Volume 10 Number 1 1982 Nucleic Acids Research

# GEL, a DNA sequencing project management system

Jan Clayton[1] and Larry Kedes[2]

[1]IntelliGenetics, Inc., 124 University Avenue, Palo Alto, CA 94301, and [2]Departments of Medicine, Stanford University and Veterans Administration Medical Center, Palo Alto, CA 94304, USA

Received 15 September 1981

## Abstract

We have developed an automated system for management of DNA sequencing projects. The system, named *GEL*, can handle data from both random sequences and from fragments whose relative positions are known. The system is highly interactive, self-documenting, and forgiving; it is designed for use by computer-naive molecular biologists. An editor designed specifically for sequences allows simple entry of data. Special functions allow direct checking and immediate editing of paired readings of the same gel. Merging of new random fragment sequences into the project as a whole is semi-automated. The user is shown probable overlaps if they exist, and can edit either the sequences or the consensus. Heuristic approaches to limiting the kinds of searches made in the merging process reduces the problem of combinatoric data overload as sequencing projects grow large. Complete histories of all entries, editing changes, and generation of consensus sequences are automatically prepared.

## Introduction

Modern DNA sequence analysis technology [1] [2] has been one of the important methodological contributions to the explosively accelerating acquisition of information about the primary structure of genes. Advanced strategies for determining the sequence of large molecules [3], [4], [5] has seen the rate of absolute sequence determination rise dramatically from a few hundred bases a week per laboratory to as much as 1000 bases per day per scientist [5]. Already, management of the information, not the actual sequence determination, may be the rate limiting step in DNA sequencing [5].

One method of approaching the problem of DNA sequence analysis of large molecules is to sequence unmapped fragments generated either by restriction digestion [6], by subcloning of fragments generated by random nuclease scission [5] or by generation of a library of overlapping clones [4] to be sequenced by primed DNA synthesis [2] using a universal primer. The kinds of random sequence information generated by these "shotgun" strategies represent a problem in information management amenable to computer assistance. Two computer programs have been reported that make substantial gains in defining the problem and some headway in solving it. Staden [7] has addressed both the difficulty in managing large amount of random sequence data and in coping with the inherent ambiguity of sequences because of shortcomings of the technology. His

system makes use of a user-driven search and compare routine designed to find regions of overlap of new sequences with older sequences in order to generate the contiguous sections of a DNA molecule. As the sequencing project matures and more overlaps are found, the contiguous segments (Staden coined the word "contig" to describe such segments and we have continued to use that nomenclature) get longer until they too merge with one another. A built-in keyboard editor allows the user to make immediate changes in either the new sequence or overlapping old ones, to realign overlaps, or to reject the program's proposed merger of contigs. The program keeps a history of all editing changes, and the origin of all information in any contig.

Gingeras et al [6] have independently written a similar program that allows the user to manually meld sequences and store them in a new file. Neither program has taken full advantage of the inherent power of computer-based automatic management of such sequencing projects or of a heuristic approach to the lengthening of contigs. Our goal was to develop such a program in an entirely interactive mode that would supply all of the housekeeping functions of a DNA sequencing project and at the same time provide the user with an easily accessible data management and retrieval system. *GEL* is a program, written in SAIL, that semi-automatically provides the management system for a DNA sequencing project. It is highly interactive and provides on-line help at every step. All entries and changes in data are automatically recorded and can be easily reviewed.

## Entering Sequences into a Project

All the sequence information pertaining to any particular DNA molecule that is being sequenced is stored in a GEL file. Any number of projects can be handled at one time. Each gel file has a separate name in the format NAME.GEL, such as ACTIN-2.GEL or LACZ.GEL. Sequences can be entered in one of two ways: either directly from the terminal by creating a new sequence with the NEW command or from an existing computer file that had been prepared with a standard text editor in the MOLGEN *SEQ* format [8]. The NEW command prompts the user to provide a name for the film on which the raw sequence data is displayed. Users should label each actual film with the identifying name. Then *GEL* prompts the user for a name for the actual film lanes being read; this might correspond to the name of a fragment e.g. *Alu*I-12 that should be written directly on the film as well. The user is then asked to enter as much textual commentary as he wants for personal future reference. Next, *GEL* asks for the name of the person who is about to read the data. After entry of the sequence data, the user will be asked if he wants to enter another new sequence. If he does, then the program will assume the name of the reader is the same unless told otherwise.

