Volume 12 Number 1 1984 Nucleic Acids Research

Computer analysis and structure prediction of nucleic acids and proteins

Minoru Kanehisa*, Petr Klein, Peter Greif and Charles DeLisi

Laboratory of Mathematical Biology, National Cancer Institute, National Institutes of Health,
Bethesda, MD 20205, USA

Received 12 August 1983

ABSTRACT

We have developed an integrated computer system for analysis of nucleic acid and protein sequences, which consists of sequence and structure databases, a relational database, and software for structural analysis. The system is potentially applicable to a number of problems in structural biology including predictive classification of the function and location of oncogene products.

INTRODUCTION

Information on DNA sequences and the protein products for which they code is being published at an explosively rapid rate. According to figures compiled at the Los Alamos nucleic acid sequence database, the number of bases determined has doubled every year since 1978 and currently totals about two million.[1] A byproduct of this information explosion is a rapidly growing number of protein sequences which have not been identified either chemically or functionally. For example, the products of a sizable percentage of recently sequenced oncogenes fall into this category. In this communication we describe a facility which combines, in a relational database, all current sequence and structural information, with software for the analysis and prediction of molecular structure. The entire system operates on the Laboratory of Mathematical Biology's VAX 11/780. Among the many potential applications of this computer system is the ability to provide clues to the structure, location, and function of DNA products, thereby guiding and speeding experimental analysis.

The computational methods fall into three categories: (i) statistical analysis, (ii) sequence analysis, and (iii) structure predictions. Among them cluster analysis and discriminant analysis based on database statistics, coupled with primary and secondary structure homology searches against the databases, are proving particularly useful for characterizing sequences of unknown function. The detailed biological applications of this system will be

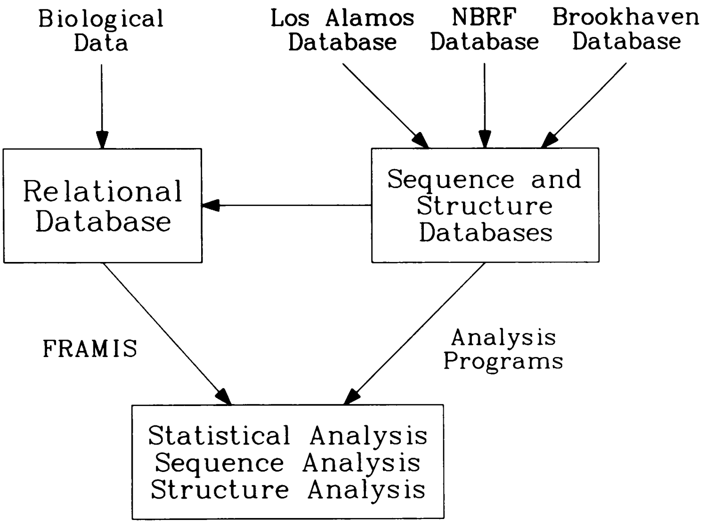


Figure 1. The overall organization of our database and analysis system.

reported elsewhere. In this communication we briefly present the logical structure of the system, its current capabilities and limitations, and some examples of how it works.

DATABASES

Primary Databases

The computer system is organized around the most recent versions of the three major databases (Fig. 1): the nucleic acid sequence database (GenBank) from the Los Alamos National Laboratory,[2] the protein sequence database from the National Biomedical Research Foundation (NBRF) at Georgetown University,[3] and the Protein Data Bank for both protein and nucleic acid structures from Brookhaven National Laboratory.[4] The contents of these databases were reorganized and integrated into a single relational database, which also contains additional information, such as cross-references of the original databases, sequence homologies, and other relevant biological data. For reasons that will become apparent shortly; structural information (sequence, bond lengths and angles, and so forth) and analytical software are stored outside the relational database under ordinary VAX files.

