Volume 14 Number 1 1986 Nucleic Acids Research

A collection of programs for nucleic acid and protein analysis, written in FORTRAN 77 for IBM-PC compatible microcomputers

B.Franz Lang[1*] and Gertraud Burger[2]

[1]Laboratoire de Biochimie, IBMC du CNRS, Strasbourg, 15, Rue Rene Descartes, France, and [2]Institut für Genetik und Mikrobiologie, Universität München Maria-Ward-Strasse 1a, D-8000 München 19, FRG

Received 18 July 1985

ABSTRACT:

We have developed a collection of programs for manipulation and analysis of nucleotide and protein sequences. The package was written in Fortran 77 on a Sirius1/Victor microcomputer which can be easily implemented on a large variety of other computers. Some of the programs have already been adapted for use on a Vax 11. Our aim was to develop programs consisting of small, comprehensible and well documented units that have very fast execution times and are comfortably interactive. The package is therefore suitable for individual modifications, even with little understanding of computer languages.

INTRODUCTION:

Most molecular biologist have or should have a personal computer in the lab, not instead of but in addition to an access to the generally less "friendly" multi-user main frame computers. Such microcomputers have the advantage to be able to serve as an always ready-to-use and inexpensive word processor, and are also available to supply the researcher with basic evaluation capabilities in a convenient form, depending on the software. Yet, the increasing power of personal computers is generally not fully exploited in molecular biology.

Although a large collection of programs already exists for manipulating and analyzing nucleic acid and protein sequence data, only a small part of the software is applicable for personal microcomputers. In addition, many of these packages are still too limited for some basic evaluations, and they are mostly very specialized with a tendency to focus on nucleotide analysis

* The compiled programs plus source code are available on diskettes for a small charge. Direct all requests to B.F. Lang (Strasbourg).

neglecting evaluation of derived protein sequences. Program collections are still offered which consist of a number of options that are trivial such as "print a file" or "delete a file", which can be done by the operating system of the computer. Indeed, the operating systems and word processors of modern microcomputers are quite versatile and as this versatility increases continuously it is profitable to know how to use them.

Many programs on microcomputers are very time consuming (e.g. have to be run overnight or even for days). However, fast executing programs are often written with assembly language subroutines and thus can not be easily changed or adapted according to the individual needs of the researcher. Even if the source programs are available, modifying them can be painful. They are often intractable either because of insufficient comments, or because they make heavy use of the particular capabilities of a certain computer, e.g. graphics. In fact, many of the program packages are only executable in a certain hardware environment. A further unpleasant experience is the confusion of different subsets of programming languages.

In order to be able to follow the fast progress in current sequence evaluation it seems indispensable (i) to learn how programs have to be modified and compiled (or to have somebody next door, who knows how to ...) and (ii) to have programs that are written in a way as to allow modifications also for a user with little knowledge of computer languages. When developing the programs described here we have focused not only on fast execution but also on comprehensible structure and the possibility to modify or extend the programs.

From the higher computer languages that are applicable on almost any microcomputer, Fortran 77 and Pascal are easy to learn and result in fast execution of programs. Our programs are exclusively written in Fortran 77 (Microsoft) on a Sirius1/Victor microcomputer which should compile with no or little changes on most other computers having such a compiler. We have started to check some of our programs on the Vax system, which will be also soon available.

MATERIALS AND METHODS:

The used hardware was the IBM-PC compatible Sirius1/Victor microcomputer (256 K memory, necessary for compilations), with two disk drives and a printer. Any other computer of this capacity should be also suitable. However, for an output of the program PSEC a printer is necessary that allows a reduction to 264 characters per line (20cpi).

The necessary software is a Fortran 77 compiler (Microsoft). We strongly recommend to take the best word processor available for the particular facility, which must be able to also produce ASCII character files. This allows writing and modifying source programs as well as data files. We have so far mainly used 'Wordstar' (non-document option) for this purpose.

GENERAL FEATURES OF PROGRAMS

File structures

There is no necessity to have the sequences written in a certain format, except that there should be no more than 130 characters per line, including blanks. When the programs read nucleotide or protein sequences from a data file, blanks and numbers will be ignored as well as unexpected characters (e.g. a '?' in a DNA sequence). Weird characters will be displayed on the screen to check possible bugs in data files.

The format of sequence files created by a program is similar to that of the University Wisconsin package (1): when a program stores sequences in a file (a translated protein sequence, for example) comments are added in the first lines, which describe source and characteristics of the data, and then two '..' indicating the end of comments and the beginning of the data. Correspondingly, reading a data file does not preclude such kind of format although it is recommended for use. If no '..' is found by the program, it will read the data as a "clean" data file without comment lines.

