Chemical and Computer Probing of RNA Structure

N. A. Kolchanov,*
I. I. Titov,* I. E. Vlassova*
and V. V. Vlassov†,[1]

*Institute of Cytology and Genetics
Siberian Division of Russian Academy
of Sciences
Novosibirsk 630090, Russia
†Institute of Bioorganic Chemistry
Siberian Division of Russian Academy
of Sciences
Novosibirsk 630090, Russia

Ribonucleic acids are one of the most important types of biopolymers. RNAs play key roles in storage and multiplication of genetic information. They are important in catalysis, RNA splicing, and the most important steps of translation. Studies in the past few years have demonstrated the possibility of developing RNA species (aptamers) that can recognize different biopolymers and synthetic organic molecules. Problems of investigation of RNA structure and functions, and recent exciting developments in the design of catalytic RNA molecules and specific RNA ligands, have been considered previously (*1*).

[1] To whom correspondence may be addressed.

The complicated natural functions of RNAs require specific interactions of these molecules with proteins and other nucleic acids. The specificity of these interactions of RNAs and their biological activities are determined by their three-dimensional structures. The three-dimensional (tertiary) structure of RNA is formed by hydrogen-bonding between functional groups of nucleosides in different regions of the molecule, by coordination of polyvalent cations, and by stacking between the double-stranded regions present in the RNA. Knowledge of the tertiary structure of RNAs and the possibility to predict RNA folding from nucleotide sequences are of key importance for understanding the principles of genetic information, for elucidation of relationships between the structure of RNA and its functions, for the design of functionally active polynucleotides, and for the selection of optimal oligonucleotide probes for the detection of specific RNAs and antisense oligonucleotides for modulating the functions of specific RNAs.

At present, the tertiary structures of only some small RNAs have been determined by high-resolution X-ray crystallographic analysis (*2–6*) and NMR analysis (*7–8*). For both of those physical methods, relatively large amounts of highly purified RNA are needed, and X-ray studies require high-quality RNA crystals, which are difficult to grow. These are serious and principal limitations, because the goal of researchers is investigation of the biologically active RNA structure, which is attained in solutions of definite composition, and specific protein factors are sometimes required for correct folding and functioning. Therefore, there is a need for methods allowing analysis of the folding of RNAs in solution and great attention has been paid to the development of approaches for prediction and investigations of RNA structure in complex biological systems.

The most widely used approach for investigation of RNA structure is chemical and enzymatic probing in combination with theoretical methods and phylogenetic studies allowing prediction of variants of RNA folding. Chemical and enzymatic probings allow determination of the reactivities of different functional groups of RNA that can be interpreted in structural terms. From such data it is possible to identify such structural features of RNA in solution as the occurrence of double-stranded regions and long-range interactions between nucleotides responsible for the three-dimensional RNA folding. Chemical methods allow detection of some specific spatial arrangements of nucleotides that bind metal ions and metal complexes. Cross-linking with chemical reagents allows the determination of intramolecular distances between certain nucleotides in the RNA tertiary structure. A key advantage of these methods is the ability to study RNAs too large for crystallographic and NMR studies and the possibility to investigate structures of nonpurified RNAs in complex systems containing various factors, or even RNA bound to specific proteins. Information from the probing experiments allows one to identify the real RNA structure among a number of potential structures that can be predicted from sequence data and free energy parameters, and from data of phylogenetic studies in which sequences of a particular RNA from diverse species are compared in order to infer the existence of base-paired regions.

In this essay we describe experimental methods for probing RNA structure and theoretical methods allowing prediction of thermodynamically favorable RNA folding. These methods are complementary, and together they provide a powerful approach to determine the structure of RNAs.

I. Probing RNA Structure by Chemical and Enzymatic Approaches

A. General Principles of Chemical and Enzymatic Probing: Analysis of Modified RNA

Most natural RNAs are globular molecules containing short single-stranded sequences and short double-stranded fragments formed by intramolecular interactions of complementary nucleotide sequences. The system of single-stranded and double-stranded regions formed by complementary nucleotide sequences of the molecule is called the secondary structure of RNA. The double-stranded helices of RNA assume the A form in physiological conditions, with 11 base-pairs per single turn. Long, regular, uninterrupted helices are rare in most RNAs. A typical element of local RNA structure is an 8- to 10-base-pair segment incorporating a bulge or mismatch (Fig. 1). Unpaired bases are often involved in interactions leading to the three-dimensional folding of the molecules. Due to these interactions between nucleotides and interactions with metal ions, the elements of the secondary structure of RNAs fold in a unique three-dimensional (tertiary) structure.

Clues for understanding the principles of the folding of RNA molecules come essentially from X-ray studies of tRNAs (2–6). tRNAs are 72–95 nucleotides long and are folded in a cloverleaf-like structure containing four stems and three loops (Fig. 2). The tertiary structure of tRNA is formed by interactions between nucleotides in the D and T loops. Besides the Watson–Crick base-pairing, a few other types of hydrogen-bonding occur in tRNAs. Thus a G15·C48 pair is formed by the bases in parallel RNA strands. m^1A58 and T54, also in parallel strands, form a reverse Hoogsteen pair.

The tertiary folding of tRNA also involves base triplets in which a third nucleotide forms hydrogen-bonds, with a Watson–Crick base-pair in the major groove of short helices. Thus, in yeast $tRNA^{Phe}$, the triplet

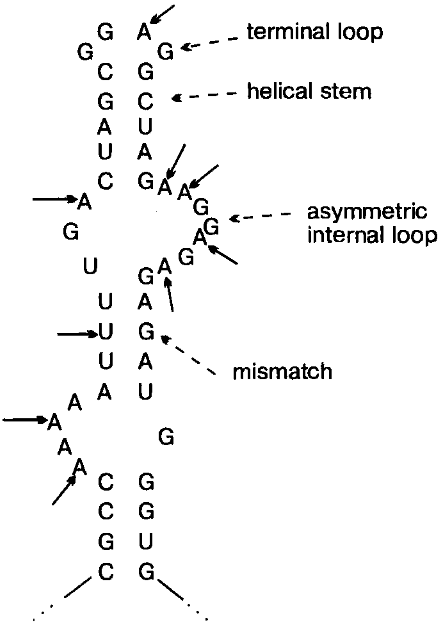

FIG. 1. A typical fragment of RNA secondary structure: a conserved RNA stem–loop structure in the packaging signal of the human immunodeficiency virus type 1. The RNA secondary structure was predicted theoretically and confirmed by probing with diethyl pyrocarbonate (reaction with adenosines in single-stranded regions of RNA) and S1 nuclease (cleavage of phosphodiester bonds in single-stranded regions of RNA). The positions attacked by the probes are indicated by arrows. It is seen that the modification and cleavage patterns are consistent with the RNA folding. Reprinted from T. Hayashi, Y. Ueno and T. Okamoto, *FEBS Lett.* **327**, 213 (1993), with kind permission from Elsevier Science—NL, Sara Burgerhart-Straat 25, 1055 KV Amsterdam, The Netherlands.

($m^2G10 \cdot C25) \cdot G45$ is formed by the interaction of the third base with a Watson–Crick pair by one hydrogen bond. Triplets $(G22 \cdot C13) \cdot m^7G46$ and $(A23 \cdot U12) \cdot A9$ are formed by two hydrogen bonds of the third base with corresponding Watson–Crick base-pairs. Due to these interactions, the cloverleaf structure is folded into a three-dimensional L-shaped structure built of two helical domains formed by the stem regions of the molecule, because of stacking interactions (2–6) (Fig. 3).

RNAs are polyanionic molecules; they bind cations (metal ions and the organic polycations, spermine and spermidine). Specific active conformations are formed by RNAs only in the presence of certain concentrations of monovalent cations and magnesium ions. In the three-dimensional structure of tRNAs, there are some sites where particularly tight binding of cations occurs. These are sites where groups capable of interacting with the ions

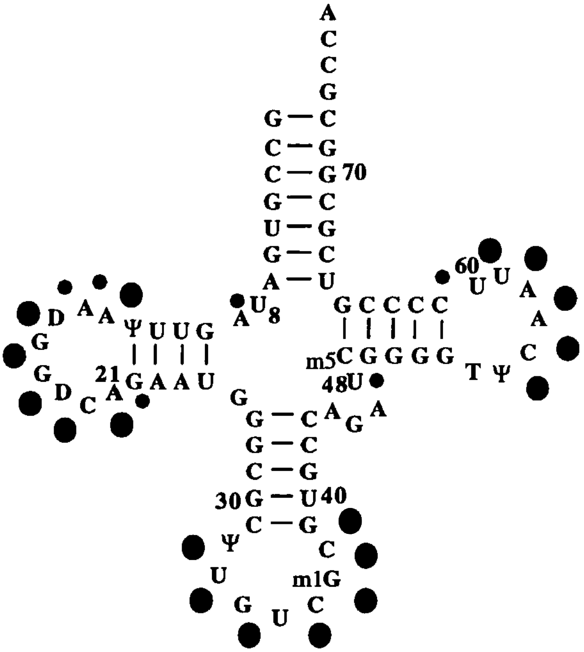

FIG. 2. Cloverleaf structure of yeast tRNA^{Asp} with imidazole-induced cleavage points. Phosphodiester linkages displaying enhanced susceptibility to hydrolysis by the imidazole buffer in conditions stabilizing the RNA structure are indicated by dots with diameters proportional to the intensity of the cuts (*46*).

(phosphates, nitrogens of heterocyclic bases, ribose oxygens) can be arranged optimally for simultaneous interaction with an ion (*11*).

RNAs are built of a large number of chemically similar monomers that possess a few chemical groups available for chemical modification or for attack by enzymes capable of hydrolyzing RNA. The microenvironment of these groups can be very different in the three-dimensional structure of RNA, and these differences can dramatically affect the reactivities of the groups toward chemical and enzymatic probes. When structural factors affecting the reactivities of specific groups to given probes are known, chemical modification data can be interpreted in structural terms.

The following main factors affect the reactivities of groups in RNAs.

1. A group can be partially or completely buried within the molecule, which interferes with reactions—in particular, with reactions with bulky probes. Stacking of heterocyclic bases decreases their reactivity toward reagents that attack the bases perpendicularly to their planes.

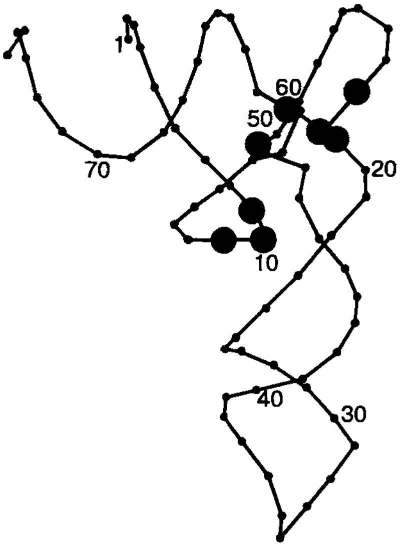

FIG. 3. Tertiary structure of yeast tRNA^{Phe}. Black circles indicate phosphates in the RNA structure protected from modification with a small probe, ethylnitrosourea. Reproduced from Ref. *10*.

2. Participation of a group in hydrogen bonding or in coordination with a metal ion affects its nucleophilicity and results in structural shielding.
3. The electrostatic environment of a group affects its ionization. Because RNAs are polyanions they repel negatively charged species and attract positively charged reagents, which results in suppression or acceleration of the corresponding reactions.

The identification of factors suppressing the reactivity of a given base is possible by investigation of patterns of reactivities of nucleotides toward different chemical probes whose chemical specificity is known. Attempts have been made to combine steric and electrostatic factors in a form of a theoretical index [Accessible Surface Integrated Field Index (*12*)], which correlates with the reactivities of functional groups of RNA—e.g., reactivities of RNA phosphates toward ethylnitrosourea (*10, 13*).

Nucleotides within single-stranded and double-stranded regions of RNA can easily be distinguished using chemical probes reacting with functional groups of the bases participating in Watson–Crick interactions. In the double-stranded regions, the groups are shielded from reagents present in solution. Similarly, nucleotides in the double-stranded regions that participate in base-triplet formation are easy to identify because of shielding of the

N-7 atoms in the major groove of the double helix. Reduced reactivities of phosphates and nitrogen atoms of heterocyclic bases can reflect their involvement in coordination of metal ions.

When performing probing experiments, one should arrange experimental conditions in which the RNA under study will assume the desired structure and retain it throughout the experiment. RNAs acquire a biologically active structure only in a relatively narrow range of conditions ("physiological" conditions). In the course of isolation, an RNA structure is often denatured and it is necessary to transfer it to conditions allowing resumption of the biologically active structure. Therefore, before RNA is subjected to probing, it is essential to ensure that the population of molecules is homogeneous and to remove traces of denaturants used in the RNA isolation. It is recommended, when possible, to perform a heat treatment followed by a slow cooling down (renaturation) to allow the RNA to assume a thermodynamically favorable structure.

An important requirement of modification experiments aimed at probing the reactivities of different nucleotides consists in ensuring that the RNA is subjected to limited chemical modification or enzymatic hydrolysis providing statistically less than one cut or modification per RNA molecule. This guarantees that the molecule under study has not been changed in the course of investigation and allows obtaining quantitative data on the reactivities of specific residues. Reactions are performed in the presence of carrier RNA for controlling the reaction conditions. Incubation of RNA in experimental conditions without the reagent is performed as a control for detection of breakage caused by nonspecific factors potentially present.

Detection of cleavage sites and modification sites can be performed by two methods, the choice determined by the size of the RNA molecule and by the nature of the modification. One method uses end-labeled RNA molecules and allows detection of cleavages in the RNA by gel-sequencing. A limitation of this method is that it detects only scissions in RNA structure. This method can be used to study RNA with at least one homogeneous end, up to 300–400 nucleotides long, or terminal sequences of large RNAs.

Labeling of the 5′ end of RNA can be performed enzymatically, using T4 polynucleotide kinase, transferring the γ-phosphate from [γ-^{32}P]ATP to the 5′-terminal ribose of RNA (*14*). If the RNA has a phosphate group at the 5′ terminus, the phosphate can be removed by alkaline phosphatase prior to labeling. Alternatively, a T4 polynucleotide kinase-catalyzed exchange reaction between the γ-phosphate of [γ-^{32}P]ATP and the phosphate of RNA can be used to substitute the labeled phosphate for the cold one (*15*). Labeling of the 3′ end of RNA can be performed by attaching [5′-^{32}P]pCp to the 3′-OH group of the RNA using T4 RNA ligase (*16*). tRNA can be 3′ end-labeled by removing the terminal CCA sequence by phosphodiesterase and restoring the CCA end using tRNA nucleotidyl transferase and labeled CTP and ATP (*17*).

The principle behind the method is outlined in Fig. 4. Cleavage of RNA with an RNase or by a chemical probe in conditions allowing one hit per molecule generates pairs of fragments. The fragments are resolved by electrophoresis in polyacrylamide gel in denaturing conditions followed by autoradiography, which allows registration of the fragments originating from the labeled end of the RNA. To determine the size of the fragments, products of limited alkaline hydrolysis of the same RNA and fragments produced by some sequencing reactions are run on the same gel.

The second method for analysis of modified RNA uses reverse transcription (Fig. 5). The attacked position is identified by a stop in reverse transcription generated from a DNA primer. This method is most generally useful and an expedient approach to probe any RNA sequence, regardless of

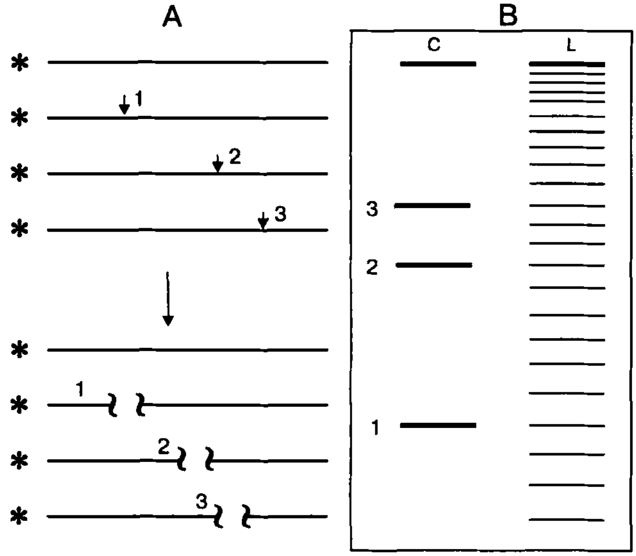

FIG. 4. Schematic representation of the method for detection of cuts in RNA structure. (A) A probe (chemical reagent or enzyme) attacks, under limiting conditions, three sites in a 5'-labeled RNA, which results in cleavage of the RNA. (B) Positions of the cleavages are mapped by electrophoresis on a denaturing polyacrylamide gel. The lengths of the produced labeled fragments are determined by comparison to the length standards, to locate precisely the reactive nucleotides. In the illustrating gel, the first line (C) might be, e.g., a partial T1 ribonuclease digest of the RNA containing three guanosine residues. The second line (L) is the ladder produced by limited alkaline hydrolysis of the RNA, providing statistical cleavages of all phosphodiester bonds.

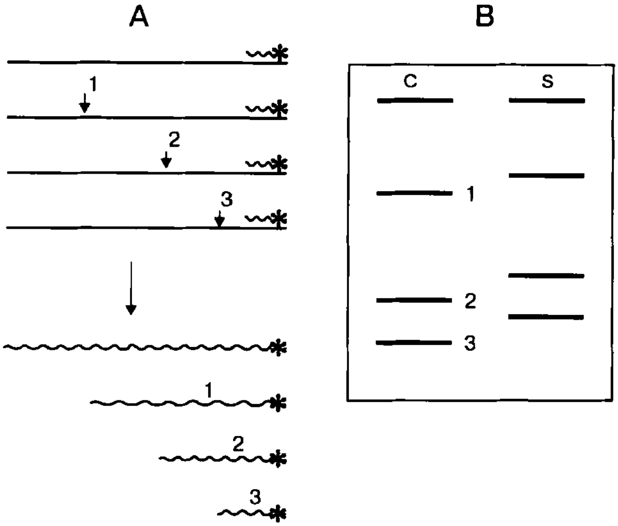

FIG. 5. The primer-extension procedure for detection of cuts and chemical modifications in RNA. (A) Arrows indicate positions of cuts or modified nucleotides. A radiolabeled oligodeoxyribonucleotide primer is annealed to the 3′ end of the RNA. Reverse transcription of the modified molecules is terminated at the modified residues and yields shortened transcripts. (B) The length of the transcripts is determined by gel electrophoresis. The left lane (C) is the schematic presentation of the bands corresponding to the fragments, which might be, e.g., fragments transcribed from an RNA with three adenosine residues modified with diethyl pyrocarbonate. The second lane (S) shows the A-specific sequencing reaction. The relative shift of the bands in the lanes is explained by termination of the reverse transcription at a nucleotide preceding the modified residue.

size; it also allows one to detect chemical modifications that do not cleave RNA. Analysis of modified RNAs using the primer extension method is performed by annealing a complementary oligodeoxynucleotide primer to the 3′ end of the RNA and synthesizing a cDNA copy of RNA using reverse transcriptase and dNTPs. Elongation proceeds from the 3′ end of the primer and it is terminated prematurely when the enzyme meets scissions or chemically modified nucleotides and stops. When transcripts are resolved by gel electrophoresis, the stops in cDNA synthesis are detected as bands corresponding to a modified position in the RNA template.

Reverse transcription is effectively arrested by modifications of nucleotides at the groups involved in Watson–Crick base pairing. This was shown for G(N-1) and G(N-2), for modification with kethoxal; for A(N-1) and C(N-3) for modification with dimethyl sulfate; and for G(N-1) and U(N-3), for reaction with carbodiimides. Carbethoxylation of adenosine at N-7 with diethyl pyrocarbonate opens the imidazole ring, and the modified residue stops the reverse transcription. Methylation of guanosine at N-7 with Me_2SO_4 does not arrest transcription, but this modification can be transformed into a cleavage by treatment with aniline. The synthesized DNA fragments can be labeled by using 5′ end-labeled primers or by the use of [γ-^{32}P]NTP. The produced fragments are identified by analyzing them in parallel with dideoxynucleotide sequencing reactions performed with the same unreacted RNA and DNA primer.

It should be mentioned that sequencing patterns obtained by using the primer extension method are shifted by one nucleotide relative to the modified residue, because the last nucleotide incorporated by the transcriptase is complementary to the one on the 3′ side of the modified residue in the template RNA. If the RNA under study is long, it is necessary to perform reverse transcription with a few different primers to explore the whole RNA molecule.

Limitations of the method are related to the sensitivity of reverse transcriptase to different modifications. Some of them do not stop the enzyme. On the other hand, some naturally occurring modified residues (e.g., m^2G, m^6A) arrest reverse transcription, and some tightly folded regions of RNA slow down the transcription process. Therefore, a control reverse transcription reaction should be run on the unmodified RNA to detect pauses of natural origin caused by nonspecific breaks of RNA in the reaction conditions and by natural modifications of nucleotides and structural elements that may affect the transcription.