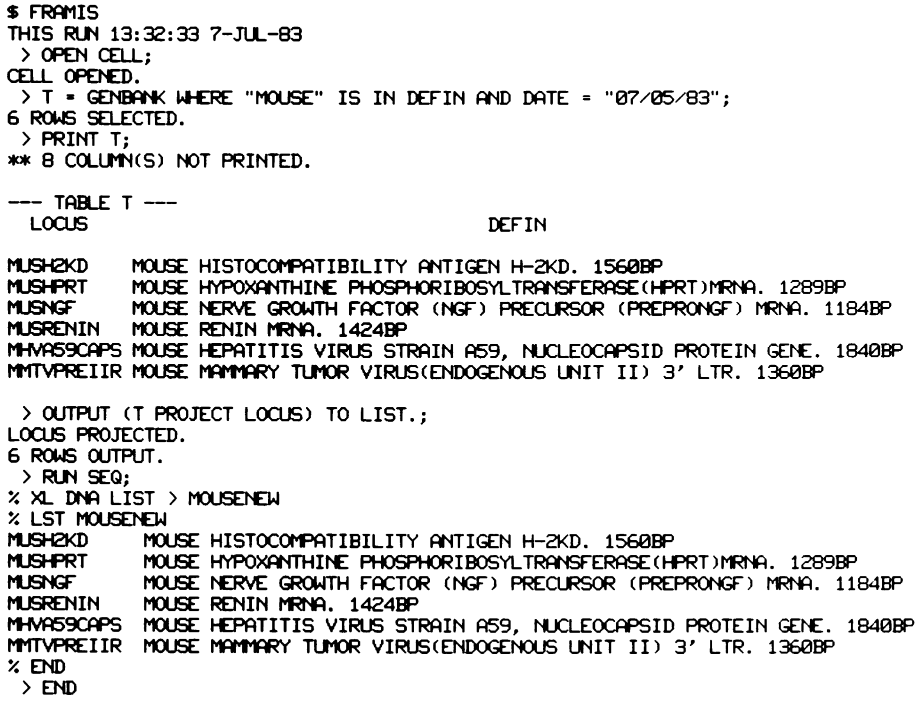
Relational Database

The relational database is a natural way of organizing various types of data so that they can be treated by a well-defined set of operations, viz, relational algebra, to perform many tasks ranging from simple retrieval to complex analysis. We have implemented our relational database, named CELL, using the FRAMIS[5] database management system. As of July 1983 CELL contained the following tables:

GENBANK	Contents of the current nucleic acid sequence database
CODONFREQ	Codon usage frequency in the nucleic acid sequence database
DAYHOFF	Contents of the current protein sequence database
AAFREQ	Amino acid frequency in the protein sequence database
PROTEINS	Classification of proteins
ENZYMES	Enzyme classification numbers
MEMBRANE	List of proteins interacting with membranes
SIGNAL	List of proteins containing signal peptides
BROOKHAVEN	Contents of the current Brookhaven Protein Data Bank
BNLDIR	Contents of our current protein structure database
SECSTR	Protein secondary structures (helix, sheet, turn, and SS-bond)

One of the main uses of the FRAMIS system is the creation and manipulation of subsets of the original databases. As an example of a simple manipulation, suppose one wished to do a homology search of all mouse nucleic acid sequences entered into the database on a particular date, against the entire DNA data bank. The first step would be to retreive the mouse sequences as shown in Fig. 2. Each row of table GENBANK in the FRAMIS database contains information on a particular sequence (unique sequence identifier, description, date of entry or update, and so on) in the original GenBank database. The first query extracts specific rows of table GENBANK, namely, new or modified entries of mouse sequences, by searching columns DEFIN and DATE, and the result is written in a temporary table named T. Then, column LOCUS of table T, which contains sequence identifiers, is projected (extracted) and the resulting table is moved to an external file named LIST. LIST is used to direct the sequence manipulation program SEQ to create a subset of sequence data, stored in file MOUSENEW, from the original sequence library DNA (sequences not shown). Such sequences can then be analyzed in any number of ways by using software that works on ordinary files--for example in homology searches against other sets of molecules. Thus, FRAMIS is useful in identifying and grouping specific sequences, or sets of sequences, for further analysis.

Although this example shows the search within one table, viz, GENBANK, the flexibility of the relational system resides in its ability to make operations on multiple tables.[1] If, for example, we enter relevant biological data in a FRAMIS table, it can be logically combined with another table in the relational system. In fact, the purpose of our relational database is to make



```
$ FRAMIS
THIS RUN 13:32:33 7-JUL-83
 > OPEN CELL;
CELL OPENED.
 > T = GENBANK WHERE "MOUSE" IS IN DEFIN AND DATE = "07/05/83";
6 ROWS SELECTED.
 > PRINT T;
** 8 COLUMN(S) NOT PRINTED.

--- TABLE T ---
  LOCUS                                DEFIN

MUSH2KD     MOUSE HISTOCOMPATIBILITY ANTIGEN H-2KD. 1560BP
MUSHPRT     MOUSE HYPOXANTHINE PHOSPHORIBOSYLTRANSFERASE(HPRT)MRNA. 1289BP
MUSNGF      MOUSE NERVE GROWTH FACTOR (NGF) PRECURSOR (PREPRONGF) MRNA. 1184BP
MUSRENIN    MOUSE RENIN MRNA. 1424BP
MHVA59CAPS MOUSE HEPATITIS VIRUS STRAIN A59, NUCLEOCAPSID PROTEIN GENE. 1840BP
MMTVPREIIR MOUSE MAMMARY TUMOR VIRUS(ENDOGENOUS UNIT II) 3' LTR. 1360BP

 > OUTPUT (T PROJECT LOCUS) TO LIST.;
LOCUS PROJECTED.
6 ROWS OUTPUT.
 > RUN SEQ;
% XL DNA LIST > MOUSENEW
% LST MOUSENEW
MUSH2KD      MOUSE HISTOCOMPATIBILITY ANTIGEN H-2KD. 1560BP
MUSHPRT      MOUSE HYPOXANTHINE PHOSPHORIBOSYLTRANSFERASE(HPRT)MRNA. 1289BP
MUSNGF       MOUSE NERVE GROWTH FACTOR (NGF) PRECURSOR (PREPRONGF) MRNA. 1184BP
MUSRENIN     MOUSE RENIN MRNA. 1424BP
MHVA59CAPS   MOUSE HEPATITIS VIRUS STRAIN A59, NUCLEOCAPSID PROTEIN GENE. 1840BP
MMTVPREIIR   MOUSE MAMMARY TUMOR VIRUS(ENDOGENOUS UNIT II) 3' LTR. 1360BP
% END
 > END
```

