Volume 10 Number 1 1982 Nucleic Acids Research

MAXAMIZE. A DNA sequencing strategy advisor

Rene Bach[1], Peter Friedland[1], Douglas L.Brutlag[2] and Larry Kedes[3]

Departments of [1]Computer Science, [2]Biochemistry, and [3]Medicine, Stanford University, Stanford, CA 94305, USA

Received 15 September 1981

Abstract

The *MAXAMIZE* advisory system determines from user-provided restriction maps an optimal strategy to do nucleotide sequencing by methods involving end-labeled fragments. The maps may be either simple linear restriction maps of fragments or complex circular maps including restriction sites of a vector. The whole system is interactive and is written in the Genetic English language provided by the *GENESIS* System, a molecular genetics knowledge representation and manipulation package. In addition, *MAXAMIZE* provides bookkeeping facilities for sequencing and offers advice on how to verify the newly obtained sequence data.

Introduction

Using the knowledge representation and manipulation tools developed by the MOLGEN project [1], we have developed a system which provides assistance in determining a nucleotide sequence using a given restriction map. The problem is to derive an efficient sequencing strategy which minimizes the number of gels run and sequencing reactions performed and maximizes the number of nucleotides read in each step. Starting from the restriction map, the sequencing advisor determines how much sequence information can be obtained considering both the average length the user can sequence from any labeled terminus and whether the fragments can be separated. Having determined the best order of successive digests, the system predicts for each digest the pattern of fragments (on a gel) and advises which of them have to be eluted and sequenced.

In addition, the sequencing advisor provides the user with bookkeeping capacities which include the ability to continually update the restriction-map with the newly gathered sequence information. Also, the user is advised:

1. how to verify the experimentally determined sequence
2. how to map RNA ends (5', 3' or both) using the nuclease protection technique [2]
3. which additional enzymes might be good cutter candidates in a region where there are no sites for enzymes already tested.

This advisor is most useful for directed sequencing performed using the Maxam and Gilbert method [3] [4]. See [5] for the description of a system which assists in providing bookkeeping for both site specific and random sequencing experiments.

Method of Solution

The sequencing advisor is written in a subset of Genetic English (*Genglish*) provided by *GENESIS* [1]. *GENESIS* is based on the Unit System [6] [7] [8] [9], a general-purpose knowledge acquisition program written in Interlisp, which runs on the Digital Equipment Corporation DecSystem 10 and 20 series of computers. *GENESIS* provides the ability to represent, store and modify molecular genetic information such as restriction maps, sequences, and restriction enzyme properties, as well as more general types of information like numbers, strings, lists, and tables. The total collection of information is known as a knowledge base, and may be easily examined, shared, and updated by a variety of users.

Furthermore, the system allows manipulation of the knowledge base by using *Genglish*, the Genetic English language. This language allows a non-programmer molecular biologist to construct sophisticated computational systems that embody domain-specific expertise. For the case of the system discussed in this paper, all developmental work was performed by one of the authors (R.B.), a molecular biologist with essentially no programming experience.

Sequencing Experiment Description

The scientist describes his problem by providing *MAXAMIZE* with a restriction map constructed with the *GENESIS* map editor. The restriction map can describe one of two types of DNA structures: either just the DNA of interest, or a vector containing that DNA. In the latter case, the user has to indicate which region is the inserted DNA by marking it as a specific region on the restriction map.

MAXAMIZE first establishes the list of restriction enzymes cutting within the inserted DNA fragment, (or, by default, the whole molecule). Editing facilities allow modification of that list by the user, for example, to avoid using a particular enzyme because he had temporarily exhausted his supply.

The user is next asked to provide several experimental parameters:

- the average number of bases the user is able to read off a gel
- the minimum size difference of two fragments which allows purification
- the coordinates of the region(s) to be sequenced
- the names of the restriction enzymes (from the list discussed above) which are relevant to the sequencing strategy problem.

