Volume 5 Number 9 September 1978 Nucleic Acids Research

Computer method for predicting the secondary structure of single-stranded RNA

Gary M.Studnicka, Georgia M.Rahn, Ian W.Cummings and Winston A.Salser

Molecular Biology Institute, and Department of Biology, University of California, Los Angeles, CA 90024, USA

Received 19 June 1978

ABSTRACT

We present a computer method utilizing published values for base pairing energies to compute the most energetically favorable secondary structure of an RNA from its primary nucleotide sequence. After listing all possible double-helical regions, every pair of mutually incompatible regions (whose nucleotides overlap) is examined to determine whether parts of those two regions can be combined by branch migration to form a pair of compatible new subregions which together are more stable than either of the original regions separately. These subregions are added to the list of base pairing regions which will compete to form the best overall structure. Then, a 'hyperstructure matrix' is generated, containing the unique topological relationship between every pair of regions. We have shown that the best structure can be chosen directly from this matrix, without the necessity of creating and examining every possible secondary structure. We have included the results from our solution of the 5S rRNA of the cyanobacterium Anacystis nidulans as an example of our program's capabilities.

INTRODUCTION

Several authors (3,4,6,7,8,13) have published results of experiments designed to measure the free energy contributions of certain configurations of polynucleotides free in defined aqueous solution at room temperature. Although there are many features of secondary structure whose effects on free energy have not yet been measured (e.g., the sequence and base composition of single-stranded loops; the location of a base pair near to or far from either end of a double-helical region; the shape and size of multiply branched internal loops), enough data have compiled to reveal strong and consistent patterns in the relationship between a given primary nucleotide sequence and the energetic stability of the various partially base paired forms it can assume. It is therefore desirable to apply the rules for secondary structure formation, which have been suggested by the data gathered thus far, to known RNA sequences of biological interest in order to test their ability to explain known relationships between structure and function, and to predict other such relationships. This sort of investigation may

lead to hypotheses concerning the utilization of various features of secondary structure for purposes of cellular genetic control, and may also help to explain the curious absence of silent mutations in relatively long stretches of some mRNA's (9).

Although the current rules for determining the free energy of a given secondary structure are straightforward and well-defined (5,6,7,8,13), it is very tedious to carry out the computations by hand even for relatively short sequences such as tRNA's, and it is practically impossible for mRNA's. Therefore, the aid of the computer has been sought to increase the speed and precision of these calculations. Previous computer programs (1,4,14,15) have been written to apply these rules within a framework of certain assumptions and limitations. They have been shown to be capable of selecting structures identical with or very similar to those obtained through experimental means such as X-ray crystallography, especially in the case of various tRNA's (1).

While we have not found that previous computer programs make any serious errors, we have devised an algorithm which takes into account a greater number of possible pairing interactions and which is not constrained by as many limiting assumptions. Also, our algorithm approaches the solution in a completely different way, utilizing certain topological properties of all secondary structures (explained more fully in Methods), and thus is able to choose the best secondary structure directly without actually having to create and examine every possible permutation. This innovation is perhaps not very important when only small sequences are considered, since the number of possible structures is also small, but it is much more significant when mRNA's or larger molecules are examined because the number of possible structures increases as an exponential function of sequence length.

No computer method thus far, including our own, makes allowances for tertiary or quaternary nucleotide interactions, nor for the effects of proteins which may bind to the polynucleotide sequence as it occurs in vivo. Also, previous computer algorithms have been designed not to select any structures which allow bases inside a hairpin loop to pair with bases outside that loop. No serious attempt has ever been made either to prove or to disprove the validity of these 'knotted' and 'pseudoknotted' structures (described more fully in Methods and in Figure 2), but we too have chosen to ignore them.

Our algorithm has improved upon previous methods by including in its analysis the possible interaction through branch migration of pairs of overlapping double-helical regions. Because two different regions which

have some of their nucleotides in common cannot both be members of the same structure, it is possible that forcing the computer to choose either one region or the other may cause it to miss a more energetically stable 'hybrid' structure containing a pair of nonoverlapping subregions formed from fragments of the original two overlapping regions. Previous computer methods have not considered these kinds of interactions. Our program examines all pairs of overlapping regions, choosing in each case the best pair of subregions if they are energetically favored.

METHODS

A. Background and Definitions

As will be described in a later section, we have found a logical system for describing certain topological relationships among double-helical regions of base pairing which facilitates the search for the most stable structure. In the following paragraphs the terms needed to describe these relationships are defined.

Members of a single-stranded polynucleotide sequence (bases) can be totally ordered by giving each one a base number consecutively in ascending order starting with the 5' end of the molecule. Thus for any two bases M and N, the statement M < N means that M is closer to the 5' end of the molecule than N.

Two single-stranded sequences (of the same length) which can combine to form an uninterrupted double helix involving GC, AU, or internal GU base pairs constitute a region. For the sake of the nomenclature, let's call a typical region A. Every region partitions the polynucleotide sequence into exactly five distinct subsequences consisting of two halfregions, one sequence of included bases, and two sequences of excluded bases. The 5' halfregion (designated A5) is base paired to a complementary and antiparallel 3' halfregion (designated A3), which together form the region A. All of the nucleotides which lie between A5 and A3 are the included bases of A, and would form a hairpin loop if left completely unpaired. The remaining nucleotides in the molecule are divided into two groups: those which are more 5' than A5 are the 5' excluded bases, and those which are more 3' than A3 are the 3' excluded bases. The diagram in Figure 1 illustrates these relationships.

Associated with each region A there are three independent values which together uniquely determine that region: 1) A5' is the base number of the 5' end of A5; 2) A3' is the base number of the 3' end of A3; and 3) AL is

