Report 81-03
Stanford -- KSL

Scientific DataLink

SEQ - Sequence Analysis System.
Jan Clayton, Peter E. Friedland,
Laurence H. Kedes, Douglas
Brutlag, Apr 1981

MEMO HPP-81-3

SEQ - Sequence Analysis System

April 1981

by

J. Clayton, P. Friedland, L. Kedes*, and D. Brutlag**
of the MOLGEN group

Department of Computer Science, Stanford University and
Departments of Medicine(*) and Biochemistry(**),
Stanford University School of Medicine

Introduction

SEQ is an interactive self-documenting program that performs many different types of nucleotide sequence analysis. SEQ prompts you with short questions for the information it needs. The program is particularly helpful for those inexperienced with the program in that typing a "?" at any point will get a description of what the program expects. The best way to learn to use SEQ is to read sections I and II of this document and then to run SEQ, typing "?" to each and every prompt for information. SEQ's responses to "?" form a short tutorial on the program itself.

While the various procedures of the SEQ program are related to many other programs of similar purpose (Korn, 1977; Staden, 1977; Staden 1978; Sege 1981; and reviewed in Gingeras 1980), SEQ has a rapid improved homology and dyad symmetry search algorithm which finds many homologies and dyad symmetries that are overlooked by earlier algorithms. SEQ also prepares restriction maps with the names and locations of the restriction sites marked on the nucleotide sequence as well as a facility for calculating the length of DNA fragments from restriction digests of any known sequence. SEQ handles circular sequences by continuing any search procedure from the end of the sequence to the beginning. Altogether SEQ has 13 different procedures with over 25 different sub-procedures many of which can be invoked simultaneously to provide different analytical methods to any sequence of interest.

This manual is designed to give a detailed explanation of the sequence analysis procedures (options) available in SEQ, and the user settable parameters that regulate these procedures.

Data Preparation

The following simple format for sequence files must be used for correct operation of the SEQ program. This format is recognized both by the SEQ program and the MOLGEN genetics knowledge bases, and it is designed to allow the user to have as much information about the sequence in the file as he may wish to have. These sequence files can be generated with any available text editors including TECO, SOS, TV, or EMACS. The format of the files is as follows:

```
; COMMENT LINE
;   "
;   "
SEQUENCE NAME1
ACGGCAAC .....
.............
.............1 or ...2
; COMMENT LINE
;   "
;   "
SEQUENCE NAME2
TTCTGGTGC.....
.............
.............1 or ...2
```

There are three sections of information for each sequence. First an unlimited number of comment lines can be present to keep information about the sequence. All comment lines must begin with a semi-colon (;), and can only come before a sequence name. Both the number of comment lines and their length is arbitrary. Comments are also optional, but they will be read by SEQ and can be typed out by the program (see Option 2). Also the first line of the comments should be descriptive of the sequence and the organism from which it came since the first comment line is printed as a descriptor of the sequences later in the program. Be careful not to put comments in the middle of a sequence or between the name and the sequence. Since blank lines in the file can be mistaken for the name of a sequence they should be avoided if possible.

The first line not starting with a ; is the name or title you wish to associate with the sequence. This name can be one word or a whole line of text, but if this line is missing, SEQ will assign the first line of the sequence to be the name of what follows.

SEQ will read subsequent lines as part of the sequence until it sees a termination character. Terminate linear sequences with "1" and circular sequences with "2". Any other characters can be used in a sequence, but only A, C, G, T, N (for any base), P (for purine), and Q (for pyrimidine) will be matched properly for homology and symmetry searches. Upper and lower case are considered equivalent, spaces, tabs, carriage returns and line feeds are ignored. Again, the length and the number of lines in a sequences is arbitrary.

Sequence files can contain as many different sequences in this format as the user wishes. For an example of a sequence file, look at one of the many sequence files at SUMEX on the <SEQUENCES> directory (e.g. <SEQUENCES>HUMAN.SEQ).

Using SEQ

Starting Up

SEQ is invoked by:

@SEQ<CR>

(where "@" is the EXEC prompt and "<CR>" represents carriage return). SEQ has been designed to be as helpful to the user as possible. Type "?" to any prompt to find out the appropriate responses.

```
Sequence file (name.ext or <CR> for none)
```

First the user is asked for the name of the file that contains the sequences that he wishes to examine. More than one sequence file can be loaded into the program. SEQ asks repeatedly for sequence files until the user responds with only a carriage return. This allows the user to enter sequences from a number of files at one time. (There is no need to start over if you forget to enter all the sequences you need at this point. Option S (see below) allows you enter more sequences at any time you see an "Option:" prompt.)

```
File for output? (<CR> for none)
```

Secondly, SEQ will ask you for a name of an output file. If you respond with a name (e.g.. SEQ.OUT), then you will be asked a few other questions about terminal output and line length for the output. Lines printed to a file can be limited to 75 characters (Short lines) or 120 characters (Long lines) in length depending on what kind of terminal or printer you will be using. (See section on Option F.) If you do not want the output to be saved in a file respond with <CR>, and the program will continue.

The Options

```
   Enter the option numbers (letters) you want, one per line.  End with
an extra carriage return.

Option: ?

    1  Prints the sequence (single or double stranded)
    2  Prints comments
    3  Nucleotide frequency tables (mono-, di- and/or trinucleotides)
    4  Rich Regions (AG and CT, AT and GC, or AC and GT)
    5  Oligonucleotide dictionaries
    6  Restriction site search, fragment list, and restriction map
    7  Translation to amino acids (3 frame, both strands, one-letter code)
```

```
 8  Homologous regions
 9  Symmetric regions
10  Regions of dyad-symmetry
11  Intersequence homologies
12  Intersequence symmetries
13  Intersequence base-pairing


 P  To change or examine PARAMETERS
 S  To input additional SEQUENCES from a file
 F  To either start or stop output to a FILE
 T  To either start or stop output to the TERMINAL
 Q  To QUIT (exit) the program
```

At this point, the user may specify as many alphabetic or numeric options as he likes in any order (one per line). After all the options for one round of analysis have been specified one begins the program with a bare carriage return with no option number or letter.