Figure 2. A sample run of FRAMIS to create a file containing selected sequences. The prompt character "$" is from the VAX/VMS operating system, ">" from FRAMIS, and "%" from the SEQ program to manipulate sequences. Each FRAMIS command is terminated by a semi-colon.

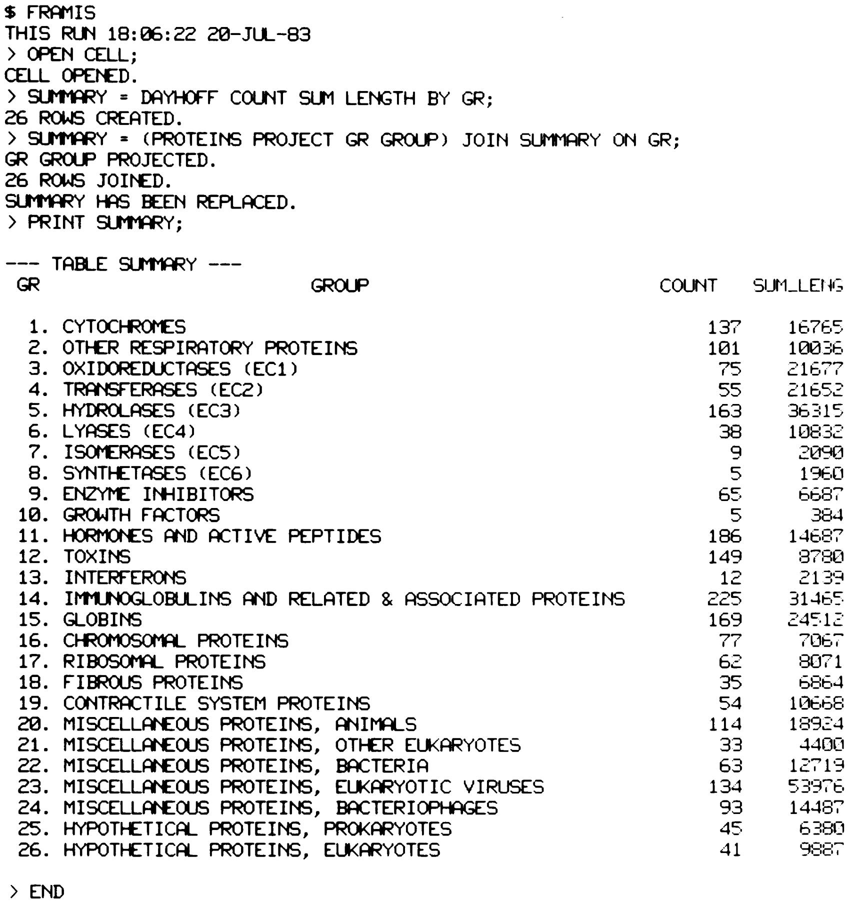
logical connections among the three primary databases and other biological information.

SOFTWARE

Discriminant Analysis

Statistical methods can be used to characterize a sequence of unknown function and/or unknown structure. The general procedure involves (i) identification of known sequences that cluster into groups according to biological, chemical, or physical parameters; (ii) finding combinations of those parameters that allow best discrimination among the clusters; and (iii) applying discriminant analysis to specific sequences in order to predict the cluster to which they are most likely to belong.

Fig. 3 shows an example of functional clustering of protein sequences into groups of superfamilies. Table DAYHOFF contains information on the

