Volume 14 Number 1 1986 Nucleic Acids Research

MBIS – an integrated system for the retrieval and analyses of sequence data from nucleic acids and proteins

C.A.Bucholtz and A.H.Reisner

CSIRO, Division of Molecular Biology, P.O. Box 184, North Ryde, NSW 2113 Australia

Received 2 July 1985

ABSTRACT

A computer-based system termed MBIS (the Molecular Biological Information Service), written in FORTRAN77 and Digital Command Language (DCL) and running on a Digital Equipment Corporation VAX computer under the VMS operating system (V4.1) is in use at the Division of Molecular Biology.

MBIS consists of three main sections: 1) The utility section, used by the system's manager to tailor the five commonly available databases so that they are useable by the applications programmes running on the system; 2) The retrieval section, used to find and extract specific sequences or bibliographic information, and 3) The analytical section, used to analyse and compare sequences either extracted from the databases or input by the user. The nucleotide databases maintained are GenBank, EMBL and PIR (Protein Identification Resource, National Biomedical Research Foundation) and the peptide databases are PIR and NEWAT. In addition, users can originate and maintain their own databases.

Those programmes which feature graphics output are compatible with most emulators of the Tektronix 4010 terminal.

INTRODUCTION

The retrieval and analysis of sequence data from nucleic acids and proteins are of major importance in molecular biology, and relatively easy to use integrated systems that assist research workers who are expert in molecular biology and biotechnology but not in the ways of computers have become necessities. The MBIS system (Molecular Biological Information Service) was placed on-line by the Division of Molecular Biology (DMB) in February 1985 [1] to meet the requirements of molecular biologists in

The source code and documentation for the MBIS system that have been developed at the DMB will be made available at approximately the cost of postage to non-commercial organizations who supply TWO 2400 ft 0.5 in. wide magnetic tapes; modified current versions of the GenBank, EMBL and NBRF databases - under the conditions stipulated by each purveyor - will be included. We can only supply tape written at a density of 1600 bpi. Unless requested otherwise the tape will be written as a VAX "backup" tape. If the complete system is requested, permission must first be obtained from third parties who have allowed us to incorporate their applications software into MBIS. Their names and addresses will be supplied on request.

Australia for information processing. It is accessible within Australia through CSIRO's CSIRONET network and can be accessed internationally through Tymnet. The system consists of just over one hundred programmes (about 100,000 lines of code), organized under a menu-driven tree structure which occupies, together with the databases, about 60 Mbytes. A running version (DCL programmes and executable binary images) together with the five databases occupies 25 Mbytes.

HARDWARE

MBIS runs on a Digital Equipment Corporation (DEC) VAX 11/750 Computer equipped with 3 Mbytes of memory, floating point accelerator, 456 Mbyte RA81 Winchester drive and TU80 industry standard tape drive (1600 bpi). Graphics are output to Tektronix 4010 or 4109 terminals and hardcopy is produced by a Tektronix 4695 ink jet printer.

SOFTWARE

Applications programmes developed at the DMB are written in either Digital Command Language (DCL), which runs under an interpreter or DEC's implementation of FORTRAN77; where appropriate, use is made of the runtime library. Currently, about half of the software running on the system has been developed at the DMB. The rest of the programmes has been contributed principally by Staden [2-8], and Kanehisa [9]. Additional programmes which are in use on the system have been donated by Novotny [10] and Novitski and Neri (Nagley, personal communication). In order to make the system appear unified to the user and to allow a common record structure for sequence data, virtually all of the programmes (with the exception of those of Staden which do not yield graphics output) required rewriting.

THE MOLECULAR BIOLOGICAL INFORMATION SERVICE (MBIS)

A user of MBIS, once he has had his password accepted, is confronted by the following output on his terminal. (Note: if there is ambiguity below, text commentary is enclosed in { }):

CSIRO, Division of Molecular Biology's
Molecular Biological Information Service - (MBIS)
(for NOTICES type NO)

{If the system's manager has made him a captive user with a special login file the following menu appears:}

```
Help level
full=1
prompt=2
none=3 or <CR>
```

{In what appears below user input is shown underlined:}

```
Enter level required: 1
```

{A novice user would be expected to type "1" and get the full menu shown below while one with more experience by typing "2" would see only a prompt line containing just the two letter commands.}