Entering new sequences from a *SEQ* [8] formatted file makes use of the ENTER command. The user is led through a series of queries about each sequence in the file. The user can modify the film name, give a name to the DNA fragment whose sequence is being entered, and enter the name of the reader. Any comments that had been stored in the file are printed and the user can modify those or add additional comments. This is done for every sequence in the *SEQ* file.

When entering sequences with the NEW command, *GEL* enters a special sequence editor phase: a scale is printed across the top of the screen (page) and as the user inputs a string of bases a space is inserted every tenth character so that the sequence is formatted in groups of 10. Only alphanumeric symbols that conform to the Stanford Code (see below and Table I) are accepted, spaces are ignored. As the user enters the fiftieth symbol, the cursor (or carriage) returns to the start of the next line for continued editing. There are only three editing commands at this point: the delete key will rub out one character at a time but only on the current line; a carriage return signifies the end of the sequence and the user is prompted for additional sequences; the ESCAPE key (called ALTMODE on some terminals) starts the sequence editor.

The sequence editor is completely interactive and help is provided for the asking. Users can insert new bases in the middle of the sequence, append new bases to the end of the sequence, replace one or more bases with any number of bases, and delete bases already entered. Any segment of the sequence can be printed out. When a gel is edited, all changes are recorded, including date and the exact modifications, in a file called the PROJECT file. PROJECT files have the same filename as GEL files, e.g. ACTIN.PROJECT or LACZ.PROJECT. The distinction between GEL and PROJECT files is simply that GEL files contain every sequence in the state it was entered and records no editing changes. The PROJECT files are the repositories of editing changes. Both kinds of files are created and modified automatically.

An example of an interactive session with the NEW option including an editing example follows. User responses are in underlined letters and explanatory annotations by the authors are in *italics*.

```
                    IntelliGenetics
              GEL - GEL Assembler System
                    August 9, 1981

        (Copyright 1981 by the Board of Trustees, Stanford University)

Name of project: late-gene     All data for the project are stored in a file with this name.

GEL: ?

        The following gel commands are available:

    CHeck <Gel1> <Gel2> - Compares two GELs and reports any base
                          mismatches
    COmments <Gel>      - Prints comments for GEL
    DElete <Gel>        - Deletes GEL from project (confirmation required)
    DIsplay <Gel>       - Prints out the PROJECT file data on GEL
    ENter <GelFile>     - Enters Gels from a file (in SEQ format)
    EDit <Gel>          - To edit a gel
    Join <Gel1> <Gel2>  - To join two sequences that have a known overlap
    List                - Prints out a list of all the gels in a project
    Merge <Gel>         - Attempt to form contigs with GEL
    New                 - To enter new data (from the terminal)
    Print <Gel>         - Prints the specified range of the gel.  If no
                          range is specified, then the whole gel is printed.
```

```
    SAve                - To save all changes made
    Undo <Contig>       - Restores member gels to gel status over the
                          specified range.  If no range is specified,
                          then the entire contig is disassociated.

GEL: new    User wants to enter a new sequence.

Name for this new GEL: film 5
The reader of film 5? [] larry

Comments (<CR> when done)
;Alu endlabelled x hae3 fragment 7
;

Please enter the new sequence.


        10         20         30         40         50
AGCGATAGAA CAGCCTCGCG AGCGCGCCCT CTCCTCCGAG CCGCGAATAT

        60         70         80         90        100
CTAATAAGGC CGCCTTATTC GCGCCTAGCT AGCTAGCCCT CCGAGCGCGA

       110        120        130        140        150
GCGATTCGCG ATTTCGCGTC GTAGCTATCC GAATCGGATC GATCGA
```