Excellent detailed experimental protocols for investigation of RNA structure with chemical and enzymatic probes and description of analytical techniques can be found elsewhere (*18, 19*).

B. Enzymatic Probes

Enzymes cleaving the ribose-phosphate backbone of RNA (ribonucleases) are the simplest and most widely used tools for probing RNA structure. Most of these enzymes attack single-stranded regions of the RNA structure showing different specificities to phosphodiester bonds adjacent to certain nucleosides. One enzyme, ribonuclease V1, cleaves RNA preferentially at double-stranded regions. Investigation of the susceptibility of different sequences within the RNA structure toward different ribonucleases allows identification of elements of the secondary structure of RNA.

Although enzymatic probes are popular because of easy handling and simplicity of detection of cleavage positions in RNA, these probes have some drav/backs. The mechanism of phosphodiester-bond cleavage is known; however, it is preceded by a step of enzyme–RNA recognition. Features of this step, involving noncovalent binding of enzyme probes with the surface of the

RNA, are not well understood. Binding of the enzyme to RNA can affect the polynucleotide structure. As a result, the produced cleavage pattern may characterize properties of a perturbed RNA structure rather than of its native structure. Thus, the small protein ribonuclease A shows a very high tendency to cleave Y–A linkages in single-stranded regions of RNA. However, the enzymes sometimes cuts at these sequences in double-stranded regions of RNA, apparently because binding of this cationic protein can unfold the substrate structure locally. Moreover, ribonuclease, noncovalently bound to RNA, can accomplish a few cuts in the same RNA molecule even under conditions of limited hydrolysis. These secondary cuts apparently do not reflect features of the native RNA structure. To detect such cuts, hydrolysis patterns of 5′- and 3′-end-labeled RNAs should be compared.

RNase U2 from *Ustilago sphaerogena* is used for probing adenines in single-stranded RNA sequences. The enzyme cleaves phosphodiester bonds adjacent to the 3′ phosphate. The order of sensitivity of phosphodiester bonds to this enzyme is $A > G \gg C > U$ (*20*). The pH optimum of the reaction is 4.5; 7 *M* urea does not stop the hydrolysis.

RNase T1 from *Aspergillus oryzae* cleaves phosphodiester bonds after the 3′ phosphate of unpaired guanosine residues. The reaction yields fragments with 3′ phosphates and proceeds via the intermediate formation of guanosine 2′:3′-cyclic phosphate (*21*). The presence of 7 *M* urea stimulates the enzyme activity, when the reaction is carried out at pH 4.5. The enzyme does not hydrolyze RNA after some naturally occurring modified guanosines (m^1G and m^7G).

RNase CL3 from chick liver is used as a probe, cleaving phosphodiester bonds after unpaired cytidines, and yields fragments with a 3′ phosphate (*22*, *23*). The enzyme activity is enhanced by spermine and magnesium ions; a pH effect depends on the nature of the buffer.

T2 RNase from *A. oryzae* cleaves RNA after unpaired adenosine residues yielding fragments with 3′ phosphate, via formation of intermediates with 2′:3′-cyclic phosphates (*21*). The enzyme has relatively low specificity and exhibits a strong activity to nucleotides at the apex of terminal loops. Internal loops are substantially less reactive. Although the pH optimum of the reaction is 4.5, the enzyme can be used at neutral pH (*20*). T2 RNase is inhibited by heavy metal ions.

S1 nuclease from *A. oryzae*, used as a probe, is capable of cleaving single-stranded regions in RNA and DNA (*24*). The enzyme yields fragments with a 5′ phosphate. The pH optimum is at 4.5, although the enzyme is still active at neutral pH. The enzyme is stimulated by Zn^{2+}.

Neurospora crassa nuclease is used as a probe cleaving single-stranded regions in RNA and DNA. The hydrolysis generates fragments terminated by 5′ phosphates. The pH optimum of the enzyme is 7.5–8 (*25*). Increasing pH and decreasing ionic strength result in decreasing sequence-specificity of the nuclease. Because the enzyme has Co^{2+} as prosthetic group, its activity is inhibited by EDTA.

RNase V1 from cobra venom cuts preferentially double-stranded and structured (stacked) regions of RNA, showing no apparent nucleotide specificity. The produced fragments contain 5′ phosphate (*26*). The enzyme needs Mg^{2+} ions; it is active in the pH range of 4–9.

C. Chemical Probes

Several chemical reagents are available to probe reactivities of functional groups of heterocyclic bases and reactivities of phosphodiester bonds and ribose (Figs. 6 and 7). Detailed protocols for probing RNA with chemical reagents can be found in Refs. *18* and *19*.

1. Alkylating Reagents

Dimethyl sulfate (Me_2SO_4) and derivatives of 2-chloroethylamine react with nucleophilic centers of heterocyclic bases. At neutral pH, the order of reactivities is G(N-7) > A(N-1), C(N-3) (*27*). The 7–8 double bond in an alkylated guanosine can easily be reduced by sodium borohydride. The resulting product provides a site for aniline-induced scission (*28*, *29*). The reaction is used for detection of guanosines with N-7 atoms involved in hydrogen bonding or in coordination with metal ions. The reactivity of guanosines is affected, to some extent, by stacking. Alkylation at adenosines and cytidines is used to detect nucleotides not involved in Watson–Crick interactions. The modified residues can be detected by the primer extension method. For detection of modified cytidines, RNA can be cleaved at the modified residues by treatments with hydrazine and then with aniline. Hydrazine reaction results in some cleavage at uridines, but this reaction is structure-independent. The modifications can be detected by the primer extension method. Modification by Me_2SO_4 can be used for detection of hydrogen-bonding of adenosines in the *syn* conformation, which occurs in GA and Hoogsteen AU pairings.

Ethyl nitrosourea (ENU) is an alkylating reagent attacking both internucleotide phosphates and nucleophilic centers of heterocyclic bases in RNA with comparable efficiency (*30*, *31*). The ethyl phosphotriesters formed are unstable; mild alkaline treatment results in breakage of the RNA chain at positions of the phosphotriesters. This reaction can be used to map phosphates not engaged in binding of metal ions or in hydrogen-bond formation (*10*, *13*, *32*).

2. Carbodiimides

For modification of RNA, 1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-*p*-toluene sulfonate (CMCT) is usually used (*33, 34*). CMCT reacts with uridine (at N-3) and less efficiently with guanosine (at N-1). CMTC reacts also with some minor RNA components (thymidine, dihydrouridine, and pseudouridine). In pseudouridine, both nitrogen atoms are reactive, and both mono- and diadducts can be formed. Reaction with CMTC is stimulated by increasing pH; usually the reaction is performed at pH 8. At pH > 10, the adducts decompose, yielding nonaltered nucleosides. Because the reactive groups of nucleotides participate in Watson–Crick base-pairing, the reaction occurs only within single-stranded regions of RNA.

3. α-Ketoaldehydes

Usually β-ethoxy-α-ketobutyraldehyde (kethoxal) is used for modification of RNA. The compound reacts with guanosine in single-stranded regions of RNA. The reaction yields a new ring involving N-1 and N-2 of the guanosine and both carboxyl groups of kethoxal (*35*). Because attack of the reagent occurs perpendicularly to the plane of the base, stacking inhibits the reaction.

Reaction is carried out at pH 7–7.5. The modification products are stable in slightly acidic medium; at basic pH, they decompose into the components. The adducts are stabilized by borate ion. The modification can be made irreversible by oxidizing the *cis*-diol group of the adducts with periodate.

4. Diethyl pyrocarbonate

This reagent carbethoxylates purines at N-7, favoring adenine over guanine in a reaction sensitive to the solvent exposure of the base (*36–39*). The reagent is used to provide information on the structural environment of purines by probing the involvement of N-7 of adenosine in tertiary interactions. DEPC is sensitive to stacking and poorly attacks purines in helical regions of RNA. The effective diameter of DEPC is 3.5 Å, which is similar to the width of the deep RNA major groove. The modification results in opening of the imidazole ring between atoms N-7 and C-8, and the RNA can be cleaved at the modified residues by aniline treatment. Minor modification of uridine (at N-3) in slightly basic medium can occur (*38*) and cytidine reacts in solutions with high concentration of salts.

5. Bisulfite

Bisulfite reacts with unpaired cytidines forming 5,6-dihydrocytidine 6-sulfonate. At pH 5–6, in the presence of high concentration of bisulfite,

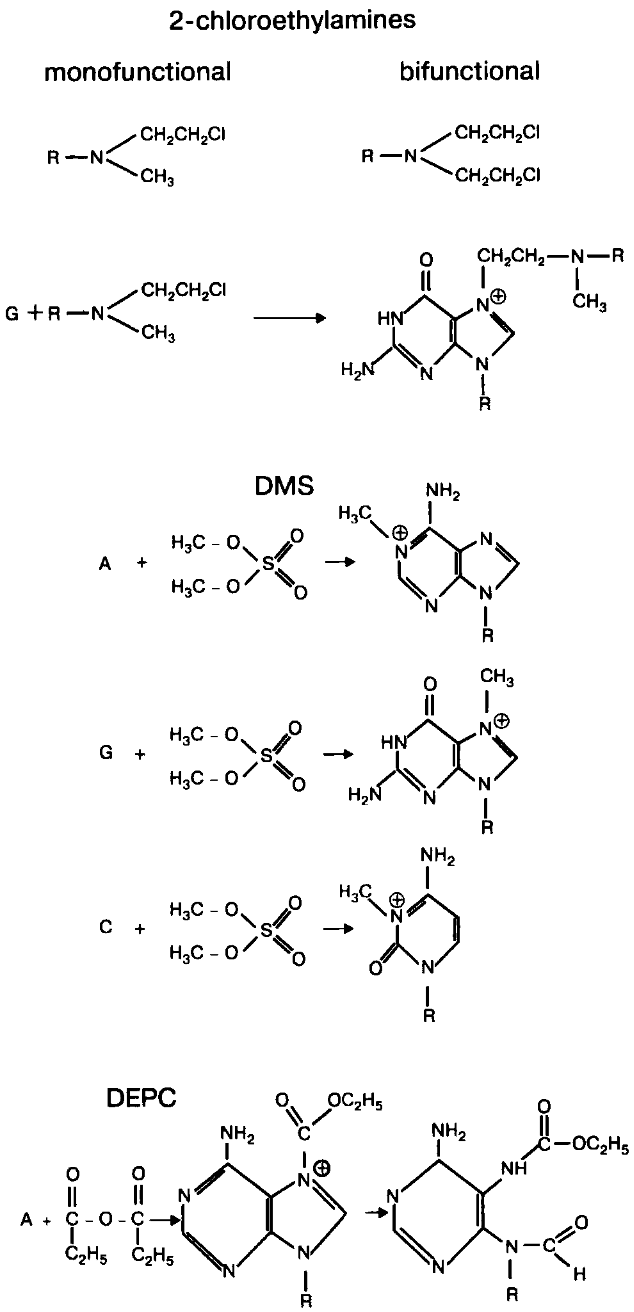

Fig. 6. Reactions of chemical probes with RNA nucleotides. DMS, Dimethyl sulfate; DEPC, diethyl pyrocarbonate; kethoxal, β-ethoxy-α-ketobutyraldehyde; CDI, carbodiimide; ENU, ethylnitrosourea.

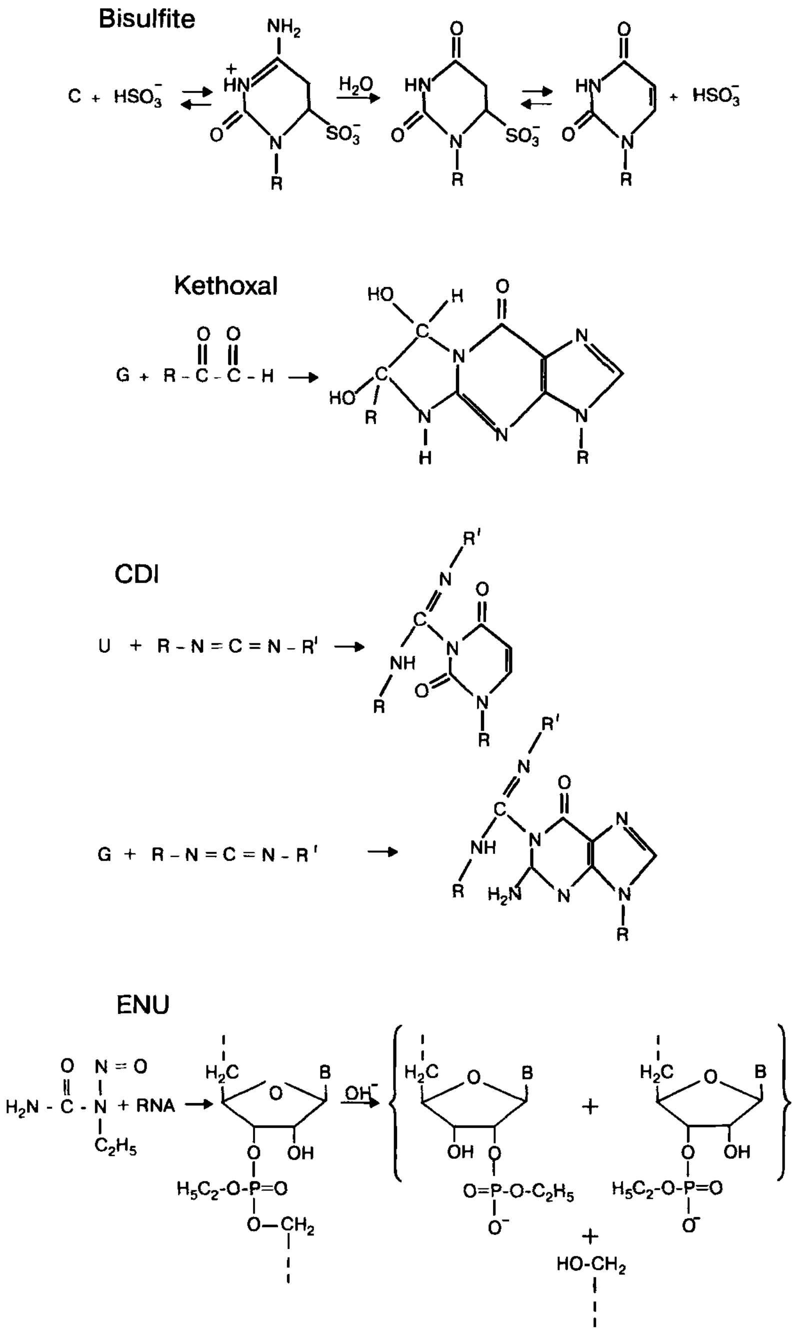

FIG. 6. *Continued*

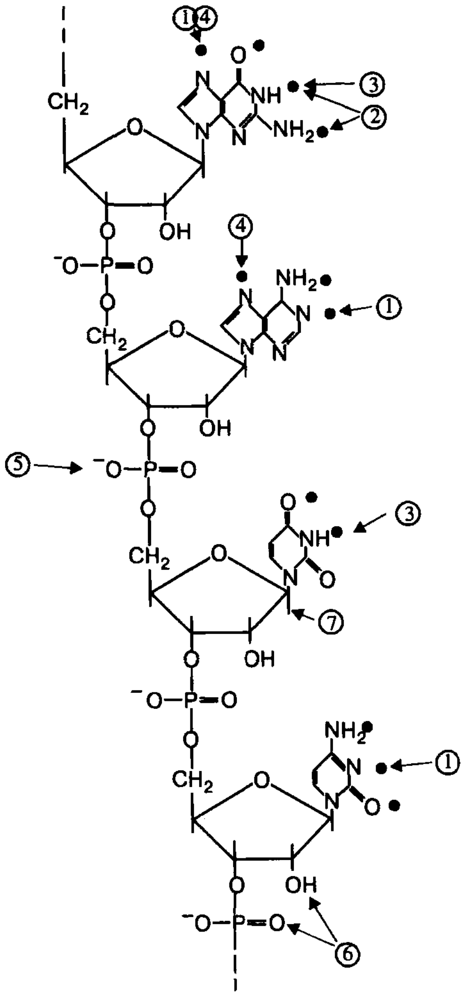

FIG. 7. Functional groups of nucleotides available for probing with chemical reagents. Black circles indicate groups of the bases participating in the Watson–Crick and Hoogsteen hydrogen bonding. Arrows indicate sites of reactions with chemical probes: 1, dimethyl sulfate; 2, kethoxal; 3, carbodiimides; 4, diethyl pyrocarbonate; 5, ethylnitrosourea; 6, imidazole; 7, OH radicals.

nucleophilic substitution at the exocyclic amino group of the cytidine derivative occurs. This results in formation of the corresponding derivative of uridine. Treatment of the compound with mild alkali removes the bisulfite moiety. Conversion of C to U can be detected by a U-specific sequencing reaction with hydrazine (38).

6. Fe(II)–EDTA

The negatively charged EDTA complex of iron(II) reacts with hydrogen peroxide in a Fenton reaction and generates hydroxyl radicals, which react with nucleic acids (*39*, *40*).

$$\mathrm{Fe(II)} + \mathrm{H_2O_2} \rightarrow \mathrm{Fe(III)} + \mathrm{OH^-} + \mathrm{OH}\cdot$$

The complex of Fe(II) with EDTA is anionic and does not bind to RNA, so the hydroxyl radicals diffuse from the generation site to RNA. Although the highly reactive hydroxyl radicals attack both heterocyclic bases and ribose, only the latter modification results in strand breaks. The cleavage is initiated by abstraction of a hydrogen from ribose. The ribose radical produced decomposes, yielding as final products RNA fragments terminated by 5′ and 3′ phosphates. Apparently, the reaction is nonspecific with respect to the nature of the nucleotide. It can be used for identification of surface residues of RNA molecules (*41–43*).

7. Methidium propyl–EDTA–Fe(II)

This conjugate contains a methidium moiety capable of intercalating in double-stranded regions of RNA and the OH-radical-generating Fe(II)–EDTA group (*44*). Due to the intercalating group, the reagent produces cleavages preferentially within the base-paired regions of RNAs.

8. Imidazole and Conjugates Bearing Imidazole

Concentrated imidazole buffer catalyzes cleavage of phosphodiester bonds in RNA (*45*). The ionized and neutral components of the buffer catalyze the reaction similarly to imidazole residues in the active center of ribonuclease. Hydrolysis of phosphodiester bonds in the single-stranded regions of RNA occurs much more rapidly compared to phosphodiester bonds in the double-stranded regions of RNA (*46*), apparently due to higher rigidity of the latter, preventing conformational changes needed for the ribosephosphate to form a reactive intermediate. Therefore the reaction is useful for mapping single-stranded regions in RNA.

Conjugates of intercalating dyes with histamine, and spermine–histamine conjugates, in the presence of imidazole cleave RNA, with a specificity similar to that of ribonuclease A (*46*, *47*). The most readily attacked are the Y–R sequences, in particular, C–A, in the single-stranded regions of RNA. The mechanism of the reaction is apparently the catalysis by imidazole residues brought into close contact to the RNA riboses and phosphodiester bonds.

D. Techniques for Probing RNA Structure

1. Identification of Nucleotide Interactions in RNA Structure and Elucidation of RNA Folding

Nucleotides located in single-stranded and double-stranded regions in RNA structure can be distinguished by testing their reactivities toward chemical probes reacting with functional groups of nucleotides participating in Watson–Crick interactions. Because the ribosephosphate backbone within the single-stranded regions is less rigid, compared to the more structured regions of RNA, single-stranded regions are more susceptible to hydrolysis by imidazole buffer (*46*). Figure 8 shows the positions of those phosphodiester bonds sensitive to hydrolysis by imidazole in a small folded RNA molecule. It is seen that the cleavages occur within the single-stranded regions of the RNA. Spermine–imidazole conjugate in the presence of imidazole buffer cleaves preferentially Y–R sequences in the single-stranded regions of the molecule.

Direct correlations between chemical reactivities of nucleotides within RNA and conformation of the nucleic acids are well established by studies on tRNAs in which chemical reactivity patterns could be explained by crystallographic structures. The knowledge obtained on the reaction specificity allowed investigations of a great number of RNAs and RNA–protein complexes by chemical mapping procedures combined with modeling. In addition to identification of single- and double-stranded regions of RNA structure, chemical and enzymatic probes allow investigation of the stability of

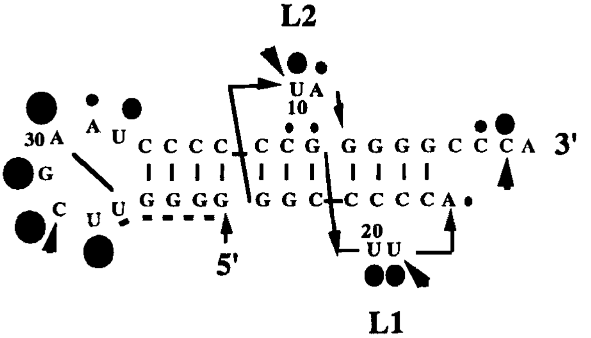

FIG. 8. Secondary structure of the RNA transcript derived from the TMV tRNA-like domain. Dots indicate nucleotides 5′ to the phosphodiester bonds susceptible to hydrolysis by imidazole buffer. The arrows indicate nucleotides 5′ to the phosphodiester bonds attacked by the binary chemical nuclease consisting of spermine–imidazole (2.5 m*M*) supplemented with 25 m*M* imidazole buffer, pH 7.0. L1 and L2 emphasize the two single strands of the pseudoknot crossing the deep and the shallow grooves, respectively. The dashed line indicates the nucleotides that could not be tested for methodological reasons (*46*).

