Computer programs for the assembly of DNA sequences

T.R.Gingeras, J.P.Milazzo*, D.Sciaky and R.J.Roberts

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, and *Computing Center, State University of New York, Stony Brook, NY 11794, USA

Received 16 May 1979

ABSTRACT

A collection of user-interactive computer programs is described which aid in the assembly of DNA sequences. This is achieved by searching for the positions of overlapping common nucleotide sequences within the blocks of sequence obtained as primary data. Such overlapping segments are then melded into one continuous string of nucleotides. Strategies for determining the accuracy of the sequence being analyzed and reducing the error rate resulting from the manual manipulation of sequence data are discussed. Sequences mapping from 97.3 to 100% of the Ad2 virus genome were used to demonstrate the performance of these programs.

INTRODUCTION

Technical advances in the procedures used for DNA sequencing have progressed rapidly. The advent of the plus-minus method (1) and the chemical modification method (2) led quickly to the determination of the complete sequences for two small viral genomes, ϕ X174 (5,386 base pairs) (3) and SV40 (5,226 base pairs) (4, 5). Recently, a new method for sequence determination has been described by Sanger and his colleagues (6). This method, which is based upon the incorporation of dideoxynucleotides as chain terminators, has proved to be more efficient and rapid than any method currently employed. The rate limiting step in the process of nucleic acid sequencing is now shifting from data acquisition towards the organization and analysis of that data.

For most sequences reported to date, a first step has been the construction of a detailed restriction enzyme map. However, as longer DNA molecules are tackled, this initial step can become a formidable task. In view of the comparative ease with which DNA sequences can now be generated, this step no longer seems necessary. Rather, it is more profitable to prepare a large number of small restriction fragments from within the DNA molecule of interest and to use each fragment either as a primer in the chain termination

reaction or as a substrate for the chemical method of sequencing. In this way a large number of short sequence stretches, each 200-300 nucleotides long, can be obtained. In principle these fragmentary sequences can then be pieced together by finding overlapping stretches and, with a sufficient number of such sequences, the complete primary sequence of the original molecule can be deduced. The assembly of such fragmentary data into a complete sequence by repetitive searching for overlaps or complementarity is a task ideally suited to computer methods. In this paper we describe a collection of such computer programs.

MATERIALS AND METHODS

A. All programs described in this report were written in ASCII Fortran and executed on a Univac 1110 computer.

B. Description of programs collectively called ASSEMBLER

MONITOR

This is an interactive program which allows the user to choose which function from the ASSEMBLER collection (Figure 1) the computer is to perform. The options include the ability: to enter new DNA sequence data into the computer; to reproduce either all or part of the previously-entered data; to determine and locate overlapping sequences between any two sets of data; to take two strings of nucleotides which share a common region of sequence and meld them into a single string. This last operation is the basic element in reconstructing the complete DNA sequence.

ALIGN

This program arranges any stored sets of sequence data into a standard format, which includes a heading for each set of sequences and their subsequent formatting into 50 characters per line.

HOMOLOGY

This program searches for strings of nucleotides which may be common to any two sets of stored sequences. The stringency by which the overlap is defined can be adjusted by the user (see Fig. 1). The program allows for overlaps which occur either by homology or by complementarity. The search for overlaps can be conducted on either strand of the DNA molecule as selected by the user (Figure 1).

Figure 1**@XQT ASSEMBLER****WELCOME TO THE ASSEMBLER****WHAT REGION ARE YOU INTERESTED IN?****PLEASE KEY IN A LETTER A, B.....S****TO TERMINATE RUN...KEY IN @EOF****> N****THE ASSEMBLER HAS THE FOLLOWING CAPABILITIES:**

- 1. LISTING OF ENTIRE MASTER TABLE**
- 2. ENTRY OF REGIONAL DATA**
- 3. LISTING OF PREVIOUS REGIONAL DATA**
- 4. ALIGNED AND MELDED DATA FOR A GIVEN REGION**

PLEASE KEY IN FUNCTION DESIRED (1..2..4)**> 4****DO YOU WANT TO ALIGN IN THE 3-5 DIRECTION?****> NO****HOW MANY IN A ROW FOR INITIAL OVERLAP?****> 11****WILL LOOK FOR 11**

Figure 1: This is an example of the interaction between the MONITOR program and a user. The program first requests the user to identify the region of the genome which is of interest and requires a single letter response. In this case, the data of REGION N is requested, which includes the sequences mapping from 97.3% to 100% on the Ad2 genome. The MONITOR program then requests which function of the ASSEMBLER the user would like to operate. If the request is to meld or align data (Option 4), the user is questioned as to (a) which strand is to be searched, and (b) what stringency is required for the definition of the initial overlap between any two sequences being compared.