*As user types in a sequence, the spaces and the scales automatically appear. A simple carriage return signifies the end of the entry.*

```
Another new GEL? N

GEL: PR

Gel to Print: ?


    Please enter the name of the GEL or CONTIG that you want to print.
Specify the name of number of a GEL (can be followed by "'" for the
inverse complement.)

    Available Gels are:

1.  Gel 1;Bgl2-1
2.  Gel 2;aluxhae3 6
3.  Gel 2;aluxhae3 5
3.  Gel 3;aluxhae3 3
4.  Gel 4;aluxhae3 9
5.  Film 9; sequence 1
6.  FiLm 9; sequence 2
7.  Late-gene-1:
        Film 9; sequence 1 {1 - 77}
        Comp-FiLm 9; sequence 2 {21 - 101}
8.  Gel 13
9.  film 5


Gel to Print: 8

Range (<CR> for entire GEL) : 37 90
```

```
Film 13 from 37 to 90


        46          56          66          76          86
CGAGCCGCGA ATATTTATAA TATAAATTCG CCGCCTAATA AGGCCGCCTT ATTC


GEL: ED
Gel Name: 8

Edit: ?


    The editing commands are:

        D <range> - to delete a range of characters
        I <place> - to insert characters before place
        A <place> - to append characters after place
        R <range> - to replace specified range of characters with others
        P <range> - to print a section or all of the Gel
        Q         - to quit (all changes will be saved)

        <range>   - can either be a single number or a range of numbers
                    designated in any of the following ways:
                        23-34, 23:34, or 23 34
Edit: R

Base(s) to replace: 37 42
Replace with: AAAAAAAAAAAAAAAAAAAAA
Edit: P2180


        30          40          50          60          70          80
AGCGCGCCCT CTCCTCAAAA AAAAAAAAAA AAAAAAAAGC GAATATTTAT AATATAAATT

Edit: Q
```

## Checking Two Readings of the Same Gel

A fast and convenient check on the identity of two independent readings of the same gel is offered by the CHECK command. The matching algorithm requires a high degree of homology. Ambiguities, whether base mismatches or small loopouts, are shown as asterisks in the consensus sequence that is presented to the reader. In practice, direct reexamination of the sequencing film should allow either an agreement as to the assignment or an agreement to retain a symbol appropriate to the degree and kind of ambiguity as provided by the Stanford code (Table I). A user interaction with the CHECK command follows:

```
GEL: check
First Gel: 12
Second Gel: 13

Stars mark mismatches.
```

```
        10         20         30         40         50         60         70
CCTTCTTTCC AGCTGGTTGT GCTTTCGCAG GGCATGATGA TTGTGATTCT CACGAATGCT GGTGTGCAAT
CCTTCTTTCC AGCTGGTTGT GCTTTCGCAG G CATGATGA YYGTGATTCT CACGAATGCT GGTGTGCAGT
                                  *         **                              *
CCTTCTTTCC AGCTGGTTGT GCTTTCGCAG GGCATGATGA TTGTGATTCT CACGAATGCT GGTGTGCA*T
Consensus

        80
TGGAAGCG
TGGAAGCG

TGGAAGCG
```

*Note how the consensus sequence accepted the unmatched G at position 32, converted a pyrimidine/T mismatch to T's and retained ambiguity at position 69.*

```
Action on Match? (C, E, D, or ?) ?

        C - Confirm match and record in project file
        D - Delete match from consideration list
        E - Edit match before confirming

Action on Match? (C, E, D, or ?) e

Edit 12, 13 or Consensus? 12

Edit: d

Delete: 32
Edit: q
```

*After direct examination of the raw data, the readers decide that base 32 in Gel 12 was not really there. Quiting the editor automatically redisplays the match being edited but with the changes in place.*

```
Stars mark mismatches.

        10         20         30         40         50         60         70
CCTTCTTTCC AGCTGGTTGT GCTTTCGCAG GCATGATGAT TGTGATTCTC ACGAATGCTG GTGTGCAATT
CCTTCTTTCC AGCTGGTTGT GCTTTCGCAG GCATGATGAY YGTGATTCTC ACGAATGCTG GTGTGCAGTT
                                          * *                               *
CCTTCTTTCC AGCTGGTTGT GCTTTCGCAG GCATGATGAT TGTGATTCTC ACGAATGCTG GTGTGCA*TT
Consensus

        80
GGAAGCG
GGAAGCG

GGAAGCG

(C, D, E or ?) e

Edit 12, 13 or Consensus? 13

Edit: replace
```

```
Base(s) to replace: 68
Replace with: A
Edit: q


Stars mark mismatches.

        10         20         30         40         50         60         70
CCTTCTTTCC AGCTGGTTGT GCTTTCGCAG GCATGATGAT TGTGATTCTC ACGAATGCTG GTGTGCAATT
CCTTCTTTCC AGCTGGTTGT GCTTTCGCAG GCATGATGAT TGTGATTCTC ACGAATGCTG GTGTGCAATT

CCTTCTTTCC AGCTGGTTGT GCTTTCGCAG GCATGATGAT TGTGATTCTC ACGAATGCTG GTGTGCAATT
Consensus

        80
GGAAGCG
GGAAGCG

GGAAGCG

(C, D, E or ?) c
```

*User chooses to save the consensus sequence.*

## Joining Two Sequences

The *GEL* command JOIN is designed to meld two sequences already known to be overlapping. Small loopouts (up to three bases) are allowed. JOIN looks at the percentage of the match. The 100 percent matches are shown first, and they are ordered by length, just in case there happens to be a case where there are two matches that have exactly the same percentage. Because JOIN *requires* that the first couple of bases of the overlap match, this command does not address the problem of accounting for ambiguous or error-prone readings near the ends of the sequencing film. Thus, if a match is not found, the MERGE command should be used.

