Volume 14 Number 1 1986 Nucleic Acids Research

Nucleic acid secondary structure prediction and display

Kurt Stüber

Max-Planck-Institut für Züchtungsforschung, Egelspfad, D-5000 Köln 30, FRG

Received 2 July 1985

ABSTRACT

A set of programs has been developed for the prediction and display of nucleic acid secondary structures. Information from experimental data can be used to restrict or enforce secondary structural elements. The predictions can be displayed either on normal line printers or on graphic devices like plotters or graphic terminals.

METHODS AND IMPLEMENTATION

Development and testing of the programs was done on a VAX 11/750 from Digital Equipment Corporation. All programs are written in the language PASCAL. The PASCAL version used was VAX-11 PASCAL V1.2. Any nonstandard PASCAL statements were avoided to insure the transportability of the programs.

INTRODUCTION

With increasing amounts of nucleic acid sequence data available the need for rapid and exact prediction of nucleic acid secondary structures increases very quickly. Many algorithms for such predictions have been proposed and implemented (6,11-15). The programs shown here combine a powerful prediction scheme with simple and direct display routines. The analysis of secondary structures should not be restricted to those users who have plotting devices at their command. Therefore the display of structures is compatible

The programs are distributed as a part of a Sequence Analysis Software Package, described in reference 1. The package is available for $150 from the Max-Planck-Gesellschaft through Garching Instruments, Königinstrasse 19, D-8000 Munich 22, West Germany.

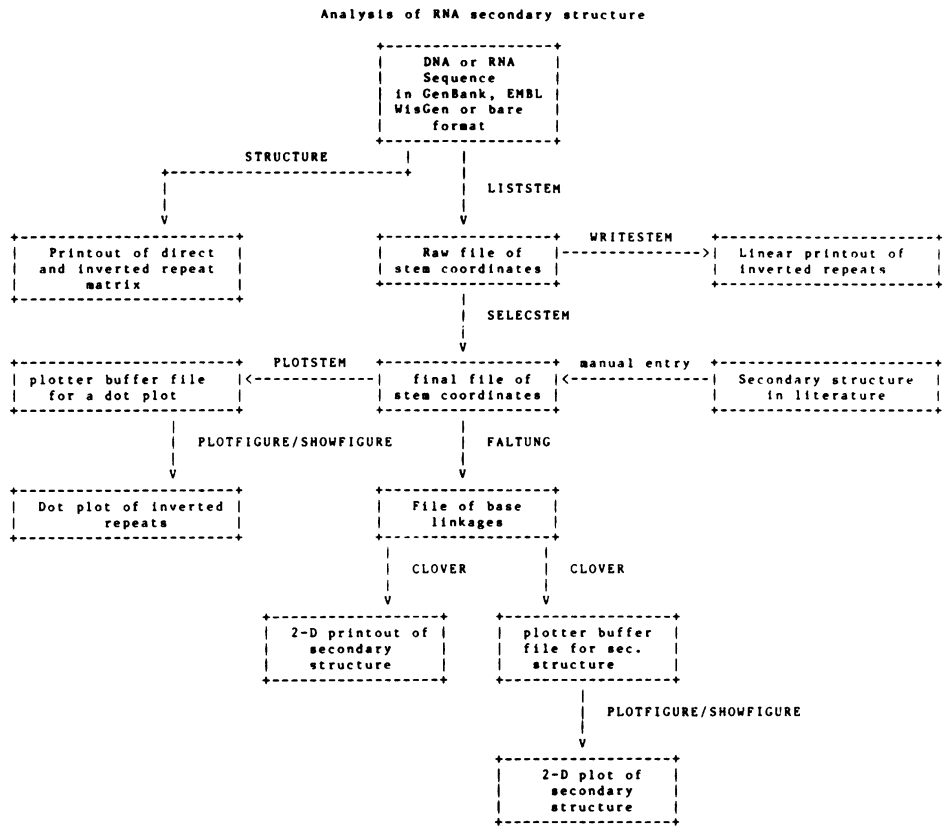


Figure 1: Schematic diagram of the flow of data during secondary structural analysis by the programs STRUCTURE, LISTSTEM, STRUCTURE, WRITESTEM, SELECSTEM, PLOTSTEM, FALTUNG, CLOVER, PLOTFIGURE, AND SHOWFIGURE. Rectangles represent data files and arrows show the action of the different programs. Only for the entry of secondary structure information from the literature manual entry is needed, all other steps are aided by programs. The final plots can either be seen on the sceen (program SHOWFIGURE) or be plotted by a mechanical plotter (program PLOTFIGURE).

