Volume 10 Number 1 1982 Nucleic Acids Research

Nucleic acid sequence database computer system

B.C.Orcutt, D.G.George, J.A.Fredrickson and M.O.Dayhoff

National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road,N.W., Washington, D.C. 20007, USA

Received 29 September 1981

On September 15, 1980, the Nucleic Acid Sequence Database Demonstration Project of the National Biomedical Research Foundation was made available to interested users through telephone access to our computer (1,2). Over two hundred user groups requested access during the ten months of the demonstration. We are continuing the database as a computer "publication" on a subscription basis. Both magnetic tapes and direct "on line" access to our computer are provided. In this paper we describe the computer system that evolved over the first year of operation. In its development, we have benefitted from the questions and comments of the users and we have endeavored to attune the system to their interests and needs. We had been using the computer system ourselves for some time and had found that a computerized management system was essential to minimize the overall cost of collecting, updating, and critically reviewing the data. In preparing the information, we were the heaviest users of the system.

For the Demonstration Project, we made a diligent effort to collect all published sequences of over 500 nucleotides in length. We also included sequences shorter than this such as the functional RNA molecules that were published in the Atlas of Protein Sequence and Structure. Over 227,000 nucleotides from 255 sequences published and reaching our offices by August 1, 1980, were included initially. This database has grown to 557,000 nucleotides and 500 sequences in less than a year. At the outset, the longest sequence was the genome of bacteriophage fd with 6,408 nucleotides. Now the whole human mitochondrion of 16,569 nucleotides is known and there are seven entries with sequences longer than bacteriophage fd (Table 1). Because of the advances in nucleic acid sequencing techniques (3,4,5), this exponential growth in sequence information is expected to continue. Table 2 shows the distribution of sequences among the major groups of species in our September 1981 database.

This is a third generation computer database available in an interactive

Table 1. Longest sequences in the September 1981 database

Total number of sequences:	500
Total number of residues:	556,828
16,569	Genome - Human mitochondrion (SGC1)
12,100	5' genetic end of genome - Bacteriophage T7
10,168	oxi 3 (cytochrome oxidase) gene - Yeast mitochondrion (SGC2)
9,302	Genome (RNA) - Rous sarcoma virus
8,332	Genome (RNA) - Moloney murine leukemia virus
8,024	Genome - Cauliflower mosaic virus
7,433	Genome (RNA) - Poliovirus
6,408	Genome - Bacteriophage fd

Table 2. Biological group summary

Group	Number of Sequences	Length of Sequences
Total	500	556,828
Eukaryotes	296	280,449
Mammals	129	146,884
Plants and fungi	80	68,107
Eukaryotic viruses	54	129,924
Prokaryotes	125	96,397
Bacteriophages	25	50,058
Animal viruses	46	117,048
Plant viruses	8	12,876
Escherichia coli	82	78,805
Fungi	60	60,510
Human	42	57,910
Mitochondria	22	62,013
Chloroplasts	8	4,509

mode. On the first level, as has been widely available for many years, the system allows the user to retrieve bibliographic information on authors, references, titles, and abstracted text words. Secondly, the user has access to scientific data that have been abstracted, correlated with available data, and critically reviewed. When redundant information is already available, the new material is added to the old. One-third of the entries in the database contain overlapping sequences from more than one reference. In the last update, 20% of the sequences that we entered initially overlapped sequences already in the database. Only the reference, a textual description of the work done, and the differences from the sequence shown appeared in the final entry. Thirdly, the user can manipulate the scientific data through computer programs and commands to understand the information content and can do computer thought experiments to gain new insight.

The database computer system is directed toward efficiently providing the

user with an awareness of what has been done, permitting the analysis of new sequences and their comparison with those already elucidated, and facilitating the transmission by telephone of selected portions of the database for further analysis on the user's own computer. Clearly, our VAX-11/780 could not support all of the computer needs of research workers in this area in the whole United States. We therefore restricted the kinds of programs available to those most closely linked to obtaining the data in computer-readable form. Operations that require a great deal of computer time, such as searches of the whole database, comparison of all segments within sequences, and statistical tests that require comparisons of randomized sequences were not permitted. Even though many people interrogated the system each day, there was seldom any overload of our telephone lines.

In designing the database, we aimed for a system from which maximal relevant information was obtainable by the user in minimal time, using minimal documentation. Human engineering factors, including how and what was done, were carefully considered. Learning the system is easy (everything works in a similar, intuitive manner) and help is readily available. User suggestions are encouraged, and new features are tested and adapted. The system provides for the efficient updating of the information and modification of the system functions and applications programs.