The user is shown the possible overlaps and can accept the consensus, discard it, or save it for future reference. As each possible match is shown, the user can enter the editor and modify either of the two sequences or the consensus. Upon exiting the editor, the match is updated and redisplayed. The program automatically considers first a meld between the two sequences; if that fails, then it uses the inverse complement of the second sequence. If a consensus sequence is selected, a new contig is created and automatically named and its history, including editing changes, is recorded in the PROJECT file. The original sequence data is stored in the GEL file. An interaction with the JOIN function follows:

```
GEL: j
Name of GEL on left: 3
Name of GEL on right: 4


The following match was found with an overlap of 93 and a percentage of 98:
```

```
        3    34 ..ATTAAAAAAAGATAACTRAAGAGCGTAATGGAATGAATCTTTCYTGAATATTT
        4     1              GATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT
                                        *           *           *
    * CON    34    ATTAAAAAAAGATAACTAAAGAGCGTAAT*GAATGAATCTTTCTTGAATATTT

           GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAACCANTTTAC
           GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAACCANTTTACTTGGAGGTGG...

           GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAACCANTTTAC


Action on Match? (C, E, D, or ?) e

Edit 1) 3
     2) 4
     3) Consensus

Edit? (1, 2, or 3) 1
Edit: p 60:70


        69
AATGGAATGA A

Edit: r

Base(s) to replace: 63
Replace with: C
Edit: q


The following match was found with an overlap of 93 and a percentage of 100:

        3    34 ..ATTAAAAAAAGATAACTRAAGAGCGTAATCGAATGAATCTTTCYTGAATATTT
        4     1              GATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT
                                        *                       *
    * CON    34    ATTAAAAAAAGATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT

           GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAACCANTTTAC
           GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAACCANTTTACTTGGAGGTGG...

           GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAACCANTTTAC


Action on Match? (C, E, D, or ?) c
```

## Merging New Sequences into Old Contigs

Random sequencing strategies generate sequences whose origins, adjacencies, and overlaps are not known in advance. The MERGE command in *GEL* takes as its argument the name of a gel or contig and systematically searches for homologies with every contig in the project including complementary sequences. Extensive pruning of the search is implemented by the heuristic requirement that sequence at or nearby one of the ends of the contig must match and that if the end matches then the downstream sequences must meet minimal standards of homology. Furthermore, if

the beginning location of a match means that one of the sequences would be entirely contained within the other, then both ends of the shorter sequence must meet the minimal matching requirements. As each candidate match is displayed (again in rank order of significance), the user again has the options of discarding, saving, or agreeing to the consensus or of editing the sequences. The safest strategy is to save the likeliest candidates for subsequent review before making a final selection.

Contigs, once assembled, might prove to be inappropriately melded. Since the history of each contig is accessible in the project files, they can be disassembled into the original contigs or sequences with the UNDO command and alternative melds evaluated.

An example of a melding session with MERGE follows:

```
@gel

                    IntelliGenetics
              GEL - GEL Assembler System
                  September 1, 1981

       (Copyright 1981 by the Board of Trustees, Stanford University)

Name of project: late

GEL: list

    Gels in this project are:

1.  Film 1
2.  Film 2
3.  Film 3
4.  Film 4
5.  Film 5
6.  Film 7
7.  FILM 8
8.  Film 9
9.  FiLm 10
10.  Film 11


GEL: m
Gel to be merged: 8

The following match was found with an overlap of 56 and a percentage of 100:

        8    12   GCTGGTTGTGCTTTCGCAGGCATGATGATTGTGATTCTCACGAATGCTGGTGT
 -FiLm 10     1              CTTTCGCAGGCATGATGATTGTGATTCTCACGAATGCTGGTGT

    * CON    12   GCTGGTTGTGCTTTCGCAGGCATGATGATTGTGATTCTCACGAATGCTGGTGT

          GCAATTGGAAGCG
          GCAATTGGAAGYGTYYGATATTT
                     *
          GCAATTGGAAGCGTYYGATATTT
```

```
Action on Match? (C, E, D, or ?) c
```

