**CABIOS**

Vol. 11 no. 4 1995
Pages 423–426

# SAMSON: a software package for the biopolymer primary structure analysis

N.N.Nazipova, S.A.Shabalina[3], A.Yu.Ogurtsov, A.S.Kondrashov[1], M.A.Roytberg, G.V.Buryakov[2] and S.E.Vernoslov

## Abstract

*The SAMSON package is a tool for advanced analysis of primary DNA, RNA and protein structures. The package consists of 16 programs performing statistical analysis and comparison of biopolymer sequences, search for homologies, translation of DNA and RNA sequences into amino acid sequences, splicing of RNA sequences and restriction map construction, recognition of functionally related sites in biopolymer molecules, textual analysis of DNA and RNA regulatory sites and prediction of intermolecular hybridization sites in DNA and RNA molecules.*

## Introduction

Since the first software tools for biopolymer sequence processing appeared in the beginning of the 1970s, nucleic acid sequence data have been accumulated at a constantly growing rate, accompanied by the growth of data processing software quality. Current primary structure software employs more abstract notions and methods of mathematics, physics, chemistry and linguistics in deciphering and interpretation of genetic texts.

The SAMSON package has been developed since 1987 (Vernoslov *et al.*, 1989). During its long lifetime the package has collected a number of programs reflecting a variety of approaches. The SAMSON package contains several traditional programs: mapping of restriction enzyme cleavage sites, open reading frame determination, and splicing and dot-matrix construction. These tools are supplemented with several possibilities for advanced theoretical research, such as recognition of functionally related sites in biopolymer molecules, textual analysis of DNA and RNA regulatory sites and prediction of intermolecular hybridization sites in DNA and RNA molecules. The package includes the FORTRAN source library containing standard input–output and simple sequence processing routines facilitating development of new programs. Below we shall describe our algorithms for DNA, RNA and protein sequences analysis in more detail.

## System

The SAMSON package requires an IBM PC-compatible computer running MS-DOS version 3.0 or higher, 512K RAM, 2.5 Mbytes of hard disk space and supports VGA or EGA graphic display adapters and Epson dot-matrix printers (optional). For development of new programs or modification of the existing ones a Microsoft FORTRAN 5.0 compiler is required.

## Algorithms and implementation

The package creates a user-friendly execution environment. The user interface is based on input and output files. The input file is a text file consisting of partitions, each containing a partition name, a parameter value and a comment line explaining the sense and the definition domain of the parameter. A user may process the existing input files or edit parameter values. Structures of input and output files are similar, so an output file from any program may be used directly as input file for another program.

The package allows one to work both with databases and with sets of individual sequences. SAMSON provides various facilities for sequence set specification, such as a list of database identifiers. Database names should be added as identifier paths when listed sequences are taken from different databases. All the programs of the package accept flat sequence format, which allows the user to process sequences extracted from any database. Databases in EMBL format may be processed without any preliminary extraction.

The SAMSON package contains 16 programs which may be divided in six different blocks: software tools, sequence statistics, sequence comparison and alignment, pattern recognition, simulation of molecular-biological processes and structural analysis of functional sites.

### *Software tools*

The following programs allow the user to obtain

*Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142292 Russia, [1]Section of Ecology and Systematics, Corson Hall, University of Cornel, Ithaca, NY 14853-2701 and [2]Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA*

[3]*To whom correspondence should be addressed. Email. shab@impb.serpukhov.su*

information from databases without exiting from the execution environment.

SEARCH search through databases for documents satisfying certain criteria.
TAKE creation of subdatabases, view of database contents.
TABL preparation of extracted subdatabases, creation of the table of pointers.

*Sequence statistics*

FREQ determination of sequence element frequencies.

Frequencies of individual symbols, pairs and triplets of symbols and textually specified fragments are calculated. Histograms of fragment frequencies along the sequence are plotted.

For any combination of symbols the expected relative frequency is calculated assuming independent symbol appearances in every sequence position.

JUMP DNA and RNA sequence composition analysis.