In the course of the first year of operation, there has been a coevolution of the computer system with the content and format of the information base. We have developed a uniform format so that frequently recurring information can be interpreted by the computer programs. From the users' point of view, the information about a sequence is stored in one entry. In addition to the sequence, each entry has the following: a code, which is the key for retrieving the entry; a title, which includes a nucleic acid name and a species name; references to papers describing the experimental work; comments; and a feature table showing the functional regions of the sequence. The material in the feature table is interpreted by several commands of the retrieval system. All of the information is carefully checked for conceptual errors and format irregularities, so that effective retrieval and manipulation can be performed using computer programs.

The main retrieval program of the system is the nucleic acid query program (NAQ). The commands of this program and other ancillary programs of the system are designed so that similar nucleic acid sequences can be aligned readily, protein sequences or the complements of nucleic acid sequences can be constructed, stored, and aligned, and feature tables can be examined. A

diagrammatic view of the features can be produced. Sequences can be searched for short segments of interest. Provision is also made so that users can enter sequences in computer-readable form and examine them in our system before submitting them for publication in journals and/or in the collection. They can see the appearance of the entry and compare the sequence against other sequences in the collection for similarity or possible overlap.

In addition to the specialized system for manipulating the information in the entries, we also have a relational database, not available to our users, in which are stored a number of additional kinds of information such as the higher taxonomic groups, the date of publication of the sequence, the date of the last update of the entry, the order of the entries in the collection, the nucleotide composition and length of each sequence, and information on the verification of the sequence in other laboratories. Information generated from this database is available to the users.

The general scheme for the flow of information in the NAQ program is shown in Fig. 1. Central to this program is the "current list" of entries. This list can be produced by the retrieval commands operating on the main information base or can be generated from an external file containing entry codes, such as those produced by the relational database. The codes of entries on the current list can be stored in a file for future use. For example, a current list of entries that were new last month, modified last month, worked on by a particular author, isolated from prokaryotes or from yeast, or the entry for a

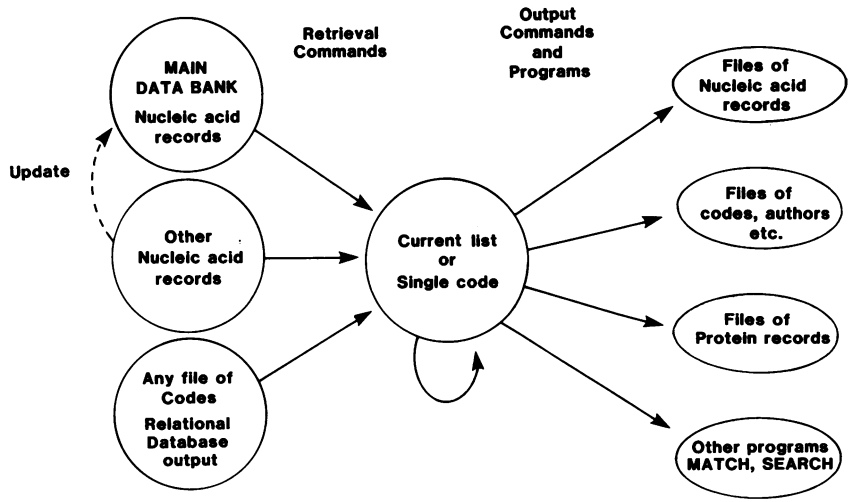


Figure 1. General scheme for the flow of information in the NAQ program.

particular ribosomal RNA sequence that might be of interest can be generated. The retrieval commands allow the database to be searched for a subset of entries with particular strings of characters in their titles, their references, their text lines, or their authors. The current list can be modified by retrieval operations that add to it or subtract from it and entries can be removed by direct editing of the list. The current list can then be used to produce output of various kinds using the NAQ commands. The total information for each entry of the current list as well as lists of titles, of authors, and of journal citations and the results of text searches and sequence matches can be produced for the entire list. Protein sequences can be translated automatically and stored in a file using information in the entry feature tables, and the results of searching sequences of the current list for matches to a short sequence can be displayed.