*The program continues to match Film 9 sequences with others but now uses the contig that has been automatically designated as "LATE-1".*

```
The following match was found with an overlap of 36 and a percentage of 94:

  Film 11    99   ATCGGGCGGACCTTCTT CCAGCTGGTTGTGCTCTCGCARRCATRR
   LATE-1     1              CCTTCTTTCCAGCTGGTTGTGCTTTCGCAGGCATGATGATTGT
                                    *                *      **    **
    * CON    99   ATCGGGCGGACCTTCTTTCCAGCTGGTTGTGCT*TCGCAGGCATGATGATTGT


Action on Match? (C, E, D, or ?) e

Edit 1) Film 11
     2) LATE-1
     3) Consensus

Edit? (1, 2, or 3) 1
Edit: r

Base(s) to replace: 131
Replace with: I
Edit: q


The following match was found with an overlap of 36 and a percentage of 97:

  Film 11    99   ATCGGGCGGACCTTCTT CCAGCTGGTTGTGCTTTCGCARRCATRR
   LATE-1     1              CCTTCTTTCCAGCTGGTTGTGCTTTCGCAGGCATGATGATTGT
                                    *                      **    **
    * CON    99   ATCGGGCGGACCTTCTTTCCAGCTGGTTGTGCTTTCGCAGGCATGATGATTGT


Action on Match? (C, E, D, or ?) c

GEL: list

    Gels in this project are:

1.  Film 1
2.  Film 2
3.  Film 3
4.  Film 4
5.  Film 5
6.  Film 7
7.  FILM 8
12.  LATE-2
        Film 11 {1 - 143}
        Film 9 {109 - 185}
        Comp-FiLm 10 {130 - 209}
```

*The names of the member sequences making up the contig named Late-2 are displayed as well as the limits of the regions they cover. The intermediate contig Late-1 has merged into contig Late-2 and is not in the list.*

```
GEL: print 12
Range (<CR> for entire GEL) :
```

```
LATE-2 from 1 to 209


        10         20         30         40         50         60         70
AGCGCTAGCT AGGGCTAGCG CTAGCGCGAT TTCGCCTCCG ATTCGCTAGG CTCCGCGCCT CGCCTTTCGC

        80         90        100        110        120        130        140
CTCTCGCGCT ATCGCGATTC GCGATTCGAT CGGGCGGACC TTCTTTCCAG CTGGTTGTGC TTTCGCAGGC

       150        160        170        180        190        200
ATGATGATTG TGATTCTCAC GAATGCTGGT GTGCAATTGG AAGCGTYYGA TATTTATATC AAACGCCGG


GEL: q

    Changes have been made to the project since it was last saved.
Do you want to save the changes? (Y or N) y
Project file LATE.PROJECT.2
  Saturday, September 5, 1981 4:36AM-PDT
```

*Editing and merging changes need not be saved, but in this case user, working diligently in the early morning, elects to save the changes.*

*At a later time user reneters the program:*

```
@gel

                        IntelliGenetics
                    GEL - GEL Assembler System
                        September 1, 1981

        (Copyright 1981 by the Board of Trustees, Stanford University)

Name of project: late

GEL: list

    Gels in this project are:

1.  Film 1
2.  Film 2
3.  Film 3
4.  Film 4
5.  Film 5
6.  Film 7
7.  FILM 8
8.  LATE-2
        Film 11 {1 - 143}
        Film 9 {109 - 185}
        Comp-FiLm 10 {130 - 209}

GEL: m
Gel to be merged: 3


The following match was found with an overlap of 90 and a percentage of 93:
```

```
       3   34   ATTAAAAAAAGATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT
  Film 4    1              GATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT

    * CON   34   ATTAAAAAAAGATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT

         GGTGGCTCTGATAAGA CGTATGTAGTAATAT  GGTATGARCANTTT AC
         GGTGGCTCTGATAAGAACGTATGTAGTAA ATATGGTATGAACCANTTTACTTGGAGGTG
                       *           *  **       * ***  *
         GGTGGCTCTGATAAGAACGTATGTAGTAATATATGGTATGAAC*ATTTTACTTGGAGGTG


Action on Match? (C, E, D, or ?) e

Edit 1) 3
     2) Film 4
     3) Consensus

Edit? (1, 2, or 3) 1
Edit: p 100:200


      109        119        129
AGACGTATGT AGTAATATGG TATGARCANT TTAC

Edit: i

Insert before base number: 115
String to insert: A
Edit: Q
```