The Numeric Options

The numeric options will be executed only once per round no matter how many times they are requested. These options will be performed on ALL the sequences that you specify subsequently.

Many of these options require that the user specify suboptions. For example, option 3 (nucleotide frequency tables) has four sub-options (base composition, nearest neighbor frequencies, total trinucleotide frequencies, and codon usage in each reading frame) from which the user can choose. As with specifying options, only one sub-option can be specified per line, and the user is reprompted for additional sub-options until <CR> is typed to the prompt. A sub-option will be executed only once during a single round no matter how may times it is requested.

Option 1 - Printing the Sequence

Option 1 prints the specified sequence(s) in groups of ten nucleotides with every tenth nucleotide labeled numerically. When you specify option 1 you will be asked whether you want the sequence printed as a single strand or with its inverse complement in double stranded form.

An example of double stranded printing:

```
         10         20         30
AACGTTGCCT AACATTCGCT CCTGGTCGTA    . . .
TTGCAACGGA TTGTAAGCGA GGACCAGCAT
```

Option 2 - Printing Comments

Option 2 prints out all the comments associated with the specified sequence(s) as they are written in the sequence file. (In the sequence file these are the lines that begin with ";" that come immediately before the name of the sequence.)

Option 3 - Nucleotide Frequencies

Option 3 tallies and prints nucleotide frequency tables.

```
Which frequency table? (M, D, T, T3, ALL, <CR> or ?)
```

The appropriate responses are:

M - M prints out the Mononucleotide frequencies both as absolute quantities and percentages of the total base composition of the sequence.

D - D prints out Dinucleotide (nearest neighbor) frequencies. Again both the quantity and percentage of each dinucleotide are listed.

T - This suboption tabulates the Trinucleotide frequencies, Percentages and associated amino acids are printed out as in the following example:

```
TTT-Phe  0 (  .0)   TCT-Ser  2 ( 2.9)  TAT-Tyr  1 ( 1.4)   TGT-Cys  4 ( 5.8)
TTC-Phe  0 (  .0)   TCC-Ser  0 (  .0)  TAC-Tyr  2 ( 2.9)   TGC-Cys  0 (  .0)
TTA-Leu  0 (  .0)   TCA-Ser  2 ( 2.9)  TAA- .  1 ( 1.4)   TGA- .   2 ( 2.9)
TTG-Leu  0 (  .0)   TCG-Ser  1 ( 1.4)  TAG- .  1 ( 1.4)   TGG-Trp  0 (  .0)

CTT-Leu  0 (  .0)   CCT-Pro  1 ( 1.4)  CAT-His  3 ( 4.3)   CGT-Arg  0 (  .0)
CTC-Leu  2 ( 2.9)   CCC-Pro  1 ( 1.4)  CAC-His  0 (  .0)   CGC-Arg  1 ( 1.4)
CTA-Leu  2 ( 2.9)   CCA-Pro  2 ( 2.9)  CAA-Gln  2 ( 2.9)   CGA-Arg  0 (  .0)
CTG-Leu  3 ( 4.3)   CCG-Pro  0 (  .0)  CAG-Gln  1 ( 1.4)   CGG-Arg  0 (  .0)

ATT-Ile  0 (  .0)   ACT-Thr  3 ( 4.3)  AAT-Asn  1 ( 1.4)   AGT-Ser  1 ( 1.4)
ATC-Ile  1 ( 1.4)   ACC-Thr  1 ( 1.4)  AAC-Asn  1 ( 1.4)   AGC-Ser  0 (  .0)
ATA-Ile  3 ( 4.3)   ACA-Thr  1 ( 1.4)  AAA-Lys  5 ( 7.2)   AGA-Arg  0 (  .0)
ATG-MET  1 ( 1.4)   ACG-Thr  0 (  .0)  AAG-Lys  1 ( 1.4)   AGG-Arg  2 ( 2.9)

GTT-Val  0 (  .0)   GCT-Ala  1 ( 1.4)  GAT-Asp  0 (  .0)   GGT-Gly  1 ( 1.4)
GTC-Val  2 ( 2.9)   GCC-Ala  1 ( 1.4)  GAC-Asp  2 ( 2.9)   GGC-Gly  1 ( 1.4)
GTA-Val  1 ( 1.4)   GCA-Ala  1 ( 1.4)  GAA-Glu  0 (  .0)   GGA-Gly  0 (  .0)
GTG-Val  3 ( 4.3)   GCG-Ala  0 (  .0)  GAG-Glu  0 (  .0)   GGG-Gly  1 ( 1.4)
```

T3 - T3 lists the codon frequencies and percentages in each of the three frames separately in the above format.

ALL - All of the frequency tables.

<CR> - When you are finished specifying sub-options for a particular option, just type <CR> to the prompt and SEQ will continue.

Option 4 - Richness Option

Option 4 finds and marks regions which are "rich" in a specific pair of nucleotides. Rich, in this case, is defined as a containing greater than 75% of the specified nucleotides. The smallest region marked must be 8 bases long in which 6 of the 8 nucleotides are those bases specified. Since these regions are extended on the left and right until just before they drop below 75% of the specified nucleotides, the nucleotides immediately to the left or right of the marked regions will be

those bases NOT specified.

```
Richness Option (AT, AC, AG, or ALL) ?
```

AG - AG will print out the sequence with areas of AG richness marked by overlining (a leftward arrow on some terminals), and will then print out areas of CT richness in the same manner.

AT - This sub-option will mark areas rich in AT and CG.

AC - This sub-option will mark areas rich in AC and GT.

