BioSOS: A program package for the analysis of biological sequences in the microcomputer

R. Bringas, R. Ricardo, J. Fernández de Cossío, M. Ochagavía, A. Suárez and R. Rodríguez

Center for Genetic Engineering and Biotechnology (CIGB), P.O. Box 6162, Havana 6, Cuba

Received in February 1992
Approved in March 1992

SUMMARY

A series of programs for the analysis and management of biological sequences in the microcomputer were grouped with a common user interface. BioSOS provides a comprehensive menu with hot keys to select options such as: a sequence oriented editor, EMBL and SWISSPROT databank access either in CD-ROMs or in a compressed hard disk version, restriction analysis, translation, dot plot sequence comparison, alignment and protein profiles, among others. Most of the results can be evaluated in a graphic form, and the charts can be printed or saved into disk files for further reading with desktop publishing programs. BioSOS was written for the midrange user, with minimum hardware requirements but achieving an optimum performance. In this paper we summarize the program options and present some of its features.

RESUMEN

BioSOS agrupa una serie de programas escritos para la manipulación y el análisis de secuencias biológicas en las microcomputadoras. Mediante un menú con "teclas calientes" se pueden seleccionar opciones como: un editor de secuencias, acceso a los bancos de datos del EMBL y el SWISSPROT en discos compactos (CD-ROM) o en una versión comprimida en el disco duro, análisis de restricción, traducción, comparación de secuencias usando matriz de puntos, alineamiento y perfiles de proteínas, entre otras. La mayoría de los resultados pueden ser evaluados mediante gráficos que pueden ser impresos o salvados en disco para su posterior lectura con otros programas de edición.

BioSOS fue escrito para el usuario promedio, tratando de lograr el máximo de prestaciones sin el uso de configuraciones muy complejas. En este artículo se trata de resumir las diferentes opciones con las que cuenta hasta ahora el programa y presentar algunas de sus características.

INTRODUCTION

For years the people involved in molecular biology research have used computers not only to store their results, but also as very powerful tools in the analysis of their data, before and after the experiments, i.e. they plan very carefully what is to be done in the laboratories and then they analyze the outcoming data, with the help of a computer system. Most of the obtained data can be related with DNA, RNA and amino acid sequences and the aim of the programmers is focused on providing efficient tools for sequence analysis.

Since the establishment in 1982 of nucleic acid databanks e.g. in the European Molecular Biology Laboratories (EMBL, Heidelberg) and in the National Institutes

Publicado por el Palacio de las Convenciones, La Habana, Cuba

of Health (NIH, USA) the accumulated data reach at present more than 70 million nucleotides. Along with this progressive accumulation of data, many mathematical, statistical and computational methods have been developed and there are many programs directed to solve very specific problems or covering a group of general procedures.

At the Center for Genetic Engineering and Biotechnology in Havana we have our own implementation of many of the algorithms used worldwide, and those programs have become the building blocks of an integrative software package: BioSOS.

BioSOS was written for the midrange user and does not require either complex and expensive hardware platforms or specialized training for its operation.

In this paper we present the main features and some of the programming engine of BioSOS, a detailed discussion of each of the algorithms and options is beyond the scope of this introductory article and can be seen in the BioSOS user's manual.

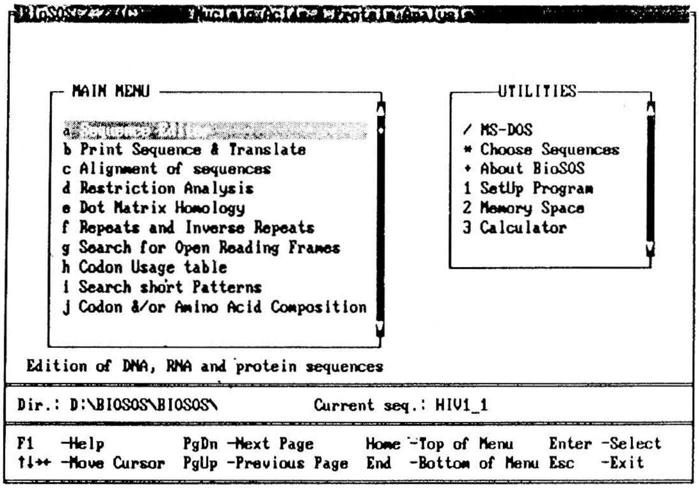
MATERIALS AND METHODS

BioSOS is presented with a main menu and a set of utilities (Fig. 1). The user can select the desired option and then the menu shell loads the specific program. We choose to keep most of the procedures compiled into separate programs because many of the algorithms employed to work with biological sequences are memory intensive and use almost the entire available memory. To keep track of the parameters utilized by all the programs e.g. the active and the secondary sequence, working directories, etc., we use disk files containing all this information.