MELD

This program condenses any two strings of nucleotides which contain overlapping sequences into a single composite sequence. From a complete set of data it would allow the reconstruction of an entire DNA sequence.

C. Description of files comprising ASSEMBLER

TRIAL Files

These files are work spaces within the computer where primary sequence data, as read from each of the autoradiographs, can be entered and edited. There is one TRIAL file for each major segment being sequenced. For example, a genome of 50,000 nucleotides might be divided into 10 major segments; data from each segment would be assigned to a different TRIAL file.

REGION Files

For every TRIAL file there is a corresponding REGION file. These files are constructed by taking the primary sequence data entered into the TRIAL file and passing it through the ALIGN program. This results in the sequences being formatted into 50 nucleotides per line and being identified by an appropriate heading (Figure 2). It is the data found in these REGION files upon which the HOMOLGY and MELD programs operate.

MASTER File

This file acts as a permanent archive to store data from every sequencing gel processed by the ASSEMBLER programs. The sequences stored in this file are an exact copy of the primary data which enter the REGION files.

TESTOUTPUT File

During the operation of the HOMOLGY and MELD programs the results are recorded in an output file, called TESTOUTPUT. The entire contents of this file can be printed after the HOMOLGY or MELD programs have functioned. In addition, the power of the editing utilities, associated with most large computers, can be used to select specific data of interest from within these files.

TEMP.

This file is an output file which receives the results of the operations derived from either the ENZYMES or RESEARCH programs.

```

BED,R REGIONN.
READ-ONLY MODE
ED 15R2-MON-04/02/79-17:19:56-(0,)
EDIT
0:>LNP!
1:N-107 (09/29/78) 99999 146
2:AGGAGGTATAACAAAATTAATAGGAGAGAAAAACACATAAACACCTGAAA
3:AACCCTCCTGCCTAGGCAAAATAGCACCTCCCCTCCAGAACACATAC
4:AGCGCTTCCACAGCGGCAGNCATCAACAGTCAGCTTACGTAAAAGC
5:N-129 (10/10/78) 99999 184
6:AGTCAGCCTTACCAGTAAAAAANCCTATTAAACACACCACTCGACAGGG
7:CACCAGCTCAATCAGTCACAGTGTAAGGAGGGCCAACTACAGAGCGAGTA
8:TATATAGGACTAAAAAATGACGTAAACGTTAAAGTCCACAAAAACACCC
9:AGAAAACCGGCANGCGAACCTACGCCCAGAACGA
10:N-144 (10/12/78) 99999 101
11:TGAAAAACCTCCTGCCTAGGCAAAATAGCACCTCCCCTGATNCAGAACCAA
12:ANTACAGCANNNTCCACANTGGNACCATAATAGTNAGCTTTACCAGTAAAN
13:A

```

Figure 2: This is the formatted sequence data as it is recorded within a REGION file. Depicted here are the results of three sequencing gels used to determine the sequence of the HindIII K fragment (97.3% to 100%) of the Ad2 genome. Each block of data contains a heading which lists the segment from which these sequences were determined, and a code number corresponding to the date the experiment was performed. The total number of nucleotides in each block is recorded after the flag 99999 (e.g., N-107 (09/29/78)...99999 146 nucleotides).

RTAB

This file contains a table which lists the name, recognition site, and length of the recognition site for every restriction endonuclease thus far determined (7). It can be easily updated to accommodate new information.

D. Additional Programs

Two other programs have also been found useful in the assembly of DNA sequences. These programs, which are briefly described below, are similar to ones presented elsewhere (8, 9, 10).