We shall now briefly describe the instructions for using the computer system and the structure of the system as it existed in September 1981. The following commands are recognized by the database monitor:

HELP	Obtain help on using the system
NAQ	Invoke the nucleic acid query program
INDEX	Display the index of system files
TYPE	Display the contents of a file
MATCH	Invoke the sequence alignment program
DIRECTORY	Display list of files owned by the user
EDIT	Invoke the editor to create or modify files
DELETE	Delete files owned by the user
MAIL	Invoke the mail program
VIDEO	Set terminal characteristics to those of a video terminal
LOGOUT	Log off the system

As a general rule, the system does not distinguish between uppercase and lowercase letters. Commands may be abbreviated to as few letters as necessary to make them unambiguous. Control characters can be used to control the flow of information to and from the computer during the execution of any command.

OBTAINING HELP

There are two major barriers to using a data retrieval system: the time required to master a large instruction manual and the time required to find the appropriate suggestion when help is needed. In our system, the instruction manual is very concise. Help may be obtained from the computer, as needed, on any of the programs or commands recognized by the database monitor.

Within a program the computer guides your usage through prompts. Invalid responses elicit a ? from the system. Each program has its own help facility that describes the possible responses to any prompt. Furthermore, the NAQ

program has many commands that accept one or more modifiers. If you type a command line containing a command or modifiers that are not valid, the computer will respond by putting a ? under any characters that are not appropriate. You can then modify your command line.

THE NAQ RETRIEVAL PROGRAM

The NAQ retrieval program recognizes twenty commands. The commands are one-letter abbreviations of the full command name. In Table 3 the commands are divided into five categories determined by the primary function of the

Table 3. Commands of the NAQ program

Text searching commands:		
F	FIND	Search for sequence and organism names
S	SEARCH	Search titles, references, comments, and feature tables for text strings
A	AUTHOR	Search references for author name
Display commands:		
T	TYPE	Display entry
D	DECIDE	Selective display or deletion from the current list
V	VIEW	Display view of sequence features
L	LIST	Display titles of entries on the current list
J	JOURNAL	Display list of journals and number of citations
I	IDENTIFY	Display citation for the database
Sequence search and manipulation commands:		
M	MATCH	Search a sequence for a string of nucleotides
U	USAGE	Compute and display codon usage tables or mono-, di-, and trinucleotide frequency tables
X	EXTRACT	Splice segments of a sequence or its complement to produce nucleic acid sequence or its protein translation
Y	HYPOTHETICAL	Searches sequence and complement for possible protein coding regions
Z	ENZYME	Searches sequence for restriction enzyme cut sites
Interface to other programs:		
C	COPY	Copy entry to an external file
G	GET	Get list of codes from an external file
Miscellaneous commands:		
H	HELP	Obtain help from the system
P	PRINT	Display the contents of a file
Q	QUIT	Terminate NAQ
O	OPTION	Set optional terminal width

Each execution of a text-searching command, A, F, or S, generates a new (or modified) current list. Most of the other commands accept a modifier that instructs NAQ to perform the command operation on each entry on the current list.

command. The sequence search and manipulation commands and the T, V, and C commands will also operate on sequences stored in user files. A single command may actually perform several functions. One-character command modifiers are used to select the function of interest. Command modifiers, given in Table 4, can be entered in any order after the command on the command line.

A brief description of the commands will now be given.

AUTHOR (A) searches the author names for all occurrences of a user-specified name or partial name. For each name found, the author, the code, and the title for each entry referencing the author are displayed. Brief output of this command is shown in Fig. 2a.

Table 4. Command modifiers of the NAQ program

Command		1	2	3	4	5	6
A	AUTHOR	C + -	K	B			P
C	COPY	C					
D	DECIDE			R S	M B		
F	FIND	C + -	K				P
G	GET	C + -	K				
H	HELP						P
I	IDENTIFY						P
J	JOURNAL	C			N		P
L	LIST	*				O	P
M	MATCH	C		B	M		P
O	OPTION						
P	PRINT						P
Q	QUIT						
S	SEARCH	C + -	K	B			P
T	TYPE	C		R S	M B		P
U	USAGE	C		B F	T N		P
V	VIEW	C					P
X	EXTRACT	C		A	T	O	P
Y	HYPOTHETICAL	C		A	T	O	P
Z	ENZYME	C		B	L S N		P

Column 1: Current list command modifiers. The modifier C, for CURRENT, instructs NAQ to carry out the specified operation on the entries on the current list; +, for ADD, causes the entries found to be added to the current list; -, for SUBTRACT, causes the entries found to be subtracted from the current list; and * causes the command to act on the entire database.

Column 2: The modifier K, for KEEP, permits execution without changing the current list.

Column 3: These modifiers either restrict the output to a subset of the normal output or cause additional output to be displayed.

