A computer program package for storing and retrieving DNA/RNA and protein sequence data

Katsumi Isono*

Max-Planck-Institut für Molekulare Genetik, Abteilung Wittmann, Ihnestrasse 63-73, 1000 Berlin 33, FRG

Received 11 August 1983

ABSTRACT

A computer program package has been made available which contains five compound programs written in FORTRAN 10 (for DECsystem-10). Program DATBAS is for storing and improving DNA sequence data, especially those obtained by the sequencing method using M13 phages. Programs NUCDAT and PROTEN are for analyzing DNA/RNA and protein sequence data, respectively. They contain various options to help users in analyzing DNA/RNA and protein sequence data. With program NUCDAT, it is also possible to get access to the EMBO Nucleotide Data Library. This can be achieved by running program COPY to create two data files from the library. Program LITRAT enables users to prepare a scientific literature file convenient for writing scientific articles, and program STRAIN for storing information concerning bacterial and/or plasmid strains.

DESCRIPTION OF THE PROGRAMS

Program DATBAS

1) General feature of the program. The program is partly based upon the "DB-series" programs written by Staden (4) and consists of a main program and a collection of 43 subroutines. In the past years I have been constantly adding extensive modifications to make the program easier to run. It is now possible, for example, to let the program perform most of the jobs (sequence alignment, sequence display, etc.) automatically.

It is also possible that at any step of execution of the program (e.g. defining the region for display, selecting options for sequence-editing, etc.), if for any reason the user does not want to proceed further, he can stop the process by typing G (for give-up) in answer to a question asked by the computer. The program aborts the on-going process and returns to the point just before thè aborted process or to the main option shown in

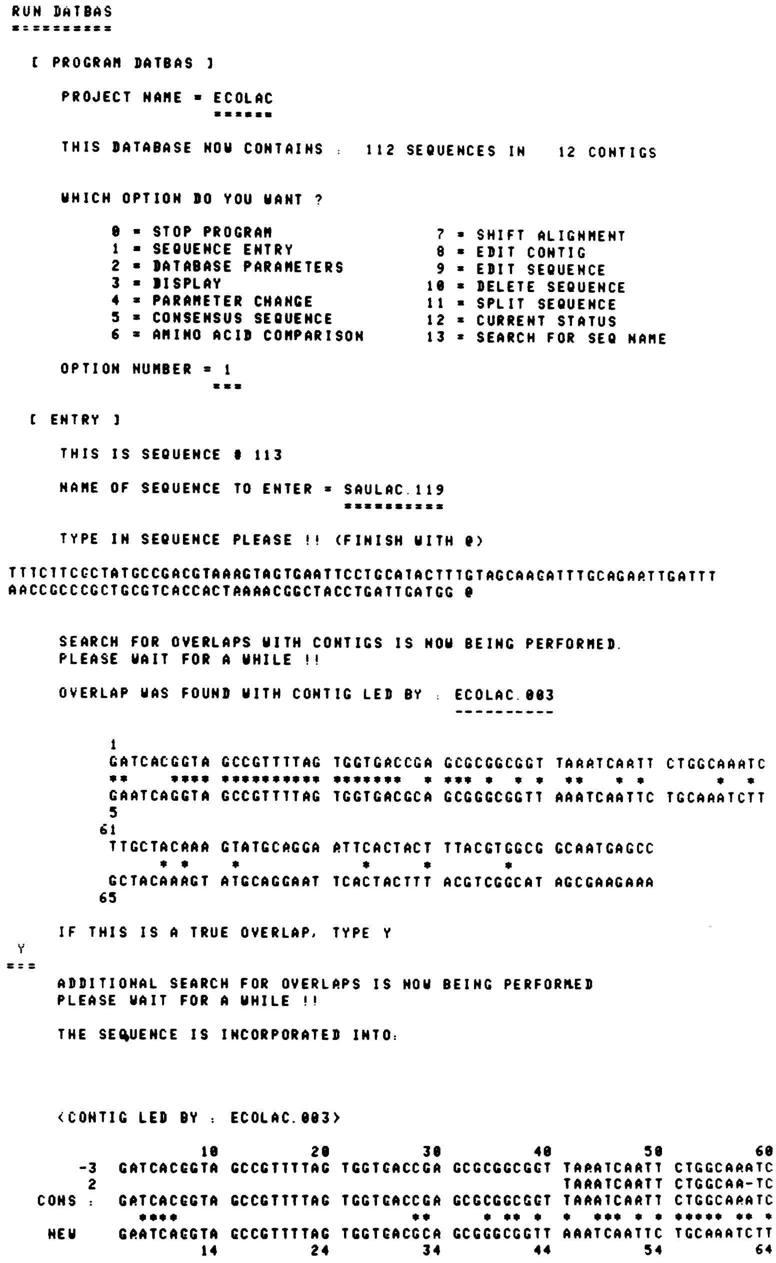
Fig. 1. This function is achieved by a subroutine termed RDOPT which first reads any numbers (either real or integer) entered by the user as alphanumerical characters, examine them and, when appropriate, converts them to numbers (real or integer) by internal data transfer.

