Computer-aided nucleic acid secondary structure modeling incorporating enzymatic digestion data

Gary J.Quigley*, Lee Gehrke+§, David A.Roth+ and Philip E.Auron*+‡

*Biology Department, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 15 August 1983

ABSTRACT

We present a computer-aided method for determining nucleic acid secondary structure. The method utilizes a program which has the capability to filter matrix diagonal data on the basis of diagonal length, stabilization energy, and chemical and enzymatic data. The program also allows the user to assign selected regions of the structure as uniquely single-stranded or paired, and to filter out "trade-off" structures on the basis of such pairing.

In order to demonstrate the utility of the program we present a preliminary secondary structure for the 3' end of alfalfa mosaic virus RNA 4 (AMV-4 RNA). This structure is based on an analysis which includes the use of in vitro partial enzymatic digestion of the RNA.

INTRODUCTION

It is a well established fact that "single-stranded" nucleic acid molecules possess the capability of condensing into compact structures by virtue of intrastrand secondary structures. This type of interaction is exemplified by the structure of transfer RNA [1], which has been determined by X-ray crystallography, and by the structures of bacteriophage MS2 RNA [2], 16S and 23S ribosomal RNA [3, 4] and β -globin messenger RNA [5], which have been structurally analyzed by less direct methods such as enzymatic digestion, chemical reactivity, and phylogenetic conservation. Recent evidence suggesting a role for secondary mRNA structure in translational control [5, 6, 7, 8] has stimulated interest in secondary structure determination for mRNA molecules. In addition, recent models proposing the mechanism for intervening sequence processing [9, 10] have generated interest in the investigation of intermolecular as well as intramolecular secondary structure. Unfortunately, no satisfactory method exists for predicting such structures. However, a bevy of techniques exists which can be used to analyze structure potential. Thermodynamic calculations [11, 12, 13] can be used to model minimum free energy configurations based on model oligonucleotide studies, and computer programs which implement these techniques have been reported [14, 15]. These approaches generally assume that only one "lowest energy" structure exists,

making no accommodation for "trade-off" structures, molecular "breathing" or local, as opposed to global, free energy minima. A recent approach to the problem of nucleic acid secondary structure prediction involves the use of a matrix analysis which displays secondary structure as a collection of diagonals on a square array plot of a sequence versus its complement [16, 17]. This approach, unlike many others, displays all potential secondary structures and permits ready examination them. This is made possible by the ability of the human eye to act as an image analyzer and to integrate over large areas when the data is suitably scaled and processed. We present here an approach to nucleic acid secondary structure prediction employing a new powerful matrix analysis system in which the filtering algorithms are based on a user-biased blend of thermodynamic parameters with data collected from studies employing chemical, enzymatic, and phylogenetic methods.

MATERIALS AND METHODS

Enzymatic Digestion of AMV-4 RNA

The radioactive labeling of AMV-4 RNA was accomplished as follows: Six micrograms of AMV-4 RNA were labeled with cytidine 3',5'-bis(phosphate), [5'-³²P] as described by Thomas *et al.* [18], and the labeled RNA was purified by preparative gel electrophoresis on a 3.6% polyacrylamide gel as described [8] with the exception that the gel was cast in TA buffer (36 mM NaOAc, 35 mM Tris, 2 mM EDTA, pH 8.3). Running buffer was TA buffer plus 0.2% SDS.

Partial enzymatic digests of AMV-4 for RNA were accomplished as follows: Enzymatic reaction buffers and digestion conditions were identical to those described previously [8] with the exception that the S1 ladder and formamide ladder reactions contained 7500 cpm while all other reactions contained 5000 cpm. Reactions performed in native buffer were ethanol precipitated and washed several times with 70% ethanol before loading onto gels in order to avoid salt effects.

Computer-Aided Modeling

The procedure for analyzing a sequence is based on the use of a computer program which allows (1) the examination of as many of the possible substructures as desired; (2) the use of filtering to incorporate information on (a) pairing length, (b) pairing and stacking energies, (c) experimental data, (d) user assumptions; (3) incorporation of related sequences; and (4) user selection and evaluation. In order to facilitate these goals, a number of inputs are read and two outputs are produced. One is a matrix showing "acceptable" secondary structure regions and the second is a detailed list of

the various paired regions.

Program Description: The program is written in FORTRAN-77 for a DEC VAX-11/780. The majority of the code is machine-independent although it does take advantage of the large addressable memory of the VAX. Some portions of the code for accessing the various plotting devices is device specific. The programs are available from the authors on an as-is basis.