The main menu tree and the utilities are structured as follows:

Main Menu

- Sequence editor
- Translation and printing
- Sequence alignment
- Restriction analysis
- Dot-plotting of homology
- Searching for repeats and inverse repeats
- Searching for open reading frames
- Codon usage table
- Pattern searching
- Aminoacids and/or codon occurrence
- Isoelectric point calculation for proteins
- EMBL and SWISSPROT databanks access
- Aminoacid data
- Protein secondary structure prediction



```
[illegible]

MAIN MENU                                         UTILITIES

a Sequence Editor                                 / MS-DOS
b Print Sequence & Translate                      * Choose Sequences
c Alignment of sequences                          + About BioSOS
d Restriction Analysis                            1 SetUp Program
e Dot Matrix Homology                             2 Memory Space
f Repeats and Inverse Repeats                     3 Calculator
g Search for Open Reading Frames
h Codon Usage table
i Search short Patterns
j Codon &/or Amino Acid Composition

Edition of DNA, RNA and protein sequences

Dir.: D:\BIOSOS\BIOSOS\              Current seq.: HIV1_1

F1    -Help          PgDn -Next Page       Home -Top of Menu     Enter -Select
↑↓→← -Move Cursor   PgUp -Previous Page   End  -Bottom of Menu  Esc   -Exit
```

FIG. 1. Main menu.

- Hydropathy, flexibility and aminoacid surface exposure profiles
- Searching for probable N-glycosylation sites

Utility Menu

- Operating system function (dir, copy, rename, type, del, mkdir)
- Sequences selection
- Hardware configuration and defaults
- Disk and memory space info
- Calculator

Sequence Editor

The BioSOS multiwindow editor can handle up to nine sequences at a time, the edition can be performed using both the three and the single letter code for amino acids or nucleotides. The windows can be resized and rearranged in the screen to the user convenience.

The editor has two sets of commands in their own menus: general editing command to search, replace, mark, copy or move blocks, and specific options to manipulate sequences, translate, complement, search for open reading frames, printing, etc.

The command interface was built keeping the menu choices along with Word-Star like keystrokes for advanced users.

Sequence Alignment

It is very important for the molecular biologist to compare two sequences, evaluating the degree of similarity or homology. The usual output of such a comparison is a sequence alignment i.e. the two sequences are written one above the other and gaps are inserted when necessary, to fill missing regions in the proteins. For BioSOS we used an algorithm that fits in a linear memory space, about the size of the largest sequence, avoiding the construction of large arrays and the consequent exhaustion of the memory arena. (Myers and Miller, 1988). In order to avoid the excessive insertion of gaps we used two penalty constraints in the algorithm: one for the insertion of a consecutive gap block and the other for the individual insertions. To evaluate if two elements of the sequences are homologous to substitution we used four cost matrixes: unit matrix (Doolitle, 1981), Genetic Code Matrix (Sellers, 1974; Smith *et al.*, 1981), Structure Genetic Matrix (Doolitle, 1979) and Dayhoff's log Odds Matrix (Dayhoff, 1972, 1978). Furthermore, the alignment can be displayed as a graphic profile (Ochagavia *et al.*, 1992), allowing the user the fast identification of any similarity region.

Restriction Analysis

Most of the manipulation of the genetic material is carried out using restriction endonucleases: enzymes that cleave DNA in specific sites. To elaborate the strategy of cleavage and further cloning we evaluate the possible cleavage sites for enzymes selected from a databank (Kessler and Manta, 1990). The selection of the enzymes can be done by different selection criteria such as: palindromic enzymes, commercially available, number of base pairs on cleavage site, and resulting end after the cleavage, among others. The database can be updated manually or using the enzyme catalog distributed by EMBL. The cleavage analysis can be extended also to proteins using a database of endoproteinases.

Dot plot homology

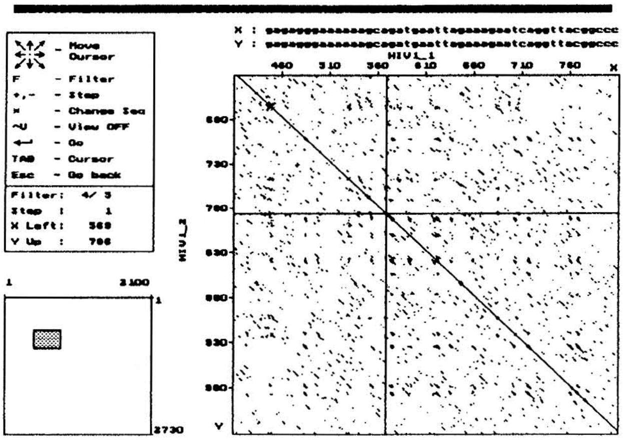
Two sequences can be compared by a dot plot, building a two dimensional array containing each one, and plotting a dot when the contiguous residues, defined by a filter, are coincident in the two sequences. The user can select both the filter and the window for displaying the entire array or only a fragment, and the maximum homology between the two sequences can be detected when densely packed dots draw a line in the display (Fig. 2).