Another modification is that the sequence data are stored after a five-fold packing as packed strings in a "random-access" file. Therefore, search for homology is now some twenty times faster than before. This makes it easier and less frustrating for the user to enter sequence data from the key-board, although ten to twenty seconds may elapse before he gets information about sequence homology when 200 sequence readings are already stored in the database.

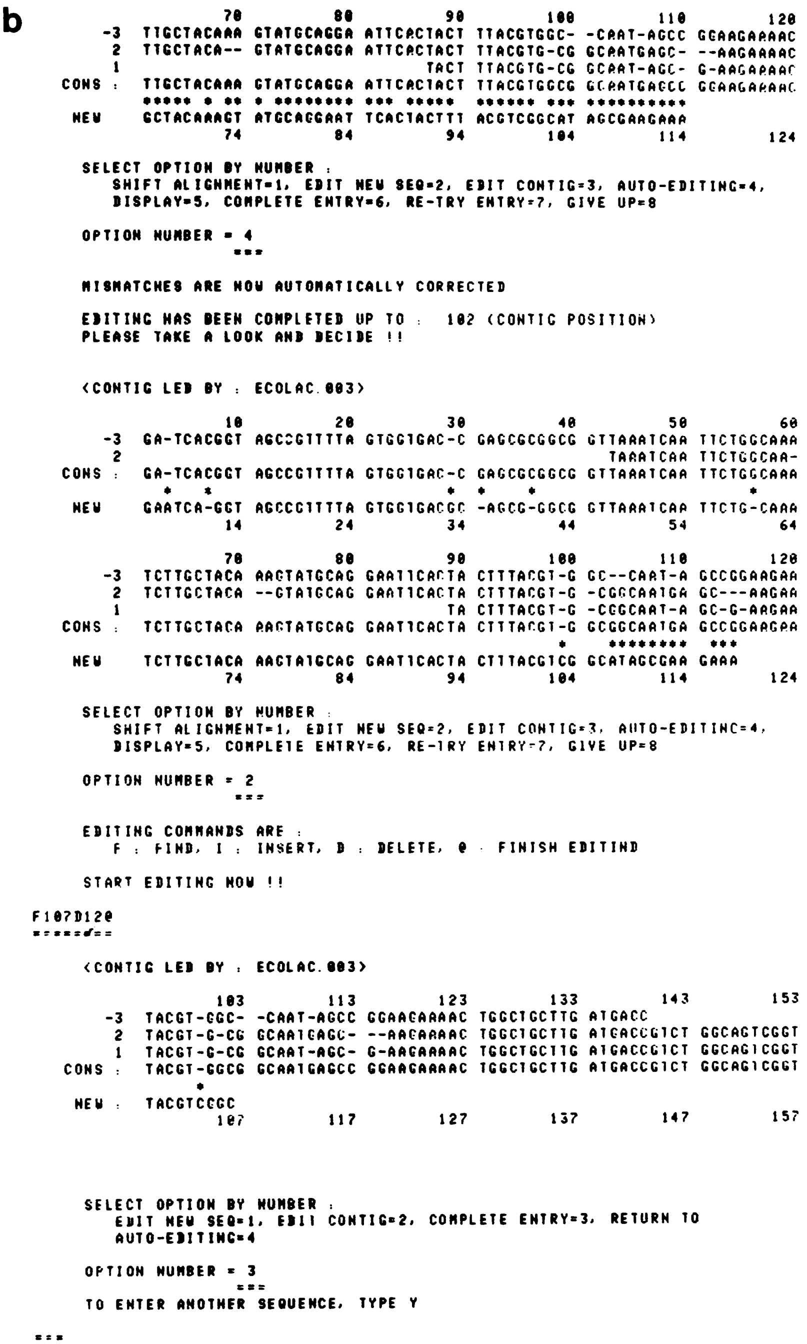
Upon execution, the program offers the user 13 options to choose (Fig. 1). By selecting one of these options, the user can store DNA sequence data (option 1), examine their quality (options 3 and 12), improve their quality (options 7 through 11) and analyze whether they code for a protein(s) or not (option 6). By selecting option 5, the user can record the consensus sequence of a contig into a disk file under the name of his choice. By running program NUCDAT (see below) he can examine this consensus sequence in various ways.

