CABIOS COMMUNICATION

Vol. 11 no. 3 1995
Pages 331-335

DNASUN: a package of computer programs for the biotechnology laboratory

A.A.Mironov[1], N.N.Alexandrov[1,2], N.Yu.Bogodarova[1], A.Grigorjev[1,3], V.F.Lebedev[1], L.V.Lunovskaya[1], M.E.Truchan[1] and P.A.Pevzner[1,4]

Abstract

The paper describes a new software package DNASUN developed for supporting gene engineering laboratories. The package provides a user-friendly interface for experimental researches and supports the traditional nucleotide/protein sequence analysis as well as physical mapping, sequencing, plasmid manipulations, optimal oligonucleotide probe selection and other common molecular biology procedures.

Introduction

Molecular biologists encounter significant difficulties trying to obtain DNA sequences and to derive biologically significant information from these data. Numerous molecular biology software tools developed mainly by computer scientists are often found unsatisfactory by biologists. To eliminate a cultural gap between computer scientists and biologists DNASUN has been implemented in the biological environment; the modules of DNASUN have been extensively tested in the biological laboratories and numerous recommendations of biologists have been taken into consideration. The DNASUN user-friendly interface and every-day-use programs have been developing since 1986 in a close collaboration with gene engeneers. Besides every-day-use programs DNASUN contains some programs which are still rare in the general purpose molecular biology software.

DNASUN (version 3.30) has been implemented on IBM PC/AT/PS2 microcomputers (640 Kb memory and 4 Mb hard disk space are required) using Microsoft C (version 5.1) and Microsoft MacroAssembler (version 5.1) (Microsoft is the registered trademark of Microsoft Corporation). The CGA, MGA, EGA, VGA, Hercules videoadapters are supported. DNASUN runs under MS DOS 3.00 (or later versions) and under MS Windows as DOS background application. The running time in the examples below is given for 40 MHz IBM PC/AT-386 (hard disk with random access time 17 ms). DNASUN (version 3.30) is available by contacting A.A.M. for the nominal cost covering shipping, copying and English translation of the manual.

All software components of DNASUN run by the management program under the same name. DNASUN has specialized menu manager and hypertext electronic manual directed towards molecular biology applications. This menu manager allows an apply programmer and even an user to customize a dialogue stream conveniently. DNASUN has his own file manager and user-directed options for viewing, analyzing and printing the results. All programs are supplied with on-line helps and descriptions. If necessary, a user can adjust help windows for his own purposes and even add additional help windows depending on personal convenience. DNASUN is an open package; it can be readily extended by new programs. DNASUN 3.30 contains about 400 menu items, 45 executable modules and provides, in particular, the following functions

- Electroforegrams/gels/sequences input
- DNA and protein sequence editing
- DNA and protein database manager
- Dot-matrices construction
- Pairwise sequence alignment
- Motif search
- Multiple sequence alignment
- Fast database search
- DNA statistics
- RNA secondary structure
- Plasmid manager
- Optimal oligonucleotide probe selection
- Small-scale DNA physical mapping
- DNA sequencing

The software implementations are based on the algorithms developed by the authors and described in more than 40 papers published in 1986-1994 and in the book Alexandrov *et al.*, 1990.

General description of the package

Electroforegrams/gels/sequences input

DNASUN facilitates the input of nucleotide and amino acid sequences and supports various instruments for gels

[1]*Laboratory of Mathematical Methods, National Center for Biotechnology NIIGENETIKA, Moscow 113545, Russia,* [2]*National Cancer Institute, Frederick, MD 21702, USA,* [3]*Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK and* [4]*Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA*

input. Numerous modes, including the input of ambiguous nucleotides, 'blind' input and arbitrary key assignments are allowed. Houston Instruments digitizer TG1017 is used as a basic device for electroforegrams/gels input.

DNA and protein sequence editing

Editing of DNA sequences is a routine and time-consuming procedure. The program for DNA sequence editing (Edit_DNA) provides a convenient way to supply a bare DNA sequence with an additional information: restriction map, open reading frames, genes descriptions, translated proteins, repeats, local similarities, codon usage, etc. All this information can be presented on the same display or output file. As well Edit_DNA models common gene engeneering manipulations: insertions, deletions, substitutions, merging of two DNA molecules, search for a site, keeping of the reading frame, etc. After these operations are done the *Feature Table* is automatically rewritten. Editing data (restriction maps, for example) are also automatically recalculated after DNA manipulations are done.

Protein sequence editing (Edit_Protein) provides all spectrum of editing operations with protein sequences in one or three-letter notation and contains a number of standard tools (hidropatic/charge profiles, various programs for predictions of protein secondary structure, chemical structure of aminoacids in different modes, etc.).

