Volume 14 Number 1 1986 Nucleic Acids Research

A dynamic programming algorithm for finding alternative RNA secondary structures

Arthur L.Williams, Jr. and Ignacio Tinoco, Jr.

Department of Chemistry and Laboratory of Chemical Biodynamics, University of California, Berkeley, CA 94720, USA

Received 6 June 1985

ABSTRACT

Dynamic programming algorithms that predict RNA secondary structure by minimizing the free energy have had one important limitation. They were able to predict only one optimal structure. Given the uncertainties of the thermodynamic data and the effects of proteins and other environmental factors on structure, the optimal structure predicted by these methods may not have biological significance. We present a dynamic programming algorithm that can determine optimal and suboptimal secondary structures for an RNA. The power and utility of the method is demonstrated in the folding of the intervening sequence of the rRNA of Tetrahymena. By first identifying the major secondary structures corresponding to the lowest free energy minima, a secondary structure of possible biological significance is derived.

INTRODUCTION

Numerous algorithms have been developed to predict RNA secondary structure by minimizing the free energy (1-9). The quality of these predictions depends upon the accuracy of the thermodynamic data which describe the free energies of various secondary structural features; the folding rules that an algorithm uses to find the lowest free energy structure; and the degree to which environmental conditions stabilize alternate structures of equivalent or higher energy.

The thermodynamic data available today and the folding rules based on those data cannot be depended on to predict a correct secondary structure. The data for base-paired helical regions are thought to have an uncertainity of $\pm$ 0.2 to 0.5 kcal, while the values for the loops have an estimated uncertainity of $\pm$ 1 to 2 kcal (10-14). Furthermore, there are secondary structural features (e.g., the sequence and base composition of loops, the size and shape of multistem loops) whose effects on the stability of a

Requests for programs, along with a self-addressed mailing label and a blank tape, should be sent to Dr. A.L. Williams, Jr., Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, NY 10029. A small charge will be requested to cover mailing and processing.

secondary structure have not been determined.

Even if a most stable secondary structure could be calculated accurately, other structures could have important biological functions. The apparent existence of alternate structures has been reported in studies of AMV-4 RNA (15) and E.coli. 23S rRNA (16). Alternate structures may be important in the initiation of translation (17), mRNA stability (18), and in transcriptional attenuation (19). Alternate equivalent and suboptimal free energy structures would be of particular interest in searches for structural isomorphism of related RNAs to identify these structures.

In recent years, a number of efficient dynamic programming algorithms have been presented (9, 20-21). However, all of these algorithms predict only one optimal free energy structure for each nucleotide sequence. Alternate equivalent and suboptimal free energy structures are not identified. Algorithms have been presented that are capable of identifying more than one secondary structure (4, 15, 22). However, these algorithms require human intervention at several stages and/or many shortcuts and compromises are made in order to arrive at a solution.

We present an algorithm that can identify the optimal and suboptimal free energy structures near the optimum for a given nucleotide sequence. We combine concepts from Sankoff et al.(9), Zuker and Stiegler (6), Comay et al.(21), and Waterman (23) into a dynamic programming algorithm designed for this purpose. The power of the algorithm is demonstrated with the folding of the intervening sequence for rRNA of Tetrahymena.

METHODS

A. Notation and Terminology

We will begin by introducing the notation and terminology that will be used to describe the algorithm presented here. The nucleotides of an RNA are numbered from the 5' to the 3' end. The sequence of nucleotides from i to j, where i (j) is the i^{th} (j^{th}) nucleotide, is written [i,j]. A base pair between nucleotides i and j is written i*j, where $1 \leq i < j \leq n$ for [1,n]. $S_{i,j}$ denotes the minimum free energy secondary structure that can form on [i,j].

A secondary structural unit, u (i.e., a hairpin loop, bulge loop, etc.), that is formed as a result of i*j is written as $u_{i,j}$. Any u may be defined by the number of base pairs, p, that close its ends and the number of unpaired nucleotides, r, between those base pairs. The base pairs other than i*j that close any $u_{i,j}$ are written as s*t and are said to be accessible from i*j. Accessible base pairs are listed s_1*t_1, s_2*t_2,....., where $s_1 < t_1 < s_2 < t_2 < \ldots$.

$s_m < t_m$ and m=p-1. For example, p=1 and $r \geq 3$ for a hairpin loop; while p=2 and r=0 for a stack or helical region. These and other examples are shown in Figure 1.

A secondary structure $S_{i,j}$ is the set of u's of which it is composed. Its free energy is written as $E(S_{i,j})$. The free energy of each type of u is written as $e_p(u^r)$. Lastly, any secondary structure which obeys the chosen base-pairing constraints is termed permissible.