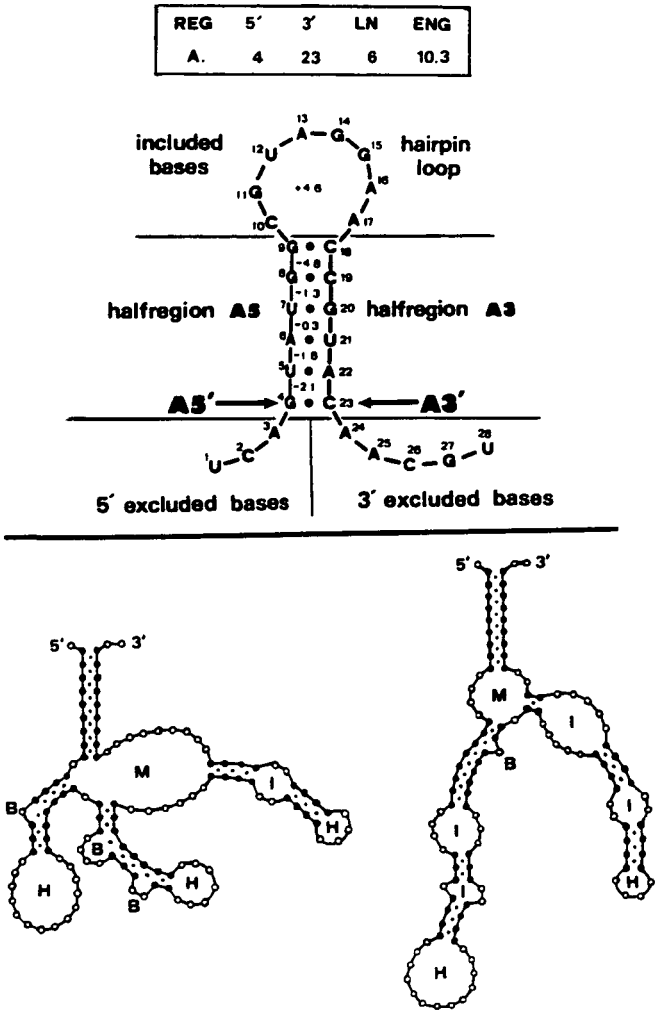

Figure 1. Terminology associated with regions and loops. TOP: Diagram of a typical region of basepairing, with A5' = 4 and A3' = 23, a length of 6 basepairs, and a stabilizing energy of -10.3 kcal/mole. The pair-of-pairs stacking energies are shown between each of the basepairs. BOTTOM: Schematic diagrams of the secondary structure of Structures 1 and 2 (Figure 6), with each loop labelled according to its type: H = hairpin loop; I = internal loop; B = bulge loop; M = multiply branched loop. Bases in double-helical regions are shown as solid circles, while single-stranded bases are open circles. Base stacking is represented by a small dot between basepairs.

the length of each of the halfregions. From these numbers, several other parameters of region A can be derived.

Another property of every region is that the sum of the base numbers of any base pair in the region (in particular A5' + A3') equals a unique

value for that region, called its group number (designated AG for region A). The regions are only partially ordered by their group numbers, since it is possible for two different regions to have the same group number. These group numbers are used in the branch migration part of the program, which is discussed below.

For any pair of regions A and B, region A overlaps region B if and only if there is at least one base which is a member of both regions. Otherwise A and B are nonoverlapping regions. There are exactly three possible mutually exclusive topological relationships between any two nonoverlapping regions A and B (named such that A5 < B5): 1) A includes B (A5 < B5 < B3 < A3) when all of the bases in each of the halfregions of B (B5 and B3) are also included bases of A; 2) A excludes B (A5 < A3 < B5 < B3) when all of the bases in B5 and B3 are also 3' excluded bases of A; 3) A knots B (A5 < B5 < A3 < B3) when the bases in B5 are included bases of A, and those in B3 are 3' excluded bases of A. Figure 2 illustrates these relationships between pairs of regions, and distinguishes 'true knots' from 'pseudoknots'.

A structure is a set of regions such that no region overlaps any other region in the structure. A structure containing at least one pair of regions which knot each other is a knotted structure, while one which has no knots is an orthodox structure. Thus, it follows from the definitions that for every pair of regions A and B (with A5 < B5) in an orthodox structure, either A includes B or A excludes B.

Associated with the base sequence of any single-stranded molecule is a unique set of all possible regions of double-helical base pairing. Every possible structure is a subset of this region set, but not every subset is a structure. The regions in the region set can be partially ordered by inclusion and also independently partially ordered by exclusion. Our method of searching for the most stable structure involves performing logical manipulations on the complete set of inclusion and exclusion relationships between all possible pairs of regions. We call this set a hyperstructure. Examples of hyperstructures will be given in detail below (and in Figure 4), where it will be demonstrated that every possible structure is identified with a unique 'pathway' through the hyperstructure. A hyperstructure which contains only inclusion relations between its member regions is called a simple inclusion hyperstructure, while one which contains only exclusion relations is a simple exclusion hyperstructure. Most hyperstructures contain both inclusion and exclusion relations, and are called complex hyperstructures.

The problem of selecting the best structure is thus reduced to selecting

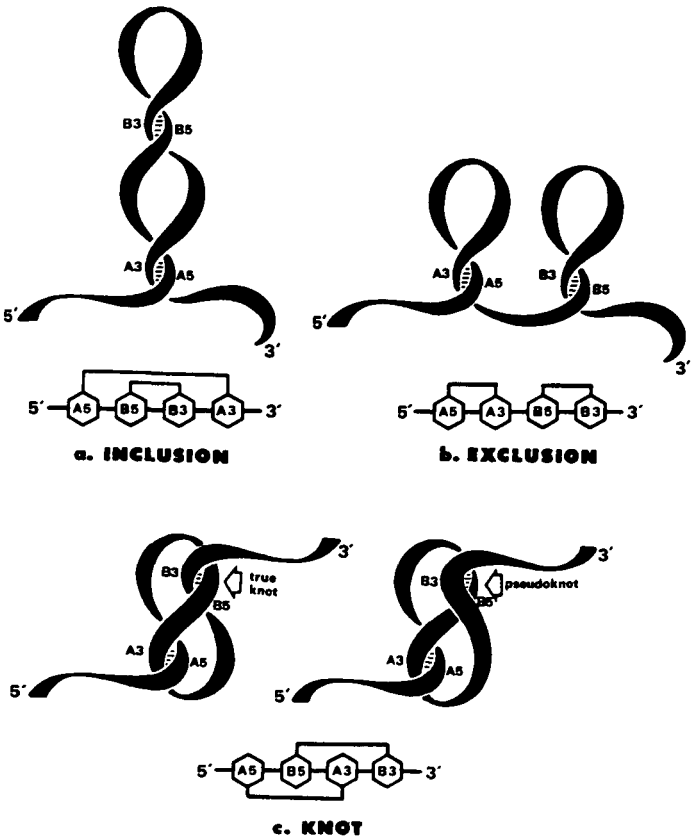

Figure 2. Perspective drawings and schematic representations of sequences of halfregions, indicating the three possible topological relationships between any two double-helical regions whose member nucleotides do not overlap. These relationships are defined to be antisymmetric (e.g., A includes B implies that B does not include A), and are always written so that A5 < B5 to preserve this directionality. (a) Region A includes region B (or B is included by A). (b) Region A excludes region B (or B is excluded by A). (c) Region A knots region B (or B is knotted by A). The diagram at the left in (c) is a 'true knot', in which the excluded bases of one region pass through the hairpin loop of the other region to form hydrogen bonds. The diagram at the right is a 'pseudoknot', in which the hydrogen bonds of region B are formed merely by twisting of the single strands without actually passing through the loop. Perhaps pseudoknots should be considered tertiary interactions involving the folding of an already formed secondary structure. True knots, however, would present an especially difficult problem to a ribosome attempting to translate such a knotted region!

the pathway through the hyperstructure which has the lowest energy. We have shown that it is not necessary to actually create and examine every possible structure in order to choose the best one. Instead, it is adequate to choose

the best pathway locally in the neighborhood of each region in the hyperstructure, and to use this information to set local pathway indicators (according to rules we will describe below) which show the energy gained by choice of each path. It is then possible to begin at the 'initial side' of the hyperstructure and to follow the best path immediately by inspection of these previously set local indicators until that path ends at the 'terminal side' of the hyperstructure. The set of regions encountered while traversing that pathway constitutes the best predicted structure. In a similar manner, other pathways can be followed to generate less favorable structures.