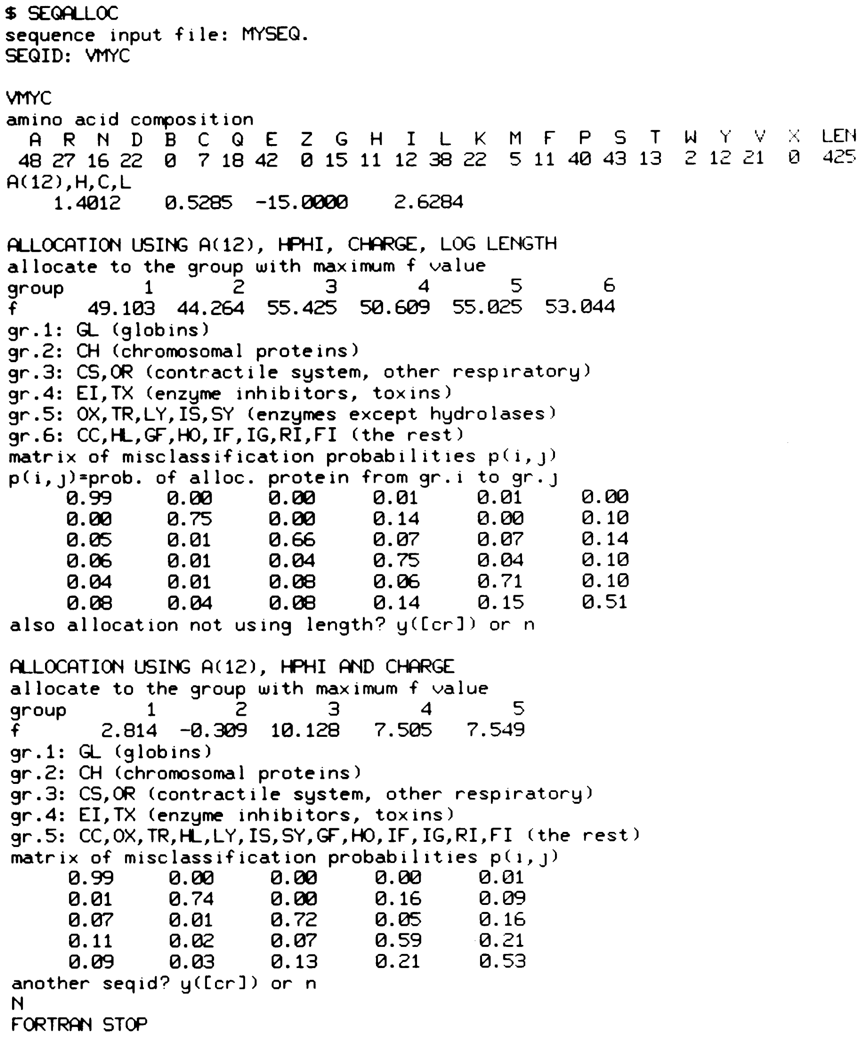


```
$ FRAMIS
THIS RUN 18:06:22 20-JUL-83
> OPEN CELL;
CELL OPENED.
> SUMMARY = DAYHOFF COUNT SUM LENGTH BY GR;
26 ROWS CREATED.
> SUMMARY = (PROTEINS PROJECT GR GROUP) JOIN SUMMARY ON GR;
GR GROUP PROJECTED.
26 ROWS JOINED.
SUMMARY HAS BEEN REPLACED.
> PRINT SUMMARY;

--- TABLE SUMMARY ---
 GR                         GROUP                              COUNT   SUM_LENG

  1. CYTOCHROMES                                                 137      16765
  2. OTHER RESPIRATORY PROTEINS                                  101      10036
  3. OXIDOREDUCTASES (EC1)                                        75      21677
  4. TRANSFERASES (EC2)                                           55      21652
  5. HYDROLASES (EC3)                                            163      36315
  6. LYASES (EC4)                                                 38      10832
  7. ISOMERASES (EC5)                                              9       2090
  8. SYNTHETASES (EC6)                                             5       1960
  9. ENZYME INHIBITORS                                            65       6687
 10. GROWTH FACTORS                                                5        384
 11. HORMONES AND ACTIVE PEPTIDES                                186      14687
 12. TOXINS                                                      149       8780
 13. INTERFERONS                                                  12       2139
 14. IMMUNOGLOBULINS AND RELATED & ASSOCIATED PROTEINS           225      31465
 15. GLOBINS                                                     169      24512
 16. CHROMOSOMAL PROTEINS                                         77       7067
 17. RIBOSOMAL PROTEINS                                           62       8071
 18. FIBROUS PROTEINS                                             35       6864
 19. CONTRACTILE SYSTEM PROTEINS                                  54      10668
 20. MISCELLANEOUS PROTEINS, ANIMALS                             114      18924
 21. MISCELLANEOUS PROTEINS, OTHER EUKARYOTES                     33       4400
 22. MISCELLANEOUS PROTEINS, BACTERIA                             63      12719
 23. MISCELLANEOUS PROTEINS, EUKARYOTIC VIRUSES                  134      53976
 24. MISCELLANEOUS PROTEINS, BACTERIOPHAGES                       93      14487
 25. HYPOTHETICAL PROTEINS, PROKARYOTES                           45       6380
 26. HYPOTHETICAL PROTEINS, EUKARYOTES                            41       9887

> END
```

Figure 3. Functional groups of protein sequences in the NBRF database based on the superfamily classification.

superfamily and group of superfamilies to which each sequence belongs, according to the compilation by Dayhoff and coworkers[3] which is largely based on sequence homologies. First, FRAMIS is used to count the number of rows (sequence entries) in each group and to sum the lengths of sequences in each group, the group identifier being in column GR. As a result of these aggregation operations, table SUMMARY containing three columns GR, COUNT, and SUM_LENGTH is created. Table SUMMARY is then combined with table PROTEINS, which contains the description of each group, by the JOIN operation. Thus, a new column GROUP is effectively added to table SUMMARY.

