Arch. Biol. Med. Exp. *17*: 15-19 (1984)
Printed in Chile

ANALYS. DNA: A Computer Program for Nucleic Acid Sequence Data Processing *

ANALYS. DNA: Un Programa de Computación para el Procesamiento de Secuencias de Acidos Nucleicos.

RODOLFO AMTHAUER[1] and ALEJANDRO ARAYA

Instituto de Bioquímica, Facultad de Ciencias,
Universidad Austral de Chile, Valdivia, Chile.
(Received December 27, 1983)

SUMMARY

A computer program written in BASIC language is described. The program allows processing and analysis of DNA data and has been designed to be used by persons with little or no computer experience.

The operator using different options can search for direct homologies with varying degrees of matching, generate complementary strands, find restriction sites, invert the polarity of the sequence and edit a print-out.

Molecular cloning and rapid sequence analysis (1,2) have provided an enormous amount of data which has to be processed rapidly and accurately. The only way to handle and analyze large sequences is by computer processing. For this purpose a number of computer programs have been written in recent years (3,4).

We are currently studying the organization of *Cyprinus carpio* mitochondrial DNA. For this purpose we have developed a simple computer program that permits the handling and analysis of the obtained sequences. We describe here an interactive program which offers different options to the user. This program permits the user: a) to compare DNA sequences recording the highest degree of direct homology; b) to create the complementary strand; c) to search for endonuclease restriction sites; d) to invert sequences; and e) to print out filed sequences in an edited form.

MATERIAL AND METHODS

The program was written in BASIC language and developed using a DEC 2020 digital computer system.

Cyprinus carpio mitochondrial DNA was cloned in pBR 325(+) and partially sequenced by the Maxam and Gilbert procedure (1). The human mitochondrial DNA sequence (5) was used for comparison purposes. All the sequences to be processed were stored in a file disk.

PROGRAM DESCRIPTION

The program offers the user different options: *Homology, Complement, Restriction, Invert* and *List.*

These options were developed as independent programs. They are commanded by a program that allows the use of any combination of the different options. Invoking the program, the user is asked for the option he will use. When the option is accessed, the user is asked for the file of the data to be processed. There is a 17000 bases limit for the nucleic acid sequence to be analyzed.

Homology. This option searches for direct homology between two DNA sequences. The sequence to which homology will be searched is given in a data file. The user can introduce the sequence to be compared either as a file previously stored on disk or by typing it directly on the remote terminal (the limit for this sequence is 500 bases). The operator has to assign the minimum percentage of bases that

* This work was supported by grant S-82-21 from Dirección de Investigación y Desarrollo, Universidad Austral de Chile.

1 To whom inquiries about the program and correspondence should be addressed. Programs will be made available for non-profit use upon written request.

must be homologous between both sequences. The regions of higher or equal homology to the value assigned between both DNA sequences are recorded on a file. An example is shown in Fig. 1. The asterisks between the two sequences indicate where matches of bases occurs. Also the percentage of homology obtained is given. If no region of homology is obtained the message "Homology not found" appears.

```
HOMOLOGY BETWEEN HUMMT.DNA  AND  INVA2.SEQ
*****************************************

SEQUENCE HUMMT.DNA  FROM :  6230   TO :  6318

PERCENTAGE ASSIGNED =  60  %

PERCENTAGE OBTAINED =  75.2809  %

      6239        6249        6259        6269        6279        6289
CACACGAGCA TATTTCACCT CCGCTACCAT AATCATCGCT ATCCCCACCG GCGTCAAAGT
**     * * ***** ** *  *** ** ** *** *****  ** ** ** * * ** *****
CAACTCGGTA TATTTTACAT TCGCAACAAT AATTATCGCA ATTCCAACAG GTGTAAAAGT
        10          20          30          40          50          60

      6299        6309        6319
ATTTAGCTGA CTCGCCACAC TCCACGGAA
**********  * ******* *** ****
ATTTAGCTGA TTAGCCACAC TCCGCGGAG
        70          80          90
```

Fig. 1: Region of highest degree of homology between human mitochondrial DNA (HUMMT. DNA) and carp mitochondrial DNA (INVA2.SEQ). The upper sequence corresponds to human mtDNA and the lower one to carp mtDNA.