Column 4: These modifiers modify the output that is normally produced or the function of the command.

Column 5: The modifier O, for OUTPUT, instructs NAQ to copy information into an output file.

Column 6: The modifier P, for PRINT, instructs NAQ to direct the terminal output into the print file. This modifier is ignored if the user is accessing the database by telephone.

```
*AB
Author: <RET>
Aaronson,S A
AB,G
Abelson,J N
Adelman,J
Adler,C J
Agarwal,K L
^O
1394 authors found
*
```

(a)

```
*JN
768 citations in 44 journals.

 117 Nucl. Acids Res.
 106 Nature
 100 Cell
  82 Proc. Nat. Acad. Sci. USA
  73 J. Biol. Chem.
^O
*
```

(b)

Figure 2. (a) The brief form of the author index, including a count of the number of authors cited in the September 1981 database. (b) Journals and the number of citations to each in the September 1981 database.

COPY (C) copies an entire entry (title, sequence, and references) to an output file. The file is independent of the database files; therefore, the information in the file can be modified for a particular use.

DECIDE (D) selectively displays or deletes entries from the current list.

FIND (F) searches the nucleic acid part of the entry titles for all occurrences of a user-specified nucleic acid name or partial name and the species part for the organism name or partial name. All entries with titles that contain both names are found and their codes and titles are displayed.

GET (G) reads a file of codes and uses these codes to create or modify the current list.

HELP (H) displays information about the NAQ commands, command modifiers, and special topics.

IDENTIFY (I) displays the citation for the current version of the database, the number of sequences, and the number of residues contained therein.

JOURNAL (J) displays the list of journals cited in the database and number of citations to each. The list is arranged in alphabetical order. Optionally, the list may be arranged in decreasing order of the number of citations. The list from the current database is shown in Fig. 2b.

LIST (L) displays the title information (code, sequence name, and organism name) of each entry on the current list. This command can optionally be used to create a user file containing the codes of the entries on the current list.

MATCH (M) searches the sequence for all segments that match a user-specified string. The display shows the number of the first nucleotide in the matching segment, the ten nucleotides preceding the matching segment, the

matching segment, and the ten nucleotides following it. The maximum length of the search string is 30. Optionally, MATCH will search for nonexact matches. In this case, the output includes the number of mismatches for each segment found. When searching for nonexact matches, the user can specify one or more portions of the search string for which an exact match is required. In addition to the nucleotides A, C, G, and T/U, the MATCH program recognizes an ambiguous code (Table 5).

OPTION (O) adjusts the length of the lines displayed by NAQ to fit the user's terminal.

PRINT (P) displays the contents of a file. User-created files as well as those stored in the database can be displayed.

QUIT (Q) terminates NAQ and returns to the database monitor.

SEARCH (S) searches the selected reference or text lines of entries for occurrences of a user-specified string. Entries with lines that contain the string are found, and the title information and the lines are displayed.

TYPE (T) displays the complete information for an entry: title, reference, comments, feature table, and sequence. The command modifiers allow selected portions of this information to be displayed.

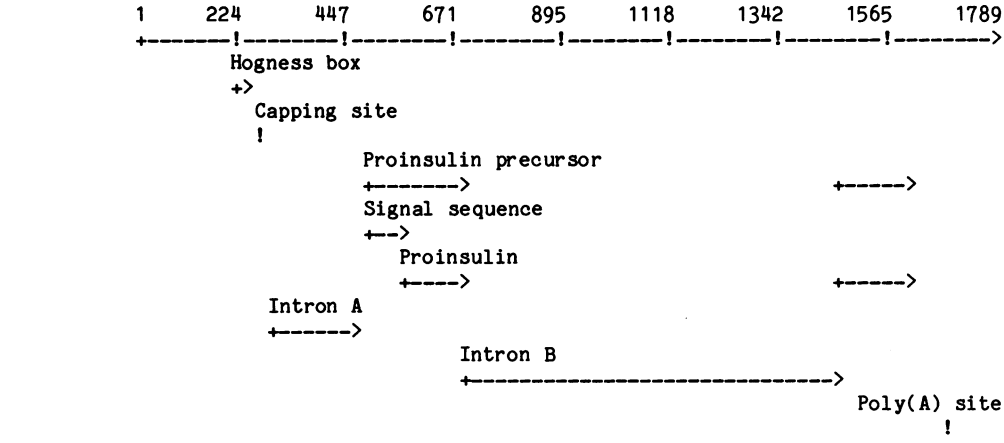
USAGE (U) displays codon usage tables. In the basic mode the coding regions are specified by the user. Optionally, the regions can be read from the entry's feature table. Three types of tables may be displayed: (1) totals for each coding region, (2) accumulated totals for all coding regions in an entry, and (3) accumulated totals for all coding regions in all entries on the current list. Codons that contain ambiguous nucleotide symbols are not counted. The amino acids corresponding to the codons reflect the genetic code used by the organism or organelle. The N command modifier displays tables of mono-, di-, and trinucleotide frequencies.

