nogalamycin and the duplex sequence dTpA (Brown *et al.*, 1984), where it has been suggested that the sugar group at each end of the drug molecule (Fig. 2) sterically blocks the free movement of the chromophore in and out of the intercalation site.

The intercalation process in solution has been extensively studied most often with reference to random-sequence DNA, and rather more rarely with respect to defined-sequence polynucleotides. For DNA itself, there are ten distinct dinucleoside intercalation sites, and two for an alternating polynucleotide such as poly(dC-dG). Most solution methods for studying drug intercalation (for example, by measurement of affinity constants), tend to average out differences in binding properties at these different sites. The molecular graphics approaches outlined above, do on the other hand, focus attention on just one site, such as the -CpG- sequence. This is likely to be a high-affinity site, especially at the low drug levels relevant to physiological conditions (reviewed in Neidle & Abraham, 1984). It will also be an important goal for the future effort in this field to relate the simulation data at these defined sites to intercalative drug behaviour in solution and in the cell as measured by fine-details methods such as stopped-flow kinetics and DNA footprinting.

I am most grateful to my colleagues and collaborators in these studies, and to the Cancer Research Campaign for continuing support.

Brown, J. R., Collier, D. A. & Neidle, S. (1984) *Biochem. Pharmacol.* in the press
Fletcher, R. & Reeves, C. M. (1969) *Computer J.* 7, 149–154
Gund, P., Andose, J. D., Rhodes, J. B. & Smith, G. M. (1980) *Science* **208**, 1425–1431
Hopfinger, A. J. (1973) *Conformational Properties of Macromolecules*, Academic Press, New York
Islam, S. A. & Neidle, S. (1984) *Acta Crystallogr.* in the press
Islam, S. A., Niedle, S., Gandecha, B. M. & Brown, J. R. (1983) *Biochem. Pharmacol.* **32**, 2801–2808
Islam, S. A., Neidle, S., Gandecha, B. M., Partridge, M., Patterson, L. H. & Brown, J. R. (1984) *J. Med. Chem.* in the press
Neidle, S. & Abraham, Z. (1984) *Crit. Rev. Biochem.* in the press
Neidle, S., Achari, A., Taylor, G. L., Berman, H. M., Carrell, H. L., Glusker, J. P. & Stallings, W. C. (1977) *Nature (London)* **269**, 304–307
Neidle, S., Berman, H. M. & Sheh, H.-S. (1980) *Nature (London)* **288**, 129–133
Quigley, G. J., Wang, A. H.-J., Ughetto, G., van der Marel, G., van Boom, J. H. & Rich, A. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7204–7208
Shieh, H.-S., Berman, H. M., Dabrow, M. & Neidle, S. (1980) *Nucleic Acids Res.* **8**, 85–97
Sobell, H. M., Tsai, C.-C., Jain, S. C. & Gilbert, S. G. (1977) *J. Mol. Biol.* **114**, 333–365
Weiner, P. K., Langridge, R., Blaney, J. M., Schaefer & Kollman, P. A. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3754–3758
White, D. N. J. (1978) in *Molecular Structure by Diffraction Methods*, vol. 6, The Chemical Society, London

The EMBL nucleotide sequence data library

G. G. KNEALE and OLGA KENNARD
Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K.

With the enormous increase in the rate of sequencing DNA fragments, it was clear some years ago that a large computerized database of sequences would be essential for research in Molecular Biology, and that such an operation should be done, at least in Europe, on an international basis. Consequently, in 1982, a nucleic acid sequence data library was set up by the European Molecular Biology Laboratory (EMBL) at Heidelberg. The first release contained 568 entries comprising over 500000 bases. The latest version (release 3.0) has 1481 entries totalling over 1.6 million bases (Cameron *et al.*, 1983).

