DNA sequence analysis: a procedure to find homologies among many sequences

Gopal Krishnan, Rajinder K.Kaul and Pudur Jagadeeswaran*

University of Illinois College of Medicine, Center for Genetics, 808 S.Wood St., Chicago, IL 60612, USA

Received 6 June 1985

ABSTRACT: SEQCMP, a program that analyzes and searches for homology among multiple nucleic acid sequences, is described. The sequences are compared by the dot matrix method and the consensus sequence is derived by superimposing all the dot matrices on one another. The program is written in MBASIC and runs on IBM-PC microcomputer. It is interactive and can be used by investigators with no computer background or experience.

INTRODUCTION

A number of computer programs for DNA sequence analysis and search for homology between nucleotide sequences have been reported in recent years (1-19). However, there is no program to compare more than two sequences of any length at the same time and obtain a consensus sequence. We have, therefore, developed a program called SEQCMP that uses the dot matrix method to compare multiple sequences of any length and derive a consensus sequence.

The program is designed to run on the IBM-PC or any other microcomputer that uses MS-DOS/PC-DOS operating system. The language is MBASIC, which is widely used in micro- and minicomputers. Very little memory is used and efficient programming minimizes running time. User friendly and interactive, the program can be used by investigators who have little or no computer experience.

Footnote: The program diskette will be available to any investigator for a nominal charge of \$25 to cover the costs of diskette and package and handling charges.

MATERIALS AND METHODS

The dot matrix method is a powerful tool for analyzing nucleic acid sequences. It provides information on the homology between DNA sequences and direct repeats within a sequence. This technique is particularly useful for analyzing long nucleic acid sequences. The principle employed to derive a consensus sequence for a known number of sequences is as follows. Each set of two sequences is compared, generating a dot matrix. When all the sequences have been compared with each other, the resulting dot matrices are superimposed on one another to identify the dot matrix they have in common, and, accordingly, the consensus sequence.

This process can pose a storage problem, because the memory capacity required for storage of dot matrices for sets of long sequences far exceeds that of a small computer. In the SEQCMP program, the sequences are compared, and the dot matrices generated are keyboarded into the diskette. This frees the computer memory from the task of retaining the matrices, and reduces the memory requirement to 64 kilobytes. If fixed-disk capacity of 20 to 40 megabytes is available, then approximately 10 sequences up to 2000 nucleotides long can be compared as a whole. Comparing small strips of the sequences at a time, rather than comparing the sequences as a whole obviates the need to use a great deal of floppy or fixed disk to keep all the dot matrices. Hence this method is followed in our SEQCMP program. The algorithm used is as follows. Each sequence is divided into strips of length of 100 bases or less. For example, to compare three sequences, each of length of 200 nucleotides, sequence 1 is divided into two strips denoted by 1 and 1a, each of 100 bases in length. Similarly, sequence 2 is divided into two strips 2 and 2a and sequence 3 is divided into two strips 3 and 3a. The individual strips thus obtained are compared and the resultant dot matrices are fused together, to obtain the consensus sequence. The exact methodology is as follows. The consensus sequence may be imagined to contain four matrices as shown in fig. 1. To obtain matrix (1,1), the strips 1 and 2, 1 and 3 and 2 and 3 are compared and the dot matrices thus generated are

superimposed on one another to identify the dot matrix they have in common. To obtain matrix (1,2), the strips 1a and 2, 1a and 3 and 2a and 3 are compared and the resultant dot matrices are superimposed on one another. In a similar manner, the strips 1 and 2a, 1 and 3a and 2 and 3a on comparison and superimposition of dot matrices generate the matrix (2,1). The strips 1a and 2a, 1a and 3a and 2a and 3a, are compared and the dot matrices obtained are superimposed to derive at the matrix (2,2).

The following steps have been taken to make the program execute fast:

- Checking for null matrix (no matches in the matrix): when superimposing one dot matrix on another, if the resultant matrix is null, then no more superimposition is done and the computer prints that the matrix is null. It then proceeds to compare the next set of strips.
- The SEQCMP program has been divided into five segments denoted by SEQCMP, SEQCMP1, SEQCMP2, SEQCMP3 and SEQCMP4. By doing so, only a segment of the whole program resides in the memory at any time. This results in faster execution of the program.

The program provides the following options:

- a. Generation of consensus sequence among many sequences, printed out as nucleotides or as dots;
- b. calculation and printout of the number and percentage of common characters found.

System Requirements

The hardware and software necessary for running the system are listed below.