A profile of the basic element composition index of a biopolymer sequence is displayed. For this purpose the user specifies a sliding window length and a set of basic elements—individual symbols, pairs and triplets of symbols composing the analyzed sequence. For each sequence position the value of the basic component composition index is calculated as:

$$P_k = \frac{1}{2LFR}\sum_{i=1}^{n}(N_i - M_i)$$

where

$k$ = current position number
$LFR$ = the window length
$n$ = basic elements number
$N_i$ = the amount of the $i$th basic element in the window on the left of the $k$th symbol
$M_i$ = the amount of the $i$th element in the window on the right of the $k$th symbol including the $k$th symbol

$P_k$ values not exceeding the user-defined threshold are presented on a histogram.

CONSFND determination of consensus sequences.

The program determines the consensus sequence and calculates the following characteristics of the sequence similarity with the consensus sequence: the number of exact matches; numbers of matches at the purine/pyrimidine level and at the level of pairs A/T, G/C, A/C, T/G; the number of matches of the types 'not A', 'not C', 'not G', 'not T'. Ambiguities are indicated in accordance with official IUPAC-IUB nomenclature.

*Sequence comparison and alignment*

DOTMAT pairwise sequence comparison.

The degree of similarity between two sequences may be determined by specifying either the number of exact matches or the minimum sum value of pairwise alignment scores per fragment of a specified length. Three algorithm modifications which may be run simultaneously are used: repeat search, palindrome search, inverse repeat search. One sequence, called basic, may be compared with any number of sequences during a single program run.

The results of sequence comparison are displayed as pairwise similarity matrices and two histograms. The first histogram presents the distribution of the sum number of dot-matrix matches containing this symbol along the basic sequence. The number of sequences containing at least one match with the corresponding position of the basic sequence is shown on the second graph.

DARWIN quick search for similar sequences in databases.

The program searches for sequences and homologies in databases using an original algorithm and technique similar to that of Sobel and Martinez (1986). In particular, it performs a trivial pairwise comparison. DARWIN finds all exact matches of a certain length for pairs of sequences. To each match, the program ascribes a weight, and successive matches are arranged in chains. The chain weight is a product of weights of the matches and spacers between them. The spacer weight is calculated as a product of its length squared and a coefficient (>1) which takes into account the difference between nucleotide composition of two corresponding fragments forming the spacer. A new match is added to the chain if the weight of the chain is changed by no less than a specified number of times. Unlike some well-known algorithms (e.g. Wilbur and Lipman, 1983), this algorithm should be used for determination of short local homologies. The program is inapplicable in the case of diffused homology, e.g. when even letters of two sequences match and odd letters mismatch.

*Pattern recognition*

SITE determination of fragments similar to assigned sites in sets of sequences.

A 'site' is a short sequence of LS length specified by the user. A symbolic sequence of an arbitrary alphabet may be presented as a string of overlapping words. For example, the sequence

$$S = a_1a_2 \dots a_{LS}a_{LS+1} \dots a_{l-LS+1} \dots a_l$$

of length $l$ one can consider as a superposition of LS-mers $a_1 \dots a_{LS}, a_2 \dots a_{LS+1}, \dots, a_{l-LS+1} \dots a_l$.

The problem of the best fit of the site into the S comes to the following: among the set of LS-mers constituting S, one should pick out those sufficiently similar to the site. Three various approaches are realized in the site definition.

The simplest approach consists in specifying the site as a contiguous string of symbols. A match of an LS-mers with the site is reported when the accumulated number of individual mismatches does not exceed a threshold value.

The second approach allows the possibility of contextual symbol replacements (as, for example, in cases where substitutions of purines to purines are more preferable than those of purines to pyrimidines). A square matrix of real weights is defined for all possible combinations of alphabet symbols. A weight of LS-mer is defined as the sum of match/mismatch weights of its symbols with those occupying corresponding positions in the site. The LS-mer is accepted as similar to the site if its weight does not exceed a desired threshold value.

The third approach considers both the location and the nature of letter ambiguity. Site is set not textually but as a matrix of symbol variants in every its position. The $(i, j)$th matrix element determines the weight of the $i$th term of the sequence alphabet at the $j$th position of the LS-mer. Individual weights ascribed to symbols comprising an LS-mer are summed up over all its positions and the match is reported when the sum does not exceed the threshold value.