*As user QUITS the editor, the MERGE routine automatically reevaluates the edited homology*

```
The following match was found with an overlap of 90 and a percentage of 96:

       3   34   ATTAAAAAAAGATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT
  Film 4    1              GATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT

    * CON   34   ATTAAAAAAAGATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT

         GGTGGCTCTGATAAGA CGTATGTAGTAAATATGGTATGARCANTTT AC
         GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAACCANTTTACTTGGAGGTGG
                       *                          * ***  *
         GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAAC*ATTTTACTTGGAGGTGG


Action on Match? (C, E, D, or ?) E

Edit 1) 3
     2) Film 4
     3) Consensus

Edit? (1, 2, or 3) 1
Edit: I

Insert before base number: 127
String to insert: C
Edit: Q


The following match was found with an overlap of 90 and a percentage of 98:
```

```
      3   34   ATTAAAAAAAGATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT
Film 4    1              GATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT

 * CON   34   ATTAAAAAAAGATAACTAAAGAGCGTAATCGAATGAATCTTTCTTGAATATTT

       GGTGGCTCTGATAAGA CGTATGTAGTAAATATGGTATGARCCANTTTAC
       GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAACCANTTTACTTGGAGGTGG
                      *                        *
       GGTGGCTCTGATAAGAACGTATGTAGTAAATATGGTATGAACCANTTTACTTGGAGGTGG
```

*User is satisfied that the consensus sequence is correct.*

```
Action on Match? (C, E, D, or ?) C
```

```
The following match was found with an overlap of 16 and a percentage of 93:

   Film 1  230   GAATGTTTAGATTATGTATGTTAAGAGTTT
   LATE-3    3             ATT TGT TGTTAA  ATTAAATGATTATA
                              *   *      ***  *
    * CON  230   GAATGTTTAGATTATGTATGTTAAGA*TT*AATGATTATA
```

*User doesn't think that match is very good and deletes it from consideration.*

```
Action on Match? (C, E, D, or ?) D

GEL: list

    Gels in this project are:

1.  Film 1
2.  Film 2
5.  Film 5
6.  Film 7
7.  FILM 8
8.  LATE-2
        Film 11 {1 - 143}
        Film 9 {109 - 185}
        Comp-FiLm 10 {130 - 209}

9.  LATE-3
        Film 3 {1 - 133}
        Film 4 {44 - 172}
```

*the contig formed by the merge of Film3 and Film 4 has been renamed as "LATE-3" and its member gels removed from future consideration.*

```
GEL: ed
Gel: 5
Edit: p


        10          20          30          40          50          60          70
ATCCAAGAAG GCTAAGGCCC CCAGACCTAG CGGGGGXXAG AAGAGGCGAA GACGCCGAAA GGAAAGCTAC

        80          90         100         110         120         130         140
GGAATCTACA TCTACAAGGT GCTGAAGCAG GTTCACCCTG ACACTGGCAT CTCCAGCCGT GCCATGTCCA

       150         160         170         180         190         200
TCATGAACAG TTTCGCAACG ATGTCTTCGA GCGCATCGCG CGAGGCTCCC GTCTTGCCCA CTACAAC
```

```
Edit: R
```

*For the sake of clear example, a number of easily visible editing changes are made here and the effect on the PROJECT file data will then be shown with the DISPLAY command.*

```
Base(s) to replace: 50:80
Replace with: NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Edit: I

Insert before base number: 20
String to insert: XXXXX
Edit: D

Delete: 110:140
Edit: A

Append after base number: END
String to append: GGGGGGGGGGGGGGGGGGGGGGGG
Edit: P


        10         20         30         40         50         60         70
ATCCAAGAAG GCTAAGGCCX XXXXCCCAGA CCTAGCGGGG GXXAGAAGAG GCGANNNNNN NNNNNNNNNN

        80         90        100        110        120        130        140
NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNTCTACAC ATCTCCAGCC GTGCCATGTC CATCATGAAC

       150        160        170        180        190        200        210
AGTTTCGCAA CGATGTCTTC GAGCGCATCG CGCGAGGCTC CCGTCTTGCC CACTACAACG GGGGGGGGGG

       220
GGGGGGGGGG GGGG

Edit: Q
GEL: DISPLAY

;;Master file for GEL project LATE
;;Edited by KEDES on Saturday, September 5, 1981 8:47AM-PDT
;;Edited by KEDES on Saturday, September 5, 1981 8:36AM-PDT
;;Edited by KEDES on Saturday, September 5, 1981 8:33AM-PDT
;;TopContigNum = 3
;;ContigList =  [Film 1] [Film 2] [Film 5] [Film 7]
;;[FILM 8] [LATE-2] [LATE-3]
;Read by rob
;Bam labelled Bam-RI containing H4 Gene
Film 1
ATCCTTGATGCATATACTTCTATTTTTTTTTTCGTCCGAACCGATAACACA......
.....AAGTAAAGAGCATTACCAGCATATTTATAATTGTAATACGCCAAAACAAA1
;Read by tim
;Edit Replace 50 to 80 (AGACGCCGAAAGGAAAGCTACGGAATCTACA) with NNNNNNNNNNNNN!
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
;Edit Insert  20 XXXXX
;Edit Delete 110 to 140 (AGGTGCTGAAGCAGGTTCACCCTGACACTGG)
;Edit Append after END GGGGGGGGGGGGGGGGGGGGGGGG
```