Hardware

1. IBM-PC computer with 64 kilobytes of memory, or PC compatible computers such as COMPAQ, ZENITH, etc.
2. Two 5 1/4" double-sided, double-density floppy diskette drives.
3. One or two fixed-disk drives with a capacity of 20 megabytes (optional).
4. Monitor with monochrome display of 80 characters in 24 lines.
5. Micromodem (Hayes Smart Com II or Cross Talk).
6. Dot matrix or letter-quality printer.

Software

1. Microsoft BASIC interpreter or compiler.
2. Communication software for micro modem.
3. PC-DOS/MS-DOS operating system.
4. Word Star or any other word processor (optional).

The optional Micromodem, can be used to convert the IBM-PC to a remote terminal, and files can be transferred from the IBM-PC to another computer or vice versa.

A fixed-disk drive, if available, can be used as a backup for storage of dot matrices as disk files.

Input to the system

The sequences to be compared comprise the input to the system. The sequences can be created as data sets in the diskette by using either the text editor EDLIN of the MSDOS operating system or any other text editor, e.g., WORD STAR or SELECT.

Description of the Program

1. To start the program, insert the operating system disk in drive A, insert the program and input sequence data

```

      1      2      3      4      5
1234567890123456789012345678901234567890
1
2
3
4
5 aa a a a cc a a a c a a cc c c a a
6
7 g gg g g g g g g g g g g
8 g gg g g g g g g g g g g
9
10 aa a a a a a a a a a a a a
11 aa a a a a a a a a a a a a
12
13 cc cc cc cc cc cc cc cc cc cc
14 cc t t t t t t t t t t t t
15
16 g gg g g g g g g g g g g
17 aa a a a a a a a a a a a a
18 g gg g g g g g g g g g g
19
20 g gg g g g g g g g g g g
21 aa a a a a a a a a a a a a
22 g gg g g g g g g g g g g
23 aa a a a a a a a a a a a a
24
25 t t t t t t t t t t t t
26 g gg g g g g g g g g g g
27
28
29 t t t t t t t t t t t t
30 g gg g g g g g g g g g g
31 g gg g g g g g g g g g g
32 aa a a a a a a a a a a a a
33
34 g gg g g g g g g g g g g
35 aa a a a a a a a a a a a a
36
37
38
39
40 g gg t g g g g g g g g g t g
41 g gg g g g g g g g g g g
42
43
44
45
46
47 t t t t t t t t t t t t
48
49 g gg g g g g g g g g g g
50 aa a a a a a a a a a a a a

```

No of Common characters = 322

Sequences used for comparison

```

aagaaaggacacctgaaagagagtgcatggaagagacctgctcatacga - 1
aagcagggaaacctggagcgagagtgacctggaggagacctgctcactaga - 2
gttcaagggaaccttgagagagaatgtatggaagaaaagtgtagttttga - 3

```

Fig 2: Computer printout of the consensus sequence matrix. The sequences compared were the 5' coding regions of mature Human factor X (sequence 1), factor IX (sequence 2) and prothrombin (sequence 3). The diagonal represents the consensus sequence.

set disk in drive B, and turn on the computer. The computer displays " A>". Enter BASIC B:SEQCMP, then press carriage return (CR).

2. A brief description of the program is then displayed.
3. After all the displayed queries are answered correctly, the program will compare all the sequences and create the consensus sequence in the diskette.
4. After the processing is completed, the program inquires whether the investigator wishes to exit or continue

processing for another set of input sequences. If "n" is pressed, the system displays "END OF JOB" and comes to a halt. Otherwise, it proceeds to accept another set of input values.

Output from the System

The printer will print out the homology between input sequences as a two-dimensional array. A sample output of the consensus sequence among multiple sequences is shown in Fig 2. Fig 2 also lists the sequences used for this comparison. We used the 5' coding sequences of factor IX, X and Prothrombin for this purpose (20,21,22).

DISCUSSION

We have developed the program in MBASIC and used micro-computers because most laboratories have their own micro computers, connected to time-sharing systems through telephone modems, and these are much less expensive to operate than mainframe computers.

Using microcomputers, however, created two problems. First, they have limited memory and back-up storage capacity. Second, they are slow.

With regard to the first problem, in the case of the IBM-PC only about 8 kilobytes out of the total memory of 128 kilobytes is available for data storage. The sequence comparison program requires $(n+1) * l^2$ bytes of memory for storage of dot matrices, where n is the number and l is the length of sequences compared. For example, to compare three sequences, each thousand characters in length, a total of four thousand kilobytes of memory is required for data storage. This problem is solved in our SEQCMP program as described below. Instead of keeping the dot matrices in the memory, SEQCMP stores them as files in diskettes. Although this slows the program, it makes comparison of longer sequences possible. This, however, creates another problem. Since the total diskette storage capacity of the IBM-PC is about 320 kilobytes, additional storage capacity is required to compare longer or more number of sequences. A fixed-disk drive of 20

to 40 kilobytes capacity is necessary for comparing 10 sequences, each about a thousand bases in length. More storage capacity is required for comparing longer sequences.

