Volume 9 Number 2 1981 Nucleic Acids Research

A conversational system for the computer analysis of nucleic acid sequences

Robert Sege[§*], Dieter Söll[†], Frank H.Ruddle* and Cary Queen[+]

*Departments of Biology and [†]Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, and [+]Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, USA

Received 28 August 1980

ABSTRACT

We present a conversational system for the computer analysis of nucleic acid and protein sequences based on the well-known Queen and Korn program (1). The system can be used by persons with only minimal knowledge of computers.

INTRODUCTION

A number of useful computer programs for analyzing nucleic acid sequences have been introduced over the past few years (1-6). Computer analysis is used to search sequences for features of biological interest, to determine possible secondary structures, and to assist in making generalizations about the relationship between nucleic acid structure and function. One of the most flexible computer programs for these purposes has been designed by Queen and Korn (1). It performs four categories of functions for nucleic acid and protein sequences: (i) counting and searching functions, (ii) examination of a single sequence for repeated, palindromic, or self-complementary regions, (iii) comparison of two or more sequences for features held in common, (iv) translation and reverse-translation using the genetic code.

A barrier to using these programs for investigators not experienced with computers is the time involved in learning computer systems. For this reason we have developed a conversational system based on the Queen and Korn program (1). The system can be used by persons with only minimal knowledge of computers. Our program is to be used interactively: the researcher sits at a terminal and the computer prompts him with questions. Here we report this conversational analysis system (written PL/I*) which can be installed at central computer facilities operating with time-sharing capabilities. The system provides a built-in teaching facility as well as error-detection and correction routines. These features enable an investigator to ask specific questions about a sequence, and receive answers quickly.

GENERAL DESCRIPTION

This program provides all the analytic capabilities for nucleic acids and proteins described by Queen and Korn (1), with the improvements discussed below. This program accepts sequences of nucleotides or amino acids as data. A nucleotide sequence is entered as a string of the letters A, C, G, T, and U. The letters T and U are not distinguished, so DNA sequences may be compared against RNA sequences. The letters P for purine, Q for pyrimidine, and N for nucleotide are accepted by procedures 4 and 14 below. An amino acid sequence is entered as a string of the standard three-letter abbreviations preceded by an asterisk.

Sixteen independent procedures (summarized in Table 1) may be used to analyze sequences, while twenty-five parameters (summarized in Table 2) provide user control of the analysis. For a detailed description of all aspects of the program see reference 1 or the user manual. Input/output format is flexible, and analysis output can be diverted to disk storage or to high speed printers. A complete tutorial mode and error-checking provisions have been supplied.

Control of the Analysis

The user controls the analytical procedures through a number of special variables, called parameters (Table 2). For example, the investigator may set the parameters that direct the computer to select only those homologies with more than a specified number of nucleotide matches, or to print only those features which would occur by chance alone with less than a specified probability. These parameters have been divided into a number of user-selectable "menus", and are available for review and change before each analysis is performed. Relevant parameter values are printed out at the end of each analysis.

Conversational Features

Program control is entirely conversational. The investigator uses a terminal to gain access to the central computer. After communication is established, all requests (parameter values, procedures, sequence manipulations, etc.) are made in response to program prompts. The whole session takes the form of a dialogue: the computer requests instructions, the user supplies these instructions, then the computer processes these instructions and requests further input.

Rather than supply instructions at any time, the investigator may request computer assistance by typing "?". The program then moves into a tutorial mode where brief explanations are supplied. This program ability enables the

