Volume 8 Number 16 1980 Nucleic Acids Research

A new computer method for the storage and manipulation of DNA gel reading data

R.Staden

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received 17 June 1980

ABSTRACT

This paper describes a new way of storing DNA gel reading data and an accompanying set of computer programs. These programs will perform all the manipulations that are required on data gained by the so-called 'shotgun' method of DNA sequencing. This system simplifies the computer processing involved with this sequencing method and also has the capability of being able at any time during a project to display, lined up in register, all the gel readings covering any section of the sequence.

INTRODUCTION

The shotgun method of DNA sequencing involves cutting the DNA into random fragments and then determining the sequence of each of these individual pieces. The relationship of the pieces is established by finding overlaps between their sequences. Eventually a sufficient number of the pieces to cover the whole of the DNA will be sequenced and joined.

In 1979 we described [1] computer programs that could be used to find overlaps between gel readings and to join sequences together to form one longer sequence. (Similar programs were described by Gingeras et al. [2]). This system of programs was used successfully during a number of large sequencing projects but it became apparent that it was also necessary to keep complicated notes in order to keep track of the origins of the evidence for each character in a sequence.. This is because after a sequencing project has been under way for some time each section of the sequence will have been determined from a number of gels and may have been revised and edited many times. It becomes increasingly more time-consuming to check the data.

The quality of a sequence can be assessed at two levels: firstly at the level of the individual nucleotide assignments as read from each gel; and secondly from the number of times each character has been given a particular assignment on different gels and whether or not it has been sequenced on both strands of the DNA. In the 1979 paper [1] we described a method of coding

for uncertainties in gel readings and this accounted for the quality of the data at the first level. In this paper we describe a computer storage and processing system that can give complete information about each character in a sequence and hence give easy assessment of the quality of data at both levels. With this system we can interrogate the computer at any stage of a project to find out for each character which gels it has been determined from, how many of the gels agree and which strand of the DNA they are from. Most importantly we can get the computer to print out all the gels covering any area of sequence lined up in register one above the other with a consensus sequence below (see Fig. 8). Having this information readily available can give quick assessment of the quality of the data, indicate problem areas and hence show which further experiments are necessary.

Description of the problem

In order to be able to have complete information about each character in a sequence we have to know how each of the gels that overlap to form any contiguous section of sequence relate to one another. We can determine these relationships initially by using a computer program to look for overlaps but the difficulty lies in being able to remember them throughout the many complicated types of manipulations that the sequence data will undergo during a project. The following are some of these manipulations: (1) calculation of a consensus sequence that can be used for finding overlaps with new data or for any other kind of analysis, i.e. all gels must be lined up and added to give a consensus; (2) addition of new gels into the system (these new gels may add new data to left or right ends of contiguous sections of sequence or simply confirm existing data by overlapping internally); (3) joining of two existing contiguous sections of sequence (a new gel may overlap with two previously unrelated sections of contiguous sequence which then need to be fused to form one contiguous section); (4) complementing of any section of contiguous data (in order that two sections of data can be joined one of them may require to have its sense reversed so that both sections are in the same sense); (5) editing of individual gel readings; (6) editing of contiguous sections of sequence; (7) examination by display of all gel readings covering any particular section of the sequence.

There are two ways of handling data of this type: one can store all the related gel readings in one large file, lined up in register one above the other (i.e. store them as they appear in Fig. 8) or one can store the individual gel readings in separate files and use an extra file to contain all the information that is required to assemble the sequences into the

correct relative positions during processing. With the first method the relationships of the data are actually contained in the structure of the file and so the file structure and relationships will change simultaneously if gel readings are altered. With the second method, if gel readings are changed then the information about their relationships (contained in a separate file) will also need to be changed.

We have chosen to use the second method for reasons of economy of disk storage and speed of processing, and also because our computer (a PDP-LSI-11) has only 32 K words of memory. Below we define the types of information required to assemble gel readings and describe how it varies during a sequencing project.

Manipulations of types 2 to 6 will change the relationships of gel readings, and manipulations 4, 5 and 6 change the gel readings themselves. Manipulations of types 1 and 7 require a knowledge of the relationships between gel readings but do not change them.

In order to perform these manipulations the following types of information are required: (1) sequences from gel readings; (2) facts about individual gel readings and their relationship (if any) to others (only those gel that overlap to form part of a contiguous section of sequence are related to others); (3) facts about individual contiguous sections of sequence; (4) general facts about the numbers of gels and numbers of contiguous sections of sequence. The amounts of these types of information vary during the course of any sequencing project, as shown in Figure 1.

The system has to be able to perform all the above types of manipulation, and efficiently allow for variation in the amounts of data.

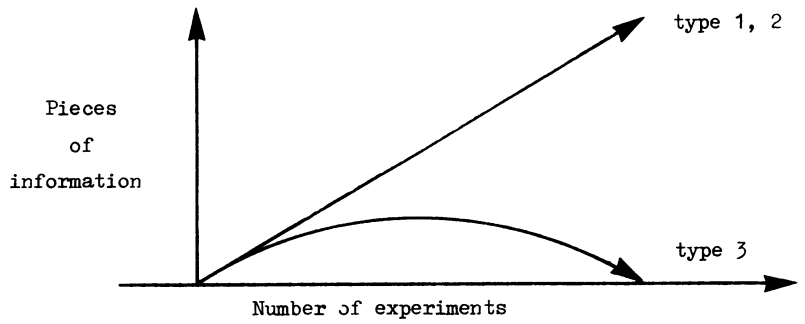


Fig. 1

Definition of a contig

In order to make it easier to talk about our data gained by the shotgun method of sequencing we have invented the word "contig". A contig is a set of gel readings that are related to one another by overlap of their sequences. All gel readings belong to one and only one contig, and each contig contains at least one gel reading. The gel readings in a contig can be summed to form a contiguous consensus sequence and the length of this sequence is the length of the contig.

DESCRIPTION OF THE SYSTEM

The description of our system is divided into two sections. The first deals with data storage and defines the types of files used and the information they contain. The second describes the programs that operate on these files and is concluded by an example of a run of the main program.

Data storage

In the system we have devised there are four types of file: (1) the original sequence from each gel - this is an archive version which cannot be overwritten; (2) a working version of each gel reading sequence which can be manipulated and edited by program; (3) a consensus sequence. This is calculated by program and is in a form suitable for input to any of the previously described programs [3-5]; (4) a master plan or database for each sequencing project. This contains all the information of types 2, 3 and 4, i.e. all the relational information about the gels. Any manipulations on the data use this file and it is automatically updated any time the relationships are changed. The actual information stored in this file is as follows:-

(1) Facts about each gel and its relationship to others:
   (a) the name of the file containing the gel reading sequence;
   (b) the length of this gel reading sequence;
   (c) the position of the left end of this gel relative to the left end of the contiguous section of sequence of which it forms a part;
   (d) the name of the next gel file to the left;
   (e) the name of the next gel file to the right;
   (f) the strandedness of this gel, i.e. which strand of the DNA it is.