Input Data: The three input files are as follows: (1) The sequence file containing a list of the sequence consisting of C, G, A and U or T; the letter X may be used as a non-pairable residue. (2) A file containing enzymatic and chemical data which specifies regions designated as non base-paired or paired to other specified regions. The chemical and enzymatic data are entered as user-assigned, double-stranded probabilities. There may be a number of different subsets of input probabilities which correspond to different reactants or strengths of reaction. These may include various probabilities for a given residue, as for example when it reacts with both single- and double-stranded nucleases, or with two different single-stranded nucleases. The various residue probabilities are combined under the assumption that each entry is independent and that only two states (paired and unpaired) exist.

$$\text{Prob.}(\text{unpaired}) = 1.0 - \text{Prob.}(\text{paired})$$

The total probability of a residue being paired is the renormalized product of the probabilities from separate reactions. (3) The final input is a set of control parameters which vary slightly from system to system, depending on the output device. It provides various necessary items of information, including the names of the files to be used and to be produced, regions of the sequence to be checked for pairing, various filter parameters (to be discussed) and symbols to be used in the matrix output (described below).

Filtering: The earliest matrix procedures consisted of simply placing a dot at each intersection element which could correspond to a Watson-Crick base pair. While these diagrams were useful they were also very dense, with roughly a quarter of the grid points marked, and possible non-standard Watson-Crick base pairs such as G-U unmarked. Several procedures have been used by various researchers to improve the situation. These generally involve accepting G-U base pairs while displaying only those pairing regions with more than some minimum number of pairs. Determination of structure often involves considering detailed stacking energies and we have included this option. We use a selection matrix which contains values of 0 or 1 for all possible two-base combinations. A zero indicates that the base combination is not

considered for potential base pairing, and a one indicates that it is. Normally the Watson-Crick and G-U pairings are included, but in certain situations G-A or other combinations might be considered acceptable.

The program then uses a length criterion to decide whether to proceed or discard a given pairing. That is, it is only accepted if some minimum number of contiguous bases fulfills the criterion for potential pairing. A minimum loop length of three is assumed. The effect of this is to produce a strip three bases wide along the matrix diagonal having no potential base pairs. If this criterion is met several other criteria (filters) are examined.

First of all, the maximum pairing length is determined by extending from the minimum until an unacceptable pair is reached. These potentially pairing fragments are then analyzed further in terms of two criteria, (a) the pairing-stacking energy of the fragment and (b) chemical/enzymatic data.

To calculate the pairing-stacking energy of the fragment, the program has a 5x5x5x5 matrix corresponding to the energies of all combinations of the five possible residue types (four bases and a blank for ends of sequences) at each of two consecutive residues on the two proposed paired chains. As a first approximation, and until reliable pairing-stacking energies can be established, we have simply assigned stacking-independent values for G-C, A-U, and G-U base-pairs of -3.0, -2.0 and 0.0 Kcal., respectively. An examination of several sample sequences indicates that this is at least a reasonable starting point. The sum of the energies for the residues along the proposed maximum pairing region is calculated, and if this energy is below some preselected value (EV1) the sequence is considered to have passed this filter and its energy value is recorded. If not, the sequence is checked for any subfragments which are greater than the minimum length, and those which are of low enough energy to pass are passed. It should be noted that the recorded energy may not be the lowest possible involving the pairing sequence since a potential pair with a positive energy contribution at or near the end might cost energy but still be passed.

Filtering by chemical/enzymatic modification is carried out as follows: The products of the normalized paired and unpaired probabilities are calculated and the overall products renormalized. In order to put the probabilities on roughly the same scale as the pairing-stacking energies, a "pseudo-energy" is calculated using a Boltzmann transformation at 300°K:

$$''\Delta E'' = -300 \cdot k \cdot \ln(\Pi(\text{Prob}(\text{Paired})_i) / \Pi(\text{Prob}(\text{Non-paired})_i))$$

Table 1: Sample pseudo-energy calculations for a hexamer with varying numbers of single- and double-stranded hits of varying strengths.

Number Double- Stranded hits	Prob. Double- Stranded	Number Single- Stranded hits	Prob. Single- Stranded	Aggregate Prob. Double- Stranded	Pseudo- Energy
12	0.7	0		0.99996	-6.0
6	0.7	0		0.994	-3.0
3	0.7	0		0.93	-1.5
5	0.7	2	0.3	0.93	-1.5
1	0.7	4	0.3	0.07	+1.5
3	0.8	0		0.98	-2.5

Table 1 shows some typical examples of pseudo-energies calculated for a possible hexamer.

As is the case with stacking energy, a check is made to determine if the pseudo-energy is below some preset value (EV2). Finally, the sum of the pseudo-energy and stacking energy is calculated and this is checked against a third preset energy value (EV3). If the sequence passes all criteria the appropriate outputs are generated as will be described. If the sequence does not pass all three then all subfragments down to the minimum length are checked and if appropriate passed.