TABLE 1

PROCEDURES OF THE PROGRAM

1.	Printing of sequence	*Prints out input.*
2.	Nucleotide and amino acid frequency	*Determines number and percentage of each kind of nucleotide or amino acid.*
3.	Dinucleotide frequency	*Frequency of occurrence of all possible dinucleotides.*
4.	AG- and CT-rich regions	
5.	AT- and GC-rich regions	*Locates regions in which 6 out of 8 consecutive bases are the ones specified.*
6.	AC- and GT-rich regions	
7.	Subsequence dictionary	*Lists all subsequences in a given sequence.*
8.	Matching subsequence dictionary	
9.	Matching subsequence dictionary, simplified	*Lists perfectly repeated oligonucleotides.*
10.	Repeated regions	
11.	True symmetries (alphabetic palindromes)	*Finds homologies which are not necessarily perfect.*
12.	Dyad symmetries (self-complementary regions)	
13.	Genetic code:	*Uses genetic code to translate nucleotide sequences and to reverse translate amino acid sequences.*
	Translation of nucleotide sequences	
	Reverse translation of amino acid sequences	
14.	Location of oligonucleotides and polypeptides:	*Searches for subsequences of various types in a given sequence (especially useful in locating restriction enzyme recognition sites).*
	Oligonucleotides in nucleic acids	
	Polypeptides coded by nucleic acids	
	Oligonucleotides reverse coded by protein	
	Polypeptides in proteins	
15.	Trinucleotide frequency	*Determines total trinucleotide frequency.*
16.	Codon frequency, separate reading frames	*Lists trinucleotide frequency in one reading frame and corresponding amino acid distribution.*

investigator to learn the system while using it, and to learn only those sections of the complete system needed to analyze the particular problem.

The program screens input for impossible values. In some cases, automatic corrections are made. In all cases, the user's attention is called to the erroneous input, and the program provides another opportunity to supply values. When an analysis does not provide the specific answers an investigator requires, the interactive system provides opportunities to perform a repeat analysis, without the considerable computer overhead involved in re-

TABLE 2

PARAMETERS OF THE PROGRAM

Parameter	Range	Default[a]	Function
NUMSEQ	1-	b	Maximum number of entered sequences
MAXLEN	1-32,000	1000	Maximum length of entered sequences
NUMRES	0-	200	Maximum number of entered subsequences for Procedure 14
MAXLENRES	1-32,000	20	Maximum length of entered subsequences for Procedure 14
SHORT	0,1	1	SHORT = 1 compresses computer output
SIMPLE	0,1	0	Controls format of sequence printing
GAP	0-	0	Number of blanks inserted between concatenated sequences
NUMAGREE	3-	c	Number of matches required by Procedures 8 & 9
MINMATCH	3-	3	Minimum number of matches in a homology
*DISTANCE	0-	e	Maximum distance between repeated regions
LOOPOUT	0-3	3	Maximum length of a loopout (used to be (LOOPLENGTH)
*EXPECT	0-	d	Sets MAXPROB to expect this number of chance holomogies
*MAXPROB	.00002	.002	Maximum probability of chance occurrence of homology
MINRATIO	.5-1.0	.75	Minimum ratio of matches to length
AFTERMISS	0-	2	The number of correct matches which have to follow a mismatch among the next 3 pairs in a homology.
LOOPDIST	0-	20	Maximum length of central loop in dyad symmetry
GTPAIR	0,1	0	GTPAIR = 1 allowes G-T matches in dyad symmetries (used to be DUBIOUS)
PHASE	1-4	1	Coding frame in Procedures 13 and 16 (4 = all frames)
MISSRES	0-	0	Number of mismatches allowed by Procedure 14
LOOPMIN	0-	0	Minimum length of central loop in dyad symmetry
FEW	0,1	0	FEW = allows interactive use of Procedure 14

a The default value is the value chosen by the program when execution begins. Use parameter menus to determine current value of a parameter.

b Set by SETUP on the assumption that each File contains one sequence

c Chosen to produce a moderate amount of output

d Chosen according to the value of MAXPROB when EXPECT = -1.

e No maximum is placed on the distance when DISTANCE = -1.

establishing the environment (reading sequences from storage, loading the program into main memory, etc.). Fast results enable the investigator to correct errors in specifying analysis procedures and to use the results of one analysis to plan another (Fig. 1A).

Improvements on the Queen-Korn Program

In addition to making the program interactive three major features have been incorporated.