MENU

Command		Description
DATABASE	DA	Enter the database system
DEL	DE	Delete a file
DIR	DI	List your files
EDIT	ED	Create text files, using the VAX editor
HELP	HE	Help
LOG	LO	Log off
MAIL or	MA	Read mail or send mail, type HELP at MAIL> for help
NOTICES	NO	Read or place notices
PASSWORD	PA	Change your password
PRINT	PR	Print a file (system line printer at North Ryde)
PURGE	PU	Delete old versions of your files, keeping the latest
SET	ST	Set help level
TRANSFER	TR	Transfers text files to and from the VAX
TYPE	TY	Type a file (terminal)

<CTRL> Y will abort the current job

```
Command: database
```

{Typing "DA" or "database" invokes the Retrieval and Sequence Analysis System shown below. Typing help wherever it is an option invokes an appropriate explanation of the options available at the level the user finds himself. A sample of the system's use is given in the next five pages.}

DATA BASE INFORMATION RETRIEVAL

AND

SEQUENCE ANALYSES

{The Retrieval Option}

```
Retrieval/comparison or Analyses/manipulation (R,AN, or HELP): r
 Enter the data base of interest (G,D,E,P,N,O or HELP): help
```

```
GENBANK NUCLEIC ACIDS     G   GenBank, Bolt Beranek and Newman Inc.
Rel. 33.0, 3 June 1985        10 Moulton St.
                              Cambridge, MA 02238 U.S.A.

PIR NUCLEIC ACIDS         D   Protein Identification Resource,
Rel. 24, 8 April 1985         Georgetown University Medical Centre
                              3900 Reservoir Rd. N.W.
                              Washington, D.C. 20007 U.S.A.

EMBL NUCLEIC ACIDS        E   European Molecular Biology Laboratory,
Rel. 5, April 1985            Postfach 10 22 09, D-6900 Heidelberg, Germany.

PIR PROTEINS              P   Protein Identification Resource,
Rel. 5, 17 May 1985               as above

NEWAT PROTEINS            N   R.F.Doolittle, Dept. of Chemistry
April 1985                    University of California,
                              San Diego, La Jolla, CA 92093.

YOUR OWN DATABASE         O
```

{The prompt line is then redisplayed:}

Enter the data base of interest (G,D,E,P,N,O or HELP): N

List menu ? (Y/N): y

{Typing "n" would result in only the prompt line being output.}

COMMANDS

Code	Name	Description
AN	Analyse	Go to the Analyses Programmes menu
AU	Author	Search for all references by a given author.
C	Compare	Compare your sequence with the chosen database, based on correlation coefficients (Qr or r).
CA	CompAll	Compare your sequence with the three nucleic acid, or two protein databases using correl. coefficients (Qr or r).
CD	CreateDB	Create a specialized database containing your own sequences.
DB	DBase	Show and/or change the current database.
E	Extract	Extract the sequence from an entry and save it as a separate file in your area for later analysis.
M	Match	Compare all sequences in the database with a given sequence; those in it with the greatest similarity are noted (Wilber & Lipman algorithm).
S	Scan	Search the chosen database DIRECTORY for a specific string.
SA	ScanAll	Search all the database DIRECTORIES for a specific string.
T	Type	Type out a database entry on the terminal, or print out a database entry on the line printer, or file the entry in your directory.
DIR	Dir	Show a listing of the files currently in your area.
MENU	Menu	Retype the main menu.
HELP	Help	Obtain help.
X	Exit	Leave the DATABASE system return to COMMAND level.

(AN,AU,C,CA,CD,DB,E,M,S,SA,T,DIR,MENU,HELP,X) Enter code:

Most of the options listed above are self explanatory, however, a few comments may be warranted. The C or CA commands (compare or compAll) make use of the quasi correlation coefficient (Qr) [11,12] to allow high speed searching of the nucleotide databases or the correlation coefficient (r) [13] for searching the amino-acid-residue databases. These options make use of secondary databases which are created whenever a new primary one is received. The option M (match) on the other hand is an implementation of the algorithm developed by Wilber and Lipman [14] and the software used is based on FORTRAN code written by Kanehisa et al., [9].