ALL - All three of the previous richness sub-options

<CR> - When you are finished specifying richness sub-options

An example of output for AT rich regions:

```
         10          20          30          40          50
  <-------- <---------- <--                <-------- <--  <---
CCACATTTTG CAAATTTTGA TGACCCCCCT CCTTACAAAA AATGCGAAAA

         60          70          80          90          100
<---------- <--------    <-------- <---------- <--
TTGATCCAAA AATTAATTTC CCTAAATCCT TCAAAAAGTA ATAGGGATCG

        110         120         130         140         150
        <-- <--------   <--------- <---------- <--------- <
TTAGCACTGG TAATTAGCTG CTCAAAACAG ATATTCGTAC ATCTATGTGA

        160         170         180         190         200
<---------    <-------- <---------- <---------- <----------
CCATTTTTAG CCAAGTTATA ACGAAAATTT CGTTTGTAAA TATCCACTTT

. . .
```

Option 5 - Oligonucleotide dictionaries (lexicography)

Oligonucleotide dictionaries are generated by looking at subsequences that begin at each character of the original sequence. For the sequence "AAGTGC" there are 6 sub-oligonucleotides, "AAGTGC", "AGTGC", "GTGC", "TGC", "GC", "G". The three types of dictionaries generated by SEQ are all variations of this complete oligonucleotide dictionary.

```
Dictionary option (1, 2, 3, ALL, <CR> or ?)
```

1 - This is the simplest type of oligonucleotide dictionary generated. Instead of printing out the entire subsequence starting at each base of the sequence, it prints only enough nucleotides to make the subsequence unique or distinguishable from all the others. The dictionary is printed in alphabetical (lexicographic) order for your convenience, and each entry is marked with the number of the first nucleotide in the string.

For the example sequence above, the Type 1 dictionary is:

```
1     AA
   2     AG
```

```
6     C
5     GC
3     GT
4     T
```

Option 1 is useful for finding the minimal length primers that would be required to initiate synthesis at any site in a given DNA template.

2 - The type 2 option prunes the Type 1 list by deletingall entries that are less than MATCHLENGTH long. (Refer to the section on parameters later in this manual.) What will be reported are sets of subsequences that are the same except for their last character. The following list could come from a Type 2 dictionary with MATCHLENGTH set to 8:

```
 26  AAAAAATC
301  AAAAAATT
 25  AAAAAAATC
300  AAAAAAATT
 24  AAAAAAAATC
299  AAAAAAAATT
```

3 - This suboption deletes the suboligonucleotides from the Type 2 dictionary list. For example, the first four entries of the Type 2 example (above) would be not be included, since they are substrings of the last two entries. This option and the previous are very useful for finding exact repeats of any specified length within a sequence.

```
 24   AAAAAAAATC
299   AAAAAAAATT
```

ALL - To print out all of the dictionaries.

<CR> - Type <CR> when finished specifying dictionaries.

Option 6 - Restriction Site Search (General String search)

Quite often one needs to look for a particular sequence of bases in a large sequence. Sometimes this small sequence will be a restriction enzyme recognition site, a Pribnow or TATA box, a capping site, etc. This option is designed to help you find a match of the specified strings. Before the procedure begins you will be asked for the character strings that you wish to use in the search:

```
String Search Option? (M, N, S, F, <CR>, or ?)
```

M - M prints out a complete nucleotide sequence with RECOGNITION sites marked and labeled with the site name.

```
        10         20         30         40         50         60         70
CCACATTTTGCAAATTTTGATGACCCCCCTCCTTACAAAAAATGCGAAAATTGATCCAAAAATTAATTTC
                              ↑
                             MnlI

        80         90        100        110        120        130        140
CCTAAATCCTTCAAAAAGTAATAGGGATCGTTAGCACTGGTAATTAGCTGCTCAAAACAGATATTCGTAC
                                                   ↑
                                                  AluI
```

. . .

N - This sub-option prints a list of all site names and sequences that were not found in the sequence (non-cutters).

S - The user can specify subsets of restriction sites for which to search. SEQ has a few predefined subsets, but if you are not happy with those, you can specify your own subset. (Note that only only one subset can be specified at a time.)

Subsets options are:

C - Restriction enzymes known to have Cohesive ends
F - Restriction enzymes that are Flush cutters
5' - Restriction enzymes that have 5' extensions
3' - Restriction enzymes that have 3' extensions

4 - Sites that are 4 nucleotides long
5 - Sites that are 5 nucleotides long
6 - Sites that are 6 nucleotides long

S - Specify your own subset from the list of restriction sites already entered. This is done in a similar manner to the specification of sites in fragment mode (see below. However, there is no limit to the number of sites that the user can request to be in the subset.

F - Specifying F puts the user in Fragment Mode. When SEQ executes this option (after all other sub-options and sequences have been specified), SEQ will start up an interactive session with the user. The user will be able to specify up to 4 enzymes from the list that he/she has already entered (see below) for generation of a fragment list (in decreasing order by fragment length). The program does not yet distinguish between restriction enzyme recognition sites and the actual cutting sites. Thus the distances calculated for fragment lengths represent the distances between recognition sites. Linear and circular molecules are correctly distinguished. After the first fragment list is printed out, SEQ will ask the user for another set of up to 4 sites for generation of a another fragment list. This process will continue until the user types <CR> to the FIRST "Site:" prompt of a given set.

In Fragment Mode, the original settings of the other string search sub-options will be retained. (Note: Sub-option "S" for subset is the only String search sub-option that cannot be used simultaneously with Fragment mode "F".)

An example of the fragment output is:

Enzyme	Site	Length	Enzyme	Site
AluI	(116)	186	AluI	(302)
MnlI	(28)	88	AluI	(116)
AluI	(302)	85	MnlI	(28)

Entering Restriction Sites (Search Strings)

SEQ does not automatically load search strings or restriction enzyme sites when you ask for Option 6. Instead, it asks the user to enter the sites from either a file or from the terminal.

```
Sequence Search file (F for file name, S for strings from the terminal,
      <CR> to user default file)
```

F - If you answer with "F", SEQ will ask you for a file name. That file should be in Restriction file format; one site name and sequence per line.

```
Site1    AACTCT
Site2    CCGTTG
Site3    CNNTG
Site4    CPQAA
.  .  .
```

A, C, G, and T will be matched directly (regardless of case), P will be matched with purines and Q will be matched with pyrimidines and N will be matched with anything. (If you have any doubts about the file format, at SUMEX look at <SEQUENCES>REST.SEQ.)