DNA and protein database manager

DNASUN database manager (Database_Manager) works with EMBL/SwissProt external formats. It stores information in a special compressed format that requires three times less space than the external format. In this compressed format both EMBL and SWISS-PROT databases (release 31) require about 190 Mb. In the case of working with a CD ROM Database_Manager uses hard disk for storing some index files. Another option of Database_Manager is an EMBL-GenBank converter for work with GenBank.

Database_Manager is a *hypertext*-like system which greatly facilitates a search in the current versions of EMBL/GenBank containing incompletely annotated entries. It uses interactive approach for search allowing a user to select keys from the key list (index file) instead of formulating queries in SQL. This approach helps a biologists to avoid common mistakes in queries and makes a search procedure simple and clear. The results of a search are presented as a set of documents to allow complex queries in a step-by-step by fashion (by supporting of different logical operations). An interactive associative search in Database_ Manager is supported by different kinds of windows: *Document List*, *Key List*, and *Document View*. From every *Key List* user may select a key and get *Document List* matching the keys. From every *Document List* a user may select a document and browse it. From every *Document View* a user may select a key for the next search, etc. Database_Manager also supports a context pattern search in sequence description that may be used for search in such 'informal' fields as COMMENTS field. It also supports context pattern search for sequences with given similarity level.

Dot-matrices

DNASUN dot-matrix program (Dot_Matrix) uses hashing technique to reduce computational time and provides flexible parameters to reveal significant similarities/reduce statistical noise. Dot_Matrix can be applied iteratively to prompt a biologists the most appropriate parameters and to reveal and zoom the most interesting regions of similarity. Dot_Matrix generated graphical cursor allows one to pick up the most interesting similarity regions. The actual sequence alignment of these regions can be viewed on the same display.

Pairwise sequence alignment

DNASUN has a variety of pairwise sequence alignment programs ranging from the classical global and local alignment algorithms to the fast program for the alignment of close sequences based on a filtration technique in a band (Needleman and Wunsch, 1970, Smith and Waterman, 1981, Lipman and Pearson, 1985, Ukkonen, 1985). The alignment programs allow one to use various modifications of gap penalties. A modification of the local alignment algorithm (Waterman and Eggert, 1987) allows fast searching for suboptimal local alignments and includes the Alignment_Editor for searching, sorting, analysing and updating of the set of found local alignments. For example a representative set of 100 best suboptimal local alignments for sequences of length 1000 bp can be found in 15 s.

Motif search

DNASUN motif search program (Motif_Search) allows a user to type or to select from the library a motif of interest and to look for this motif in a sequence/database. The gaps between various parts of a motif are allowed (for example, ATttrGT (15 ...25) GTCCGA). The Motif_Search draws the plot of the motif score distribution based on a weight matrix.

Multiple alignment

DNASUN contains two programs for multiple alignment: local multiple alignment by consensus matrix (Barton and

Sternberg, 1987, Alexandrov, 1992) and a fast modification of the Vingron–Argos method for local alignment assembly (Vingron and Argos, 1989). The output of the program contains both a multiple alignment (in a format convenient for further editing) and an evolutionary tree constructed by a clustering method.

Fast database search

DNASUN fast database search (Fast_Search) is an implementation of Roytberg (1992) algorithm based on a combination of filtration technique and dynamic programming on l-tuples. Fast_Search draws a plot indicating for each position in the target sequence the number of similar regions in the database. A user can view a list of similar sequences, analyse this list and mark the descriptions of the sequences he is interested in. In this mode a user is able to browse the descriptions that refer to the similar fragments and to analyse the descriptions of (parts of) genes where the similarity is located.

Import from this program to the Database_Manager allows further analysis of the marked set. The actual searching time depends on the parameters, in particular, on the size of l-tuple in the Roytberg (1992) technique. The searching time for a query 1000 bp long and all bacteriophage sequences from EMBL database is 3 min ($l = 7$).

DNA statistics

DNA statistics program (DNA_Statistics) allows a user to plot profiles of different parameters such as GC-content, Purine/Pyrimidine content, codon frequency, etc. The program is very flexible: a user can type a formula for a parameter of interest (for example ratio of start-codon frequency and stop-codon frequency in the first frame). DNA_Statistics plots the profile of this parameter as well as its mean value and standard deviation. In particular, the user can fastly construct various plots related to the functional regions of DNA (coding regions, promoters, TATA boxes) and check his own hypothesis on the relations between the parameters and functional regions.