VIEW (V) displays a graphic view of the sequence features mapped from the feature table of the entry. The view of the human proinsulin sequence entry is shown below.

Table 5. Ambiguous nucleotide code

R = A,G	S = A,C	J = C,G,T/U	N = A,C,G,T/U
Y = C,T/U	W = G,T/U	K = A,G,T/U	
Z = C,G	Q = A,T/U	L = A,C,T/U	
		M = A,C,G	

IPHU
Proinsulin precursor gene - Human



EXTRACT (X) constructs sequences by splicing segments from a sequence or from its complement. The nucleic acid sequence constructed or its protein translation can be obtained. In the basic mode the segments to be spliced are specified by the user. Optionally, the segments that code for proteins can be read from the entry's feature table and their translations can be displayed. Protein translations reflect the genetic code used by the organism or organelle.

HYPOTHETICAL (Y) searches a sequence for hypothetical proteins, coding regions that begin with an initiator codon and end with a terminator codon. The minimum acceptable protein length is specified by the user. The program also identifies fragments of acceptable length that are coded from the ends of the nucleotide sequence.

ENZYME (Z) searches a sequence for one or a list of restriction endonuclease recognition sites and displays the length and the first and last residue numbers of each fragment for each enzyme. Output can be limited to enzymes giving a specified number (or a range) of cuts using the N modifier and to enzymes that do not cut in a specified area using the U modifier.

THE INDEX AND TYPE COMMANDS OF THE SYSTEM MONITOR

The INDEX command results in a display of the names and a brief description of the contents of the files that are stored in the system. Many of these are produced by the relational database, to which the user has no direct access.

The TYPE command followed by the name of a file results in a display of the contents of the file. User-created files as well as those stored in the database system can be displayed.

THE MATCH PROGRAM

The MATCH program is useful for comparing sequences that are quite similar. The best alignment of the two sequences is calculated in sections of 400 nucleotides. The first 400 nucleotides of each sequence are aligned to give the best score (Needleman-Wunsch algorithm), giving positions with identical residues (or ambiguous residues that could match) a score of 1 and other positions zero. Each break in either sequence contributes a negative score equal to the gap penalty. A position where a residue matches a break in the other sequence contributes zero score. For very similar sequences, this alignment is usually unique. After a section of 400 residues is completed, the program looks back from the end of the alignment and finds the first place where a designated number of contiguous residues (typically 7 or more) match. Starting with the first of these contiguous residues that match, the next 400 residues are selected from each sequence and the segments are aligned as before, but with the condition that the first residues match. This process is repeated until all nucleotides of one of the sequences are exhausted. Because the algorithm is not performed over the entire two sequences as a single process, it does not guarantee the best match score. Typical output of this program is shown below, where the 5' end of the 16S ribosomal RNA sequence from Escherichia coli (RR16EC) is compared with that from maize chloroplast (RD16MC).

```
      261       270        280        290        300        310        320        33
        !         !          !          !          !          !          !
RR16EC  AACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACAC
          **   *   *       *    **         *          *          *
RD16MC  AATAGCTTACCAAGGCGATGATCAGTAGCTGGTCCGAGAGGATGATCAGCCACACTGGGACTGAGACAC
        !         !          !          !          !          !          !         !
      232       240        250        260        270        280        290       300

                340        350        360        370        380        390       40
                  !          !          !          !          !          !
RR16EC  GTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGC
         *                                  *  *  *          *          *  *   *
RD16MC  GCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATTTTCCGCAATGGGCGAAAGCCTGACGGAGCAATGC
                !          !          !          !          !          !         !
              310        320        330        340        350        360       370
```

18 mismatches out of 140 possible matches between residues

THE EDIT PROGRAM, YOUR PERSONAL ENTRY FILES

The user can create permanent files of entries not in the collection, which can be examined with some of the programs and commands of the system. The X command of NAQ can be used to extract and concatenate pieces of the sequence, of the complement, or of the protein translation. These sequences can be stored and examined and can be used by the NAQ and MATCH programs. The EDIT command can also be used to transmit files from the user's computer storage to our computer.