S - If you specify "S", SEQ will ask you to enter site names and sequences from the terminal. One name and one sequence should be be typed per line just as in the file format described above. When finished entering the sites, type <CR> to the "Name & string:" prompt.

<CR> - A single carriage return will load the default file REST.SEQ, which contains a list of restriction enzyme names and recognition sites. Only one prototype enzyme for each isoschizomer is present in the list in order to reduce redundancy. This file currently contains about 100 restriction enzyme names and sites. This default file must be loaded if you wish to use 5'-extension, 3'-extension, flush-cutters, or cohesive-end subsets.

After you finish loading in a set of restriction sites, SEQ will ask you if you wish to enter any additional sites. If so, just type "Y" and SEQ will reprompt you with "Sequence search file?...". This feature allows a user to enter restriction sites (or any other set of sequences) from a number of different sources.

Option 7 - Translation

This options prints out a translation of the sequence. Termination codons are signified by a period (.). Unambiguous codons are translated even if they contain an unknown base.

```
Translation option? (M, 1, 2, 3, P, F, <CR>, or ?)
```

M - (mitochondrial) This suboption translates the sequence using the mitochondrial genetic code.

1 - (One letter code) This suboption translates the codons into the one letter amino acid code.

```
                                      27                                    54
AAA AAA GGT GTG TAC CCA TAC ATG TCT GAC TGA CTA GGG CTC TAN NPQ PCC TGT
K   K   G   V   Y   P   Y   M   S   D   .   L   G   L                   C

                                      81
CAA TAT CGC CAG TGN GCA TAA CTC AAA
Q   Y   R   Q       A   .   L   K
```

2 - (2 stands) Suboption 2 translates both strands. The inverse complement is printed and numbered in the 5' to 3' direction.

```
                                        27                                  54
AAA AAA GGT GTG TAC CCA TAC ATG TCT GAC TGA CTA GGG CTC TAN NPQ PCC TGT
Lys Lys Gly Val Tyr Pro Tyr MET Ser Asp  .   Leu Gly Leu              Cys

                                        81
CAA TAT CGC CAG TGN GCA TAA CTC AAA
Gln Tyr Arg Gln      Ala  .  Leu Lys

Numbering increases in 5' to 3' direction

                                        27                                  54
TGA GTT ATG CNC ACT GGC GAT ATT GAC AGG QPQ NNT AGA GCC CTA GTC AGT CAG
 .  Val MET     Thr Gly Asp Ile Asp Arg         Arg Ala Leu Val Ser Gln

                                        81
ACA TGT ATG GGT ACA CAC CTT TTT TTA
Thr Cys MET Gly Thr His Leu Phe Leu
```

3 - (3 frames) This suboption will print out the translation for all three frames of the sequence.

```
                                        27                                  54
AAA AAA GGT GTG TAC CCA TAC ATG TCT GAC TGA CTA GGG CTC TAN NPQ PCC TGT
Lys Lys Gly Val Tyr Pro Tyr MET Ser Asp  .  Leu Gly Leu              Cys
 Lys Lys Val Cys Thr His Thr Cys Leu Thr Asp  .  Gly Ser          Pro Val
  Lys Arg Cys Val Pro Ile His Val  .  Leu Thr Arg Ala Leu          Leu Ser

                                        81
CAA TAT CGC CAG TGN GCA TAA CTC AAA
Gln Tyr Arg Gln      Ala  .  Leu Lys
 Asn Ile Ala Ser      His Asn Ser
  Ile Ser Pro Val      Ile Thr Gln
```

F - (Full) This is shorthand for suboptions 2 and 3 combined.

P - (Partial) The Partial translation suboption allows the user to specify only certain regions (exons or coding regions) within the sequence to be translated. Any part of the sequence can be printed, but only the coding segments requested will be translated. The segments can be specified in any order, but if two segments overlap, the one that was specified last will be ignored. This suboption cannot be used with TRANSLATION suboptions 2, 3 or F.

Homology and Symmetry Options

The same basic algorithm is used in all homology, symmetry, and dyad-symmetry searches. This algorithm is an improvement of the Korn-Queen algorithm since it can extend homologies past a greater variety of mismatches. The algorithm starts at a given position in the two sequences or 2 positions in the same sequence, and finds the longest exact match from those positions. It then tries to extend the match farther by looking for a mismatch with a significant region of exact match beyond the region of mismatch. The algorithm tries to extend in set order the following arrangements of mismatched nucleotides:

(Number of nucleotides in mismatched region)

Strand 1	0	0	1	1	0	2	1	2	0	3	2	1	3	2	3	3
Strand 2	0	1	0	1	2	0	2	1	3	0	2	3	1	3	2	3

In other words, when a mismatch is found, first a deletion of one base in the first strand is tried to see if the match can be extended, if not then a one base deletion is tried in the other strand. Following that, a one base mismatch is tried, etc. This process is continued until either an acceptable extension is found, or all the possible mismatches have failed to be extended. Whether a given match is acceptable or not depends upon the values of a number of user settable parameters. Five parameters effect all of the homology and symmetry search procedures. AFTERDIS (AFTER DIScontinuity) is the number of matches out of 3 which must follow a mismatch or loopout in a homology or symmetry search. LOOPOUT is the maximum length of any loopout in a dyad symmetry or the maximum insertion-deletion loop size in a homology. MINMATCH is the minimum number of exact nucleotide matches. PERCENTMATCH is the minimum percent of exact matches to the total length of the match. And lastly, EXPECT is the approximate number of matches to be reported. Detailed explanations of all these parameters can be found in the Section on Option P.