ENZYMES

This program searches through the data stored in any of the REGION files for the presence of known restriction endonuclease recognition sites as stored in the RTAB file.

RESEARCH

This program allows the user to search the REGION files for the presence of any specified nucleotide sequence of unrestricted length. Unidentified nucleotides (signified by the letter N) can be included within the specified sequence. The program will search simultaneously for up to ten such sequences of interest.

E. Summary Flow Diagram

A flow diagram indicating the interrelationship of the programs and files of the ASSEMBLER is shown in Figure 3.

Figure 3: This is a flow diagram illustrating the strategy employed by the ASSEMBLER collection of programs in order to map and analyze newly derived sequence data. A genome or DNA fragment is divided into large discrete segments (usually corresponding to restriction enzyme sites) and each major segment is assigned a TRIAL file and a REGION file. The sequence data as it is derived from each major segment of the genome is

recorded directly into the TRIAL file. Thus, data from segment A on the genome enters the corresponding TRIAL A file. The sequence from this TRIAL file can be entered into the ASSEMBLER collection of programs as new data. The MONITOR program greets the user with questions (see Figure 1) which determine from which major segment of the genome the sequence was derived. If new data is to be entered, then the MONITOR program activates the program, ALIGN, which formats the data into 50 characters per line and copies the sequences into the MASTER file and a REGION file (i.e., REGION A). The MASTER file acts as an archive for all data entered into this ASSEMBLER collection. The REGION file contains blocks of data as determined from a particular segment of the genome. It is this formatted sequence data that can be analyzed by the remaining programs of ASSEMBLER.

Sequence data stored in a REGION file can be processed by the MONITOR program to activate (a) the HOMOLOGY program, to find overlapping sequences recorded within a particular REGION file, (b) the MELD program, to condense the overlapping sequences into one continuous string of nucleotides so that the entire genome can be reconstructed, (c) the ENZYMES program, so that all restriction sites present in the sequences recorded in a REGION file can be located, and (d) the RESEARCH program so that sequences other than restriction enzyme recognition sites can also be located within a REGION file.

The results from the HOMOLOGY and MELD programs are sent to an output file called TESTOUTPUT, while the results from the ENZYMES and RESEARCH program are sent to a file called TEMP.

USE OF THE ASSEMBLER PROGRAMS

A. Objectives

These programs have been written to aid in a specific project aimed at the determination of the complete sequence of the Adenovirus-2 (Ad2) genome. However, they are of more general application when they are required to assemble a complete DNA sequence from a large mass of unordered primary data. These programs were designed with a particular strategy in mind as described below.

The Ad2 genome is approximately 35,000 nucleotide pairs in length and has been arbitrarily divided into fragments of 1,000 to 5,000 nucleotides as defined by two sets of restriction enzyme cleavage sites (EcoRI and HindIII). Each fragment is considered a separate region whose sequence is to be

deduced. In each case the primary fragments are cut, using other restriction endonucleases (e.g., HhaI, HpaII, HaeIII), into many subfragments of much shorter length (20 to 300 nucleotides) and each, in turn, used as a primer in the chain-termination sequencing procedure (6). From each primed reaction, a sequence of 100 to 300 nucleotides is obtained. As many such sequences accumulate from a given region, they will contain homologous or complementary stretches which ultimately will allow them to be fused into a continuous sequence. In this way the complete sequences of the regions can be determined without the necessity for prior mapping of restriction enzyme cleavage sites. Finally, the regions themselves can be fused together by an analogous process. The programs of the ASSEMBLER allow the entry of primary sequence data and subsequent manipulation of that data into continuous strings. This is achieved by searching for overlapping sequences and their subsequent melding. In addition, the programs provide an archive for both the primary sequence data and the manipulative procedures by which they are combined into a continuous string. Access to these various facets of the ASSEMBLER is controlled by the user-guided interactive program called MONITOR.

