Methylation blockage and other improvements to a comprehensive DNA analysis program

Terry W. Stone¹ and Kathleen N. Potter²

Alberta Research Council, ¹Oil Sands Research Department, and ²Biotechnology Department, 11315-87 Ave., Edmonton, Alberta, T6G 2C2, Canada

Received 17 July 1985

ABSTRACT

A comprehensive DNA analysis computer program was described in the second special issue of Nucleic Acids Research on the applications of computers to research on nucleic acids by Stone and Potter (1). Criteria used in designing the program were user friendliness, ability to handle large DNA sequences, low storage requirement, migratability to other computers and comprehensive analysis capability. The program has been used extensively in an industrial-research environment. This paper talks about improvements to that program. These improvements include testing for methylation blockage of restriction enzyme recognition sites, homology analysis, RNA folding analysis, integration of a large DNA database (GenBank), a site specific mutagenesis analysis, a protein database and protein searching programs. The original design of the DNA analysis program using a command executive from which any analytical programs can be called, has proven to be extremely versatile in integrating both developed and outside programs to the file management system employed.

INTRODUCTION

A laboratory environment in which DNA-RNA manipulation is carried out is incomplete without state-of-the-art computer assistance. This assistance often involves searching large DNA or RNA strings for either a single substring or groups of substrings. The most common occurrence of this type of search is the search for all occurrences of enzyme specificities. However, the user should not be restricted by the number of substrings, such as enzymes, searched for. Nor should the user be restricted by the ability to search for just one category of substrings like enzymes.

A computer program, called SEQ, designed with user friendliness and

1 This program may be obtained by writing the authors. No charge is asked for the software but users are asked to reference the authors if their software is used in publications. Some of the integrated analysis programs are proprietary (folding analysis and site specific mutagenesis analysis) and permission would have to be gained from those authors before we could release their software. See also the discussion below on availability of software.
complete searching versatility was described by Stone and Potter (1). A basic DNA-RNA sequence management system was set up with the ability to store and retrieve arbitrary numbers of sequences. These sequences were filed in user specified categories. No limit was placed on number of sequences stored within a category or number of categories. Two types of sequences could be stored. One was long sequences, called SEQ sequences, with a current limit on size of 100,000 base pairs, and the other was shorter sequences called SITE sequences with a size limit of 100 base pairs. A user could search any one SEQ sequence within a certain category with any single or group of SITE sequences from another category. An example of this is a complete enzymes recognition search. The ability to create, combine, delete and insert new sets of base pairs easily was also included. This versatility has worked well in the industrial-research environment for which it was designed.

Within this basic file management system was included various analysis programs. The user could call any SEQ or SITE sequence stored in any category and then subject it to certain types of analysis. Included was the ability to analyse for direct and inverted repeats, display all reading frames of amino acid sequences for both the input sequence and reverse complement strand and prediction of mapping order from the results of two single and one double enzyme digest.

The analytical ability of the program SEQ has been enhanced considerably in the intervening year. Much of the enhanced ability derives from integrating state-of-the-art programs that have been published in the literature. These include RNA folding analysis, homology analysis and site specific mutagenesis analysis. A DNA database (GenBank) and the NBRF protein database have been integrated. Finally, the ability to detect methylation blockage of restriction enzyme cut sites has been developed and integrated.

DESCRIPTION OF THE PROGRAM

Previous Capability

A basic criterion in the original design of the program SEQ was user friendliness. To this end, many features are included to facilitate usage. An executive section asks the user for any of 13 commands. Extensive on-line help is available on all commands. Within each command section, questions are asked of the user and only known responses are accepted with prompting upon entering an erroneous response. The user can back up within a section if a previous response needs to be changed. If the user is unsure of a response, for example the user cannot remember what categories are available
or what sequences are available within those categories, the user is told what is available. Input is accepted in upper or lower case. A small character editor allows the user a high degree of flexibility for creating or changing sequences. Sequences from any categories can be combined and changed at will.

The executive section of the program SEQ currently makes use of the following commands:

HELP
BUILD
FILES
SEARCH
REPEATS
AMINO
DIGEST
BACKUP
GENBANK
FOLD
HOMOLOGY
SNIPPER
PROTEIN

The commands up to BACKUP were described in Reference (1) and will be described only briefly here for convenience.