*The editing changes made above have been entered in the PROJECT file as comments. The changes appear in the sequence of Film 5.*

```
Film 5
ATCCAAGAAGGCTAAGGCCXXXXXCCCAGACCTAGCGGGGGGXXAGAAGAG
GCGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNTCTACACATCTCCAGCCGTGCCATGTCCATCATGAACAGTTTCGCAA
CGATGTCTTCGAGCGCATCGCGCGAGGCTCCCGTCTTGCCCACTACAACG
GGGGGGGGGGGGGGGGGGGGGGG1
.......
```

*The two contigs, Late-2 and Late-3, are stored in the PROJECT file with the names of their member sequences listed as a comment. The original member sequences are stored intact in a separate, pristine, GEL file.*

```
;GelList [Film 11 {1 - 143}] [Film 9 {109 - 185}]
;[Comp-FiLm 10 {130 - 209}]
LATE-2
AGCGCTAGCTAGGGCTAGCGCTAGCGCGATTTCGCCTCCGATTCGCTAGG
CTCCGCGCCTCGCCTTTCGCCTCTCGCGCTATCGCGATTCGCGATTCGAT
CGGGCGGACCTTCTTTCCAGCTGGTTGTGCTTTCGCAGGCATGATGATTG
TGATTCTCACGAATGCTGGTGTGCAATTGGAAGCGTYYGATATTTATATC
AAACGCCGG1
;GelList [Film 3 {1 - 133}]  [Film 4 {44 - 172}]
LATE-3
CTATTTGTTGTTAAATTAAATGATTATAAATTAATTAAAAAAAGATAACT
AAAGAGCGTAATCGAATGAATCTTTCTTGAATATTTGGTGGCTCTGATAA
GAACGTATGTAGTAAATATGGTATGAACCANTTTACTTGGAGGTGGTGTA
CTTGGTGACNNCCTTGGTACCC1

GEL: q

    Changes have been made to the project since it was last saved.
Do you want to save the changes? (Y or N) y
Project file LATE.PROJECT.3
  Saturday, September 5, 1981 8:48AM-PDT
```

## Ambiguous Base Assignments

We have tried to construct *GEL* so that it can deal with imprecise or ambiguous sequencing data. Our code (Table I) is somewhat simplified from that described by Staden [9]. Three problems regarding ambiguity are:

1. Base mismatches. Readings of the same fragment or of overlapping fragments may generate uncertain assignments at a particular position. If two gels are involved, the appropriate Stanford code assignment (Table I.) is made in the consensus sequence.

   If a third gel is melded into a contig that has a base mismatch ambiguity, *GEL* will, with user approval, make the appropriate assignment. For example, if the third gel has a C at a position where the contig had a G/C mismatch, *GEL* will attempt to convert the mismatch to a C.

2. Ambiguous assignments. The Stanford code provides for definite assignment of one of a pair of possibilities, for probable assignment of a base, and for an unknown base. When a meld takes place, if the ambiguous base is compatible with the new sequence, the definitive assignment is made after approval. Thus if the new gel has a T at the position that had a pyrimidine, the consensus would be that T is correct. Similarly, if the new gel had a purine at a position occupied by a G/C mismatch, the user would be asked to

TABLE I

AMBIGUOUS AND TENTATIVE SEQUENCE ASSIGNMENTS

| SYMBOL | DEFINITION |
|---|---|
| C, T, A, G | as usual |
| 3, 4, 5, 6 | Probably C, T, A, and G respectively |
| 7, 8, 9, 0 | Maybe C, T, A, and G respectively |
| P or R | A or G |
| Q or Y | C or T |
| J | C or A |
| K | T or G |
| L | T or A |
| M | C or G |
| N | something there, but identity unknown |
| ? | maybe something there |

consider a G as being correct.