An example of the experiment description phase is presented below. All user responses are shown in underlined letters; we have added comments in *italics*. <CR> stands for carriage-return.

```
ARE YOU FAMILIAR WITH MAXAMIZE? :  N

MAXAMIZE will determine an optimal strategy to sequence using any  defined
restriction-map. The restriction-map is described by using the GENESIS map
editor.  If you are providing  a  vector map  which  includes  an inserted
DNA sequence, then you should  define a region named INSERT  on  the  map.
```

```
If you want to define more than one INSERT, label them  INSERT1,  INSERT2,
etc..

WHAT KIND OF OUTPUT WOULD YOU LIKE :
C   : FOR COMPLETE (VERY DETAILED) OUTPUT
D   : FOR DESCRIPTIVE OUTPUT
S   : FOR SHORT OUTPUT  (JUST MAPS)
DEB : FOR DEBUGGING OUTPUT
:   C

DO YOU WANT TO :
C    : CREATE A NEW RESTRICTION MAP
E    : EDIT THE CLONE5 MAP
U    : USE A PREVIOUSLY DEFINED MAP
<CR> : CONTINUE, USING THE CLONE5 MAP
```

The user is given the option of creating a new map or working with previously defined ones. The system keeps track of the last map with which it worked, in this case CLONE5, and allows the user to modify that map or simply continue its analysis with that map.

```
WHAT IS YOUR CHOICE ? :   E

Map Editor
MapEd: length 3000
MapEd: circular
MapEd: region insert 450 2050 <ins ins>
```

The above specifies that a region called INSERT be marked on the map from 450 to 2050 with the symbols <INS and INS> indicating the position of the region on the map.

```
MapEd: enter ecor1 0 ecor1 800 ecor1 1300 ecor1 1900 sal1 450 sal1 2050
MapEd: enter bamh1 300 bamh1 600 bamh1 1500
```

The above commands specify the location of restriction sites on the map.

```
MapEd: print

Circular
Length 3000 base pairs

INSERT region from 450 TO 2050 indicated by <INS and INS>

      300    600                  1500         2050
 0       450     800         1300          1900
|--------<INS-------------------------------INS>---------------------|
0↑    ↑  ↑   ↑   ↑           ↑    ↑        ↑  ↑                      3000
 ECOR1   SAL1    ECOR1       ECOR1          ECOR1
      BAMH1  BAMH1                BAMH1        SAL1

MapEd:  done


YOU MAY NOW PROCEED WITH ONE OF SEVERAL OPTIONS:

 S :    START A SEQUENCING ANALYSIS
CO :    CONTINUE AN EXISTING ANALYSIS WITH NEW INFORMATION
M3 :    MAP THE 3' END OF AN RNA (IF YOU KNOW THE APPROXIMATE LOCATION)
        USING THE S1 NUCLEASE TECHNIQUE
M5 :    MAP THE 5' END OF AN RNA
MB :    FOR BOTH OF THE MAPPING OPTIONS
```

```
CA :    CHECK FOR ABUTTED FRAGMENTS
```

MAXAMIZE has several auxiliary features, such as RNA mapping and site uniqueness checking. This latter feature searches for abutted fragments that have been sequenced in a head-to-head manner and whose connecting restriction site has not been sequenced across. When this option is selected, MAXAMIZE tries to find the simplest experimental path to selecting a fragment that will contain the site in a position that can be easily reached by sequencing.

```
WHAT IS YOUR CHOICE? S

DEFINE THE REGION(S) TO BE SEQUENCED IN CLONE5
THESE ARE DEFAULT VALUES, OR VALUES PREVIOUSLY DEFINED :

HOLE-START      HOLE-END

    500            2K
```

The default coordinates represent that portion of the non-vector part of the map that remains to be sequenced (the "hole" remaining in the map). In this case the "hole" is the entire INSERT region.

```
YOU WILL NOW BE ASKED TO CONFIRM OR CORRECT EACH VALUE.
TYPE <CR> IF THESE VALUES ARE CORRECT,
K (FOR KILL) IF NO VALUES ARE CORRECT,
C (FOR CHANGE) IF SOME VALUES ARE CORRECT.
:  K

YOU MUST NOW DEFINE THE REGIONS YOU WISH TO SEQUENCE.  FOR EACH REGION
PROVIDE A START AND END.  WHEN YOU ARE FINISHED, TYPE AN EXTRA <CR>.

START:  450
END: 2060

START: <CR>

HOLE-START      HOLE-END

    450            2060
```

Now the system asks the user to estimate first, the number of bases he hopes to read from a sequencing gel autoradiograph, and second, the minimum percent length difference between two DNA fragments required before they can be separated by gel electrophoresis.

```
TYPE THE NUMBER OF BASES YOU CAN READ OFF A GEL (THE DEFAULT IS 300)
:

TYPE THE SIZE DIFFERENCE FOR SEPARATION OF TWO FRAGMENTS
(THE DEFAULT IS 2.0%)
: <CR>
```

The user picked the default values for both reading length and size difference for separation.