(2) Facts about each contig:
   (a) the length of this contig;
   (b) the name of the gel at the left end of this contig;

(c) the name of the gel at the right end of this contig.

(3) General facts:

(a) the number of gels in the system;

(b) the number of separate contigs in the system.

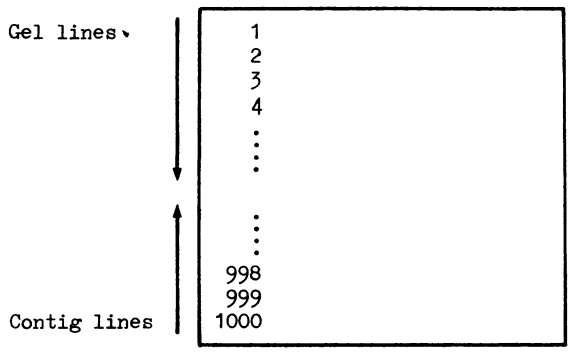
In order to use storage space efficiently this file has the following structure. The file is divided into 1000 lines of information; the general facts are stored on line 1000; facts about gels are stored from line 1 forwards and facts about contigs are stored backwards from line 999. This is shown in Figure 2.

As each new gel is introduced into the system a new gel line is added onto the end of the list of gels. If this new gel does not overlap with any contig already in the system a new contig descriptor line is added to the list. If it overlaps with one existing contig no extra lines need be added, but if it overlaps with two contigs, then these need joining and the number of lines describing contigs will be reduced by one. This list is then compressed to leave the empty line at the top.

Initially the two types of lines will move towards one another in the file but eventually, as the contigs are joined, the contig descriptor lines will move in the same direction as the gel lines. At the end of any project there will, of course, only be one contig descriptor line so with this size and structure the system is capable of handling a sequencing project requiring 998 gel readings.

The programs

The structure and content of the systems files need not normally concern the sequencer as he extracts, uses and alters the data via a few easily used programs. There are four regularly used programs: (1) CONSEN is the



There is one line used for each gel descriptor line and one for each contig descriptor line

Fig. 2

consensus calculation program. It will calculate a consensus sequence for all or selected regions of the database. This consensus sequence is in a form suitable for input to any of the previously published programs [3-5] but it is mainly used for searching for overlaps with new gel readings: (2) BATIN is a program that is used to store the archive versions of new gel readings in the computer. It will take any number of new gel readings and store each of them in separate files. It also stores the name of each of these files in a further file so that all these new gels can be compared with the consensus sequence; (3) DBCOMP is the program that searches for overlaps between sequences. It uses the file of file names set up by BATIN and compares each of the new gels (and their complementary sequences) with the consensus sequence and then each of the gels with each of the other new gels. If it finds an overlap it reports its length, the section of the database involved and displays the match. It will also compare each of the new gels with any other sequences - for example, the sequence of the cloning vector; and (4) DBUTIL. This is the most important and often used program as it performs all the manipulations which can change the data in the database. It is an interactive program and is best operated from a visual display unit with a keyboard. All options are prompted for and all operator input checked for errors. There are six functions available from the main program, each of which then contains its own options. The functions are as follows:-

(1) PRINT will print the contents of the database.

(2) DISPLAY will display on the printer or screen all the gels covering any region of the sequence, lined up in register showing their gel numbers, their strandedness and, underneath, their consensus.

(3) EDIT will allow editing of any contiguous section of sequence. It maintains alignments by making the same number of insertions or deletions in all the gels covering the edit position.

(4) COMPLEMENT will complement and reverse all the gels covering any contiguous section of sequence. It automatically reverses and complements each gel sequence, reorders left and right neighbours, recalculates relative positions and changes each strandedness.

(5) JOIN will allow joining of any two contiguous sections of sequence. The options available are:
   (a) movement of join;
   (b) editing of either the left or the right contig;
   (c) display of the overlap between the two sections of sequence.

(6) ENTER allows entry of a new gel into the database system. It is

necessary to know beforehand if the new gel overlaps with any sequences already in the database. It automatically names and writes the working version of the new gel and then offers the following options:-

(a) complementing and reversing of the new gel (to add it to existing data it may need its sense changing);

(b) adding to an existing contig either at a left or right end or internally;

(c) movement of overlap;

(d) display of overlap;

(e) editing of the new gel. (It is only during entry into the system that gels can be edited individually. Once entry is completed all gels covering any position are edited together to maintain the alignments of the sequences.)

(f) editing of the contig.

A basic requirement is that the data must be correctly aligned by the operator when he enters it into the system. If the gels do not line up the operator can use the editing functions to improve their alignment and this can always be achieved by making only insertions in the sequences. In this way no data need ever be deleted until sufficient agreement is found between the gels covering every sequence position. We use X or spaces as padding characters to achieve alignment so that problem areas stand out clearly in the lined up sequences.