RNA structures in different conditions and detection of interactions important for folding. Probing experiments can be performed under different conditions: under conditions providing the native structure, under various semidenaturing conditions (low salt conditions, variation of temperature), and under conditions wherein RNA is unstructured, at high temperature in the absence of salts. Tertiary interactions are destroyed in the semidenaturing conditions, whereas more stable elements of the secondary structure remain unchanged. Comparison of the data allows identification of elements of the tertiary folding of the molecule and provides information concerning stabilities of different elements of the secondary structure.

An example of a detailed investigation of the structure of an RNA molecule using chemical and enzymatic probes is the study of a 3′-terminal sequence of genomic brome-mosaic-virus RNAs (*48*). The terminal part of these RNAs can be specifically charged with tyrosine by tyrosyl-tRNA synthetase and it is recognized by other proteins interacting with tRNA. However, the proposed structural models of this RNA deviated considerably from the cloverleaf structure of canonical tRNAs. The 201-nucleotide RNA representing the 3′ terminal part of the viral RNA was investigated (*48*) in solution using chemical and enzymatic probes (Figs. 9 and 10). Bases were probed with Me_2SO_4, CMCT, and DPC. Ribonucleases T1, U2, and V1 and nuclease S1 were used for detection of double-stranded and single-stranded regions of the molecule. Modifications and cleavages were detected by both the primer extension method and the direct gel-electrophoretic analysis of cleaved end-labeled RNA. In these experiments all base-paired nucleotides were identified in the double-stranded regions and long-range interactions between a number of bases in the single-stranded regions were identified. The results obtained on reactivities of various atomic positions toward chemical and enzymatic probes provided information needed for building a detailed structural model of the RNA.

In the model, a domain mimicking the shape and dimensions of tRNA were identified, which explains the ability of the RNA to interact with tRNA-related proteins. The model was built as follows. First, potential elements of the secondary structure were built using a computer program (*49*) capable of predicting helices, loops, and different folding motifs. The elements were then assembled in a global structure and atomic accessibilities for target sites for chemical reagents were calculated according to Richmond (*50*). Then the model was refined taking into account the results of the probing experiments. Figure 9 shows the data of the enzymatic mapping. It is seen that the nuclease-cleavage pattern is in good agreement with the shown secondary structure of the molecule.

The results presented in Fig. 9 support the existence of helices B1, B2, B3, and C, and D in the structure. The double-stranded regions are readily

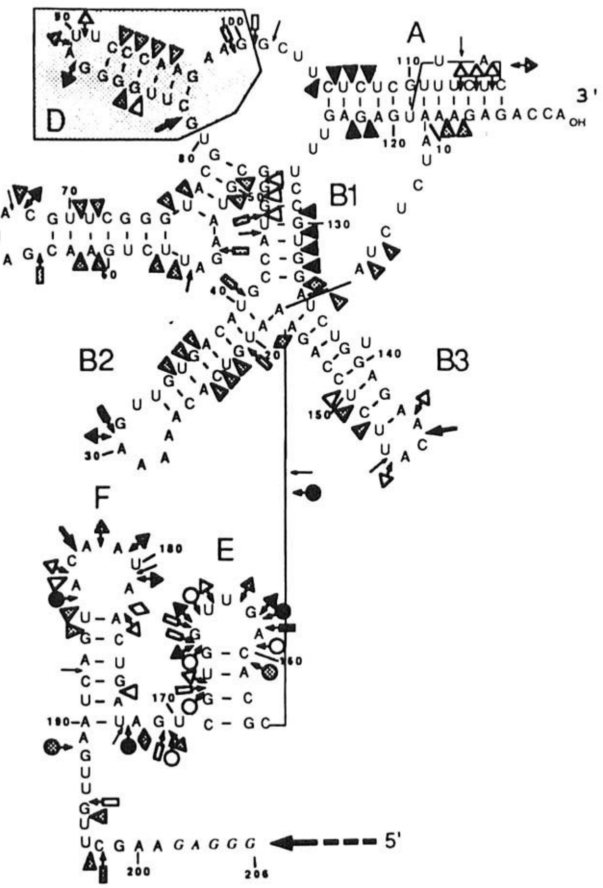

FIG. 9. Results of nuclease mapping of 3'-terminal sequence of the brome mosaic virus RNA (reproduced from Ref. *48*). Indication of cuts: ◆, RNase U2; ▶, V1; ▬►, T1; ◆►, S1 at acidic pH; and ●►, S1 at neutral pH. Open, stippled, and filled symbols correspond to weak, medium, and strong cuts, respectively. Arrows indicate fragile sites of RNA where spontaneous breakage was observed.

attacked by RNase V1. Nuclease S1 and ribonucleases T1 and U2 cut the RNA essentially within single-stranded regions. Some loops are cut less efficiently than others, apparently because of differences in accessibility to relatively bulky enzymes and differences in stability and compactness of the loops. Additional information about the mutual arrangement of the RNA domains was obtained in experiments with RNase V1 (Fig. 11). Continuous V1 cuts in a sequence U45–U51 suggested a stacking between the B1 and C arms of the structure. A symmetrical cleavage pattern in the two strands of

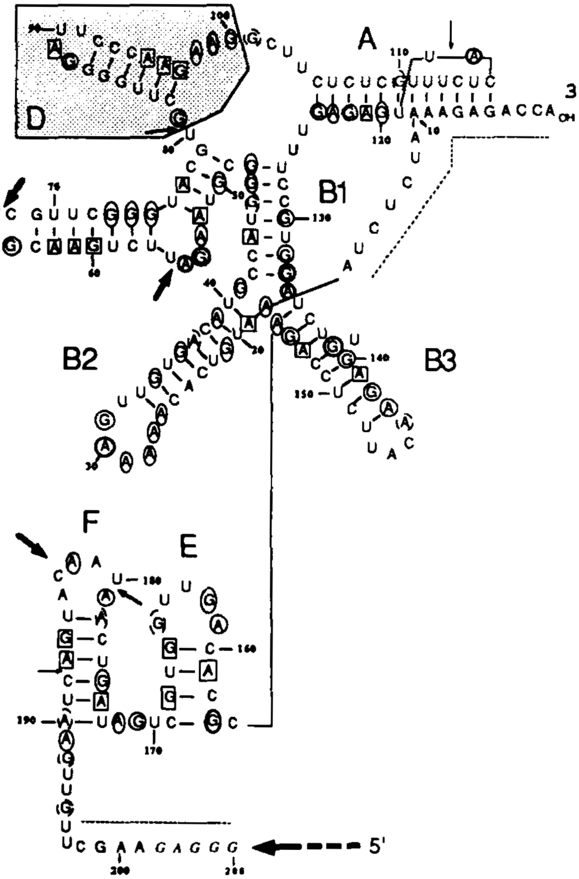

FIG. 10. Mapping of N-7 positions of purines in the 3′ end of the brome mosaic virus RNA with dimethyl sulfate and diethyl pyrocarbonate (reproduced from Ref. *48*). O, Reactive positions under native conditions; 0, nonreactive positions under native conditions, but reactive under semidenaturing conditions; □, nonreactive positions in both semidenaturing and native conditions. Bold, thin, and broken symbols correspond to strong, moderate, and marginal reactivities of the positions, respectively. Arrows indicate fragile sites in the RNA structure.

helices B and C indicated that both strands of the helices are accessible to the enzyme. In stem B3, only one strand was cut by RNase V1, which implies that another strand is not accessible for the enzyme. Figure 11 shows that the strongest V1 cuts are located at the most accessible external domains of the RNA model.

The results of probing the N-7 positions of the purines are shown in Fig. 10. All purines were reactive under semidenaturing and denaturing

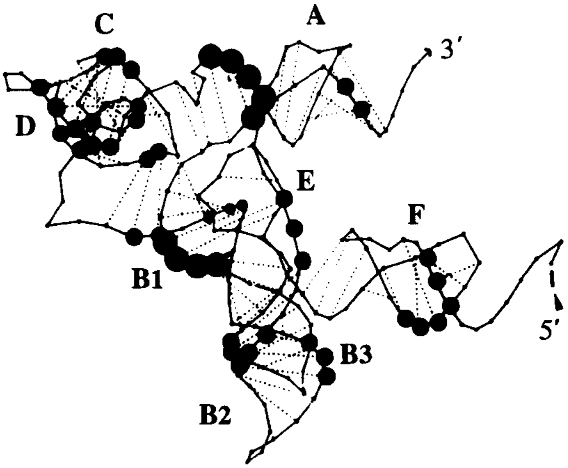

FIG. 11. Three-dimensional model of the 3′ end of the brome mosaic virus RNA. The dots indicate sites of cuts by ribonuclease V1. Sizes of the dots correspond to the intensities of the cuts.

conditions. In conditions stabilizing the native structure, the purines presented the reactivity pattern that in general fits the proposed structure. Exceptions were G13, G132, and G133, which were reactive in native conditions, indicating that helix B1 has a distorted conformation. The presence of V1 cuts between nucleotides G195 and U196 and between U197 and C198, together with the data on protection of A181 and A182 in the hairpin loop and U194 and G195 in the 5′ end of the RNA, provided evidence of the presence of a pseudoknot in which bases in the loop 181–184 bind to complementary sequence 194–197 in domain F, which was predicted from phylogenetic studies.

An example of tertiary long-range interactions identified in the molecule is a triple-helical region involving interactions (G41·A143)·A18 and (C42·G133)·A17, which include Hoogsteen binding of A17 and A18 with G133 and G41 in the major-groove side of the base-pairs C42·G133 and G41·A134. The data on the A18, G41, and A134 reactivities confirm the existence of these interactions.

2. APPROACHES FOR INVESTIGATION OF ELEMENTS OF TERTIARY STRUCTURE OF RNA

A straightforward approach to investigation of the tertiary structure of RNA is the identification of residues accessible at the surface of the molecule and residues buried within the structure. For such studies, reagents with broad specificity are needed that allow one to probe all types of nucleotides and that can potentially react with nucleotides irrespective of their involvement in the secondary interactions. A few reagents do allow probing the accessibility of universal constituents of RNA structure. Ethylnitrosourea reacts with phosphates in single-stranded and in double-stranded regions of RNA, if they are not buried in the structure (*10, 13, 32*). Figure 12 shows a good correlation between the experimentally determined reactivities of phosphates in $tRNA^{Phe}$ and the theoretically calculated availabilities of the phosphates to small probes, taking into account the geometry of the molecule and electrostatic factors (*12, 51*). Positions of the most well-protected phosphates are shown in Fig. 3.

Another small probe that allows investigation of the accessibility of a universal constituent of the RNA structure, ribose, is the OH radical produced by the Fe(II)–EDTA complex. This probe has little specificity for RNA sequence or secondary structure, making it an attractive probe for the tertiary structure of RNA. It was tested on yeast $tRNA^{Phe}$ (*41*) and it was shown that, in the native tRNA, riboses in the core of the molecule are protected from the modification. The reaction was used for investigation of structure of the self-splicing intron of *Tetrahymena thermophila* (*41*) and for monitoring the folding process of this catalytic RNA by probing the RNA structure in solutions containing different concentrations of metal ions (*42, 43*). From the modification data, an interior–exterior surface map of the folded RNA was constructed.

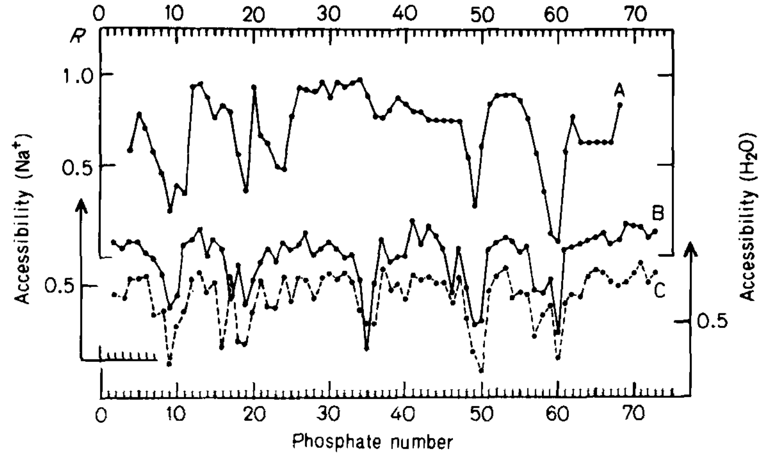

FIG. 12. Comparison of reactivities of phosphates in $tRNA^{Phe}$ toward ethylnitrosourea (A) with calculated accessible areas of the anionic oxygens of the phosphates for Na^+ ions (B) and for water (C) in the tRNA structure. Reproduced from Ref. *32*.

The determination of positions where the phosphodiester backbone of the RNA is on the inside or on the outside of the molecule provides constraints for modeling the three-dimensional structure of the ribozyme. It was found that the overall tertiary structure of this RNA forms cooperatively with the uptake of at least three magnesium ions, and that the high-order RNA foldings produced by Mg, Ca, and Sr ions are similar. Also, local folding transitions display different metal-ion dependencies, suggesting that the RNA tertiary structure assembles through a specific folding intermediate before the final structure is formed. The Fe(II)–EDTA cleavage was also used for probing the structures of mutated *Tetrahymena* ribozymes and to explore the role of individual structural elements in the tertiary folding of the RNAs. The results have allowed identification of different mutations that destabilize folding of the RNA and shift the optimal conditions of folding of the catalytic core to higher $MgCl_2$ concentrations (*43*).

Quantitative characterization of the reactivities of nucleotides toward a chemical agent when the chemistry is well-known provides high-resolution data on aspects of RNA structure such as the involvement of specific bases in stacking, local distortion of double helices, or even variations of the parameters of the RNA helices. Studies on RNA modification with DEPC have established a correlation between DEPC reactivity that reflects the accessibility of the major groove and the presence of either true or effective helix ends (*38*). In unstructured RNA, reaction with DEPC yields an even pattern of cleaving at each purine position, with adenosine being more reactive than guanosine. Purines in the uninterrupted helix are essentially unreactive toward DEPC.

Major groove inaccessibility is a result of the close approach of the phosphoribose backbone for helices of six or more base pairs. The minimum distance between phosphates across the major groove is approximately 10 Å for 7 bp, yielding a 4-Å groove width. This size of groove is not sufficient for the reagent to approach the reactive centers in a proper way. When the regularity of the helix is interrupted at the duplex termini, accessibility can be increased.

In the RNA duplex in the folded RNAs, susceptibility of each of the purines to the modification is strongly affected by position relative to helix termini, bulge, or internal loops. Asymmetric internal defects disrupt stacking regularity more than symmetric loops. Strongly coupled helices incorporate interhelix defects into a regular helix stacking geometry and are inaccessible to DEPC. A single-nucleotide bulge enhances accessibility in the major groove only modestly. Reactivity toward DEPC increases smoothly near helix termini and adjacent to the mentioned internal defects of the helix. Positions around large bulges are readily modified by DEPC. The accessibility extends one or two nucleotides further in the 5′ direction relative to a loop (equivalent to the 3′ helix end) than in the 3′ direction. However, the most reactive position is the purine 3′ to the loop (equivalent to the 5′ helix end), which is only half as accessible as the purines in the single-stranded RNA.

These differences in reactivities of purines can be explained by the geometry of the RNA helix. In the duplex terminus, the distance between phosphates across a single base-pair is 18 Å, which means a 12-Å groove width. Bases in an RNA duplex are tilted approximately 19° relative to the helix axis. Therefore, bases near the 3′ end of the helix protrude from the major groove envelope, which enhances their accessibility (38). The observed stronger modification at the 5′ base can be attributed to differences in stacking interactions, which makes the 5′ base relatively more accessible.

3. Cross-linking and Intramolecular Modifications

An efficient experimental method for determining structural relationships between different parts of an RNA chain is chemical cross-linking. Sometimes cross-linking can be obtained by direct photoactivation of juxtaposed residues of RNA. In two cases, the exact chemical nature of the cross-links has been determined. Thus, in the case of bacterial tRNAs containing a s^4U residue in position 8, photoinduced cross-linking[2] between this minor nucleoside and cytidine C-13 occurs under irradiation with UV light (330 nm) (52, 53). The reaction is facilitated by a favorable relative orientation of the two residues in the tRNA structure. Another example is the UV-induced formation of the C48-U59 cyclobutane dimer in yeast $tRNA^{Phe}$ (54). In the folded structure of this tRNA, the two pyrimidine rings are adjacent to one another in the folded structure and their 5,6 double bonds are nearly parallel and juxtaposed. This allows efficient cross-linking to occur following irradiation with short-wavelength UV light. The efficiency and kinetics of these cross-linking reactions provide a simple approach for comparison of the state of structure around the reacting residues in different tRNAs and the effect of different conditions on tRNA structure. Experiments with different tRNAs show that the reaction is very sensitive to structural changes involving the nearby pyrimidines and therefore can be used for analysis of conformational state (53).

Cross-linking of proximal groups in RNA structure can be achieved by use of bifunctional chemical reagents (Fig. 13). Bifunctional reagents of variable sizes can be used as "molecular rulers" for the identification of groups at specific locations in an RNA. Bifunctional reagents that have been used for probing RNA structure include derivatives of psoralen (55, 56), *N*-acetyl-*N′*-(*p*-glyoxylylbenzene)cystamine (56), and bis-(2-chloroethyl)methylamine (57,

[2] See essay by E. I. Budowsky and G. G. Abdurashidava in Vol. 37 of this series [Eds.].

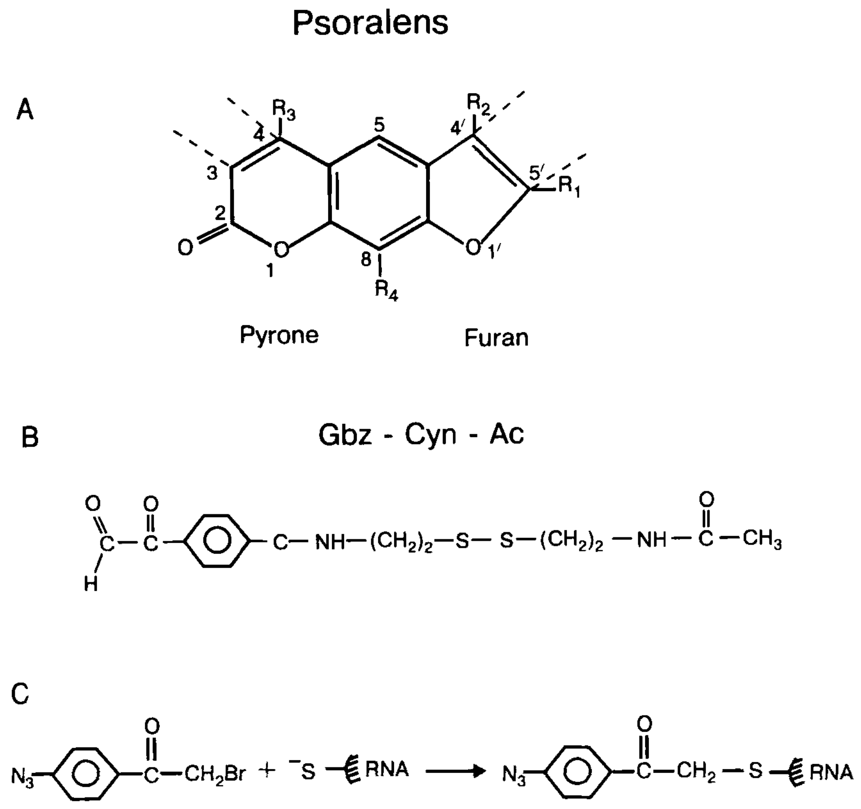

FIG. 13. Bifunctional reagents for cross-linking RNA. (A) Psoralens: 4,5′,8-trimethylpsoralen, $R_1 = R_3 = R_4 = CH_3$, $R_2 = H$; 4′-(hydroxymethyl)-4,5′,8-trimethylpsoralen, $R_1 = R_3 = R_4 = CH_3$, $R_2 = CH_2OH$; 8-methoxypsoralen, $R_1 = R_2 = R_3 = H$, $R_4 = OCH_3$. (B) *N*-Acetyl-*N*′-(*p*-glyoxylylbenzoyl)cystamine (Gbz-Cyn-Ac). (C) Attachment of a photoreactive group to the phosphorothioate at the 5′ terminus of an RNA transcript (*66*).