with even simple alphanumeric printers. For more sophisticated displays a straight forward representation was chosen that can be visualized by almost all plotters or graphical terminals.

The prediction of a secondary structure is done in five independant steps. Figure 1 shows the chematic flow of data

through the various steps. Every rectangle represents a data file and the program names beside the arrows show which program is used to transform the data from one stage to the next.

First all possible inverted repeats are calculated and inspected with the programs STRUCTURE, LISTSTEM, WRITESTEM or PLOTSTEM. STRUCTURE shows all possible inverted and direct repeats of a sequence using a format which is convenient for direct visual inspection. PLOTSTEM may similarly be used for long sequences and produces only graphical output. All inverted repeats are extracted from a nucleic acid sequence and written to a file by the program LISTSTEM. This file contains every possible stem that can be formed.

In a second step the program SELECSTEM allows the selection of stems from this starting list. Here the set of stems is restricted so that only structures that agree with experimental data are produced. It is possible to exclude stems that form within regions known to be single stranded. In addition stems can be enforced by the exclusion of alternatives. WRITESTEM sorts the final list of stems and displays it for inspection and correction.

The third step is the actual prediction of a secondary structure using the program FALTUNG. The final list of stems and the nucleic acid sequence are used for calculating a secondary structure that is thermodynamically optimized using energy parameters published by Salser (9). The search for an optimal structure is done by an algorithm that is similar to the one presented by Martinez (3).

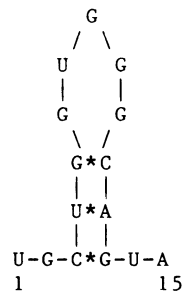
The fourth step is the production of a two dimensional layout of the predicted structure by the program CLOVER so that it can either be printed by a normal line printer or by a graphic output device.

Finally, in a fifth step the graphic output of CLOVER may either be displayed on a video screen (using the program SHOWFIGURE) or on a mechanical plotter (using PLOTFIGURE).

ALGORITHMS

LISTSTEM: The program LISTSTEM calculates a list of all the inverted repeats of a given sequence and encodes them with

three numbers, the 5 prime base of the stem, the 3 prime base of the stem and the length of the stem. These numbers are output to a file named by the user. The RNA sequence UGCUGGUGGGCAGUA could form a stem like this:



This stem would then be encoded by the numbers 3, 13 and 3 representing the G in position 3 , the C in position 13 and the 3 base pairs of the stem.

WRITESTEM : WRITESTEM sorts the stems that have been calculated by LISTSTEM and writes the individual bases for each stem in a linear layout. The stems can be sorted by stem length or by thermodynamical stability. Manually entered stems can also be processed if they are in the same format as the output of LISTSTEM. After the sorting a limited number of stems can be written to an output file, which will contain the individual bases and the coordinates of the stem together on a single line. The stem is shown with a variable number of bases before and after it. The user can choose between different sorting schemes and determine the number of stems shown and the number of bases surrounding each stem. If the input file has the stem coordinates: 9 19 3, the format of the output file is:

9 ..GGGTGGGA CCC.. 5bp .. GGG GTCCTGCTCA 19

9 is the position of the first base of the stem (5 prime end) and 19 is the last position of the stem (3 prime end) and 3 is the number of bases in the stem. The user must specify the names of the input files.