B. Choosing Regions of Uninterrupted Double-Helical Base Pairing

Every base in the primary sequence is compared with every other base to determine whether they can form a valid hydrogen-bonding base pair. In addition to the two classical Watson-Crick base pairs (AU and GC), the GU pair is allowed whenever it is not the first or last base pair in a region. Also, regions which enclose relatively small hairpin loops (consisting of fewer than six bases) are examined to determine whether a more energetically favorable configuration of region + loop can be formed by opening one or more base pairs adjacent to the hairpin loop. Usually this shorter region replaces the original region, but in some cases it must be included in the complete list of regions along with the first one. This process of opening hairpins to achieve the most energetically favorable configuration always ensures that no region can enclose a hairpin loop of fewer than three bases, and in rare cases results in the elimination of all base pairs in a region because the destabilizing energy of its hairpin loop totally overwhelms the stabilizing energy of the double-helical region.

All of these primary regions (consisting of two or more consecutive basepairs) are printed out, and the user then has the option either to keep all of them or to select only certain ones to participate in forming the best overall structure. This selection process may be repeated as often as desired, and any user-defined functions of the locations, lengths, or energies of the regions (or any combination of these parameters) may be employed to determine which regions are to be kept. Thus it is possible to obtain a partial solution for a very large RNA sequence whose complete solution would require prohibitive amounts of time or space for computation.

C. Creating Pairs of New Subregions by Branch Migration

Two regions which overlap cannot both be members of the same structure. But often a pair of nonoverlapping subregions (each of which is a subsequence

of one of the original pair of overlapping regions) will be more stable together than either member of the original pair taken separately. For any pair of overlapping regions A and B (named such that AG > BG) there are exactly three independent elementary overlap relations: 1) 5-5 overlap, where A5 overlaps B5; 2) 3-5 overlap, where B3 overlaps A5; and 3) 3-3 overlap, where A3 overlaps B3. One, two, or even all three of the overlap relations may be present in some overlapping region pairs, but because they combine independently each elementary overlap can be solved separately in the more complex cases. Figure 3 shows an example of a pair of regions A and B which are related by 5-5 overlap. In every elementary overlap, the two halfregions which have nucleotides in common interact to form the common strand, while the interfering strand is formed from the other two halfregions which are separated by a sequence of bases called the overlap loop.

Once the general form of this configuration has been established, the position of the overlap loop is systematically varied with respect to the common strand. In each position, the energy of the entire configuration is calculated, including an estimate (adjustable by the user) of the destabilizing energy of the overlap loop. Of course, it is not possible to determine the exact value of this destabilizing influence without knowing in advance the most stable base pairing configuration of the nucleotide sequence of this loop. No attempt is made to evaluate this in the branch migration portion of the program because the purpose of branch migration is not to choose the best ultimate structure (which is done later in the program) but is to select subregions which may compete to form that structure.

The pair of subregions which form the best configuration are saved and added to the complete list of base pairing regions if the energy of the entire configuration is more stable (by a user-specified amount called the 'threshold for branch migration') than that of the most energetic member of the original pair of regions. Before execution of the branch migration portion of the program, the user is requested to set the value of this threshold, whose purpose is to adjust the stringency of selection of new subregions. For large sequences where only a partial solution is carried out, we have found that it increases the efficiency of the program to set the threshold at a value comparable to the energy of the least stable region retained from the primary region table.

Of course, branch migration could be avoided simply by including in the region table every possible subregion (in addition to the larger regions which contain them). However, for every region containing N basepairs, there are

REG	5′	3′	LN	ENG
A.	8	61	7	11.6
B.	10	36	9	11.3

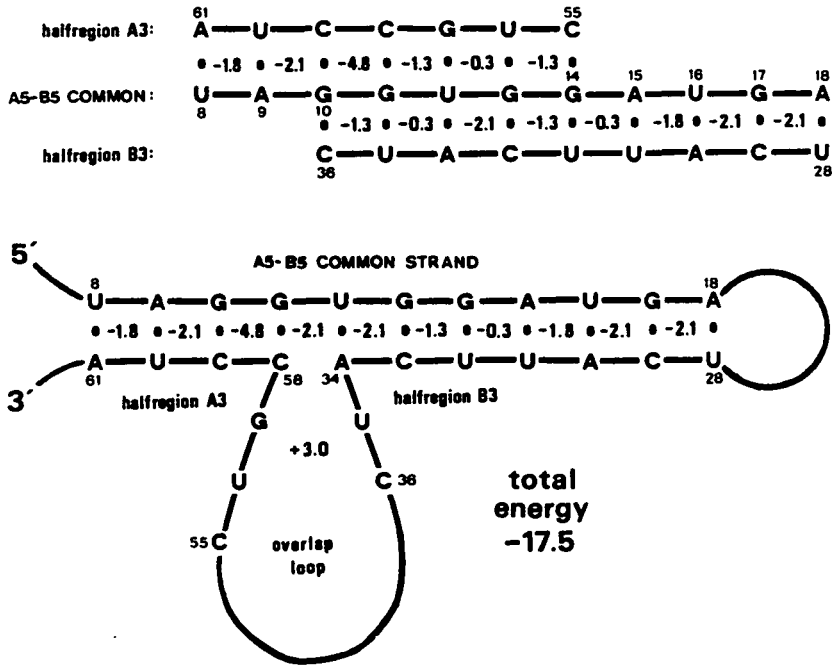

Figure 3: Example of branch migration process for 5-5 overlap. TOP: The 5' and 3' endpoints of regions A and B, their lengths, and energies. Region A's group number AG is 69, while BG is 46. MIDDLE: Diagram of nucleotides in conflict, showing the A5-B5 common strand containing five overlapping bases, and the other two halfregions A3 and B3. Some of the base numbers are shown, and each pair-of-pairs energy is indicated. BOTTOM: The optimum configuration with a total energy of -17.5 kcal/mole (including the +3.0 kcal/mole estimate for the overlap bulge loop) is formed from a four basepair subregion of A and a seven basepair subregion of B, separated by a 23 base overlap loop. Base stacking occurs across the bulge loop, but would not occur across an internal loop. Thus the energy estimate we have used for an internal overlap loop is +1.0 kcal/mole. The -17.5 kcal/mole represents a 5.9 kcal/mole improvement over the -11.6 kcal/mole for region A alone.