RNA secondary structure

RNA secondary structure program (RNA_Structure) is based on the kinetics approach for the prediction of RNA secondary structure (Mironov and Kister, 1985; 1986; Mironov and Lebedev, 1993). RNA_Structure uses Monte-Carlo simulation for RNA folding and models the kinetics of RNA secondary structure formation. As a result a user can observe the changes in the RNA secondary structure in the time scale. The transitions between various RNA foldings are calculated from energy considerations. RNA_Structure computes the probabilities of different RNA secondary structures (the *kinetic ensemble of RNA secondary structures*) and provides a user with a few ranked versions of RNA secondary structure all of which, perhaps, exist *in vivo*. It also supports flexible options for viewing the results. In particular, a biologist can find the probability of nucleotide to be paired on the sequence, observe kinetics of pairing for different nucleotides, and compute the probabilities of suboptimal RNA secondary structures. RNA_Structure allows a user to conveniently view different RNA structures in the graphic mode. An example of application of RNA_Structure in the project on the optimization of human interleukin expression in *E.coli* is given in Lomakin *et al.*, 1993.

Plasmid manager

Plasmid manager program (Plasmid_Manager) allows one to build, store and view the genetic and restriction maps of plasmids. As well Plasmid_Manager simulates electrophoresis experiments and compares the expected results with the experimental data. A user can model a wide range of gene engeneering experiments on the computer. For example, a user can select a set of restriction enzymes, cut the plasmid, select a fragment and link it with other fragment. During these operations the program automatically updates the genetic and restriction maps and provides a user with the expected results of the electrophoresis for step by step control. The scenario of Plasmid_Manager user-friendly interface was designed by a gene engineer and was extensively tested in biological laboratories.

Optimal oligonucleotide probe selection

The program for optimal oligonucleotide probe selection (Probe_Selection) designs probes to efficiently select a region (gene) of interest from a clone library in a given vector. This program searches similarities between a gene and a vector and label high similarity segments on gene sequence as unappropriate for probe design. The program uses information from a DNA database about other genes from the same organism and the regions of high similarity with other genes from this organisms are also marked as unappropriate. As well all self-complementary segments on gene are excluded from the consideration. This information and various probe characteristics are displayed on the screen to help in selection of optimal probes.

Double Digest and Multiple Digest DNA physical mapping

In spite of a large number of algorithms for Double Digest Problem (DDP) this problem still cannot be considered

resolved. The major difficulty in DDP is the search and analysis of a large number of hypotheses about the ordering of restriction fragments.

DNASUN physical mapping program (Map_Sun) uses the method of *equivalent transformations* combined with a *consistency checking algorithm* (Pevzner, 1995). Numerous combinatorial procedures (Pevzner and Mironov, 1987) and the method of equivalent transformations allow one to decrease the computational time significantly. If the experimental errors are low MAPSUN constructs physical maps with 10–15 sites for each restriction enzyme. It is well-known that the number of potential physical maps grows exponentially with the number of sites growing (Goldstein and Waterman, 1987). That is why in practice biologists usually get information about a number of double digests (Multiple Digest Problem) and use this information to reduce the combinatorial explosion. Map_Sun allows one to built multiple maps with up to eight enzymes. The resulting maps can be graphically presented by Plasmid_Manager.

DNA sequencing

The program for DNA sequencing (DNA_Sequencing) constructs the sequence assemblies for shotgun sequencing using files from Applied Biosystems automatic sequencer or other source. DNA_Sequencing allows biologists to combine fragments from different gels, to manipulate with a library of gels, search similarity between gels, assembly contigs, edit fragments and contigs, view and print maps of contigs. DNA_Sequencing handles DNA stretchs up to 32 000 nucleotides long and supports a step-by-step strategy allowing a biologist to update the contigs after new fragments have been read. Numerous options saving biologist's efforts were implemented in DNA_Sequencing in a close collaboration with DNA sequencing laboratory of the National Center for Biotechnology, Moscow, Russia. DNA_Sequencing can run in either automatic or semi-automatic regime allowing a biologist to check every step in the sequencing project and to add a new fragment after it has been read. Automatic assembly of 300 fragments with an average length 300 bp takes about 0.5 h (total length of the assembled sequence is 20 000 bp). DNA_Sequencing was extensively used in Russian *Human Genome* program and in French *Bacillus subtilis* project. For example, in the framework of this project Sorokin *et al.*, 1993 used DNASUN for sequencing 28 206 bp region of *B.subtilis* chromosome between the *spoVAF* and *serA* genes. In this study the estimate of the average error rate for sequence assembly provided by DNA_Sequencing was 10^{-4} bp.

Programs in progress

Currently we are developing user-friendly versions of the programs for promoter recognition (Alexandrov and Mironov, 1990), DNA linguistics (Kozhukhin and Pevzner, 1991; Pevzner, 1992b), fast statistical database search for weak similarities (Mironov and Alexandrov, 1988; Pevzner, 1992a), kinetics of protein folding, 3-D structures database manager, etc.