B. The Minimization Recursion

Since dangling single-stranded regions neither add nor subtract from the stability of any secondary structure, of all the possible structures that are permissible for any [i,j], only those structures in which i and j are base-paired need to be calculated. There are two types of base-paired structures that can be formed - closed or open. A closed structure on [i,j] is any permissible secondary structure closed by i*j. An open structure on

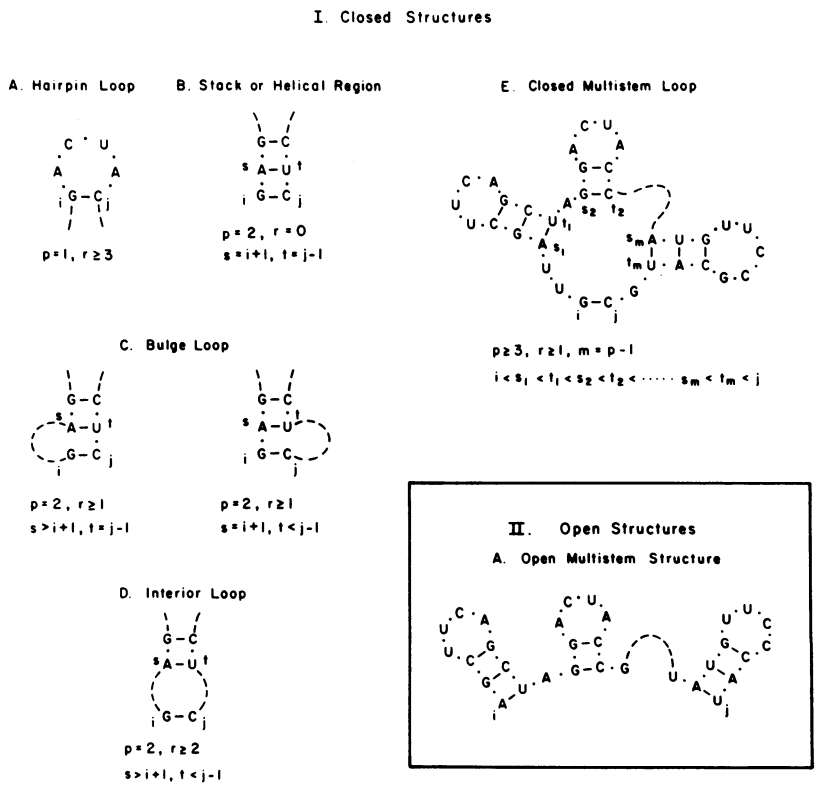

Figure 1. The closed and open secondary structural units, see text for explanation of nomenclature.

[i,j] is any permissible structure in which i and j are base-paired, but not to each other.

We define $C_{i,j}$ and $O_{i,j}$ to be the minimum free energy secondary structure of all the closed and open structures on [i,j], respectively. Their free energies are identified as $E(C_{i,j})$ and $E(O_{i,j})$. $C'_{i,j}$ and $O'_{i,j}$ identify the alternate optimal and suboptimal closed and open structures that can form on [i,j]. Only one structure will be identified as the optimal closed or open structure. All other secondary structures of equal free energy as the optimal are identified as alternate optimal closed or open structures.

The minimum free energy secondary structure, $S_{i,j}$, and its free energy, $E(S_{i,j})$, on [i,j] can be evaluated by the following recursion:

$$E(S_{i,j}) = \min \begin{cases} \text{infinity} \\ E(S_{i,j-1}) \\ E(S_{i+1,j}) \\ E(C_{i,j}) \\ E(O_{i,j}). \end{cases} \quad (1)$$

If the optimal secondary structure on [i,j] is a closed or open base-paired structure, then an optimal subregion, $R_{i,j}$, has been found. So that,

$$R_{i,j} = \begin{cases} C_{i,j}, \text{ if } E(S_{i,j}) = E(C_{i,j}); \text{ or} \\ O_{i,j}, \text{ if } E(S_{i,j}) = E(O_{i,j}). \end{cases}$$

So then,

$$S_{i,j} = \begin{cases} 0, \text{ or} \\ S_{i,j-1}, \text{ or} \\ S_{i+1,j}, \text{ or} \\ R_{i,j}. \end{cases}$$

C. Programming Considerations

Any dynamic-programming algorithm that attempts to identify alternate structures must accomodate additional computational back tracking and maximize its use of dynamic memory. We introduce here those procedures that address these problems. The procedures are: (1) the order of the search, (2) the use of vectors that order back tracking and, (3) the compacting and limiting storage of information.

(1) The Search Procedure

There are three mathematically equivalent search procedures that can be employed (20). However, in the algorithm that follows, the search starts with

short sections at the 5' end of the sequence and proceed to the 3' end. For each 3' nucleotide j, the programs evaluate the structures for each 5' nucleotide i, where $1 \leq i \leq j-4$, in the order of decreasing length. The advantage of using this particular search procedure will become evident in section (3).

(2) Optimization Vectors

To facilitate the back tracking necessary to determine both $C_{i,j}$ and $O_{i,j}$, as well as, the additional back tracking necessary to find the $C'_{i,j}$ and $O'_{i,j}$ structures for any [i,j], we make use of two sets of vectors.