$\frac{1}{2}(N^2-N)-1$ additional subregions, and most of those subregions will not participate in the final structures because the larger regions containing them will almost always compete better energetically. For any given nucleotide sequence, the cost of a solution increases roughly as the square (R^2) of the number of regions (R) considered. Thus, our approach greatly reduces the

cost of the calculation by eliminating from consideration a large number of subregions which can be shown to have no chance of occurring in the most stable overall structures.

D. Simple Hyperstructures

Because we have chosen not to consider knotted structures, it follows that the secondary structure of the hairpin loop included by any region can be solved independently of the secondary structure of the 5' and 3' excluded bases associated with that region. This is the fundamental theorem of the logical system we have devised.

Since the inclusion and exclusion relations each partially order the set of regions, the relations between some pairs of regions can be derived by transitivity theorems (2) using another region which lies 'between' them in the hyperstructure. For example, for any three regions A, B, and C, if A includes B and B includes C, then A must include C. The same sort of reasoning applies to exclusion relations. Thus, the entire set of relations between all region pairs can be derived from a smaller subset of local relations between region pairs which are especially 'close together' in the hyperstructure.

Because simple hyperstructures are partially ordered sets, we can define a pathway through the hyperstructure as any set of regions, totally ordered by local relations, which begins at an initial element and ends at a terminal element of the hyperstructure. An initial element is a region which has no other regions 'before' it, and a terminal element is a region with no other regions 'after' it. Before and after, of course, have to do with inclusion and exclusion relations which define order and direction in the hyperstructure.

Figure 4A shows a schematic diagram of a simple exclusion hyperstructure containing seven regions which are interconnected by eight local exclusion relations symbolized by arrows. Every pathway through this hyperstructure is a valid orthodox structure for the corresponding molecule (although not necessarily the best structure). Note that by following only the pathways designated in the hyperstructure, the computer would never postulate any structures containing both regions B and C, because there is no unidirectional exclusion pathway which contains them. This is a useful property of the hyperstructures we define here, since any structure containing both of those regions would be knotted. Thus, it can be shown to be a general property of all simple hyperstructures that, by choosing only those pathways which conform to the transitivity theorems, all knotted structures and all 'impossible' structures (where some regions overlap) are automatically eliminated from consideration.

E. Choosing the Energetically Most Favorable Structure

If the base pairing energy of every region in a simple hyperstructure is known, then the best pathway through that hyperstructure can be selected by applying mathematical induction to the relations between the member regions. Because simple hyperstructures are partially ordered sets, there is always at least one terminal element (2) which, by definition, does not have any double-helical regions 'after' it. The terminal elements are thus the starting points for the induction process, since the secondary structure 'after' them is already known by definition. For simple inclusion hyperstructures the portions of the molecule 'after' the terminal regions are the hairpin loops associated with these regions, while for simple exclusion hyperstructures they are the 3' excluded bases of these regions.

Once the secondary structure (and the associated energy) of any portion of the molecule is known, each cycle of the induction process involves examining all of the regions connected by local relations to those known portions, and calculating the energy of the secondary structure formed by adding each of the new regions to the already computed secondary structure.

Every step in this inductive process increases the length of each subpathway (which started out simply as a terminal element) until it becomes a complete pathway by ending at an initial element. Because the computation proceeds backward along the local relations, much computing time is saved by calculating the energy of each subpathway only once, and then including its energy in the computation later on if a new (longer) subpathway contains the old one.

Whenever there are several subpathways leading away from the same region, the computer sets a local pathway indicator designating the subpathway with the best energy so that this information does not have to be recomputed at the time structures are actually built.

The energies are computed using the data in references 6-9. Our choices of how best to apply these data have been summarized by Salser (reference 9, see the legend to Figure 3 of that paper). The computer program has used these criteria for calculating energies, with one exception: because there are no biochemical data on the appropriate energy contributions of multiply branched loops (loops connecting three or more stems, described in Figure 1), both our program and the Pipas and McMahon program (1) assign zero energy to all such loops regardless of shape, size, or nucleotide composition. The destabilizing effects of these loops are normally taken into account by hand according to the criteria described by Salser (9) once the computer has