Description of the example of program DBUTIL

Figures 3-8 show a sample run of program DBUTIL in which two new gel readings are added to the database. These are simply photographs of the computer output for a typical run of the program, except that all the typing done by the operator has been underlined to make explanation easier. Before processing of the data reached this stage both of these new gel readings were stored on the computer using program BATIN and then they were compared with all of the data already in the database using program DBCOMP. DBCOMP reported that one gel did not overlap at all and that the other overlapped with two separate contigs. The overlap was complicated by the fact that the gel overlapped with one contig in one sense but in the opposite sense with the other. This means that before these two contigs can be joined by this new gel the sense of one must be changed. All these operations are shown in the example.

```
RU DBUTIL

   DBUTIL

    PROJECT NAME =LAMBDA.RS1

   NUMBER OF GELS=   280  NUMBER OF CONTIGS=    66
   FOR 120 CHARACTER LINES TYPE Y
Y

   SELECT OPTION BY NUMBER
   STOP=0,ENTER=1,PRINT=2,DISPLAY=3,JOIN=4,COMPLEMENT=5,EDIT=6
    OPTION NUMBER = 1

   ENTER

    NAME OF ARCHIVE = LAD99

   WORKING NAME FOR THIS GEL = LAMBDA.281
   IF THIS GEL OVERLAPS TYPE Y


   SELECT OPTION BY NUMBER
   STOP=0,ENTER=1,PRINT=2,DISPLAY=3,JOIN=4,COMPLEMENT=5,EDIT=6
    OPTION NUMBER = 1

   ENTER

    NAME OF ARCHIVE = LAD14

   WORKING NAME FOR THIS GEL = LAMBDA.282
   IF THIS GEL OVERLAPS TYPE Y
Y

   TO REVERSE AND COMPLEMENT THIS GEL TYPE Y
Y

    NUMBER OF LEFT GEL THIS CONTIG =4
```

```
    IF THIS NEW GEL EXTENDS THE CONTIG LEFTWARDS TYPE Y
Y

     POSITION IN NEW GEL OF LEFT CHARACTER IN OLD CONTIG =156

              10.       20.       30.       40.       50.       60.       70.       80.       90.      100.      110.      120.
    4  GATCCCTTCTGATACTGTCATCAGCATTACGTCATCCACCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACAXTCCGXCAAHCCCT4CC
   -1  GATCCCTTCTGATACTGTCATCAGCATTACGTCATCCTCCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACACTCCGGCAAGCCCTGCC
    2                                  CATCCTCCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACACXXCGGCAAGCXCTGCC
       GATCCCTTCTGATACTGTCATCAGCATTACGTCATCC-CCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACACTCCGGCAAGCCCTGCC
       GATCCCTTCTGATACTGTCATCAGCATTAACGTCATCCTCCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACAACCTCCGGCAAGCCCT
                                  ******* ** ******** ** ***** *   * ***   *** ** *** *  *** *  * *  *** *****  ** ******** *
             165.      175.      185.      195.      205.      215.      225.      235.      245.      255.      265.      275.
    IF JOINT CORRECT TYPE Y
Y

    SELECT OPTION BY NUMBER
    EDIT NEW GEL=1,DISPLAY=2,COMPLETE ENTRY=3,GIVE UP =4,EDIT CONTIG=5
     OPTION NUMBER = 5

    INSERT=1,DELETE=2,CHANGE=3,RETURN=4
     OPTION NUMBER = 1

     POSITION =30

     NUMBER OF CHARS =1

     CHARS TO INSERT INTO GEL    4 =

     CHARS TO INSERT INTO GEL    1 =

    INSERT=1,DELETE=2,CHANGE=3,RETURN=4
     OPTION NUMBER = 4

    SELECT OPTION BY NUMBER
    EDIT NEW GEL=1,DISPLAY=2,COMPLETE ENTRY=3,GIVE UP =4,EDIT CONTIG=5
     OPTION NUMBER = 2

    DEFINE REGION
     RELATIVE POSITION OF LEFT END=

     RELATIVE POSITION OF RIGHT END=
```

Fig 3

Nucleic Acids Research

In Figure 3 the operator has started the program and supplied the project name which tells the program which database to use. The program reports that so far there are 280 gel readings and 66 contigs already in the database. The operator has selected to have 120 character line printout and then the program has asked him to select from six functions. The first thing the operator wants to do is to add the new gel reading that does not overlap to the database. He therefore selects the ENTER option and the program responds by requesting the name of the file containing the gel reading. It then copies this gel reading (which is kept as an archive) into another file that is used as a working version of this gel. This is the 281st gel reading to be entered into the database for this project and so the working file is automatically named LAMBDA.281. The program then asks if this gel overlaps and the operator signifies "no" by typing only "carriage return". This is all the operator input required when entering a non-overlapping gel reading: all the data needed for the database will be set up by the program. It will have written a gel descriptor on line 281 and a new contig descriptor on line 999-66 = 933.

As the operator input is complete the program leaves the ENTER option and again offers the main options. The operator now wants to enter the gel that does overlap so first he needs to add it to one of the contigs using ENTER and then join the two contigs using JOIN. He therefore goes through the same process as before but this time he tells the program the gel does overlap and so the program gives him further options to allow him to define the overlap and align the new gel. The first option he is offered is to reverse and complement the gel reading sequence which he does by typing "Y". This means that the working version of this gel will now be in the opposite sense to the original gel. Then the program needs to know how the gel overlaps and so it asks if the new gel extends the contig leftwards (the program needs to fix either the position of the left end of the gel relative to the left end of the contig or the position of the left end of the contig relative to the left end of the gel and so it separates overlaps into two classes - those that extend the contig leftwards and those that do not, i.e. those that either overlap wholly internally or extend the contig rightwards). This gel reading does extend the contig leftwards and so the operator types Y. The program therefore asks him for the position in the new gel of the left character in the contig. The operator knows this from the previous run of program DBCOMP so he types the position and the program displays the join.

The display consists of: on the top line the relative numbers for the

contig characters and then on succeeding lines the sequences from each of the gels that are in this part of the contig. At the left end of each sequence is its gel number and the sign of this number shows whether or not a gel reading is in the same sense as the original gel. A minus sign indicates the opposite sense and so it can be seen at a glance if a section has been determined on both strands: if two gels of opposite sign cover a section it has been determined on both strands. On the next line down is shown the consensus sequence for the gels lined up above and then below that is the sequence of the new gel. Below that asterisks show mismatches between the new gel and the consensus and then the new gel numbers are written.

The program then asks if the joint is correct which refers to the positioning of the left end. It is, so the operator types Y and the program then offers him five options to help completion of the entry of the gel. If the joint is incorrect the operator types only carriage return and the program asks him to redefine the join. It is clear from the display that some editing is required to improve the alignment of the new gel with those already in this contig.

Alignment can be achieved by making insertions or deletions but to delete data would require checking gels so we usually make only insertions until such time as sufficient of the gels agree. The new gel has an extra base at position 30 in the contig and so the operator has selected to insert one character at this point. The program then prompts him for the characters to insert in each of the gels that cover this position. If the operator types only carriage return as he has done in this case, the program automatically inserts the requested number of spaces. The operator has then displayed the whole of the overlapping region (this appears in Fig. 4) and then made further insertions in the contig at position 107. After displaying again he has selected to edit the new gel which he does using a different type of editing program which allows multiple edits to be made at once. A further display shows the alignment to be good and so the operator has selected the "complete entry" option. It is only at this stage that the database on the disk is altered which is why we can also have a "give-up" option which will leave the database unchanged if a problem arises.