Output Files: Two output files are produced, each containing a record of the input for the run. The first file then contains an ordered list of all the possible fragments. A sample is shown in Figure 1. For each fragment the following data are listed: first pairing residue of initial sequence, last pairing residue of the complementary sequence, pairing length, loop length, pairing and stacking energy, reactivity pseudo-energy and the sum of these two energies. In addition, the pairing region is displayed with the three flanking residues on either side and with symbols above or below each residue indicating its reactivity, as shown in Table 2. This allows for a detailed analysis of each potential pairing region.

The second output contains a matrix displaying the various fragments along with several indications as to their reactivity. The outputs vary slightly depending on the attributes of the output device being used. Currently, versions exist for the following devices: a standard line printer which has a limit in the range of one to two hundred nucleotides/frame but can put any available symbol at each grid point; an N.E.C. Spinwriter which can go up to approximately 400 nucleotides/frame, again with any character at each

495	632	11	116	CCA UGC GGGAUUC AAA	621	869	7	235	AAA GGCGUUC CCG
-25.0	-1.3	-26.3	CGA GCGCCUUGCG GAA	-16.0	-5.6	-21.6	UCC CCGUAG UAC	...	...
			...				...	...	...
622	505	11	116	AAG GCGUCCCGCG AGC	643	666	9	6	CCC CUCCGAGU UUG
-25.0	-1.3	-26.3	AAA CGUAAGGCGU ACC	-19.0	-3.3	-22.3	GUC GGGGUGUA UUA	...	...
			...				...	...	...
658	651	9	6	AUU ACUGGGG CUG	681	698	7	4	AGC UGACGAU CUU
-19.0	-3.3	-22.3	GUU UGACGCCUC CCC	-17.0	-3.0	-20.0	GUU ACUGCUA GUU	...	...
			...				...	...	...
692	687	7	4	UUG AUCGUCA UUG	699	763	11	43	UUA UUGAUUACCC CAU
-17.0	-3.0	-20.0	UUC UAGCAGU CGA	-19.0	-1.4	-20.4	UGG AAUAGGUGG UCA	...	...
			...				...	...	...
753	709	11	43	ACU GGGUGGAUUA GGU	770	812	11	21	CAA GGUUGAAGUC CUA
-19.0	-1.4	-20.4	UAC CCCAUUAGUU ACU	-21.0	-4.6	-25.6	UCG UUAUACUUCAG CUA	...	...
			...				...	...	...
802	780	11	21	AUC GACUCCAUAU GCU	805	865	5	51	GAC UUCAU AUU
-21.0	-4.6	-25.6	AUC CUGAAGUAGG AAC	-11.0	-10.4	-21.4	CGU AAGUA CGU	...	...
			...				...	...	...
815	842	12	4	UCC UUAUUAUUCU UAA	818	841	8	8	UUA UUAUUCU CUA
-27.0	-5.9	-32.9	GUA AAUAUAUACCG CAA	-15.0	-6.9	-21.9	UAA AAUAUAUAC ACG	...	...
			...				...	...	...
831	826	12	4	AAC GCACAUUAUAA AUG	834	825	8	8	GCA CAUAUAUA AAU
-27.0	-5.9	-32.9	AAU CGUGUAUAUAU CGU	-15.0	-6.9	-21.9	AUC GUGUAUAU AUU	...	...
			...				...	...	...
847	864	7	4	UCC UCAUGCA AAA	858	853	7	4	AAC UGCAUGA AUG
-17.0	-12.0	-29.0	GUA AGUACGU CAA	-17.0	-12.0	-29.0	AAA ACGUACU CGU	...	...
			...				...	...	...
861	809	5	51	UCC AUGAA UGC	863	627	7	235	CAU GAAUGCC CCU
-11.0	-10.4	-21.4	UUA UACUU CAG	-16.0	-5.6	-21.6	GCC CUUGCGG AAA	...	...
			...				...	...	...

Figure 1: Sample Fragment List from AMV-4 RNA. The fragment list shown here corresponds to the matrix shown in Figure 7. The twenty fragment entries correspond to the twenty diagonals shown in the matrix plot. Each entry includes a schematic representation of the paired regions and flanking nucleotides as well as special symbols which refer to double-stranded probabilities (see text).

grid point (at the higher density, however, characters larger than a dot run together and become uninterpretable); and a Printronix type printer-plotter in plot mode which allows only dots (or spaces) to be plotted but gives up to 750 nucleotides/frame.

Table 2: Symbols associated with ranges of double strand probability in fragment output.