PLOTSTEM : PLOTSTEM gives a two dimensional display of the stem positions. The input for PLOTSTEM is a LISTSTEM output

file. The output of PLOTSTEM is a plotter buffer file that may either be displayed on a graphic terminal with SHOWFIGURE or on a HP7475 plotter (Hewlett-Packard) using the program PLOTFIGURE. Although the program STRUCTURE will show direct as well as inverted repeats, PLOTSTEM gives only the positions of inverted repeats. PLOTSTEM produces a plotter buffer file using a stem coordinate file (see figure 1). The stems are shown in a dot plot fashion in a rectangular frame. Only one side (above the diagonal) is shown. The region of display can be restricted so that only the stems from part of the sequence are seen.

SELECSTEM : The raw list of stem coordinates can be sorted and refined with SELECSTEM. In many instances some additional information about the secondary structure of a certain RNA molecule is known. Single strand or double strand specific enzymes often cut at specific sites. In some cases other homologous sequences are known where mutations suggest certain base pairings. Even by inspection with the electron microscope some conspicious secondary structural elements may be seen. All these experimental data can be taken into account by selecting the raw list of stem coordinates. Single stranded regions can be enforced by the exclusion of stems formed from them. If a stem is known or thought to exist all other alternative stems from the same region may be taken out of the list. If the user is only interested in local secondary structure he may exclude stems that have hairpin loops greater than a limit size. To see non-local structures a minimal hairpin loop size may be required. SELECSTEM reads the stem coordinate file as produced by LISTSTEM and prompts the user for the experimental boundary conditions. Then the list of coordinates is investigated and all those stems not in aggreement with the experimental conditions are excluded. This refined list is then used by the program FALTUNG for the prediction of an energetically minimized secondary structure. FALTUNG does not calculate stem coordinates by itself but is restricted to the coordinate list as provided by SELECSTEM. If stems have been excluded from the list they will not appear in the final prediction.

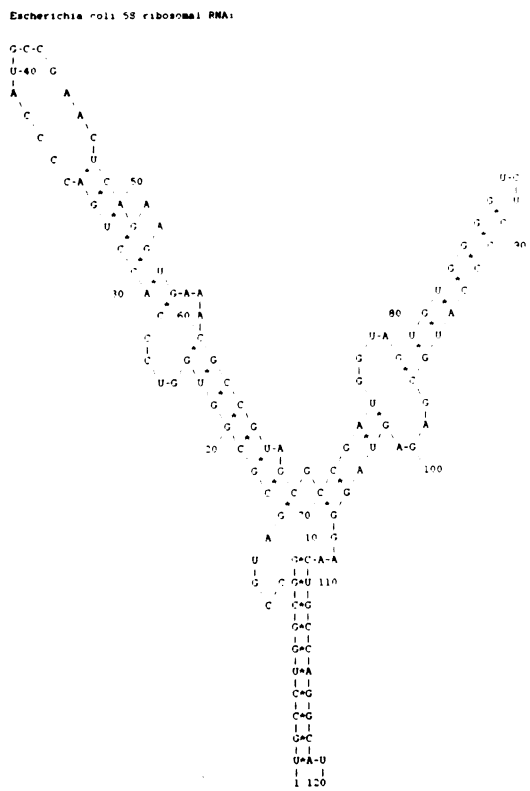


Figure 2: Secondary structure for 5S ribosomal RNA from Escherichia coli as predicted by Noller (5). The layout of the structure has been calculated by the programs FALTUNG and CLOVER. The free energy of this structure is -51.1 kcal according to Salser (9).

SHOWFIGURE : The program SHOWFIGURE allows the display of plotter buffer files on a video screen (for example VT240 and VT241 from Digital Equipment Corporation). It will take as input any plotter buffer file containing per line 3 or 4 integer numbers signifying x- and y-coordinates, a penstatus (1 for down, 0 for up) and a pen color (1 or 4 for red, 2 and 5 for blue and 3 and 6 for green). No other data are allowed in a plotter buffer file. The figure is not distorted and if one dimension is shorter than the height or width of the screen then the figure is shifted to the middle. The user can give a reduction factor. For example a factor of 2.0 will