Acknowledgements

We would like to thank A.Bolotin, A Gladkii, Yu.Kozlov, S.Mashko, A Sorokin, and V.Veiko for useful suggestions. The DNASUN project was supported by the National Center for Biotechnology and by the Russian 'Human Genome' program P.A.P. was supported in part by the NSF Young Investigator Award, by the NIH under the grant 1 R01 HG00987-01 and by the DOE under the grant DE-FG02-94ER61919.

References

Alexandrov,A.A., Alexandrov,N.N., Borodovsky,M.Yu., Kister,A.E., Kalambet,Yu.A., Mironov,A A, Pevzner,P.A., Shepelev,V.V. (1990) *Computer analysis of genetic texts* Moscow, Nauka, (Russian), 263 pp.

Alexandrov,N.N., Mironov,A.A (1990) Application of a new method of pattern recognition in DNA sequence analysis: a study of *E.coli* promoters. *Nucleic Acid Res.*, **18**, 1847–1852

Alexandrov,N.N. (1992) Local multiple alignment by consensus matrix. *Comput. Applic. Biosci.*, **8**, 339–345

Barton,G.J. and Sternberg,M J.E. (1987) A strategy for the rapid multiple alignment of protein sequences. *J. Mol Biol.*, **198**, 327–337

Goldstein,L and Waterman,M S. (1987) Mapping DNA by stochastic relaxation. *Advances Appl. Math.*, **8**, 194–207.

Kozhukhin,C.G., Pevzner,P.A. (1991) Genome inhomogeneity is determined mainly by WW and SS dinucleotides *Comput. Appl Biosci.*, **7**, 39–49.

Lipman,D.J and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Lomakin,I.B., Lebedeva,M.I., Konusova,V.G., Velichko,E.V., Lebedev,V.F., Ketlinsky,S.A., Vinetsky,Y.P, Mashko,S.V (1993) T7 RNA Polymerase derived expression of the human interleukin-8 cDNA in *Escherichia coli*. Purification and initial characterization of recombinant hIL-8., *Biotechnol*, **10**, 10–15.

Mironov,A.A., Alexandrov,N N. (1988) Statistical method for rapid homology search *Nucleic Acids Res.*, **16**, 5169–5173.

Mironov,A., Kister,A. (1985) A kinetic approach to the prediction of RNA secondary structures *J. Biomol. Dynam. Struct.*, **2**. 953–961.

Mironov,A, Kister,A. (1986) RNA secondary structure formation during transcription *J. Biomol. Dynam. Struct.*, **4**. 1–9

Mironov,A A., Lebedev,V.F (1993) A kinetic model of RNA folding. *BioSystems*, **30**, 49–56

Needleman,S. B., and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of 2 proteins *J Molec. Biol* **48**, 443–453.

Pevzner,P.A (1992a) Statistical distance between texts and filtration methods in rapid similarity search algorithm. *Comput Appl Biosci.*, **8**, 1992, 121–127.

Pevzner,P.A. (1992b) Nucleotide sequences versus Markov models. *Chemistry and Computers*, **16**, 103–106

Pevzner,P.A. (1995) DNA physical mapping and alternating Eulerian cycles in colored graphs. *Algorithmica*, **13**, 77–105.

Pevzner,P.A., Mironov,A.A. (1987) An efficient method for DNA physical mapping. *Molec. Biol*, **21**, 788–796.

Roytberg,M.A. (1992) Fast algorithm for optimal alignment of symbol sequences. *DIMACS Series in Discrete Mathematics and Computer Science*, **8**, pp. 113–126

Smith,T.F. and Waterman,M.S (1981) Comparison of biosequences. *Adv. Appl. Math.* **2**, 482–489.

Sorokin,A.V., Zumstein,E., Azevedo,V., Ehrlich,S.D., Serror,P. (1993) The organization of the *Bacillus subtilis* 168 chromosome region between *spoVA* and *serA* genetic loci, based on sequence data. *Mol. Microbiol.*, **10**, 385–395

Waterman,M S. and Eggert,M. (1987) A new algorithm for best subsequence alignments with application to tRNA – rRNA comparisons. *J. Mol. Biol.*, **197**, 723–728.

Ukkonen,E. (1985) Finding approximate patterns in strings. *J Algorithms*, **6**, 132–137.

Vingron,M and Argos,P. (1989) A fast and sensitive multiple sequence alignment algorithm. *Comput. Applic. Biosci.*, **5**, 115–121.

Received May 6, 1994, accepted January 10, 1995