58). This latter reagent can alkylate heterocyclic bases within the same polynucleotide chain or bases in two juxtaposed chains, when the distance between the reactive centers is less than 12–15 Å. The reagent shows little sequence specificity; it can cross-link residues located in single-stranded and double-stranded regions and form cross-links that are stable during analysis. An example of application of the reagent is the investigation of the structure of the small nuclear RNAs U1 and U2 (*57*, *58*). In both RNAs, formation of two intramolecular cross-links was observed between residues for a part in the primary sequences of the molecules. The identification of the positions of the reactive residues was performed using the following procedure. After the reaction, individual forms of the cross-linked RNAs were isolated by electrophoresis. The RNAs were 3′ end-labeled and the position of the modified residue closest to the 3′ end was identified by comparing the products of partial enzymatic digests of the modified and the intact RNAs. Similar experiments with the 5′ end-labeled RNAs allowed localization of the second modified residue, which is closer to the 5′ end of the molecule. These studies have provided the information needed for reconstruction of the tertiary structures of the U1 and U2 RNAs.

Derivatives of psoralen are widely used for photocrosslinking of nucleic acids (*59*). These three-ring heterocyclic compounds (Fig. 13) intercalate into double-stranded regions of RNA; on irradiation with UV light (365 nm), they undergo a photochemical addition to heterocyclic bases of RNA located at a distance of about 8 Å. The intercalated psoralen attaches covalently to pyrimidine nucleosides, especially to uridine, by cyclobutane linkages to one nucleotide, producing a monoadduct, or to two nucleosides, producing cross-links. On activation with UV light, either the 3,4-pyrone double bond or the 4′,5′-furan double bond of psoralen photoreacts with the 5,6 double bond of a pyrimidine to form a pyrimidine–psoralen monoadduct. Reaction of the 3,4-pyrone double bond destroys the coumarin nucleus and leads to monoadduct formation. If the reaction occurs with the 4′,5′-furan double bond, the coumarin nucleus of the compound remains intact and can absorb light at 365 nm for its 3,4-pyrone double bond to react with the 5,6 double bond of another pyrimidine to form a cross-link.

Psoralen derivatives prefer to react with uracils near internal loops but not in loops or within perfect double helices (*60*). RNA molecules cross-linked by psoralen can be fractionated by gel electrophoresis in denaturing conditions. Although psoralen does not modify positions participating in Watson–Crick pairing, the cyclobutane adduct terminates reverse transcription because of the change in the uracil geometry. Therefore the sites of cross-links can be determined by primer-extensions with reverse transcriptase.

An example of cross-linking with psoralen for structural studies is the investigation of secondary structure of the SP6/mouse insulin precursor RNA (*55*). The RNA was treated with psoralen under conditions providing statistically less than one cross-link per RNA molecule, and individual fractions of the cross-linked RNA were isolated by gel electrophoresis in denaturing conditions. These RNAs were used as templates for reverse transcription to identify the cross-linking sites. A series of long-range contacts were detected within the 5′-half of the pre-mRNA that contains the intervening sequence. Because some of the interactions showed common sites, it was concluded that the RNA exists as a mixture of conformers. The pre-mRNAs with psoralen cross-links in different positions were used as substrates for *in vitro* splicing, and it was found that psoralen cross-linking of nucleotides in any of the double-stranded regions of RNA inhibited splicing, suggesting that a destabilization of secondary structure of the RNA precursor is required for splicing to occur *in vitro*. It was concluded that a melting of double-stranded regions within the pre-mRNA occurs before the endonucleolytic cleavage.

Cross-linking with psoralen has been used for investigation of the structure of 16-S RNA (*56*). Cross-linkage maps were generated for isolated 16-S RNA and for 16-S RNA within a 30-S ribosomal subunit. It was concluded that in both cases the RNA has equivalent regions of secondary structure. The data of cross-linking with different reagents were used for building a detailed model of 16-S RNA (*61*).

Cross-linkable groups can be introduced artificially in the single-stranded regions of RNA. This technique uses as a reagent *N*-acetyl-*N'*-(*p*-glyoxylylbenzoyl)cystamine (Gbz-Cyn-Ac; Fig. 13). This compound reacts with accessible guanosine residues in the single-stranded regions of tRNA by its glyoxal group, as shown in Fig. 6 for α-ketoaldehyde compounds. The adducts are stabilized by oxidation of their *cis*-diol groups to form *N*-acylguanosine derivatives; the disulfide bond of the derivatives is then reduced with sodium borohydride. After the reduction, each derivatized guanosine carries a free SH group (*62*). Treatment with hydrogen peroxide leads to formation of disulfide bonds between the modified guanosines, which are within a distance of approximately 17 Å. Cross-linking achieved by using this procedure allows identification of guanosines in single-stranded regions of RNA that are near one another in the folded molecule.

Identification of cross-linking sites is simplified considerably when a bifunctional reagent is attached in a specific position of an RNA. This approach is less general, but it allows a detailed study of the geometry of a specific part of an RNA. The technique was tested first on tRNAs. An aromatic 2-chloroethylamine residue, chlorambucil, was attached to the amino group of the amino-acid residue in aminoacylated yeast $tRNA^{Val}$ (*63*). Intramolecular alkylation in this modified tRNA occurred within its acceptor stem: at the 5' phosphate and at residues of the CCA end in accordance with the solution structure of tRNA in which the CCA stem does not contact other parts of the molecule.

The approach was developed further in experiments with $tRNA^{Phe}$, to which a chlorambucil residue was conjugated via linkers of different lengths (*64*). A rigid linker in the constructs of the general formula chlorambucil-(prolyl)$_n$-[^{3}H]phenylalanyl-$tRNA^{Phe}$ allowed variation of the probing group by changing the number of the prolyl residues. In the constructs with maximal length of the "molecular ruler" ($n = 15$, the distance between the 3' end of the molecule and the alkylating group is 62 Å), intramolecular alkylation of guanosine G20 in the D loop (located 60 Å from the 3' end of the molecule, according to the X-ray data) and rare nucleoside W (wyosine) in the anticodon was observed. The results are consistent with the known parameters of the tRNA structure.

Probing the tertiary structure of RNAs using autocleavage by the Fe(II)–EDTA group attached to a specific position has been tested in experiments with the yeast tRNA^{Phe} (*65*). The modified molecule was constructed by chemical incorporation of an EDTA-linked uridine into the 3′ half-fragment of the tRNA at position 47. This modified 3′ half of the tRNA was ligated enzymatically to the 5′ half of the tRNA by T4 DNA ligase. The produced molecule was cleaved by lead ions (this cleaving probe is in Section I,D,4) similarly to the intact tRNA^{Phe}, which indicated that the uridine modification did not disturb folding of the molecule. Autocleavage of the molecule by the tethered group in the presence of Fe(II) and a reducing agent produced a set of fragments that were in general agreement with the three-dimensional structure derived from X-ray analysis. Because the cleavage is produced by diffusing species and because not every ribose at a fixed distance has identical reactivity toward the radicals, quantitative characterization of reactivities of individual riboses was not possible. However, for large RNAs the low-resolution information available from such experiments may be sufficient to discriminate between different structural models.

A method has been elaborated for the attachment of a photoactivatable cross-linking agent, the azidophenacyl (APA) group, to RNA molecules (*66*). To prepare the 5′ end-labeled RNA, the RNA transcript was prepared using T7 RNA polymerase in the presence of guanosine monophosphorothioate. Inclusion of guanosine monophosphorothioate in a transcription reaction results in its incorporation only at the 5′ end of the transcripts, because nucleoside monophosphates can initiate transcription but cannot be incorporated in the growing RNA chain. The phosphorothioate provides a unique site in the RNA for the conjugation of the APA residue (Fig. 13). This method was used first for conjugation of APA to the 5′ terminus of tRNA, to prepare an affinity reagent for determination of sites in RNase P RNA that interact with tRNA substrate. Later it was shown that attachment of the APA group to specific sites in any part of RNA molecules can be achieved by tethering the group to the 5′ terminus of circularly permuted RNA analogs (cpRNAs) (*66, 67*).

This methodology has been used for investigation of the three-dimensional structure of the catalytically active RNA component of ribonuclease P from *Escherichia coli* involved in tRNA maturation and for investigation of the tertiary structure of the RNase P RNA complexed to the tRNA^{Asp}.

The circularly permuted RNA analogs of RNase P were molecules represented by RNase P RNA with 5′ and 3′ ends connected with a nonnative

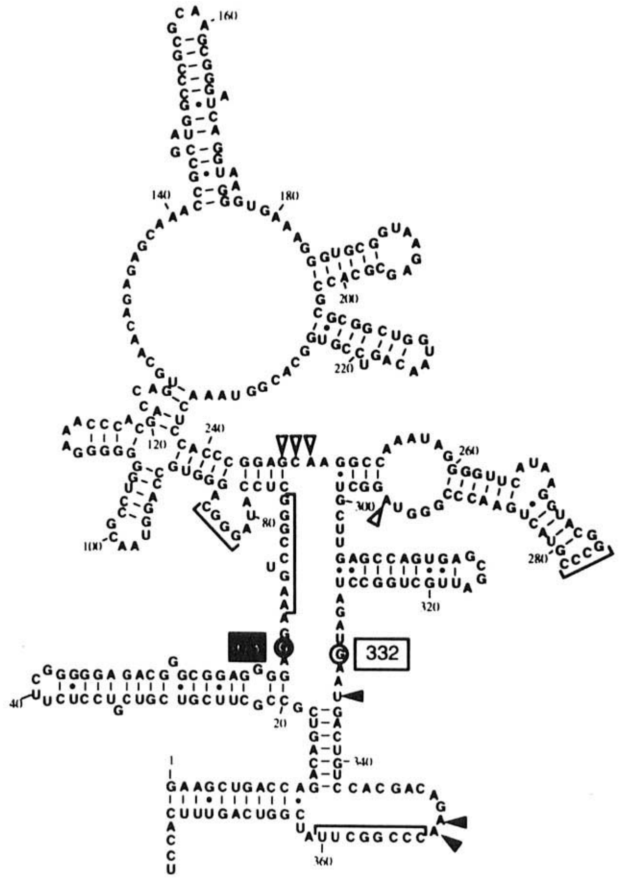

FIG. 14. Mapping of cross-links in circularly permuted RNase P RNA analogs. Reproduced from Ref. *68*, M. E. Harris, J. M. Nolan, A. Malhotra, J. W. Brown, S. C. Harvey and N. R. Pace, *EMBO J.* 13, 3953 (1994), by permission of Oxford University Press. Filled arrowheads indicate the positions of nucleotides attacked by the reactive group attached to the nucleotide shown by the filled circle. Open arrowheads indicate the residues modified by the group attached to the residue indicated by the open circle.

oligonucleotide linker (Fig. 14). The molecules contained discontinuities in the ribose phosphate backbone. Positions of the discontinuities were dictated by the DNA templates from which the RNAs were transcribed by T7 RNA polymerase. The end points were the specific photoreagent attachment sites for intramolecular cross-linking. Several cpRNase P RNAs with photoreactive groups located in different positions of the structure were prepared. The modified RNAs retained catalytic activity comparable to that of the natural RNase P RNA, proof that single interruptions introduced in the phosphoribose backbone and attachment of the APA did not alter significantly the RNA structure.

The modified RNAs were subjected to irradiation with UV light to convert the azido groups to nitrenes and produce cross-links. The cross-linking resulted in formation of "lariats," which were separated from the noncross-linked RNAs by gel electrophoresis. The particular nucleotides cross-linked to the 5′ ends of the cpRNase P RNAs were determined by the primer-extension method. Investigation of the cross-linking has allowed determination of orientation and distance constraints between elements in the RNase P RNA and within the RNase P RNA–pretRNA complex. The cross-linking data together with the established secondary structure of RNase P RNA and the tertiary structure of tRNA were used with a molecular mechanism protocol to develop a model of the global structure of the core of the RNase P RNA–pretRNA complex (*68*).

Reactive groups can be introduced in any selected position of RNA by means of affinity modification with reactive derivatives of corresponding complementary oligonucleotides (*69*). RNA is alkylated with oligonucleotide derivatives bearing an aromatic 2-chloroethylamine at the terminal phosphate (Fig. 15). After the reaction, the modified RNA is incubated in mild acid conditions in which the phosphoramide bond between the reactive group and the oligonucleotide is hydrolyzed. The result is that a residue with aliphatic amino group is introduced in a specific position of the RNA structure. These groups can be reacted with the bifunctional reagent 2,4-dinitro-5-fluorophenylazide to attach the photoreactive azido group.

4. Probes Sensitive to Specific Elements of RNA Folding

Some probes allow testing the state of the global structure of RNAs. One example is the photocrosslinking between pyrimidines in specific positions of some tRNAs (*52–54*), which occurs only in the native tRNA structure and can serve as a test for maintenance of the biologically active conformation. Another example of such probes is represented by cleavage of RNA with certain metal ions. Scission of RNA by coordinated metal ions is a simple and sensitive test for detection of the cation-binding regions and for probing the state of the RNA structure. Thus, highly specific hydrolysis of some tRNAs by Pb^{2+} occurs due to the presence of tight metal-binding sites in the RNA (*70–73*). The cleavage results in the formation of 2′:3′-cyclic-phosphate and 5′-hydroxyl termini. Cleavage of $tRNA^{Phe}$ between nucleotides U17 and G18 was a sensitive way to identify and correctly position the two lead-coordinated pyrimidines. Nucleotide substitutions that disrupted the tertiary interactions of $tRNA^{Phe}$ reduced the rate of cleavages dramatically. This

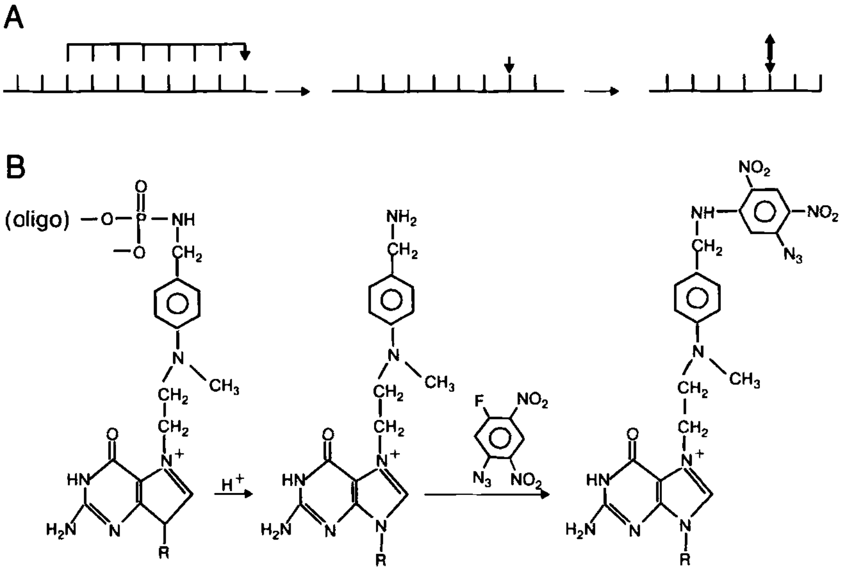

FIG. 15. Attachment of reactive groups to arbitrary sites in RNA by means of affinity modification with derivatives of complementary oligonucleotides. (A) Schematics of the procedure. (B) Chemical reactions used for attachment of photoreactive groups to specific guanosine residues in RNA.

cleavage reaction has been exploited as a sensitive probe for the tertiary folding of RNA variants (*71*, *72*).

An *in vitro* selection method has been developed to obtain RNA molecules that specifically undergo cleavage by Pb^{2+} ions (*74*). This selection method was applied for identification of different RNA motifs sensitive to cleavage by Pb^{2+} ions (*75*). The ability of $tRNA^{Phe}$ to undergo a specific cleavage in the presence of Pb^{2+} ions was also used as a selective pressure for isolation of RNA molecules having a core part similar to that of natural $tRNA^{Phe}$. In these experiments, tRNA molecules with the anticodon hairpin replaced by some artificial sequences were constructed. From the rates of the site-specific cleavage by Pb^{2+} and the formation of specific UV-induced cross-links, it was concluded that certain tetranucleotide sequences can allow proper folding of the rest of the tRNA molecule (*76*).

The 5-S RNAs from a few bacterial species have been characterized by Pb(II)-induced hydrolysis. Investigation of the cleavages has allowed a refinement of the secondary structure model of 5-S RNA. The effect of binding ribosomal proteins L18 and L25 to the *E. coli* 5-S RNA on RNA cleavage was also investigated. Besides the shielding effect of the bound proteins, a highly enhanced cleavage in the RNA, between A108 and A109, was detected. This finding has supported the concept that the major L18-induced conformational change involves portions of helices A, B, and D of the RNA (*77*).

Cleavage of the RNase P RNA with different metal ions has been investigated in detail (*78*).-A number of cations hydrolyze the RNA and five preferential cleavage sites have been characterized. Pb^{2+}-induced hydrolysis was suitable to sense different conformations of RNase P RNA (*79*). Good correlation of susceptibility to Pb^{2+} cleavage with catalytic activity was shown for the *T. Thermophilus* RNase P RNA under activity-assay conditions. This allows use of the test for studying conformation states of the RNA, to probe enzyme–substrate complexes, and to evaluate different salt and temperature conditions in reactions catalyzed by RNase P RNAs. The Pb^{2+}-cleavage assay was also applied to probe the tertiary structure of mutant RNase P RNAs. RNase P RNAs from three phylogenetically disparate organisms, *Chromatium vinosum, Bacillus subtilis*, and a few mutants from *E. coli* with deletions, were studied. Investigation of the patterns revealed some regions of identical structure that provide evidence for several ubiquitous metal-ion binding-sites in eubacterial RNase P subunits (*79*). Two cleavage sites occur at homologous positions in all the native RNAs regardless of sequence variations, suggesting common tertiary structural features. Such conservation in structure suggests that these regions are involved in some specific role of the RNA, for instance in substrate binding or catalysis. The cleavage sites in four deletion mutants of *E. coli* RNase P RNA differed from the native patterns, indicating alterations in the tertiary structures of the mutant RNAs.

Some complexes of transition metals can shape selective photoinduced cleavage of structured RNAs (*80*). Tris(1,10-phenanthroline)ruthenium(II) [$Ru(phen)_3^{2+}$], tris(3,4,7,8-tetramethylphenanthroline)ruthenium(II) [$Ru(TMP)_3^{2+}$], tris(4,7-diphenyl-1,10-phenanthroline)ruthenium(II) [$Ru(TMP)_3^{2+}$], tris(4,7-diphenyl-1,10-phenanthroline)rhodium(III) [$Rh(DIP)_3^{3+}$], and bis(phenanthroline)(9,10-phenanthrenequinonediimine)rhodium(III) [$Rh(phen)_2phi^{3+}$] are complexes that bind to RNA at some sites matching their shape; on photoactivation, they induce RNA strand scission, thereby marking sites of specific structural features.

Cleavage of a few complexes has been assayed on yeast $tRNA^{Phe}$ and a distinctive diversity in site-selective cleavage was shown. The RNA was irradiated with UV light in the presence of the complexes and then subjected to aniline treatment. Reactions with $Ru(phen)_3^{2+}$ and $Ru(TMP)_3^{2+}$ resulted in cutting preferentially at guanosine residues and formation of RNA fragments with terminal 5′ and 3′ phosphates. In these cases, the proposed mechanism of the reaction was attack on the nucleic acid base in a reaction mediated by singlet oxygen generated by photoexcitation of the ruthenium complex.

A different cleavage chemistry was observed for rhodium complexes. No preferred base composition for the attack was observed and aniline was not required for fragmentation. It was concluded that, in this case, the photoinduced cleavage occurs through a direct oxidation path and the target of the reaction is the RNA sugar. Different patterns of cleavage were observed for complexes with different ligands, which apparently reflect differences in their binding characteristics governed by their different molecular shapes. The rhodium complexes demonstrate a pronounced preference for some sites in the central part of the RNA. This structural preference was governed not by the cleavage chemistry, but rather by the presence in this part of appropriate binding sites fitting the shape of the compounds. $Rh(DIP)_3^{3+}$ induces strong cleavages at residues Ψ55 and C70 with other weaker sites present at T54 and C56. The most interesting probe, $Rh(phen)_2phi^{3+}$, induces strong cleavages at residues G22, G45, U47, and U59. Under denaturing conditions, these sites were relatively unreactive, suggesting that the complex binds in the folded molecule to a unique region of RNA organized by parts of the D stem, T loop, and variable loop. The unusual reactivity pattern is consistent with recognition of a widened RNA A-like helix distorted by local formation of a base triplet that is open to permit intercalation by the bulky complex. $Rh(Phen)_2phi^{3+}$ was suggested to be a potential shape-selective probe targeting triple binding sites in RNA. This and other complexes may become useful for deducing tertiary structure features of RNA molecules.

II. Computer Analysis of the Secondary Structure of RNA

The problem of predicting secondary structure from nucleotide sequence dates back to the 1960s (*81*). The question was how to determine a folding that provides the largest number of complementary bases. As the number of known RNA sequences increases the matter of prediction of secondary structure becomes of particular concern.

Up-to-date algorithms relevant to the subject fall into three basic groups: (1) those searching for the lowest energy secondary structures with the use of the thermodynamic parameters (thermodynamic approach); (2) those determining RNA secondary structure by simulation of folding (kinetic approach); (3) those searching for invariant secondary structures by comparing sets of homologous sequences (comparative approach).

The thermodynamic approach has been well-described (e.g., see Refs. *82–84*). Thus we will only outline it and mention its recent modifications. What we focus on are the kinetic and comparative approaches to the prediction of RNA secondary structure. The statistical aspects of organization and evolution of RNA secondary structure are discussed as well.