2) Entry of sequence data into a database. By selecting option 1 (sequence entry), the user can enter new sequence data from the key-board. The program then automatically searches for sequence homology with the stored sequence data unless the sequence is the first one of the database. When it does not find any homology, it prints out a message notifying the user that the new sequence does not overlap any of the "contigs" (or, groups of sequences which are related to one another; see ref. 4). When it finds one, the program displays the homology and asks the user whether it is a true match or not (Fig. 1). Here, matches (indicated by asteriks) not shorter than 12 consecutive bases are reported by the program. If it is a true match and the user answers by typing a Y, then the data are recorded in a temporal buffer and the program further searches for homology with the rest of the sequence data in the database.

a



```
RUN DATBAS
==========

  [ PROGRAM DATBAS ]

     PROJECT NAME = ECOLAC
                    ======

     THIS DATABASE NOW CONTAINS :  112 SEQUENCES IN   12 CONTIGS

     WHICH OPTION DO YOU WANT ?

          0 = STOP PROGRAM                 7 = SHIFT ALIGNMENT
          1 = SEQUENCE ENTRY               8 = EDIT CONTIG
          2 = DATABASE PARAMETERS          9 = EDIT SEQUENCE
          3 = DISPLAY                     10 = DELETE SEQUENCE
          4 = PARAMETER CHANGE            11 = SPLIT SEQUENCE
          5 = CONSENSUS SEQUENCE          12 = CURRENT STATUS
          6 = AMINO ACID COMPARISON       13 = SEARCH FOR SEQ NAME

     OPTION NUMBER = 1
                     ===

  [ ENTRY ]

     THIS IS SEQUENCE # 113

     NAME OF SEQUENCE TO ENTER = SAULAC.119
                                 ==========

     TYPE IN SEQUENCE PLEASE !! (FINISH WITH @)

TTTCTTCGCTATGCCGACGTAAAGTAGTGAATTCCTGCATACTTTGTAGCAAGATTTGCAGAATTGATTT
AACCGCCCGCTGCGTCACCACTAAAACGGCTACCTGATTGATGG @


     SEARCH FOR OVERLAPS WITH CONTIGS IS NOW BEING PERFORMED.
     PLEASE WAIT FOR A WHILE !!

     OVERLAP WAS FOUND WITH CONTIG LED BY :  ECOLAC.003
                                             ----------

          1
          GATCACGGTA GCCGTTTTAG TGGTGACCGA GCGCGGCGGT TAAATCAATT CTGGCAAATC
          **    **** ********** *******  * *** *  * *  **   * *       *  *
          GAATCAGGTA GCCGTTTTAG TGGTGACGCA GCGGGCGGTT AAATCAATTC TGCAAATCTT
          5
         61
          TTGCTACAAA GTATGCAGGA ATTCACTACT TTACGTGGCG GCAATGAGCC
               * *    *            *     *       *
          GCTACAAAGT ATGCAGGAAT TCACTACTTT ACGTCGGCAT AGCGAAGAAA
         65

     IF THIS IS A TRUE OVERLAP, TYPE Y
 Y
===
     ADDITIONAL SEARCH FOR OVERLAPS IS NOW BEING PERFORMED
     PLEASE WAIT FOR A WHILE !!

     THE SEQUENCE IS INCORPORATED INTO:



     <CONTIG LED BY : ECOLAC.003>

                  10         20         30         40         50         60
     -3   GATCACGGTA GCCGTTTTAG TGGTGACCGA GCGCGGCGGT TAAATCAATT CTGGCAAATC
      2                                               TAAATCAATT CTGGCAA-TC
  CONS :  GATCACGGTA GCCGTTTTAG TGGTGACCGA GCGCGGCGGT TAAATCAATT CTGGCAAATC
            ****                       **     * ** *  *  *** * * ***** ** *
   NEW    GAATCAGGTA GCCGTTTTAG TGGTGACGCA GCGGGCGGTT AAATCAATTC TGCAAATCTT
                  14         24         34         44         54         64
```