The order of checking mismatches described above is not necessarily optimal for every homology or symmetry. We do believe, however, that this is one of the best orderings to pass through regions of extreme mismatch to other more highly paired regions. Unlike the Needleman and Wunsch (1971) or the Sellers algorithms (1974, 1979), SEQ will not extend a homology with the minimum number of mismatches, since it does not look ahead and then make up its mind on what mismatch to select. Don't be surprised if at times you can see, by a quick glance, that a better arrangement of mismatches would have allowed a homology to be extended further or if some other arrangement of mismatches would result in a higher overall degree of matching. The major advantage of the SEQ algorithm is speed; this makes it useful interactively. Comparison of the output from both this new SEQ algorithm and the Needleman and Wunsch or Sellers algorithm applied to identical sequences has shown that the alignments produced by those procedures are subsets of the alignments found by SEQ.

Option 8 - Homologous Regions and Option 11 - Intersequence Homology

Options 8 and 11 look for homologous regions within a single sequence and between two sequences respectively. Besides the parameters mentioned above, internal homologies are regulated by the parameter MAXDIST, which is the maximum distance between the starting nucleotide of the matching regions.

```
                 *  **    *  *
   235       AAAT AATCATTTATTTTGCCACA       257
   344       AAATTAA  ATTT TTTGGCCACA       361
% = 79.2
P( 24, 19) = .1510-4 E = .236
```

The output consists of a list of the homologies in the order they are found. The beginning and last base of each strand are numbered, and the sequences are printed so that insertions (deletions) are easily seen. Asterisks (*) mark all unmatched bases. The percentage displayed is the percentage of exact matches in the total length of this particular homology. The value P(N,M) is the absolute probability of a match N nucleotides long with M nucleotides matched or something more significant would occur in comparing two sequences of the given length and base composition assuming a random distribution of the bases (see Appendix A to see how P(N,M) is calculated). Thus this probability is the sum of the probabilities of a series of homologies which are at least as significant as the one reported. When this probability is multiplied by the total number of comparisons made during this search then you get the value of E, the expected number of matches of this significance or greater that you would find in a random sequence of the same length. Since the values of the expectation frequencies depend on the algorithm used to find the homologies (which is determined by values of AFTERDIS, LOOPLENGTH, MINMATCH, MINDIST, and

PERCENTMATCH; see Appendix A), the determination of the expectation frequency of each homology should be measured using the most stringent search parameters possible. For instance, if a homology is found which has no insertion-deletion loops but LOOPLENGTH was set so as to allow a loopout of 3 bases, then the value of E would overestimate the expected frequency of such a homology and not be indicative of the true significance of the homology. The search should be rerun with the most stringent values for all of the parameters (in the case just cited, LOOPLENGTH should be reset to 0 base loopouts allowed). A homology is considered to be STATISTICALLY significant when the value of "E" under the most stringent search conditions is less than 0.05.

Do not be surprised if many biologically significant homologies are not statistically significant. Unfortunately biological significance does not imply statistical significance. However, the converse is usually true and that is what makes SEQ valuable, i.e. those sequences which are statistically significant are usually biologically significant.

Option 9 and Option 12 - Symmetric Regions

Options 9 and 12 are similar to 8 and 11 except that they search for symmetric regions instead of homologous regions. These procedures search for true alphanumeric palindromes (NOT dyad symmetries) and the biological significance of such symmetries has yet to be demonstrated. In addition to the five parameters mentioned in the introduction to this section, Option 9 is regulated by MAXLOOP and MINLOOP, the maximum and minimum distances (respectively) between the starting nucleotides of the
matches.

```
                  *    * * *    *
     305      TCGTAATAAA ATT TCCAATCAA     326
      96      TCGTACTAAAGAGTCTCCACTCAA      73
% = 79.2
P( 24, 19) = .974@-6 E = .209@-1
```

The printout of the symmetric regions is equivalent to that of homologies (Note, however, that the numbering of the second strand decreases instead of increases). Asterisks (*) mark the mismatched nucleotides, the "P" value is the probability that a match of this type would occur, and the "E" value is the number of similar matches that one would expect to find in a search of a random sequence of the same length and base composition.

Options 10 and 13 - Inter- and Intrasequence Dyad-Symmetries

These options find areas where base paring can occur. Option 10 finds regions of dyad symmetry within a given sequence and procedure 13 finds regions that would hybridize between two sequences. During these comparisons A's match with T's (or U's), C's with G's, and if the parameter GUPAIR is set to 1, G's will also pair with U's (T's). Thus Option 10 can be used both for looking for staggered symmetries typical of protein binding sites with GUPAIR set to 0, and for hairpin loops in single-stranded DNA or RNA sequences with GUPAIR set to 1. As with option 9, option 10 is governed by MAXLOOP and MINLOOP, the maximum and minimum distances between starting nucleotides of the two strands.

```
   36      CAAAAAATGCGAAAAT TGATCCAAAAA TT AATTTCC 71
           |||||  ||| ||||| ||  ||||||| || |||||||
   19      GTTTTAAACG TTTTACAC CGGTTTTTTAAATTAAAGG  342

% = 79.5
P( 39, 31) = .122@-7 E = .191@-3 G = -28.9
```

For options 10 and 13, SEQ marks all Watson-Crick base pairs with "|". GU-pairs (when GUPAIR is set to 1) are marked with ":". In addition to the percent matching, the probability of the match and the expectation value for the match, the Delta-G (total free energy of base pairing) is calculated according to the rules presented in Tinoco (1971, 1973) and the free energies calculated by Borer (1974). These values are determined for hairpin loops and stems in RNA only and hence are not necessarily valid for DNA. Furthermore, these values are very imprecise and should only be used to give relative values for the stabilities of such hairpins. Only values of delta-G less than about -15 kcal would normally be considered stable under physiological conditions.