A separate "directory", made up of several lines of descriptive information about each sequence entry, is created for every database and cross-indexed to them. The "S" and "SA" options are used to search these directories, and Boolean operators may be used. The utility is based on a programme supplied by the PIR. An example of its use for the NEWAT database is given below:

```
(AN,AU,C,CA,CD,DB,DIR,E,M,S,SA,T,MENU,HELP,X)  Enter code:s
     TYPE ©S TO HALT THE SCREEN, ©Q TO RESTART, AND ©Y TO ABORT
{© = Control key}
Enter string (U or L case) - sheep or " ovine"
You have typed - sheep or " ovine"
Is this correct (Y/N)? y
Do you want to print (P), view (V) or file (F) the searches (P/V/F)? v
                    Searching the NEWAT data base directory
              Title                                                    Start
CRFX      CORTICOTROPIN RELEASING FACTOR PRECURSOR, SHEEP              4633
 .
 .
 .
RHOV      RHODOPSIN, OVINE (C-TERM FRAG)                               6293
 .
 .
 .
(AN,AU,C,CA,CD,DB,DIR,E,M,S,SA,T,MENU,HELP,X) Enter code: t
```

The option T (type) allows the user to have information output to his terminal, to a file in his disc area or to the VAX's line printer. In addition, he may chose to have output just the bibliographic information available for a given entry or that information together with the sequence itself. The sequence, however, is not output in a computer readable form but rather in one easy to read by individuals. The response to having typed "t"

```
is shown below:
Do you want to Print (P), view (V) or file (F) the output (P/V/F)?: v
     TYPE ©S TO HALT THE SCREEN, ©Q TO RESTART, AND ©Y TO ABORT
   THIS PROGRAMME LISTS OUT SECTIONS OF THE NEWAT PROTEIN DATA BASE
Whole entry printed (information + sequence) (W)
or just the information (I): (W/I) w

Enter record start position (Enter -1 to finish) : 6293
Enter record start position (Enter -1 to finish) : -1

RHOV      RHODOPSIN, OVINE (C-TERM FRAG)
REF:      Findlay et al (1981) Nature, 293, 314.

            5        10        15        20        25        30
  1 S A T T Q K A E K E V T R M V I I M V I A F L I C W L P Y A
 31 G V A F Y I F T H Q G S D F G P I F M T I P A F F A K S S S
 61 V Y N P V I Y I M M N K Q F R N C M L T T L C C G K N P L G
 91 D.D E A S T T V S K T E T S Q V A P A

                          Composition of fragment

     10 Ala A      4 Gln Q      5 Leu L      8 Ser S
      2 Arg R      4 Glu E      6 Lys K     11 Thr T
      4 Asn N      5 Gly G      6 Met M      1 Trp W
      3 Asp D      1 His H      8 Phe F      4 Tyr Y
      4 Cys C      9 Ile I      6 Pro P      8 Val V

                                Number of residues = 109
```

The E (extract) option creates the 60 character records used by the analysis programmes. Only the sequence is transcribed into a file designated by the user.

The CD (CreateDB) option allows the user to create databases of his own in his disc area. They are written as EMBL formatted entries, are accessible only by the individual creating them, unless he specifically requests the system manager to make them public, and he may create as many different ones as he has space for.