Repeats and inverse repeats

With this program the user can evaluate any repeat, inverse repeat or palindrome in a given sequence. The algorithm used in the search (H. Martinez, 1983) is very fast and accurate.

Searching for open reading frames and patterns

Searching for any open reading frame (fully translatable DNA fragment) is very useful to look at any coding region within a DNA or RNA sequence. This option shows the position and length of possible open reading frames from 5' to 3' and from 3' to 5'. The minimum length and the inclusion or not of stop codons in the generated map is selected by the user.



X : gagaggggaaaaaagcagatgaattagaaagaatcaggttacggccc
Y : gagaggggaaaaaagcagatgaattagaaagaatcaggttacggccc

Filter: 4/ 5
Step : 1
X Left: 568
Y Up : 786

FIG. 2. Dot-plotting of homology.

Another program was written to search for any defined sequence pattern, using a wide variety of wildcards. The user can browse in the map using the cursor to see each found fragment and its coincidence with the defined template.

Aminoacid composition and data, codon occurrence and codon usage tables

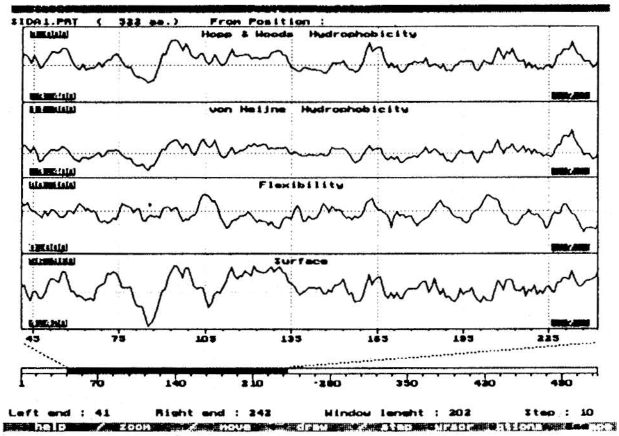
This program calculates the aminoacid composition and occurrence of each codon in a given sequence, including also the average molecular weight of the proteins and their atomic composition.

The main features of the aminoacids: chemical structure, polarity, isoelectric point and hydrophobicity, among others, can be also displayed.

The codon usage tables contains the frequency of occurrence of each codon and the number of genes and bases compiled for a great number of species. With the program it is possible to have access to and do some operations with the existing tables, introduce new tables and edit the already existing.

Protein profiles and N-glycosylation sites

The identification of regions with certain features that could make, for example, potential targets to antibodies, and the prediction of secondary structures, are very important when only the sequence data is known for the protein. We provide programs to plot the hydrophobicity profiles using 12 different scales (Cornette *et al.*, 1987; Thornton



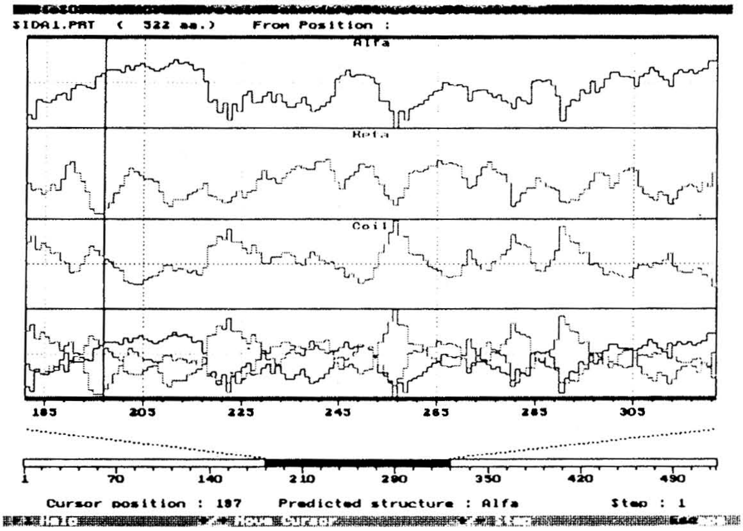
SIDA1.PRT (522 aa.) From Position :

Left end : 41 Right end : 242 Window lenght : 202 Step : 10

FIG. 3. Protein profiles.

and Taylor, 1989) as well as profiles of the flexibility and the surface exposure probability of each aminoacid (Ragone *et al.*, 1989) (Fig. 3).