The program allows a simultaneous search for a group of sites and their complementary sequences in a set of sequences.

| | |
|---|---|
| REST | construction of restriction site maps. |
| SYMMET | search for sequence regions possessing internal symmetry. |

The program determines direct repeats, inverted repeats and palindromes in DNA and RNA sequences.

*Simulation of molecular-biological processes*

HYBRID localization of potential stable hybrid complexes.

HYBRID searches for sites forming stable hybrid complexes and secondary structures in one or two molecules of single-stranded nucleic acids. The program reveals complementary oligonucleotides or sites capable of forming stable intermolecular hybrid structures. The free energy of an intermolecular hybrid structure is calculated as the sum of energies of all its consecutive nucleotide pairs (Frier *et al.*, 1986). The program constructs chains of possible perfect helices using the modified parameters of Papanicolaou *et al.* (1984).

Intermolecular hybrid structures with total free energy lower than a threshold value and consisting of three or more complementary pairs are registered. The results of the program are presented in dot-matrix form. On the basis of these data the histograms of distribution of complementary site frequencies of one sequence along another are created (Matveeva and Shabalina, 1993). Summarized distribution histograms for groups of related sequences (i.e. mRNA sequences belonging to one functional group, intron sequences, randomly generated sequences) may be created. In this case values of complementary site frequencies determined for pairs of analyzed sequences are summed up.

TRANS translation.

The program performs a search for open reading frames, and decodes nucleotide sequences of amino acid sequences.

SPLICE splicing.

The program constructs a nucleotide sequence from various sequence fragments.

*Structure analysis of functional sites*

CODESER decoding of DNA primary structure into different physico-chemical parameters of the DNA helix.

Symbol sequences are decoded into series of real numbers representing different qualitative and quantitative characteristics of nucleotide sequences. The numbers characterize quantitatively physical or geometrical DNA properties, such as melting energy, energy of transformation of B-DNA to A-DNA, turn angles according to Trifonov (1983), and 'twist' and 'roll' angles (Calladine, 1982; Dickerson, 1983). Ten different ways of sequence decoding are realized.

ACTIVITY calculation of contributions of individual monomeric units to biopolymer function.

The performance of various functional sites on DNA, RNA and protein molecules depends on sequences of their monomeric elements. Biopolymer functional activity (such as the binding constant, the interaction energy, the reaction product output) may be regarded as a complex function of its monomeric elements. Linear functions are customarily used for describing the dependence of biopolymer activity on the contributions of its monomeric elements. It is known that even the most complete data set does not allow one to determine the absolute contributions of monomeric units, while their relative contributions may be determined (Ogurtsov *et al.*, 1993).

The program determines the relative contributions of individual monomeric elements to biopolymer function based on the data on primary structures and functional activities for sets of related biopolymer sequences. It is assumed that individual sequence elements contribute additively to the activity of the whole sequence. The series of the sequences are considered as vectors in multidimensional space.

The relative weights of nucleotides are calculated by minimization of the sum of distances between sequences and the basic one. The functional activities of new nucleotide sequences may be predicted. The program may also be applied in the case of multiplicative contributions of separate sequence positions in the activity of the entire site. In this case the logarithms of the activities will be determined.

## Discussion

The program package described has a number of apparent advantages over existing general-purpose commercial products.

The main features of the SAMSON package are as follows:

- Processing of any sequences is provided, including nucleotide and amino acid or any symbolic sequences from different sources (file, console, or directly from databases).
- Each program can process both the sense and complementary DNA sequences and their fragments.
- The absence of limitations on the length of analyzed sequences, the possibility to analyze sets of sequences in a single program run, temporary storage of accumulated data to preserve them from an unexpected program interruption, rerunning interrupted programs with the automatic data reconstruction.
- Highly developed conception of input and output file compatibility allows interaction between package components. Software tools and application programs have similarly designed transparent interfaces. A reasonable level of program interface standardization is attained despite the complexity of such attempts associated with a diversity of studied problems. We have probably reached the optimal level of integration.