B. Entering Primary Sequence Data.

The first option available through the MONITOR is the entry of new sequence data. This data is identified by a heading which contains the following information: (a) the region from which the sequence was derived (e.g., sequences from HindIII K, coordinates 97.3% to 100%--the right hand end of the Ad2 genome--are assigned to Region N); (b) a code number identifying the particular reaction from which the data were obtained; and (c) the date of the experiment. The sequence is entered as a continuous string (Figure 4) and occupies a work file (TRIAL file) within the computer. It is then automatically processed through the ALIGN program which formats the data and records it in two separate files. The first is a permanent archive called MASTER which contains a cumulative record of all primary sequence data. The second is a working file--the REGION file--which provides a data base for further manipulation of the primary sequences. It should be noted that only one MASTER file exists but there are many different REGION files. The contents of these files can be printed out either in part or in whole, with access being controlled through either the MONITOR or through the editing functions of the computer.


```

@ED,U TRIALN.
ED 15R2-MON-04/02/79-16:09:10-(0,1)
EDIT
0:>LNP!
1:N-107 (09/29/78):
2:AGGAGGTATAACAAATTAATAGGAGAGAAAA
3:CACATAAACACCTGAAA
4:AACCCCTCCTGCCTAGGCAAAATAGCACCCCTCCCGCTCCAGAACACATACAGCGCTTCC
5:ACAGCGGCGAGNCATCAACAGTCAGCTTACGTAAAGC
6:N-129 (10/10/78):
7:AGTCAGCCTTACCACTAAAAAANCCTATT
8:AAAACACACCACTCGACAGG
9:CACCACTCAATCAGTCACAGTGTAAAAAGGGCCAACTACAGAGCGAGTATATATAGGACT
10:AAAAAATGACGTACGGTTAAAGTCCACAAAAACACCCAGAAAACCGGCAN
11:GCGAACCTACGCCCAAGACGA
12:N-144 (10/12/78):
13:TGAAAAACCTCCTGCCTAGGCAAAATAG
14:CACCCCTCCGATNCAGAACCAANTACAGCANNNTCCACAN
15:TGGNACCATAATAGTNAGCTTTACCACTAANA

```

Figure 4: An example of data input into a TRIAL file. The primary sequence as derived from an autoradiograph is typed directly into a TRIAL file without any prior formatting. A heading (see Figure 2 for details) precedes each block of data.

C. Identification of overlapping sequences.

Another option provided by the MONITOR is access to the program called HOMOLOGY. This program searches for the presence of overlapping strings of nucleotides within the data base stored in the REGION files. An overlap may be homologous, in which case an exact match of say 10 nucleotides is found in two different sequence entries or it may be complementary, in which case the match may be between a string of 10 nucleotides from one sequence entry and its inverse complement from another entry. The exact length of the matching string can be chosen by the user. Clearly, the longer the string, the less the likelihood of a chance match, but the greater the chance of sequence errors preventing two overlapping strings from being discovered. In order that true overlaps may be found efficiently, using a minimal matching string, the program contains certain restraints. Having identified two sequences which both contain the matching string, the program searches for a continuation of this overlap to the left, as well as to the right, of the first match. Once a mismatch occurs between the two sets of sequences being compared, the HOMOLOGY program allows for continued mismatching for a total of three nucleotides. If a match occurs within this three nucleotide limit, then the program continues to scan the pair of sequences, noting the

position of the insertions, deletions, or discrepancies which comprise the mismatches. The continued common sequence must persist for at least an additional five nucleotides from the last point of mismatch to ensure that the genuine continuation has been discovered. If common sequences are not found within a three base limit allowed for mismatches, or if the continued overlapping sequence after a mismatch does not exceed five nucleotides, then another continuous string of the length specified by the user must be found elsewhere along the two sequences being compared.

The results of the HOMOLGY program are summarized immediately in table form for the user (Figure 5) while a detailed listing of these results is transferred to an output file called TESTOUTPUT which can be printed upon request. The listings logged in TESTOUTPUT are in two parts; the first part is a table which maps the positions of the overlapping nucleotides within the two sequences being compared (Figure 6). The second part consists of the two sequences positioned one over the other with the common nucleotides aligned. The nature of the overlaps and the extent of the differences are easily observed as the areas of difference are highlighted by placing them in brackets.