b

```
                   70         80         90        100        110        120
       -3  TTGCTACAAA GTATGCAGGA ATTCACTACT TTACGTGGC- -CAAT-AGCC GGAAGAAAAC
        2  TTGCTACA-- GTATGCAGGA ATTCACTACT TTACGTG-CG GCAATGAGC- --AAGAAAAC
        1                                TACT TTACGTG-CG GCAAT-AGC- G-AAGAAAAC
   CONS :  TTGCTACAAA GTATGCAGGA ATTCACTACT TTACGTGGCG GCAATGAGCC GGAAGAAAAC
           ***** * ** * ******** *** *****  ****** *** **********
    NEW    GCTACAAAGT ATGCAGGAAT TCACTACTTT ACGTCGGCAT AGCGAAGAAA
                   74         84         94        104        114        124

     SELECT OPTION BY NUMBER :
        SHIFT ALIGNMENT=1, EDIT NEW SEQ=2, EDIT CONTIG=3, AUTO-EDITING=4,
        DISPLAY=5, COMPLETE ENTRY=6, RE-TRY ENTRY=7, GIVE UP=8

     OPTION NUMBER = 4
                   ===

     MISMATCHES ARE NOW AUTOMATICALLY CORRECTED

     EDITING HAS BEEN COMPLETED UP TO :  102 (CONTIG POSITION)
     PLEASE TAKE A LOOK AND DECIDE !!


     <CONTIG LED BY : ECOLAC.003>

                   10         20         30         40         50         60
       -3  GA-TCACGGT AGCCGTTTTA GTGGTGAC-C GAGCGCGGCG GTTAAATCAA TTCTGGCAAA
        2                                                    TAAATCAA TTCTGGCAA-
   CONS :  GA-TCACGGT AGCCGTTTTA GTGGTGAC-C GAGCGCGGCG GTTAAATCAA TTCTGGCAAA
             *   *                       *  *    *                     *
    NEW    GAATCA-GGT AGCCCTTTTA GTGGTGACGC -AGCG-GGCG GTTAAATCAA TTCTG-CAAA
                   14         24         34         44         54         64

                   70         80         90        100        110        120
       -3  TCTTGCTACA AAGTATGCAG GAATTCACTA CTTTACGT-G GC--CAAT-A GCCGGAAGAA
        2  TCTTGCTACA --GTATGCAG GAATTCACTA CTTTACGT-G -CGGCAATGA GC---AAGAA
        1                                TA CTTTACGT-G -CGGCAAT-A GC-G-AAGAA
   CONS :  TCTTGCTACA AAGTATGCAG GAATTCACTA CTTTACGT-G GCGGCAATGA GCCGGAAGAA
                                                    *    ********  ***
    NEW    TCTTGCTACA AAGTATGCAG GAATTCACTA CTTTACGTCG GCATAGCGAA GAAA
                   74         84         94        104        114        124

     SELECT OPTION BY NUMBER :
        SHIFT ALIGNMENT=1, EDIT NEW SEQ=2, EDIT CONTIG=3, AUTO-EDITINC=4,
        DISPLAY=5, COMPLETE ENTRY=6, RE-TRY ENTRY=7, GIVE UP=8

     OPTION NUMBER = 2
                   ===

     EDITING COMMANDS ARE :
        F : FIND, I : INSERT, D : DELETE, @ : FINISH EDITIND

     START EDITING NOW !!

F107D120
=====/==

     <CONTIG LED BY : ECOLAC.003>

                  103        113        123        133        143        153
       -3  TACGT-GGC- -CAAT-AGCC GGAAGAAAAC TGGCTGCTTG ATGACC
        2  TACGT-G-CG GCAATGAGC- --AAGAAAAC TGGCTGCTTG ATGACCGTCT GGCAGTCGGT
        1  TACGT-G-CG GCAAT-AGC- G-AAGAAAAC TGGCTGCTTG ATGACCGTCT GGCAGTCGGT
   CONS :  TACGT-GGCG GCAATGAGCC GGAAGAAAAC TGGCTGCTTG ATGACCGTCT GGCAGTCGGT
                *
    NEW :  TACGTCGGC
                  107        117        127        137        147        157



     SELECT OPTION BY NUMBER :
        EDIT NEW SEQ=1, EDIT CONTIG=2, COMPLETE ENTRY=3, RETURN TO
        AUTO-EDITING=4

     OPTION NUMBER = 3
                   ===
     TO ENTER ANOTHER SEQUENCE, TYPE Y

===
```

If the program finds more than one contig with which the newly entered sequence shows a true match, then the program sorts the data out for merging of these contigs and performs the merge. The new sequence is incorporated into the contig during this process.

If a part of the newly entered sequence shows an overlap with the M13 phage sequence, the program reports this finding to the user so that the user can delete that part of the sequence.

When a new sequence is incorporated into a contig with which it shows a true overlap, the program displays the alignment and asks the user to choose one of the suboptions (Fig. 1). If the alignment of either the left end of the new sequence and the contig, or the left end of the contig and the new sequence (depending upon the type of overlap), is not correct, then the user must shift the aligment to the left or to the right by selecting suboption 1 (shift alignment). If it is correct, the user can choose either suboption 2, 3 or 4 for refining the data. If suboption 4 is selected, the control goes to a subroutine termed AUTOED which performs automatic refinement of the sequence data by comparing the new sequence and the consensus sequence of the contig (the sequence headed by CONS: in Fig. 1) and by inserting hyphens at positions at which the two sequences do not match.

The program first tries to skip two bases and examine whether the mismatch is a simple exchange type (such as GC versus CG, etc. which occurs rather frequently according to our experience) or not. If it fails to regain sequence matching in this way, the program further tries to shift one or two bases of either the new sequence or the contig and select the one which shows a longest stretch of matches after that point. If none of these trials is successful, the program performs the refining editing only upto that point, prints out a message and passes the control to the user. In the example shown in Fig. 1, the user decides to delete the tail (12 bases) of the new sequence, because apperently it has come from a different region or from the M13 phage genome.

Fig. 1: Entry of sequence data by running program DATBAS. All entries performed by the user are doubly underlined. For details, see text.

I have tried to perform this sequence alignment using the known diagonal comparison algorism (e.g. ref. 5), but I found it not necessary, because in most cases mistakes in sequence readings were of the types described above.