As a simple example of higher order clustering, the first 19 groups of



```
$ SEQALLOC
sequence input file: MYSEQ.
SEQID: VMYC

VMYC
amino acid composition
  A  R  N  D  B  C  Q  E  Z  G  H  I  L  K  M  F  P  S  T  W  Y  V  X  LEN
 48 27 16 22  0  7 18 42  0 15 11 12 38 22  5 11 40 43 13  2 12 21  0  425
A(12),H,C,L
    1.4012    0.5285  -15.0000    2.6284

ALLOCATION USING A(12), HPHI, CHARGE, LOG LENGTH
allocate to the group with maximum f value
group        1       2       3       4       5       6
f      49.103  44.264  55.425  50.609  55.025  53.044
gr.1: GL (globins)
gr.2: CH (chromosomal proteins)
gr.3: CS,OR (contractile system, other respiratory)
gr.4: EI,TX (enzyme inhibitors, toxins)
gr.5: OX,TR,LY,IS,SY (enzymes except hydrolases)
gr.6: CC,HL,GF,HO,IF,IG,RI,FI (the rest)
matrix of misclassification probabilities p(i,j)
p(i,j)=prob. of alloc. protein from gr.i to gr.j
      0.99      0.00      0.00      0.01      0.01      0.00
      0.00      0.75      0.00      0.14      0.00      0.10
      0.05      0.01      0.66      0.07      0.07      0.14
      0.06      0.01      0.04      0.75      0.04      0.10
      0.04      0.01      0.08      0.06      0.71      0.10
      0.08      0.04      0.08      0.14      0.15      0.51
also allocation not using length? y([cr]) or n

ALLOCATION USING A(12), HPHI AND CHARGE
allocate to the group with maximum f value
group         1       2       3       4       5
f         2.814  -0.309  10.128   7.505   7.549
gr.1: GL (globins)
gr.2: CH (chromosomal proteins)
gr.3: CS,OR (contractile system, other respiratory)
gr.4: EI,TX (enzyme inhibitors, toxins)
gr.5: CC,OX,TR,HL,LY,IS,SY,GF,HO,IF,IG,RI,FI (the rest)
matrix of misclassification probabilities p(i,j)
      0.99      0.00      0.00      0.00      0.01
      0.01      0.74      0.00      0.16      0.09
      0.07      0.01      0.72      0.05      0.16
      0.11      0.02      0.07      0.59      0.21
      0.09      0.03      0.13      0.21      0.53
another seqid? y([cr]) or n
N
FORTRAN STOP
```

Figure 4. Functional discrimination of an oncogene protein according to periodicity, hydrophobicity, net charge, and sequence length.

superfamilies, excluding miscellaneous and hypothetical proteins, cluster into six large groups when periodicity of nonpolar amino acids, hydrophobicity, net charge, and sequence length are used as parameters. The first parameter was defined as the average amplitude of 3.6 residue periodicity of nonpolar amino acids taken for the segment of 12 residues long at a time, which is representative of an amphipathic helix. Thus, as indicated in Fig. 4, this parameter set places globins in one group (primarily by high periodicity), chromosomal proteins in another (by positive charges), contractile proteins and some respiratory proteins in a third (by negative charges), enzyme inhibitors and toxins in a fourth (by low hydrophobicity and small length), and enzymes except hydrolases in a fifth (by long sequences). If discriminant analysis is now used to ask, for example, in which cluster the v-myc oncogene sequence (which was not part of the original group of proteins) is most likely to belong, one finds that it is a member of the third cluster (contractile system proteins and respiratory proteins other than cytochromes). Even though the overall probability of belonging to one group was over 70% (average along the diagonal in Fig. 4), more than double the probability of belonging to all other groups combined, this is not one of the more precise classifications. However, the probability of misclassifying a globin was 1% using the above four parameters as the basis for discrimination; specifically, only two out of 169 sequences were misclassified. It is interesting to note that discrimination was in general better when a few suitable variables were chosen, as in Fig. 4, than when clustering was attempted on the basis of unbiased residue composition. This seems to suggest intrinsic degeneracy of the properties of amino acids (Klein, Kanehisa and DeLisi, in preparation). The methodology is currently being extended to include structural parameters for finer discrimination, and the class of macromolecules is being expanded to include nucleic acids.