Sequencing Analysis Phase

MAXAMIZE first calculates the fragment sizes which would be generated by all restriction enzyme single or double digests and checks to see if any fragments would co-migrate (and therefore not be

separable) under the experimental criterium provided. For each single or double digest, the system provides a table giving the name of the enzymes, the size of the fragments (sorted by decreasing size), the coordinates of the sites, and whether the fragment can be purified (if the map includes the vector then all vector fragments are considered as well).

MAXAMIZE then determines the maximum number of nucleotides that could be read with single and double digests using strand-separation and double digests using secondary restriction enzyme cleavage. In determining this number, *MAXAMIZE* takes into account the length of each fragment and does not count contributions by fragments which cannot be separated. The highest ranking digest will be the one that provides for the greatest number of bases read.

Next, the system indicates for each digest what the expected fragment separation gel should look like in terms of relative mobilities. In addition, the user is advised which fragment(s) must be eluted and sequenced, since the map may include vector fragments as well. Any discrepancy between the experimentally determined pattern of the gel and the predicted fragment pattern may serve to indicate a mapping error and would be a useful check on the accuracy of the starting map.

Currently, *MAXAMIZE* assumes that the selected regions are capable of being sequenced, marks them as such on the sequencing map of the structure, and then recursively attempts to provide advice about how best to sequence the remaining unknown regions of the map. In the near future, the system will allow the user to interrupt this process to indicate experimental failure to carry out a projected sequencing step. Then *MAXAMIZE* will suggest the next best route as an alternate.

An example of the analysis phase of the sequencing advisor follows.

```
THESE ARE THE ENZYMES CUTTING IN THE INSERT REGION OF THE RESTRICTION-MAP:
SAL1 BAMH1 ECOR1

DO YOU WANT TO EDIT THIS LIST ? :  N
```

The user was given the option of pruning restriction enzymes from the list of known cutters.

The following are typical examples of those tables that the user would see. These tables contain single and double enzyme digest data that are determined and used by MAXAMIZE. The tables include cutting site locations, fragment lengths, and a determination of whether the fragment can be separated by gel electrophoresis. It should be noted that circular sequences are handled properly.

ENZYME	SITE	NEXT-SITE	FRAGSIZE	PURIFIABLE
SAL1	450	2050	1600	YES
SAL1	2050	450	1400	YES

ENZYME1	ENZYME2	SITE1	SITE2	FRAGSIZE	PURIFIABLE
SAL1	ECOR1	2050	0	950	YES
ECOR1	ECOR1	1300	1900	600	YES
ECOR1	ECOR1	800	1300	500	YES
ECOR1	SAL1	0	450	450	YES
SAL1	ECOR1	450	800	350	YES
ECOR1	SAL1	1900	2050	150	YES

The next class of tables shown to the user deals with predictions about possible strategy outcomes as if strand separation methods will be used in the sequencing experiments. The two typical tables shown here display the number of labelled fragments within the regions of interest (#-FRAGMENTS); the total number of bases pairs that can be read for each digest (TOTAL-READ); the number of overlapping, complementary nucleotides read on both strands for each digest (OVERLAP); the difference between TOTAL-READ and OVERLAP (BP-READ); and the number of non-separable fragments (#-NONPURIF). The table is ordered by the TOTAL-READ column. The first table assumes that ALL fragments can be separated while the second table makes use of the information about the percent length difference required to separate fragments of similar length. The second table is the one used for further analysis by the system. If the BP-READ values in the first table are very much greater than those in the second table, the user might decide to attempt to separate fragments of similar length. MAXAMIZE assumes that all fragments can be strand-separated. In the examples shown, the best values in the two tables are similar.

STRAND SEPARATION RESULTS :

THIS TABLE ASSUMES ALL FRAGMENTS ARE SEPARABLE

ENZYME1	ENZYME2	#-FRAGMENTS	BP-READ	OVERLAP	#-NONPURIF	TOTAL-READ
SAL1	ECOR1	4	1100	500	0	1600
ECOR1	BAMH1	6	750	850	0	1600
ECOR1		4	1450	100	0	1550
SAL1	BAMH1	3	1100	200	0	1300
BAMH1		3	900	150	0	1050
SAL1		1	600	0	0	600

