

Los Alamos sequence analysis package for nucleic acids and proteins

Minoru I. Kanehisa

Theoretical Biology and Biophysics Group, University of California, Los Alamos National Laboratory,
Los Alamos, NM 87545, USA

Received 10 August 1981

ABSTRACT

An interactive system for computer analysis of nucleic acid and protein sequences has been developed for the Los Alamos DNA Sequence Database. It provides a convenient way to search or verify various sequence features, e.g., restriction enzyme sites, protein coding frames, and properties of coded proteins. Further, the comprehensive analysis package on a large-scale database can be used for comparative studies on sequence and structural homologies in order to find unnoted information stored in nucleic acid sequences.

INTRODUCTION

For biological sequences, perhaps because there is neither the natural order of numerical data nor the patterns of natural language text, the computer becomes an especially potent aid to interpretation and analysis.¹⁻⁷ Korn, Queen, and Wegman^{1,2} developed a package of computer programs that assist by performing a variety of useful tasks, extending from such routine operations as tallying base composition and translating between nucleotide and amino acid sequences to complicated searches for approximate homologies, including direct and inverted repeats within a single sequence.

The everyday utility of such programs has been enhanced by widespread adoption of time-sharing computer operating systems which effectively execute programs that at any point can send output to a controlling terminal, then request and await further input or instruction, and immediately upon receiving it resume execution taking many-time-per-second turns at occupying the computer's central processor. An investigator at his terminal can now execute a series of tasks, letting each reflect the results of earlier ones, in immediate and continuing response to the natural flow of reflection and question. Clayton et al.³ and Sege et al.⁴ have cast the Korn-Queen program into this "conversational" form.

We shall briefly describe a comprehensive, interactive program package that we have developed. It considerably extends the inventory and power of

options offered, including the new algorithms for homology-symmetry search and secondary structure prediction, which are the subjects of separate papers.^{8,9} It also has a much more extensive capability for analysis of translated protein sequences than any other existing programs of similar purpose. Being programmed in Fortran with all of the calls for operating system action that confer interactivity clearly identified (so that they can be replaced as appropriate to another operating system), it should be highly portable.

THE LOS ALAMOS DNA DATABASE

At Los Alamos National Laboratory a computerized nucleic acid sequence library has been established by Walter Goad and his colleagues. At the moment (July 1981) it contains about 280 published sequences of 370,000 bases. Most of them have been double checked with independently entered sequences in other computer based libraries. Each of our sequence entries is annotated with biologically relevant site information. The analysis package can be called by a general executive program for retrieving, editing, displaying, and, performing various other analytical tasks on the database. A sample run of the analysis package on our database is shown in Fig. 1, where direct and inverted repeats are found around J gene of phage G4,¹⁰ and they are annotated on the sequence together with initiation and termination codons and a restriction enzyme site.

FEATURES OF THE ANALYSIS PACKAGE

The execution of the analysis package is based on a set of commands which specifies routine selection, input/output file definition, and request for helpful information. As shown in Table 1 a command is an alphabetic mnemonic of one or two letters. A command is common to both nucleic acid and protein sequences, although the actual calculation is usually done in two different routines. Depending on the type of the input file one of them is automatically selected. For example, one can search homology of two genes by FH command, translate them to protein sequences and store in a file by TS command, redefine this file as the input file by I command, and find homology of the translated sequences by (the same) FH command. It is this parallel capability for both nucleic acid and protein sequences that makes the analysis package very powerful.

Self-documentation

The analysis package is an interactive self-explanatory program; a user is always prompted specifically what the computer wants, and he can request help messages if the prompt is not clear. To get started use R command which

TABLE 1
Commands of the Analysis Program

<u>Information</u>	
H	Help (summary of commands)
R	Read on-line documents
L	List directory of input file
<u>File Definition</u>	
I	Input file definition (default: Los Alamos DNA Database)
O	Output file definition (default: user's terminal)
<u>Routine Selection</u>	
A	*Annotate sequence (function sites, sequence patterns, G+C rich regions)
P	Print sequence (also composition and molecular weight for proteins)
CN	*Count Nucleotide, dinucleotide, trinucleotide frequencies
CC	*Count Codon usage frequency
T	Translate to amino acids (or reverse translate to nucleotides)
TS	*Translate and Store in a file (for further analysis of protein sequences)
M	Matching subsequence pattern search (eg., restriction enzyme sites)
FH	Find Homology of two sequences
FS	*Find Symmetry (repetition, dyad and mirror symmetries)
FA	Find Alignment of two close sequences and count replacements
D	evolutionary Distance between two sequences
SC	*Search protein Coding regions
SS	Search Secondary structures
E	End of execution