The Alphabetic Options

Notice that in addition to the numbered options, there are also options that can be specified with alphabetic characters. These options perform tasks that alter the performance of the program (e.g. changing parameters, stopping and starting output to a file). Unlike the numbered options, SEQ takes action as soon as you request the option. Because of this, you can change your mind about parameter settings or output files, as many times as you wish, when ever you see the "Option:" prompt.

Option P - Examining and Changing Parameters

Option P calls a procedure that enables the user to examine and change the user settable parameters.

```
Change parameter (Parameter name or number, ? for a description,
        L for a list, or <CR> when done)
```

If you type "L", SEQ will give you a list of the parameters and their current values. A "?" will give you a short description of each of the parameters along with the "L" list. If you type the name or number of a parameter, SEQ will show you the current setting of that parameter and ask you for a new value. When you have finished changing parameters, type <CR> to the prompt, and SEQ will return with...

```
Create a new parameter profile file? (Y or N)?
```

(NOTE: This is not available to GENET users!)

If you wish to save the current parameter settings for future SEQ sessions, answer "Y". This will make a new SEQ.PROFILE file in your directory, and the values that are saved will become the default values for all subsequent sessions. If you just want to use the altered settings for the current sessions, answer "N".

The following is a list of the user settable parameters and their default values:

```
1.  AfterDis = 2              7.  GUPair = 0
2.  LoopOut = 3               8.  MaxDist = INF
3.  MinMatch = 5              9.  MatchLength = 0
4.  PercentMatch = 75        10.  Expect = 10.0
5.  MinLoop = 0              11.  RestFile = <SEQUENCES>REST.SEQ
6.  MaxLoop = 20
```

AFTERDIS -

This parameter determines how many nucleotides out of the next three must be matched in order to continue a homology or symmetry past a mismatch or loopout. This parameter may be set between 1 and 4. We suggest that the value of this local matching not be set lower than 2 because at a setting of 1 an excessive amount of marginally significant matches will be generated. If set to four, then no mismatches will be allowed in the comparison.

LOOPOUT -

This parameter controls the length of internal bulge loops that are allowed in dyad symmetries. It also controls the length of insertion or deletion loops allowed in comparing two homologous sequences. If set to zero it prevents the introduction of bulge loops or insertions-deletion loops. The maximum size of bulge loops or insertions and deletions is three bases.

MINMATCH -

MINMATCH is the minimum number of exact matches in a homology or symmetry that must be met before SEQ will print out the match. If set to 10 for example, only homologies that have at least 10 nucleotides matched will be reported. MINMATCH effects the number of comparisons made near the ends of sequences. If set to 10, for instance, the first 10 nucleotides of one sequence will not be paired with the last ten of the other.

PERCENTMATCH -

PERCENTMATCH is the minimum ratio of matched nucleotides to total nucleotides in a homology or a symmetry in order for that match to be printed. However, it has an additional consequence due to the fact that the SEQ searching algorithm proceeds unidirectionally along a sequence. The homologies that are generated by sequence comparisons can never be less matched than PERCENTMATCH at any point in the homology. This means that the homologies that are printed will start off well paired and will then degenerate, sometimes precipitously, near the end of the match. The final few matched and mismatched nucleotides at the end of the homology should usually be ignored. In order to overcome this polarity effect it is usually wise not only to compare one sequence to another but to also compare the complements of the two sequences, thus examining each homology from both directions. Similarly it is wise to search for dyad symmetries in both the sequence and in its inverse complement and to look for internal homologies within a single sequence and its complement.

MINLOOP -

MINLOOP is the minimum distance allowed between starting nucleotides of symmetric and dyad symmetric regions. MINLOOP can be used effectively in the dyad-symmetry procedure to look for hairpin loops greater than some minimum size, which will decrease the total number of comparisons made during the search.

MAXLOOP -

MAXLOOP determines the maximum distance allowed between symmetric and dyad symmetric regions. Therefore, MAXLOOP can be used with the dyad symmetry procedure to look only for hairpins with loops less than some specified maximum size. Decreasing MAXLOOP decreases the total number of comparisons made during the search. Looking for hairpins of small loop size results in more statistically significant dyad symmetry than a similar inverted repeat in which the elements of symmetry are widely spaced.

GUPAIR -

GUPAIR determines whether GU-pairing will be considered as a match for the dyad-symmetry search procedures. If set to 1, G's will be allowed to match with U's (or T's) which is useful for looking at hairpin stems in RNA. Setting GUPAIR to 0 prevents GU or GT pairing and is most useful when scanning for structures similar to the staggered dyad symmetries found in many protein binding sites. Setting GUPAIR to 0 reduces the probability that any two nucleotides chosen at random will match and, hence, increases the STATISTICAL significance of dyads discovered.

MAXDIST -

MAXDIST determines the MAXimum DIStance allowed between the initial nucleotides of regions in an internal homology (Option 8). In other words, MAXDIST sets the upper limit on the distance between homologous regions. Decreasing MAXDIST decreases the total number of comparisons made during the search.

EXPECT -

EXPECT limits the number of homologies or dyad symmetries printed to only the most STATISTICALLY significant. For each match that is found, the expectation value is determined as described in Appendix A. This is the number of times that one would expect a comparison of this statistical significance or greater to occur in a random sequence of nucleotides of the same base composition and length as the sequence(s) being examined. If one sets the value of expect to 10, on the average ten homologies will be printed, and the expectation values for each of the homologies will be between 0 and 10. If there is a considerable amount of internal redundancy in the sequence under consideration, more output should be expected. Only those homologies whose expectation values are substantially lower than 1.0 are statistically significant.

The values of the expectation depend critically on the nature of the algorithm being used in the search (determined by the settings of the parameters). This is because there are many more ways to find a homology, say 30 nucleotides in length with only 20 nucleotides matched if one allows loopouts and mismatches, than there are if one merely allows only mismatches or loopouts.