THE DIRECTORY, DELETE, AND VIDEO COMMANDS AND THE MAIL PROGRAM

The DIRECTORY command results in a display of the name and date of creation of each file created by the user, the DELETE command is used to delete user-created files that are no longer being used, and the VIDEO command sets the characteristics of the terminal to those of a video terminal rather than a printing terminal. The MAIL program enables a user to communicate with other users and with the system staff.

UPDATING THE SYSTEM

In entering new data and in making corrections, each scientist analyzing data works with complete entries, creating new ones, correcting old ones, or deciding to delete old ones. Periodically, an update file is created by copying into one file the update entries that have been prepared by many individuals. Updates of the information base are made in a batch for economy both in the computer time involved in creating the inverted files and the personnel time in checking against the original information sources and in correlating new entries with old for redundancy, style, and content. Because of the nonuniform nature of the information, we discuss each update file to formulate additional editorial policy. At this point, changes to the text of the update file can be accomplished by editing this file directly.

The database consists of three primary files (a file containing code, title, and sequence for each entry; a file containing code, title, references, comments, and the feature table for each entry; and the relational database master file) and three auxiliary files (an index file and two inverted files).

During an update, the information in the primary files is modified and the index and inverted files are generated for the NAQ retrieval program. Also, a log file is created containing a record of all changes made during the update.

EXAMPLES OF MANIPULATION OF THE NUCLEIC ACID DATABASE INFORMATION

The NAQ retrieval program was designed for ease in exploring the nucleic

acid database and to allow easy manipulation of the data in order to answer important scientific questions. A few simple examples will be given to demonstrate the use of this system.

The nucleotide sequences of E. coli RNA polymerase promoter sites have been well studied. The promoters appear to have a recognition site specific for the sigma subunit of RNA polymerase with a consensus sequence of TG(TTG)ACA, where the nucleotides in parentheses are highly conserved. About 16 to 19 nucleotides to the 3' end of this sequence, the RNA polymerase core binds at a site with the consensus sequence (TA)QRQ(T)K (6,7). (See Table 5 for definitions of the ambiguous nucleotides Q, R, and K.)

The entry for the rpsL gene of E. coli contains the promoter region. The MATCH command can be used to search for these recognition sequences. The M command modifier searches for sequences that do not match exactly.

```
*MM
Code: RP12EC
RP12EC
rpsL and fragment of rpsG (S12 and S7 ribosomal proteins)
    genes - Escherichia coli
Search string (or E to exit): TG(TTG)ACA
Number of mismatches allowed: 1
Exact matches at       !!!
Matching-----------TGTTGACA
residue            !                    No.
     25  GTTGTATATT TCTTGACA CCTTTTCGGC  1
Search string (or E to exit): E
*
```

Since translation of the rpsL gene begins at residue 130, the matching sequence is a good candidate for the sigma recognition site. The unmodified MATCH command can be used to search for the core recognition site.

```
*M
Code: RP12EC
RP12EC
rpsL and fragment of rpsG (S12 and S7 ribosomal proteins)
    genes - Escherichia coli
Search string (or E to exit): TAQRQTK
Matching-----------TAQRQTK
residue            !
     19  GGGATCGTTG TATATTT CTTGACACCT
     50  GGCATCGCCC TAAAATT CGGCGTCCTC
    630  TTGGTCAGCG TAAAATT CTGCCGGATC
    672  AACTGCTGGC TAAATTT GTAAATATCC
Search string (or E to exit): E
*
```

The sequence starting at nucleotide 50 is 17 nucleotides from the end of the putative sigma recognition sequence and could well be the core recognition site.

Recently, there has been considerable interest in the dramatic difference

in codon usage patterns among different species (8). There are 11 yeast mitochondrial coding regions in the database. Use of the FIND command allows a current list of all mitochondrial sequences to be generated. Control O (^O) stops transmission of data to the user's terminal but allows the command to continue execution.

```
*F
Nucleic acid:
Species: Mitochondrion
HUMIT Genome - Human mitochondrion (SGC1)
^O
*
```

The FIND command with the C modifier produces a new current list containing only yeast mitochondrial sequences. These are translated using special genetic code 2 (SGC2), as noted in the title.

```
*FC
Nucleic acid:
Species: Yeast
S1MTBY  Segment I of repeated DNA sequences - Yeast
            mitochondrion (SGC2)
^O
*
```