Program Package for Sequence Homology Search

In the next level of analysis we search sequence homologies. The repertoire of homology search programs is shown below:

SEOH	Local homology search in nucleic acid sequences
SEQHP	Local homology search in protein sequences
SEQDP	Significance of homology in protein sequences
SEQP	Local secondary structures in nucleic acid sequences
SEOA	Global homology alignment of two close sequences

These programs, which are available upon request, are adopted from the sequence analysis package originally written for the Los Alamos sequence library.[6] All of them employ dynamic programming algorithms based on base by base (or amino acid by amino acid) comparisons. Therefore, the number of operations required is proportional to the product of sequence lengths being compared, except in the SEQP and SEQA programs where the user can limit the search size and increase the efficiency. As shown in Table 1 these programs meet most of the requirements of searching global homology, local homology,

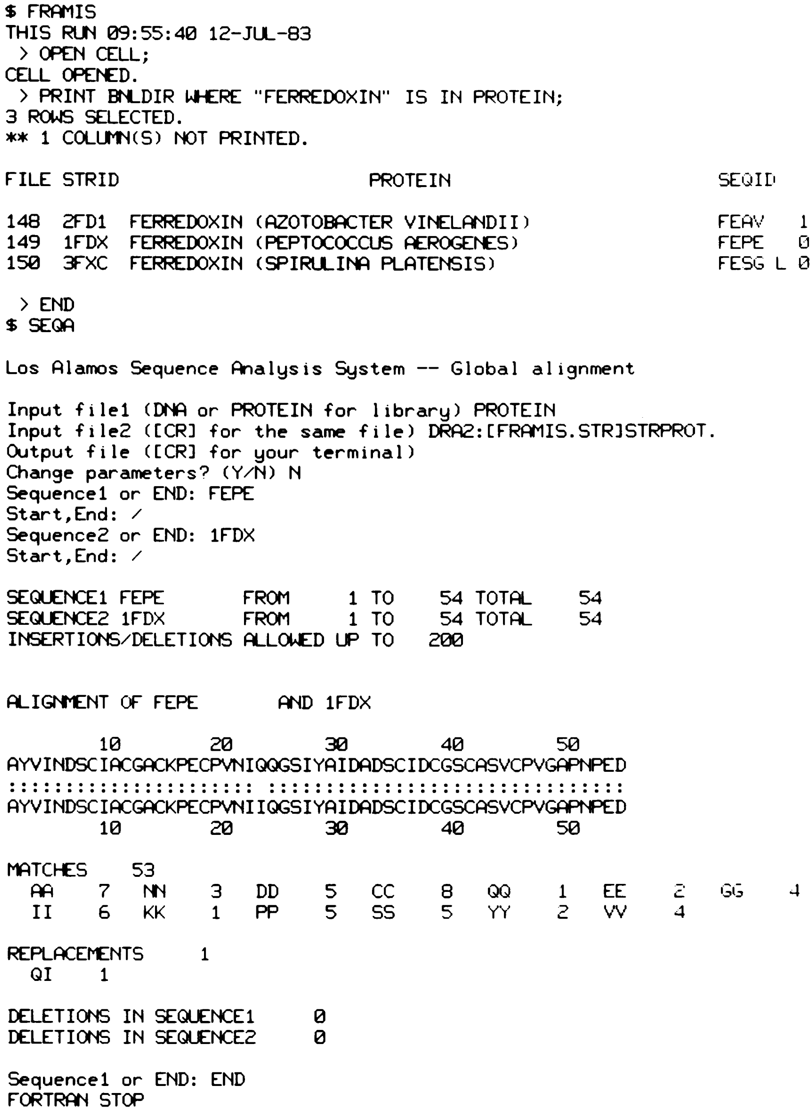
Table 1. A package of sequence homology search programs

Sequence	Global homology	Local homology	Internal homology	Significance check
Nucleic acid sequences	SEQA	SEQH	SEQH	SEQH
Nucleic acid sequences in free energy values	——	——	SEOP	——
Protein sequences	SEQA	SEQHP	SEQHP	SEQDP

and internal homology (direct repeats and inverted repeats) with different criteria of similarity, i.e., mutation data and free energies, as well as checking the statistical significance of homologies found. Fig. 5 is an example of using one of the homology search programs. Table BNLDIR in our FRAMIS database contains the cross-reference of Brookhaven sequence identifier STRID and NBRF sequence identifier SEQID, which was created by identifying the same journal reference in the two databases. Using this information the SEQA program can identify any discrepancy in the two databases.