Probable assignments carry less force and only become considered as certain when melded with a certainty or when matched exactly with another probable assignment of the same type.

3. Possible (extra) assignments. Often it may be difficult to decide on a sequencing gel whether there is or is not a base present at a certain position, particularly in a homopolymeric stretch, e.g are there one or 2 T's there. The Stanford code allows for reserving a potential space for such a base. The user would assign a certain T and a "maybe" T. At MERGE time the user will be asked to consider dropping the uncertain base if the second gel has no such assignment.

The ambiguous assignments are carried permanently in the history file on each sequence. In the case that an ambiguity is not resolved, the contig will continue to carry the appropriate Stanford code symbol. The researcher will have to decide whether the information to be gained by resequencing in order to remove the ambiguity would be worth the effort.

## Concluding Comments

*GEL* is capable of handling the often overwhelming task of housekeeping on a DNA sequencing project. It will not eliminate paper and pencil and careful annotation, but it does provide an automated milieu for ensuring that all required information is recorded, easily accessible, and legible. Users do not have to worry about the orientation of sequences since *GEL* automatically consideres merges and joins of both the sequence and its complement. Keeping detailed histories of all sequences entered is currently allowing us to develop a sequencing strategy map generator that will show in simple graphics each sequence, the end at which it was labelled or from which it was extended, and its orientation relative to the final contig map.

One concern that still eludes our capability is the problem of internally repetitive sequences. When

one sequence ends in the middle of a repetitive stretch and another sequence begins in the middle of that stretch, long exact overlaps will be found that may be incorrectly situated. Users must be aware of this problem and make merges in such regions only with caution.

The project files of *GEL* are written in the MOLGEN *SEQ* format. This means that as they are updated the new information and the new contigs can be analyzed by *SEQ* [8], the DNA sequencing advisor, *MAXAMIZE* [10], and the simulation system, *GENESIS* [11].

## Acknowledgments

This work is a part of the MOLGEN project, a joint research effort among the Departments of Computer Science, Medicine, and Biochemistry at Stanford University. The research has been supported under NSF grant MCS80-16247. Coputational resources have been provided by the SUMEX-AIM National Biomedical Research Resource, NIH grant RR-00785-08, and by the Department of Computer Science.

We are extremely grateful to our many enthusiastic collaborators in the MOLGEN project, and in particular to Doug Brutlag and Peter Friedland.

*MOLGEN*, *GEL*, and *SEQ* are registered trademarks of the Board of Trustees of Stanford University. *GENESIS* is a registered trademark of IntelliGenetics Inc.

Address all correspondence to: Dr.L.H.Kedes, 151M, VA Hospital, Miranda Avenue, Palo Alto, CA 94304, USA

## References

1. Maxam, A.H., and Gilbert, W., Proc. Nat. Acad. Sci. USA, Vol. 74, 1977, pp. 560-574.
2. Sanger, F., Nicklen, S. and Coulson, A.R., Proc. Nat. Acad. Sci. USA, Vol. 74, 1977, pp. 5463-5467.
3. Heidecker, G., Messing, J. and Gronenborn, B., Gene, Vol. 10, 1980, pp. 69-73.
4. Messing, J., Crea, R. and Seeburg, P.H., Nuc. Acids Res., Vol. 9, 1981, pp. 309-321.
5. Anderson, S., Nuc. Acids Res., Vol. 9, 1981, pp. 3015-3027.
6. Gingeras, T.R., Milazzo, T.J.P., Sciaky, D. and Roberts, R.J., Nuc. Acids Res. , Vol. 7, 1979, pp. 529- 545.
7. Staden, R., Nuc. Acids Res., Vol. 8, 1980, pp. 3673-3694.
8. Brutlag, D. L., Clayton, J., Friedland, P., and Kedes, L. H., Submitted to Nucleic Acids Research
9. Staden, R., Nuc. Acids Res., Vol. 6, 1979, pp. 2601-2610.
10. Bach, R., Friedland, P., Brutlag, D. L., and Kedes, L. H., Submitted to Nucleic Acids Research
11. Friedland, P., Kedes, L., Brutlag, D., Iwasaki, Y. and Bach, R., Submitted to Nucleic Acids Research