Long sequences can be compared by analyzing only small strips of about 100 to 200 nucleotides at a time. This method does not require a great deal of storage capacity, and so we have used this method in our SEQCMP program. This brings us to the second problem created by the use of micro-computers: their slowness. Running time is not a big constraint, however, considering the relatively high cost of time-sharing systems. For most of the problems, an overnight or weekend run is sufficient. The program can be started by feeding the input values, and the investigator can look after other things, while the computer does the processing.

The SEQCMP program can be processed on a variety of micro- and minicomputers. For systems other than the ones described here, the program has to be slightly modified to reflect the differences between the versions of BASIC used.

To compare long sequences, it is advisable to keep the program and the sequence data sets in the A drive, which also contains the operating system programs. This will reserve drive B entirely for the diskette holding the consensus sequence, which the program generates. If a fixed-disk drive is used, the program and sequence data sets can be kept in the B drive and the fixed disk can be set aside entirely for storing the consensus sequence generated by the program.

This program can also be used to generate the inverted repeats. In this case the program will not search for matching nucleotides, but will search for complementary bases. The principles of this have been described before (1). The superimposition of matrices containing information of inverted repeats will give the common secondary structures. For example, multiple 5s RNA sequences and tRNA sequences can be used to generate common secondary structures.

Conclusion

Because no existing program could compare more than two DNA sequences of any length at a time to obtain a consensus sequence, the SEQCMP program was developed to compare multiple

sequences of any length. A microcomputer is used to translate multiple sequences into dot matrices, which are superimposed to obtain the consensus sequence. Continued use of this program has proven SEQCMP to be not only effective but minimal in its cost and memory capacity requirements. SEQCMP uses MBASIC, but may be translated into FORTRAN for mainframe processing when speedier results are needed.

ACKNOWLEDGEMENT

This work was supported by a NIH grant HL31458 and a grant from March of Dimes 5-459. We thank Linda Cardenas for preparing the manuscript.

*To whom correspondence should be sent

REFERENCES:

1. Jagadeeswaran, P. and McGuire, P.M (1982) Nucl. Acids Res. 10, 433-447.
2. Stone, T.W. and Potter, K.N. (1984) Nucl. Acids Res. 12, 367-378.
3. Larson, R. and Messing, J. (1982) Nucl. Acids Res. 10, 39-49.
4. Maizel, J.V., Jr., and Lenk, R.P. (1981) Proc. Natl. Acad. Sci. USA (78), 7665-7669.
5. Queen, C., Wegman, M.N. and Korn, L.J. (1982) Nucl. Acids Res. 10, 449-456.
6. Mount, D.W. and Conrad, B. (1984) Nucl. Acids Res. 12, 811-817.
7. Needleman, S.B. and Wunsch, C.D. (1970) J. Mol. Biol. 48, 443-453.
8. Collins, J.F. and Coulson, A.F.W. (1984) Nucl. Acids Res. 12, 181-192.
9. Novotny, J. (1982) Nucl. Acids Res. 10, 127-131.
10. Kroger, M. and Kroger-Block, A. (1984) Nucl. Acids Res. 12, 193-201.
11. Kanehisa, M. (1984) Nucl. Acids Res. 12, 203-213.
12. Staden, R. (1977) Nucl. Acids Res. 4, 4037-4051.
13. Staden, R. (1978) Nucl. Acids Res. 5, 1013-1015.
14. Staden, R. (1979) Nucl. acids Res., 6, 2601-2610.
15. Staden, R. (1980) Nucl. Acids Res., 8, 817-825.
16. Korn, L.J., Queen, W. (1977) Proc. Natl. Acad. Sci. USA 74, 4401-4405.
17. Zweig, S.E. (1984) Nucl. Acids Res. 12, 767-776.
18. Gingeras, T.K. and Roberts, R.J. (1980) Science 209, 1322-1325.
19. Gingeras, T.R., Milazzo, J.P., Sciaky, D. and Roberts, R.J. (1979) Nucl. Acids Res. 7, 529-545.
20. Anson, D.S., Choo, K.H., Rees, D.J.G., Giannelli, F., Gould, K., Huddleston, J.A. and Brownlee, G.G. (1984) The EMBO J. 3, 1053-1060.
21. Kaul, R.K., Hildebrand, B., Roberts, S. and Jagadeeswarn, P. (1985) Gene (in press).
22. Degen, S.J., Friezner, M.R.T.A. and Davie, E.W. (1983) Biochemistry 22, 2087-2097.