#	> 0.8	Probability double-stranded
+	0.6 - 0.8	" "
.	0.4 - 0.6	" "
-	0.2 - 0.4	" "
=	< 0.2	" "
*	Specific pairing or non-pairing	

An example of output from the Printronix is shown later in this paper (Figure 7). The actual plot is outlined by a double border; the outer border is a histogram of the double-stranded probability derived from reactivity data. The orientation of the histogram is such that an increase in double-stranded probability for the side border is toward the left, whereas for the top and bottom borders the increase in double-stranded probability is directed upward. The base line represents a 40-60% probability, while 2 dots or 1 above or below represent >80%, 60-80%, <20% or 20-40% respectively. The inner border consists of a dot at the corners and for every twentieth residue. Each grid location in the matrix which corresponds to the potential pairing of the respective nucleotides down and across has a number of attributes which determine the character to be displayed at that location. The attributes are: whether from reactivity data, the residue in the first strand is double-stranded (>70% prob.), uncertain (30-70% probability of being double-stranded) or probably single-stranded (<30% prob. of being double-stranded); the same determination for the second strand; and finally, whether the pair has passed the criteria for forming a double-stranded region. Since the program does not distinguish which chain has a given attribute, the total number of attributes is reduced to 12, six paired possibilities and six unpaired. On those printers which allow different characters, each attribute can be given a different symbol. In the case of the dot matrix shown, only a dot or space is available. In this example all paired regions have dots and all nonpaired are blank. One could redo the run to show only those portions of the pairing diagonals indicated by reaction data to be probably double helical, or one could use the similar but slightly less restrictive option of employing blanks for those portions which reactivity data indicate are probably single-stranded.

Matrix Interpretation: In Figure 2a a nucleotide primary sequence is shown which can potentially form the secondary structure shown in Figure 2b. This structure contains both long-range and short-range complementary base-pair interactions. The types of structures indicated include all the types defined by Zuker and Steigler [14]. Figure 2c is a matrix representation of the sequence shown in Figure 2a. A minimum filtering level of length 5 was used in producing this plot. The potential secondary structures are represented in the matrix as diagonals perpendicular to the long diagonal which represents the 2-fold symmetry reflection axis for the matrix plot. A base-pairing relationship exists between the sequences represented by the two symmetrically disposed diagonals. The 2-fold reflection axis does not

Figure 2: Sample secondary structure. a) Primary structure. b) Hypothetical folding. c) Complete matrix of sequence shown in (a). d) Matrix corresponding to the hypothetical folding shown in (b).

represent any secondary structure potential, but is merely a reference line drawn by the program in order to facilitate the association of any diagonal with its symmetrical counterpart. The complementary relationship between the sequences represented by any pair of symmetrical diagonals has been described elsewhere [16].

Any unbroken diagonal represents an uninterrupted region of base-paired

Figure 3: Sample secondary structures as they appear in a matrix. a)-d) Simple secondary structures. e)-f) higher-order secondary structures.

nucleotides. A break in a diagonal, such as that indicated for diagonals E-F and E'-F' (Fig. 2d), shows a symmetrical interior loop in which an equal number of nucleotides is looped out of the pairing region on either strand between two paired sequences. If the interior loop is not symmetrical (i.e., if the number of nucleotides looped out is greater on one side than the other), the two interrupted diagonals are displaced either vertically or horizontally with respect to an uninterrupted diagonal. In the example shown in Figure 2d diagonals B-C and B'-C' are examples of horizontal displacement which results from a loop-out on the 3' side of the paired nucleotides. If the 5' side of the paired region contained the loop-out, the displacement would have been vertical with respect to an uninterrupted diagonal (Fig. 3).

All the diagonals contained within the matrix can be related to any other diagonal in the matrix by connecting series of neighboring parallel lines extending vertically from the end elements of any diagonal in a consecutive fashion across the matrix (Fig. 2d). The net result is a consecutive folding map which describes the spatial relationship of any portion of the secondary structure with respect to the entire structure.

Figure 2 contains some secondary structures which are composites of stem and loop structures held together by long-range interactions. For example, in

Figure 2b, the secondary structure consisting of stem G:G' encloses a "domain" consisting of two short-range stem and loop structures H:H' and I:I'. Similarly, stem J:J' encloses the structures K:K', L:L', and M:M'. These bipartite and tripartite (as well as higher order) structures are readily identified on the matrix plot. This is because any diagonals contained within a domain can be bounded by vertical and horizontal lines extending from a diagonal which represents paired nucleotides involved in long-range interactions closing that domain (Fig. 2b). Long-range interactions which do not enclose any higher-order domain structures do not enclose diagonals within extended line boundaries (Fig. 2d, diagonals N:N'). Note the complex domain structure enclosed by stem D:D' in Figure 2.