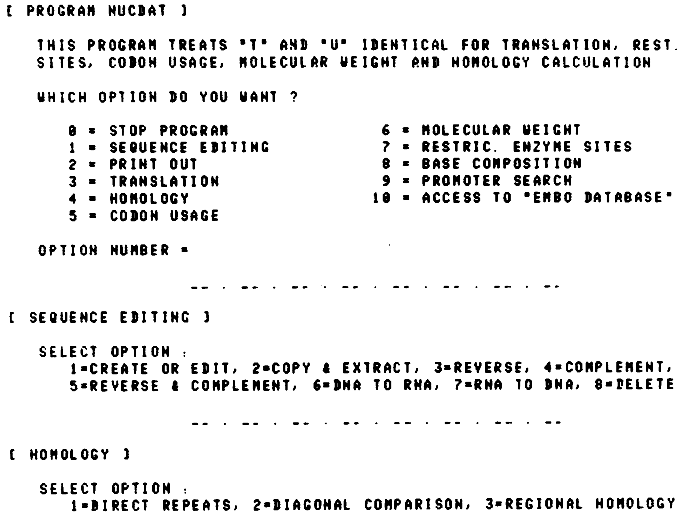
If by any reason the user has lost correct alignment between the newly entered sequence and the contig during this process, he can re-try the whole entry process described above by selecting suboption 8 (Fig. 1). All the editings performed either manually by the user (suboptions 2 or 3) or by the subroutine AUTOED (suboption 4) are canceled and the new sequence as well as the contig go back to their original state.

I have tested the above-mentioned entry step for a large number of sequence in one of our real projects and found that in most cases (48 out of 50 cases examined) they could be automatically edited to the end. With the previous programs (1,4), this process was very tedious and mistake-prone. One of the difficulties was that the information concerning sequence homology one could obtain by running a program such as DBCOMP (4) became out-dated as soon as a new sequence was incorporated into a contig, because the parameters of that contig was altered by the incorporation of the new sequence and by the various editing processes including reversion and complementation of the new sequence. Thus, display of the contig and a lot of recalculation were necessary. Another difficulty was that we had to sit in front of a teletype screen for a long time and had to carefully read the sequences on the screen to find the positions for editing. After incorporation of 10 sequences at a time, we were quite exhausted.

Program NUCDAT

This is a collection of various programs some of which were originally written by Staden (2,3). As shown in Fig. 2, this program treats T's and U's identical for various calculations.

It offers the user 10 options to select (Fig. 2). The first option (sequence editing) is further divided into eight suboptions as shown. The sequences created by the user will be recorded in disk files in much the same way as those in the EMBO Nucleotide Data Library, i.e. the sequence name is written in a



```
[ PROGRAM NUCDAT ]

   THIS PROGRAM TREATS "T" AND "U" IDENTICAL FOR TRANSLATION, REST. ENZYME
   SITES, CODON USAGE, MOLECULAR WEIGHT AND HOMOLOGY CALCULATION

   WHICH OPTION DO YOU WANT ?

      0 = STOP PROGRAM                 6 = MOLECULAR WEIGHT
      1 = SEQUENCE EDITING             7 = RESTRIC. ENZYME SITES
      2 = PRINT OUT                    8 = BASE COMPOSITION
      3 = TRANSLATION                  9 = PROMOTER SEARCH
      4 = HOMOLOGY                    10 = ACCESS TO "EMBO DATABASE"
      5 = CODON USAGE

   OPTION NUMBER =

                      -- . -- . -- . -- . -- . -- . -- . --

[ SEQUENCE EDITING ]

   SELECT OPTION :
      1=CREATE OR EDIT, 2=COPY & EXTRACT, 3=REVERSE, 4=COMPLEMENT,
      5=REVERSE & COMPLEMENT, 6=DNA TO RNA, 7=RNA TO DNA, 8=DELETE

                      -- . -- . -- . -- . -- . -- . -- . --

[ HOMOLOGY ]

   SELECT OPTION :
      1=DIRECT REPEATS, 2=DIAGONAL COMPARISON, 3=REGIONAL HOMOLOGY
```

Fig. 2: Options of program NUCDAT. For details, see text.

line headed by "ID" which is followed by a brief feature description written in a line headed by "DE", and the sequence data which are recorded in a format of 60 nucleotides per line and in groups of ten nucleotides for better visibility. Names of the user-created sequences will be also recorded in a 'random access' file termed NUCDAT.NAM and the sequence data will be recorded individually in files named NUCDAT.nnn (nnn is a three digit number). Thus, users can search for sequences of their own by names (partial or full) as well.

With the "PRINT OUT" option, users can format sequences and their translations both vertically and horizontally. If a sequence is too long to print out in one page, then it can be split into several pages (for an actual example, see ref. 6).