D. Melding overlapping sequences

The linking of overlapping sequences, identified by the HOMOLGY program, into one continuous string of nucleotides is performed by a program called MELD. This program allows the two sequences containing the overlap to be fused into a single continuous sequence. At those positions where discrepancies occur, diacritical marks are placed above the nucleotides concerned (Figure 7). This serves to identify those positions within the sequence where either new data is needed or some re-evaluation of old data is required. Thus, the output from this program gives some quantitative measurement of the consistency and accuracy of the sequence generated.

In the instance where no discrepancy exists between the two sequences being compared, the two sequences can be fused by the MELD program. The original sequences can then be manually deleted from the REGION file and replaced by the new melded product. However, when discrepancies do arise at any position between two sequences being compared, the user must resolve this difference manually before or after the meld is done. Several cycles of melding blocks of larger and larger sequences are required to reconstruct the sequence of each main region and finally these regions can be fused to reconstruct the entire genome sequence.

```

@XQT    ASSEMBLER.RUN
THE ASSEMBLER HAS THE FOLLOWING CAPABILITIES
1. LISTING OF ENTIRE MASTER TABLE
2. ENTRY OF NEW REGIONAL DATA
3. LISTING OF PREVIOUS REGIONAL DATA
4. ALIGNED AND MELDED DATA FOR A GIVEN REGION
PLEASE KEY IN FUNCTION DESIRED (1..2...4)
>4
DO YOU WANT TO ALIGN IN THE 3-5 DIRECTION?
>NO
HOW MANY IN A ROW FOR INITIAL OVERLAP?
>12
WILL LOOK FOR          12

OVERLAPPING SEGMENTS N-107 (09/29/78)          AND    N-144 (10/12/78)
  4
 46   1   7
 54   8  31
 86  40   8
 98  53  16

OVERLAPPING SEGMENTS N-129 (10/10/78)          AND    N-144 (10/12/78)
  2
  1  81   7
  9  89  13
COMPLETED ALIGNMENT FUNCTION
THE ASSEMBLER HAS THE FOLLOWING CAPABILITIES
1. LISTING OF ENTIRE MASTER TABLE
2. ENTRY OF NEW REGIONAL DATA
3. LISTING OF PREVIOUS REGIONAL DATA
4. ALIGNED AND MELDED DATA FOR A GIVEN REGION
PLEASE KEY IN FUNCTION DESIRED (1..2...4)
>#EOF

```

Figure 5: The immediate results of the HOMOLOGY program. After choosing option 4 of the MONITOR, overlapping sequences are identified by heading and the position and extent of overlaps are indicated in tabular form. Thus, within data N-107 and N-144 there are four regions of overlap. The next line indicates that nucleotide 46 of N-107 and nucleotide 1 of N-144 are identical and match for the next 7 nucleotides. The homologies beginning at nucleotides 54 or 98 of N-107 would have been responsible for activating this program, since at both positions, overlaps exceeding 12 nucleotides occur. The precise sequences involved are transferred to the TESTOUTPUT file and can be examined separately (see Figure 6).

The feasibility of using these programs to reconstruct a DNA sequence has been assessed by using primary data we have accumulated while sequencing the right hand end of the Ad2 genome. The region, mapping from 97.3 to 100% on the genome, is about 1,000 nucleotides long. Data resulting from

```

1:      ALIGNED AND MELDED DATA
2:
3: OVERLAPPING SEGMENTS N-107 (09/29/78)                AND   N-144 (10/12/78)
4:      4
5:      46      1      7
6:      54      8     31
7:      86     40      8
8:      98     53     16
9: *
10: * <AGGAGGTAT AACAAAATTA ATAGGAGAGA AAAACACATA AACACC>TGA *
11: * <                                     >TGA *
12: * * * * *
13: * AAAA<C>CCT CCTGCCTAGG CAAAATAGCA CCCTCCCG<C >TCCAGAAC< *
14: * AAAA< >CCT CCTGCCTAGG CAAAATAGCA CCCTCCCG<A >TNCAGAAC< *
15: * * * * *
16: * AACA >TACA GCGCTTCCAC AG<CGGCAGN CATCAACAGT CAGCTTACGT *
17: * CAAAN>TACA GCNNNTCCAC AN<TGGNACC ATAATAGTNA GCTTTACCAG *
18: * *
19: * AAAAGC>
20: * TAAANA>

30: OVERLAPPING SEGMENTS N-129 (10/10/78)                AND   N-144 (10/12/78)
31:      2
32:      1      81      7
33:      9      89     13
34: *
35: * <
36: * <TGAAAAACC TCCTGCCTAG GCAAAATAGC ACCCTCCCGA TNCAGAACCA *
37: * * * * *
38: * * * * *
39: * AANTACAGCN NNTCCACANT GGNACCATAA T>AGTNAGC< C>TTACCACT *
40: * * *
41: * AAAA<ANCC TATTAAACA CACCACTCGA CAGGCGACCA GCTCAATCAG *
42: * AAANA<
43: *
44: * TCACAGTGTA AAAAGGGCCA ACTACAGAGC GAGTATATAT AGGACTAAAA *
45: *
46: * * * * *
47: * AATGACGTAA CGGTAAAGT CCACAAAAA CACCCAGAAA ACCGGCANGC *
48: *
49: *
50: * GAACCTACGC CCAGAACGA>
51: *