It should be obvious by now that horizontal and vertical axial relationships figure prominently in matrix interpretation. A useful interpretation which takes advantage of this is the realization that certain diagonals cannot coexist within any unique overall structure. Therefore any two diagonals may be mutually inclusive (i.e., both are compatible with a unique structure) or mutually exclusive (i.e., one or the other, but not both, may occur within the same structure) with respect to any unique folding pattern. The trade-off structures contained within the attenuator regions of mRNAs coding for amino acid biosynthetic enzymes in *E. coli* [19, 20] are examples of mutually exclusive secondary structures. Mutually exclusive diagonals are readily pinpointed on any matrix plot because they are co-intersected by a horizontal or vertical line passed through the matrix (Fig. 4). If for some reason the secondary structure associated with a particular diagonal is considered to have a high certainty associated with its existence in the folding scheme, any diagonals which intersect horizontal or vertical lines passing through its elements can be eliminated from consideration in the final structure.

Knots: Any set of mutually inclusive diagonal pairs may form a knot, provided that each diagonal in the set fulfills all of the following requirements:

1. Each diagonal is not contained within a bounded domain closed by the other diagonal in the set.
2. The bounded domains for the diagonals intersect.
3. The diagonals are sufficient to describe at least one complete turn of the helix (an RNA helix theoretically should be 11 nucleotides per rise).

Figure 4: Mutually exclusive structures. Trade-off structures cannot simultaneously coexist within a single strand of RNA. The diagonal pair $a:a'$ defines an area which excludes other diagonals.

Fulfillment of all three requirements above categorizes the structure as a "potential" knot. "Real" knots are impossible to determine from the matrix analysis since they require threading a molecular end through a loop (see Fig. 5). A structure which looks locally identical to that described by Cantor [21] (see Fig. 5) may be only pseudo knotted, i.e., the structure may locally appear knotted but may not have resulted from the threading through of an end. Fig. 5 shows an example of how such a pseudo-knot can be generated from a theoretical, circular molecule which is incapable of being knotted (at least in the simple sense described here).

General Considerations: The selection of probabilities associated with

Figure 5: Potential knot structures. a) Potential knot structure is represented in the matrix by diagonals which bound overlapping areas (shaded). b) "Real" knot. c) Formation of a "pseudo" knot.

reactivities and of filter levels are clearly somewhat arbitrary and are not independent of each other. If only single-stranded reactivity data are available, for example, the pseudo-energy cut-off must be greater than or equal to zero. Reactivity data may not be uniform. Indeed, it may be quite good for some portions and nonexistent for others. These facts need to be taken into account by the user, both in selecting filters and in interpreting the results.

When interpreting the pairing and stacking free energies one must keep in mind that these are at best a good averaged estimate for the energy in the particular sequence. In any given real structure, however, various environmental factors such as tertiary structure, protein interactions, etc., could easily shift the stacking and affect the interactions. The user should not take the specific energies too literally and certainly should not select among otherwise more or less equivalent structures on the basis of small differences in stacking/pairing energies alone.

Given the above limitations, how can one proceed? We have chosen to accept the bias that hairpins with a short loop of 5 to 9 nucleotides are preferred over long-range pairing. This assumption can be justified in a number of ways [11, 22]. It should be noted that for large loops the nucleotides located between paired regions could be involved in additional structures, thus effectively shortening the distance between the nucleotides at the ends of the loop. This possibility must be carefully considered, since it may define structural domains. For these reasons we tend to select initially only the best short-range interactions.

As one proceeds through a structure, certain trade-off structures will be discovered which are mutually exclusive and appear to be equally good. Several approaches are possible at this stage. One could ignore both and see if other indications arise. Alternatively, one could fork the interpretation by trying both and determining if either evolves to a more obviously correct structure. Good trade-off structures should certainly be noted and examined, since they could be active regions of the molecule which change its folding as a result of various environmental factors.

General Procedure for Analysis of Secondary Structure

1. Enter sequence or series of several related sequences.
2. Enter reactivity data for each residue as double-stranded probabilities.
3. Enter regions which have been selected as paired or non-paired.

4. Run two or three passes using high, medium and low filter levels.
 - a. high: length = 4, EV1 = -12.0, EV2 = -2.0, EV3 = -20.0
 - b. medium: length = 4, EV1 = -8.0, EV2 = 0.0, EV3 = -16.0
 - c. low: length = 4, EV1 = -4.0, EV2 = +5.0, EV3 = -5.0
5. Examine high filter for good, short hairpins and other outstanding features.
6. Select best as tentatively paired structures.
7. Repeat 3 - 6 with particular note of:
 - a. those pairing regions brought close by previous pairing,
 - b. domain structures (i.e., structure closing off a number of paired and possibly unpaired regions).