Homology Search against the Database

The most direct way to identify functional and structural properties of a sequence is to search for homologies of well-characterized sequences stored in the database. The homology search against the protein sequence database, which contains about 0.4 million residues, by the SEQHP program requires several hours of CPU time for a query sequence of a few hundred residues on VAX 11/780. However, the search against the nucleic acid sequence database is not practical by the SEQH program. We have therefore implemented a faster version of the database search routine for nucleic acid sequences. The SEQH program or the Goad-Kanehisa algorithm[7] is a powerful, and probably the most sensitive, method to locate all local homology alignments in two sequences. It is, however, a time-consuming method because of its sophisticated path pruning procedure. It is also memory-consuming because it stores all alignment paths. The fast search program reports only whether the two sequences being compared contain a candidate of good alignment or not, based on the approach used by Smith and Waterman,[8] and if it does, the alignment score and the starting position of the alignment. It runs roughly ten times faster than the SEQH program and can handle much longer sequences. In order to achieve even faster execution, one has to resort either to an approximate procedure[9] or to a machine with parallel processing capabilities. We take the



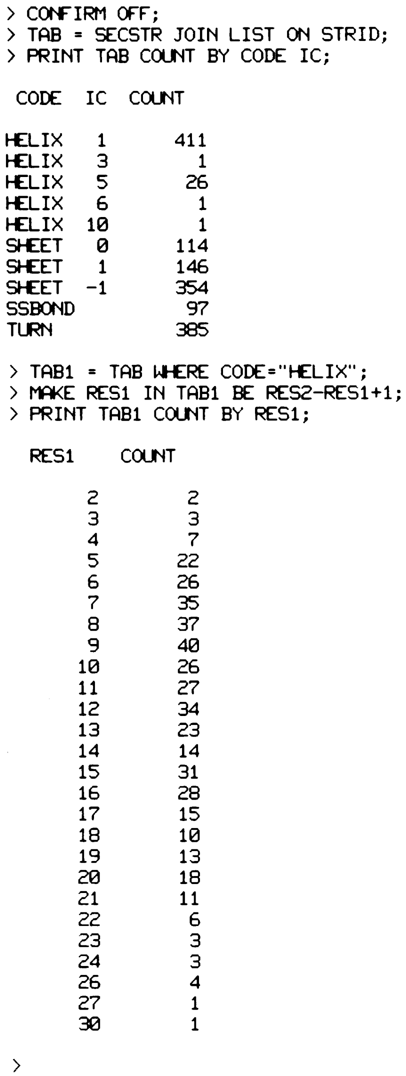
```
$ FRAMIS
THIS RUN 09:55:40 12-JUL-83
 > OPEN CELL;
CELL OPENED.
 > PRINT BNLDIR WHERE "FERREDOXIN" IS IN PROTEIN;
3 ROWS SELECTED.
** 1 COLUMN(S) NOT PRINTED.

FILE STRID                    PROTEIN                        SEQID    DATE

148  2FD1  FERREDOXIN (AZOTOBACTER VINELANDII)               FEAV   12-NOV-81
149  1FDX  FERREDOXIN (PEPTOCOCCUS AEROGENES)                FEPE   01-AUG-76
150  3FXC  FERREDOXIN (SPIRULINA PLATENSIS)                  FESG L 07-DEC-81

 > END
$ SEQA

Los Alamos Sequence Analysis System -- Global alignment

Input file1 (DNA or PROTEIN for library) PROTEIN
Input file2 ([CR] for the same file) DRA2:[FRAMIS.STR]STRPROT.
Output file ([CR] for your terminal)
Change parameters? (Y/N) N
Sequence1 or END: FEPE
Start,End: /
Sequence2 or END: 1FDX
Start,End: /

SEQUENCE1 FEPE       FROM     1 TO     54 TOTAL     54
SEQUENCE2 1FDX       FROM     1 TO     54 TOTAL     54
INSERTIONS/DELETIONS ALLOWED UP TO    200


ALIGNMENT OF FEPE        AND 1FDX

         10        20        30        40        50
AYVINDSCIACGACKPECPVNIQQGSIYAIDADSCIDCGSCASVCPVGAPNPED
:::::::::::::::::::::: :::::::::::::::::::::::::::::::
AYVINDSCIACGACKPECPVNIIQGSIYAIDADSCIDCGSCASVCPVGAPNPED
         10        20        30        40        50

MATCHES     53
  AA    7   NN    3   DD    5   CC    8   QQ    1   EE    2   GG    4
  II    6   KK    1   PP    5   SS    5   YY    2   VV    4

REPLACEMENTS      1
  QI    1

DELETIONS IN SEQUENCE1      0
DELETIONS IN SEQUENCE2      0

Sequence1 or END: END
FORTRAN STOP
```

Figure 5. A sample run of one of the sequence homology search programs. The discrepancy shown here is remarked in the original Brookhaven file.

latter course and the source codes are being vectorized for execution on a Cray.