```

Figure 6: This is a printout from the TESTOUTPUT file recorded as a result of the operation of the HOMOLOGY program. A table of the overlaps (line 4-8) is recorded above each set of sequences (cf., Fig. 5).

From lines 10 to 20, a detailed listing of the sequences present in N-107 and N-144 are printed. Areas of homology are printed one over another in order to facilitate the easy recognition of such sites. Any areas of disagreement in the two sequences are placed into brackets < >. An asterisk (*) placed above the sequence indicates positions where N (an unidentified nucleotide) occurs in either or both sequence elements.

the extension of twenty different primers were processed through the ASSEMBLER before the complete sequence could be reconstructed. Each base

```

21: * ----- *
22: * <AGGAGGTAT AACAAAATTA ATAGBAGAGA AAAACACATA AACACC>TGA *
23: * ----- *
24: * AAAA<C>CCT CCTGCCTAGG CAAAATAGCA CCCTCCCG<A >TCCAGAAC< *
25: * . . . *** * . * . * ..... * ..... *
26: * AAAAN>TACA GCGCTTCCAC AG<CGGCACC AAAAAACAGA CAGCTAACAG *
27: * . * *
28: * AAAAGA>

52: * ----- *
53: * <TGAAAAACC TCCTGCCTAG GCAAAATAGC AGCCTCCCGA TNCAGAACCA *
54: * ----- *
55: * AANTACAGCN NNTCCACANT GGNACCATAA T>AGTCAGC< C>TTACCACT *
56: * * - - - - - *
57: * AAAAA<ANCC TATTAACA CACCCTCGA CAGGGCACCA GCTCAATCAG *
58: * ----- *
59: * TCACAGTGTA AAAAGGGCCA ACTACAGAGC GAGTATATAT AGGACTAAAA *
60: * ----- *
61: * AATGACGTAA CCGTTAAAGT CCACAAAAA CACCCAGAAA ACCGGCANGC *
62: * ----- *
63: * GAACCTACGC CCAGAACGA>

```

Figure 7: This is an example of a set of results which are a product of the MELD program as recorded into the TESTOUTPUT file. Lines 21 to 28 illustrate the condensed version of the overlapping sequences recorded in Figure 5 for N-107 and N-144. Lines 52 to 63 illustrate similar melded data condensed from gels N-144 and N-129. The sequences within the brackets < > signify nucleotides which are different from the ones present at the same position as printed above. This mismatch can occur either because there exists a genuine disagreement between two sets of data (see Figure 5), or because there is no corresponding data in one set to be compared (see lines 52-55). In the former of these cases, the MELD program chooses a nucleotide to be placed at a position of disagreement based on the following arbitrary rule: A over C over G over T over N. When this rule of precedence is invoked, the nucleotide has a dot (·) placed over it. This tells the user that this base(s) requires re-evaluation. If a dash (-) appears over a base, the user is notified that there was no corresponding nucleotide at that position within the other set of data in order for a comparison to be made. Thus, additional sequence data is needed for this segment of sequence in order that at least two sets of data be used for corroboration.

from the r strand† appeared an average of 5 times within this data. Discrepancies which occurred during multiple readings of any given segment of the DNA sequence were discovered and indicated by the align and melding programs (see Figures 6 and 7). Such discrepancies are caused by: (1) insertions/deletions produced through faulty reading of autoradiographs; (2) positions at which a specific nucleotide is read from one gel and at which no decision (N) could be made from a second gel; (3) positions at which two different specific assignments were made from different readings.