RESULTS: ANALYSIS OF AMV-4 RNA

In order to test the usefulness of the program in studying RNA secondary structure employing chemical/enzymatic data, we proceeded to collect such data on an RNA molecule using some currently available techniques. Figure 6a shows a polyacrylamide gel which was used to display structural information on 3'-labeled AMV-4 RNA. The experimental conditions are described in the figure legend. The end-labeled RNA was subjected to limited digestion by enzymes which are sensitive to specific types of structures within the RNA molecule. A band on the gel locates a site where the structure-specific enzyme cleaves the RNA molecule. Such methodology has been used previously for nucleic acid structure probing, as has been described elsewhere [23, 24]. The enzymes used in our experiments are sensitive to single-stranded (S1 and MB1) as well as double-stranded (V1) regions of RNA molecules. The location of the digestion site within the RNA molecule is recorded in a constraint file as described in the 'Methods' section. An example of a constraint file used in the analysis of AMV-4 RNA is shown in Figure 6b. Each site is entered into the file in association with a probability value which describes the intensity of the cleavage. As mentioned above in 'Methods,' the probability recorded is that of the residues being double-stranded. Therefore, strong S1 and MB1 sites are given probabilities approaching 0.0 whereas strong V1 sites are given probabilities approaching 1.0. It is therefore a reasonably straightforward procedure to enter the gel data into the constraint file for use in filtering the matrix.

Figure 6: Enzymatic digestion data for 3'-labeled AMV-4 RNA. a) Electrophoretic analysis of partial enzymatic digests. The conditions for radioactive labeling and for the enzymatic digestions are described in 'Materials and Methods' and in an earlier publication [8]. The polyacrylamide gel is a 6% polyacrylamide, 8.3M urea gel in TBE buffer (90 mM Tris, 2.5 mM EDTA, 90 mM boric acid), and was electrophoresed at 65 watts constant power. The samples on the left were run for 3 h; the samples on the right were run for 1 h. Lane 1 is non-digested; 2 and 3, MB1 nuclease; 4 and 5, V1 ribonuclease; 6, T1 ribonuclease (denatured RNA); 7, formamid; 8, S1 nuclease (denatured RNA); and, 9 and 10, S1 nuclease. b) Data constraint file.

Figure 7: Matrices for AMV-4 RNA. a) Matrix resulting from free energy filtering without enzymatic data weighting (length = 4, EV1 = -9, EV2 = 20, EV3 = 20). b) Composite matrix showing two different enzymatic data filtering levels. The lower left half is from a matrix filtered at length = 4, EV1 = -9, EV2 = 3, EV3 = -12. The upper right half of (b) is from a matrix filtered at length = 4, EV1 = -9, EV2 = 3, EV3 = -16. c) Matrix with high stringency enzymatic filtering (Length = 4, EV1 = -9, EV2 = 3, EV3 = -20). A through G indicate mutually exclusive diagonals which are in good agreement with the enzymatic data. d) Composite matrix using exclusivity filter. The lower left half of the matrix shows output when the program was run with the parameters set identically to those shown in (a) except that the structures shown in (c) were specified as being paired. The upper right half of the matrix is similar to the lower left except that enzymatic data was included (length = 4, EV1 = 20, EV2 = 20, EV3 = -9).

Figure 7 shows filtered matrices corresponding to nucleotides 400 through 881 of AMV-4 RNA. Each matrix is bordered by a histogram plot which displays the position and magnitude of the summed double strand probabilities entered into the program via the constraint file. Figure 7a is a complete matrix (showing both symmetric halves) in which the AMV-4 RNA sequence is analyzed for potential intramolecular base-pairing without regard for the enzymatic data contained within the constraint file. The level of filtering is such that only contiguous pairs of four or greater, which possess a free energy of -9 Kcal or less, are displayed. In Figure 7b two half-matrices are shown representing two different levels of constraint-file filtering. In Figure 7c the stringency of the filter is sufficiently high that only a small number of diagonals remains in the matrix. These diagonals (labeled A through G in Figure 7c) or any other diagonals can be selected as matched structures during the program execution simply by indicating within the constraint file the nucleotide positions of the match. This is accomplished by assigning the code M for match, followed by the length and position of the paired region. An example of this is shown at the bottom of Figure 6b where the notations M-A through M-G are matched regions representing the diagonals indicated in Figure 7c. When diagonals A through G are required to match in this way, the program automatically filters the matrix diagonals on the basis of exclusivity and removes any other diagonals that lie along the areas excluded by the 'matched' nucleotides (see Figure 4). Figure 7d shows two half-matrices filtered in this way. The lower left half-matrix was calculated using the same parameters as in Figure 7a, but with exclusivity filtering. Constraint file data filtering was not used. It is interesting to note that increasing the stringency of the data filtering not only eliminates some diagonals, but results in the appearance of new diagonals which would have otherwise been eliminated by filtering. This is because in some instances the data-derived pseudo-energies can lower the sum energy with respect to that which would be calculated by simple analysis of the nature of the base pairs without regard for the chemical/enzymatic data.