A. Thermodynamic Parameters of RNA Secondary Structure

Understanding the RNA folding code requires determination of its thermodynamic parameters. It is equally important to know both the parameters of the helices and the parameters of loops of any kind. The respective contributions of helices and loops to structure stability are opposed (*82–84*): the formation of a helix lowers free energy, and the closing of a loop usually increases free energy owing to entropy losses. The resultant stability depends on the balance between these two opposing factors.

Helix stability strongly depends on the ratio of A–U to C–C. However, it was shown as early as 1963 (*85*) that more precise estimates of helix stability require accounting for the sequence of bases in the RNA primary structure. This effect is due to the contribution of stacking interactions between the neighboring base-pairs within the helix that exceeds that of H bonds (*86*). That is why helix energy is evaluated under the nearest-neighbor model (*87*). According to the model, helix energy is postulated to be the sum of the energies of nearest complementary pairs within the helix.

Traditionally the thermodynamic parameters of neighbor pairs are estimated from melting experiments on duplicates of short synthetic oligonucleotides of varying base content (*88*). Spectroscopic and calorimetric methods are used (*87*). Thermodynamic parameters are inferred from the experimental data by using a two-state model. Under this model, only the completely melted or helical states are considered, the intermediate pairing states being ignored (*87*). Optical and calorimetric data are identical unless the model falls short of applicability. Today there is a great variety of sets of thermodynamic parameters of canonical complementary pairs in the RNA helix with different neighborhoods (*82*, *88–93*). Table I presents the commonly used compilation of these parameters (*88*). Besides, thermodynamic parameters have been identified for the nearest-neighbor model of the noncanonical G·U pairs (*82*, *89*), as well as the stabilization parameters of the nucleotides adjacent to the helix (*82*).

Despite unquestionable attainments, the errors of determination of the thermodynamic parameters (entropy and enthalpy) are still high. Free-energy assessments are much more accurate because the errors of entropy and the errors of enthalpy to some extent compensate for each other. The errors result from the limited applicability of the statements used in assessment of the parameters, and a rather broad melting profile of short helices. Stacking interactions in single-stranded oligonucleotides may also contribute to the error (*94*, *95*).

TABLE I
FREE ENERGY INCREMENTS FOR RNA HELIX PROPOGATION[a]

	⟶	⟶	⟶	⟶	⟶	⟶	⟶	⟶	⟶	⟶
Propagation	AA	AU	UA	CA	CU	GA	GU	CG	GC	GG
sequence	UU	UA	AU	GU	GA	CU	CA	GC	CG	CC
	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵
ΔG°_{37} (kcal mol^{-1})	−0.9	−0.9	−1.1	−1.8	−1.7	−2.3	−2.1	−2.0	−3.4	−2.9

[a] In 1 *M* NaCl (*82*).

Reproduced, with permission, from the *Annual Review of Biophysics and Biochemistry*, Volume 17, © 1988, by Annual Reviews Inc.

The thermodynamic parameters of loops are also determined from calorimetric and spectroscopic measurements (*87*). Loop parameters are considerably less studied than those of helices. Studies on multiloops are few (*96*, *97*). Little is known of the effect the nucleotide context has on loop energy. Calculations are based on averaged data not dependent on nucleotide context. The data on very stable loops, the so-called tetraloops (*98*, *99*), are the only exception. Experimental estimates for loop free energy have been obtained only for those loops not longer than $n_{\max}$, where $n_{\max} = 5$, 6, and 9 for bulge, internal, and hairpin loops, respectively (*92*). For longer loops, the following approximation of free energy changes is used:

$$G(n) = G(n_{\max}) + \alpha RT \ln(n/n_{\max}),$$

where α is a parameter depending on which model of polymer chain is used [$\alpha = 1.5$ for a phantom chain (*100*)].

Table II[3] presents a compilation of loop free energies (*82*). There is a continuous updating of thermodynamic parameters for helices and loops; to highlight details, a special review is required. The available thermodynamic parameters of helices and loops allow prediction of low-energy secondary structure, their statistical analysis, and the simulation of RNA folding.

B. Thermodynamic Approach

The thermodynamic approach is based on the assumption that the native secondary structure of RNA is the lowest energy form or one of the suboptimal forms of secondary structure of this molecule. In the programs implementing this approach (*82*, *83*, *91*, *92*, *101–107*), the free energy of RNA is calculated by use of the above-described thermodynamic parameters characterizing helix and loop formation. Two main directions of searching for low-energy secondary structure have been suggested: combinatorial and recursive (*82*).

[3] Table II is on page 188.

As has been noted (*108*), the number of possible RNA secondary structures increases with sequence length N as $1.85^N/N^{3/2}$. Hence, any algorithm for predicting low-energy secondary structure is faced with a problem: how to examine all the possible secondary structures, because the sequences of just about 100 nucleotides in length produce a very large number of potential secondary structures.

A quantitative prediction of RNA secondary structure by searching for the lowest energy secondary structure was pioneered by Tinoco *et al.* (*109*). Thermodynamic parameters were first used to predict the secondary structure of the short fragment of R17 RNA. The change of free energy was taken as -1.2 kcal/mol for the formation of A·U pairs, and -2.4 kcal/mol for G·C pairs. The change of free energy for loop formation was also taken into account. The lowest energy secondary structure was determined by trying all the complementary pairs obtained from complementarity matrix analysis. Further development of new methods for the prediction of RNA secondary structure and for the determination of the thermodynamic parameters of secondary structure was quite explosive, and resulted in a great variety of methods for the prediction of low-energy secondary structure (*101–107, 110–112, 114–118*).

An efficacious step was made when it was decided to take a helix, not a pair of complementary nucleotides, as an elementary object for analysis (*101*). The routine included (1) the research for the longest potential helices, (2) analysis for their compatibility by applying stereochemical constraints (Fig. 16), (3) reconstruction of the secondary structure from compatible helices, and (4) calculation of secondary structure energy in accordance with thermodynamic parameters. The first rule (Fig. 16A) forbids the concurrent presence of any RNA fragment at two helices; the other disallows pseudoknot formation (Fig. 16B). The time of calculation depends on the nucleotide

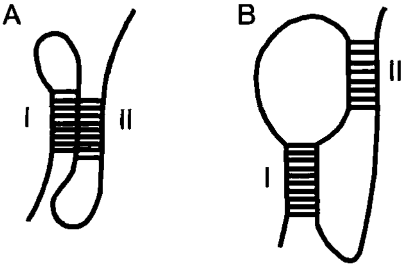

FIG. 16. Rules of stereochemical compatibility for helix pairs in the RNA secondary structure. (A) Ban on the concurrent presence of an RNA fragment in two helices; (B) ban on pseudoknots.

sequence length N as $N2^N$, which implies that the algorithm would not be applicable if sequences were longer than 70–80 nt. It was this algorithm that gave a start to the combinatorial approach to determination of the low-energy secondary structure of RNA by trying stereochemically compatible combinations of helices.

Use of the two rules of stereochemical compatibility accelerates examination of secondary structure remarkably, which is why they are now commonly exploited and form part of many later algorithms. However, they do not always reflect the features of real secondary structure. For example, if the termini of helices are partly open, two shorter nonoverlapping helices may exist (in violation of Rule 1).

This constraint was initially forestalled by regarding a "sliding" boundary between competing helices (*102*). Besides, examination of secondary structure was optimized by using routines of graph theory, which reduced time consumption to N^5 and allowed analysis of sequences of up to 150 nt in length.

Rule 2 (no pseudoknotting is allowed) is widely used in the most effective current means of secondary structure determination, the recursive algorithms (*103–107*). Pseudoknots are not allowed in these algorithms, which restricts their applicability. How to handle this restriction is described in Section II,E, where pseudoknot formation in RNA is discussed.

The key stage of secondary structure prediction by application of graph theory should be the construction of a graph of stereochemical compatibility, in which a helix is related to a vertex, two compatible helices being connected by an edge (*110, 111*). Finding low-energy secondary structure is equivalent to finding cliques. The method is good for prediction of secondary structure for RNA molecules up to 150 nt in length.

Another algorithm (92, *112*) for prediction of low-energy secondary structures applying graph theory was based on the well-known method of branches and boundaries. This method was applicable to RNAs up to 200 nt in length.

One of the advantages of the combinatorial algorithms is that it is possible to determine not only the lowest energy secondary structure, but also a set of suboptimal secondary structures. What makes the combinatorial approach difficult is the necessity of examining an enormously large number of structures, which is rapidly rising with RNA nucleotide length. This was what encouraged the development of the currently most effective and rapid methods for RNA low-energy secondary-structure prediction by the recursive approach.

The recursive algorithms (*102–105, 107*) use the ideas of dynamic programming. A recursive approach was initially applied in 1966 (*113*) to determine the RNA secondary structure with the maximum number of complementary pairs. As there were no thermodynamic parameters of RNA available at that time, further development of the method was suspended. Later use of the recursive algorithm for maximization of number of complementary pairs was independently suggested and mathematically substantiated (*106*). Further development resulted in the first recursive algorithm for determination of the lowest energy secondary structure of RNA, using thermodynamic parameters (*103*).

The basic principle of the algorithm can easily be understood from the following example. Let us define a structure with the maximum number of pairs. Let the nucleotide sequence be presented as a circle (Fig. 17), with the possible complementary pairs as the arcs connecting the respective nucleotides. Now consider a circle section B_xB_y of length p and define the maximum number, $M(x, y)$, of pairs in it:

$$M(x, y) = \max\begin{cases} M(x, k - 1) + M(k + 1, y - 1) + 1, \\ M(x, y - 1) \qquad x \leq k < y = x + p. \end{cases} \tag{1}$$

Increasing x and y, we try all the sections of the entire sequence. Then the routine is iterated on the p values with an increment of 1. It is important that $M(x, y - 1)$, $M(x, k - 1)$, and $M(k + 1, y - 1)$ from Eq. (1) be evaluated at the previous step, thus providing easy calculation of the matrix element $M(x, y)$. Calculations terminate when the section B_xB_y corresponds to the entire sequence B_1B_n. By then, a matrix K has been filled. The element K_{ij} is the count number of the nucleotide that, when paired with the nucleotide B_j, provides the optimal folding of the sequence B_iB_j. The structure with the maximum number can readily be deduced from the matrix K.

The algorithm for searching for the structure with minimum energy is in

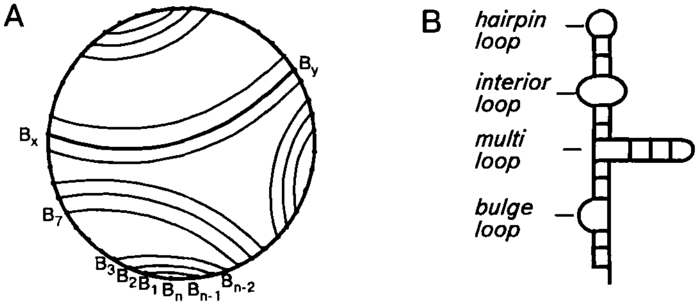

FIG. 17. Base-pairing in the planar secondary structure of an RNA. (A) Extended form; complementary base pairs are represented by arcs; nucleotides are designated by dots. (B) Condensed form of RNA secondary structure (*103*).

general the same, except that it is aimed at the lowest energy of an RNA section, not at the number of complementary pairs.

The same idea of recursive approach is exploited elsewhere. The suggested and widely used variants of the recursive algorithms (*104*) take account of the energy of helices as calculated by the nearest-neighbor method and the destabilizing energy of loops in accordance with thermodynamic parameters—those mentioned above and a range of others (*82*). On the whole, the works mentioned have given rise to a wide range of current recursive methods for determination of RNA secondary structure (*105, 107, 114–117*).

The main advantage of the recursive algorithms is that they are fast (*82*). Consumption time depends on sequence length as N^3; the secondary structure has been predicted for sequences up to 4217 nt in length (*118*). A remarkable advantage of the algorithms is that they easily make use of additional biological information, in particular the experimentally determined location of some nucleotides in single-stranded or double-stranded regions of the secondary structure (*104*). Thus, in this algorithm the thermodynamic parameters can be combined with various types of experimental data, which notably raises prediction accuracy. The common drawback of the recursive algorithms is that pseudoknots are not allowed in them, whereas they are a feature of real RNA secondary structure.

The characteristic feature of dynamic programming methods is that each step of analysis gives the only optimal structure. As a result, the initial versions of the above recursive algorithms produced only one optimal energy secondary structure. At the same time, the lowest energy secondary structure may not necessarily be the only functionally significant secondary structure for a given RNA molecule. There may also exist an ensemble of functionally important alternative secondary structures in dynamic equilibrium as, for example, the attenuators (*119*). It is also of importance that the RNA functioning in the "cellular context" interacts with other macromolecules (RNA, proteins, RNP particles, etc.), which can result in RNA refolding.

Finally, it is noteworthy that all the thermodynamic parameters used for determination of secondary structure are estimated with considerable errors. With loops, the error sometimes ranges between 15 and 50% (*120*).

It therefore seems reasonable to find a set of secondary structures within a certain energy window rather than the lowest energy secondary structure. Thus the probability of revealing the native secondary structure can be enhanced. Accordingly, the initial algorithm for prediction of secondary structure (*104*) was modified to target suboptimal secondary structures. The modified versions of the algorithm (*105, 114*) can bring up secondary structures within the window of energy defined by the user. Base-pairing probabilities can be determined with the energy weights of optimal and suboptimal structures (Fig. 18), the melting profile, and the RNA molecule specific heat (*121*). All of it can be effectively achieved by applying another dynamic programming approach (*122*), which calculates the secondary structure partition function and the probabilities of substructures. The partition function describes completely the equilibrium ensemble of secondary structure. Base-pairing probabilities, the melting profile of the RNA molecule, and other equilibrium parameters are also evaluated through the partition function (Fig. 19). The time consumed by the algorithm is estimated as N^3.

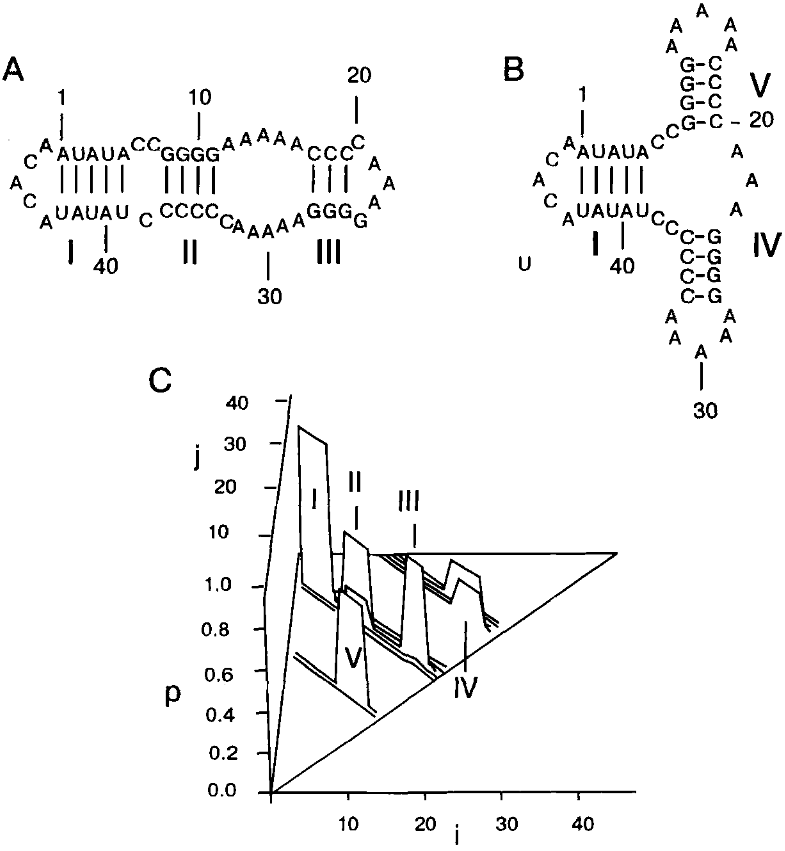

FIG. 18. Representation of the equilibrium distribution of secondary structures of a circular single-stranded RNA, 48 nucleotides long. Roman numbers I–V represent helices identical in each representation. (A) Optimal secondary structure at 35°C. (B) First suboptimal secondary structure. (C) Three-dimensional base-pairing plot at 35°C. Reproduced from Ref. *121*, M. Schmitz and G. Steger, *CABIOS* **8**, 389 (1992), by permission of Oxford University Press.

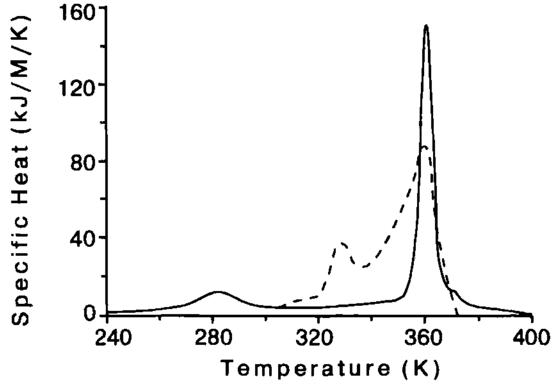

FIG. 19. The specific heat of *E. coli* 5-S RNA as a function of temperature [solid line represents the calculated curve (*122*), dashed line shows the experimental curve for the A form for 5-S RNA (*123*)].

C. Simulation of RNA Folding

The kinetic approach is based on the assumption that native secondary structure is the most kinetically attainable state of the RNA molecule and results from a multistage folding process (*124–126*). Levinthal was the first to deal with the problem of self-directed folding of a biopolymer into a unique spatial conformation (*127*). He noted that the total number of conformations of an N-monomer chain is g^N, where g is the number of stable conformations of the monomer. It is therefore clear that the native conformation of a long biopolymer cannot be achieved for any biologically reasonable time by randomly trying all the conformations. Initially, the idea was suggested for proteins (*128*). As first postulated (*129, 130*) and then demonstrated (*131, 132*), there must be a process underway such that the protein molecule passes a sequence of kinetically attainable states toward native spatial structure. This may also apply to RNA molecules. Presumably, during the formation of secondary structure the RNA molecule passes through kinetically attainable states with a gradual (in most cases) reduction in energy, certain elements of secondary structure being formed at each step (*133*). Generally, it is not yet clear if the secondary structure thus formed complies with the global minimum of energy.

The idea of secondary-structure prediction via folding simulation was initially realized in the simplest version (*134*). The suggested algorithm of predicting a secondary structure with the maximum number of complementary pairs was based on the step-by-step addition of new helices to the existing secondary structure. At each stage, there was added a helix with the maximum number of complementary pairs. A later algorithm (*132*) was based on the step-wise formation of secondary structure, the energy of which was calculated through thermodynamic parameters. At any step the structure would acquire such a helix from among all possible ones, whose formation would provide the maximum gain in free energy. Of a total of 80 tRNAs that had passed through the algorithm, 55% resulted in secondary structures of the "cloverleaf" type. A similar folding algorithm (*133*) used improved thermodynamical parameters. Its speed was N^2, which is much faster than the combinatorial and the recursive algorithms. Testing showed a good correspondence between the secondary structure predicted by this algorithm and the data on cleavage of RNA molecules by nucleases (*133*). Other algorithms have also been proposed for RNA secondary structure determination, on the basis of simulations of folding by step-wise addition of helices with the use of thermodynamic parameters (*135–138*).

Note that to add helices subsequently with the maximum gain in energy to secondary structure is in fact to select a way of folding characterized by maximum values of the equilibrium constants for the reactions of helix adding. Thus, under these algorithms, the travel of RNA through kinetically attainable states is substituted by a travel through local minima of energy.

In the mid-1980s an algorithm of RNA folding was suggested on the basis of the rates of helix formation and decay (*124–126*). Other algorithms of that kind have been implemented since (*139–143*). Under these algorithms, secondary structure formation is simulated by a Markovian random process. In the algorithm of Mironov and co-workers (*124–126*), the secondary structure transition from state to state leads either to the formation or decay of a helix. Which of the events will take place depends on the probabilities of the transitions, i.e., on the rates of the corresponding processes. It is assumed that the rate of decay of a helix of length N base pairs depends on its energy ΔG_h (*144*), which is calculated according to the nearest-neighbor model (*87*); the rate is given by Eq. (2):

$$k_d = Nk_0 \exp(-\Delta G_h/RT), \tag{2}$$

where k_0 ($= 10^{-6}$–10^{-8} sec^{-1}) is the rate of pairing of two bases adjacent to the helix (*145*). Note that the lifetime of a typical helix in the secondary structure is great. For example, Eq. (2) sets it at nearly one second for a helix of four G·C pairs at 37°C.

Helix formation is a kinetically controlled two-stage process. The rate-limiting stage is helix nucleus formation. First (after a spatial proximity has been achieved), a nucleus forms that is in fact one or several complementary pairs providing for sufficient stability of the intermediate structure. Second, helix formation proceeds rapidly by the "zipper" mechanism. That is why the rate of formation of a helix of length N is determined by losses of entropy

(ΔG_l) during the loop closure, which is calculated from tabulated values (*82*, *89*):

$$k_f = Nk_0 \exp(-\Delta G_l/RT). \qquad (3)$$

This way of determining the kinetic constants of helix formation or decay is the most reasonable because it satisfies the mass-action law.