The operator now needs to change the sense of all the gels in one contig so that it can be joined to another. In Figure 5 the operator has first selected the PRINT option and asked the program to print out the contents of the database for the contig he wishes to complement. Contigs are always referenced by the name of their leftmost gel reading. The program first

```
          10.       20.       30.       40.       50.       60.       70.       80.       90.      100.      110.      120.
 4  GATCCCTTCTGATACTGTCATCAGCATTA CGTCATCCACCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACAXTCCGXCAAHCCCT4C
-1  GATCCCTTCTGATACTGTCATCAGCATTA CGTCATCCTCCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACACTCCGGCAAGCCCTGC
 2                               CATCCTCCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACACXXCGGCAAGCXCTGC
    GATCCCTTCTGATACTGTCATCAGCATTA-CGTCATCC--CCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACACTCCGGCAAGCCCTGC
    GATCCCTTCTGATACTGTCATCAGCATTAACGTCATCCTCCGTCATGTCCGCCACATCCGGGGAAGCGGGGATAACTTCATTCCCGTCCGGGCCAAAGCGGACAACCTCCGGCAAGCCCT
                                 *         *                                                                  ** ********* ***
          165.      175.      185.      195.      205.      215.      225.      235.      245.      255.      265.      275.

          130.      140.      150.      160.      170.      180.      190.      200.      210.      220.      230.      240.
 4  C4CTTTCTXCATCAGCACATXCATCTTCAGGCTCTTCGTCCAGCCTXG-GCCGGTTCAGCAGACTGAAATCXAGCGGAGTCATATXDGGATC
-1  CGCTTTCTGCATCAGCACAT CATCTTCAGGCTCTTCG TCAGCCTCGCGCCGGTTCAGCAGACTGAAATCCAGCGGATGCATATCCGGATC
 2  CGCTTTCTGCATCAGCACATTCATCTTCAGGCTCTTGC TCAGCCTCGCGCCGGTTCAGCAGACTGAAATCCAGCGGATGCATATCCGGATCGCTGAAAAACAGGCTGAGCACGGTGTAC
    CGCTTTCTGCATCAGCACATTCATCTTCAGGCTCTT--T-CAGCCTCGCGCCGGTTCAGCAGACTGAAATCCAGCGGA--CATATCCGGATCGCTGAAAAACAGGCTGAGCACGGTGTAC
    GCCGCTTTCTGCATCAGCACATCATCTTCAGGCTCTTCGTCAGCCTCGCG
    ** ** * *********  ******* *** **** ****          ************************************************************************
          285.      295.      305.      315.      325.      335.      345.      355.      365.      375.      385.      395.
SELECT OPTION BY NUMBER
EDIT NEW GEL=1,DISPLAY=2,COMPLETE ENTRY=3,GIVE UP =4,EDIT CONTIG=5
 OPTION NUMBER = 5

INSERT=1,DELETE=2,CHANGE=3,RETURN=4
 OPTION NUMBER = 1

 POSITION =107

 NUMBER OF CHARS =2

 CHARS TO INSERT INTO GEL    4 =

 CHARS TO INSERT INTO GEL    1 =

 CHARS TO INSERT INTO GEL    2 =

INSERT=1,DELETE=2,CHANGE=3,RETURN=4
 OPTION NUMBER = 4

SELECT OPTION BY NUMBER
EDIT NEW GEL=1,DISPLAY=2,COMPLETE ENTRY=3,GIVE UP =4,EDIT CONTIG=5
 OPTION NUMBER = 2
```

```
    DEFINE REGION
     RELATIVE POSITION OF LEFT END=100

     RELATIVE POSITION OF RIGHT END=170

              109.      119.      129.      139.      149.      159.      169.      179.      189.      199.      209.      219.
   4  GGACAXT  CCGXCAAHCCCT4CC4CTTTCTXCATCAGCACATXCATCTTCAGGCTCTTCGTCCAGCCTXG
  -1  GGACACT  CCGGCAAGCCCTGCCGCTTTCTGCATCAGCACAT CATCTTCAGGCTCTTCG TCAGCCTCG
   2  GGACACX  XCGGCAAGCXCTGCCGCTTTCTGCATCAGCACATTCATCTTCAGGCTCTTGC TCAGCCTCG
      GGACACT--CCGGCAAGCCCTGCCGCTTTCTGCATCAGCACATTCATCTTCAGGCTCTT--T-CAGCCTCG
      GGACAACCTCCGGCAAGCCCTGCCGCTTTCTGCATCAGCACATCATCTTCAGGCTCTTCGTCAGCCTCGCG
           ****                                 ***** *** **** ********* *
              264.      274.      284.      294.      304.      314.      324.      334.      344.      354.      364.      374.
    SELECT OPTION BY NUMBER
    EDIT NEW GEL=1,DISPLAY=2,COMPLETE ENTRY=3,GIVE UP =4,EDIT CONTIG=5
     OPTION NUMBER = 1

         TYPE EDITS NOW

/F/298/I/X/F/316/I/X//0

    SELECT OPTION BY NUMBER
    EDIT NEW GEL=1,DISPLAY=2,COMPLETE ENTRY=3,GIVE UP =4,EDIT CONTIG=5
     OPTION NUMBER = 2

    DEFINE REGION
     RELATIVE POSITION OF LEFT END=100

     RELATIVE POSITION OF RIGHT END=___

              109.      119.      129.      139.      149.      159.      169.      179.      189.      199.      209.      219.
   4  GGACAXT  CCGXCAAHCCCT4CC4CTTTCTXCATCAGCACATXCATCTTCAGGCTCTTCGTCCAGCCTXG-GCCGGTTCAGCAGACTGAAATCXAGCGGAGTCATATXDGGATC
  -1  GGACACT  CCGGCAAGCCCTGCCGCTTTCTGCATCAGCACAT CATCTTCAGGCTCTTCG TCAGCCTCGCGCCGGTTCAGCAGACTGAAATCCAGCGGATGCATATCCGGATC
   2  GGACACX  XCGGCAAGCXCTGCCGCTTTCTGCATCAGCACATTCATCTTCAGGCTCTTGC TCAGCCTCGCGCCGGTTCAGCAGACTGAAATCCAGCGGATGCATATCCGGATCGCTGA
      GGACACT--CCGGCAAGCCCTGCCGCTTTCTGCATCAGCACATTCATCTTCAGGCTCTT--T-CAGCCTCGCGCCGGTTCAGCAGACTGAAATCCAGCGGA--CATATCCGGATCGCTGA
      GGACAACCTCCGGCAAGCCCTGCCGCTTTCTGCATCAGCACATXCATCTTCAGGCTCTTCGTXCAGCCTCGCG
           ****                                 *                  ** *          **********************************************
              264.      274.      284.      294.      304.      314.      324.      334.      344.      354.      364.      374.
    SELECT OPTION BY NUMBER
    EDIT NEW GEL=1,DISPLAY=2,COMPLETE ENTRY=3,GIVE UP =4,EDIT CONTIG=5
     OPTION NUMBER = 3
```