Structure Prediction

The prediction of RNA secondary structures is, basically, a homology



```
> CONFIRM OFF;
> TAB = SECSTR JOIN LIST ON STRID;
> PRINT TAB COUNT BY CODE IC;

 CODE   IC  COUNT

HELIX   1     411
HELIX   3       1
HELIX   5      26
HELIX   6       1
HELIX  10       1
SHEET   0     114
SHEET   1     146
SHEET  -1     354
SSBOND         97
TURN          385

> TAB1 = TAB WHERE CODE="HELIX";
> MAKE RES1 IN TAB1 BE RES2-RES1+1;
> PRINT TAB1 COUNT BY RES1;

  RES1     COUNT

        2        2
        3        3
        4        7
        5       22
        6       26
        7       35
        8       37
        9       40
       10       26
       11       27
       12       34
       13       23
       14       14
       15       31
       16       28
       17       15
       18       10
       19       13
       20       18
       21       11
       22        6
       23        3
       24        3
       26        4
       27        1
       30        1

>
```

Figure 6. Analysis of secondary structure types and helix length distribution in the Protein Data Bank.

search problem. It is a better defined problem than the prediction of protein structures, in the sense that it is simply a search for complementary base pairs with somewhat complex weighting schemes representing free energies of base pairs and loops. For most practical purposes we want to know candidates of good, or stable, hairpin structures, rather than to calculate further the best structure when all possible hairpins are assembled in a single molecule. The former requires, at least, the order of N^2 operations while the latter N^3 operations, where N is the sequence length. In addition, if we are only interested in hairpin structures up to a specified length, say L, the former can be done by the order of L x N operations. Namely, the number of operations required to calculate all hairpin structures is linearly proportional to the sequence length. This is the approach taken by Kanehisa and Goad[10] or in the SEQP program in Table 1.

The development of our predictive methods for protein structures is still in a preliminary stage. We are currently taking various statistics from the Brookhaven database and developing the statistical mechanical methods necessary for converting the observed frequencies into thermodynamic parameters. It is generally believed that most of the existing methods based on database statistics would not be improved very much by simply increasing the size of the database. However, this does not necessarily mean intrinsic problems of statistical approaches. The number of entries in the Protein Data Bank has doubled in the past four years, and we are in the process of critically reviewing various predictive methods thus far proposed and strengthening them in the areas indicated by the analysis.

As a final example of information retrieval in FRAMIS, Fig. 6 shows an analysis of secondary structures in a set of 87 proteins selected from the structure database. A list of sequence identifiers STRID was prepared for this set in table LIST (contents not shown). It is then used to extract corresponding entries from table SECSTR, which contains secondary structure information for all proteins in the Brookhaven database. The first tabulation shows different types of helices, α helices (IC=1), 3_{10} helices (IC=5), etc., and different types of β sheets, parallel (IC=1) and anti-parallel (IC=-1). The second tabulation is the distribution of helices against the helix length. Columns RES1 and RES2 of table TAB contains start and end residue numbers of each helical segment. Therefore, after the MAKE operation RES1 contains the helix length. Thus, analytical features of FRAMIS, coupled with powerful relational operations, can free us from writing many special purpose programs.

DISCUSSION

Database technology is one of the most important areas in the application of computers. Unfortunately, however, many existing databases are often mere repositories of un-organized data; replacements of dictionaries and encyclopedia. In such instances, computers are used for storage and retrieval of information, which can only be parsed and understood by humans. The relational database on the other hand, allows us to use computers to make the logical inferences, thus increasing by many orders of magnitude the capability for data retrieval and analysis. The simplicity of the relational database is due to the well-defined mathematical framework of relational algebra. Given the exponential growth of data in all areas of biology, it is unlikely that any individual knows more than a small fraction of the literature, even in his or her own specialty. Under these circumstances, the relational database, with appropriately integrated software, offers the best opportunity for making logical connections among datasets that might not otherwise have been fully assimilated.

*On leave from Theoretical Division, University of California, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

References

1. Kanehisa, M., Fickett, J.W., and Goad, W.B. (1984) Nucl. Acids Res. this issue.
2. Fickett, J., Goad, W., and Kanehisa, M. (1982) Los Alamos Sequence Library: A Database and Analysis System for Nucleic Acid Sequences, LA-9274-MS, Los Alamos National Laboratory.
3. Dayhoff, M.O., Hunt, L.T., Barker, W.C., Orcutt, B.C., Yeh, L.S., Chen, H.R., George, D.G., Blomquist, M.C., and Johnson, G.C. (1983) Protein Sequence Database, National Biomedical Research Foundation, Washington, D.C.
4. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) J. Mol. Biol. 112, 535-542.
5. Jones, S.E., Ries, D.R., Lyles, L., Dittli, A.L., and Johnson, K.W. (1981) FRAMIS Reference Manual, LCSD-554, Lawrence Livermore National Laboratory.
6. Kanehisa, M.I. (1982) Nucl. Acids Res. 10, 183-196.
7. Goad, W.B. and Kanehisa, M.I. (1982) Nucl. Acids Res. 10, 247-263.
8. Smith, T.F. and Waterman, M.S. (1981) J. Mol. Biol. 147, 195-197.
9. Wilbur, W.J. and Lipman, D.J. (1983) Proc. Natl. Acad. Sci. USA 80, 726-730.
10. Kanehisa, M.I. and Goad, W.B. (1982) Nucl. Acids Res. 10, 265-278.