HELP provides detailed on-line help for all of the commands, including HELP itself. HELP on HELP describes the basics of the program SEQ including any internal limits. Help on any other command describes what the command section does and any limitations imposed. This feature lets new users become acquainted with the program without outside assistance and is helpful when a program is being run that was programmed by another programmer. The user is warned of differences in style that may be present (although when an outside program is integrated, operating strategy is changed as much as possible to conform to the strategy employed by SEQ).

BUILD is a small program for creating, combining, inserting and deleting base pair strings from any category. When creating new strings, any number of base pairs can be entered on one line at the terminal. The program checks for any errors in the input line, and will only accept base pairs entered up to the error. A star, '*', entered at any time will cause a printout of all basepairs entered up to that time after which entry is resumed. Sequences obtained from a database (GenBank) or from any other category can also be
Reconstruction of a Gene from Two Single and One Double Digest

<table>
<thead>
<tr>
<th>double digest</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18</td>
<td>17</td>
<td>18</td>
<td>25</td>
<td>27</td>
<td>34</td>
<td>154</td>
<td>211</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>First Digest</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15</td>
<td>17</td>
<td>238</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Second Digest</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18</td>
<td>34</td>
<td>44</td>
<td>170</td>
<td>238</td>
</tr>
</tbody>
</table>

Results of Gene Reconstruction

<table>
<thead>
<tr>
<th>Compatible Set 1</th>
<th>Order uncertain:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>* *</td>
</tr>
<tr>
<td>First Digest</td>
<td>1 1 1 1 3 3 2 2 1</td>
</tr>
<tr>
<td>Double Digest</td>
<td>1 3 6 2 5 8 4 7 1</td>
</tr>
<tr>
<td>Second Digest</td>
<td>4 1 2 3 3 5 5 4 4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Compatible Set 1, reverse direction</th>
<th>Order uncertain:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>* *</td>
</tr>
<tr>
<td>First Digest</td>
<td>1 2 2 3 3 1 1 1 1</td>
</tr>
<tr>
<td>Double Digest</td>
<td>1 7 4 8 5 2 6 3 1</td>
</tr>
<tr>
<td>Second Digest</td>
<td>4 4 5 5 3 3 2 1 4</td>
</tr>
</tbody>
</table>

Note: Only adjacent starred fragments in the double digest need be interchanged for other possible solutions.

Figure 1. Results of a gene reconstruction from two single and one double digest.

altered by combining, inserting and deleting.

FILES allows the user to examine the contents of any category of sequences, print out a particular sequence from that category or delete a sequence from that category.

SEARCH lets the user search any SEQ sequence from any category with either a single member or a whole category of SITE sequences. The most common usage of this command is an enzyme recognition search with enzymes from the category 'Enzymes'. With the use of the BUILD command, new enzymes or other SITE sequences in SITE categories can be updated at will. Currently, there are over two hundred enzymes in the 'Enzymes' category.

REPEATS searches for direct and inverted repeats. The program for detecting inverted repeats was obtained from a set of programs from the Cold Spring Harbour laboratory by Gingeras, Rice and Roberts (2). This search for
inverted repeats has, however, been partially superseded by the folding analysis program described below.

AMINO employs two programs from the Cold Spring Harbour set of programs by Gingeras, Rice and Roberts (2). One of the programs presents an amino acid reading frame display for a quick overview of the entire molecule and the other program translates any portion of that DNA-RNA sequence in any reading frame for both the input and reverse complement sequences.

DIGEST uses the results of two single enzyme digests and a double digest using both of these enzymes. It then predicts the mapping order of the restriction fragments of circular or linear plasmid molecules. The technique is based on the approach of Fitch, Smith and Ralph (3). An example of a reconstruction is shown in Figure 1.

The BACKUP command will copy all SEQ or SITE categories to a different disk directory. This provides added protection against inadvertent changes to a sequence.

**New Capability**

The National Institute of Health's DNA database called GenBank (4) has been integrated with SEQ. This database comes with a computer file called a SHORT DIRECTORY which contains all sequences contained in the database together with a few word description of the sequence. Also included with the database is a computer file containing a list of all authors contributing sequences to the database together with the sequences that they contributed. A command in our SEQ program called GENBANK will search either the SHORT DIRECTORY for any key word associated with a sequence, or will search for any author name. The user can then request that the sequence associated with the key word or the author's name be copied into the SEQ sequence management system as described above. This database currently has over 4.3 million base pairs of DNA sequences.
RNA folding analysis is carried out with a command called FOLD. A program developed by Zuker and Stiegler (5) at the National Research Council of Canada folds an RNA molecule by finding a configuration of minimum free energy using published values of stacking and destabilizing energies. Folds of molecules that are approximately 1000 base pairs long take roughly 1.5 cpu hours on a VAX 11/780 computer. Figure 2 shows an example of a folded molecule predicted with FOLD.

