Volume 12 Number 1 1984 Nucleic Acids Research

ACNUC: a nucleic acid sequence data base and analysis system

M.Gouy[1,2], F.Milleret[3,2], C.Mugnier[4], M.Jacobzone[1] and C.Gautier[1,2]

[1]Laboratoire de Biométrie, 43 bd du 11 nov., [2]Institut d'Evolution Moléculaire, Université Lyon 1,69622, [3]Laboratoire d'Informatique Appliquée, INSA, 20 av Albert Einstein, 69621 Villeurbanne, and [4]Citi 2, 45 rue des Saints Pères, 75006 Paris, France

Received 19 August 1983

ABSTRACT

Structured as a data base and associated with data analysis tools, ACNUC allows both on-line access to a central computer and local exploitation of published nucleotide sequences. Its data retrieval capabilities seem to be presently the most powerful available.

INTRODUCTION

For ten years, we have been searching for general structures in nucleic acid sequences (1-5). This led us to construct a nucleic acid sequence bank (ACNUC) and to publish several compilations of codon frequencies (1-4) and a sequence handbook (6). More recently, both complexity and volume of the data has required a complete modification of the bank structure. Furthermore ACNUC has been chosen for the French project of computerisation of nucleic acid sequences. This project emphasizes on-line access to software on a central computer (CITI 2 in Paris). However a transportable bank has also been developed to allow creation of new programs, modification of data structure, etc, which may be inconvenient on the central computer. We want to here show the efficiency of data base structures which take into account important biological concepts and allow users to develop their own research programs. We shall use Backman's diagrams, well known tools in Computer Sciences (see for example (7)) that seem neglected in Biology where they can be helpful in modelisation.

I. BANK STRUCTURE

Among biological data, nucleic acid sequences show specific features related to the central place these molecules occupy in living systems. They relate to several biological phenomena, for example: phylogeny, protein function (for coding sequences) and expressivity level, cellular mechanisms such as replication, transcription, translation, mutation

and base modification,or to technical aspects (vectors, restriction enzyme sites). It was not possible to integrate all these relationships and we had to make choices between them. Consequently, other structures of the bank could have been created. Our main choice has been the definition of a "sequence". Nucleotide sequencing furnishes genome segments whose ends do not always coincide with biological limits. We define a sequence as a region of nucleotides having homogeneous function and not only as the succession of bases as published (cf fig. 1).

The ACNUC structure is represented in Figure 1 in a Backman diagram. This easily understandable diagram identifies and coordinates the data. We note that the hierarchical relationships between organisms and categories (fig. 1) allow the use of connected trees.

II. DATA RETRIEVAL

Based on this logical structure two implementations of the bank with their associated retrieval software have been developed. The first, for on-line use, is installed on a central computer (Citi 2, HB 66) and uses a data base management system (IDS2). Data retrieval uses the "menu principle", which constantly guides the user by a dialogue in natural language. The system directs this dialogue, users can only answer "yes" or "no" or, more generally, choose a number in a list of proposals. This query language is easy to use, but because search is hierarchically organized every node must be run without any shortcut. This software allows the extraction of a sequence subset on a multi-criterion basis (organism, category, type, author).

We have recently developed a second transportable implementation of the same logical data structure using Fortran 77 programs and direct access files. This new software is devoted to local exploitation of subsets of the whole bank. Based on a syntax more formal than that of the on-line software, it offers real query language. The query program REQUETE works by defining sequence subsets of the bank. A sequence subset is a combination of predefined subsets associated to all entry points in the bank (e.g. sequences of any species, category, type or reference) with the 3 standard logical operators AND, OR and NOT, the length operator and operators giving the mother or the daughter sequences of a subset. At least 15 sequence subsets can be simultaneously handled by the program (the exact number depends on total sequence number in the bank and on available computer memory), and previously defined subsets can be used in the definition of new ones. We will

not detail here the query syntax which is developed at length in the user's manual (+), but give two examples of possible queries :

"E=HOMO" ET "C=IMMUNOGLOBULINE" ET "T=INTRON"

defines the subset of all introns in human immunoglobulins.

"E=PROCARYOTES" ET "C=NAR" ET "C=1982"

defines the subset of all prokaryotic sequences published in Nucleic Acids Research in 1982.

After a subset has been defined, the mnemonics, references and comments of the sequences contained can be examined. The sequences can be printed out with various formats or extracted into working files for further analysis (see part III) . Moreover site-directed extractions are available as : find all sequences around exon-intron functions including 10 bases on each side. The software manages all console outputs on video display or hardcopy terminals. Access times are generally small because few disk readings are necessary. This is achieved by a hashing method on sequence mnemonics and by a wide use of bit tables for the internal storage of links of sequences, species and categories with sequence subsets.