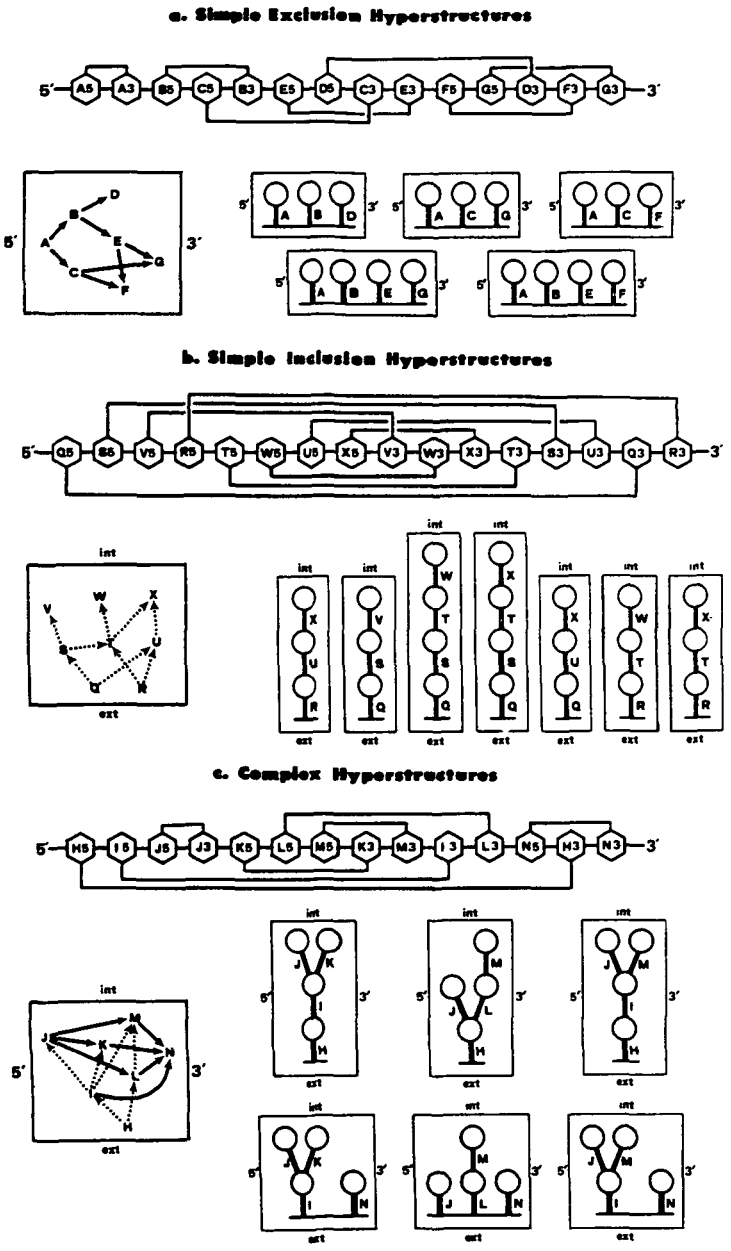

Figure 4. Three sets of schematic diagrams, each consisting of a polynucleotide sequence with labelled halfregions, the hyperstructure corresponding to this sequence, and a set of all the possible structures which can be derived from the given configuration. In each polynucleotide sequence, pairs of halfregions which hydrogen bond to form double-helical regions have been connected by lines to give some idea of how the molecule would have to fold to form these regions.

In the drawings of hyperstructures, the solid lines represent exclusion relations, while the dotted lines represent inclusion relations. Molecular structures are shown schematically with hairpins and internal loops represented as circles, and with helical regions represented as heavy lines with the name of the region immediately adjacent.

(a) Diagram of a polynucleotide sequence containing seven regions named A through G. Because no region includes another in this sequence, their mutual interrelationships can be represented by the simple exclusion hyperstructure to the left, in which eight local exclusion relations (solid arrows) partially order the regions in the hyperstructure from the 5' end of the molecule to the 3' end. Thus, according to the diagram, all of the bases of region A are more 5' than those in region B. Since B excludes E, it is also known by transitivity that A is more 5' than E. The initial element is A, and the terminal elements are D, G, and F. There are five possible pathways (always following local exclusion relations) which connect the initial elements to the terminal elements: 1) ABD; 2) ACG; 3) ACF; 4) ABEG; 5) ABEF. Each of these pathways corresponds to one of the structures shown to the right.

(b) Diagram of a polynucleotide sequence with eight regions named Q through X. Because there are no exclusion relations among these regions, they can be represented by the simple inclusion hyperstructure shown to the left, in which nine local inclusion relations (dotted arrows) partially order the regions in the hyperstructure from the external portions of the molecule (near the 5' and 3' ends) to the internal portions. Thus, all of the bases in region T and in the hairpin loop associated with region T are internal to (i.e., included bases of) region R. Since T includes W, it is known by transitivity that R includes W. The initial elements are Q and R, while the terminal elements are V, W, and X. There are seven possible pathways (following local inclusion relations) which connect initial elements to terminal elements: 1) RUX; 2) QSV; 3) QSTW; 4) QSTX; 5) QUX; 6) RTW; 7) RTX. Each pathway corresponds to a structure.

(c) Diagram of a polynucleotide sequence with seven regions named H through N. Because both inclusion and exclusion relations exist among these regions, they cannot be represented by any simple hyperstructure, but instead correspond to the complex hyperstructure shown to the left. Seven local exclusion relations (solid arrows) partially order the regions from the 5' end of the molecule to the 3' end. Also, six local inclusion relations (dotted arrows) independently partially order the regions from external portions of the molecule to internal portions. Transitivity can be applied to complex hyperstructures in the same manner as with simple hyperstructures. The initial exclusion elements are regions J, I, and H, while the terminal exclusion elements are H and N. (Since region H is not related by exclusion to any of the regions in this hyperstructure, it is both an initial and a terminal exclusion element.) The initial inclusion elements are regions H and N, while the terminal inclusion elements are J, K, M, and N. Six possible structures can be derived from this configuration: 1) HIJK; 2) HJLM; 3) HIJM; 4) IJKN; 5) JLMN; 6) IJMN.

calculated and drawn the secondary structures.

Thus, after all of the possible pathways (and their associated energies) in the hyperstructure have been calculated, the computer examines all of the initial elements and selects the one whose pathway has the best energy. Then by stepping forward along the local relations in the hyperstructure, always choosing the subpathway leading away from each region in accordance with the

previously set local indicators, the set of regions which form the most energetically favorable structure is determined.

For short RNA sequences the number of pathways in the hyperstructure would be small, and it would probably be equally efficient simply to create and examine all of the possible secondary structures and to compare their energies directly. But the number of possible structures increases as an exponential function (2^R) of the number of regions (R), and because the computation required by our algorithm increases approximately as the square (R^2) of the number of regions, it is clear that for large values of R (as would be encountered in the solution of mRNA's) the method of creating and examining every possible permutation would require much more computer time to achieve the same result.

Another feature of our algorithm is that it has been designed not to generate substructures, which are almost always less stable than their corresponding structures, and which can be derived from those structures by opening one or more individual basepairs (or entire regions) in the structure. For every structure containing N basepairs, there are 2^N different possible substructures that can be formed. Since these kinds of computations are easily done by hand once the computer has calculated and drawn the best predicted structures, we felt that it would be inefficient to have the program also consider such a large number of substructures. This has resulted in a substantial increase in the relative morphological diversity of the set of predicted structures which can be generated by our program within a limited amount of computer time.

F. Complex Hyperstructures

As discussed above, the hyperstructures corresponding to most sequences of halfregions contain both exclusion and inclusion relations among their member regions, and thus by definition are complex. Figure 4C shows a schematic diagram of a complex hyperstructure containing seven regions named H through M. Both of the transitivity theorems can be applied to this complex hyperstructure: 1) J excludes K and K excludes N implies that J excludes N; 2) H includes I and I includes K implies that H includes K. There are two additional theorems which apply only to complex hyperstructures: 3) J excludes L and L includes M implies that J excludes M; 4) I includes K and I excludes N implies that K excludes N.

Although complex hyperstructures are solved by an inductive method which is analogous to the backward scanning of simple hyperstructures, the algorithm used by the computer is of necessity somewhat more complicated and involves

the generation of a simple exclusion sub-hyperstructure for every region in the complex hyperstructure. These sub-hyperstructures are interrelated only by inclusion, and therefore they are themselves 'elements' of a simple inclusion hyperstructure. Thus, using the theorems described above, every complex hyperstructure can be made equivalent to some heirarchy of simple hyperstructures, and having done this, the solution is obtained in the same way as with ordinary simple hyperstructures.