The layout of the database is designed for both human and computer readability. Each entry in the EMBL library is labelled by a short (6–8 character) sequence identifier, providing a unique reference for each entry which is used in the indexes. The structure of a typical entry is shown in Fig. 1. The nature of the information on each line is specified by a two-letter line code at the front of each line and includes a description of the sequence, the source organism, keywords, references, a table of any interesting features and the sequence itself. The feature table is especially useful as it defines regions of biological significance, as well as sequence ambiguities. It lists regions of transcriptional significance (e.g. the extent of primary transcripts, introns, exons, tRNA genes); control regions (e.g. promotors, ribosome-binding sites and origins of replication); transposons and inverted repeats; and mutations, conflicts and uncertain assignments. The feature table is easily read by computer for automatic processing of sequences, e.g. 'splicing' of mRNA and 'translation' to protein sequences (Bishop, 1984).

The EMBL library is being maintained on the IBM 3081 at Cambridge, along with the manual and seven indexes. These, together with a wide variety of programs for sequence analysis and search and retrieval of sequences from the database, are available for interactive use. The system is an integrated one in which users can key in one line to receive information on the facilities available and access documentation for the program. The programs may be run on-line simply by typing in the program name. Examples of the facilities available are shown in Figs. 2 to 5. Fig. 2 shows the 'primary menu' that is used to select, for instance, a description of the files (Fig. 3) and facilities (Fig. 4) of the system. Alternatively a menu of programs for DNA sequence analysis can be selected (Fig. 5).

The Los Alamos library, GenBank, is also available to users of the system as it contains, at present, many sequences not in the EMBL library. The GenBank sequences can be used for input to the standard analysis programs. A very comprehensive database is thus available to users. In addition there are programs for searching protein sequence databases, and for analysis of protein sequences. The system is presently being used by over 30 researchers in eight departments in the University and in local research institutes. These users can keep in touch with each other via the MAIL command. The M.R.C. is supporting the extension of these facilities to researchers at Universities and institutes throughout the U.K. on the Joint Academic network (JANET) and those interested should contact one of the authors for further details of the facilities.

A further development of the system will be the centralized collection of nucleic acid sequences from research groups in the U.K. The data will be transferred to EMBL at appropriate intervals. Direct input of sequences in machine readable form will be of considerable benefit, since it will minimize transcription errors and will reduce

```
print embl.dna:hsagl1.
ID   HSAGL1     HOMO.SAP.GERM.GLOBIN.ALPHA;  DNA;  1138 BP.
XX
AC   V00488;
XX
XX
DT   14-SEP-1981  (first entry)
DT   08-APR-1983  (feature table corrected)
XX
DE   Human alpha-globin germ line gene.
XX
KW   germ line; globin; alpha-globin.
XX
OS   Homo sapiens (man, homme, Mensch)
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda; Mammalia;
OC   Eutheria; Primates.
XX
RN   [1]   (bases 1-1138)
RA   Liebhaber S.A., Goossens M.J., Wai Kan Y.;
RT   "Cloning and complete nucleotide sequence of human 5'-alpha-
RT   globin gene";
RL   Proc. Natl. Acad. Sci. USA 77:7054-7058(1980).
XX
CC   KST HSA.ALPGLOBIN.GL [1138]
XX
FH   Key         From      To        Description
FH
FT   TRANSCR       98     929        primary transcript
FT   MSG           98     230        exon 1
FT   CDS          135     230        reading frame (part 1)
FT   MSG          348     551        exon 2
FT   CDS          348     551        reading frame (part 2)
FT   MSG          692     929        exon3
FT   CDS          692     817        reading frame (part 3)
XX
SQ   Sequence   1138 BP;  183 A;  412 C;  193 T;  350 G.
     AGGCCGCGCC CCGGGCTCCG CGCCAGCCAA TGAGCGCCGC CCGGCCGGGC GTGCCCCCGC
     GCCCCAAGCA TAAACCCTGG CGCGCTCGCG GCCCGGCACT CTTCTGGTCC CCACAGACTC
     AGAGAGAACC CACCATGGTG CTGTCTCCTG CCGACAAGAC CAACGTCAAG GCCGCCTGGG
     GTAAGGTCGG CGCGCACGCT GGCGAGTATG GTGCGGAGGC CCTGGAGAGG TGAGGCTCCC
     TCCCCTGCTC CGACCCGGGC TCCTCGCCCG CCCGGACCCA CAGGCCACCC TCAACCGTCC
     TGGCCCCGGA CCCAAACCCC ACCCCTCACT CTGCTTCTCC CCGCAGGATG TTCCTGTCCT
     TCCCCACCAC CAAGACCTAC TTCCCGCACT TCGACCTGAG CCACGGCTCT GCCCAAGTTA
     AGGGCCACGG CAAGAAGGTG GCCGACGCGC TGACCAACGC CGTGGCGCAC GTGGACGACA
     TGCCCAACGC GCTGTCCGCC CTGAGCGACC TGCACGCGCA CAAGCTTCGG GTGGACCCGG
     TCAACTTCAA GGTGAGCGGC GGGCCGGGAG CGATCTGGGT CGAGGGGCGA GATGGCGCCT
     TCCTCTCAGG GCAGAGGATC ACGCGGGTTG CGGGAGGTGT AGCGCAGGCG GCGGCGCGGC
     TTGGGCCGCA CTGACCCTCT TCTCTGCACA GCTCCTAAGC CACTGCCTGC TGGTGACCCT
     GGCCGCCCAC CTCCCCGCCG AGTTCACCCC TGCGGTGCAC GCTTCCCTGG ACAAGTTCCT
     GGCTTCT
```