Program interaction

Much effort has been invested into a convenient and self-explaining interactive setting of options and parameters. Questions (in english) are comprehensive without studying the program manuals. Multichoice answers are either in numbers or

characters, valid as capitals and in lower case. Default values suitable for most calculations have been carefully selected and are always indicated. For a fast standard calculation, it is quite convenient to zip through most answers by carriage returns. The logic of answers and their dimensioning are checked routinely to avoid 'bombing' of the program or producing absurd results. For time consuming operations (i.e. more than about 1 minute), a prediction of the execution time is displayed together with the question whether to continue or not. If extended output of a program is expected (especially printing) an estimation of its volume is indicated. It is then possible to sort out a rational number of the best results.

Use of graphics

We have not implemented any kind of graphic analysis, because the usual dot plots can reach either enormous square meter dimensions if sequences of 10.000 or more bases are analyzed; or when reduced in size their informational content is then also reduced (e.g. 2,3). This is especially critical when RNA secondary structures are analyzed and the pairing energies are displayed or when protein sequences are compared using the mutation data matrix of Dayhoff (4). We prefer to leave the job of finding relevant diagonals to the computer and to summarize the results in a list showing the matching stretches and their positions (see Figure 1).

Using the microcomputer for a databank search

The size of the nucleotide sequence collections seems to increase faster than the power of microcomputers. Therefore one should attempt to hook up to a mainframe computer via a modem and let this time consuming job be performed with more sensitive programs and with the most updated versions of data collections.

AVAILABLE PROGRAMS:

FORM

For basic data- and file operations we recommend the text editor(s) supplied with your computer (e.g. 'Word star'), as already outlined above. More advanced formatting can be performed with program FORM which extracts a subsequence, joins, complements or inverts sequences, reformats into various formates

```
.....................................................................................
SEQUENCE 1 = PCOXI2.PEP     ( 323 A.A.), SEQUENCE 2 = CCOBI4.PEP     ( 385 A.A.)
MIN. SUMM. VALUE =  41, MAX. SUM NEG. VALUES =  -5, MAX. LENGTH NEG. VAL.=   3
.....................................................................................


  SUMMATION VALUE =    47,     QUALITY =     1.62,     DIAGONAL =    46

  198 N H W L A G F S D A D A S F
      2 312 6 2 5 2-1 4 1 4 1-3 9
  152 N Q W L A G L I D G D G Y F


  SUMMATION VALUE =    50,     QUALITY =     1.24,     DIAGONAL =   -59

  238 L S L I K D N L G G N I G Y R K S Q D T Y Y
      2 0-1-1 3 3-2 2 5 5 2 5-5 7-1 5 1 1 2 010 7
  297 V E Y Y R E V F G G N I Y F D K A K N G Y F


  SUMMATION VALUE =    41,     QUALITY =     1.18,     DIAGONAL =   -60

  197 N N H W L A G F S D A D A S F Q
      2 2-112 2-1 5 9-3 4 2 4 1 1 1 1
  257 D N A W F M G F F D A D G T I N


  SUMMATION VALUE =    66,     QUALITY =     2.31,     DIAGONAL =   -65

   82 N D S Q F G H Y L A G L I D G D G H F
      1 2-1 1 9 0 3 0 6 2 5 6 5 4 5 4 5 0 9
  147 S N I R F N Q W L A G L I D G D G Y F
```

Figure 1: Output example of program PSCH. Two polypeptides encoded by evolutionary distant intronic reading frames have been compared (the second cox1 intron of S. pombe , and the fourth cob intron of S. cerevisiae, (9,10)). The respective data matrix values are indicated between the two homologous sequence stretches.

of choice, with comment lines at the beginning, followed by sequence data. For the input file no special format is required.

NUC

Makes the fundamental evalutations of nucleotide sequences (RNA- is converted to a DNA-sequence). It checks all characters in a given file, tells you about AT-content and length. It is able to complement and invert sequences, to search strings (e.g. restriction sites, promoter motifs, Z-DNA) , search open reading frames with the genetic code of your choice, to assemble codon usage tables, to translate into polypeptides and to calculate their molecular weight. For further extras, particularly concerning output, programs FORM, TRANS and PUB have been written.

