Volume 5 Number 3 March 1978 Nucleic Acids Research

Further procedures for sequence analysis by computer

R. Staden

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received 30 January 1978

ABSTRACT

A previous paper[1] described programs for sequence data handling and analysis by computer. The facilities of this basic set are extended by further easily used programs.

INTRODUCTION

In this paper are described programs to

(1) search for palindromes, inverted repeats and hairpin-loops;

(2) search for base-pairing in blocks of complementary bases;

(3) produce a translation of a DNA sequence in all three phases and write it onto a magnetic disk using the one letter code ready for use by other programs;

(4) print out an amino acid sequence in both one and three letter codes;

(5) compare amino acid sequences on the properties of the individual amino acids;

(6) print a DNA sequence with its overlapping genes.

A magnetic tape containing these programs, written in FORTRAN is available on request.

Program Descriptions

HAIRPN

This program searches for palindromes, inverted repeats and potential hairpin-loops in nucleic acid sequences. The operator defines the types of sequences searched for and can select from different ways of sorting the output.

BPFTBK

This program is identical to program BPFIT[1], which searches for regions of sequences that could base-pair, but is improved in that it allows the operator to specify a further parameter. This parameter 'the minimum block length' defines a minimum number of consecutive complementary bases. Only

those bases which are contained in a block of length at least equal to this number will contribute to the score. The regions of possibly base-pairing found in this way are more likely to be stable than those found by BPFIT[1]. If a minimum blocklength of 1 is specified by the operator the results obtained will be the same as those found by BPFIT[1]. This program therefore replaces the latter but at the cost of increasing the running time by 50%.

TRANDK

A program to translate regions of a DNA sequence chosen by the operator into the one letter amino acid codes and write the translation to a disk file. This gives a permanent record of amino acid sequences independent of their DNA sequence. Written in this form the data is ready for input to other programs such as AAFIT. The program can be used for picking out the individual genes from a DNA sequence or to give a complete three phase translation of the whole sequence.

FMT1T3

This is a program for printing copies of amino sequences on the keyboard. Input to the program is a file containing the one letter amino acid codes for the sequence. Output is in both one and three letter codes.

AAFIT

This program is used for comparing amino acid sequences. Operator input is identical to that for program SEQFIT[1] but comparison is done on the properties of the individual amino acids contained in the sequences. For the purposes of the program each amino acid is assigned membership of one of four classes: hydrophobic, basic, acidic or polar (uncharged). When two sequences are compared they are lined up alongside one another in every possible position. For each position the program then looks at pairs of adjacent amino acids and asks if they belong to the same class. If for any position a sufficiently high percentage of the amino acids belong to the same classes their scores are printed out as for SEQFIT[1] with stars marking amino acids belonging to the same class. The classification of the amino acids, both in the number of groups and individual assignments is easily changed.

TRNTRP

A program to print a DNA sequence and its genes in a form suitable for publication. The operator input is identical to that for TRANSQ[1] but the output needs a 132 character line keyboard.

Summary

The programs described in this paper and in an earlier paper[1] form a comprehensive set of computer programs for simple handling and analysis of sequence data. The programs are easy to use and are written in the form of small subroutines which can easily be reassembled to carry out further variations of this kind of analysis.

Reference

1. Staden, R. (1977) Nucleic Acids Research Vol. 4, pp. 4037-4051.