Complement. This program provides two choices to generate the complementary strand of the input sequence. One records the complementary strand in a file as data to be used for another option of the program. The other records the original sequence and its complementary in a file with 60 bases per line, with empty spaces every ten bases, each block of 10 boxes being correlatively numbered (Fig. 2). The operator is asked what choice to use. The name of the file where the sequences has to be recorded must be given.

Restriction. This program permits the identification of the recognition sites for any restriction enzyme in a given DNA sequence. The program is complemented with a file *(Restriction)* that contains the restriction enzyme and their corresponding recognition sequences according to the last compilation by Roberts (6). For restriction endonucleases that recognized the same sequence (isoschizomers), only the first enzyme isolated is included in the file.

The operator is asked for the name of the restriction enzyme. Then the user is asked if he wants to look for another site and so on. A complete identification of all the restriction sites in a given sequence can be made by typing "All" instead a name of a restriction enzyme. For restriction enzymes that have more than one recognition sequence, the following key was used: J correspond to nucleotides A or C; K to nucleotides G or T; N to nucleotides A, C, G or T; R to nucleotides A or G; Y to nucleotides C or T; X to nucleotides A or T; and Z to nucleotides C or G.

For restriction endonucleases that recognize non palindromic sequences, both possible, the 5'→ 3' and the 3' → 5' recognition

```
SEQUENCE : BAMA1.SEQ  AND HIS COMPLEMENTARY
**********************************************

        10          20          30          40          50          60
GATCCCTCCT AGGACTATGC TTAATTACCC AAATTTTAAC CGGCCTATTC CTAGCCATAC
CTAGGGAGGA TCCTGATACG AATTAATGGG TTTAAAATTG GCCGGATAAG GATCGGTATG

        70          80          90          100         110         120
ACTACACCTC AGACATCTCA ACCGCATTCT CATCTGTTAC CCACATCTGC CGAGACGTAA
TGATGTGGAG TCTGTAGAGT TGGCGTAAGA GTAGACAATG GGTGTAGACG GCTCTGCATT

        130         140
ATTACGGCTG ACTAATCCGT AATG
TAATGCCGAC TGATTAGGCA TTAC
```

Fig. 2: Complementary strand of a region (BAMA1.SEQ) of carp mitochondrial DNA created by the option complement.

sequences were considered, e.g. Hga I. However as with all the restriction endonucleases recognition sequences, the program searches only in the 5' → 3' sense (Figure 3).

The output with this option is recorded on a file and gives a list with the name of the enzyme used, its recognition sequence, the number of sites found, and the localization indicating the number where the recognition sites began in the sequence (Fig. 3).

Invert. This option allows the operator to convert the original sequence written 5' → 3' to 3' → 5' and viceversa. The inverted sequence is stored as a data that can be used in another option. The operator is asked for the name of the file where the inverted sequence will be recorded. *Invert* is useful for processing complementary strands obtained by the *Complementary* option. Sequences obtained after 3' end labeling (7, 8) can be introduced in the data file by direct reading of the gel and then, inverted for processing.

DISCUSSION

The program *Analys. DNA* was designed in the interactive form to facilitate its use, without need of previous computer experience. It offers the possibility to handle efficiently the sequence data and allows the screening of restriction sites and direct DNA homology analysis. Furthermore, the homology option is also suitable to search for direct homology between two proteins. In this case the protein sequence must be expressed in the one letter amino acid code.

Although BASIC is not the most efficient computer language, is one of the most widely used in microcomputers. In addition, all computers allow its use. This is a distinct feature of this program, because it can be potentially used in any computer with only minor modification, The latter can be performed by any person with knowledge in BASIC language.

The fact that each option is an individual program offers the possibility that the options or programs can be adapted for use on a microcomputer. A big package of programs is not always easy to compatibilize with other computers or microcomputers (9).

Essentially, the program described here will be compatible with any DEC (Digital) system. Care must be taken with the memory capabilities of each system. The size of the memory to be used is related to the length of the sequence to be processed.

This program was successfully used in the analysis and processing of the carp mitochondrial DNA sequences* obtained in our laboratory.