*Not available for protein sequences.

makes access to various on-line documents (we do not provide printed manuals). Each command is named after its action, so it is easy to remember; but it can be reminded by H command. A list of sequence identifiers and their definitions can be re-examined by L command or typing HELP at all input requests for sequence identifiers.

Input/output file manipulation

We are trying to make the program as flexible as possible in handling

Nucleic Acids Research

LOS ALAMOS SEQUENCE ANALYSIS SYSTEM 07/14/81 MON 16:29:02 PRIME TIME
A NEW USER SHOULD USE R COMMAND (TYPE R AFTER NEXT ?) AND READ DOCUMENT #1.

ENTER COMMAND (H FOR COMMAND LIST), OR (CR) TO EXIT ANALYSIS.

? FS
(SYMMETRY)
DO YOU WANT TO CHANGE PARAMETERS? (Y OR N)

? Y
ENTER MATCHM, MDL, DEL, DMAT, DREP, LEN, LOV, OR (CR) FOR DEFAULT.
? 12/

PARAMETERS ARE:

MATCHM - 12 MDL - 1 DEL - 3.0 DMAT - -1.0 DREP - 2.0
SEARCH-LENGTH - 300 OVERLAP-LENGTH - 100

ENTER SEQUENCE IDENTIFIER, OR (CR) TO QUIT.

? PHAGEG4
ENTER START, END OF SEQUENCE, OR (CR) FOR ENTIRE SEQUENCE.

? 2381, 2660
ENTER OPTION CHARACTERS, R(REPEAT), D(DIAD SYMMETRY), AND/OR M(MIRROR SYMMETRY)
? RD

SEQUENCE PHAGEG4 FROM 2381 TO 2660 TOTAL 280

OPTIONS RD
CONFIRM BY (CR)

?

REPEATED REGIONS IN PHAGEG4

*1(14) 9.28E-07
2425 2435
AGGTGTCATGTAAAGA
:::: :: :::::: ::
AGGAGTTATGTATGA
2468 2478

DIAD SYMMETRY (INVERTED REPEATED) REGIONS IN PHAGEG4

*1(18) 1.95E-11	*2(16) 4.85E-06	*3(12) 6.55E-05
2567 2577	2604 2614	2406
GTGGACCCGCGTCCAC	ATGTCT AACGTTCAACAT	CGTGACAGTT ACG
::::::::::	:::: :::::: ::	::: ::: ::: :::
GTGGACCCGCGTCCAC	ATGTTTGAACGTT AACAT	CGT AACTGTTCCAG
-2576 -2566	-2614 -2604	-2411

ENTER SEQUENCE IDENTIFIER, OR (CR) TO QUIT.

?
(END OF SYMMETRY)

ENTER COMMAND (H FOR COMMAND LIST), OR (CR) TO EXIT ANALYSIS.

? A
(ANNOTATION)
ENTER SEQUENCE IDENTIFIER (LOCUS), OR (CR) TO QUIT.

? PHAGEG4
DO YOU WANT TO PRINT OUT IN THE DOUBLE-STRANDED FORM? (Y OR N)
? N

DO YOU WANT TO ANNOTATE FUNCTIONAL SITES? (Y OR N)

? Y
DO YOU WANT TO ANNOTATE SPECIFIC PATTERNS? (Y OR N)