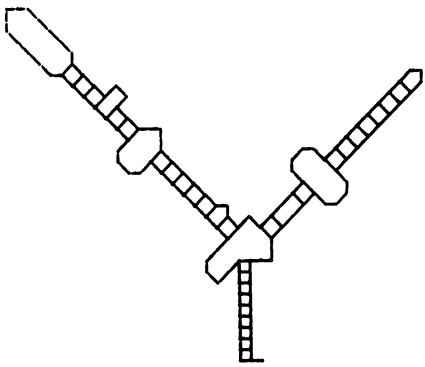


Figure 3: The same secondary structure as shown in figure 2. (5S ribosomal RNA from Escherichia coli (5) formatted by CLOVER). This is a schematic plot done by the HP7475 plotter using the output file of CLOVER. The program used for the plotting is PLOTFIGURE.

reduce the size of the figure to 1/2 of the full screen size.

PLOTFIGURE : The program PLOTFIGURE allows the plotting of plotter buffer files on the HP7475 plotter. It will take as input any file containing per line 3 or 4 integer numbers signifying the x- and y-coordinates, a penstatus (1 for down, 0 fur up) and a pen color (1 for black 0.7 mm, 2 for black 0.3mm, 3 for red 0.3mm, 4 for green 0.3mm, 5 for blue 0.3mm and 6 for violet 0.3mm). No other data are allowed in a plotter buffer file. The data are shifted and reduced as by the program SHOWFIGURE.

STRUCTURE, LISTSTEM, FALTUNG, CLOVER : The algorithms for the programs STRUCTURE, LISTSTEM, FALTUNG and CLOVER have been described previously (1). Since then the sequence data input has been improved considerably. The programs now automatically identify the input file format. Currently four different formats are recoqnized: EMBL Nucleotide Sequence Data Library entry format (2), Genbank Nucleotide Sequence Data Library format (2), Wisconsin Genetics Computer Group data file format (7) and a format free of comments.

DISCUSSION

The calculation of the best (minimal free energy) secondary structure of a nucleic acid sequence has been solved by Zuker and Stiegler (6) but these structures still do not represent those found in vivo. This may be accounted for in part by missing or inexact parameters for the calculation of the free energy (4,9) but also by the fact that only the secondary

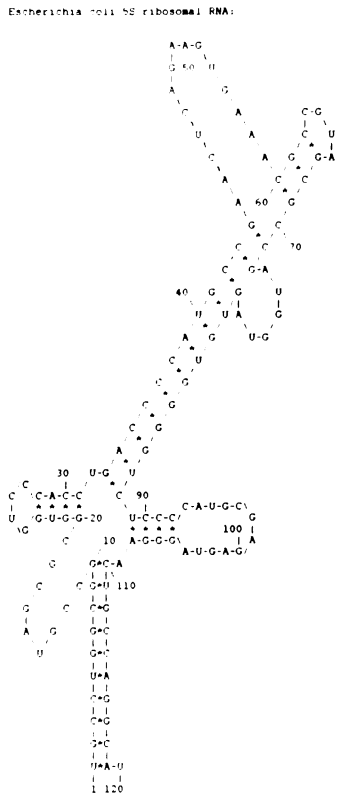


Figure 4: Secondary structure predicted by the program FALTUNG for the 5S ribosomal RNA from Escherichia coli (sequence taken from ((5)) and formatted by CLOVER. The free energy of this structure is -53.6 kcal according to Salser (9).

interactions of a molecule are predicted. In solution the molecules surely will also have tertiary and quarternary interactions (loop to loop and intermolecular contacts) and are often found in physical contact with proteins or other biological molecules that change their secondary structures.

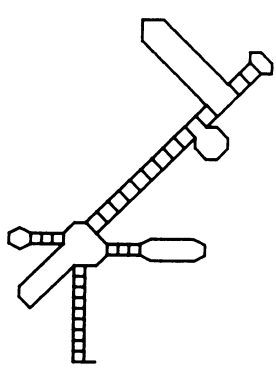


Figure 5: Secondary structure of the 5S ribosomal RNA from Escherichia coli (sequence taken from (5)). Same structure as shown in figure 4 but alternatively plotted with the HP7475 plotter.