The Analysis Option

If the AN (analysis) option is chosen when entering the DA area or from the menu above the following output to the terminal occurs:

```
Do you wish to see the menu (Y/N): y
                ANALYSIS SECTION PROGRAMMES AVAILABLE - SUMMARY

STADEN programmes :-

ANALYSEQ DIAGON SEQTREE SEQFIT MWCALC HYDROPLOT BACKTRAN
FILINS SEQLST CUTOUT GETFRQ
```

```
STADEN dbsystem programmes: - (for handling shotgun sequencing projects)

DBUTIL77 GELIN DBCOMP77 DBAUTO77 SCREENA SCREENB SCREENR77 SCREENV77
DBTOTAPE7 TAPETODB7 DBSTART77 HIGHLT GELSOUT77 ENDSOUT77

NIH Programmes: -  SEQA  SEQDP  CHOFAS  HPLOT  SEQP  SEQH  SEQHP

DMB programmes: -  SEQFIX  COUNT  MTX

Other programmes: - CHOU  CHOUDOT  ENRGFIT

Commands: - HELP  MENU  DIR  DOC  R = return to retrieval section  X = exit
Enter command (MENU,HELP,DIR,DOC,R,X or programme name): help
Type ©S TO HALT THE SCREEN, ©Q TO RESTART, AND ©Y TO ABORT
Information available:
ANALYSEQ    BACKTRAN    CHOFAS    CHOU      CHOUDOT   COUNT     CUTOUT
DBSYSTEM    DIAGON      DOC       ENRGFIT   FILINS    GETFRQ    HPLOT
HYDROPLOT   MENU        MTX       MWCALC    OVERVIEW  R         SEQA
SEQDP       SEQFIT      SEQFIX    SEQH      SEQHP     SEQLST    SEQP
SEQTREE     SIGNAL_SEARCH         X

Topic? mtx

MTX

     The MTX programme package, designed to analyze sequences of nucleotides
and amino acid residues, was written at CSIRO's Div. of Mol. Biol.
  .
  .
  .
Press RETURN to continue - Press ? to select topics ... ?

  Additional information available:

  MTXDOT    MTXLIN    MTXPLOT    MTXANL    MTXRAN    MTXAACOR    MTXNUCCOR

MTX Subtopic? ©Z    {©Z = exit from the HELP utility}

Enter command (MENU,HELP,DIR,DOC,R,X OR PROGRAMME NAME): doc

Documents available: -

1.   Introduction to MBIS                              (file name = MBIS.DOC)
2.   Staden programmes and the DBSYSTEM programmes     (file name = STADEN.DOC)
3.   ANALYSEQ                                          (file name = ANALYSEQ.DOC)
4.   NIH programmes                                    (file name = IDEAS.DOC)
5.   Div. Molecular Biology MTX programmes             (file name = MTX.DOC)
6.   Quit - no document required
Enter document number required: 1
Do you want the document printed now or filed in your directory (P/F)?: f
Enter command (MENU,HELP,DIR,DOC,R,X or programme name): x

                          LEAVING DATA BASE SYSTEM

                          TO RE-ENTER TYPE DATABASE
```

Command:

At this point the command level menu will reappear if the help level was set to "1".

DISCUSSION

In designing MBIS we have not tried to reconfigure the five databases to a common single format which would tend to restrict the use of some of the features present in one or another of them. In addition when changes in format for a given database occur, it is relatively simple to alter those parts of the utility section which have to be modified. GenBank is a special case in that between full quarterly issues of the database, monthly updates are supplied which are used to replace entries which have been found to be in error as well as adding new entries. The MBIS software contains the utilities to perform the "interleaving".

To the user the system appears unified, i.e. he is shielded from the changes required to deal with the different databases, and this reduces the number of commands that are needed to drive the system. Finally, the use of the DCL shell allows the amalgamation and ready addition of software without altering the apparent fabric of the system.

ACKNOWLEDGEMENTS

We thank those individuals and groups who have sent copies of source code to us; in particular R. Staden, M. Kanehisa, R.F. Doolittle, J.Novotny, the Protein Identification Resource and P. Nagley.

REFERENCES

1. Reisner, A.H. and Bucholtz, C.A. (1985) Nature 314, 310.
2. Staden, R. (1984) Nucleic Acids Res. 12, 499-503.
3. Staden, R. (1980) Nucleic Acids Res. 8, 817-825.
4. Staden, R. (1982) Nucleic Acids Res. 10, 2951-2961.
5. Staden, R. (1984) Nucleic Acids Res. 12, 551-567.
6. Staden, R. (1984) Nucleic Acids Res. 12, 505-519.
7. Staden, R. (1984) Nucleic Acids Res. 12, 521-538.
8. Staden, R and McLachlan, A.D. (1982) Nucleic Acids Res. 10, 141-156.
9. Kanehisa, M., Klein, P., Greif, P. and DeLisi, C. (1984) Nucleic Acids Res. 12. 417-428.
10. Novotny, J. and Auffray, C. (1984) Nucleic Acids Res. 12, 243-255.
11. Reisner, A.H. and Bucholtz, C.A. (1983) EMBO J. 2, 1145-1149.
12. Reisner, A.H. and Bucholtz, C.A. (1984) Nucleic Acids Res. 12, 409-416.
13. Reisner, A.H. and Westwood, N.H. (1982) J. Mol. Evol. 18, 240-250.
14. Wilber, W.J. and Lipman, D.J. (1983) Proc. Natl. Acad. Sci. USA 80, 726-730.