New probability routine. The Queen-Korn program has the capability of selecting homologous regions in two sequences, palindromic subsequences, self-complementary regions and repetitions (see Table 1). These sequence features may all be imperfect, i.e. need not match perfectly. With each feature located, the program prints the probability that it would occur by chance alone. In order to reflect the actual probability of a given sequence match occurring by chance alone, a new procedure for computing probabilities has been incorporated into the present program. The probability routine used in the original program reported a probability based on an even distribution of nucleotides, regardless of the base composition of the sequences or of the stringency of search parameters. The current version incorporates a routine developed at Stanford University (D. Brutlag, personal communication) which uses a more sophisticated approach. Search parameters and sequence features are used to determine the transition matrix of a Markov chain. Probabilities are generated for each search performed. The following factors influence the transition matrix used: the base composition of each sequence, the number of consecutive misses allowed, the length of allowed loopouts, the minimum ratio of misses to total length, the length of the total match and the number of misses. No value is reported for extremely long matches, as the probability is extremely low.

Minimum loopsize. Self-complementary regions are potentially capable of forming a stem and loop structure. A new parameter, LOOPMIN (see Table 2), enables the investigator to specify the minimum length of the loop thereby supplementing the parameter LOOPDIST, which specifies its maximum length. Thus it is now possible to search a sequence for self-complementary regions with a precisely defined loop size (Fig. 1B).

Location of oligonucleotides and oligopeptides. Procedure 14 (see Table 1), which scans nucleic acid and protein sequences for short oligomer sequences, is especially useful for simulating restriction endonuclease digests. Three major changes have been made in this routine in order to make it more useful in a conversational environment. First, the code which per-

```
SEQUENCE:test1

PROCEDURES:14 0

ENZYME:'psr 1'


PSR 1 DOES NOT MATCH THE NAME OF ANY OLIGONUCLEOTIDE SPECIFIED.
WOULD YOU LIKE TO SEE A LIST?:no

ENZYME:'pst 1'

                    TEST1                                          PAGE  1

                                  # OF
                                  SITES   SITES          FRAGMENTS     FRAGMENT ENDS

PST 1 (CTGCAG)                      1
                                            85          177C (99.9)      85     85
                                                         92L (52.0)       85    177
                                                         85L (48.0)        1     85


SHOW RESTRICTION SITES?:yes


PST 1
        10          20          30          40          50          60
ATTGGAATAA TATATATAGT GCAATTATAA GAACAAGTCG TCTACGGCCA TACCTAGGCG
        70          80          90         100         110         120
AAAACACCAG TTCCCGTCCG ATCACTGCAG TTAAGCGTCT GAGGGCCTCG TTAGTACTAT
       130         140         150         160         170
GGTTGGAGAC AACATGGGAA TCCGGGGTGC TGTAGGCTTC TTTTTTTTAA ATTCCAA
```

(A)

```
SEQUENCE:fiction

PROCEDURES:1 12 0

                    FICTION                                        PAGE  1

THE NUCLEOTIDE SEQUENCE IS:


        10          20          30          40          50          60
AGCTCCAGAA ACCTGTGTGG GTTTACGTCG TAGCGTAGCA GTPAPCGGAT GCGATGATGC
        70          80          90         100
GGPGGPAAAA CGTCGTGCAT TTTTCCCQCC GCATCATGCA TCACTG

                    FICTION                                        PAGE  2

THE DYAD SYMMETRIES ARE:



    20      GGTTT      24
    13      CCAAA       9
RATIO                1E+00
PROBABILITY          1E-03
EXPECTED NUMBER      6E-01

    81      TTTTCCCQCCGCATCAT GCATC    102
    70      AAAAPGGPGGCGTAGTAGCGTAG     48
RATIO                9E-01
PROBABILITY          6E-09
EXPECTED NUMBER      2E-06

 THE NUMBER OF MATCHES IS   2

                              ANALYSIS PARAMETERS

10: MINMATCH(  3)                        11: DISTANCE( 106)
12: LOOPOUT(   3)                        15: MINRATIO(0.750)
13: EXPECT(    0)                        14: MAXPROB(0.00199)
17: LOOPDIST( 20)                        18:  GTPAIR (0)
23:  LOOPMIN( 0)
```

(B)

Figure 1

This represents portions of DNASYS sessions. Program output is upper case, user response is lower case.