One set of vectors identifies all the optimal subregions in which a particular nucleotide is part of a dangling single-stranded end, as well as those in which it forms one of the closing base pairs. These vectors are important in identifying the optimal subregions that will combine in a pair-wise fashion to form closed multistem loops. The vector S5(E,S,i,h,r) identifys all the optimal subregions of which nucleotide i is a 5' element. The vector S3(E,S,h,j,r) identifys all the optimal subregions of which nucleotide j is a 3' element. The elements of these vectors are assigned during the recursion (equation 1), in the following manner:

$$S5(S) = \begin{cases} S_{i+1,j}, \text{ if } S_{i+1,j} = S_{i,j-1} \text{ and } E(S_{i,j}) = E(S_{i+1,j}) \\ \qquad\qquad \text{or} \\ R_{i,j}, \text{ if } S_{i,j} = R_{i,j} \end{cases} \tag{2}$$

and,

$$S3(S) = \begin{cases} S_{i,j-1}, \text{ if } S_{i,j-1} = S_{i+1,j} \text{ and } E(S_{i,j}) = E(S_{i,j-1}) \\ \qquad\qquad \text{or} \\ R_{i,j}, \text{ if } S_{i,j} = R_{i,j}. \end{cases} \tag{3}$$

Another set of vectors identify only the optimal subregions in which a particular nucleotide forms one of the closing base pairs. These vectors are important in identifying the optimal subregions that will combine in a pair-wise manner to form open multistem structures. All the optimal subregions in which a nucleotide i is base-paired and closes the 5' end are identified by the vector R5(E,R,h,r). All the optimal subregions in which a nucleotide j is base-paired and closes the 3' end are identified by the vector R3(E,R,h',r). The elements of these vectors are assigned during the recursion (equation 1), in the following manner:

$$R5(R) = R3(R) = R_{i,j}, \text{ if } S_{i,j} = R_{i,j}. \tag{4}$$

(3) Information storage

We will now see how the search procedure [section (1)] enables us to store the vector information [section (2)] in an ordered and compact manner.

Instead of an (i,j) storage matrix, we have two types of vectors. Each of the vectors is a series of one dimensional arrays containing information on the nature of the optimal subregions: their free energy, their 5' and 3' closing base-pair nucleotides, their structural type ($C_{i,j}$ or $O_{i,j}$), and the number of unpaired nucleotides between their stems. One of these vectors is for fixed 3' nucleotides, j, where the 5' nucleotide ,i, varies. From equation (1), it is evident that for these vectors, "S3" and "R3", we need only the information for the j and j-1 nucleotides. Therefore, the "S3" and "R3" vectors contain only the information on the optimal subregions of those nucleotides. The second type of vector is for fixed 5' nucleotides, i. These vectors, "S5" and "R5", contain all the information on the optimal subregions for every 5' nucleotide, i.

The storage requirements for the "S5" and "R5" vectors can be quite substantial. However, from equations (2-4), it is clear that vector "R5" is a subset of "S5". The information in the one dimensional arrays are ordered, so that, the R5 data is stacked above the rest of the S5 data for each nucleotide i. Thus, there are two sets of "pointers" that identify in the one dimensional arrays where information on the optimal subregions of each i begins and where the R5 data for each i ends. The information is stored sequentially for each i in reverse order from n-4 to 1. This method of ordered storage results in a reduction in the number of page faults on virtual memory computers (i.e., VAX 11/780) and allows for greater vectorization of DO loops on CRAY computers; in both cases, drastically reducing the execution (CPU) time.

D. The algorithm

In the algorithm that follows, the evaluation of $S_{i,j}$ for each subsequence leading to the optimal fold proceeds in three stages. During the first stage, it is determined whether or not a permissible base-pair can form. If so, then $E(C_{i,j})$ is computed and $C_{i,j}$ determined. In the second stage, $E(O_{i,j})$ is calculated if an open multistem structure can be formed from two non-overlapping optimal subregions. In the third stage, $E(S_{i,j})$ is evaluated for every [i,j] by the recursion, equation (1).

This algorithm is similar to those of Zuker and Stiegler (7) and Sankoff et al.(9), but does differ in several important ways. We have introduced procedures that, along with the ability to identify alternate structures, make the algorithm computationally more efficient in its evaluation of multistem and loop secondary structures. These procedures are (1) the use of pointers that identify the base-pair(s) that are acessible from a closing base-pair,

(2) the use of vectors to order the search, (3) limiting the search procedure and (4) limiting the information kept in dynamic memory. These procedures will be illustrated in the following sections.

(1) The Evaluation of $C_{i,j}$ and $C'_{i,j}$

(a) The pointers

Pointers are used in tracing back the optimal and alternate structures following the minimization recursion. Every base-pair has associated with it a pointer which identifies the other base-pair(s) that together close a secondary structural unit (i.e., an interior loop or closed multistem loop). There are two sets of pointers. One set identifies only those base-pair(s) that form the optimal closed structures ($C_{i,j}$) for each subsequence in which nucleotides i and j form a base-pair. The other set identifies the base-pairs that form the alternate optimal and suboptimal closed structures ($C'_{i,j}$).