? Y
ENTER PATTERN, #NAME, ALL, OR HELP AFTER ? OR (CR) WHEN ALL DONE.
? #HHA1

?

```

DO YOU WANT TO ANNOTATE SYMMETRY REGIONS? (Y OR N)
? Y
DO YOU WANT THE PROFILE OF LOCAL G+C CONTENT? (Y OR N)
? N
DO YOU WANT TO PRINT OUT ENTIRE SEQUENCE OF 5577 BASES? (Y OR N)
? N
ENTER START,END OF SEQUENCE, OR (CR) FOR ANOTHER LOCUS.
? 2381,2660

INPUT SEQUENCE PHAGE64 FROM 2381 TO 2660 TOTAL 280

      2390      2400      2410      2420      2430      2440      2450
GGTCACGCTGAAACAAACATCCGTGAAACATTTACGCGCTCAGGTTGTCATGTAAAGACCTTTGATTTTAT
                AAAAAAAAAAAAAA
                DYAD
                        AAAA
                        H#RI
                                AAAAAAAAAAAAAAAAAA
                                REPT #1
                                        AAA
                                        D END
                                                AAA
                                                E END

      2460      2470      2480      2490      2500      2510      2520
CGTCTTCACTTTTAAAGGATTTATGTATGAAATTCATTTCCGCCGCTCTGGTGGCAATCTAAGGGTGC
                AAAAAAAAAAAAAAAAAA
                REPT #1
                        AAA
                        J START

      2530      2540      2550      2560      2570      2580      2590
CCGTCTCTGGTATGTAGCCGAAACACAACTACTATCTTTTATGTGGAAACCCCGGTCCTCACTTATTTAG
                        AAA
                        J END
                                AAAAAAAAAAAAAAAAAA
                                DYAD

      2600      2610      2620      2630      2640      2650      2660
GATCAAAATGTCTTACGTTCAACATCTGCGGACCGGTACCTCATGACTTATCTCACCTTGTCTTTG
                AAA
                F START
                AAAAAAAAAAAAAAAAAA
                DYAD

ENTER START,END OF SEQUENCE, OR (CR) FOR ANOTHER LOCUS.
?
ENTER SEQUENCE IDENTIFIER (LOCUS), OR (CR) TO QUIT.
?
(END OF ANNOTATION)

ENTER COMMAND (H FOR COMMAND LIST), OR (CR) TO EXIT ANALYSIS.
?

END OF SEQUENCE ANALYSIS LTSS TIME 8.428 SECONDS

```

Figure 1. A terminal session of the analysis package at the Los Alamos DNA Sequence Database. The question mark in column 1 represents a prompt for user input.

different data formats. In addition to the standard data format of our database, any sequence file may be read in the program if one adds Fortran format specification at the top of the file. A protein file is also recognized by a special key word at the top of the file, which is attached when creating a

protein file by TS command. Therefore, all one needs to do to access a package of protein analysis routines is to redefine a protein file as the input file by I command. Some of the stored protein sequence files, such as structurally resolved globular proteins with secondary structure entries and membrane and other hydrophobic proteins, are also accessible in our database. Regular outputs may be directed to a disk file instead of terminal by O command and then sent to a high speed line-printer.

Annotation

This is designed to visualize and help make comparison of experimentally determined function sites which are stored in our database and various features found in the analysis package, by marking them on the sequence as shown in Fig. 1. In addition to biological function sites, A command can annotate subsequence patterns by invoking pattern search routine (M command), symmetry regions found in the last FS command, and hairpin structure regions found in the last SS command. It can also plot the profile of local G+C content, which is the average G+C content over a fixed sequence length.

Sequence printing and frequency counting

For nucleic acid sequences P command prints the sequence in the single- or double-stranded form, CN command counts mono-, di-, and tri-nucleotide frequencies, and CC command counts codon frequency. CN command also identifies shortest oligonucleotides not present in the sequence, which often shows the deficiency of CG containing patterns in higher eukaryotic sequences. For protein sequences P command prints the sequence, counts amino acid composition and calculates molecular weight.

Translation

Translation regions, which may be spliced, in CC, T, and TS commands are usually specified by start and end numbers of the base position. Each DNA sequence in our database is numbered in the 5' to 3' direction starting with one. There is a special numbering convention in the analysis package. A negative number represents the complementary base position (see Fig. 1). Thus, the increasing negative number means the 5' to 3' direction on the complementary strand. (A pair of decreasing numbers is usually taken to imply the opposite 3' to 5' direction except when ambiguity arises in circular sequences.) If the end number is zero translation stops at the first termination codon. Genes of the known complete genomes, such as phages and animal viruses, can be specified by symbolic identifiers; for example, LT specifies the spliced large tumor antigen gene in animal viruses. T command for protein sequences reverse translates to nucleotides.