In order to determine the actual expectation frequency of a particular homology that has been found, one should repeat the search and set the most stringent values of the search parameters that are possible that still find the homology. For instance if, for a particular homology pair, no loopouts longer than 2 base pairs occur then looplength should be reset to 2, or if all mismatches in the homology are followed by three matched nucleotides then AFTERDIS should be reset to 3 (or to 4 if there are no mismatches). Furthermore, one should also set the value of PERCENTMATCH to the highest value possible which still allows the particular homology to be found. Thus, the expectation for a homology printed under the most stringent search algorithm gives a good estimate of the number of times one would expect to encounter that homology in a random sequence of identical base composition and length. Values of .05 or less are usually considered statistically significant.

MATCHLENGTH -

This parameter is used in the Type-2 and Type-3 dictionary procedures. It is the minimum subsequence length that will be reported. If a value of 0 is specified, then the program will calculate a value which will produce a limited amount of output.

RESTFILE -

The RESTFILE parameter stores the name of a file that SEQ will look at when it is directed to do a general sequence search (Option 6). The data in the file named in RESTFILE is read when the user types <CR> to the prompt:

```
Sequence search file? (F for file name, S for strings from the terminal,
        <CR> to use default file)
```

Usually RESTFILE is set to the name of a file that the MOLGEN group maintains and updates regularly. This file contains essentially all known restriction enzyme names and recognition sites. On SUMEX the file is named <SEQUENCES>REST.SEQ. This file has a special format which is discussed in the Section on Option 6. REST.SEQ has special codes that mark each site to tell SEQ whether that site is a flush-end, a cohesive-end, a 5'-extension or a 3'-extension cutter. If you wish to use the Subset sub-option of Option 6, the file REST.SEQ must be used.

Option S - Reading in More Sequences

Quite often users forget to enter the names of all the sequence data files that they want to analyze. Or, in the middle of a session, the user might decide to look at a sequence in a file that is not loaded into the program. Option S allows a user to enter the name of a file that contains sequence data whenever options can be requested. Options S will immediately prompt with "Sequence file (name.ext or <CR> for none) " which is the same as the first prompt that SEQ gives. Treat it the same way by giving the name of a sequence file. SEQ will continue to prompt for the name of additional files until <CR> is typed to the prompt.

Option F - To Stop or Start Output to a File

This option will let you stop or start entering SEQ output into a file. If you are currently writing output to a file. SEQ will ask you to confirm the closure of the output file ("End output to file? (Y or N)"). If you type "N", no action will be taken.

If data is not currently being stored in a file then SEQ will ask you to create the name of a file into which the output will flow for future reference ("File for output? (<CR> for none)"). Once the file has been opened, SEQ will then ask you several questions about terminal output and line length:

```
Terminal output also? (Y or N)
```

When you open a file for output you can stop the output to the terminal if you you like. Remember, however, this will stop all the output until you either close the output file or request Option T (See below).

`Short lines? (Y or N)`

SEQ allows the use to choose between 2 output line lengths. If the user answers "Y" to the short lines prompt, SEQ will limit the line length to approximately 75 characters (standard for 8 1/2 X 11 paper). The line length is set to 120 characters for a "N" response.

Option T - To Start or Stop Output to the Terminal

A user may want to turn terminal output on and off regardless of whether the output is going to a file. Option T acts as a switch to do this. If terminal output is turned on, SEQ will turn it off and conversely. The user is always asked to confirm the change.

Option Q - To Quit (Exit) the Program

Option Q is the only safe way to exit the program! If you type ↑C to exit the program, you will lose the last partial page of your output. If for some reason you have to exit with a control-C (to prevent a runaway program), type CLOSE ALL to the EXEC prompt to close the output files. All but the last page of your output will be retained.

Requesting Sequences

After a user has finished specifying options, SEQ will ask for the sequence(s) to be examined. For options 1-10, a single prompt (Sequence:) is given since those options only analyze one sequence at a time. For options 11-13, however, prompts for two different sequences are given (Sequence 1: or Sequence 2:). As with the option and suboption prompts, SEQ allows you to request as many sequences (or pairs of sequences) as you wish. When you have finished specifying sequences, type <CR> to the prompt.

Sequences may be specified in several ways. Giving the name or number of a sequence, the word ALL for all the sequences that have been read into SEQ, the word COMBINED for all the sequences concatenated into one long sequence or following any of the above with "'" for the inverse complement(s) are all legal ways to request sequences. Either entire sequences or parts of sequences (such as single genes) may be analyzed. To analyze part of a sequence you merely follow the name (or number) of the sequence with a single space and then the lower and (optionally) the upper coordinate of the nucleotides you which included in the region. If no upper limit is given then the end of the sequence in the file is assumed. On a circular sequence the coordinate of the lower limit may be greater than the upper limit in which case the region will extend from the location of the lower limit, cross the beginning of the sequence, and end at the location of the upper limit.

SEQ does not expect users to remember the names of all the sequences that they have entered, nor the order in which they were entered. So, if you type "?" to the "Sequence: " prompt, SEQ will respond with some general information and then will give a list of all the sequences that have been entered. For each entry, the sequence number, the sequence name, the first line of the comments for that sequence, and the length of the sequence are given.

As mentioned above, inverse complement sequences can be analysed by following a sequence name or number with "'" (an apostrophe). There is no need to store the complements in files as SEQ does it on its own when you make this request.

SEQ also allows you to combine sequences. When you specify "COMBINE" SEQ will ask you for the names (or numbers) of the two or more sequences that you wish to put together. SEQ appends these sequences to each other in the order that you give them, and from then on this combination is treated as a new sequence (the joined regions are not distinguished in any way). Regions of sequences can be specified in the way described previously by giving a lower and (optionally) an upper limit. Thus the COMBINED option allows the user to insert or delete regions within a single sequence or insert one piece of sequence into another. When you have finished specifying the sequences for the combination, SEQ will ask you to create a name for the new sequence. From then on you will be able to refer to that combination with that name or its new sequence number.

