same strand sequence if necessary. To facilitate editing, the sequence displayed will be that from the autoradiograph in both cases. (B) After printing (A), the program's editing functions were used to insert a loop-out into the master sequence. Alternatively, another base could have been inserted or the C deleted from the new sequence. For DNA sequencing, this would normally be done with reference to the original autoradiographs. (C) After printing (B), a loop-out was added to the new sequence. The only non-homology is now the T - G pair at position 56 in the master sequence. This could be resolved using the X(change command to exchange the T for a G or vice versa). A new master sequence could then be generated.

!Demonstration sequence.

-29

ACCCGGCATT AGGTATAACG ATTTGCCGGA AAAGTGGATG GCCTATGAAC

H		H	S	H	F
p		p	t	a	o
a		a	a	e	k
N		2	r	3	1
c			t		
i					

22

CACGTCAGGT AAGTAACCAAG ACCTGGGTC AATGATTATA GGAATACTTG

EI	E B A	IE
xn	c s v	nx
ot	o t a	to
nr	R N 2	rn
o	2 1	o
n		n

72

TACCAAGGTC CCGACAACATG ACACGAGCCA GTAAATAAGC GTCGGATCCG

R	A	H	S	B D
s	v	g	t	a p
a	a	a	o	m n
1	2	1	p	H 1
				1

122

TTGATC

M D
b p
o n
1 1

Figure 2: An example of the output from MARKSITE. The sequence shown was searched using BACHREST (1) and a file of commercially available restriction enzymes. The output was sent to a file and then used as input for MARKSITE. The restriction enzymes always cut the sequence immediately to the right of their names. Note the following: -(1) The numbering begins at -29. The user can define the first number written and whether to number in ascending or descending order. Zero is always skipped. In the example shown, the numbering is such that the start of transcription is 1. (2) Hpa2 and NciI both cut at position -27. Since they do not always cut at the same site the names are fused. (3) User supplied markers have been inserted at positions 1, 29, 30, 62, 63 and 111. (4) FokI cuts nine bases downstream of the sequence GGATG. The recognition sequence is at position 7. (5) EcoR2 and BstNI share the same recognition sequence but cut it in different places. They are therefore printed separately. (6) MboI and Sau3AI always cut at the same site as BamHI, therefore the BamHI site is printed and the others suppressed. In the case of an MboI (and hence Sau3AI site) which is not a BamHI site, MboI is printed and Sau3AI suppressed.

is also made for the user to indicate other sites, such as the start of transcription, splice sites, etc.

MARKTRAN: This program is essentially the same as MARKSITE except that it permits the user to mark particular segments of the sequence, such as

;Demonstration sequence.

```

-29
ACCCGGCAIT AGGTATAACG AITTGCCGGA AACCTGGATG GCCATGAAC
  H          ^^^^^          H  -  -----  H-----F-
  p          p          S
  a          a          t          +++  e+++++***k*
  N          2          a          3          1
  c          r
  i          t

22
CACGTCAGGT AAGTAACCG ACCCTGGGTCC AATGATTATA GGAATACTTG
----- E.B..A. ....
##### ##### c s v ##### ##
***** o t a *****
          R N 2
          2 1

72
TACCAAGGTC CGCACAACCTG ACACGAGCCA GTAAATAAGC GTCGGATCCG
R-----A-----H-----B D
s v g S a p
a*****a*** *****== = a t m n
1 2 1 o H 1
p 1

122
TTGATC
M D
b p
o n
1 1

```

Figure 3: Output from MARKTRAN, derived from the same data as figure 2. Note the following: -(1) The output is essentially similar to that from MARKSITE. (2) Non-overlapping regions can be marked with different symbols on the same line - e.g. the TATA box (A), exons (-), introns (.), and the poly A addition sequence (~). (3) Overlapping regions can be marked on different lines - e.g. the splice consensus sequences are marked (#) on the second line and the protein coding region on the third, with amino-acids removed during processing (+) and those found in the mature protein (*) distinguished. The stop codon (=) is also marked. Up to six lines can be marked. There are limits to the number of regions and markers that can be added to the sequence and the program checks to prevent these from being exceeded. However, these limits will only be encountered in extremely complex diagrams.

transcribed regions (Figure 3). It is intended primarily for the preparation of figures for publication.

SEQCHANGE: This program takes a sequence and numbers and formats it (Figure 4). It will also generate the opposite strand of a sequence and number and format it. This is necessary if one is to use FASHOM or GRAHOM (1)

```
;WRITESEQ-50
;Demonstration sequence.

1
ACCCGGCAIT AGGTATAACG ATTGCGCGGA AAACGGGATG GCCTATGAAC

51
CACGTCAGGT AAGTAACGAG ACCTGGGTCC AATGATTATA GGAATACTTG

101
TACCAAGGTC CGCACAACCTG ACACGAGCCA GTATAAAGC GTCGGATCCG

151
TTGATC
```

Figure 4: Output from SEQCHANGE. The program is designed to be very simple to use and so the format (five groups of ten bases per line, numbered from 1) is fixed. The heading (;WRITESEQ-50) indicates that the sequence conforms to this format. Programs under development will allow files in this format to be accessed more rapidly.

to search for inverted repeats. SEQCHANGE is more convenient than NUMSEQ (1) for these tasks in that one does not have to deal with the large (and somewhat complex) parameter menu of NUMSEQ. Further, the output from SEQCHANGE is much more compact than that from NUMSEQ, greatly increasing the amount of numbered sequences which can be stored on a diskette.

WRITESEQ: This program is designed to facilitate the entry of sequences into the computer. It will only accept the letters "A," "C," "G," "T" and "N" and will generate a different tone for each, if desired. If another character is entered by error the program will reject it and signal the error. If an incorrect base is entered it can be corrected by backspacing. This program is useful if one cannot touch-type and will help to reduce errors even if one can. Once the sequence is entered it is automatically numbered and formatted using the same format as SEQCHANGE.

Improvements to Existing Programs

The existing programs are sensitive to input/output errors: for example entering an incorrect file name or failing to have the correct diskette in the appropriate drive unit will cause most of them to fail. This can be extremely irritating if the program has just read in a long sequence. Modifications have been written which circumvent these problems. In addition, the reading of sequences into memory has been made up to 40% faster by omitting the echoing of the sequence to the monitor.

Availability of New Programs and Improvements to Existing Programs

The new programs and instructions for adding the improvements to the existing programs are available to anyone requesting them. Send three 5.25 inch diskettes (preferably formatted for Apple Pascal--please indicate whether formatted or blank). The programs are not available on tape. All requests should be addressed to Dr. John Lis. Requests for technical information should be addressed to Dr. John de Banzie.

ACKNOWLEDGEMENTS

We would like to thank Bill Noon for helpful discussions regarding ALIGNSEQ and Brian Fristensky for his constructive criticism of the manuscript. This work was supported by a NATO Postdoctoral Fellowship to J.S. de B. and NIH research grant GM25232 to J.T.L.

REFERENCES

1. Fristensky, B., Lis, J., and Wu, R. (1982) Nucleic Acids Res. 10, 6451-6463.
2. Staden, R. (1979) Nucleic Acids Res. 6, 2601-2610.