TRANS

This program is a version of the NUC program spezialized on

```
TRANS
TRANSLATE WHICH SEQUENCE                    = POM.SEQ
.....................................................................

THE SEQUENCE POM.SEQ          CONTAINS 19430 NUCLEOTIDES

IT STARTS WITH:   AATGTGTAAT... AND ENDS WITH ...TTATATATGT
IS THAT O.K.  (*YES*)  IF NOT, TYPE N       : <RETURN>
.....................................................................

STORE RESULTS ON DISK ?                     : <RETURN>
PRINT DIRECTLY                              :    P
SCREEN OUTPUT ONLY                          :    S

                                    ANSWER = <RETURN>
.....................................................................

NAME OF OUTPUT FILE                         : POM.PEP
TRANSLATE WITH UNIVERSAL CODE (*YES*)  N?   : <RETURN>
SEARCH THE POSITIONS OF ORFS  (*YES*)  N?   : <RETURN>
MINIMUM LENGTH OF ORFS  (*48*)              = 200
DOES THIS MINIMUM LENGTH COUNT FROM 1st AUG (*YES*)  N ? : <RETURN>
PRINTER ON-LINE ??

SEQUENCE = POM.SEQ          (19430 NUCLEOTIDES)
ORIGINAL STRAND IN ANALYSIS
MINIMUM LENGTH OF ORF WAS SET TO 200 TRIPLETS,
COUNTED FROM 1st AUG
.....................................................................

ORFs ON CURRENTLY ANALYZED STRAND  :

   1st CODON      1st AUG     STOPCODON    TOTAL PROTEIN LENGTH   LENGTH FROM AUG

      4869         4885         6444              531                  519
      6890         6965         7717              275                  250
      7866         7938         8789              307                  283
      8946         8958         9767              273                  269
     10133        10172        10921              262                  249
     11366        11591        13279              637                  562
     14755        14755        15528              257                  257
     18545        18560        19306              253                  248

AND ON ITS COMPLEMENTARY STRAND  :

   1st CODON      1st AUG     STOPCODON    TOTAL PROTEIN LENGTH   LENGTH FROM AUG

NO ORF OF THIS LENGTH PRESENT
.....................................................................

ORIGINAL STRAND IN ANALYSIS

TRANSLATE CURRENT DNA STRAND (*YES*)   N ? : <RETURN>

TRANSLATE FROM (*1*)                        = 10592
TO POSITION  (MAXIMUM IS:15091)             = 11098

                              ETC.
```

Figure 2: Demonstration of a session using program TRANS. The screen and printer output (boxed lines) of part of a translation routine is shown. The interactive responses are also indicated by boxing.

the translation of nucleic acid sequences into proteins. From sequence files up to 40.000 nucleotides it searches for open reading frames on both strands, sorts them according to their starting position, gives you first a clean list of all open reading frames with lengths, start and stop positions and the position of the first initiation codon, then translates the desired frames with any variant of genetic code, calculates the codon usage and the molecular weight of the putative polypeptides and offers a very broad variation of output formats (e.g. 1- and 3-letter code, in capital or lower case letters, with or without corresponding nucleotide sequence, numbering etc.). The output described here (Figure 2) contrasts with the widely used presentation of a sequence translated in all three frames and from both strands at once, which gives stacks of printer paper and is virtually unreadable.

CONS

Detects common nucleotide sequence elements which may serve as important sites for various processes in DNA or RNA metabolism. Consensus sequences are searched in two different sequence files with up to 12.000 nucleotides each. The minimum length of the motif and the percentage of matching positions can be set. If motifs longer than the minimum size are found, only the longest one is shown, in contrast to programs using a 'window' which produce a vast number of redundant matches. The homologous regions can be sorted out either by position or by their length. As this program takes considerable execution times, if two sequences with more than 5.000 bases each are analyzed (see Table 1), a prediction of the execution time is displayed. Also the number of matches is indicated before starting output, and the user can select how much of the best results he wants to see.

REP

Searches direct repeats in a given nucleotide sequence of up to 24.000 bases. It is basically the same program as CONS, but the maximal size is twice as large.