The USAGE command computes codon usage. The C modifier restricts operation to the current list, the T modifier causes the protein coding regions to be read from the feature tables, and the B modifier displays only the cumulative data. The following table is produced in less than 15 seconds.

```
*UCBT
Cumulative statistics from 9 entries
Codon usage: 2975 codons from 11 coding regions.

    123 UUU Phe     53 UCU Ser    161 UAU Tyr     32 UGU Cys
     75 UUC Phe      3 UCC Ser     15 UAC Tyr      2 UGC Cys
    353 UUA Leu    107 UCA Ser      0 UAA Ter     59 UGA Trp
      3 UUG Leu      0 UCG Ser      0 UAG Ter      2 UGG Trp

     12 CUU Thr     58 CCU Pro     70 CAU His      4 CGU Arg
      1 CUC Thr      3 CCC Pro      4 CAC His      1 CGC Arg
     22 CUA Thr     44 CCA Pro     63 CAA Gln      0 CGA Arg
      2 CUG Thr      0 CCG Pro      8 CAG Gln      0 CGG Arg

    251 AUU Ile     58 ACU Thr    176 AAU Asn     31 AGU Ser
     32 AUC Ile      3 ACC Thr     13 AAC Asn      1 AGC Ser
     43 AUA Ile     70 ACA Thr    134 AAA Lys     63 AGA Arg
     93 AUG MET      2 ACG Thr      7 AAG Lys      0 AGG Arg

     64 GUU Val     97 GCU Ala     75 GAU Asp    149 GGU Gly
      4 GUC Val     10 GCC Ala      5 GAC Asp      1 GGC Gly
    112 GUA Val     67 GCA Ala     71 GAA Glu     41 GGA Gly
      8 GUG VAL      2 GCG Ala      5 GAG Glu      7 GGG Gly
*
```

Protein coding regions have been found in the introns between coding

regions in yeast mitochondria (9,10). Yeast mitochondria use a special genetic code. To distinguish these entries from those using a different genetic code the titles of all yeast mitochondrial entries contain "(SGC2)". This string can also be used in the FIND command to generate a current list of all yeast mitochondrial sequences.

```
*F
Nucleic acid:
Species: (SGC2)
S1MTBY Segment I of repeated DNA sequences - Yeast
          mitochondrion (SGC2)
^O
*
```

All sequences on the current list that contain introns can be found using the SEARCH command with the C modifier.

```
*SC
Line types: C
Search string: INTRON
CBBY Cytochrome b (short) gene - Yeast (Saccharomyces
        cerevisiae) mitochondrion (SGC2)
         899-3212       Cytochrome b intron
^O
*
```

Using the TYPE command the segments where the introns are located in these entries can be determined. The R modifier restricts ouput to the reference information of each entry.

```
*TR
Code: CBBY
CBBY
Cytochrome b (short) gene - Yeast (Saccharomyces cerevisiae)
   mitochondrion (SGC2)

Map position: 71.4-80.2
Nobrega, F.G., and Tzagoloff, A., J. Biol. Chem. 255, 9828-
9837, 1980
    The sequence between residues 4957 and 5064 may contain
    errors. A sequence of about 22 nucleotides between
    residues 4689 and 4690 was not determined. The amino
    acids coded by the nucleotides at the exon-intron
    boundaries are uncertain.
     Residues            Feature
     140-898,            Protein: Cytochrome b
     2316-2366,
     3100-3444
     899-2312            Cytochrome b intron
     2367-3096           Cytochrome b intron
     902-2053            Protein: Hypothetical

Composition:  2501 A,   582 C,   619 G,   2539 T
Length:       6241
*
```

Three introns were found. Using the X command the introns can be isolated from

the nucleic acid sequences. The O command creates a user file (INTRON.FIL) containing the three introns found. For example:

```
*XO
Code: CBBY

CBBY    6241 bases
Cytochrome b (short) gene - Yeast (Saccharomyces cerevisiae)
    mitochondrion (SGC2)
Complement sequence (Y/N/E): N
Segments: 899-2312
Translate to protein (Y/N): N

Composition:   598 A,   111 C,   154 G,   551 T
Length:       1414

                10         20         30         40         50
   1  AATATGGCCT TATTATTAAT TACATATGTA ATTAATATTT TATGTGCTGT
^O
Copy entry to output file (Y/N): Y
Output file: INTRON.FIL
Complement sequence (Y/N/E): E
*
```