In general, most of the errors arose through attempting to read portions of autoradiographs which were innately difficult to analyze, i.e., the extreme bottom of the gel where artifactual bands often occur or the top of the gel where compression can easily lead to missed residues. Because the decision to replace primary data by melded sequences in the REGION file is effected manually, it is rather easy to assess the quality of the melded product and hence to discard from consideration those regions of sequence which are necessarily error prone. During the course of this evaluation of the ASSEMBLER, we observed no instance of the incorrect overlapping of sequence stretches.

E. Additional programs provided by the ASSEMBLER

The program called ENZYMES is similar to others described elsewhere (8, 9, 10). It identifies and locates all known restriction endonuclease recognition sites (7) from within the sequences stored in the REGION file. It can be used as a check on the final sequence by predicting the number of fragments and their sizes to be expected for all restriction enzymes which cleave the region sequence. These may then be compared with experimental values as illustrated in Table 1. In the case of discrepancies further evaluation of the final sequence is required.

DISCUSSION

The programs described in this paper were designed to assist in the assembly of long DNA sequences from the much shorter sequences obtained as primary data. They are tailored for use in a sequencing strategy which avoids the mapping of restriction enzyme cleavage sites. This is accomplished

† The r strand of Ad2 is the strand which is transcribed rightwards on the conventional map (11).

Table 1: List of predicted and observed restriction endonuclease recognition sites present between map positions 97.3 to 100%^a in the Ad2 genome.

<u>Enzyme</u>	<u>Occurrences</u>	<u>Occurrences</u>	<u>Recognition</u>
	<u>Predicted</u>	<u>Observed^b</u>	
<u>AluI</u>	2	2 ^c	AGCT
<u>AsuI</u>	1	nt	GGNCC
<u>AvaI</u>	1	nt	CPyCGPuG
<u>BbvI</u>	3	nt	GC ^A _T GC
<u>EcaI</u>	1	nt	GGTNACC
<u>EcoRI</u>	2	nt	PuPuATPyPy
<u>FnuDII</u>	3	3	CGCG
<u>HaeII</u>	1	1	PuGCGCPy
<u>HaeIII</u>	2	2	GGCC
<u>HhaI</u>	3	3	GCGC
<u>HindII</u>	1	1	GTPyPuAC
<u>HindIII</u>	1	1 ^d	AAGCTT
<u>HinfI</u>	2	1 ^e	GANTC
<u>HpaI</u>	1	1	GTTAAC
<u>HpaII</u>	4	4	CCGG
<u>HphI</u>	2	nt	GGTGA
<u>MboII</u>	1	1	GAAGA
<u>MnII</u>	6	nt	CCTC
<u>TaqI</u>	1	1	TCGA
<u>XmaI</u>	1	1	CCCGGG

^a This region of Ad2 contains 1,009 nucleotides.

^b The occurrences observed shown in this column not only agree in number, but also in the size of fragments predicted by a computer analysis of the sequences in Region N. nt = not tested.

^c One AluI site is from within the HindIII site and will not generate a new fragment.

^d This HindIII site is the site used to generate the fragment from 97.3-100%.

^e One HinfI fragment would not be expected to be observed because it is predicted to be only 14 base pairs in length.

by defining a region of DNA which is of manageable size (say, 1,000-5,000 nucleotides long) and then obtaining sequence information from that region in an arbitrary fashion. For instance, the region may be cleaved by many

different restriction enzymes and the resulting fragments used as primers in the chain termination procedure. The sequences produced are then searched for overlapping stretches and sequences combined until all data is accommodated within a final continuous string. By using the computer to both store the initial data and carry out the subsequent processing, the complete sequence can be reconstructed in a manner which is faithful and efficient. This approach is most important when rather long DNA sequences are studied because the problems of data management begin to rival those of data collection and a serious source of error arises when sequences are copied by hand from one sheet of paper to another.