Fig 4

```
SELECT OPTION BY NUMBER
STOP=0,ENTER=1,PRINT=2,DISPLAY=3,JOIN=4,COMPLEMENT=5,EDIT=6
 OPTION NUMBER = 2

NUMBER OF GELS=   282  NUMBER OF CONTIGS=    67
TO SELECT CONTIGS TYPE Y
Y

 NUMBER OF LEFT GEL THIS CONTIG =282

CONTIG LINES

 999        661.        0     282     232
DEFINE REGION
 RELATIVE POSITION OF LEFT END=___

 RELATIVE POSITION OF RIGHT END=___

GEL LINES

LAD14        282          1.    -327       0       4
LAMA9          4        156.     214     282       1
LAD9           1        156.    -214       4       2
MLD5           2        189.     331       1     232
LAR16        232        452.     210       2       0
TO SELECT ANOTHER CONTIG TYPE Y
—

SELECT OPTION BY NUMBER
STOP=0,ENTER=1,PRINT=2,DISPLAY=3,JOIN=4,COMPLEMENT=5,EDIT=6
 OPTION NUMBER = 5

 NUMBER OF LEFT GEL THIS CONTIG =282

COMPLEMENT

SELECT OPTION BY NUMBER
STOP=0,ENTER=1,PRINT=2,DISPLAY=3,JOIN=4,COMPLEMENT=5,EDIT=6
 OPTION NUMBER = 2

NUMBER OF GELS=   282  NUMBER OF CONTIGS=    67
TO SELECT CONTIGS TYPE Y
Y

 NUMBER OF LEFT GEL THIS CONTIG =282

THIS IS NOT A LEFT END, RELATIVE POSITION =      335.
TRUE LEFT END =   232
TO CONTINUE TYPE Y
Y

CONTIG LINES

 999        661.        0     232     282
DEFINE REGION
 RELATIVE POSITION OF LEFT END=___

 RELATIVE POSITION OF RIGHT END=___

GEL LINES

LAR16        232          1.    -210       0       2
MLD5           2        143.    -331     232       1
LAD9           1        293.     214       2       4
LAMA9          4        293.    -214       1     282
LAD14        282        335.     327       4       0
TO SELECT ANOTHER CONTIG TYPE Y
```

Fig 5

prints the contig descriptor line which shows that this line is on 999 of the database, that the contig is 661 characters long, that the left gel is gel 282 and the right gel is 232. It then asks the operator which region of the contig he wants to have printed and by typing only carriage return the operator signifies he wants to see it all. The program then prints the gel descriptor lines. Each line describes one gel and the lines are printed in the order in which they overlap in the contig. For example, the first line (which is the gel just entered) says that the gel whose archive name is LAD14 is gel number 282, that its left end is at position 1 in the contig; the gel is 327 characters long, the minus sign indicates that it is in the opposite sense to its archive; it has no gel to its left and gel number 4 is to its right. The next line describes the gel whose archive is LAMA9, whose number is 4; has its left end at position 156 in the contig; is 214 characters long; its left neighbour is gel 282 and right neighbour is gel 1. Of course, the last gel in the list LAR16 has no right neighbour. In order to reverse the sense of a contig the operator has selected the COMPLEMENT function. The only operator input required by this function is the name of the contig to complement and this is the number of the contig's left gel which is typed by the operator. The program then reverses and complements the working version of each gel in the contig, reorders the database and changes the strandedness of each gel. The operator has then selected the PRINT option to demonstrate the effect of this function. He does not know which gel is now at the left end of the contig but as long as he knows the number of any gel in the contig the program will find the true left end. The printout shows the effect of the complement function.

The two contigs to be joined are now in the same sense and so in Figure 6 the operator has selected the JOIN option. He has been asked to define by their left gels the two contigs to join and their relative position. The program then displays the join in much the same way as for ENTER. This consists of all the gels overlapping the right end of the left contig (in this case only gel 282) with their consensus below and then, un-numbered the leftmost gel of the right contig. Again mismatches are denoted by asterisks. Although the program only displays the leftmost gel from the right contig any edits made to this contig will again be performed on all gels. This can be seen in Figure 6 when the operator has selected option 3 (edit right contig) and inserted at positions 56 and 53: the program has prompted for characters for both gels 5 and 217 and therefore maintained the alignment. In Figure 7 the operator has completed the editing required to give a well-aligned join

```
 SELECT OPTION BY NUMBER
 STOP=0,ENTER=1,PRINT=2,DISPLAY=3,JOIN=4,COMPLEMENT=5,EDIT=6
  OPTION NUMBER = 4

 LEFT CONTIG
  NUMBER OF LEFT GEL THIS CONTIG =232

 RIGHT CONTIG
  NUMBER OF LEFT GEL THIS CONTIG =5

 JOIN

RELATIVE POSITION IN LEFT CONTIG OF LEFT CHAR OF RIGHT CONTIG =596

          605.      615.      625.      635.      645.      655.      665.      675.      685.      695.      705.      715.
282  GAGTCAGGCGTCATTTTT1TGGTACGGAAAGTGATGCGAAAAABCAGCGGCCAGTTVCGTVGAACAG
     GAGTCAGGCGTCATTTTTCTGGTACGGAAAGTGATGCGAAAAACAGCGGCCAGTTTCGTTGAACAG
     GAGTCAG4CGT1ATTTTTCTGGTACGGAAAGTGATGCGAARAAAACAGCGGCAGTCGTTGAACAGTC0CTGAGCCGACAGGCT6GCTGCACA6AAAGCHGG A2TTCCGTC0GGCA2TAT
           *   *                          *  * ** ****** *** ** ****
            10.       20.       30.       40.       50.       60.       70.       80.       90.      100.      110.      120.
 SELECT OPTION BY NUMBER
 MOVE JOIN=1,EDIT LEFT CONTIG=2,EDIT RIGHT CONTIG=3,DISPLAY=4,COMPLETE JOIN=5,GIVE UP=6

  OPTION NUMBER = 2

 INSERT=1,DELETE=2,CHANGE=3,RETURN=4
  OPTION NUMBER = 1

  POSITION = 639

  NUMBER OF CHARS =2

  CHARS TO INSERT INTO GEL  282 =__

 INSERT=1,DELETE=2,CHANGE=3,RETURN=4
  OPTION NUMBER = 4

 SELECT OPTION BY NUMBER
 MOVE JOIN=1,EDIT LEFT CONTIG=2,EDIT RIGHT CONTIG=3,DISPLAY=4,COMPLETE JOIN=5,GIVE UP=6
```

```
 OPTION NUMBER =4
DEFINE REGION
 RELATIVE POSITION OF LEFT END=596

 RELATIVE POSITION OF RIGHT END=

          605.      615.      625.      635.      645.      655.      665.      675.      685.      695.      705.      715.
282  GAGTCAGGCGTCATTTTTTTGGTACGGAAAGTGATGCGAAAAB  CAGCGGCCAGTTVCGTVGAACAG
     GAGTCAGGCGTCATTTTTCTGGTACGGAAAGTGATGCGAAAAA--CAGCGGCCAGTTTCGTTGAACAG
     GAGTCAG4CGT1ATTTTTCTGGTACGGAAAGTGATGCGAARAAAACAGCGGCAGTCGTTGAACAGTCG
           *   *                                 *  **       ***** * *** ***
            10.       20.       30.       40.       50.       60.       70.       80.       90.       100.      110.      120.
SELECT OPTION BY NUMBER
MOVE JOIN=1,EDIT LEFT CONTIG=2,EDIT RIGHT CONTIG=3,DISPLAY=4,COMPLETE JOIN=5,GIVE UP=6

 OPTION NUMBER =3

INSERT=1,DELETE=2,CHANGE=3,RETURN=4
 OPTION NUMBER = 1

 POSITION =56

 NUMBER OF CHARS =2

 CHARS TO INSERT INTO GEL    5 =

 CHARS TO INSERT INTO GEL  217 =

INSERT=1,DELETE=2,CHANGE=3,RETURN=4
 OPTION NUMBER = 1

 POSITION =53

 NUMBER OF CHARS =1

 CHARS TO INSERT INTO GEL    5 =

 CHARS TO INSERT INTO GEL  217 =

INSERT=1,DELETE=2,CHANGE=3,RETURN=4
 OPTION NUMBER = 4
```