With option 3 users can not only translate the whole or defined regions of DNA/RNA sequences in one, two or three reading frames, but also search for "open reading frames" by limiting the translation from the initiator codons to the terminator codons. The default for the latter option are AT(U)G and GT(U)G for the initiator and T(U)AG, T(U)AA and T(U)GA for the terminators, but they can be re-defined by users if necessary. When the latter option is selected, the "open reading frames" stand out nicely in the print without the translation of unnecessary regions. Users can further examine the "open reading frames" by analyzing their codon usage with option 5, and the presence of possible promoters by selecting option 9 (see below).

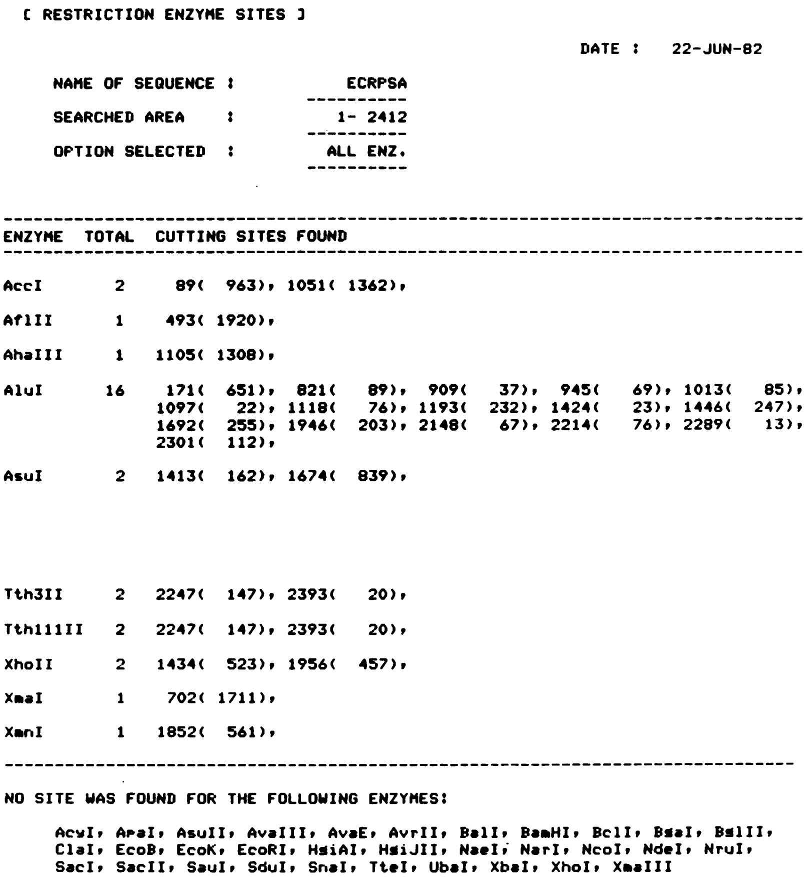
The "HOMOLOGY" option is subdivided into three further options as shown in Fig. 2. With suboption 2 (diagonal comparison), one can search for complementary regions between two sequences or within one sequence. By increasing the 'minimum stability value' for such complementary stretches, users can limit the search only for longer complementary sequences. The 'stability values' are set at 2 for a G-C pair, 1 for a A-T pair and 0 for a G-T(U) pair. The last pair only contributes to the continuation of a stretch. The results are put out in diagonal arrays of two sequences.

Fig. 3 shows an example of search for restriction enzyme cutting sites. The data stored in a file termed RESENZ.DAT which is included in the package and is accessed by the program contain cutting sites for 98 different enzymes. If necessary (e.g. when selecting enzymes by name), users can get information for enzyme names while running the program. The cutting sites in Fig. 3 are listed as distances from the 5'-end of the sequence with the distances to the next cutting sites or to the 3'-end in the parentheses.

With option 9 users can look for promoter-like sequences. The search is performed purely by statistical scorings based on the promoter sequences compiled in ref. 7. Here, the frequency of occurrence of each one of the six bases at the "-35 region" and of those at the "-10 region" was calculated and weighted statistically. The distance between the two regions was also taken into consideration for calculation. When the calculated scores are higher than the scoring limit set by the user, then the likely promoter sequences are printed out together with the scores as shown in Fig. 4. I tested about twenty real sequences with this option and found that it always picked up the actual promoters mostly at the highest scores.