Using the Monte Carlo method and Eqs. (2) and (3), secondary structure formation can be simulated for a time interval T. Multiple iterations would give the probability of certain secondary structures for the time $t < T$ (Fig. 20).

With this method, folding has been simulated for a number of tRNAs (*124*, *126*), leader and intercistronic regions of mRNA transcribed from the *E. coli atp* operon (*124–126*), and the self-splicing YC4 intron of mitochondrial RNA of fungi (*141*). In this case (*141*), the experimentally predicted and estimated RNA secondary structures were in a good agreement. With the aid of the kinetic approach, a correlation was revealed (*124*, *126*) between gene expression and RNA secondary structure. The correlation was found (*140*) between the refolding events of the RNA growing chain and the location of replication pause sites in MDV-1 RNA (*146*).

As is seen from Eqs. (2) and (3), the kinetic constants of helix formation or decay are strongly dependent on the length of helices and loops. Computer simulation of the folding of random RNA chains by the method described above demonstrated a power-law growth of the average time of conformational rearrangements (*142*). In fact, this implies that, in the course of folding, the secondary structure becomes more and more "frozen," which shows up in difficulties of large conformational rearrangements.

Accordingly, the kinetic algorithm (*139*, *143*) was modified so as to account for a division of the kinetic ensemble of secondary structures into clusters. A cluster involves the conformations that are similar topologically and apt to undergo rapid transition into one another. From this viewpoint, the simulation of folding reduces to the description of kinetics of intercluster transitions. This simplified algorithm is more effective and allows calculation of the secondary structure of RNA sequences of hundreds of nucleotides in length, using personal computers (*139*).

D. Comparative Approach

Under the comparative approach, the matter of the native secondary structure of RNA correspondence to the global minimum of energy or to a kinetically attainable state is not considered. However, it is assumed that a functionally significant form of secondary structure must be evolutionarily conservative. This approach was applied to reconstruction of the secondary structure of practically all the classes of RNA shown experimentally to pos-

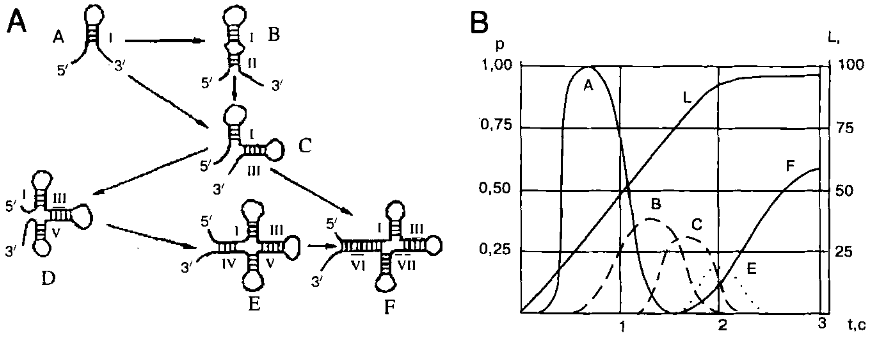

FIG. 20. The kinetic ensemble of RNA secondary structures of the pre-tRNA$^{\text{Ala}}$ of *Bombyx mori*. (A) Secondary structure folding in the course of RNA chain elongation. (B) Probability distribution on the kinetic ensemble. The *L* curve designates the increase of RNA length with time during transcription (*124*).

sess a secondary structure: tRNA (*147*), 5-S RNA (*148*), 16-S RNA, 23-S RNA (*149*, *150*), and many others.

The programs implementing the comparative approach are aimed at determination of evolutionarily invariant secondary structures in the families of isofunctional molecules of RNA. The known basic mechanism of secondary structure conservation in the course of RNA evolution is the fixation of so-called compensatory substitutions in the sequences, that is, nucleotide substitutions retaining helices.

With three types of complementary pairs (A·U, G·C, and G·U), 15 variants of compensatory substitutions are possible, i.e., substitutions of one complementary pair for another (*151*). Of them, 11 are double substitutions (for example, A·U to G·C) and 4 are single (for example, A·U to G·U).

The probability of a double compensatory substitution in the helical part of RNA due to random combination of single nucleotide substitutions is 0.256; the probability of a single compensatory substitution under the same condition is 0.125. With these estimates, 28-S ribosomal RNAs of vertebrates were studied (*151*). It was shown that the observed numbers of single and double compensatory substitutions in the helical parts of these RNAs were significantly higher than could be attributed to chance. Thus, fixation of compensatory substitutions in the helical parts of RNA is a mechanism that keeps RNA secondary structure evolutionarily conserved. Studies of 5-S and 5.8-S RNA molecules from different organisms (*152–154*) also provide evidence for the elevated number of compensatory substitutions within the helical parts of RNA compared with a random level.

Thus, the compensatory substitutions are the basic mechanism of maintaining the evolutionary conservatism of RNA secondary structure and to detect these substitutions is a common approach to predicting RNA secondary structure elements. The choice of algorithm for prediction of secondary structure essentially depends on the degree of homology between the sequences under comparison.

At a high level of homology, the method based on examining the aligned sequences is effective (*148*). All the possible locations of a pair of windows of size *W* are tried at a fixed distance *R* between them. For each location of windows, a helicity index is calculated. This index is defined as the number of the aligned sequences that have mutually complementary fragments corresponding to some helix of length not exceeding *W* in the given pair of windows. The maximum values of the helicity index indicates that there are invariant helices formed by complementary fragments about *R* bases apart. By interactively locally improving the alignment of the fragments comprising the helices, it is possible to increase the value of the helicity index. On the helicity plot (Fig. 21), it is signaled by a less fuzzy peak corresponding to the helix in question. By varying the distance *R* it is possible to reveal all the RNA regions that contain highly conserved helices. Figure 21 presents the superposition of the main peaks of the helicity index obtained by analyzing the 5-S RNA from *E. coli* and related organisms. Peaks I–IV correspond to four canonical helices of this RNA secondary structure that are invariant in most of the sequences studied. Besides, less clear-cut peaks (A–C) have been revealed. They correspond to the helices that are invariant within separate subgroups of the set of sequences in question.

There are also available methods for the automated secondary structure reconstruction based on alignment data. With these methods (*155*), a matrix of complementarity is constructed for the family of *M* aligned sequences of length *N*. One of the following symbols is assigned to an element BP(*i*,*j*) of the matrix: **o** stands for nucleotides *i* and *j* that are complementary in all the sequences and these positions are absolutely conservative; * stands for those that are complementary in all the sequences, and compensatory substitutions are observed at these positions; w indicates those forming a G·U pair in most of the sequences; + is for those forming a complementary pair in most of the sequences (over some threshold), but not in all of them; and • indicates those with a considerable number of noncomplementary pairs (over some threshold).

Information on all the invariant helices for the family of sequences under consideration is coded in the matrix BP (*i*,*j*). To each invariant helix *S* in this matrix, there corresponds a continuous track of symbols **o**, *, w, or + normal to the diagonal. On calculating helix length *h*, loop length *l*, and defining the number of symbols * (*A*), w (*B*), and + (*C*), it is possible to estimate the "pseudopotential" *F* for the helix S. In general, the longer the helix and the higher the number of positions with compensatory substitu-

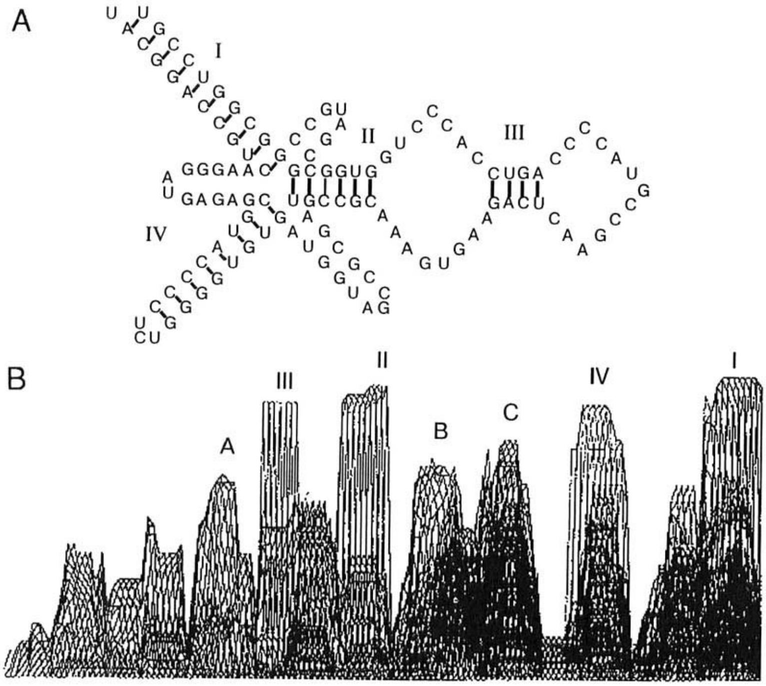

FIG. 21. Reconstruction of the secondary structure of 5-S RNA by comparison of the sequences from *E. coli* and related organisms (*148*). (A) The secondary structure of *E. coli* 5-S RNA. (B) The superposition of helicity peaks for the set of aligned sequences.

tions in it, the greater the value of pseudopotential $F(S)$. On the other hand, the higher the number of G·U pairs or noncomplementary pairs and the longer the loop formed, the less the pseudopotential $F(S)$. For example, $F(S) = h + A - 0.5B - 2C - 0.051$ proved good for revealing invariant helices (*155*).

Then an invariant secondary structure is constructed by step-wise selection of the helices with the maximum value of $F(S)$. In fact, this is a sort of simulation of RNA folding with only invariant helices allowed. The procedure follows the standard rules of stereochemical compatibility of helices (Fig. 16). A testing of the algorithm on the known secondary structures as tRNA, 5-S RNA, and 16-S RNA showed a good agreement between the predicted and canonical structures.

In fact, the method at issue (*155*) is the first practically effective comparative approach regarding storage capacity and time consumption. The time required for finding the invariant secondary structure for M sequences of length N is here $MN^2 + N^3$, and storage capacity is N^2 (*155*). Interestingly, the invariant RNA secondary structure reconstructed with this method (*155*) for the TAR fragment of 200 nt from type 1 human immunodeficiency virus (HIV-1) coincides with the lowest energy structure predicted for this fragment by the FOLD program (*155*). This result exemplifies that in some cases evolutionarily invariant secondary structure corresponds to the lowest energy form.

If homology of the considered RNA molecules is not high, analysis of multiple alignment is not applicable to determining the invariant secondary structure. If so, the following approach may be effective: (1) search for a set of suboptimal secondary structures for each RNA under study by one of the thermodynamic or kinetic methods; (2) compare the sets and the choice of secondary structure that is invariant for all sequences under comparison. Many methods have been proposed for comparison of secondary structures within the frame of this approach. All of them depend on the manner in which the secondary structure is coded and on further comparison of the resulting codes.

Under one of the versions, hairpin, internal, bulge, and multiple loops relate to tree vertices, and helices to edges (*157–159*). Under another version (*160*), helices relate to vertices, and loops to edges. Thus constructed, the trees can be compared by one of the methods.

For example, RNA secondary structure can be translated into a symbol string (*158*). In particular, the symbol of the corresponding local secondary structure can be assigned to each nucleotide. The resulting symbol strings can be aligned by one of the procedures of pairwise or multiple alignment. One of the advantages of this approach is that it is possible to use the routine of alignment. A significant drawback is that this approach provides a rough description of the compared secondary structures neglecting their individual features. This is not a serious problem when analyzing highly homologous sequences, but it surely is when the sequences are evolutionarily remote.

Another group of methods for RNA secondary-structure coding and comparison is based on tree-editing algorithms (*158*, *159*, *161*). These methods allow a detailed consideration of the RNA secondary structure. In particular, the method based on the construction of the so-called tree of cycles for RNA secondary structure provides its detailed description in linear or nonlinear code (*162*). With this method, types and the mutual disposition of loops, helices, and even the pairs of complementary nucleotides in helices can be described. To save computation time, at the next stage a condensed tree is constructed. In this tree a set of helices split by internal or bulge loops is regarded as a nonbranching helix, and a separate vertex is assigned to such a helix. With this approach, the reconstruction of an invariant secondary structure was performed for a highly variable domain D3 of the rRNA of large ribosomal subunit (*162*). The secondary structure of this domain has been reconstructed for each of the three major kingdoms: prokaryotes, eukaryotes, and archeaebacteria. What is essential is that, apart from having the common features (helices A, B, C, and D), these domains contain specific elements of secondary structure that are invariant only within a kingdom (Fig. 22).

The last result clearly demonstrates how the crucial limitation of any

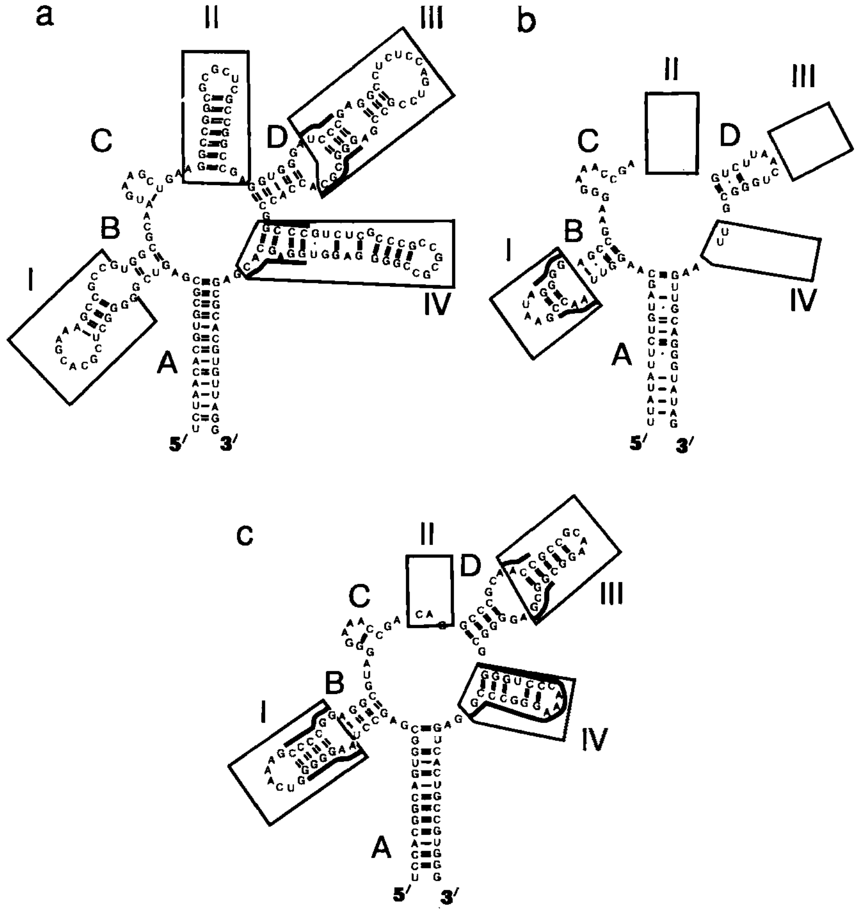

FIG. 22. Reconstruction of the invariant secondary structure of the D3 domain of large ribosomal RNA for three major kingdoms (*162*). (a) eukaryotes (*H. sapiens*); (b) eubacteria (*E. coli*); (c) archeaebacteria (*Desulfurococcus mobilis*). The helices invariant within all the three kingdoms are denoted by capital letters (A–D). Regions presenting structural variation among the three kingdoms are boxed and indicated by roman numbers. Secondary structure features conserved in each kingdom are shown by solid lines. Reproduced from Ref. *162*, C. Chevalet and B. Michot, *CABIOS* 8, 215 (1992), by permission of Oxford University Press.

comparative approach works. This is the assumption that RNA secondary structure is invariant within any given group of sequences. Whether secondary structure is invariant should be checked in each individual case, because while the primary structure of RNA was evolving, so was the secondary structure. Nevertheless, this approach has a sound motivation, a pronounced hierarchy of the variability levels corresponding to the levels of RNA organization: primary structure is the most variable, secondary structure is less variable, and tertiary structure is highly conserved. This is well exemplified by the fragments of genomic RNA from some plant viruses considered in Section II,E. These fragments have a tRNA-like tertiary structure that displays a high structural similarity with the L-shaped tRNA (*163*). At the level of secondary structure, similarity is slight, and none of it can be detected at the level of primary structure.

Each of the above approaches relies on definite features of structure, function, and evolution of RNA secondary structure and, therefore, has certain advantages. Nonetheless, the most reliable results appear to be expected of a combination of experimental and computer methods. Such a combined approach was applied (*164*) to identify the packaging signal of HIV-1. Loops were localized using diethylpyrocarbonate (DEP) and RNase S1. The former is specific to the A sites of the single-stranded RNA, the latter to the A and U sites. Low-energy secondary structures were determined by the program FOLD of the software package GCG. The homologous sequences of HIV-1 strains were compared for the presence of evolutionarily invariant secondary structure. Indeed, the lowest energy secondary structure, namely, a stem of three helices separated by loops (Fig. 1), was revealed by DEP and S1-RNase cleavage of RNA.

The algorithms for prediction of RNA secondary structure by comparative analysis of sequences are additionally described in Section II,E, where pseudoknots are discussed.

E. Prediction of Pseudoknots

There is still a quite vague understanding of the factors accounting for tertiary structure formation. The conformations of single-stranded regions are one of them. The structure of these regions, which may be stabilized by stacking interactions, has been studied intensively since the 1960s (*87*), with synthetic oligonucleotides. The thermodynamics of such other features of tertiary structure as nucleoside triplets (*165*) and helix–helix interactions (*97*) have been less studied. Below we consider pseudoknots, one of the most important features of tertiary structure (*166*, *167*).

A pseudoknot forms when the single-stranded region of a loop forms Watson–Crick pairs with a complementary fragment not within the loop (*166*, *167*). Pseudoknots were discovered in studies of turnip yellow mosaic virus (TYMV) RNA (see *166*). As it turned out, this RNA resembles tRNA in several respects (Fig. 23). In particular, it is able to accept Val-tRNA at its 3′ end. However, further investigations (*168*) showed that the secondary structure of the 3′ end of this RNA, constituted by four helices (I–IV), is quite unlike a "cloverleaf." Meanwhile, it was possible to regard the 3′ end structure as a three-dimensional L-shaped tRNA by assuming that the pseudoknot forms owing to complementary interaction between the CCC fragment near helix II and the GGG fragment in the hairpin loop of helix I. Computer simulations showed that the tertiary structure of this fragment of RNA with a pseudoknot is similar to the L-shaped tertiary structure of tRNA (*163*). With the use of pseudoknot structure, spatial models were also suggested for the His- and Tyr-accepting domains at the 3′ ends of tobacco mosaic virus (TMV) RNA and brome mosaic virus (BMV) RNA (*170–172*).

There are four types of loops that may contain a single-stranded region involved in the formation of a pseudoknot, namely, hairpin, bulge, internal, and multiple. Note that the complementary region to be paired with may either be related to one of the loop types mentioned or be in the single-stranded region of RNA. Thus 14 different types of pseudoknots may exist, depending on the location of the complementary regions by which they are formed (*166*).

An H-type (hairpin) pseudoknot (Fig. 23A) that involves a hairpin loop is perhaps the best studied. It is formed by two helices, S1 and S2, and two

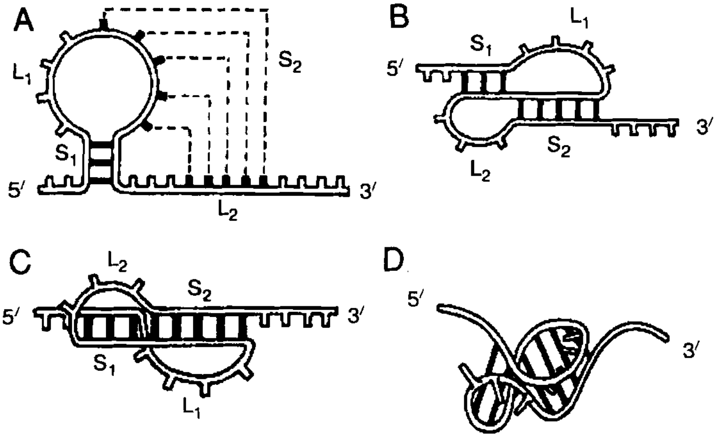

FIG. 23. Schematic representation of an H-type pseudoknot (*166*). (A) Basic elements of a pseudoknot: S1 and S2 are helices; L1 and L2 are connecting loops; (B) planar representation of a pseudoknot; (C) coaxial stacking of helices S1 and S2; (D) pseudoknot: three-dimensional schematic representation.

connecting loops, L1 and L2 (*166*). The helices are stacked coaxially in this structure to form a longer quasicontinuous helix S1 + S2 (*166*). While loop L1, which is the first from the 5′ end, crosses the major groove of the quasicontinuous helix, loop L2 crosses the minor. It is supposed that stacking interactions between the termini of the two helices provide an additional gain in energy on the formation of pseudoknots.