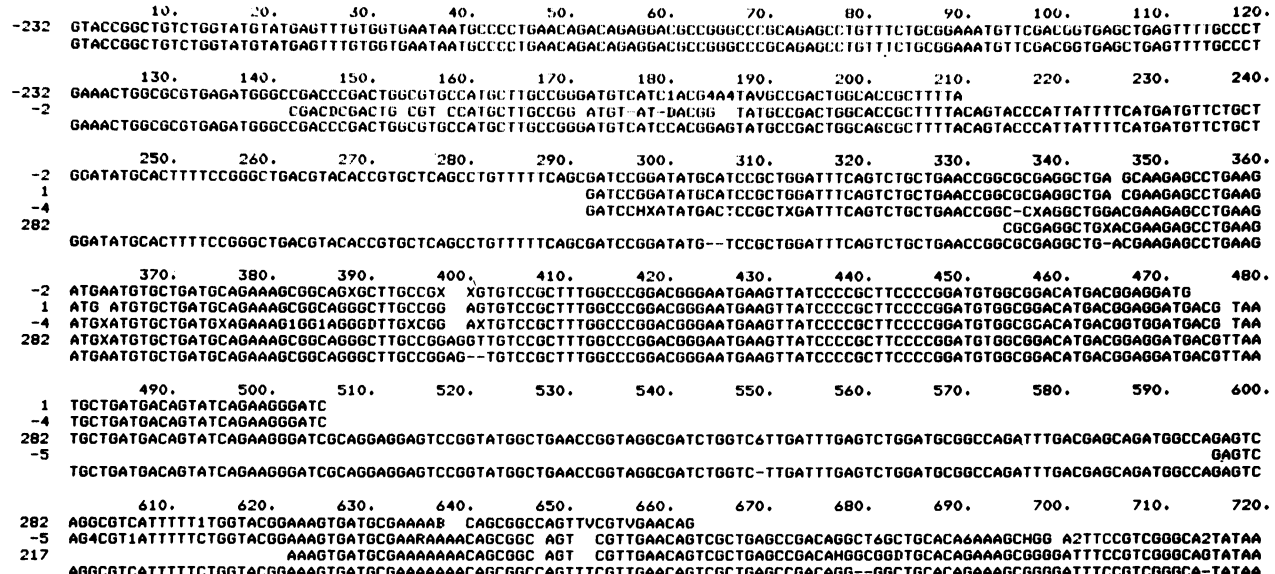
Fig 6

Nucleic Acids Research

```
SELECT OPTION BY NUMBER
MOVE JOIN=1,EDIT LEFT CONTIG=2,EDIT RIGHT CONTIG=3,DISPLAY=4,COMPLETE JOIN=5,GIVE UP=6

 OPTION NUMBER =4

DEFINE REGION
 RELATIVE POSITION OF LEFT END=596

 RELATIVE POSITION OF RIGHT END=___

          605.      615.      625.      635.      645.      655.      665.      675.      685.      695.      705.      715.
282  GAGTCAGGCGTCATTTTT1TGGTACGGAAAGTGATGCGAAAAB  CAGCGGCCAGTTVCGTVGAACAG
     GAGTCAGGCGTCATTTTTCTGGTACGGAAAGTGATGCGAAAAA--CAGCGGCCAGTTTCGTTGAACAG
     GAGTCAG4CGT1ATTTTTCTGGTACGGAAAGTGATGCGAAKAAAACAGCGGC AGT  CGTTGAACAG
           *   *                              *  **       *   **
            10.       20.       30.       40.       50.       60.       70.       80.       90.      100.      110.      120.
SELECT OPTION BY NUMBER
MOVE JOIN=1,EDIT LEFT CONTIG=2,EDIT RIGHT CONTIG=3,DISPLAY=4,COMPLETE JOIN=5,GIVE UP=6

 OPTION NUMBER =5


SELECT OPTION BY NUMBER
STOP=0,ENTER=1,PRINT=2,DISPLAY=3,JOIN=4,COMPLEMENT=5,EDIT=6
 OPTION NUMBER = 2


NUMBER OF GELS=   282  NUMBER OF CONTIGS=    66
TO SELECT CONTIGS TYPE Y
Y

 NUMBER OF LEFT GEL THIS CONTIG =232
```

```
CONTIG LINES

 999       1295.       0     232     240
DEFINE REGION
 RELATIVE POSITION OF LEFT END=___

 RELATIVE POSITION OF RIGHT END=___


GEL LINES

LAR16         232          1.    -210       0       2
MLD5            2        143.    -331     232       1
LAD9            1        293.     214       2       4
LAMA9           4        293.    -214       1     282
LAD14         282        335.     329       4       5
LAFB12          5        596.    -232     282     217
LAR29         217        623.     271       5       6
LAZA68          6        770.    -121     217      10
LAAD4          10        802.     148       6       7
LAZA31          7        897.     268      10       8
LAZA46          8        897.     149       7       9
LAZA65          9        897.     237       8     240
LZA65         240       1100.     196       9       0
TO SELECT ANOTHER CONTIG TYPE Y


SELECT OPTION BY NUMBER
STOP=0,ENTER=1,PRINT=2,DISPLAY=3,JOIN=4,COMPLEMENT=5,EDIT=6
 OPTION NUMBER = 3

 NUMBER OF LEFT GEL THIS CONTIG =232

DEFINE REGION
 RELATIVE POSITION OF LEFT END=___

 RELATIVE POSITION OF RIGHT END=___
```