Figure 8 is a schematic representation of the 3' terminus of AMV-4 RNA as modeled by our interpretation of the filtered matrices described above. Superimposed upon the structure are arrows indicating the sites of enzymatic digestion determined by visual inspection of the gel shown in Figure 6. The structure shown appears to be in good general agreement with the data. Two notable exceptions are shown as shaded regions in Figure 8. These regions are located between positions 710 and 752 (adjacent to stem D) and between

Figure 8: Possible structure for the 3' end of AMV-4 RNA. The structure is shown schematically. Dashed lines connect adjacent nucleotides. Bold arrows indicate single-stranded (MB1 and S1) digestion sites and light arrows indicate double-stranded (VI) digestion sites. No distinction is made for digestion intensity. Capital letters adjacent to certain helices correspond to similarly designated diagonals show in Figure 7d. The remaining helices correspond to the diagonals designated by lower case letters in Figure 7. Shaded regions interconnected by double-headed arrows indicate trade-off structures which are consistent with the data.

positions 781 and 801 (adjacent to stem C). These regions, unlike the other positions shown, contain common sites for both single and double strand specific nuclease digestion. This result could be due to a lack of structure specificity by the nucleases. However, the fact that the digestion in the remainder of the molecule appears consistent with the model suggests that a lack of nuclease specificity is not the problem. Another possible explanation might be that either some sort of tertiary structure or structure trade-off exists which causes the observed pattern. The latter explanation is reasonable provided that the potential for forming such trade-off structures

could be located within the AMV-4 RNA sequence. Examination of Figure 7d reveals that proximal to diagonals C and D are several mutually exclusive trade-off structures that could exist within this region of the sequence (m' and n' versus k and l near stem D; and i' versus h' near stem C). These structures are also shown as shaded areas in Figure 8. It is obvious that either a heterogeneous population or an equilibrium between the two structures shown for each region would yield the observed digestion pattern.

DISCUSSION

We present a method involving a novel computer program for predicting and analyzing nucleic acid secondary structures. This method invokes a series of filter algorithms which utilize chemical/enzymatic data and produce a reasonable fit of a sequence to a secondary structure. In addition to the AMV-4 RNA data presented here, we have also used the program for analyzing the published chemical/enzymatic data reported for E. coli 16S rRNA, and we have found that the program is capable of predicting a structure which is similar to that described by Noller and Woese [3] and Brimacombe et al. [4]. We are currently in the process of testing the program on other RNA sequences which are of biological interest, such as α - and β -globins as well as other rRNAs.

A striking feature of most of these studies (AMV-4 RNA and E. coli 16S rRNA included) is that the majority of the stable helices seem to represent short-range interactions (diagonals located close to the symmetry diagonal). This is consistent with the model of RNA folding proposed by Auron et al. [22] in which kinetic folding of proximal complementary nucleotides to form stems precedes the formation of long-range interactions. The preponderance of such short-range structures is obvious in AMV-4 RNA and has also been shown to be the case for E. coli 16S and 23S rRNA [3, 4, 25, 26].

Another interesting observation made as a result of the experiments with AMV-4 RNA is the apparent existence of alternate structures representing either static heterogeneity or dynamic molecular "breathing." We have reported this type of phenomenon previously for the 5' end of AMV-4 RNA [8]. Noller et al. have similarly described this phenomenon in E. coli 23S rRNA [26].

Since alternate structures seem to exist in RNA molecules and may be biologically important, the structure modeling method should take this possibility into consideration. Our program, unlike some other approaches [14] and [15], has the ability to display many different, mutually exclusive trade-off structures. The secondary structure data derived from

experimentation can then be used to prune these structures and thereby yield a unique or several alternate results.

Although the examples of nucleic acids referred to in this paper are RNA molecules, it is also possible to analyze DNA sequences in a similar way. The base-pairing rules in the program are easily redefined. We have developed a similar and closely related program for nucleic acid sequence homology determination which allows for varying degrees of mismatching (to be presented elsewhere).

We have made use of a suggestion forwarded by Maizel and Lenk [16] in which color transparencies are used to analyze matrix data. By adjusting the program matching parameters appropriately, we can produce output matrices which contain a dot element wherever there is double strand or single strand data coincident with a diagonal. If these plots are converted into color transparencies of different colors and overlaid, the result is a multicolored plot which clearly indicates data agreement or disagreement for individual diagonals. Routinely, we show double-stranded dot elements in blue, single-stranded dot elements in red and non-data elements in green. The result is that diagonals which agree well with the data are primarily blue in color, those that do not agree with the data are red, and those that are neutral with respect to the data are green. In this way, very low length and free energy filter levels can be used to show virtually all potentials and thus allow more diagonal alternatives to be simultaneously analyzed. Of course, the color plots can be filtered more stringently to yield additional data on individual diagonal stability on the basis of free energy.