Subsequence pattern search

M command finds given oligonucleotide patterns or translated oligopeptide patterns in a nucleic acid sequence. An oligonucleotide pattern may contain special characters representing multiple nucleotides. Alternatively, restriction enzyme names can be given to find corresponding oligonucleotide patterns (see Fig. 1). For protein sequences M command finds oligopeptide patterns allowing a given number of mismatches.

Homology-symmetry search

This employs a powerful new algorithm⁸ that finds homologous subsequences of two sequences. The method, which is an extension of the Sellers method,¹¹ examines all possible alignments of two sequences including any number of gaps by the order of $N \times M$ operations where N and M are the sequence lengths being compared. The significance of each local homology found is checked against random sequences of the same composition according to a simple probability calculation formula.⁸ Homology is based on assignment of weights for nucleotide (or amino acid) matches, replacements, and deletions.¹² We use Dayhoff's amino acid mutation data¹³ in related protein families for a set of weights (metric) in protein alignment. It is also possible to define a different metric based on, for example, chemical or structural similarity of amino acids, or minimum necessary base changes in genetic codes. For nucleotides we have a tentative metric which gives alignments of nucleic acid sequences similar to alignments of the translated protein sequences with Dayhoff's data. Although the method is flexible enough to accommodate different weights for transition and transversion, or the same weights for a block of nucleotide deletions, more accumulation of data is required to establish a suitable metric. The basic FH command finds local homologies of two sequences, but it can be applied to the same sequence. FS command is more convenient to search internal homologies within one sequence. It finds direct repeats (homologies on the same strand), inverted repeats or dyad symmetries (homologies on the complementary strands), and mirror symmetries (homologies on the same strand in opposite directions).

Evolutionary distance

The best alignment (minimum distance) of two entire sequences is calculated by the Needleman-Wunsch algorithm¹² according to the given metric. In D command the significance of the distance is represented by the standard deviation unit of the distances of a given number of random sequences of the same composition.¹³ The actual alignment of two entire sequences can be displayed by FA command, which also summarizes the numbers of all observed matches,

replacements, and deletions. This command can be applied only to close sequences because it searches in a limited band near the diagonal element of the Needleman-Wunsch matrix.

Protein coding region

At the moment FC command just tabulates open reading frames. We plan to improve the routine by incorporating other features: for example, statistical distinction between coding and noncoding regions by the codon usage, search for possible ribosome binding sites, and analysis of the properties of coded protein molecules.

RNA secondary structure

This is also based on a new efficient algorithm⁹ that finds all possible hairpin structures, each of which may contain bulges and internal loops, by the order of $N \times N$ operations where N is the sequence length. The method is similar to the homology search method,⁸ but here the metric for closeness is a set of free energy values for base pairing (including base stacking) and for various loops with different lengths and base pairs closing the loop. The free energy calculation is based on the modified Tinoco rule¹⁴ with the free energy data compiled by Salser.¹⁵ SS command is designed to catalogue all candidates for secondary structures with free energies below a threshold set by the user, which may then be screened by other experimental and theoretical considerations. Thus, the assembly of overall structure is left to the user's discretion. Another way to search possible secondary structures in the analysis package is to use FS command and look for dyad symmetries, which of course do not include free energy values.

Protein secondary structure

Although there may be better methods we use the Chou-Fasman method¹⁶ to predict helices and beta-segments because it is the simplest one to program. In addition, SS command for protein sequences displays the hydrophobicity profile¹⁷ which reveals the clustering of hydrophobic amino acids, and the periodicity of 3.6 residues in the hydrophobicity profile which locates the candidates for amphipathic helices.¹⁸ SS command also annotates known secondary structures if the datafile contains such information.

DISCUSSION

One of the most challenging problems in developing pattern analysis of nucleic acid sequences is how to extract regulatory information in gene expression. Often conspicuous subsequence patterns are observed, for example, around promoter sites and splicing sites, but they are apparently not suffi-

cient information for recognition by other functional molecules. An extensive analysis package on a large-scale database may help find less conspicuous, but equally important, information stored in nucleic acid sequences. In the present analysis package this can be done by applying match, find, and search commands to related sequences or related regions which are known to exercise similar biological functions.