Fig 7


10. 20. 30. 40. 50. 60. 70. 80. 90. 100. 110. 120.
-232 GTACCGGCTGTCTGGTATGTATGAGTTTGTGGTGAATAATGCCCCTGAACAGACAGAGGACGCCGGGCCCGCAGAGCCTGTTTCTGCGGAAATGTTCGACGGTGAGCTGAGTTTTGCCCT
GTACCGGCTGTCTGGTATGTATGAGTTTGTGGTGAATAATGCCCCTGAACAGACAGAGGACGCCGGGCCCGCAGAGCCTGTTTCTGCGGAAATGTTCGACGGTGAGCTGAGTTTTGCCCT
130. 140. 150. 160. 170. 180. 190. 200. 210. 220. 230. 240.
-232 GAAACTGGCGCGTGAGATGGGCCGACCCGACTGGCGTGCCATGCTTGCCGGGATGTCATC1ACG4A4TAVGCCGACTGGCACCGCTTTTA
-2 CGACDCGACTG CGT CCATGCTTGCCGG ATGT-AT-DACGG TATGCCGACTGGCACCGCTTTTACAGTACCCATTATTTTCATGATGTTCTGCT
GAAACTGGCGCGTGAGATGGGCCGACCCGACTGGCGTGCCATGCTTGCCGGGATGTCATCCACGGAGTATGCCGACTGGCAGCGCTTTTACAGTACCCATTATTTTCATGATGTTCTGCT
250. 260. 270. 280. 290. 300. 310. 320. 330. 340. 350. 360.
-2 GGATATGCACTTTTCCGGGCTGACGTACACCGTGCTCAGCCTGTTTTTCAGCGATCCGGATATGCATCCGCTGGATTTCAGTCTGCTGAACCGGCGCGAGGCTGA GCAAGAGCCTGAAG
1 GATCCGGATATGCATCCGCTGGATTTCAGTCTGCTGAACCGGCGCGAGGCTGA CGAAGAGCCTGAAG
-4 GATCCHXATATGACTCCGCTXGATTTCAGTCTGCTGAACCGGC-CXAGGCTGGACGAAGAGCCTGAAG
282 CGCGAGGCTGXACGAAGAGCCTGAAG
GGATATGCACTTTTCCGGGCTGACGTACACCGTGCTCAGCCTGTTTTTCAGCGATCCGGATATG--TCCGCTGGATTTCAGTCTGCTGAACCGGCGCGAGGCTG-ACGAAGAGCCTGAAG
370. 380. 390. 400. 410. 420. 430. 440. 450. 460. 470. 480.
-2 ATGAATGTGCTGATGCAGAAAGCGGCAGXGCTTGCCGX XGTGTCCGCTTTGGCCCGGACGGGAATGAAGTTATCCCCGCTTCCCCGGATGTGGCGGACATGACGGAGGATG
1 ATG ATGTGCTGATGCAGAAAGCGGCAGGGCTTGCCGG AGTGTCCGCTTTGGCCCGGACGGGAATGAAGTTATCCCCGCTTCCCCGGATGTGGCGGACATGACGGAGGATGACG TAA
-4 ATGXATGTGCTGATGXAGAAAG1GG1AGGGDTTGXCGG AXTGTCCGCTTTGGCCCGGACGGGAATGAAGTTATCCCCGCTTCCCCGGATGTGGCGGACATGACGGTGGATGACG TAA
282 ATGXATGTGCTGATGCAGAAAGCGGCAGGGCTTGCCGGAGGTTGTCCGCTTTGGCCCGGACGGGAATGAAGTTATCCCCGCTTCCCCGGATGTGGCGGACATGACGGAGGATGACGTTAA
ATGAATGTGCTGATGCAGAAAGCGGCAGGGCTTGCCGGAG--TGTCCGCTTTGGCCCGGACGGGAATGAAGTTATCCCCGCTTCCCCGGATGTGGCGGACATGACGGAGGATGACGTTAA
490. 500. 510. 520. 530. 540. 550. 560. 570. 580. 590. 600.
1 TGCTGATGACAGTATCAGAAGGGATC
-4 TGCTGATGACAGTATCAGAAGGGATC
282 TGCTGATGACAGTATCAGAAGGGATCGCAGGAGGAGTCCGGTATGGCTGAACCGGTAGGCGATCTGGTC6TTGATTTGAGTCTGGATGCGGCCAGATTTGACGAGCAGATGGCCAGAGTC
-5 GAGTC
TGCTGATGACAGTATCAGAAGGGATCGCAGGAGGAGTCCGGTATGGCTGAACCGGTAGGCGATCTGGTC-TTGATTTGAGTCTGGATGCGGCCAGATTTGACGAGCAGATGGCCAGAGTC
610. 620. 630. 640. 650. 660. 670. 680. 690. 700. 710. 720.
282 AGGCGTCATTTTT1TGGTACGGAAAGTGATGCGAAAAB CAGCGGCCAGTTVCGTVGAACAG
-5 AG4CGT1ATTTTTCTGGTACGGAAAGTGATGCGAARAAAACAGCGGC AGT CGTTGAACAGTCGCTGAGCCGACAGGCT6GCTGCACA6AAAGCHGG A2TTCCGTCGGGCA2TATAA
217 AAAGTGATGCGAAAAAAACAGCGGC AGT CGTTGAACAGTCGCTGAGCCGACAHGGCGGDTGCACAGAAAGCGGGGATTTCCGTCGGGCAGTATAA
AGGCGTCATTTTTCTGGTACGGAAAGTGATGCGAAAAAAACAGCGGCCAGTTTCGTTGAACAGTCGCTGAGCCGACAGG--GGCTGCACAGAAAGCGGGGATTTCCGTCGGGCA-TATAA