ACKNOWLEDGMENTS

We would like to thank Dr. John Bol of the State University of Leiden, The Netherlands, for the generous gift of the AMV-4 RNA. We also thank Dr. Nahum Sonenberg, McGill University, for helpful discussions. Special thanks to Dr. William Gilbert and the Whitaker College Computer Facility for discussions and services. Pam Alexander provided invaluable help in preparing the manuscript. This work was supported by grants from the N.I.H. (2-S07-RR07047-18; CA0-4186-25), the U.S. Department of Agriculture (83-CRCR-1-1275), the Whitaker Health Science Fund (84-02), and MIT UROP. We wish to thank Dr. Katsumi Isono of the Max-Planck Institute of Molecular Genetics, Berlin, GFR, for providing us with a copy of his homology program, MATCH, which gave us the inspiration to use matrix programs in our research.

⁺Harvard-M.I.T. Division of Health Sciences and Technology, Cambridge, MA 02139, USA

[§]Department of Anatomy, Harvard Medical School, Boston, MA 02115, USA

[#]Department of Experimental Medicine, Tufts University Medical School, Boston, MA 02111, USA

REFERENCES

1. Quigley, G.J. and Rich, A. (1976) *Science* 194, 796-806.
2. Fiers, W., Contreras, R., Duerinck, F., Haeghebaert, G., Merregaert, J., MinJou, W., Raeymakers, A., Volckaert, G., Ysebaert, M., Van deKerckhove, J., Nolf, F. and vanMontagu, M. (1975) *Nature* 256, 273-278.
3. Noller, H.F. and Woese, C.R. (1981) *Science* 212, 403-411.
4. Glotz, C., Zwieb, C., Brimacombe, R., Edwards, K. and Kossel, H. (1981) *Nuc. Acids Res.* 9, 3287-3306.
5. Pavlakis, G.N., Lockhard, R.E., Vavakopoulos, N., Rieser, L., RajBhandary, U. and Vournakis, J.N. (1980) *Cell* 19, 91-102.
6. Kozak, M. (1980) *Cell* 19, 79-90.
7. Morgan, M.A. and Shatkin, A.J. (1980) *Biochemistry* 19, 5960-5966.
8. Gehrke, L., Auron, P.E., Quigley, G.J., Rich, A. and Sonenberg, N. *Biochemistry*, in Press.
9. Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L., and Steitz, J.A. (1980) *Nature* 283, 220-224.
10. Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E. and Cech, T.R. (1982) *Cell* 31, 147-158.
11. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) *Nature New Biology* 246, 40-41.
12. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biology 42, 985-1002.
13. Gralla, J. and Crothers, D.M. (1973) *J. Mol. Biol.* 73, 497-511.
14. Zuker, M. and Stiegler, P. (1981) *Nucleic Acids Res.* 9, 133-148.
15. Nussinov, R., Tinoco, I. and Jacobson, A.B. (1982) *Nuc. Acids Res.* 10, 351-363.
16. Maizel, J. and Lenk, (1981) *Proc. Natl. Acad. Sci., USA* 78, 7665-7669.
17. Pastell, J. and Kafatos, F.C. (1982) *Nuc. Acids Res.* 10, 4764-4782.
18. Thomas, N.S.B., Butcher, P.D., and Arnstein, H.R.V. (1983) *Nuc. Acids Res.* 11, 1-10.
19. Lee, F. and Yanofsky, C. (1977) *Proc. Natl. Acad. Sci., USA* 74, 4365-4369.
20. Gardner, J.F. (1979) *Proc. Natl. Acad. Sci., USA* 76, 1706-1710.
21. Cantor, C.R. (1980) in *Ribosomes: Structure, Function and Genetics*, Chamblis, G., Craven, G. R., Davies, J., Davis, K., Kahan, L. and Nomura, M., Eds., pp. 23-49, University Press, Baltimore, Md.,
22. Auron, P.E., Rindone, W.P., Vary, C.P.H., Celentano, J.J. and Vournakis, J.N. (1982) *Nuc. Acids Res.* 10, 403-419.
23. Wurst, R.M., Vournakis, J.M. and Maxim, A.M. (1978) *Biochemistry* 17, 4493-4499.
24. Vournakis, J.M., Celentano, J., Finn, M., Lockard, R.E., Mitra, T., Pavlakis, G., Troutt, A., van der Berg, M. and Wurst, R.M. (1981) in *Gene Amplification and Analysis*, Chiekjian, J.G. and Papas, T., Eds., pp. 267-298, Elsevier Press, North Holland,
25. Steige, W., Glotz, C. and Brimacombe, R. (1983) *Nuc. Acids Res.* 11, 1687-1706.
26. Noller, H., Kop, J., Wheaton, V., Brosius, J., Gutell, R.R., Kopylov, A.M., Dohme, F., Herr, W., Stahl, D.A., Gupta, R. and Woese, C.R. (1981) *Nuc. Acids Res.* 9, 6167-6189.