Program COPY and Access to the EMBO Nucleotide Data Library

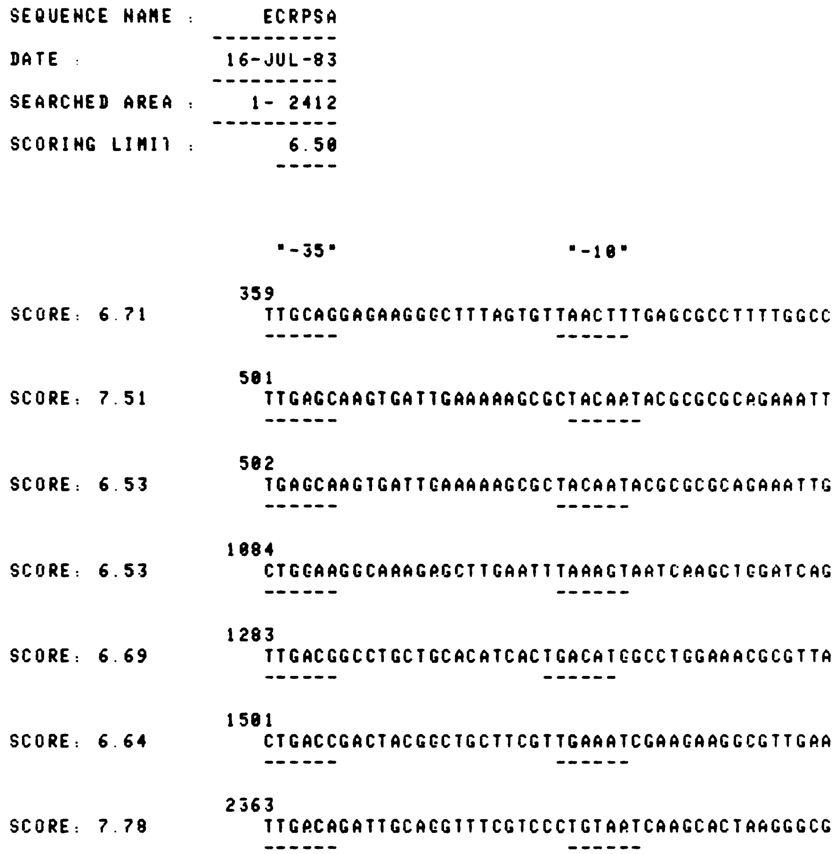
The option 10 of program NUCDAT (Fig. 2) exists for users



```
 [ RESTRICTION ENZYME SITES ]

                                                              DATE :    22-JUN-82

    NAME OF SEQUENCE :          ECRPSA
                              ----------
    SEARCHED AREA    :          1- 2412
                              ----------
    OPTION SELECTED  :         ALL ENZ.
                              ----------

-------------------------------------------------------------------------------
ENZYME  TOTAL  CUTTING SITES FOUND
-------------------------------------------------------------------------------

AccI       2     89(  963), 1051( 1362),

AflII      1    493( 1920),

AhaIII     1   1105( 1308),

AluI      16    171(  651),  821(   89),  909(   37),  945(   69), 1013(   85),
               1097(   22), 1118(   76), 1193(  232), 1424(   23), 1446(  247),
               1692(  255), 1946(  203), 2148(   67), 2214(   76), 2289(   13),
               2301(  112),

AsuI       2   1413(  162), 1674(  839),




Tth3II     2   2247(  147), 2393(   20),

Tth111II   2   2247(  147), 2393(   20),

XhoII      2   1434(  523), 1956(  457),

XmaI       1    702( 1711),

XmnI       1   1852(  561),
-------------------------------------------------------------------------------

NO SITE WAS FOUND FOR THE FOLLOWING ENZYMES:

     AcyI, ApaI, AsuII, AvaIII, AvaE, AvrII, BalI, BamHI, BclI, BsaI, BslII,
     ClaI, EcoB, EcoK, EcoRI, HsiAI, HsiJII, NaeI, NarI, NcoI, NdeI, NruI,
     SacI, SacII, SauI, SduI, SnaI, TteI, UbaI, XbaI, XhoI, XmaIII
```

Fig. 3: Search for restriction enzyme cutting sites. Only the beginning and the end are listed.

to get access to the EMBO Nucleotide Data Library (abbreviated as EMBOL). The latest release (i.e. Release 2.0, May 1983) of EMBOL contains 786 entries and 1,086,352 nucleotides. I obtained the entire data library through the courtesy of Dr. G. Cameron et al. of the EMBO Laboratory in separate files in the VAX format. I then grouped the sequence data into 10 sequencial files of approximately equal length so that program NUCDAT can read in the data faster than when all data are stored in separate files or when all data are stored in just one file. These files were named EMBNUC.01 through EMBNUC.10. The EMBNUC.01 file contains



```
SEQUENCE NAME :      ECRPSA
                   ----------
DATE :             16-JUL-83
                   ----------
SEARCHED AREA :     1- 2412
                   ----------
SCORING LIMIT :        6.50
                       -----


                      "-35"                   "-10"

                    359
SCORE: 6.71           TTGCAGGAGAAGGGCTTTAGTGTTAACTTTGAGCGCCTTTTGGCC
                      ------                 ------

                    501
SCORE: 7.51           TTGAGCAAGTGATTGAAAAAGCGCTACAATACGCGCGCAGAAATT
                      ------                  ------

                    502
SCORE: 6.53           TGAGCAAGTGATTGAAAAAGCGCTACAATACGCGCGCAGAAATTG
                      ------                 ------

                   1084
SCORE: 6.53           CTGGAAGGCAAAGAGCTTGAATTTAAAGTAATCAAGCTGGATCAG
                      ------                 ------

                   1283
SCORE: 6.69           TTGACGGCCTGCTGCACATCACTGACATGGCCTGGAAACGCGTTA
                      ------                ------

                   1501
SCORE: 6.64           CTGACCGACTACGGCTGCTTCGTTGAAATCGAAGAAGGCGTTGAA
                      ------                 ------

                   2363
SCORE: 7.78           TTGACAGATTGCACGTTTCGTCCCTGTAATCAAGCACTAAGGGCG
                      ------                  ------
```