As noted above, the existence of pseudoknots was initially demonstrated in plant viruses (*166*). Today there is experimental evidence for the existence of these structures in the RNA genomes of plant and some other viruses (see *166*, *167*).

Pseudoknots have been revealed in small subunit rRNAs (*169*, *173*, *174*). One of theme is H-type; another is formed by complementary pairing of the fragments of a hairpin loop and a multiple loop, and the last one by complementary pairing between a hairpin loop and a bulge loop of 16-S RNA.

Evidence has been gathered for an important role of pseudoknots in the autoregulation of prokaryotic mRNA expression (*166*). A pseudoknot has been found in the 5′ noncoding leader region of the bacteriophage T4 gene-32 mRNA (*175*). Comparison of the operator sequences of phages T2, T4, and T6 containing this gene provides phylogenetic evidence for the conservation of this pseudoknot. The binding of the proteins encoded by the corresponding mRNA to the pseudoknots in the 5′ leader nontranslated regions of phage mRNA can provide the autoregulation of translation of the corresponding mRNA (*166*).

Site-directed mutagenesis of *E. coli* operon mRNA (*176*) provides evidence for the existence of the pseudoknot, which is of importance for the binding of ribosomal protein S4, one of four protein products encoded by this operon.

Finally, there is evidence that pseudoknots play an important role in providing a translation frameshift for a range of eukaryotic mRNAs, in particular Rous sarcoma virus RNA (*177*) and avian coronavirus RNA (*178*). In the latter, the H-type pseudoknot is located six nucleotides away from a translation frameshift site.

It is also suggested that pseudoknots should be found in some ribozymes. In particular, type-I self-splicing introns carry the pseudoknots that provide the spatial proximity of the RNA regions involved in formation of the active center (*166*, *179*).

In general, the available data provide evidence for the existence of pseudoknots in various RNAs and for their functional significance.

The discovery of pseudoknots as new elements of RNA structure brought about a determination of their thermodynamic parameters. These parameters should be determined as dependent on the type of the pseudoknot, on the length of its helices and loops, and on their nucleotide sequences. Today, it is only being considered how to deal with the problem.

Pseudoknot formation was shown experimentally to depend on a number of factors, such as the content of the solution, the lengths of the involved loops and helices, the energies of the helices, the coaxial stacking energy of the helices, and the mutual arrangement of pseudoknot elements (*180–183*). The stability of such structures essentially depends on the conformation of the bases forming complementary pairs and on the steric constraints on the loops. It has been shown (*180, 181*) that the gain in enthalpy observed on pseudoknot formation is less than that calculated applying the nearest-neighbor model (*88*) to the helices involved. This effect can be due to unfavorable interactions between helices and loops of the pseudoknot or violation of conformation of complementary pairs from that typical for an A helix.

Studies of the pseudoknots formed by synthesized oligonucleotides demonstrated (*180, 181*) that some of them melt in a multistep way, whereas others can be described by the all-or-none model (cooperative melting). Such pseudoknots are just a little more stable (by 1.5–2 kcal/mol) than the most stable of the involved hairpins. Because there is no complete energy table for pseudoknots, simplified approximations are used in computer calculations. On the basis of experimentally confirmed pseudoknots, parameters were introduced (*137*) for the destabilizing energy of loops less than 15 nt in length for H-type pseudoknots. Constraints regarding the sizes of loops and helices and their mutual arrangement were also imposed on the pseudoknots. If the pseudoknot was allowed, the destabilizing energy of the loop was set at 4.2 kcal/mol. Another way to describe the energy of pseudoknots has also been suggested (*138*). The destabilizing energy of the resulting loops was set equal to the mean value of the energy of the loops involved in the pseudoknot (two loops if the helices were stacked coaxially, and three loops if the helices were slightly remote from each other). The energy of the helices within a pseudoknot was determined here by using standard parameters (*88*).

Allowance was made for the entropic interaction between the loops (*141, 184*). In this case closure of a loop facilitates formation of the next loop due to the cooperative effect. Cooperativity improved the simulation of kinetics of secondary structure formation for the self-splicing group-1 intron RNA (*141*).

Today there are suggested several approaches to pseudoknot detection in RNA secondary structure. One is based on comparison of homologs (*185*) and it was applied to pseudoknot detection in the V4 variable region of RNA of the small ribosomal subunit of eukaryotes. As in any comparative method the functionally significant pseudoknots are assumed to be evolutionarily conserved.

Analysis of 13 aligned sequences allowed the invariant helices to be listed. Then the compensatory substitutions were counted for each of them. Helices with the maximum number of compensatory substitutions were then selected into a subset. As a result, there was obtained a set of helices of maximal functional significance. These helices were used in developing a model of RNA secondary structure. Figure 24 exemplifies invariant pseudoknots revealed in the V4 region. That these pseudoknots are functionally significant is supported by the fact that the sequences belong to phylogenetically remote species (e.g., *Homo sapiens, Drosophila melanogaster,* and *Saccharomyces cerevisiae*).

Another empirical approach for predicting the location of H-type pseudoknots in RNA is also of interest (*186*). This one is based on assessment of the indices of statistical (St) and thermodynamic (Th) significance for a local region of RNA containing a potential pseudoknot. The index St shows to what extent the energy of a given short fragment containing a potential pseudoknot differs from the energy of random sequences of the same length. The index Th shows to what extent the energy of a given short fragment containing a potential pseudoknot differs from the energy of the other real fragments of the same RNA sequence. By surveying the whole sequence through a sliding window, one selects RNA fragments with the maximum value of the indices. Then all the potential pseudoknots for the selected regions are constructed. Further selection among them is performed by a range of empirical criteria. Let us consider a fixed pair of helix (S1) and loop

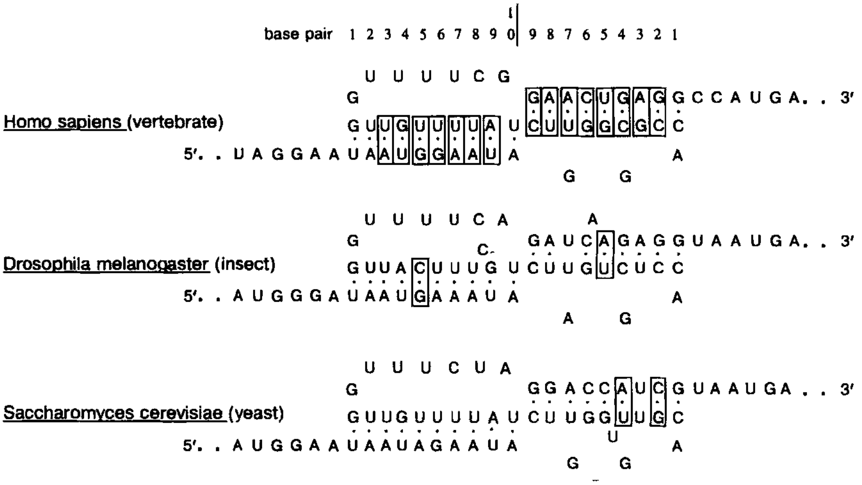
base pair 1 2 3 4 5 6 7 8 9 1 0 | 9 8 7 6 5 4 3 2 1
Homo sapiens (vertebrate)
U U U U C G
G GAACUGAG G C C A U G A . . 3′
G U U G U U U U A U C U U G G C G C C
5′. . U A G G A A U A A U G G A A U A A
G G
Drosophila melanogaster (insect)
U U U U C A A
G C GAUCAGAGGUAAUGA. . 3′
GUUACUUUG UCUUGUCUCC
5′. . A U G G G A U A A U G A A A U A A
A G
Saccharomyces cerevisiae (yeast)
U U U C U A
G GGACCAUCGUAAUGA . . 3′
GUUGUUUUAU CU UGGUUGC
5′. . A U G G A A U A A U A G A A U A U A
G G

FIG. 24. Evolutionarily invariant pseudoknots in the variable region of eukaryotic small ribosomal subunit RNA. Reproduced from Ref. *185*, J.-M. Neefs and R. De Wachter, *NARes* **18**, 5695 (1990), by permission of Oxford University Press.

(L1). To simulate pseudoknot formation, it is necessary to indicate another helix (S2) and another loop (L2) within the preselected fragment (Fig. 23). Chosen out of all those available is the longest helix, S2, combined with loop L2 such that the minimum loss in free energy is provided. This method (*186*) was tested on RNA with pseudoknots in known locations. Its identification capacity proved to be quite high (*186*).

Most of the methods for predicting the RNA secondary structure described in Sections II,B–II,D forbid pseudoknot formation. This limitation was introduced in the first algorithms for prediction of RNA secondary structure. At that time there was no evidence of pseudoknots in RNA. It should be emphasized that this limitation is crucial for the most popular algorithms calculating low-energy secondary structures (*104*). Accordingly, it is not easy to consider pseudoknots in the framework of these algorithms. In particular, allowing pseudoknots makes combinatorial algorithms enormously complicated, and recursive algorithms impossible to be implemented.

This stimulated the development of algorithms that allow pseudoknots (*136–138*). They are based on a simulation of the step-wise folding of RNA. For example, RNA folding was simulated by step-wise selection of the helices compatible with those having been formed at the previous stages (*138*). Detected at each step were all the helices compatible with those previously formed and providing the energy gain higher than a threshold. Then a set of 100 random secondary structures was generated with these helices by the Monte Carlo method. The frequency of each helix involved in these structures was estimated. The product of the frequency and the free-energy gain at the helix formation was calculated. The helix with greatest product was included in the current secondary structure at the given simulation step. Energy parameters were taken from Freier *et al.* (*88*). The destabilizing energy of the pseudoknot was set equal to the average of the energies of the loops forming the pseudoknot.

This algorithm provided a correct prediction of 66% of phylogenetically conservative helices in the secondary structure of the small subunit rRNA from *E. coli* (*138*). Interestingly, allowing pseudoknots does not bring about any great number of them in the predicted secondary structures. As a rule each secondary structure involves three to five pseudoknots.

There is another similar approach to predicting RNA secondary structure with pseudoknots (*137*). Here, again, RNA folding is simulated via step-wise addition of a helix to the structure that has already been formed. The empirical values of the thermodynamic parameters of the pseudoknot are as described above (*137*). The predicting ability of this method was checked on a range of objects. With the LSU intron, 95 of 127 complementary pairs (75%) were identified correctly. As for the small subunit rRNA of *E. coli*, 26 of 65 (40%) phylogenetically conservatized helices were predicted correctly.

Using this method, the pseudoknot at the 3′ end of TYMV RNA was predicted and good predictions were obtained for the TMV RNA 3′ end secondary structure with five pseudoknots (*137*). It is noteworthy that this program is faster than those based on the recursive algorithms. Time consumption is about N^2 (where N is the sequence length), whereas the recursive algorithms (*104*) usually take about N^3.

One more program (*136*) for predicting the RNA secondary structure with pseudoknots was developed as a modification of the program RNAFOLD (*135*), initially intended for simulating step-wise RNA folding. Interestingly, although the algorithms for RNA folding simulations are not aimed at searching for the lowest energy structures, the resulting structures are often just like those corresponding to the global minimum of energy.

F. Statistical Analysis of RNA Secondary Structure

To understand the basic features of RNA secondary structure, it is of interest to analyze the dependence of fundamental characteristics (total energy, number of complementary base pairs, number of loops and helices, and some others) on the length and nucleotide content of the RNA molecule. The features of evolution of RNA secondary structure are of interest, too. This is of special importance because the comparative methods are highly effective means of RNA secondary structure prediction. An understanding of the matter may be achieved through analysis of secondary structure predicted for a variety of random RNA sequences (*108, 187–190*). By applying methods based on free energy minimization, it is possible to analyze the dependence of characteristics of secondary structure on RNA sequence length or on the mode of evolution. Such analyses usually cover sequences no longer than 150 nt (*108, 187–190*). With these lengths a correct prediction of low-energy RNA secondary structure is possible.

The totality of potential RNA sequences of length N forms an N-dimensional hypercube. The total number of M of RNA sequence variants depends on length L exponentially: $M = 4^N$, whereas the estimated number of different secondary structures of RNA is $S = N^{-3/2}(1.85)^N$ (*108*). This formula is applied to the planar secondary structures with hairpin loops more than 2 nt in length and helices not shorter than 2 bp. As length N approaches infinity, the M/S ratio also approaches infinity. This implies that at large N there is, on the average, a great many RNA sequences corresponding to any secondary structure. In other words, the RNA folding code is strongly degenerate (*108, 189–191*).

Prediction of low-energy secondary structure for random RNA sequences with homogeneous nucleotide content allowed the dependencies of the mean values of secondary structure characteristics on sequence length to be determined (*187, 188*). As it turned out, characteristics such as the total energy of secondary structure, the number of base pairs in helices, and the number of helices and loops depend linearly on sequence length (*187*). The dependence becomes linear for sequences with lengths exceeding 20 to 30 nt. On the average, a fragment of this length contains 1 to 3 helices and loops. At the same time, the mean length of loops and helix size tend to constant values as sequence length grows. The same is typical of the branching degree (the number of helices closed by a loop). The average helix length is 3 to 7 bp. The average loop length is not less than 3 to 5 nt, and the average branching index is 1.5 (*187*). What can be inferred from these results is that there must be some basic "module" of RNA secondary structure folding characterized by the parameters presented above.

The study of low-energy secondary structure formed by random RNA sequences characterized by different alphabets (*187, 188*) is also of concern. Considered were both the real alphabet {A, U, G, C} and various model alphabets: {A, U}; {G, C}; {G, C, X, K}; {A, B, C, D, E, F}. In the last two, complementarity was set up as X·K, A·B, C·D, E·F (in any, the energy was as in G·C pairs). Model alphabets help discriminate between effects of different nature and aid understanding the interference of the respective contributions to the natural RNA folding (alphabet volume, pairing rules, and base-pairing strength).

Any alphabet is characterized by two essential parameters: strength and "stickiness." Strength is defined by the energy of complementary base pairing. Thus {G, C} is a stronger alphabet than {A, U}. Calculations show that a stronger alphabet favors the formation of more low-energy secondary structure. Besides, a stronger alphabet provides higher compactness of the resulting secondary structure (the less the number of loops and the higher the number of paired bases, the higher compactness). The reason here is that, *ceteris paribus*, the helices based on a strong alphabet are more stable than those on a weaker. High stability of helices provides possibilities of short-loop closure. As was indicated, loop formation entails energy loss (Table II). That is why loop closure is impossible unless there are highly stable helices involved. Besides, the stronger the alphabet, the higher the number of helices in RNA secondary structure, because even short helices become stable enough to compensate for the energy loss owing to the loop closure.

"Stickiness" characterizes the probability of two randomly selected bases forming a complementary pair in a homogeneous sequence. Thus the stickiness of the real alphabet {A, U, G, C} with two complementary pairs (A·U and G·C) is $1/4$, the same for the model alphabet {G, C, X, K}. Including G·U complementary pairs in the real alphabet gives stickiness as 0.375. Stickiness of the model alphabets {A, U} and {G, C} is 0.5.

At a fixed alphabet strength, the higher the stickiness, the higher the probability of complementary pairing for any nucleotide within the se-

TABLE II

FREE ENERGY INCREMENTS FOR LOOPS[a]

Loop size	Internal loop	Bulge loop	Hairpin loop
1	—	+3.3	—
2	+0.8	+5.2	—
3	+1.3	+6.0	+7.4
4	+1.7	+6.7	+5.9
5	+2.1	+7.4	+4.4
6	+2.5	+8.2	+4.3
7	+2.6	+9.1	+4.1
8	+2.8	+10.0	+4.1
9	+3.1	+10.5	+4.2
10	+3.6	+11.0	+4.3

[a] In units of kcal mol^{-1}, in 1 *M* NaCl, at 37°C (*82*).

Reproduced, with permission, from the *Annual Review of Biophysics and Biophysical Chemistry*, Volume 17, © 1988, by Annual Reviews Inc.

quence. Thus the mean helix length grows, the probability of new helices formation rises, and the loops shorten. That is, on average, as compactness of secondary structure increases, its energy decreases. The above estimates of the characteristics of the RNA secondary structure were obtained for random sequences, and it was interesting to compare them with the estimates for real DNA sequences.

The low-energy secondary structures of real RNA calculated by recursive algorithm (*104*) were compared (*187*) with the secondary structure of random sequences that had equal nucleotide frequencies. Mitochondrial, eubacterial 16-S rRNA and β-globin mRNA were analyzed. All the real sequences considered had approximately the same frequencies of every nucleotide. Five secondary structures with the lowest energies were considered for each of the RNA sequences.

With these structures the mean values of the following parameters were estimated: size of a loop, length of a helix, branching degree and helicity, i.e., the ratio of the number of complementary pairs, and sequence length (Table III). As is seen, the real and random RNA sequences have a similar mean branching degree, helicity, and helix size. The mean loop lengths of the real and random sequences differ by about 20%, whereas the respective mean values of the remaining parameters do not differ by more than 14%.

The results obtained suggest that the concept of "random edited biopolymer" proposed for natural aminoacid sequences (*192*) could be applied to RNA, too. According to the concept, globular proteins arose in the course of evolution from random aminoacid sequences due to the fixation of a limited number of aminoacid substitutions ("evolutionary editing"). This concept

TABLE III
MEAN VALUES FOR RANDOM AND NATURAL SEQUENCES[a]

Source	n_{BP}/N	n_{st}	n_{lp}	n_{BD}
Random sequences	0.29	4.57	5.42	1.82
β-Globin mRNAs	0.31	4.49	4.42	1.89
Mitochondrial rRNAs	0.26	4.44	6.53	1.74
Eubacterial rRNAs	0.33	4.59	4.62	1.92
Mitochondrial rRNAs	0.24	3.76	6.00	1.90
Eubacterial rRNAs	0.28	4.35	5.81	1.93

[a] n_{BP}, Mean number of base pairs; n_{st}, mean helix size; n_{lp}, mean loop size; n_{BD}, mean branching degree. Reproduced from Ref. *187*. Copyright © 1993. Reprinted by permission of John Wiley & Sons, Inc.

is supported by the similarity between mean values of different structural parameters of real globular proteins and the proteins formed from random sequences (*192*).

Again, the similarity of secondary structure parameters of random sequences and real RNA suggests that they can also be considered as random edited polymers. Computer experiments (*108*, *189*, *190*) provide a sound argument favoring evolutionary editing as a mechanism of incidence of the current RNA secondary structure. It was concluded that within a small neighborhood of any random point in the multidimensional space of RNA sequences there exists such a set of RNAs that can form all the possible low-energy secondary structures. In fact, it means that any random sequences can be transformed into a sequence with definite low-energy secondary structure via fixation of a limited number of certain mutations. The number of mutations to provide "editing" is 15–20 in a length of 100 nt (*108*, *189*, *190*). Thus, mutational editing was proved effective to obtain RNA sequences with given secondary structure from random sequences.

That secondary structures are vulnerable to mutations (*108*, *189*, *190*) is an important factor facilitating such editing. Experimental study of secondary structure of threonyl-tRNA synthetase mRNA from *E. coli* shows that this secondary structure can be drastically affected by a single nucleotide substitution (*193*). This is in a good agreement with the results of a computer simulation of mutation effects on secondary structure of random RNA sequences of 100 nt. It was shown (*108*) that even a few mutations (one to three) can result in significant alteration of the most low-energy secondary structure.

It is noteworthy that, if there were more than three random substitutions in the sequence, the probability of folding into the initial secondary structure was low. At 15–20 random mutations, the probability of sequence folding into the initial secondary structure was the same as the probability of two random sequences folding identically. It means that at this number of mutations, the sequence memory about the initial secondary structure is totally erased.

Meanwhile, analyses of isofunctional RNA molecules of different taxa show that a functionally significant secondary structure is, as a rule, highly conservative, whereas primary structures are quite variable. In some cases, there is the lowest similarity between primary structures (30–40%), yet the secondary structures of RNA have a similar pattern. As was pointed out in Section II,D, this must be due to a particular mode of RNA evolution, by which the helical regions of secondary structure were shown to fix the compensatory substitutions—that is, substitutions retaining complementarity. This feature of RNA evolution is taken into account for prediction of secondary structure using the methods described above.

Secondary structure conservation in the course of the evolution of an RNA sequence has been studied by many authors (*108*, *190*), and was considered for various types of mutations (point substitutions, recombinations, deletions, and insertions) (*190*). As it turned out, deletions and insertions are the most effective in changing secondary structure; they thus provide the fastest possible evolutionary optimization of secondary structure. On the other hand, recombinations between strongly homologous RNAs causes only local changes in secondary structure. Synonymous substitutions cause fewer alterations in secondary structure than random point mutations.

It has been demonstrated (*108*, *189*) that there are routes in the multidimensional space of sequences in which the initial low-energy secondary structure remains unaltered, no matter what the primary structure. Thus, in 22% of the cases, the route reaches an end point where the distance between the current and initial sequence is as great as possible, equal to the length of the sequence (zero homology). Such routes are implemented by the fixing of single mutations in nonhelical regions and compensatory mutations in helices. It implies that RNAs with similar secondary structures can be slightly homologous (*108*, *189*). Approaches to revealing the invariant secondary structures for slightly homologous RNAs were discussed in Section II,D.