RESULTS

To illustrate the capabilities of our program, we have used it to study the possible secondary structures of the 120 nucleotide sequence (10) of the 5S ribosomal RNA of the cyanobacterium Anacystis nidulans. Other sequences have been analyzed by our program, in particular the complete messenger RNA of rabbit alpha hemoglobin, the results of which are discussed elsewhere (11).

The data shown in Figures 5 and 6 represent a 'complete' solution for this sequence, meaning that all 169 of the primary regions and all 25 of the subregions generated by branch migration were saved (for a total of 194 regions consisting of two or more basepairs) and allowed to compete to form the best structures. Because the CPU time (and therefore the cost of computing) required by our algorithm increases roughly as the fifth power of sequence length, it becomes impractical to attempt a complete solution for sequences greater than about 250-300 nucleotides. In such cases (for example with rabbit alpha hemoglobin messenger RNA, which produced 3686 primary base pairing regions), many of the less energetic regions are not included in the first cycle of structure building. Then, two or three long and strong regions which are found to be common to many of the best structures generated from this first cycle are used to partition the nucleotide sequence into domains consisting of the included bases of these regions. Incorporated into our program is a routine which chooses, according to parameters supplied by the user, all the regions whose halfregions are contained within any of the specified domains. A complete solution is then easily obtained for these domains independently, and the regions participating in the best structures generated from these subsequences can then be added to the set of regions from the first attempt, thus generating a more refined overall structure.

This method will arrive at the most stable structure if the first cycle is able to predict two or three strong regions which are actually present in the best structure. However, there is always some chance that even the best regions from a partial solution will not occur in the complete solution, so that the correct convergence will not be obtained. To guard against this, the

user may break the sequence into smaller domains in several alternative ways and ascertain which gives the best structure after the next cycle of refinement.

Our program has been designed to provide great flexibility by permitting the user to use a variety of criteria for the selection of regions. The user may then judge for himself whether the kinds of structures vary significantly as a function of those criteria. Previous computer algorithms have not provided this kind of flexibility, for example the program written by Pipas and McMahon (1) was designed to be capable of a complete solution only for sequences that generate 240 or fewer primary regions consisting of three or more basepairs (no branch migration is carried out), which corresponds to a sequence length of about 150 nucleotides. Although their program is able to give partial solutions for larger sequences, it was written to save only the 240 longest regions in these cases, independent of their energies, and has no provisions for user modification of these selection criteria. Our algorithm is limited only by the size of the computer's memory and by the money which the user is willing to spend for computing time, and thus in principle is capable of giving a complete solution for a molecule of any size. For the rabbit alpha hemoglobin messenger, the partial solution we obtained with the Pipas and McMahon program yielded a best structure of -159.6 kcal/mole, while the most energetic structure from our own partial solution had an energy of -255.6 kcal/mole, representing a 60% improvement. (Upon correction for the energy of multi-branch loops, both values decrease by a similar amount. Our adjusted value is -224.6 kcal/mole (11)).

Figures 5 and 6 show the final output from the complete solution for the 5S rRNA of Anacystis nidulans. Many intermediate results printed by the computer have not been included here, consisting of the table of 169 primary regions, the 25 regions generated by branch migration, the final set of 194 regions, and some information concerning the specific pathways the computer followed to compute the best structures. Once the computer had carried out the computations to assign energies to the subpathways in the hyperstructure (by induction, travelling backward along the local relations and setting the local pathway indicators), the energy of the best structure was then known and printed out. However the actual structure (set of regions) corresponding to that energy was not yet known. After receiving input from the user which indicated how many structures should be computed (in this case 30 were requested), the program proceeded (following the previously set local pathway indicators) to generate the 30 best structures. When this was completed, a

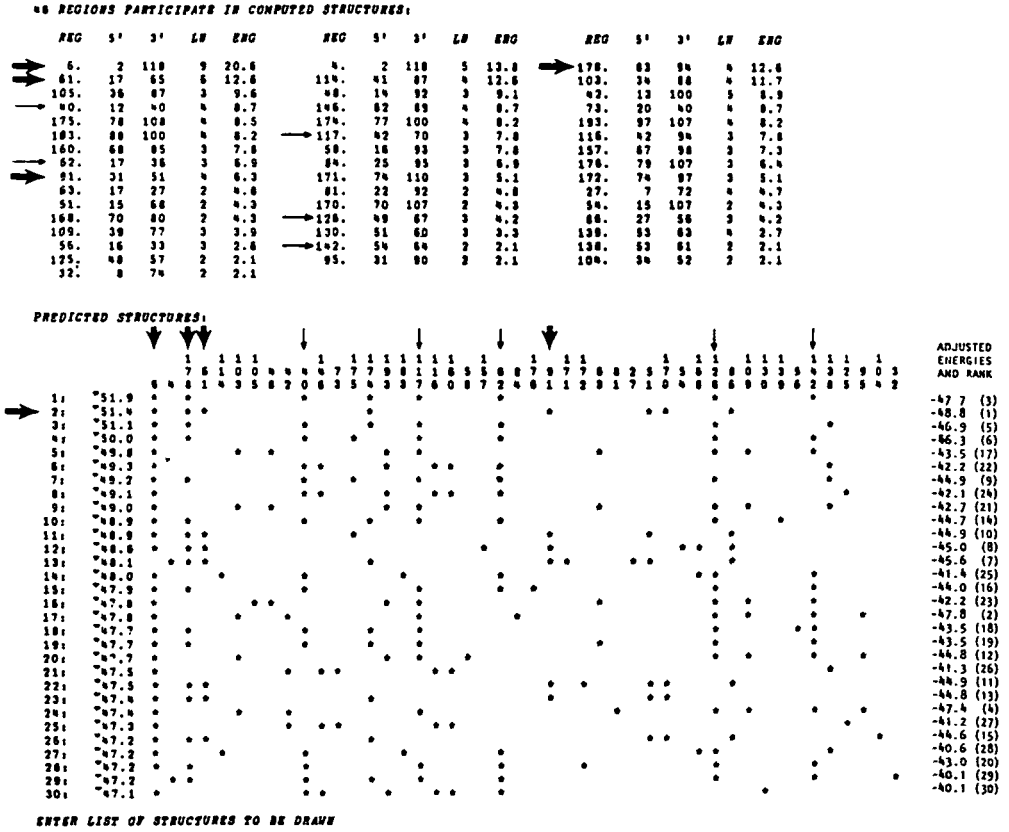

Figure 5. Output from the computer program describing the 30 best predicted structures, with regions corresponding to the Fox and Woese structure and to our Structure 2 emphasized by heavy arrows, and with regions important to our Structure 1 highlighted by thin arrows. TOP: Table of the 46 double-helical regions which actually participate in the 30 predicted structures. BOTTOM: The 30 X 46 matrix, in which the rank of each structure and the energy assigned to it by the computer have been listed down the left side, while the region numbers (written vertically, and corresponding to the table above) have been listed across the top of the matrix. The asterisks indicate which regions are members of each of the 30 predicted structures. In accordance with Salser (9) the free energy value for each of the structures has been adjusted by hand to take into account the destabilizing influences of multi-branch loops (which are described in the text), and these new energies have been listed along the right side of the matrix with their new overall ranks among the 30 structures. Note that Structures 17 and 24 (whose adjusted ranks are now second and fourth) are noteworthy, since they are unbranched structures morphologically quite different from those drawn in Figure 6.