(A) Restriction enzyme analysis. This session shows error correction (the enzyme name was misspelled) and tutorial functions (following the ?).

(B) Dyad symmetries. Program output includes current values of all relevant variables. Probability calculation is adjusted for these values.

forms the actual comparisons has been re-written to require significantly less processing time. Second, an interactive facility (parameter FEW, see Table 2) has been incorporated which enables the investigator to specify that a sequence be searched for the sites of only one or a few recognition sequences out of a list, rather than performing an exhaustive search for all defined sites. Third, some enzymes have specificity for two different nucleotides at a particular position. This program has expanded the vocabulary of nucleotides to include new special characters, R, S, V, and W, which specify either A or T, either G or C, either A or C, and either G or T, respectively. When combined with the special characters P (purine), Q (pyrimidine) and N (any nucleotide), it is now possible to specify all symmetrical endonuclease restriction sequences.

Input and Output Flexibility

Sequence information can be prepared using any standard editor. (An editor is a program maintained by the host installation which provides a facility for entering, manipulating and staring data. Editors also allow correction of typographical errors in this input. The user manual includes instructions for the IBM TSO editor.) Sequences need only be entered once. The system will catalog and store the information indefinitely.

Output of the analysis is handled separately from control dialogues. It is possible to route analysis to the terminal, to a high-speed printer or to a disk dataset for subsequent examination. Thus it is feasible to use a video screen for controlling the session and a separate printer for hardcopy printouts of the analysis results. This feature is particularly valuable when large sequences are analyzed, producing long output listings. Analysis output format can also be controlled by the user through the use of control parameters.

Portability

The main program is written in PL/I. Any computer which has a PL/I compiler available (e.g. IBM/370 and 4300-series computers) should be able to support it.

A short supplementary program (also in PL/I) which handles TSO file allocation has also been written. Minor modifications will allow this second program to be used with the VM/CMS operating system. It is independent and may be replaced for use on other computers without revising the main program.

Efficient compilation requires in excess of one million bytes of main storage, but compilation need only be done once. For those investigators with compatible systems, compiled code is available. Execution requires

350K bytes (IBM/370 OS/VS 2). The user manual includes some suggestions for decreasing storage requirements, if this is needed.

Distribution

Copies of this program and a User Manual are available from Dr. F. Ruddle, Department of Biology, Yale University, P.O.Box 6666, 260 Whitney Avenue, New Haven, Connecticut 06511 USA.

ACKNOWLEDGEMENTS

We are grateful to Dr. D. Brutlag at Stanford University for supplying the algorithm for sequence probability computation and for helpful ideas and critical discussions. We are also grateful to the User Services staff at the Yale Computer Center, particularly G. Moss, for technical advice and critical comments during program development. We also acknowledge the use of the SUMEX-AIM facility. We thank Dr. T. Platt for his helpful suggestions for writing the manuscript.

* An interactive sequence analysis system based on Queen and Korn (1) written in SAIL and running on the SUMEX-AIM facility at Stanford has been devised by Dr. D. Brutlag.

§ Current Address: Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, MA 02139.

REFERENCES

1. Queen, C.L. and Korn, L.J. (1980) Methods in Enzymology 65, 595-609.
2. Korn, L.J., Queen, C.L. and Wegman, M.N. (1977) Proc. Nat. Acad. Sci. 74, 4401-4405.
3. Staden, R. (1977) Nucl. Acids Res. 4, 4037-4051.
4. Staden, R. (1978) Nucl. Acids Res. 5, 1013-1015.
5. Staden, R. (1980) Nucl. Acids Res. 8, 817-825.
6. Gingeras, T.K. and Roberts, R.J. (1980) Science 209, 1322-1325.