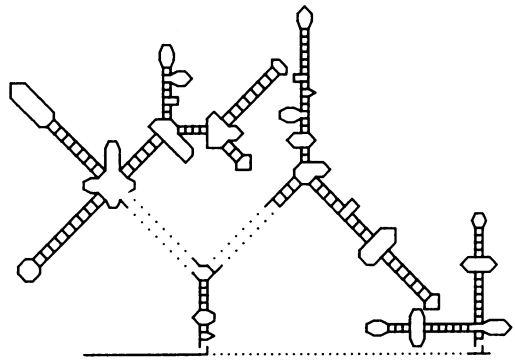


Figure 6: Secondary structure of the intron from Tetrahymena thermophila (sequence taken from Cech et al. (10)) as predicted and drawn by FALTUNG and CLOVER. The free energy of this structure is -116.13 kcal according to Salser (9).

Despite these limitations the predicted structures show many of the features that have also at least partially been verified experimentally (5). Instead of using large amounts of computer time for the calculation of the very best structures, the programs presented here allow the fast calculation of approximated secondary structures. As an example of the use of the programs presented here the 5S ribosomal RNA from Escherichia coli was investigated. The prediction based on experimental evidence shows a free energy of -51.1 kcal calculated according to Salser (9) (see figures 2 and 3) and the structure predicted by the programs (figures 4 and 5) has a free energy of -53.6 kcal. The predicted structure is therefore theoretically slightly more stable, but only the major stem (coordinates: 1, 119, 10) is identical with the structure proposed by Noller (5). With the use of the program SELECSTEM the list of stem coordinates was restricted to be in aggreement with experimental data. The programs then also predicted the structure shown in figures 2 and 3.

In addition the programs were used to predict the structure of the self splicing Tetrahymena thermophila pre-ribosomal RNA (10). The result is shown in figure 6. The free energy of this structure (-116.13 kcal) calculated according to Salser

(9) is similar to the free energy of the structure as predicted by Cech et al. (10) (-115.7 kcal) using the program of Zuker and Stiegler.

REFERENCES

1. Stüber K. (1985) Compl. Appl. Biol. Sci. 1, 35-42.
2. EMBL Nucleotide Sequence Data Library, EMBL, Pf. 10.2209, D-6900 Heidelberg, West Germany and GenBank (TM) Los Alamos National Laboratory, T-10, Mail Stop 465, Los Alamos, NM 87545 U.S.A.
3. Martinez H.M. (1984) Nucl. Acids Res. 12, 323-334.
4. Tinoco I., Borer P., Dengler B., Levine M.D., Uhlenbeck O.C., Cech T.R., (1973) Nature New Biology 246, 40-41.
5. Noller H.F. (1984) Ann. Rev. Biochem. 53, 119-162.
6. Zuker M., Stiegler P. (1981) Nucl. Acids Res. 9, 133-148.
7. Devereux J., Haeberli P., Smithies O. (1984) Nucl. Acids Res. 12, 387-395.
8. Feldmann R.J. (1981) NASSD - Nucleic Acid Structure Synthesis and Display, Version 1.5 (Manual for programs NUCSHO and NUCGEN).
9. Salser W. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 985-1002.
10. Cech T.R., Tanner N.K., Tinoco I., Weir B.R., Zuker M., Perlman P.S. (1983) Proc. Natl. Acad. Sci. USA 80, 3903-3907.
11. Nussinov R., Pieczenik G., Griggs J.R., Kleitman D.J. (1978) SIAM J. Appl. Math. 35, 68-82.
12. Studnica G.M., Rahn G.M., Cumming I.W., Saiser W.A. (1978) Nucl. Acids Res. 5, 3365-3387.
13. Waterman M.S., Smith T.F. (1978) Mathematical Biosciences 42, 257-266
14. Rogers J., Clarke P., Salser W. (1979) Nucl. Acids Res. 6, 3305-3321.
15. Yamamoto K., Kitamura Y., Yoshikura H. (1984) Nucl. Acids Res. 12, 335-346.