table of the 46 regions which actually participate in the 30 best structures was printed, and a 30 X 46 matrix showing which of the 46 regions are members of each of the 30 predicted structures was also printed (Figure 5). At this

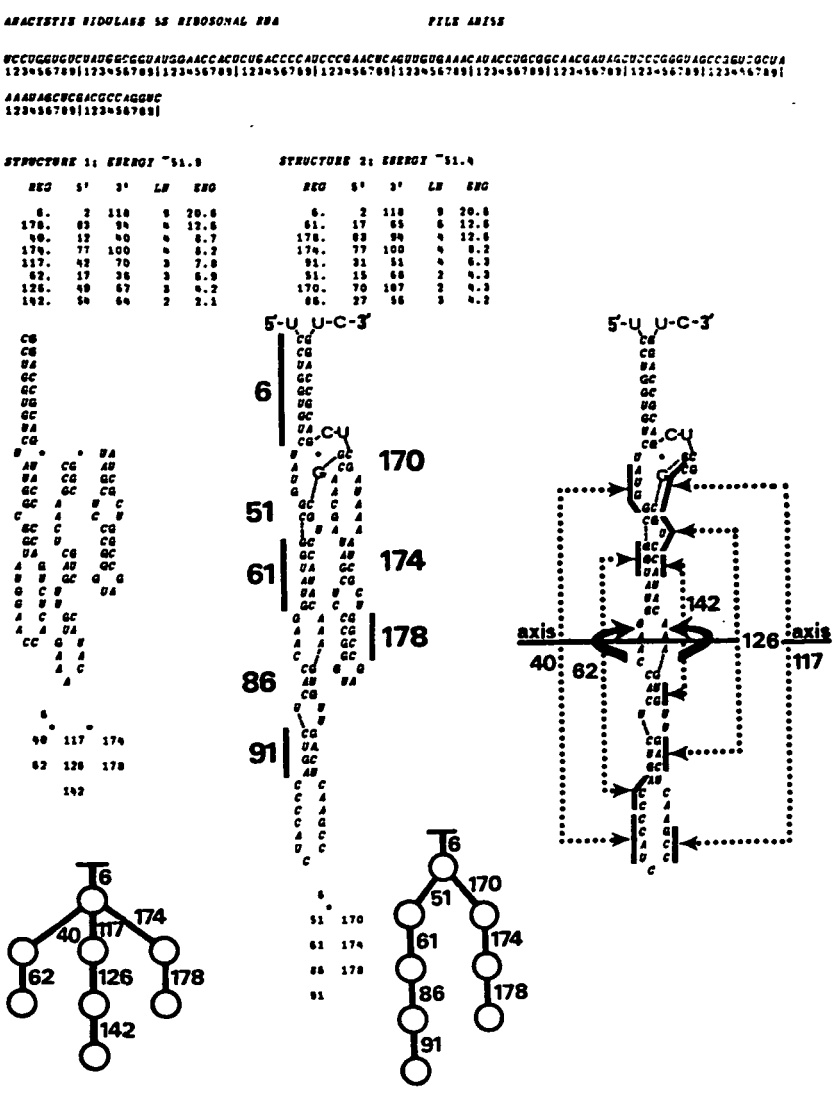

Figure 6. Output from the structure drawing routines of our computer program, and schematic diagrams of the canonical form of each structure. LEFT: Structure 1 exactly as drawn by the computer. CENTER: Structure 2, in which the unpaired bases that were not drawn by the program have been added by hand, along with region numbers (from table above) in bold print. RIGHT: An extra copy of Structure 2 with information added describing a possible mechanism for morphological interconversion between structures similar to 1 and 2. Note that when the destabilizing effects of multi-branched loops were taken into account by hand (according to Salser (9)), Structure 2 in fact became the best overall with an energy of -48.8 kcal/mole, while Structure 1 received -47.7 kcal/mole. Refer to the text for a more complete description of this figure.

point, the computer asked for a list of the structures to be drawn. Because many computed structures are often similar to each other except perhaps for one or two different regions, the user need not wait to see all of them drawn (especially if several hundred structures were requested), but may inspect the displayed matrix to select the ones which are most diverse in morphology. Although in this case all of the 30 structures were drawn by the computer, only Structures 1 and 2 have been included in Figure 6 because of space limitations.

The energy of each structure and its rank among all possible structures (not including substructures) are shown in Figure 6, and a region table for each structure is given. Every region in the region table has five parameters: 1) a region number 'REG' assigned in ascending order of the A5' base numbers; 2) the base number (A5' in Figure 1) of the most 5' nucleotide in the region; 3) the base number of the most 3' base (A3' in Figure 1); 4) the length 'LN' of the region in basepairs; and 5) the stabilizing energy 'ENG' in kcal/mole.

Below each region table the computer has drawn schematic representations of the physical morphology of the folded RNA molecule. The program also has the capability of drawing much larger structures, such as those generated from mRNA sequences. Because it is difficult to represent these structures on paper in a reasonably compact form, some artistic license was taken in our design of the routines used by the computer in drawing structures. All double-helical regions are oriented vertically with the external ends toward the top of the page and the internal ends toward the bottom. Hairpin loops of any size are drawn in their entirety, but internal loops whose unpaired bases are not distributed equally on each side had to be drawn with gaps left on the short side. Three of these gaps have been filled in by hand with dotted lines in Structure 2 (Figure 6). Also, the single stranded sequences at the extreme 5' and 3' ends of the molecule are not drawn by the computer, but can easily be supplied by hand.

Below each drawing is a matrix of numbers and asterisks, which is called the canonical form of the structure. (For every set of regions which comprise an orthodox structure, there is a unique canonical form which is topologically equivalent to the molecular structure itself, and which permits display of these structures in a graphically consistent manner.) The numbers are region numbers from the table above, and make it easy to associate the double-helical regions in the drawing with their parameters in the region table. The asterisks, both in the drawings and in the canonical forms, indicate the presence of a multiply branched loop whose single stranded bases may not have

been drawn due to artistic license. By mapping the region numbers from the canonical form onto the drawn structure, the base numbers of the ends of the regions bordering on the multi-branch loop can be determined and the missing bases can easily be supplied by hand. All six of the missing bases in Structure 2 have been added by hand in larger print and joined to the rest of the structure by dotted lines. Also, the region numbers have been added in bold print. Below the canonical forms, schematic diagrams of the topological shape of the predicted structure, similar to the ones in Figure 4 (except upside down, to agree with the orientation of the drawings), have also been added.