All this makes the package a convenient tool in everyday research work. Adaptation of the package according to the user's needs is possible. The FORTRAN source library containing standard input–output and simple processing routines facilitating development of new programs is included into package.

A number of important theoretical and experimental results have been obtained with the help of the SAMSON package (e.g. Kaliman *et al.*, 1988; Kondrashov *et al.*, 1990; Shabalina *et al.*, 1991; Matveeva and Shabalina, 1993; Ogurtsov *et al.*, 1993).

The entire package with the FORTRAN source library is available free of charge on written request to the authors.

## Acknowledgements

We thank Dr O. Yuryeva for her valuable contribution to formulation of some algorithms, programmer R. Khunafina for aid in package maintenance, Dr N. Spiridonov for helpful discussions and critical comments and O. Ivanova for careful reading of this manuscript. This work was supported by grants from the Russian State Programs 'Frontiers in Genetics', 'Biological Diversity' and from the Russian Foundation of Fundamental Research (grant no. 95-04-12083a).

## References

Calladine,C.R. (1982) Mechanism of sequence-dependent stacking of bases in B-DNA. *J. Mol. Biol.*, **161**, 343–352.

Dickerson,R.E. (1983) Base sequence and helix structure variation in B- and A-DNA. *J. Mol. Biol.*, **166,** 419–441.

Frier,S.M., Kierzek,R., Jaeger,J.A. *et al.* (1986) Improved free-energy parameters for prediction of RNA duplex stability. *Proc. Natl. Acad. Sci. USA*, **83,** 9373–9377.

Kaliman,A.V., Zimin,A.A., Nazipova,N.N., Krayev,A.S., Mironova,M.V., Kryukov,V.M., Skryabin,K.G., Tanyashin,V.I. and Bayev,A.A. (1988) The comparative study of DNA-ligase genes of the bacteriophages T4 and T6. *Dokl. Akad. Nauk SSSR*, **299,** 737–742.

Kondrashov,A.S., Beridze,T.G. and Chiaureli,N B. (1990) Two regions of M13 phage genome hybridizing with human DNA are similar to several keratin genes. *Biochimie*, **72,** 867–871.

Matveeva,O.V and Shabalina,S.A. (1993) Intermolecular mRNA–rRNA hybridization and the distribution of potential interaction regions in murine 18S rRNA. *Nucleic Acids Res.*, **21,** 1007–1011.

Ogurtsov,A.Yu., Elkin,Yu.E. and Shabalina,S.A. (1993) Calculation of contributions of individual monomeric units to biopolymer function. *J.Theor. Biol.*, **161,** 395–401.

Papanicolau,C., Gouy,M. and Ninio,J. (1984) An energy model that predicts the correct folding of tRNA and the 5S RNA molecules. *Nucleic Acids Res.*, **12,** 31–44.

Shabalina,S.A., Yuryeva,O.V. and Kondrashov,A.S. (1991) On the frequencies of nucleotide substitutions in conservative regulatory DNA sequences. *J. Theor. Biol.*, **149,** 43–51.

Sobel,E. and Martinez,H.M. (1986) A multiple sequence alignment program. *Nucleic Acids Res.*, **14,** 363–374.

Trifonov,E.N. (1983) Sequence-dependent variations of B-DNA structure and protein-DNA recognition. *Cold Spring Harbor Symp. Quant. Biol.*, **47**(1), 271–278.

Vernoslov,S.E., Kondrashov,A.S., Roytberg,M.A., Shabalina,S.A., Yuryeva,O.V. and Nazipova,N.N. (1990) The software package 'SAMSON' for the analysis of primary structure of biopolymers. *Mol. Biol. (USSR)*, **24**(2), 524–529.

Wilbur,W.J. and Lipman,D.J. (1983) Rapid similarity searches of nucleic and protein data banks. *Proc. Natl. Acad. Sci. USA*, **80,** 726–730.

*Received on January 24, 1995; revised on April 7, 1995; accepted on April 12, 1995*