THIS TABLE MAKES USE OF SIZE DIFFERENCE PERCENTAGE IN DETERMINING NON-SEPARABLE FRAGMENTS

ENZYME1	ENZYME2	#-FRAGMENTS	BP-READ	OVERLAP	#-NONPURIF	TOTAL-READ
SAL1	ECOR1	4	1100	500	0	1600
ECOR1		4	1450	100	0	1550
SAL1	BAMH1	3	1100	50	1	1150
BAMH1		3	1050	0	0	1050
ECOR1	BAMH1	6	750	300	3	1050
SAL1		1	600	0	0	600

The next two tables provide analogous information for each digestion by each pair of enzymes as if the secondary cleavage method rather than the strand separation method will be used in the sequencing experiment. The values in the column ENZYME1 represent the restriction enzyme site at the radiolabeled end of the DNA fragment. The values in the column #-2-LABEL are the numbers of fragments NOT cut by ENZYME2. Such fragments are therefore not usable by a secondary cleavage method. The fragment sets in the categories #-NON-PURIF and #-2-LABEL are not necessarily overlapping.

```
SECONDARY-CLEAVAGE RESULTS :
**************************

THIS TABLE ASSUMES ALL FRAGMENTS ARE SEPARABLE

ENZYME1      ENZYME2    #-FRAGMENTS    TOTAL-READ    #-NONPURIF    #-2-LABEL

BAMH1        ECOR1          4             700            0            1
ECOR1        BAMH1          6             700            0            2
SAL1         BAMH1          2             450            0            0
BAMH1        SAL1           3             450            0            1
SAL1         ECOR1          2             450            0            0
ECOR1        SAL1           4             450            0            1

THIS TABLE MAKES USE OF SIZE DIFFERENCE PERCENTAGE IN DETERMINING
NON-SEPARABLE FRAGMENTS
ENZYME1      ENZYME2    #-FRAGMENTS    TOTAL-READ    #-NONPURIF    #-2-LABEL

SAL1         ECOR1          2             450            0            0
ECOR1        SAL1           4             450            1            2
SAL1         BAMH1          2             300            1            0
BAMH1        SAL1           3             300            1            1
BAMH1        ECOR1          4             300            3            1
ECOR1        BAMHI          5             300            3            2
```

The information shown in the tables above is now used to explain more clearly each single and double digest result. Typical examples of such explanations follow.

```
STRAND-SEPARATION RESULTS :

ECOR1 - SAL1 DIGEST.
USING 8 SEQUENCING REACTIONS :
1600 BASES TOTAL CAN BE READ
1100 BASES ARE READ ONCE
 500 BASES ARE READ TWICE

...

SECONDARY-CLEAVAGE RESULTS :

LABEL SAL1 - ECOR1 DIGEST.
USING 2 SEQUENCING REACTIONS :
450 BASES TOTAL CAN BE READ
450 BASES ARE READ ONCE

...
```

The user is now given his choice of sequencing strategies.

```
WHAT KIND OF SEQUENCING STRATEGY DO YOU PREFER
STRAND-SEPARATION (S), SECONDARY-CLEAVAGE (C) OR BOTH (B)
: C
```

At this point MAXAMIZE simulates the selected sequencing experiment path. First it displays the best primary digest, invokes a radiolabeling step and suggests which enzyme should be used to cut the labeled fragments. No intermediate purification step is assumed. The program does not discard

useful fragments even if their length is shorter than those that can be read in the users laboratory. Such fragments are displayed next. Then the program submits a prediction of the radiolabeled fragment lengths useful in interpreting fragment separation gels and autoradiographs. MAXAMIZE makes a judgment about each radiolabeled fragment and advises the user whether that fragment should or should not be sequenced. A fragments is not a candidate for sequencing either when it maps outside of the region of interest (e.g. when a fragment maps entirely within the vector or has already been sequenced) or it CANNOT BE SEQUENCED *(i.e. when the fragment cannot be cut by a second enzyme or cannot be separated from another fragment by gel electrophoresis).*

```
                DIGEST SAL1  (LABELED) CUT WITH   ECOR1
                =======================================
SHORT FRAGMENTS ARE USEFUL :

LABELED-SITE CUTTER FRAGSIZE  SITE1   READ-TO
  SAL1        ECOR1   150     2050    1900

LEFT-END  RIGHT-END   FRAGSIZE   *-SITE    READ-TO    COMMENT

SAL1       ECOR1       950        2050      2350       NOT TO BE SEQUENCED
ECOR1      SAL1        450        450       150        NOT TO BE SEQUENCED
SAL1       ECOR1       350        450       750        TO BE SEQUENCED
ECOR1      SAL1        150        2050      1900       TO BE SEQUENCED

THE REGIONS TO BE SEQUENCED :

START  and  END   at :

 450       2050

 AFTER THIS EXPERIMENT, THE REGIONS REMAINING TO BE SEQUENCED

START  and  END   at :

 750       1900

AS SHOWN BY THE FOLLOWING MAP :

Circular
Length 3000 base pairs

INSERT region from 450 TO 2050 indicated by <INS and INS>
SHORT READING region from 1900 TO 2050 indicated by <.
SEQUENCED region from 450 TO 750 indicated by >.

      300    600                 1500         2050
 0       450     800          1300          1900
|-------->>>>>>>>>-------------------------<<<<<----------------------|
0↑    ↑  ↑   ↑  ↑           ↑    ↑        ↑   ↑                       3000
 ECOR1   SAL1    ECOR1       ECOR1         ECOR1
      BAMH1  BAMH1                BAMH1        SAL1
```