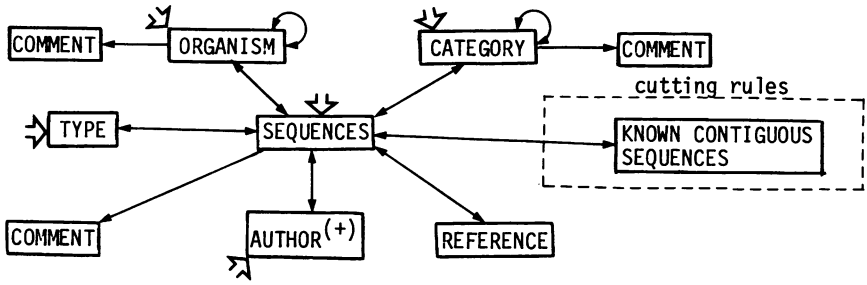
Although technically different, these two softwares are based on the same internal logic, that described in fig. 1. Nevertheless the Fortran package requires more practice for efficient use but allows more sophisticated queries. We emphasize that in both cases subsequences are directly available. Hence, one can for example get without any programming effort, the protein coding region of a gene and consequently the protein sequence.

III. STUDYING SEQUENCES

We consider ease of implementing new programs as a major feature in a scientific data base, although the complexity of the logical structure and the necessity of specialized software are somwhat contradictory to this ease. Moreover, it is important that modifying the structure of the base or the software (such modifications will be necessary to follow developments in molecular biology and computer science) does not imply modification of every utility or research program.

For these reasons, we chose a two-step process for studying sequences: First, sequence selection by a query program for the present bank structure and extraction of the selected sequences into a simply structured file (either sequential or direct access). Second, running utility programs on these new files. This choice requires more disk space,particularly for work on large sequence subsets,but this requirement should become

BACKMAN DIAGRAM



(+) This feature is not available in the FORTRAN software

Boxes represent the data and arrows the links between boxes.

→ One-way association : an element of the set of departure can be associated to one or several elements of the arrival set (example : several comments are associated to a sequence).

↔ Two-way association (example : a sequence has been worked on by several authors and an author has studied several sequences).

Association between elements of the same set. This symbol represents hierarchy, one arrow indicates descent in the tree, the other ascent.

Access point in the bank, that is,criterion for sequence selection.

DATA AND LINKS DESCRIPTION

SEQUENCE Each sequence is named by a mnemonic code which is its unique identifier. This mnemonic comprises three regions separated by the character /. The first identifies the organism and is a contraction of its Latin name, the second is the code of the "gene" and the third is the type of sequence. For example HOMO/A1BGL/PC is the coding part (PC) of the α1 globin (A1BGL) of man (HOMO). For this same gene and organism the following sequences exist : HOMO/A1BGL/I1 first intron; HOMO/A1BGL/I2 second intron; HOMO/A1BGL/M mature messenger RNA; HOMO/A1BGL/T primary transcript.

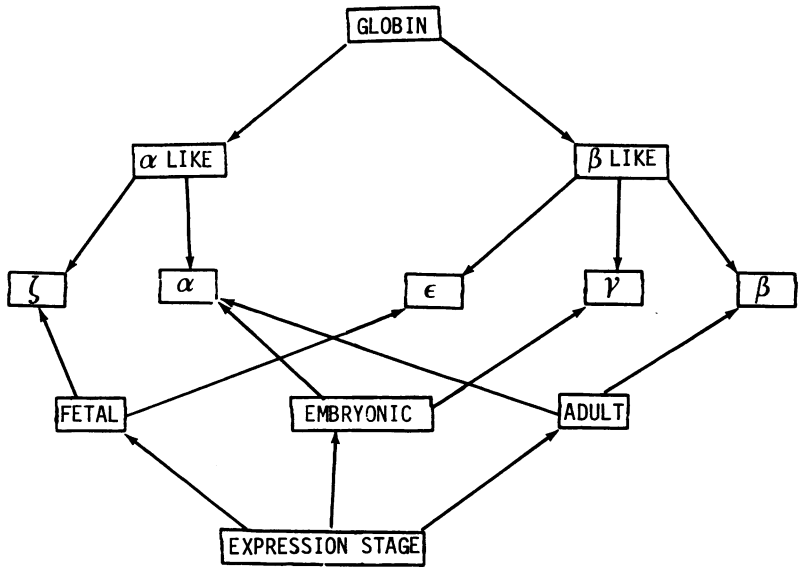
To avoid increasing the bank volume, the longest continuous sequences known are the only ones actually written in the bank. These sequences are named mother sequences. All sequences in the bank (daughter sequences) are defined from mother sequences by the use of cutting rules. Cutting rules allow :

- extraction of fragments from one or several mother sequences
- concatenation of extracted fragments
- complementation and inverse reading

ORGANISM Organism here stands for any level in the tree of species classification. So mammal is an "organism". The arrow represents the hierarchical classification of species. The level numbers in each branch do not need to be equal and combination of several trees is allowed (see CATEGORY for example). Specifying any point in the species tree selects all sequences attached to the lineage of this point in the tree.

CATEGORY Categories are keywords assigned to the sequences to allow their rapid retrieval. Like species, categories are hierarchically structured. Moreover since many characterizations are possible for a sequence, connected trees are allowed.