DISCUSSION

Fox and Woese (12) performed a comparative analysis on ten complete 5S rRNA sequences from various species of procaryotes and eucaryotes. Their results showed that four strong double-helical regions were common to all procaryotic sequences studied, and that three of these four were also common to eucaryotic sequences. These regions are all present in Structure 2, and have been emphasized by vertical bars (Figure 6). According to the Fox and Woese nomenclature, our region number 6 is the molecular stalk, 61 is the tuned helix, 91 is the common arm base, and 174 is the procaryotic loop. Our program was not constrained in any way that might artificially favor the prediction of such a structure, and the fact that many of our top 30 structures are similar to the Fox and Woese structure is quite remarkable. In fact, analysis of Structure 1 shows that it contains a large multiply branched loop of 17 bases, while the corresponding loop in Structure 2 contains only 7 bases. Because there are no biochemical data on the energy contributions of multiply branched loops, both the Pipas and McMahon program (1) and our own program have assigned zero energy to all such loops regardless of their shape, size, or nucleotide composition. However, like ordinary internal loops, these single stranded portions of the molecule are expected to contribute a destabilizing energy whose magnitude increases with the length of the loop. Thus, when these contributions were taken into account by hand in accordance with Salser (9), Structure 2 became the best overall, with an energy of -48.8 kcal/mole, which is 1.1 kcal/mole more stable than the -47.7 kcal/mole for Structure 1. The extreme right hand column of Figure 5 shows these adjusted energies and ranks for the 30 computed structures. It should be emphasized that we do not regard the top position in this energy list as enormously significant, because such a molecule will not adopt a 'single most stable' configuration, but instead will show a distribution among a number of different conformations of very

similar stabilities, some of which may be freely interchanging in a rapid equilibrium, while others may not. Thus many of these 30 predicted structures are likely to have important biological significance.

In accordance with this, Fox and Woese discuss the biochemical characterization of two forms of the E. coli 5S rRNA: the native form, which is active in reconstitution of 50S ribosome subunits, and the B form which is inactive. (There is also an intermediate A form.) These studies suggest that the tuned helix becomes uncoiled during the transition from the native form to the B form. In Structure 1, the tuned helix and the common arm base are not coiled, and the single long arm (from Structure 2) consisting of regions 51, 61, 86, and 91 has been replaced by two shorter arms containing regions 40, 62, 117, 126, and 142. In the diagram to the right in Figure 6, a possible mechanism for this conformational change is shown. Since regions 174 and 178 (including the procaryotic loop) are common to both structures, they have been removed from this extra copy of Structure 2 to allow space for the arrows which indicate the single-stranded sequences that can pair to form the regions in Structure 1. By undergoing a 180 degree rotation about the axis shown, the lower half of the long arm (regions 86 and 91) comes into register with the multi-branch loop near the molecular stalk. In addition, all of the halfregion pairs corresponding to Structure 1 line up and the internal loop between regions 61 and 86 (the axis of rotation) can transform into the two hairpin loops of regions 62 and 142. This configuration is a very frequent alternative in the 30 X 46 matrix (Figure 5) to shapes like Structure 2. In fact, all of the predicted structures which contain region 62 also contain 40, and many of them also contain 117, 126, and 142.

In support of this hypothetical mechanism for conformational change is the fact that all six of the procaryotic species discussed by Fox and Woese (E. coli, A. nidulans, Photo. 8265, P. fluorescens, B. megaterium, and B. stearothermophilus) contain a region equivalent to 86 in Structure 2 whose sequence is conserved as CAC paired to GUG (except for a G to U transversion in B. megaterium) and that even the lengths and part of the sequence of the single-stranded loops adjacent to region 86 are conserved among these six species (except for P. fluorescens, which suffered a deletion of the base corresponding to 26 in Structure 2.

Our computer program was written in the APL Language, and two versions exist, one which has been implemented on the APL*PLUS system running on an IBM 360/91 at UCLA, and another which was implemented on the VS-APL system running on an IBM 370/158 computer at UCSF. The program was debugged and

its accuracy was verified by obtaining complete solutions for relatively short (70-100 nucleotide) RNA sequences both from our program and from the Pipas and McMahon program (1), and comparing the predicted structures. We are indebted to Pipas and McMahon for the use of their program and for advice concerning its use, without which the debugging of our own program would have been exceedingly more difficult. Copies of our program will be made available by the first author upon written request.

ACKNOWLEDGEMENTS

We thank James Adams for advice on the design of some of our algorithms, and we thank Martha Wynne for her help in preparing the manuscript. We thank Fred Eiserling for his patience during the development of our computer program, and we again thank James McMahon for many helpful discussions about the use of his program. This work was supported in part by PHS grants CA15940 and GM18586. G.M.S. was supported by USPHS genetics training grant GM07104.

REFERENCES

1. Pipas, J., and McMahon, J. (1975) Proc. Nat. Acad. Sci. USA 72, 2017-2021.

2. Dugundji, J. (1966) in Topology (Allyn and Bacon, Boston), 15 & 29-31.

3. Delisi, C., and Crothers, D.M. (1971) Proc. Nat. Acad. Sci. USA 68, 2682-2685.

4. Tinoco, I., Uhlenbeck, O., and Levine, M. (1971) Nature 230, 362-367.

5. Jacobson, H., and Stockmayer, W. (1950) J. Chem. Phys. 18, 1600-1606.

6. Gralla, J., and Crothers, D.M. (1973) J. Mol. Biol. 73, 497-511.

7. 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.

8. Borer, P.N., Dengler, B., Tinoco, I., and Uhlenbeck, O.C. (1974) J. Mol. Biol. 86, 843-853.

9. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. XLII, 987-1004.

10. Corry, M.J., Payne, P.I., and Dyer, T.A. (1974) FEBS Lett. 46, 63-66.

11. Heindell, C., Liu, A., Paddock, G., Studnicka, G., and Salser, W. Cell, in press.

12. Fox, G.E., and Woese, C.R. (1975) J. Mol. Evol. 6, 61-76.

13. Gralla, J., and Crothers, D.M. (1973) J. Mol. Biol. 78, 301-319.

14. Korn, L.J., Queen, C.L., and Wegman, M.N. (1977) Proc. Nat. Acad. Sci.

USA 74, 4401-4405.

15. Fitch, W.M. (1972) J. Mol. Evol. 1, 185-207.