After specifying the sequences, SEQ will ask the user to fill in the information regarding the SUB-options for those OPTIONS previously requested by the user. SEQ then will generate the output. When it has completed its tasks, SEQ will return control to you, and will prompt for another round of options.

Control Characters

Most control characters are ignored by SEQ, that is they do not perform any special function. However, the following control characters can be used to advantage.

<ESC> (or ALTMODE on some terminals) will aid you in filename recognition. After typing the first few characters of a filename followed by <ESC>, SEQ will spell out as much of the filename as is unambiguous, and then will "beep" at you if there are several files with similar names. If a given file is unambiguously identified by those first few characters then <ESC> will spell out the complete filename and ask for confirmation (carriage return confirms).

↑F will also aid you in file recognition. Control-F is similar to <ESC> except it will only complete recognition of one field of the file name at a time (<DIRECTORY> or FILENAME or EXTENSION or VERSION).

↑O means "shut up" and will temporarily discontinue output to the terminal. If you know that a printout of a translation or a list of homologies is going to be extremely long, you may not want to have it printed out to the terminal. (Note that this will not stop output to a file!) In SEQ ↑O will only stop the output for a single option. So if you have specified a number of options at one time, one ↑O will only stop the output of the option that is currently being printed. Especially on slow terminals, ↑O can speed up the clock time that elapses while SEQ is performing its tasks. While "running in the dark" in response to ↑O, the time SEQ takes to complete its task is limited by the speed of the computer, not by the printing speed of your terminal.

↑C BEWARE! Control-C should not be used as a normal exit from the program. Use ↑C only if you wish to abort the program, and if you do not want to save the recent output that has been sent to a file. Please use Option Q for exiting the program. (See section on Option Q.)

Appendix A

Calculation of probabilities

The probabilities and expectation frequencies for homologies and symmetries are calculated according to Markov methods as originally suggested by Korn et al. (1977). The SEQ search procedure can be represented as a Markov Chain, i.e. a series of events, the probability of whose next step depends only on the present state and not on the series of steps that lead up to that state. SEQ first calculates the probability P of any two nucleotides matching given the base compositions of the two sequences being compared. This is the probability of the algorithm extending the current homology by one base pair which is matched. Then SEQ calculates the probability of the algorithm not finding a match (1-P) and extending the homology by a insertion or deletion loop of one nucleotide in either strand. This probability depends on the values of the parameters AFTERDIS and LOOPLENGTH. SEQ then calculates the probability for the next possible extension. This gives a series of probabilities for the extension of any homology by N nucleotides with M of them matched at any step of the homology pairing routine.

These probabilities then serve as the transition probabilities in a transition matrix (See Feller 1968 or Karlin and Taylor, 1975). The transition matrix has to be modified slightly to take into account that the algorithm can not reach any homology with M/N being less than PERCENTMATCH. SEQ then calculates the equilibrium value for the state vector governed by this transition matrix. This equilibrium vector gives the absolute probability of reaching any given state of homology (the absolute probability of obtaining a homology N nucleotides long with M nucleotides paired). These probabilities are then ranked in descending order and for each final state (N,M) SEQ sums all of the probabilities less than or equal to that for (N,M). This gives the probability of reaching the state (N,M) or any state less likely than (N,M). Thus SEQ prepares a matrix P(N,M) which gives this probability for each possible final state. These probabilities are then multiplied by the total number of comparisons made (L*(L-1)/2 for a sequence L long versus itself for instance) in order to calculate the number of homologies of that significance or greater that would be expected to be found for this particular run.

Finally, the actual calculation of the steady state values of the state vector are performed by a highly efficient approximation. To make this calculation without this approximation would take 1000 times longer than the current code. The approximation gives probabilities good to 1 part in 10,000 which is more than satisfactory for this work.

These probabilities are calculated for each and every run of the dyad symmetry, homology or symmetry options taking into account for each comparison the base composition, the length of the sequences, and the algorithm used (i. e. values of AFTERDIS, PERCENTMATCH, LOOPLENGTH, MINMATCH, and MAXDIST or MAXLOOP). This means that with EXPECT set to 10, SEQ will print about 10 homologies or symmetries regardless of the length or base composition of the sequences being compared.

References

Borer, P. N., Dengler, B. and Tinoco, I. Jr. (1974). Stability of ribonucleic double-stranded helices. *J. Mol. Biol. 86*, 843-853.

Feller W. (1968). "An Introduction to Probability Theory and its Applications." John Wiley and Sons, Inc. New York, pp. 372-424.

Gingeras, T. R. and Roberts, R. J. (1980). Steps toward computer analysis of nucleotide sequences. *Science 209*, 1322-1328.

Karlin, S. and Taylor, H. M. (1975). "A First Course in Stochastic Processes." 2nd Edition, Academic Press, New York.

Korn, L. J., Queen, C. L. and Wegman, M. N. (1977). Computer analysis of nucleic acid regulatory sequences. *Proc. Natl. Acad. Sci. U.S.A. 74*, 4401-4405.

Needleman, S. B. and Wunsch, C. D. (1969). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol. 48*, 443-453.

Sege,R., Soll, D., Ruddle, F. H. and Queen, C. (1981). A conversational system for the computer analysis of nucleic acid sequences. *Nucleic Acids Res. 9*, 437-444.

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math. 26*, 787-793.

Sellers, P. H. (1979). Pattern recognition in genetic sequences. *Proc. Natl. Acad. Sci. USA 76*, 3041.

Staden, R. (1977). Sequence data handling by computer. *Nucleic Acids Res. 4*, 4037-4051.

Staden, R. (1978). Further procedures for sequence analysis by computer. *Nucleic Acids Res. 5*, 1013-1015.

Tinoco, I., Uhlenbeck, O.D., and Levine, M.D., (1971). Estimation of Secondary Structure in Ribonucleic Acids. *Nature 230*, 5293.

Tinoco, I., et. al., (1973). Improved Estimation of Secondary Structure in Ribonucleic Acids. *Nature New Biology 246*, 40.