The Y command is finally used on each of the isolated introns to search for hypothetical proteins. The A and T modifiers allow the translated hypothetical proteins and their compositions to be displayed.

```
*YAT
Code: RM21S=INTRON.FIL

RM21S 21S rRNA gene fragment - Yeast (Saccharomyces
          cerevisiae) mitochondrion, 3 strains (SGC2)
1143 nucleotides
Minimum protein length: 70
1 protein found
Search complementary strand (Y/N): Y
No proteins found on complementary strand

     Residues          Feature
     298-1002          Protein: Hypothetical protein A-235

1     143      286      429      572      714      857     1000     1143
+-------!--------!--------!--------!--------!--------!--------!-------->
              Hypothetical protein A-235
              +---------------------------------------->

      Hypothetical protein A-235 - Yeast (Saccharomyces cerevisiae)
          mitochondrion, 3 strains (SGC2)

Composition 235 residues
     7 Ala A    3 Cys C   4 His H    4 Met M   10 Thr T
     4 Arg R   10 Gln Q  23 Ile I   11 Phe F    5 Trp W
    25 Asn N   12 Glu E  26 Leu L    9 Pro P   15 Tyr Y
     8 Asp D   10 Gly G  26 Lys K   12 Ser S   11 Val V
```

```
                5         10        15        20        25        30
  1+M K N I K K N Q V I N L G P N S K L L K E Y K S Q L I E L N
 31 I E Q F E A G I G L I L G D A Y I R S R D E G K T Y C+M Q F
 61 E W K N K A Y+M D H V C L L Y D Q W V L S P P H K K E R V N
 91 H L G N L V I T W G A Q T F K H Q A F N K L A N L F I V N N
121 K K T I P N N L V E N Y L T P I S L A Y W F I D D G G K W D
151 Y N K N S T N K S I V L N T Q S F T F E E V E Y L V K G L R
181 N K F Q L N C Y V K I N K N K P I I Y I D S I S Y L I F Y N
211 L I K P Y L I P Q+M I Y K L P N T I S S E T F L K
*
```

Two proteins were found. These proteins correspond to the hypothetical proteins in the feature tables of the entries.

Finally, as an example of the Z command, we will look for a set of restriction enzymes suitable for removing intact the human hemoglobin epsilon chain gene from the 3,919-nucleotide genome segment in the database. We search for enzymes that do not cut in the coding or intron regions but do cut in from 2 to 15 other places.

```
*ZUNV
Code: HEHU
Number of sites: 2-15
Uncut segment: 2079-3496

HEHU    3919 bases
Hemoglobin epsilon chain gene - Human

                        1      970      1940      2910      3919
Enzyme      No. Cuts +--------!---------!---------!---------->
Cau II           2.                     1                    1
EcoRI            2.1                                         1
FnuD III         2.      1              1
Hap II           3.                     1  1                 1
Hha I            2.      1              1
Hpa II           3.                     1  1                 1
Mno I            3.                     1  1                 1
Pvu II           2.       1                    1
Sau I            2.                       11
SciN I           2.      1              1
Sph I            2.      2
Tth111 II        3.  1   1        1
12 enzymes found
*
```

ACKNOWLEDGMENTS

This research was partially supported by NIH Grant GM-08710 from the National Institute of General Medical Sciences and by Contract NASW3317 from the National Aeronautics and Space Administration.

REFERENCES

1. Dayhoff, M.O., Schwartz, R.M., Chen, H.R., Hunt, L.T., Barker, W.C., and Orcutt, B.C. (1980) Science 209, 1182.

2. Dayhoff, M.O., Schwartz, R.M., Chen, H.R., Hunt, L.T., Barker, W.C., and Orcutt, B.C. (1981) Nucleic Acid Sequence Database, Vol.1, National Biomedical Research Foundation, Washington D.C.
3. Sanger, F. and Coulson, A.R. (1975) J. Mol. Biol. 94, 411-448.
4. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. USA 74, 5463-5467.
5. Maxam, A.M. and Gilbert, W. (1977) Proc. Nat. Acad. Sci. USA 74, 560-564.
6. Rosenberg, M. and Court, D. (1979) Annu. Rev. Genet. 13, 319-353.
7. Pribnow, D. (1978) in Biological Regulation and Development, Goldberger, R.F., Ed., Vol.1, pp.219-277, Plenum Press, New York and London.
8. Grantham, R. (1978) FEBS Lett. 95, 1-11.
9. Nobrega, F.G. and Tzagoloff, A. (1980) J. Biol. Chem. 255, 9828-9837.
10. Dujon, B. (1980) Cell 20, 185-197.