One aspect of the approach described here is of considerable importance. No restriction enzyme mapping is undertaken prior to sequence determination and so a major time-consuming step is omitted. However, this does mean that the location of restriction enzyme sites can be used as an independent means of checking the sequence. In its simplest form this could be achieved by digesting the DNA, whose sequence has been deduced, with a number of different restriction endonucleases and comparing the digestion patterns with those predicted to occur within that sequence. This results in a rather arbitrary check for the presence of short specific sequences at intervals along the final sequence. We have calculated that as much as 20% of the total Ad2 sequence could be independently checked in this manner with the restriction endonucleases now available. A more rigorous procedure, which might be useful occasionally, would be to actually map the restriction enzyme sites using the method of Smith and Birnstiel (12).

One final check on the sequence accuracy is also possible through the programs of the ASSEMBLER. The heading which precedes each piece of input data contains a symbol (* or -) which indicates the manner of preparation of the template (either a 5'→3' exonuclease or a 3'→5' exonuclease). These markings have no significance during the operation of the HOMOLOGY and MELD programs and so the computer takes no account of the strandedness of each individual sequence. Nevertheless, the strandedness is known by the user and therefore it becomes straightforward to check that each piece of sequence is actually placed within the correct strand.

The programs described above represent a first step in the development of software to aid in DNA sequence determination. Much of the processing still needs user intervention and a next step will require the definition of algorithms to combine these procedures into a single automatic operation. Ultimately, it seems likely that all of the data processing, including the

reading of sequence gels, can be automated.

A copy of these ASCII Fortran programs, along with a complete documentation package, is available from the first author. With slight modifications, these programs have been executed on a PDP 11/60 as well as the Univac 1110 and they should prove adaptable to many other computers as well.

ACKNOWLEDGEMENTS

We thank R.E. Gelinas for his helpful comments and useful discussion. A special thanks to R. Yaffe, C. Carpenter, and M. Moschitta for their help in preparing this manuscript.

The work was supported by grants to RJR from the National Science Foundation (PCM76-82448) and to TRG and REG from the Whitehall Foundation. TRG was supported by a Postdoctoral Fellowship from the National Institutes of Health. DS was a Fellow in Cancer Research supported by Grant DRG-119-F of the Damon Runyon-Walter Winchell Cancer Fund.

REFERENCES

1. Sanger, F. and Coulson, A.R. *J. Mol. Biol.* **94**: 441 (1975).
2. Maxam, A.M. and Gilbert, W. *Proc. Nat. Acad. Sci.* **74**: 560 (1977).
3. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A. III, Slocombe, P.M. and Smith, M. *Nature* **265**: 687 (1977).
4. Reddy, V.B., Thimmappaya, R., Dhar, K.N., Subramanian, B., Zain, S., Pan, J., Ghosh, P.K., Celma, M.L., and Weissman, S.M. *Science* **200**: 494-502 (1978).
5. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G., and Ysaebaert, M. *Nature* **273**: 113-120 (1978).
6. Sanger, F., Nicklen, S. and Coulson, A.R. *Proc. Nat. Acad. Sci. USA* **74**, 5463-5467 (1977).
7. Roberts, R.J. *Methods in Enzymology*, Vol. 65 (in press) (1979).
8. Korn, L.J., Queen, C.L. and Wegman, M.W. *Proc. Nat. Acad. Sci. USA* **74**: 4401-4405 (1977).
9. Staden, R. *Nuc. Acids Res.* **4**: 4037-4051 (1977a).
10. Staden, R. *Nuc. Acids Res.* **4**: 1013-1015 (1977b).
11. Mulder, C., Arrand, J.R., Delius, H., Keller, W., Pettersson, U., Roberts, R.J. and Sharp, P.A. *Cold Spring Harbor Symp. Quant. Biol.* **39**, 397-407 (1975).
12. Smith, H.O. and Birnstiel, M.L. *Nuc. Acids. Res.* **3**, 2387-2398 (1976).