Fig. 4: Search for promoter sequences by option of program NUCDAT. The two stretches of highest scores (starting residue #501 and #2267, respectively) are real promoters.

sequences whose names begin with A, the EMBNUC.02 file contains sequences whose names begin with B through D, and so on. (These data files are not included in this program package, however. Therefore, users have to prepare these files themselves).

The program package contains two files termed EMBNUC.FTR and EMBNUC.NAM which have been created by running program COPY and are based on the EMBOL data(which have been devided into ten groups as described above). Users can use these files directly if they have also the above mentioned EMBNUC.01 through EMBNUC.10 files prepared in exactly the same way. Otherwise, they can create these two files based on their own EMBOL data by running program COPY.

Upon execution program COPY reads individual sequence names recorded in the "ID lines" of EMBOL and transfer them into the EMBNUC.NAM file (a 'random access' file). At the same time it records the number (nn) of the data file (i.e. EMBNUC.nn; nn is from 1 to 10) and the line number of the "ID line" of each sequence counted from the beginning of that EMBNUC.nn file. The last information is used by program NUCDAT to reach a desired sequence quickly by the FORTRAN expression "SKIP RECORD n" (n is an integer).

Program COPY also transfers the brief feature descriptions written in the "DE lines" of EMBOL to create the EMBNUC.FTR file. This file is accessed also by program NUCDAT when users want to search for sequence names by some key-words. I found that the "key-words" provided in EMBOL are not so informative as compared to the feature descriptions in the "DE lines". Therefore, I decided to make the EMBNUC.FTR file by collecting the "DE line" information.

When the user wants to search for sequence names by the words appearing in the EMBNUC.FTR file, he can do so within program NUCDAT (option 10, Fig. 2). Here, he can type in any one or more words (either partial or full) he can think of (such as: coli, ribosom, prot), then the sequence names will be printed out that contain all of these words in their feature descriptions. This search is performed after converting all alphabetical characters into upper case, because the feature descriptions of the sequences in EMBOL which were entered earlier are written exclusively in upper case.

Program PROTEN

The program contains five options: sequence editing, translation from DAN/RNA sequences, sequence homology analysis, molecular weight and amino acid composition analysis, and print out of sequences in either one-letter or three-letter codes. These options have been adopted and modified from the corresponding ones in program NUCDAT.

Programs LITRAT and STRAIN

In addition to the programs described above, I have written a program termed LITRAT for storing scientific literature. It is not meant for general reference search, however. Its main purpose is to store the references which are very frequently quoted when users write scientific papers. The stored rerences can be printed out in the format of one of the 12 leading journals in biology-biochemistry fields or in the format of the users own specifications. In addition, search for stored references can be made by authors' name(s) or by any words (partial or full) in the reference titles. The total length of a reference is limited to 400 characters (including spaces). As in the case of program DATBAS, each reference is stored in a 'random access' file after being packed five-fold. A subroutine termed REFCOR allows users to correct any references after entry using six letter-oriented editing commands.

Program STRAIN (8) has also been extensively modified. One of the major modifications is the use of packed strings as in the case of programs DATBAS and LITRAT. As a result, the disk space needed for data storage can now be decreased some five-fold with a minimum loss of time in searching strains by markers.

Program distribution

The latest version of the program package described above will be distributed upon request. The package includes a detailed instruction of the programs. There will be a charge for the magnetic tape and postage. In addition users will be requested to agree to help me to further distribute the program package to other people in future.

*Present address: Department of Biology, Faculty of Science, Kobe University, Rokkodai, Kobe 657, Japan

REFERENCES

1 Isono, K. (1982) Nucleic Acids Res. 10, 85-89.
2 Staden, R. (1977) Nucleic Acids Res. 4, 4037-4051.
3 Staden, R. (1979) Nucleic Acids Res. 6, 2601-2610.
4 Staden, R. (1980) Nucleic Acids Res. 8, 3673-3694.
5 Queen, C.L., Korn, L.J. (1980) Methods in Enzymology, 65, 595-609.
6 Schnier, J., Isono, K. (1982) Nucleic Acids Res. 10, 1857-1865.
7 Siebenlist, U., Simpson, R.B., Gilbert, W. (1980) Cell 20, 269-281.
8 Isono, K. (1982) Molec. Gen. Genet. 186, 493-496.