RSEC

Helical regions in a nucleotide sequence are calculated by RSEC, providing a basis to predict RNA folding or DNA signal structures. Different energy values are assumed for G-C, A-U and

Table 1: Brief summary of programs and their execution times. The times indicated do not include input of data and printer output, as both are largely dependent on the facilities. Reading from a diskette is for technical reasons slow, e.g. approximately 10 seconds for a 4.000 bases long sequence (hard disks are much faster).
(1) needs up to a few seconds
(2) main time is needed for interaction

Program	Purpose	Maximal length of manipulated sequences	Specification of test run	Execution time (seconds)
FORM	formatting, joining and output of sequence	40.000 b	-	(1)
NUC	basic nucleotide analysis and translation, string search	40.000 b	complement sequence of 20.000 b search string of 30 b	3 12
TRANS	translation	40.000 b	search open reading frames from both strands (20.000 b), then translate a protein of 800 residues	30 4
CONS	nucleotide homologies	2 files with 12.000 b each	search in two files of 1.000 b each	250
REP	direct repeats	24.000 b	1.000 b	125
RSEC	analysis of helical regions in nucl. sequ.	12.000 b	search in 1.000 b	85
PSCH	protein homology	2 sequences with 3000 residues each	analyze homology between two proteins with 400 a.a	40
PSEC	protein struct. analysis	3000 residues	protein with 300 residues	80
IDENT	aligns protein and nucl. sequ., identity matrix	18 sequences with 900 residues each	-	(2)
PUB	output in publication form	40.000 b	-	(2)

G-U pairs. In this program version no mismatches or loops are allowed. An extended version considering dinucleotide stacking energies (5,6) and unpaired regions, will be soon available. RSEC in its present form is a very fast program that operates with up to 12.000 nucleotides.

PSCH

This program helps to identify the potential functional and evolutionary relationships between different polypeptides. It uses the mutation data matrix (4) as a very sensitive means for detecting local, very low protein homologies (Figure 1). It is fast enough to be suitable for a limited data base search even on a microcomputer, especially after deleting the largely redundant information (e.g. cytochrome c sequences from closely related

```
-164  TCTATGTAAGTTCGAATCTTACCATCCCATTACTTAAACATCTTTAAGATGTTCCTAAACCATCCAAAAAAAAAACACTTTATTAGTGGGATGGTCTCTT

                                                                    M  Q  K  N  N  L  K  N
 -64  TTTTATTTAAGTATTTCCAAAATCAATTAGGAACCAATTAGGGTTTCTTAAATTTCTAATCAAA    ATG CAA AAA AAT AAT TTA AAA AAT

       L  I  T  T  I  V  T  N  A  F  F  N  Q  K  A  N  F  S  M  P  T  K  G  V  I
  25   TTA ATT ACT ACT ATT GTA ACT AAT GCT TTT TTT AAT CAA AAA GCT AAT TTT TCA ATA CCT CTT AAA GGT GTT ATT

       G  E  K  R  P  S  I  L  M  G  N  I  N  M  N  F  K  S  D
 100   GGA GAA AAA AGA CCA TCA ATC TTA ATA GGA AAT ATT AAT ATA AAT TTT AAA AGT GAT TCTTTAATTGAAGTATCTTTCCCT

          pro leu thr asn lys asn tyr pro asn pro ser ile met ser asn ile met gln lys ala thr ser asn
 181   TATTAT  CCT TTA CTT AAT AAA AAT TAT CCA AAT CCT TCA ATT ATA AGT AAT ATT ATA CAA AAA GCT CTT TCA AAT

       his thr leu tyr ser ser lys asn tyr ser phe ile val asn ile arg ala thr pro ile ser thr pro tyr gly
 256   CAT CTT TTA TAT TCA TCT AAA AAT TAC TCT TTT ATT GTT AAT ATT AGA GCT CTT CCT ATT TCA ACT CCT TAT GGA

       ser ser                                MET MET MET GLY SER ASN PRO LYS MET ALA SER THR LEU TRP MET ASP
 331   AGT AGT TTAATCTTTTCAAAATATATAGCA       ATA ATA ATA GGA TCA AAT CCA AAA ATA GCT TCT ACA TTA TGG ATA GAT

       PRO LYS ARG PHE ILE ASN LEU PRO LYS LEU GLN SER ASP SER MET PHE LYS ILE LEU GLY LEU ASN VAL PRO LYS
 409   CCT AAA CGA TTT ATT AAT TTA CCC AAA TTA CAA AGT GAT TCT ATA TTT AAA ATT TTA GGA TTA AAT GTA CCA AAA

       GLY TRP LYS GLY ILE HIS ILE SER LEU ASN LEU ILE LYS TRP ASN SER THR SER SER ARG GLY ARG MET THR ASN
 484   GGA TGG AAA GGT ATT CAT ATC TCA TTA AAT TTA ATT AAA TGA AAT AGT CTA TCA TCA AGA GGA CGA ATA ACT AAT

       MET ILE LYS GLY SER VAL PRO LEU THR ASN ASN SER ASN GLY TYR ASP GLU SER SER LEU ALA ILE TYR SER LYS
 559   ATA ATT AAA GGT AGT GTA CCA TTA ACA AAT AAT TCT AAT GGA TAT GAT GAA AGT TCT TTA GCT ATT TAT TCT AAA

       MET GLY THR ILE GLN ILE LYS VAL ARG LEU SER TYR SER SER ASN THR TER
 634   ATG GGT ACT ATT CAA ATT AAG GTT AGA TTG TCT TAT TCT TCA AAT CTT TAA TTGTTTCCCCCTGCAATCATTTAAAGTTCTTT
```