(b) The Recursion

The free energy of $C_{i,j}$; the free energy of the alternate closed structures on [i,j], $C'_{i,j}$; and their associated pointers are computed recursively in the following manner.

For p=1, $E(p1) = e_1(u^r)$

$\text{pointer}(p1) = 0$ (5a)

For p=2, $E(p2) = \min \{E(C_{s,t}) + e_2(u^r)\}$

$\text{pointer}(p2) = C_{s,t}$ (5b)

For $p \geq 3$, $E(p3) = \min \left\{ E(S_{i+1,h}) + E(S_{h',j-1}) + e_3(u^r) \right\}$,

where $i+6<h<h'<j-6$ (5c)

$\text{pointer}(p3) = \left\{ 0_{s,t} \text{ or } 0'_{s,t} \right\}$

So that,

$$E(C_{i,j}) = \min \begin{cases} E(p1) \\ E(p2) \\ E(p3) \end{cases} \quad (5d)$$

$$\text{pointer}(C_{i,j}) = \begin{cases} \text{pointer}(p1), \text{ or} \\ \text{pointer}(p2), \text{ or} \\ \text{pointer}(p3) \end{cases}$$

The pointers identifying the alternate closed structures are assigned during the recursion steps as improvements of E(p2), E(p3), and $E(C_{i,j})$ are computed.

(c) Optimization Vectors

The non-overlapping optimal subregions that form the multistem loops (equation 5c) are identified by the vectors "S5" and "S3". We have already detailed how these vectors are created and accessed in section C. In a

similiar manner, we make use of a vector C3(E,C,i) to facilitate the back tracking necessary to determine E(p2) for interior and bulge loops (equation 5b). This vector identifies only those accessible closed structures that can form permissible structures with a closing base-pair i*j. The elements of this vector are assigned as each $C_{i,j}$ is determined in the following manner:

$$C3(C) = C_{i,j}, \text{ if } C_{i+1,j-1} \text{ exists.} \quad (6)$$

(2) The Evaluation of $O_{i,j}$ and $O'_{i,j}$

(a) The pointers

Just as each closed structure has associated with it pointers that identify the previous optimal closed structures that make up the optimal and suboptimal closed structures, each open structure has associated with it a similiar set of pointers. One set identifies only those base pairs that close the stems that make up the optimal open structure, $O_{i,j}$. The other set identifies the base pairs that close the stems of the suboptimal open structures, $O'_{i,j}$.

(b) The recursion

The free energy of $O_{i,j}$; the free energy of the alternate open structures on [i,j], $O'_{i,j}$; and their associated pointers are computed recursively in the following manner:

$$E(O_{i,j}) = \min \{E(R_{i,h}) + E(R_{h',j})\} \quad \text{where } i+5<h<h'<j-5 \quad (7)$$

$$\text{pointer}(O_{i,j}) = R_{i,h}, R_{h',j}$$

The pointers that identify the alternate open structures are assigned during the recursion, equation 7, as improvements of $E(O_{i,j})$ are computed.

(c) Optimization Vectors

The vectors "R5" and "R3" identify the non-overlapping optimal subregions that form the open multistem structures (equation 7). How these vectors are created and accessed has been detailed in Section C.

(3) Computational Considerations

The computational time for the evaluation of $S_{1,n}$ and the alternate structures grows very quickly as a function of n. The number of alternate structures that are evaluated, even for short sequences, is very large. Therefore, it is necessary to incorporate several limits into the algorithm, that do not sacrifice the optimality of the results, to decrease their computational time.

The thermodynamic data for bulge and interior loops may have an uncertainity of as much as 2 kcal. So, only information on those bulge and interior loop alternate structures that are within 5 kcal (more than double

their possible uncertainity) of each optimum $C_{i,j}$ are written on disk.

Currently, there is not any thermodynamic data available on the effect of unpaired nucleotides on the stability of closed and open multistem structures. The alternate multistem structures are more globally different structures than the alternate loop structures. The alternate loop structures are variations of local structure, whereas alternate multistem structures are variations of global structure. Given these facts, information on alternate multistem structures that are within at least 5 kcal and as high as 15 kcal of the optimum free energy of each [i,j] is written on disk.

Limiting the search range for alternate structures reduces both the number of computations needed to determine them and the amount of disk space needed to store them. However, by setting a limit of 30 unpaired nucleotides permitted for interior and bulge loops we can reduce the amount of information needed in dynamic memory as well. With this limit, only the optimal closed structures that are within a 30 nucleotide separation from the 3' nucleotide under consideration need to be kept in dynamic memory. This information is stored in the vector "C3" (equation 6) and is continually updated as each additional 3' nucleotide is considered. Furthermore, the value of 30 is consistent with the search range of 5 kcal for these structures.