For example, the ϵ globin in man is a β like globin and is an embryonic protein. Part of the hierarchy of globin is described in the following diagram.



Two connected trees with roots "GLOBIN" and "EXPRESSION STAGE" are possible.

TYPE Sequence types characterize the sequence functions in cellular processes. Nine types are currently defined : primary transcript, mature mRNA, protein coding part, intron, rRNA, tRNA,snRNA, sequences with no unique type (these have daughter sequences with specific type), other cases (e.g. pseudogenes).

Figure 1 Logical structure of ACNUC bank

less important as disk cost diminishes. Also, this strategy allows any user to treat personal or unpublished sequences with any utility program in access-protected disk files.

Staden's program library and that developed by the "Laboratoire de Génétique Moléculaire des Eucaryotes" in Strasbourg, are presently available on the central computer for working on sequence files extracted from ACNUC and EMBL nucleotide sequence banks.

The Fortran query program copies the sequence subsets from the base into direct access working files. Although these files are simply sufficiently structured to be directly used, utility program ANALSEQ has been developed to facilitate the implementation of new research tools. ANALSEQ frees the programmer from handling data in memory and managing console output and can receive new subprograms for new sequence treatments. This program comprises a sequence editor (extraction, concatenation, inversion, complementation), allows sequence printing,insertion and deletion in or from the working file and various standard sequence analysis routines. Moreover, all sequence reading is handled via an input buffer that allows manipulation of sequences of any length, even on machines with little memory such as small 16-bit computers. Currently available standard analysis routines are : gene translation (with the adequate genetic code); numeric indexing of nucleotides, codons and aminoacids; restriction site mapping; open reading frame searches; base and codon frequencies; sequence comparison by the dot-matrix method (written for a SECAPA, Techtronix compatible, video display).

Other more statistically oriented routines developed by our group are available : interface with multivariate analysis programs (correspondance analysis, automatic classification (1-3)); classical non parametric tests (rank sum and variance, run number, multiple groupings and center grouping); specific non parametric tests (7); sequence simulations (conserving the codon content or the protein sequence); computation of p1 and p2 indexes connected to bacterial gene expressivity (4).

IV. SYSTEM DIFFUSION AND ON-LINE ACCESS

Users can access the data-base on-line through the central computer (+) by telephone line or Modem or by Transpac in France or Euronet elsewhere in Europe. When the communication is established a menu appears that gives access to every program described here. Hence this on-line access does not require computer expertise.

The Fortran 77 software, described in a user's manual (in French) and the sequence data are available upon request on magnetic tape (++).

V. CONCLUSION

ACNUC provides a data base with a relatively rich logical structure allowing multiple access, and sequence retrieval and extraction into simply structured working files. Various sequence analysis programs are currently gathered and interfaced with the structure of the working files. This organization guaranties wide possibilities for software evolution without greatly modifying user and programer habits. Both software versions, one on-line from a central computer and the transportable one written in Fortran 77, have the same logical structure and sequence data. They comprise currently (August 1983) the sequences entered at Lyon (731418 nucleotides). By September 1983, the new sequences in the EMBL nucleotide library should begin to enter the data base. New entries will be supervised by scientific experts in order to achieve greater precision in the data, to provide a commentary for each sequence and to maintain a logical structure current with the latest developments in the field.

+ Please contact C. Mugnier at Citi 2 for information.
++ Request should be sent to F. Rodier, ACNUC Users Club, IRBM, Université Paris 7, Tour 43, 2 pl. Jussieu, 75251 Paris Cedex 5 FRANCE.

ACKNOWLEDGEMENTS

We thank J. Favrel for help in data bank structuration. S. Louail, D. Mouchiroud and P. Perrin are scientifically responsible for large data domains in the bank, their work has been essential for building ACNUC. We are indebted to all molecular biologists who have shown an interest in the bank and particularly to P. Kourilsky, who initiated the present French project, and F. Rodier who is responsible for the "Club d'utilisateurs". The computer facilities of the Laboratoire de Biométrie (Lyon I) and the Laboratoire d'informatique appliquée (INSA Lyon) have been used.

REFERENCES

1. Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pavé, A. (1980) Nucleic Acids Res. 8, r49-r62.
2. Grantham, R., Gautier, C., Gouy, M. (1980) Nucleic Acids Res. 8, 1893-1912.
3. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R. (1981) Nucleic Acids Res. 9, r43-r74.
4. Gouy, M., Gautier, C. (1982) Nucleic Acids Res. 10, 7055-7074.
5. Grantham, R. (1980) FEBS Lett. 121, 193-199.
6. Gautier, C., Gouy, M., Jacobzone, M., Grantham, R. (1981) Nucleic Acid Sequences Handbook, vol 1,2, Praeger Publishers, London.
7. Tricot, J. (1982) L'informatique professionnelle 2, 35-65.
8. Gautier, C. (1982) XI^{e} International Biometric Conference, Toulouse 6-11 Sept..