* Araya, A.; Amthauer, R.; León, G. and Krauskopf, M. submitted to Mol. Gen. Genet.

```
SEQUENCE : C2A1.SEQ
**********************

        10          20          30          40          50          60
TGACTTGAAG AACCACCGTT GTTATTCAAC TACAAGAACC ACTAATGGCA AGCCTACGAA

        70          80          90         100         110         120
AAACACACCC TCTCATTAAA ATCGCTAACG ACGCACTAGT TGACCTACCA ACACCATCCA

       130         140         150         160         170         180
ACATCTCAGC ATGATGAAAC TTTGGATCCC TCCTAGGACT ATGCTTAATT ACCCAAATTT

       190         200         210         220         230         240
TAACCGGCCT ATTCCTAGCC ATACACTACA CCTCAGACAT CTCAACCGCA TTCTCATCTG

       250         260         270         280
TTACCCACAT CTGCCGAGAC GTAAATTACG GCTGACTAAT CCGTAATG
```

```
RESTRICTION ANALISIS IN SEQUENCE : C2A1.SEQ
*******************************************

ENZYME      SEQUENCE       N* SITES  POSITIONS
******      ********       ********  *********

ALUI         AGCT                     NO  SITE   FOUND

AVAI         CYCGRG                   NO  SITE   FOUND

AVRII        CCTAGG            1      152

BAMHI        GGATCC            1      144

DDEI         CTNAG             2      125    212

HAEIII       GGCC              1      186

HGAI         GACGC             1       90

HGAI         GCGTC                    NO  SITE   FOUND

HINDII       GTYRAC            1       99

HPAII        CCGG              1      184

MBOI         GATC              1      145

MBOII        GAAGA             1        7

MBOII        TCTTC                    NO  SITE   FOUND

MNLI         CCTC              3       69    149    211

MNLI         GAGG                     NO  SITE   FOUND

XHOII        RGATCY            1      144
```

Fig. 3: Carp mitochondrial DNA region (C2A1.SEQ) obtained in edited form by list option and its restriction analysis.

ACKNOWLEDGEMENTS

We are grateful to Dr. Manuel Krauskopf, in whose laboratory this work was carried out, for his suggestions and encouragement. We also thank Nelson Parra and Raimundo Vega from Direccion de Computación e Informática (UACH) for assistance and helpful discussions. Thanks are also fue to Dr. Clifford Cox for excellent help in the preparation of this manuscript and to Mrs. M. Angélica Espinoza for secretarial assistance.

REFERENCES

1. MAXAM, A. and GILBERT, W. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74:* 560-564.
2. SANGER, F.; NICKLEN, S. and COULSON, A.R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74:* 5463-5467.
3. SOLL, D. and ROBERTS, R.J. (1982) eds. (Compilation of papers) in *Nucleic Acid Res. 10* (1).
4. STADEN, R. (1983) in *Laboratory Techniques in Biochemistry and Molecular Biology* (Work, T.S. and Burdon, R.H., eds.) Amsterdam, Elsevier Biomedical Press, Vol. 10, pp. 311-368.
5. ANDERSON, S.; BANKIER, A.T.; BARRELL, B.G.; DE BRUIJN, M.H.L.; COULSON, A.R.; DROUIN, J.S.; EPERON, I.C.; NIERLICH, D.P.; ROE, B.A.; SANGER, F.; SCHREIER, P.H.; SMITH, A.J.H.; STADEN, R. and YOUNG, I.G. (1981) *Nature 290,* 457-465.
6. ROBERTS, R.J. (1983) *Nucleic Acid Res. 11,* r135-r167.
7. CHALLBERG, M.D. and ENGLUND, P.T. (1980) in *Methods in Enzymology* (Grossman, L. and Moldave, K., eds.) New York, Academic Press, Vol. 65, pp. 39-43.
8. ROYCHOUNDHURRY, R. and WU, R. (1980) in *Methods in Enzymology* (Grossman, L. and Moldave, K., eds.) New York, Academic Press, Vol. 65, pp. 43-62.
9. LARSON, R. and MESSING, J. (1983) *DNA 2,* 31-35.