E. The Traceback

We will now outline a general procedure to determine the major RNA secondary structures corresponding to the lowest distinct free energy minima. We call these major minima secondary structures. Following the completion of the minimization procedure, four files (which were written on disk) have been created. These four files contain information regarding the $C_{i,j}$, $O_{i,j}$, and $O'_{i,j}$ secondary structures, and the $C'_{i,j}$'s, the alternate branches of each $C_{i,j}$. From these four files, the major minima secondary structures within the search range from the optimum free energy can be traced back in the following manner:

(1) Compile a list of structures to traceback

All structures whose free energy are within a given search range of the optimum free energy are combined into a list of structures to be traced back. The elements of the list are ordered from the most to the least stable structure. The list contains information regarding the type of structure for each element (closed or open), its free energy, and what the initial base pair(s) are for each stem(s). This list is compiled from the information on

the $C_{i,j}$, $O_{i,j}$, and $O'_{i,j}$ files. These files contain information on all the optimal closed structures, optimal open multistem structures, and all the alternate open multistem structures that have been calculated.

(2) Traceback the Secondary Structures

All the $C_{i,j}$ data (free energy, 5' and 3' nucleotides, and pointers), along with the list of structures to be traced back, and the $C'_{i,j}$ data, are used to traceback the optimal, alternate optimal and suboptimal structures.

(a) Follow the pointers

After the type of structure and initial base pair(s) are obtained from the list of structures, the traceback procedure begins. The pointer(s) of the initial base pair(s) are followed to the next base pair, whose pointer(s) are followed to the next, and so on until all the base pairs of an RNA secondary structure are identified.

(b) Select the Major Minima Structures

After the optimal structure has been identified, selection of subsequent major minima structures is made on the basis of their structural difference from previously identified (lower energy) major minima structures. Structural difference is determined by the following equation:

$$\% = \frac{\sum_{s=1}^{m} D(n,s) \,/\, M(n,s) + D(n,s)}{m} \times 100$$

where,

D(n,s) = the number of base pairs in the new structure, n, that are different than major minima structure, s;

M(n,s) = the number of base pairs that are common to structure n and s;

m = the number of major minima structures previously identified.

When this percentage is greater than 15%, a new major minima structure has been identified. Simply, this means that 15% or more of the new structure's base pairs must be different from all of the previous major minima structures when a one-to-one analysis is performed.

(c) Check for Alternate Branches

When a closed structure is identified as a major minima structure, each base pair of the closing stem (helical regions + interior and bulge loops) is checked against the $C'_{i,j}$ list for possible alternate branches. The alternate branches are substituted for the optimal ones and their free energy

determined. All of the alternate structures whose free energy are found to be within the search range are then added to the list of structures to be traced back according to their stability.

Step (2) is then repeated until a given number of structures has been traced back or until the list has been exhausted.

RESULTS

To illustrate the utility of our algorithm, we have folded the 414 nucleotide intervening sequence (IVS) of Tetrahymena. This is the same sequence folded by Cech et al.(24) using the program of Zuker and Stiegler (7). The secondary structures were calculated using the free energy values found in Jacobson et al.(20); except that the destabilizing effect, $e_3(u^r)$, due to the number of unpaired nucleotides, r, for closed multistem loops were calculated in the following manner: $e_3(u^r) = r \times 0.2$ kcal.

Ten major minima secondary structures of IVS were determined. The optimal structure (not shown) had a calculated free energy of -145.9 kcal/mol. A structure similiar to that identified by Cech et al.(24) was found to be the fourth lowest-energy major minima structure, ΔG = -142.4 kcal/mol. (The free energy of the Cech structure, Fig. 2 of ref. 24, calculated by the above energy for closed multistem loops gives -139.3 kcal/mol. Cech et al. reported -131.7 kcal/mol using a different energy.) Our structure, however, was obtained without any constraints on what bases should be unpaired or followed by unpaired bases. Cech et al. obtained their structure only after the RNase T1 cleavage data were included in the calculation.

The major minima secondary structures were examined to determine which one or ones were consistent with the enzymatic cleavage data (24), phylogenetic data (25-27), and the insertion/deletion data (28). The structure proposed by Cech et al. has been shown to be consistent with most but not all of these data (24,28). The sixth lowest-energy major minima structure, shown in Fig. 2, was found to be consistent with all these data. It has a ΔG = -141.7 kcal/mol, 4.2 kcal higher in energy than the optimal structure and 2.4 kcal lower in energy than the Cech structure (ΔG=-139.3 kcal/mol). Other major minima structures were found that were as (or nearly as) consistent with the data as the Cech structure. The structure shown in Fig. 2 was selected because its secondary structure has functional implications.

The major minima structures may reveal unique local and/or core secondary structural features that are suggestive of function, while other regions of their structure may be inconsistent with experimental or phylogenetic data.