An example of searching symmetry regions within a sequence by FS command is shown in Fig. 1. For homology of two sequences we have analyzed by FH command the 5' end of gene A in phages ϕx^{19} and G4¹⁰ from the initiation codon ATG to the Thr codon ACT, one codon before the internal restart codon for gene A'. There are 172 codons (516 nucleotides) in ϕx and 213 codons (639 nucleotides) in G4. The numbering starts in both sequences at the initiation codon, therefore, nucleotides 286-288 correspond to codon 96. The upper part of Fig. 2 shows the locally homologous regions in the nucleotide sequences, when a set of weights is assigned to nucleotide matches, mismatches, and deletions.⁸ The longest alignment #1 contains 57 base matches and the expectation value of finding this particular type of alignment in random sequences is $2.88 \times 10^{-24} \times 516 \times 639 = 9.5 \times 10^{-19}$. The lower part shows the locally homologous regions in the translated amino acid sequences based on Dayhoff's mutation data.¹³ It is apparent that the amino acid sequences give longer alignments. Amino acid sequences are usually better conserved because the genetic codes are redundant, i.e., a change in the third position of a codon does not often change the amino acid, and also because the codon changes that do not alter drastically the properties of amino acids are presumably favored in evolution.

Even if no sequence homologies can be observed, structural correspondence may still be used to search common regulatory information. Figure 3 shows output examples of SS commands, possible secondary structures of potato spindle tuber viroid (PSTV)²⁰ and yeast phenylalanine tRNA,²¹ where a hairpin loop is represented at the right end of each alignment. PSTV is a covalently closed RNA of 359 nucleotides, but we have analyzed it as a linear sequence from base 1 to base 359 according to the numbering convention used by the original authors.²⁰ The best single hairpin loop structure, which is very similar to the one reported by them, contains 125 base pairs with the free energy -224 kcal/mol. On the other hand, the output for tRNA needs further considerations. Note that the three hairpin loops in the cloverleaf are included in the structures #1, #3, and #4, and the acceptor stem in #1. The hairpin loops in #2 and #4 are not topologically compatible; either one of

PARAMETERS ARE:

MAXIMUM FREE ENERGY FOR PRINT OUT -200.0
 MAXIMUM LENGTHS OF HAIRPIN, INTERNAL AND BULGE LOOPS 20 10 5
 SEARCH LENGTH 0

LOCALLY STABLE SECONDARY STRUCTURES IN PSTV

```
*1(125) -224.
  5          15          25          35          45          55          65
GGACUAAACUCGUGGUCCUGUGGU CACACCUGACCUCUGA GC AGAAAAGAAAAGAG GCGG C
::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: :::
CCUUGGUU GA CGCCAGGUUCCGAAUUGUG GGAGCGG GG CUUCGUUCA UUCUA UC UCUUUUCGCCAG
 355      345      335      325      315      305      295
  73      83      93      103      113      123      133
UCGGAGGAGCGCUUCAGGGA UCC CCGGGGAA CCUGGAGCGARCUG GCAAAAAGGACGGUG GGGAGUG CC
::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: ::: :::
AGCC CUCG AAGUCAACAAAGGUGGCCCAUCAUCGG CUUCGC UGUCGCGUUUC CC CCGUCCC CACCAGG
 285      275      265      255      245      235      225
 144      154      164      174
CAGCG GC CGAC AGGAGUAAUCCGCCGAAC AGGGUUU
::: ::: ::: ::: ::: ::: ::: ::: ::: :::
A CGCCCGCGCUCCUCCUGU GGGCUUCUUCUCCCAU
 216      206      196      186
```

PARAMETERS ARE:

MAXIMUM FREE ENERGY FOR PRINT OUT -5.0
 MAXIMUM LENGTHS OF HAIRPIN, INTERNAL AND BULGE LOOPS 20 10 5
 SEARCH LENGTH 0

LOCALLY STABLE SECONDARY STRUCTURES IN TRNAPHE

```
*1(21) -21.8
  4          14          24          34
GCGAAUUAGCUCAGUUG GGA GAGCG CCAAGCUG
::: ::: ::: ::: ::: ::: ::: :::
CGCUAAG ACACCUAGCUUGUGUCCUGGAGGUCUAGA
  69      59      49      39

*2(17) -16.7
  6          16
GGA UUUAG CU CAGU UGGGAG
::: ::: ::: ::: ::: ::: A
CCUGGAGGUCUAGAGUCAGACCCGG
  46      36      26

*3(8) -7.23
  48
GGUC CUGUGUUC
::: ::: ::: ::: G
CCACGCUAAGACACCUA
  72      62

*4(6) -6.20
  6          16
GGAUUUAGCUCAGUU
::: :::
CCG CGAGAGGG
  25
```

Figure 3. Locally stable secondary structures in potato spindle tuber viroid and yeast phenylalanine tRNA.

them may exist in the assembled overall structure.

Figure 4 is the result of using SS command for the amino acid sequence of bacteriorhodopsin.²² Clustering of hydrophobic amino acids and periodic appearance of hydrophobic amino acids are consistent with the seven trans-

namely, it is based on the structural data of water-soluble globular proteins, the examination of hydrophobicity profiles may be more widely applicable to different types of proteins.

CONCLUDING REMARKS

The analysis package consists of over 5000 Fortran statements in nine programs: one highly interactive controller program and eight non-interactive programs for large calculations. Although some of the features we use, such as the communication between two programs and the dynamic expansion of core size, are beyond the capability of regular Fortran, the entire package may be transportable to other systems (we use CDC 7600) if appropriate routines are written in different languages. Alternatively, a separate program may be run as a stand-alone program. The complete source codes are available upon request.

ACKNOWLEDGMENTS

I am grateful to Dr. Walter Goad for help and encouragement. This work was performed under the auspices of the U.S. Department of Energy.

REFERENCES

1. Korn, L.J., Queen, C.L. and Wegman, M.N. (1977) Proc. Nat. Acad. Sci. USA 74, 4401-4405.
2. Queen, C.L. and Korn, L.J. (1980) Methods Enzym. 65, 595-609.
3. Clayton, J., Friedland, P., Kedes, L. and Brutlag, D. (1981) MOLGEN Report, Stanford University.
4. Sege, R., Söll, D., Ruddle, F. and Queen, C. (1981) Nucl. Acids Res. 9, 437-444.
5. Staden, R. (1977) Nucl. Acids Res. 4, 4037-4051.
6. Staden, R. (1978) Nucl. Acids Res. 5, 1013-1015.
7. Gingeras, T.R. and Roberts, R.J. (1980) Science 209, 1322-1328.
8. Goad, W.B. and Kanehisa, M.I., submitted for publication.
9. Kanehisa, M.I. and Goad, W.B., submitted for publication.
10. Godson, G.N., Barrell, B.G., Staden, R. and Fiddes, J.C. (1978) Nature 276, 236-247.
11. Sellers, P.H. (1980) J. Algorithms 1, 359-373.
12. Needleman, S.B. and Wunsch, C.D. (1970) J. Mol. Biol. 48, 443-453.
13. Dayhoff, M.O. (1978) Atlas of Protein Sequence and Structure, Volume 5, National Biomedical Research Foundation, Washington, D.C.
14. Tinoco, I., Jr., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) Nature New Biol. 246, 40-41.
15. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. 62, 985-1002.
16. Chou, P.Y. and Fasman, G.D. (1978) Advan. Enzymol. 47, 45-148.
17. Rose, G.D. (1978) Nature 272, 586-590.
18. Kanehisa, M.I. and Tsong, T.Y. (1980) Biopolymers 19, 1617-1628.
19. Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison, C.A., III, Slocombe, P.M. and Smith, M. (1978) J. Mol Biol. 125, 225-246.

20. Gross, H.J., Domdey, H., Lossow, C., Jank, P., Raba, M., Alberty, H. and Sanger, H.L. (1978) *Nature* 273, 203-208.
21. RajBhandary, U.L. and Chang, S.H. (1968) *J. Biol. Chem.* 243, 598-608.
22. Ovchinnikov, Yu.A., Abdulaev, N.G., Feigina, M.Yu., Kiselev, A.V. and Lobanov, N.A. (1979) *FEBS Lett.* 100, 219-224.