Fig. 1. *The structure of a typical entry in the EMBL library (the human α-globin gene)*

```
c embl.help:start

The following files contain information to help you use the EMBL library
and associated programs.There are 6 such files:
0) STOP -exit to command level
1) PHX  -for elementary use of the IBM 3081
2) GEN  -for general EMBL facilities
3) FILES  -for files belonging to fileowner EMBL
4) DNAMENU  -for documentation on DNA sequence analysis programs
5) GELMENU  -for DNA sequencing projects
6) PEPMENU  -for peptide analysis programs
Press space bar to halt scrolling; CTRL-X to recommence scrolling.
In response to 'page waiting' press RETURN.
Type the number you require (0-6)  2
```

Fig. 2. *Primary 'menu' of options for the EMBL library and associated programs*

```
The following files belong to fileowner EMBL (whose manager is MJB1):

EMBL.DNA        The EMBL nucleotide sequence data library.This file is a PDS file
                the members of which are individual sequences and can be accessed
                directly by their sequence identifiers (eg. embl.dna:mmig10 )
                The present version is release 3.0 (december,1983)
EMBL.TXT        A pds file containing the manual and release notes for EMBL.DNA
EMBL.INDEX      This contains a variety of indexes to the contents of EMBL.DNA
EMBL.HELP       Contains useful information about EMBL facilities
EMBL.CL         Command sequences to execute programs.
EMBL.NEWS       News of additions or changes to the EMBL files
EMBL.LIB        The library of programs whose member DNA should be loaded by
                EMBL users by entering "LIBRARY EMBL.LIB:DNA". This will be
                done automatically by some of the utility programs.

Also available:
The Dec 1983 release of GENBANK from Los Alamos which is split into files:
EMBL.GENBANK.BACTERIA
EMBL.GENBANK.PHAGE
EMBL.GENBANK.INVERT
EMBL.GENBANK.VIRUS
EMBL.GENBANK.MAMMAL    (corrupt on tape sent to us)
EMBL.GENBANK.ORGANELL
plus the following indexes and manual:
EMBL.GENBANK:LDIRECT     "long directory" i.e. detailed index
EMBL.GENBANK:SDIRECT     "short directory" i.e concise index
EMBL.GENBANK:MANUAL

Conversion of the GENBANK sequences to EMBL format is in progress.
EMBL.DNAG       contains the DNA sequences
EMBL.GENINDEX   contains indexes to the contents of EMBL.DNAG
                LOWER VERTEBRATE,
                PLANT,STRUCTURAL RNA,SYNTHETIC SEQUENCES are included
                (note that some identifiers have been shortened
                 as recorded in :CHANGES)


The Doolittle peptide database (NEWAT) in EMBL format is in the file
EMBL.PEP  (as a PDS)
the index for these sequences being in the file
EMBL.PEPINDEX:ID
A peptide dictionary (accessed by command ZDICTPEP) is in the file
EMBL.DOOLDICT
```