For the prediction of the secondary structure we used an algorithm that evaluates the tendency of each aminoacid to be in an alpha helix, a beta sheet or random coiled (Garnier *et al.*, 1978) (Fig. 4).



```
SIDA1.PRT   (  322 aa.)     From Position :
```

```
Cursor position : 187    Predicted structure : Alfa          Step : 1
```

FIG. 4. Protein secondary structure prediction.

The consensus sequence Asn-X-Ser/Thr, (X≠Pro), was used to identify the probable N-glycosylation sites.

Sequence banks data handling

BioSOS may have access to EMBL and SWISS PROT databanks, allowing a fast search by entry id, keywords or taxonomic index, the result is an EMBL format file.

Those databanks are distributed usually in compact disks and the programs can use the CD ROM data or a compressed version that can be created in a hard disk. The latter is very convenient in the laboratories that do not have a CD ROM drive or a huge hard disk.

Other features of current and future releases

BioSOS was written in Turbo Pascal 6.0 for IBM compatible microcomputers and the minimum hardware requirements are: 640 kb memory, 20 Mb hard disk and EGA adapter. The current version can handle Epson, HP Laserjet series II and HP Deskjet compatible printers, the new release will include a graphical user interface with icons and full mouse control, multiple sequence alignment, RNA secondary structure prediction and new methods for secondary structure prediction.

RESULTS AND DISCUSSION

BioSOS has been employed for more than three years in the laboratories of the Center for Genetic Engineering and Biotechnology as a tool in many research projects.

With BioSOS we intented to provide our researchers with tools for the most common sequence analysis in our labs in a form of a user friendly program with minimum hardware requirements but with optimum performance.

Having our own implementation of the most commonly used algorithm gives us the possibility to tailor the systems to the specific needs of our molecular biology scientists.

To conduct the performance tests of BioSOS and evaluate the effectiveness of many of the programs, we used many of the already reported experimental data.

The methodology used in each particular program for the calculations and many of the package performance data will be published elsewhere.

ACKNOWLEDGMENTS

The authors are committed to thank many of our colleagues from the CIGB, specially Dr. S. Pérez Talavera and R. Domínguez Pérez for their cooperation, critics and suggestions that helped us in the development of BioSOS since its early beginnings.

REFERENCES

CORNETTE, J.L.; K.B. CEASE; H. MARGALIT; J.L. SPOUGE; J.A. BERZOFSKY and C. DELISI (1987). *J. Mol. Biol.* **195**: 659-685.

DAYHOFF, M.O. (1972). *A model of evolutionary change in proteins. Detecting distant relationships: computer methods and results. Atlas of protein sequence and structure.* Ed. M.O. Dayhoff. National Biomedical Research Foundation, Washington D.C., **5**: 89-110.

DAYHOFF, M.O. (1978). *A. model of evolutionary change in proteins. Matrices for detecting distant relationships. Atlas of protein sequence and structure.* Ed. M.O. Dayhoff. National Biomedical Research Foundation, Washington D.C., **5**: 345-358.

DOOLITTLE, R.F. (1979). *Protein evolution. The proteins.* Eds. H. Neurath, R.L. Hill. Academic Press, New York, **4**: 1-118.

DOOLITTLE, R.F. (1981). *Science* **214**: 149-159.

FENG, D.F.; M.S. JOHNSON and R.F. DOOLITTLE (1985). *J. Mol. Evol.* **21**: 112-125.

GARNIER, J.; D.J. OSGUTHORPE and B. ROBSON (1978). *J. Mol. Biol.* **120**: 97-120.

KESSLER, C. and V. MANTA (1990). *Gene* **92**, 1,2: 1-248.

MARTINEZ, H. (1983). *NAR* **11**: 4629.

MYERS, E.W. and W. MILLER (1988). *CABIOS* **4**: 11-17.

OCHAGAVIA, M.E.; R. RICARDO; J. FERNANDEZ DE COSSIO and R. BRINGAS (1992). PROFALIGN: una representación gráfica del alineamiento de dos secuencias biológicas. *Biotecnología Aplicada* **9**(2): 174-179.

RAGONE, R.; F. FACCHIANO; A. FACCHIANO; A.M. FACCHIANO and G. COLONNA (1989). *Protein Engineering* **2**,(7): 497-504.

SELLERS, P.H. (1974). SIAM. *J. Appl. Math.* **26**: 787-793.

SKOOG, B. and A. WICHMAN (1986). *Trends in Analytical Chemistry* **5**(4): 82-83.

SMITH, T.F.; M.S. WATERMAN and W.M. FITCH (1981). *J. Mol. Evol.* **18**: 38-46.

THORNTON, J. and W. TAYLOR (1989). *Protein Sequencing: a practical approach.* Edited by J.B.C. Findlay & J.M. Geisow, pp. 147-190.