Figure 3: Output example of program PUB. This sequence has been translated into arbitrary small open reading frames, using different options to demonstrate the capacity of the program. Note also that in this example the genetic code has been changed (CTN = thr, ATA = met and TGA = trp).

species) in the NBRF collection, or by creating ordered subsets of sequences.

PSEC

In order to get information about structural characteristics of a (potential) protein, PSEC evaluates (i) the hydropathy pattern according to Kyte and Doolittle (7) using individual settings of span and average values. This is the only program in our collection that makes use of a pseudo-graphic display on the Sirius1 printer, reducing the output to 264 characters per line. (ii) The distribution of alpha-helical, beta-sheet, coiled regions and beta-turns in the polypeptide (local secondary structure) is calculated according to (8). This subroutine gives you either a listing of data or integrates the results into the hydropathy graph, together with an indication of positive and negative charged residues.

IDENT

For the purpose of evaluating evolutionary distances, IDENT generates and compares alignments of up to 18 protein or nucleotide sequences, respectively. Prealigned sequences are sequentially read in, aligned and the number of matches per

column indicated. The percentage of identity of all sequence combinations in this alignment is also calculated. Mistakes in the alignment can be easily corrected (by using an editor, saving the corrections by a subroutine and subsequent realigning the corrected sequences).

PUB

This program makes a fast and attractive output of any DNA sequence in a form, suitable for publication with no or very little further editorial modifications. Following the recommendations of journals to document sequences as condensed and as readable as possible, the following presentation has been chosen: only one strand of the sequence is shown with optionally 60, 80, 100, 120 or 160 characters per line and an enumeration on the left side only. In translated regions triplets are separated by single spaces, the amino acids are indicated above the nucleotide sequence either in 1- or 3-letter code, and the 3-letter translation optionally in capitals or lower case (Figure 3).

ACKNOWLDGEMENTS:

We thank Drs. R.J. Cedergren (Montreal), F. Michel (Gif-sur-Yvette), I. Schmidt (Munich) and P. Blantz (Tübingen) for discussion and for copies of programs that were helpful during development of the collection presented here. We are also grateful to Dr. J-L. Risler (Gif-sur-Yvette) for providing a revised and simplified Fortran 77 version of the protein secondary structure prediction (8), the basic algorithm of which was integrated into our program package and to C. Poret (Orion company, Illkirch, Strasbourg) for transfering the programs from Sirius/Victor to IBM-PC formatted diskettes. Finally we would like to express our gratitude to the members of the IBMC in Strasbourg for the access to a VAX 11 computer, for the help to use it and Dr. D. de Marcillac for providing the possibility to further develop some of the programs on a Sirius1 microcomputer. At an initial stage, this work was supported by grants from the "Deutsche Forschungsgemeinschaft" to B.F.L. .

*To whom correspondence should be addressed

REFERENCES

1. Devereux, J., Haeberli, P. and Smithies, O.(1984) Nucleic Acids Res. **12**: 387-395
2. Mount, D.W. and Conrad B.(1984) Nucleic Acids Res. **12**: 819-823
3. Zweig, S.E.(1984) Nucleic Acids Res. **12**: 767-776
4. Dayhoff, M.O. (1979) Atlas of Protein Sequence and Structure Vol. 5, NBRF, Wahington D.C.
5. Jacobsen, A.B., Good, L., Simonette, J. and Zuker, M.(1984) Nucleic Acids Res. **12**: 45-52
6. Salser, W.(1977) Cold Spring Harbor Symposium on Quantitative Biology **42**: 985-1002
7. Kyte, J. and Doolittle, R.F.(1982) J. Mol. Biol. **157**: 105-132
8. Garnier,J., Osguthorpe, D.J. and Robson,B.(1978) J. Mol. Biol. **120**: 97-120
9. Lang, B.F. (1984) EMBO J. **3**: 2129-2136
10. Nobrega, F.G. and Tzagoloff,A.(1980) J. Biol. Chem. **255**: 9828-9837