The maps show the progress of the sequencing strategy in a very visual way. *MAXAMIZE* continues to develop tables and maps for either of the methods and each of the digests. Each digest, in turn,

will fill in the restriction map with symbols and indicate what regions are expected to be sequenced after that particular digest experiment is performed. The program stops when all regions have been sequenced or when no more sites are available to fill in the holes.

At that stage, the system gives the user some choices. He may actually go to the laboratory bench and begin sequencing. In that case the sequence information will be used by *MAXAMIZE* to generate a cutting frequency table for each restriction enzyme site discovered in the newly sequenced region. This table will be then be used by the program to suggest alternate sequencing strategies helpful for filling in the remaining holes.

Discussion

This paper describes a system which uses the techniques of symbolic computation and artificial intelligence to provide assistance in the design of a class of laboratory experiments in molecular biology. *MAXAMIZE* does not provide the initial criteria for selection of restriction enzymes which will be used to make the initial map; this choice tends to be one of individual laboratory preference. It does provide considerable guidance in the management of a problem which becomes combinatorially awkward for all but very simple restriction maps. It also provides considerable information for monitoring and verifying the progress of sequencing experiments.

The Unit System is a powerful and versatile knowledge base development and manipulation system which has been applied to a variety of problems both within molecular biology (*GENESIS* [1]) and in other scientific domains. It can serve as an "intelligent encyclopedia" or a teaching tool in addition to the problem-solving functions that *MAXAMIZE* typifies. The Genetic English language in *GENESIS*, (*Genglish*), helps to free the domain expert, the molecular geneticist, from the effort of teaching a domain-ignorant computer scientist the details of his area of expertise. In addition, it allows continual modification and improvement of problem-solving systems without the presence of the original system designer.

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. Computational resources have been provided by the SUMEX-AIM National Biomedical Research Resource, NIH grant RR-00785-08, and by the Department of Computer Science.

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

References

1. Friedland, P., Kedes, L., Bruttag, D., Iwasaki, Y. and Bach, R., "Genesis, a Knowledge-Based Genetic Engineering Simulation System for Representation of Genetic Data and Experiment Planning", Submitted to *Nucleic Acids Res.*
2. Berk, A.J. and Sharp, P.A., *Cell*, Vol. 12, 1977, pp. 721-732.
3. Maxam, A.H., and Gilbert, W., *Proc. Nat. Acad. Sci. USA*, Vol. 74, 1977, pp. 560-574.
4. Maxam, A.M. and Gilbert, W., "Sequencing end-labeled DNA with base specific cleavage," *Methods in Enzymology*, L. Grossman and K. Moldave eds., Academic Press, 1980, pp. 499-560.
5. Clayton, J. and Kedes, L., "Gel, a DNA Sequencing Project Management System", Submitted to *Nucleic Acids Res.*
6. Friedland, P., "Knowledge-Based Experiment Design in Molecular Genetics," Computer Science Department Report CS-79-771, Stanford, October 1979.
7. Stefik, M. J., "An Examination of a Frame-Structured Representation System," Proceedings of the Sixth International Joint Conference on Artificial Intelligence, IJCAI, 1979, pp. 845-852.
8. Smith, R. G. and Friedland, P., "A User's Guide to the Unit System," Heuristic Programming Project Memo HPP-80- 28, Stanford, December 1980.
9. Friedland, P., "Acquisition of Procedural Knowledge from Domain Experts," Proceedings of the Seventh International Joint Conference on Artificial Intelligence, IJCAI, 1981, pp. 856-861.