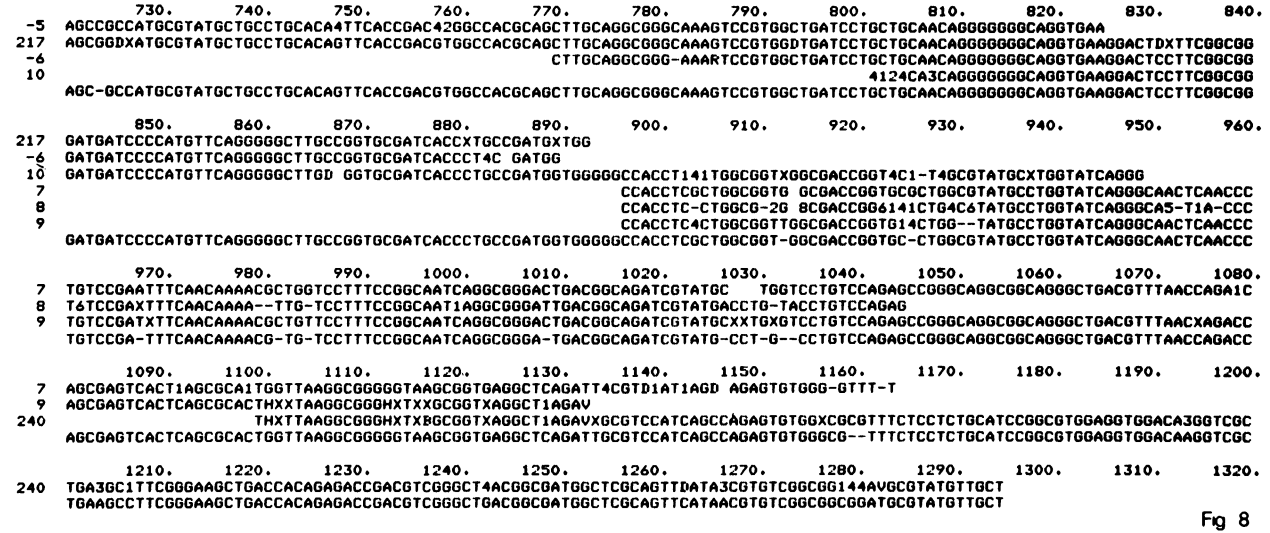

730. 740. 750. 760. 770. 780. 790. 800. 810. 820. 830. 840.
-5 AGCCGCCATGCGTATGCTGCCTGCACA4TTCACCGAC42GGCCACGCAGCTTGCAGGCGGGCAAAGTCCGTGGCTGATCCTGCTGCAACAGGGGGGGCAGGTGAA
217 AGCGGDXATGCGTATGCTGCCTGCACAGTTCACCGACGTGGCCACGCAGCTTGCAGGCGGGCAAAGTCCGTGGDTGATCCTGCTGCAACAGGGGGGGCAGGTGAAGGACTDXTTCGGCGG
-6 CTTGCAGGCGGG-AAARTCCGTGGCTGATCCTGCTGCAACAGGGGGGGCAGGTGAAGGACTCCTTCGGCGG
10 4124CA3CAGGGGGGGCAGGTGAAGGACTCCTTCGGCGG
AGC-GCCATGCGTATGCTGCCTGCACAGTTCACCGACGTGGCCACGCAGCTTGCAGGCGGGCAAAGTCCGTGGCTGATCCTGCTGCAACAGGGGGGGCAGGTGAAGGACTCCTTCGGCGG
850. 860. 870. 880. 890. 900. 910. 920. 930. 940. 950. 960.
217 GATGATCCCCATGTTCAGGGGGCTTGCCGGTGCGATCACCXTGCCGATGXTGG
-6 GATGATCCCCATGTTCAGGGGGCTTGCCGGTGCGATCACCCT4C GATGG
10 GATGATCCCCATGTTCAGGGGGCTTGD GGTGCGATCACCCTGCCGATGGTGGGGGCCACCT141TGGCGGTXGGCGACCGGT4C1-T4GCGTATGCXTGGTATCAGGG
7 CCACCTCGCTGGCGGTG GCGACCGGTGCGCTGGCGTATGCCTGGTATCAGGGCAACTCAACCC
8 CCACCTC-CTGGCG-2G 8CGACCGG6141CTG4C6TATGCCTGGTATCAGGGCA5-T1A-CCC
9 CCACCTC4CTGGCGGTTGGCGACCGGTG14CTGG--TATGCCTGGTATCAGGGCAACTCAACCC
GATGATCCCCATGTTCAGGGGGCTTGCCGGTGCGATCACCCTGCCGATGGTGGGGGCCACCTCGCTGGCGGT-GGCGACCGGTGC-CTGGCGTATGCCTGGTATCAGGGCAACTCAACCC
970. 980. 990. 1000. 1010. 1020. 1030. 1040. 1050. 1060. 1070. 1080.
7 TGTCCGAATTTCAACAAAACGCTGGTCCTTTCCGGCAATCAGGCGGGACTGACGGCAGATCGTATGC TGGTCCTGTCCAGAGCCGGGCAGGCGGCAGGGCTGACGTTTAACCAGA1C
8 T&TCCGAXTTTCAACAAAA--TTG-TCCTTTCCGGCAAT1AGGCGGGATTGACGGCAGATCGTATGACCTG-TACCTGTCCAGAG
9 TGTCCGATXTTCAACAAAACGCTGTTCCTTTCCGGCAATCAGGCGGGACTGACGGCAGATCGTATGCXXTGXGTCCTGTCCAGAGCCGGGCAGGCGGCAGGGCTGACGTTTAACXAGACC
TGTCCGA-TTTCAACAAAACG-TG-TCCTTTCCGGCAATCAGGCGGGA-TGACGGCAGATCGTATG-CCT-G--CCTGTCCAGAGCCGGGCAGGCGGCAGGGCTGACGTTTAACCAGACC
1090. 1100. 1110. 1120. 1130. 1140. 1150. 1160. 1170. 1180. 1190. 1200.
7 AGCGAGTCACT1AGCGCA1TGGTTAAGGCGGGGGTAAGCGGTGAGGCTCAGATT4CGTD1AT1AGD AGAGTGTGGG-GTTT-T
9 AGCGAGTCACTCAGCGCACTHXXTAAGGCGGGHXTXXGCGGTXAGGCT1AGAV
240 THXTTAAGGCGGGHXTXBGCGGTXAGGCT1AGAVXGCGTCCATCAGCCAGAGTGTGGXCGCGTTTCTCCTCTGCATCCGGCGTGGAGGTGGACA3GGTCGC
AGCGAGTCACTCAGCGCACTGGTTAAGGCGGGGGTAAGCGGTGAGGCTCAGATTGCGTCCATCAGCCAGAGTGTGGGCG--TTTCTCCTCTGCATCCGGCGTGGAGGTGGACAAGGTCGC
1210. 1220. 1230. 1240. 1250. 1260. 1270. 1280. 1290. 1300. 1310. 1320.
240 TGA3GC1TTCGGGAAGCTGACCACAGAGACCGACGTCGGGCT4ACGGCGATGGCTCGCAGTTDATA3CGTGTCGGCGG144AVGCGTATGTTGCT
TGAAGCCTTCGGGAAGCTGACCACAGAGACCGACGTCGGGCTGACGGCGATGGCTCGCAGTTCATAACGTGTCGGCGGCGGATGCGTATGTTGCT

Fig 8

and so has selected the "complete join" option. Then he has selected to print the whole of this contig. After this he has selected to display the whole of this contig, which is shown in Figure 8. It can be seen how easily the quality of the data for each part of the sequence can be assessed from this type of printout.

SUMMARY

This paper describes a new computer storage and processing system which has particular application to data gained by the shotgun method of DNA sequencing. Its advantage lies in its ability to establish, maintain and display the relationships between gel readings and in its ease of use. The programs written in FORTRAN run on the same PDP-11 computer mentioned in previous papers [3] and are available on request.

ACKNOWLEDGEMENTS

I should like to thank Dr. T.S. Horsnell for discussions and Drs. F. Sanger and P.J.G. Butler for critical reading of this manuscript.

REFERENCES

1. Staden, R. (1979) Nucleic Acids Research 6, 2601-2610.
2. Gingeras, T.R., Milazzo, J.P., Sciaky, D. and Roberts, R.J. (1979) Nucleic Acids Research 7, 529-545.
3. Staden, R. (1977) Nucleic Acids Research 4, 4037-4051.
4. Staden, R. (1978) Nucleic Acids Research 5, 1013-1015.
5. Staden, R. (1980) Nucleic Acids Research 8, 817-825.