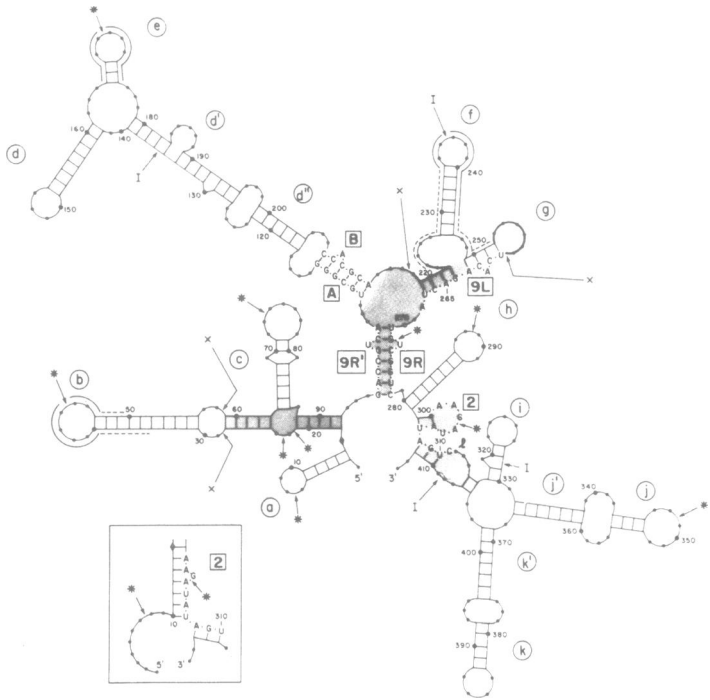

Figure 2. The predicted Secondary Structure of the IVS of T. thermophila rRNA. (*) RNase T1 cleavage sites; data from Cech et al.(24). (I) Locations of insertions generated by recombinant DNA manipulations, (————) regions which can be deleted with little effect on excision activity, (- - - -) regions whose deletion reduces but does no eliminate activity, (X) endpoints of deletions for which no activity can be detected; data from Price et al.(28). Sequences are given for those that are conserved in sequence or position between nuclear rRNA introns and mitochondrial introns. (♦) Site of the addition of the 3' end to form circular IVS with release of hairpin a.

Such a partially consistent structure can provide a starting point for a further search for alternate structure or modification. We have developed an interactive program that for a given sequence region will find alternate stems, locate alternate multistem structures, or eliminate a stem altogether allowing its separated strands to reform alternate local structures. All of these possibilities are easily accomplished without further calculations; all the information needed is found in the four files ($C_{i,j}$, $C'_{i,j}$, $O_{i,j}$, and $O'_{i,j}$) written previously on disk. Shown in Fig. 2 is a structure generated from such a procedure.

In Fig. 2, the RNase T1 cleavage sites (indicated by an asterisk followed

by an arrow), the insertion (I followed by an arrow), and deletion (X followed by an arrow) data have been mapped on the structure. The shaded areas highlight the stems and loops that are not found in the Cech structure. The stems and hairpins that are found in both structures are labelled with the same letters (circled) used by Price et al.(28). The boxed numbers and letters next to listed sequences denote the sequences that are conserved either in position or nucleotide sequence between nuclear rRNA introns and mitochondrial introns.

The original major minima structure contained the closing stem shown in the inset of Fig. 2. This stem was broken and the separated strands formed stem a and the stem formed from the conserved sequence 2. This structure has a ΔG=-135.1 kcal/mol; 4.2 kcal higher in energy than the Cech structure, 6.6 kcal higher in energy than the original major minima structure it was derived from.

There are two functional domains present in this structure. The first of these is the domain closed by the base-paired stem between sequences 9R and 9R'. There is genetic evidence for the existence of this stem. However, in the Cech structure these sequences are predominantly single stranded. Within this domain, the structure formed by the base paired region of the highly conserved sequence 9L is most notable. The nucleotide positions that are in bold print in this region are the nucleotides that Price et al.(28) have identified as being involved in some structure required for autoexcision activity.

The deletion derivatives of Price et al.(28) which retain activity have one common structure with the one shown in Fig. 2. Namely, a stable structure closed by a base-pair stem formed with nucleotides 264 to 257 of 9L followed by a loop on the 5' side and another stem, formed from the 9L nucleotides 259 to 263, which close a hairpin in which the sequence 255 to 258 (CAGU) is part of. These four nucleotides are among those indicated by Price et al. as having some functional significance. It may be that part of or all of this tetranucleotide is involved in some tertiary interaction required for activity. Furthermore, none of the deletion derivatives in which activity is lost can be folded into this structure.