Fig. 3. *Option 3: a listing of files associated with the libraries*

```
This file lists the general purpose utility programs and files
MAIL      Enables us to send messages to other EMBL users.The messages are
          filed in the file EMBL.NEWS
NEWS      Messages from the mail command ,identified by date.
PACKAGE   An interactive program for newcomers to the system which allows easy
          access to the EMBL library and programs.
START     An introduction to the facilities of the EMBL files.
RETRIEVE  A program for retrieving sequences from the EMBL library.The sequences
          are extracted on the basis of keywords etc. and can be put on the
          users own filespace for further analysis.
          Seven indexes to the EMBL library are also kept:
          EMBL.INDEX:ID          an alphabatic list of sequence identifiers
          EMBL.INDEX:SPECIES     listed according to species
          EMBL.INDEX:AUTHOR      listed according to author of publication
          EMBL.INDEX:KEYWORD     listed according to keywords on the KW line
          EMBL.INDEX:JOURNAL     an index of the literature cited in EMBL.DNA
          EMBL.INDEX:SHORT       a one line description of each entry
          EMBL.INDEX:ACCESSNO    an index of accession numbers
```

Fig. 4. *Option 2: a listing of general facilities available*

```
                              DNAMENU

The following programs will be available for DNA sequence analysis
on the IBM 3081 under MVS (! indicates not yet implemented):

OPERATIONS ON A SINGLE SEQUENCE
 1.Sequence manipulation (!ANALYSEQ,ZSEQ)
   a) Verification and conversion (including genbank to embl conversion etc.)
   b) Gene extraction/splicing (with optional translation,complementation etc.)
   c) Listing of sequences (with optional translation,complementation etc.)
   d) Counting (base composition,di- or trinucleotide frquencies)
   e) Searching (probes, restriction sites etc.)
 2.Repeats and inverted repeats (!SEQTREE,ZDICT,XDRLEN)
 3.Secondary structure and hairpins (!ANALYSEQ)
 4.Gene identification (!ANALYSEQ)
     a) t-rna genes
     b) protein genes

OPERATIONS ON MORE THAN ONE SEQUENCE
 1.Sequence alignment (!RBSA)
 2.Database searching for homology (EMBLSCAN,!RBSA)
 3.Dot matrix plot (DIAGON)

MISCELLANEOUS PROGRAMS
 1. Formatting and file conversion (MAKEPDS,MAKEPS,MKSEQ,MKSTADEN)
 2. Randomisation of sequences (SHUFFLE)
 3. Statistics (DOUBLETS)

Do you want to read the documentation for one of these? n
```

Fig. 5. *Option 4: a 'menu' of programs for analysing DNA sequences*

the workload at EMBL. This workload is likely to increase greatly with the increase in the number of sequence determinations. To aid this collection, a program is being developed which prompts the user for text and outputs the sequence in EMBL format.

We are grateful to Dr. M. J. Bishop for writing and implementing many of the programs, to the EMBL database staff for their co-operation and to the M.R.C. for financial support.

Bishop, M. J. (1983) *Biochem. Soc. Trans.* **12**, 1015–1017

Cameron, G., Hamm, G., Nial, J., Rudloff, A., Stoesser, G. & Stueber, K. (1983) *EMBL Nucleic Acid Sequence Data Library User Manual*, European Molecular Biology Laboratory, Heidelberg