Primary structure homology between two sequences may be examined with the HOMOLOGY command. Programs by Novotny (6) (see also Novotny and Auffray (7)) to analyze homology of letter strings (nucleotide or amino acid sequences) and to display the result in the form of a dot matrix have been integrated with SEQ. The use of colour has also been programmed for use with DNA or RNA strings. Assigning a different colour to each of the four base pairs allows the user an extra degree of capability in understanding the homology of two strings. An example of a dot-matrix homology scan is shown in Figure 3.

A site specific mutagenesis analysis is accomplished with the command SNIPPER. SNIPPER is a program by Sprang (8) that reads a DNA sequence and attempts to generate additional restriction sites by silent point mutations. It accesses the enzymes contained within the SEQ 'Enzymes' category. An example of this analysis on a DNA string is shown in Figure 4.
Figure 4. Results of a site specific mutagenesis analysis.

A complete protein sequence database can be called from SEQ with the command PROTEIN, but shares no interface with it. It is the Protein Sequence Database of the Protein Identification Resource (supported by the Division of Research Resources of the National Institute of Health) (9). This is a stand-alone database system that is capable of construction of a peptide from segments of a sequence, back and forward translation, search of a translation for restriction enzyme cut sites and computation and display of amino acid composition tables, as well as other non-analytical features. The file structure is different than that employed in SEQ. It is intended that parts of this database be incorporated with SEQ in the future as the amino acid capabilities of SEQ are increased.
The ability to detect methylation blockage of restriction enzyme recognition sites has been added to SEQ. This has been accomplished as follows. After an enzyme search of a molecule has been completed, if the user desires detection of methylation blockage with all known methylases, the program then scans a file of known methylases and the sequences they are known to block. If the cut site is potentially blocked, it is noted at the terminal and in the output file. Data for this file comes from Nelson, Christ and Schildkraut (10) and McClelland and Nelson (11). Figure 5 shows an example of predicted blockages of restriction enzyme cuts.

**AVAILABILITY OF SOFTWARE**

As noted in the footnote of the first page, most of the software described is available with few restrictions except that the source be acknowledged wherever used. However, for the FOLD and SNIPPER programs, those authors have requested direct contact before their software can be distributed. In order that SEQ be released with the FOLD software intact, please contact

Dr. Michael Zuker,
Division of Biological Sciences
National Research Council of Canada
Ottawa, Ontario, Canada
K1A 0R6.
tel. (613) 992-4182
In order that SEQ be released with the SNIPPER software intact, please contact

Dr. Steve Sprang,
Department of Biophysics and Biochemistry
University of California, San Francisco,
San Francisco, CA U.S.A.
tel. (415) 666-5051

The programs have all been written in FORTRAN-77 on the VAX 11/780 computer with a VMS operating system. Migration of this software to other machines might involve some changes in the OPEN statements. In one of the program sections (DIGEST), there has been some use of the (nonstandard VAX FORTRAN) DO-END DO construct. Movement to another operating system would require that numbers be put on these do loops. Finally, some computers (CDC) do not support the LOGICAL*1 characteristics that VAX FORTRAN does, so that these arrays might have to be changed to CHARACTER*1 arrays.

DISCUSSION

The ability to access state-of-the-art programs for DNA, RNA and amino acid analysis and integrate them into SEQ has been shown to be versatile in designing the computational analysis needs of an industrial-research laboratory around its research needs. This process shall continue. Future directions are further enhancement of the DNA-RNA analysis programs as the need arises, as well as the design and implementation of a similar protein file management system.

REFERENCES

4. GenBank (1985), Genetic Sequence Data Bank created in 1982 by the National Institute of General Medical Sciences of the National Institutes of Health, distributed by the Research Systems Division, Bolt Beranek and Newman Inc., 10 Moulton Street, Cambridge, MA, 02238, U.S.A.
8. Sprang S., SNIPPER was obtained by writing the author c/o Dept. of Biophysics and Biochemistry, University of California, San Francisco, San Francisco CA 94143, U.S.A., tel. (415) 666-5051.
9. Protein Sequence Database of the Protein Identification Resource, obtained by writing the National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Rd., N.W., Washington, D.C. 20007, U.S.A., tel. (202) 625-2121