The second of the functional domains occurs at the 3' end. The existence of stem 1 has implications for the mechanism by which the 3' splice site is chosen and how the active site for cyclization is positioned. It occurs only 2 nucleotides upstream from the 3' splice site and could serve to help identify the 3' spice site and possibly position the 3' hydroxyl of guanosine

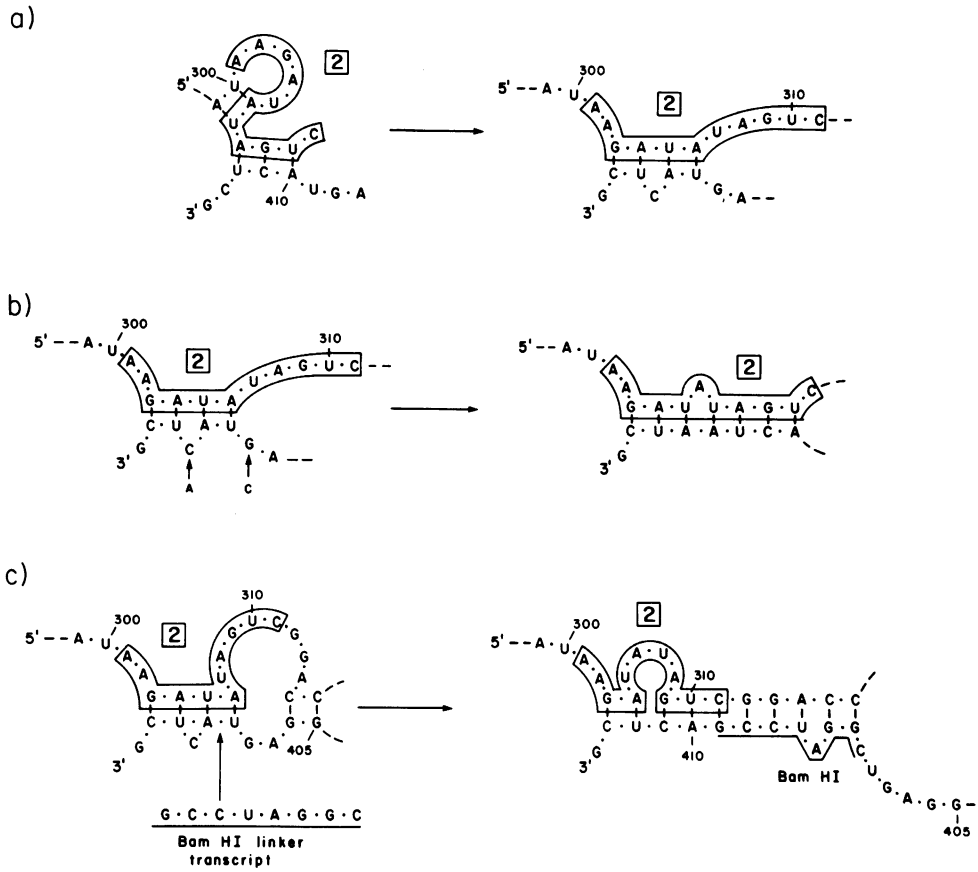

Figure 3. a) The formation of a more stable 1 stem. b) The single base substitutions which give rise to the 1 stem of T. pigmentosa. c) The extended 1 stem formed as a result of insertion of the Bam HI transcript into the IVS of T. thermophila.

residue 414 for attack on the bond between nucleotides 15 and 16 (see diamond at nucleotide 16 in Fig. 2).

The hairpin formed from part of the conserved sequence 2 which is just upstream from the 5' end of this stem is unstable (ΔG=2.3 kcal). We, therefore, searched for a stable alternate stem that would close the i, j, and k stems that was composed of these nucleotides. The alternate stem shown in Fig. 3a was found to be more stable by 1.6 kcal than the structure for this same region shown in Fig. 2. This stem is only one nucleotide upstream from the 3' splice site. The existence of this stem is supported by the single

base changes at positions 398 and 411 of the IVS of T. pigmentosa, resulting in the very stable stem shown in Fig. 3b. The insertion of the Bam HI linker (CGGATCCG) at position 409 does not eliminate splicing acitivity. Shown in Fig. 3c is the extended 1 stem that can be formed between the linker and 1 stem sequences; thus, the structure is preserved.

CONCLUSION

To date, dynamic programming algorithms that predicted RNA secondary structure could only solve for one optimal structure for each RNA sequence studied. To obtain alternate structures, it was necessary to do multiple runs using different constraints to identify possible alternate structures of functional significance. We have described a dynamic programming method that can determine the optimal and alternate suboptimal secondary structures in a single computational run. The power and utility of this method has been demonstrated with the folding of the IVS rRNA sequence of Tetrahymena. In this example, we show how, starting with one of the major minima structures, a secondary structure of possible biological significance can be constructed. Two functional domains are identified in this structure. Within one domain, closed by a stem formed from the conserved sequences 9R and 9R', a hairpin is required for excision of the IVS. Also, the existence of a 3' domain in this structure has implications for the mechanism by which the 3' splice site is chosen.

The method is computationally more efficent and requires less dynamic memory than other procedures that have appeared in the literature. However, the overall computer time and memory requirements depend on the amount of information that must be saved. The search range for alternate structures above the optimum will determine those requirements. The factors which influence the search range will be discussed elsewhere.

The procedures, presented here, are a powerful tool for the biochemist to use along with other experimental methods to elucidate the two dimensional structure of an RNA. Data regarding specific double-or-single-stranded regions from chemical modification or enzymatic studies and data from phylogenetic studies can be incorporated into the minimization procedure by use of appropriate bonus and penalty weights in a manner described by Zuker and Stiegler (7). However, it is possible, without constraints on the minimization procedure, to search for structures (consistent with those data) in the traceback procedure. Furthermore, by modification of the traceback procedure, each secondary structure can be represented by a unique sequence

and/or tree structure; so that, searches for specific secondary structural patterns or structures (e.g., cloverleafs) among the alternate structures is possible. The details of these and other studies will be presented elsewhere.

The programs for these procedures are writtern in FORTRAN for a VAX 11/780 running under VMS and in CRAY fortran, CFT, for a CRAY-1-S and CRAY X-MP running under CTSS at the MFE Computer Center at the Lawerence Livermore National Laboratory.

ACKNOWLEDGEMENTS

This work was supported in part by a grant from the National Institutes of Health (GM10840) and by the U.S. Department of Energy Office of Energy Research under Contract 03-82ER60090.000.

REFERENCES

1. Tinoco, I., Uhlenbeck, O.C., and Levine, M.D. (1971) Nature 230, 362-367.
2. Pipas, J.M. and McMahon, J.E. (1975) Proc.Natl.Acad.Sci. 72, 2017-2021.
3. Waterman, M.S. and Smith, T.F. (1978) Mathematical Biosciences 42, 257-266.
4. Studnicka, G.M., Rohn, G.M., Cummings, I.W., and Salser, W.A. (1978) Nucleic Acids Research 5, 3365-3387.
5. Nussinov, R., Piecznik, G., Grigg, J.R., and Kleitman, D.J. (1978) SIAM Journal of Applied Mathematics 35, 68-82.
6. Nussinov, R. and Jacobson, A. (1980) Proc.Natl.Acad.Sci. 77, 6309-6313.
7. Zuker, M. and Stiegler, P. (1981) Nucleic Acids Research 9, 133-148.
8. Dumas, J.P. and Ninio, J. (1982) Nucleic Acids Research 10, 197-206.
9. Sankoff, D., Kruskal, J.B., Mainville, S., and Cedergreen, R.J. (1983) in Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, Sankoff, D. and Kruskal,J.B. Eds.,pp. 93-120, Addison-Wesley, New York.
10. Uhlenbeck, O.C., Borer, P.N., Dengler, B., and Tinoco, I. (1973) J.Mol.Biol. 73, 483-496.
11. Gralla, J. and Crothers, D.M. (1973) J.Mol.Biol. 73, 497-511.
12. Gralla, J. and Crothers, D.M. (1973) J.Mol.Biol. 78, 301-319.
13. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M., and Gralla, J. (1973) Nature New Biol. 246, 40-41.
14. Salser, W. (1977) Cold Spring Harbor Symp.Quant.Biol. 42, 985-1002.
15. Quigley, G.J., Gehrke, L., Roth, D.A., and Auron, P.E. (1984) Nucleic Acids Research 12, 347-366.
16. 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) Nucleic Acids Research 9, 6167-6189.
17. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B.S., and Stormo, G. (1981) Ann.Rev.Microbiol. 35, 365-403.
18. von Gabin, A., Belasco, J.G., Schottel, J.L., Chang, A.C.Y., and Cohen, S. (1983) Proc.Natl.Acad.Sci. 80, 653-657.
19. Kolter, R. and Yanofsky, C. (1982) Ann.Rev.Genet. 16, 113-134.
20. Jacobson, A.B., Good, L., Simonetti, J. and Zuker, M. (1984) Nucleic Acids Research 12, 45-52.
21. Comay, E., Nussinov, R., and Comay, O. (1984) Nucleic Acids Research 12, 53-66.

22. Martinez, H. (1984) Nucleic Acids Research 12, 323-334.
23. Waterman, M.S. (1983) Proc.Natl.Acad.Sci.USA 80, 3123-3124.
24. Cech, T.R., Tanner, N.K., Tinoco, Jr., I., Weir, B.R., Zuker, M., and Perlman, P.S. (1983) Proc.Natl.Acad.Sci. 80, 3903-3907.
25. Michel, F., Jacquier, A., and Dujon, B. (1982) Biochmie 64, 867-881.
26. Waring, R.B., Scazzocchio, C., Brown, T.A., and Davies, R.W. (1983) 163, 595-605.
27. Michel, F. and Dujon, B. (1983) EMBO Journal 2, 33-38.
28. Price, J.V., Kieft, G.L., Kent, J.R., Sievers, E.L., and Cech, T.R. (1985) Nucleic Acids Research 13, 1871-1889.