III. Concluding Remarks

Investigations of RNA structures with different enzymatic and chemical probes can provide detailed data allowing identification of double-stranded regions of the molecules and nucleotides involved in tertiary interactions. Combining results of probing experiments, and taking into account thermodynamic data, the data of phylogenetic studies, and the geometry of RNA units, it is possible to build models of RNA structures at the nucleotide level of resolution. Cross-linking approaches together with RNA shape-sensitive chemical probes and cleaving reagents that recognize specific structural features of RNA molecules provide easy means for monitoring conformational changes in RNA under different conditions or on the binding of various factors. Intrinsically, chemical probing is a high-resolution method because it allows investigation of the reactivities of individual functional groups of the RNA. It does not allow discovery of novel elements in an RNA. This can be done only by physical methods (X-ray analysis and NMR). However, when a structural element is discovered and characterized, it can be detected by probing techniques in novel RNA species and thus taken into account when building high-resolution models. Enzymatic and chemical probes are becoming increasingly more important for detecting structural variations, for monitoring conformational changes of RNA, for investigating effects of mutations on the RNA structure, and for investigating RNA–protein complexes. A great advantage of probing techniques is the possibility of investigating RNA structure in complex systems regardless of the presence of other biopolymers, when other methods are not applicable. Further development of the chemistry of probes and progress in computer modeling will provide researchers with new, simple, and reliable methods for investigation of RNAs at any level of complexity and for investigations of the dynamics of RNA–protein complexes in the cell.

Acknowledgments

The work of N.A.K. and I.I.T. was supported by the Russian Human Genome Project, the Russian Ministry of Science and Technique Politics, and by Russian Fundamental Research Foundation Grant No. 12757. I.E.V. and V.V.V. acknowledge the support of the Centre National de la Recherche Scientifique (CNRS), the Ministère de l'Education Nationale et de la Culture and Université Louis Pasteur (Strasbourg, France). We are grateful to V. Filonenko for translating Section II of this essay from Russian to English.

References

1. R. F. Gesteland and J. F. Atkins, eds., "The RNA world." Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1993.
2. G. J. Quigley, A. H. J. Wang, N. C. Seeman, F. L. Suddath, A. Rich, J. L. Sussman and S. H. Kim, *PNAS* **72,** 4866 (1975).
3. R. Giege, D. Moras and J. L. Thierry, *JMB* **115,** 91 (1977).
4. A. Jack, J. E. Ladner and A. Klug, *JMB* **108,** 619 (1976).
5. J. L. Sussman, S. R. Holbrook, W. R. Wade, G. M. Church and S. H. Kim, *JMB* **123,** 607 (1978).
6. E. Westhof, P. Dumas and D. Moras, *JMB* **184,** 119 (1985).

7. P. B. Moore, *Curr. Opin. Struct. Biol.* **3,** 340 (1993).
8a. J. K. James and I. J. Tinoco, *NARes* **21,** 3287 (1993).
8b. H. W. Pley, K. M. Flaherty and D. B. Mckay, *Nature* **372,** 68 (1994).
9. T. Hayashi, Y. Ueno and T. Okamoto, *FEBS Lett.* **327,** 213 (1993).
10. V. V. Vlassov, R. Giege and J. P. Ebel, *FEBS Lett.* **120,** 12 (1980).
11. A. Jack, J. E. Ladner, D. Rhodes, R. S. Brown and A. Klug, *JMB* **111,** 315 (1977).
12. R. Lavery and A. Pullman, *Biophys. Chem.* **19,** 171 (1984).
13. P. Romby, D. Moras, H. Bergdoll, P. Dumas, V. V. Vlassov, E. Westhof, J. P. Ebel and R. Giege, *JMB* **184,** 455 (1985).
14. M. Silberklang, A. M. Gillam and U. L. RajBhandary, *NARes* **4,** 4091 (1977).
15. G. Chaconas, J. H. Van de Sande and R. B. Church, *BBRC* **66,** 962 (1975).
16. A. G. Bruce and O. C. Uhlenbeck, *NARes* **4,** 2427 (1978).
17. B. Rether, J. Bonnet and J. P. Ebel, *EJB* **50,** 281 (1974).
18. C. Ehresmann, F. Baudin, M. Mougel, P. Romby, J. P. Ebel and B. Ehresmann, *NARes* **15,** 9109 (1987).
19. A. Krol and P. Carbon, *in* "Methods in Enzymology," Vol. 180, p. 212. Academic Press, San Diego, 1989.
20. T. Uchida, T. Arima and F. Egami, *J. Biochem.* **67,** 91 (1970).
21. T. Uchida and F. Egami, *in* "Methods in Enzymology," Vol. XII, p. 228. Academic Press, New York, 1967.
22. M. S. Boguski, P. Hieter and C. C. Levy, *JBC* **255,** 2160 (1980).
23. C. Florentz, J. P. Briand, P. Romby, L. Hirth, J. P. Ebel and R. Giege, *EMBO J.* **1,** 269 (1982).
24. T. Ando, *BBA* **114,** 158 (1966).
25. S. Linn and T. R. Lehman, *JBC* **240,** 1287 (1965).
26. S. K. Vassilenko and V. C. Ryte, *Biokhimiya* **40,** 578 (1975).
27. P. D. Lawley and P. Brookes, *BJ* **89,** 117 (1963).
28. N. Wintermeyer and H. G. Zachau, *FEBS Lett.* **58,** 306 (1975).
29. V. S. Zueva, A. S. Mankin, A. A. Bogdanov and L. A. Baratova, *EJB* **146,** 679 (1985).
30. B. Singer, *Nature* **264,** 333 (1976).
31. B. Singer and H. Fraenkel-Conrat, *Bchem* **14,** 772 (1976).
32. V. V. Vlassov, R. Giege and J. P. Ebel, *EJB* **119,** 51 (1981).
33. B. J. Van Stolk and H. F. Noller, *JMB* **180,** 151 (1984).
34. R. Naylor, N. W. Y. Ho and P. T. Gihlam, *JACS* **87,** 4209 (1966).
35. R. Shapiro, B. I. Cohen, S. J. Shiuey and H. Maurer, *Bchem* **8,** 238 (1969).
36. D. A. Peatty, *PNAS* **76,** 1760 (1979).
37. D. A. Peatty and W. Gilbert, *PNAS* **77,** 4679 (1980).
38. K. M. Weeks and D. M. Crothers, *Science* **261,** 1574 (1993).
39. T. D. Tullis and B. A. Dombroski, *Science* **230,** 679 (1985).
40. T. D. Tullis and B. A. Dombroski, *PNAS* **83,** 5465 (1986).
41. J. A. Latham and T. R. Cech, *Science* **245,** 276 (1989).
42. D. W. Celander and T. R. Cech, *Science* **251,** 401 (1991).
43. B. Laggerbauer, F. L. Murphy and T. R. Cech, *EMBO J.* **13,** 2669 (1994).
44. R. P. Herztberg and P. B. Dervan, *Bchem* **23,** 3934 (1984).
45. R. Breslow, *Accts. Chem. Res.* **24,** 317 (1991).
46. V. V. Vlassov, G. Zuber, B. Felden, J.-P. Behr and R. Giege, *NARes* (1995). In press.
47. M. A. Podyminogin, V. V. Vlassov and R. Giege, *NARes* **21,** 5950 (1993).
48. B. Felden, C. Florentz, R. Giege and E. Westhof, *JMB* **235,** 508 (1994).
49. E. Westhof, P. Romby, C. Ehresmann and B. Ehresmann, *in* "Theoretical Biochemistry

and Molecular Biophysics" (D. Beveridge and R. Lavery, eds.), p. 399. Adenine Press, Guilderland, New York, 1990.

50. T. J. Richmond, *JMB* **173,** 63 (1984).
51. P. Thiyagarajan and P. K. Ponnuswamy, *Biopolymers* **18,** 2233 (1979).
52. D. E. Bergstrom and N. J. Leonard, *Bchem* **11,** 1 (1972).
53. A. Favre, R. Buckingham and G. Thomas, *NARes* **2,** 1421 (1975).
54. L. S. Behlen, J. R. Sampson and O. C. Uhlenbeck, *NARes* **20,** 4055 (1992).
55. P. L. Wollenzien, P. Goswami, J. Teare, J. Szeberenyi and C. J. Goldenberg, *NARes* **15,** 9279 (1987).
56. P. L. Wollenzien, R. F. Murphy and C. R. Cantor, *JMB* **184,** 67 (1985).
57. B. Datta and A. M. Weiner, *JBC* **267,** 4497 (1992).
58. B. Datta and A. M. Weiner, *JBC* **267,** 4503 (1992).
59. G. D. Cimino, H. B. Gamper, S. T. Isaaks and J. E. Hearst, *ARB* **54,** 1151 (1985).
60. J. Christiansen, *NARes* **16,** 7457 (1988).
61. R. Brimacombe, J. Atmadja, W. Stiege and D. Schuler, *JMB* **199,** 115 (1988).
62. A. Expert-Bezancon and D. Hayer, *EJB* **103,** 365 (1980).
63. M. A. Grachev and M. I. Rivkin, *NARes* **2,** 1237 (1975).
64. E. Wickstrom, L. S. Behlen, M. A. Renben and P. R. Ainpour, *PNAS* **78,** 2082 (1981).
65. H. Han and P. B. Dervan, *PNAS* **91,** 4955 (1994).
66. A. B. Burgin and N. R. Pace, *EMBO J.* **9,** 4111 (1990).
67. J. M. Nolan, D. H. Burke and N. R. Pace, *Science* **261,** 762 (1993).
68. M. E. Harris, J. M. Nolan, A. Malhotra, J. W. Brown, S. C. Harvey and N. R. Pace, *EMBO J.* **13,** 3953 (1994).
69. M. Zenkova, C. Ehresmann, J. Caillet, M. Springer, G, Karpova, B. Ehresmann and P. Romby, *EJB* (1995). In press.
70. R. S. Brown, J. C. Dewan and A. Klug, *Bchem* **24,** 4785 (1985).
71.W. J. Krzyzosiak, T. Marciniec, M. Wiewiorowski, P. Romby, J. P. Ebel and R. Giege, *Bchem* **27,** 5771 (1988).
72. L. S. Behlen, J. R. Sampson, A. B. DiRenzo and O. C. Uhlenbeck, *Bchem* **29,** 2515 (1990).
73. J. R. Rubin and M. Sundaralingam, *J. Biomol. Struct. Dyn.* **1,** 639 (1983).
74. T. Pan and O. C. Uhlenbeck, *Bchem* **31,** 3887 (1992).
75. T. Pan, B. Dichtl and O. C. Unlenbeck, *Bchem* **33,** 9561 (1994).
76. B. Dichtl, T. Pan, A. B. DiRenzo and O. C. Uhlenbeck, *NARes* **21,** 351 (1993).
77. J. Ciesiolka, S. Lorenz and V. A. Erdmann, *EJB* **204,** 575 (1992).
78. S. Kazakov and S. Altman, *PNAS* **88,** 9193 (1991).
79. J. Ciesiolka, W.-D. Hardt, J. Schlegl, V. A. Erdmann and R. K. Hartmann, *EJB* **219,** 49 (1994).
80. C. S. Chow and J. K. Barton, *JACS* **112,** 2839 (1990).
81. J. R. Fresco, B. M. Alberts and P. Doty, *Nature* **188,** 98 (1960).
82. D. H. Turner, N. Sugimoto and S. M. Freier, *Annu. Rev. Biophys. Biophys. Chem.* **17,** 167 (1988).
83. M. Gouy, *in* "Nucleuc Acid and Protein Sequence Analysis: A Practical Approach" (M. J. Bishop and C. J. Rawlings, eds.), p. 259. IRL Press, Oxford, 1987.
84. A. Wada and A. Suyama, *Prog. Biophys. Mol. Biol.* **47,** 113 (1986).
85. M. Chamberlin, R. L. Baldwin and P. Berg, *JMB* **7,** 334 (1963).
86. H. De Voe and I. Tinoco, Jr., *JMB* **4,** 500 (1962).
87. C. R. Cantor and P. R. Schimmel, "Biophysical Chemistry." W. H. Freeman, San Francisco, California, 1980.

88. S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Nelson and D. H. Turner, *PNAS* **83,** 9373 (1986).

89. W. Salser, *CSHSQB* **62,** 985 (1977).

90.G. Steger, H. Hoffman, J. Fortsch, H. J. Gross, J. W. Randles and H. L. Saenger, *J. Biomol. Struct. Dyn.* **2,** 543 (1984).

91. J. Ninio, *Biochimie* **61,** 1133 (1979).

92. C. Papanicolaou, M. Gouy and J. Ninio, *NARes* **12,** 31 (1984).

93. J. A. Jaeger, D. H. Turner and M. Zuker, *PNAS* **86,** 7706 (1989).

94. V. V. Filimonov and P. L. Privalov, *JMB* **122,** 465 (1978).

95. D. W. Appleby and N. R. Kallenbach, *Biopolymers* **12,** 2093 (1973).

96. L. A. Marky, N. R. Kallenbach, K. A. McDonough, N. C. Seeman and K. J. Breslauer, *Biopolymers* **26,** 1621 (1987).

97. A. E. Walter and D. H. Turner, *Bchem* **33,** 12715 (1994).

98. H. Heus and A. Pardi, *Science* **253,** 191 (1991).

99. G. Varani, C. Cheong and I. Tinoco, Jr., *Bchem* **30,** 3280 (1991).

100. H. Jackobson and W. H. Stockmayer, *J. Chem. Phys.* **18,** 1600 (1950).

101. J. M. Paipas and J. A. McMagon, *PNAS* **72,** 2017 (1971).

102. G. M. Studnicka, J. M. Rahu, I. M. Cummings and W. A. Salser, *NARes* **5,** 3365 (1978).

103. R. Nussinov and A. B. Jackobson, *PNAS* **77,** 6309 (1980).

104. M. Zuker and P. Stiegler, *NARes* **9,** 133 (1981).

105. M. Zuker, *in* "Mathematical Methods for DNA Sequences" (M. S. Waterman, ed.). CRC Press, Boca Raton, Florida, 1987.

106. R. Nussinov, G. Pieczenik, J. R. Gribbs and D. J. Kleitman, *SIAM J. Appl. Math.* **35,** 68 (1978).

107. M. Zuker and D. Sankoff, *Bull. Math. Biol.* **46,** 591 (1984).

108. P. Schuster, W. Fontana, P. F. Stadler and I. L. Hofacker, *Proc. R. Soc. Lond. Ser. B* **255,** 279 (1994).

109. I. Tinoco, Jr., O. C. Uhlenbeck and M. D. Levine, *Nature* **230,** 362 (1971).

110. L. V. Omelyanchuk, Yu E. Bessonov and N. A. Kolchanov, *in* "Computer Systems" (N. G. Zagoruiko, ed.), p. 135. Institute of Mathematics, Novosibirsk, 88, 1981 (in Russian).

111. N. A. Kolchanov and L. V. Omelyanchuk, *Stud. Biophys.* **87,** 115 (1982).

112. J.-P. Dumas and V. P. Ninio, *NARes* **10,** 197 (1982).

113. V. G. Tumanyan, L. E. Sotnikova and A. E. Holopov, *Dokl. Akad. Sci. USSR* **166,** 1465 (1966) (in Russian).

114. A. L. Williams and I. Tinoco, Jr., *NARes* 14, 299 (1986).

115. E. Comay, R. Nussinov and O. Comay, *NARes* **12,** 53 (1984).

116. P. Hogeweg and B. Hesper, *NARes* **12,** 67 (1984).

117. A. B. Jackobson, L. Good, J. Simonetti and M. Zuker, *NARes* **12,** 45 (1984).

118. A. B. Jackobson and M. Zuker, *JMB* **233,** 261 (1994).

119. R. Kolter and C. Yanofsky, *ARGen* **16,** 113 (1982).

120. I. Tinoco, R. N. Borer, B. Dengler, M. D. Levine, O. C. Uhlenbeck, D. N. Grothers and J. Gralla, *Nature* **245,** 40 (1973).

121. M. Schmitz and G. Steger, *CABIOS* **8,** 389 (1992).

122. J. S. McCaskill, *Biopolymers* **29,** 1105 (1990).

123. S. V. Matveev, V. V. Filimonov and P. L. Privalov, *Mol. Biol.* **16,** 990 (1982).

124. A. A. Mironov and A. E. Kister, *Mol. Biol.* **19,** 1350 (1985) (in Russian).

125. A. A. Mironov, L. P. Dyakonova and A. E. Kister, *J. Biomol. Struct. Dyn.* **2,** 953 (1985).

126. A. A. Mironov and A. E. Kister, *J. Biomol. Struct. Dyn.* **4,** 1 (1986).

127. C. Levinthal, *J. Chim. Phys. (Paris)* **65,** 44 (1968).

128. O. B. Ptitsyn, *Usp. Sovr. Biol.* **69,** 26 (1970) (in Russian).

129. O. B. Ptitsyn, *Izv. Acad. Nauk SSSR* **5,** 57 (1973) (in Russian).

130. D. B. Wetlaufer, *PNAS* **70,** 697 (1973).

131. T. E. Kreighton, *Prog. Biophys. Mol. Biol.* **53,** 231 (1978).

132. L. M. Gierasch and J. King, eds., "Protein Folding, Description of the Second Half of the Genetic Code." Am. Assoc. Adv. Sci., Washington, D.C., 1989.

133. V. B. Bokhonov and N. A. Kolchanov, *in* "Mathematical Models of Molecular Genetic Systems Controls" (V. A. Ratner, ed.), p. 124. Institute of Cytology and Genetics, Novosibirsk, 1979.

134. B. R. Jordan, *J. Theor. Biol.* **34,** 363 (1972).

135. H. M. Martinez, *NARes* **12,** 323 (1984).

136. H. M. Martinez, *in* "Methods in Enzymology," Vol. 183, p. 306. Academic Press, San Diego, 1990.

137. J. P. Abrahams, M. van den Berg, E. van Batenburg and C. W. A. Pleij, *NARes* **18,** 3035 (1990).

138. A. P. Gultyaev, *NARes* **19,** 2489 (1991).

139. A. A. Mironov and V. F. Lebedev, *Biosystems* **30,** 49 (1993).

140. A. Fernandez, *EJB* **182,** 161 (1989).

141. A. Fernandez, *Chem. Phys. Lett.* **183,** 499 (1991).

142. A. Fernandez, *Phys. Rev. Lett.* **64,** 2328 (1990).

143. A. Fernandez, *Phys. Rev. A* **43,** 1138 (1991).

144. V. V. Anshelevich, A. V. Vologodskii, A. V. Lukashin and M. D. Frank-Kamenetskii, *Biopolymers* **23,** 39 (1984).

145. D. Porschke, *Biophys. Chem.* **2,** 97 (1974).

146. D. R. Mills, C. Dobkin and F. R. Kramer, *Cell* **15,** 541 (1978).

147. M. Levitt, *Nature* **224,** 759 (1969).

148. M. S. Waterman, *in* "Methods in Enzymology," Vol. 164, p. 765. Academic Press, San Diego, 1989.

149. H. F. Noller and C. R. Woese, *Science* **212,** 403 (1984).

150. H. F. Noller, *ARB* **53,** 119 (1984).

151. M. T. Dixon and B. M. Hills, *Mol. Biol. Evol.* **10,** 256 (1993).

152. W. C. Wheeler and R. L. Honeycutt, *Mol. Biol. Evol.* **5,** 90 (1988).

153. F. Michel and B. Dujon, *EMBO J.* **2,** 33 (1983).

154. W. C. Curtiss and J. N. Vournakes, *J. Mol. Evol.* **20,** 351 (1984).

155. K. Han and H.-J. Kim, *NARes* **21,** 1251 (1993).

156. D. Sankoff, *SIAM J. Appl. Math.* **45,** 810 (1985).

157. D. Sankoff, J. B. Kruskal, S. Mainville and R. J. Cedergen, *in* "Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison" (D. Sankoff and J. B. Kruskal, eds.), p. 93. Addison-Wesley, Reading, Massachusetts, 1983.

158. B. A. Shapiro, *CABIOS* **4,** 387 (1988).

159. H. Margalit, B. A. Shapiro, A. B. Oppenheim and J. V. Maizel, Jr., *NARes* **17,** 4829 (1989).

160. S. Y. Le, J. Owens, R. Nussinov, J. H. Chen, B. A. Shapiro and J. V. Maizel, *CABIOS* **5,** 205 (1989).

161. A. S. Noetzel and S. M. Selkow, *in* "Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison" (D. Sankoff and J. B. Kruskal, eds.), p. 237. Addison-Wesley, Reading, Massachusetts, 1983.

162. C. Chevalet and B. Michot, *CABIOS* **8,** 215 (1992).

163. P. Dumas, D. Moras, R. Giege, P. Verlaan, A. van Belkum and C. W. A. Pleij, *J. Biomol. Struct. Dyn.* **4,** 707 (1987).

164. T. Hayashi, Y. Ueno and T. Okamoto, *FEBS Lett.* **327,** 231 (1993).

165. M. Chastain and I. Tinoco, Jr., *Bchem* **32,** 14220 (1993).

166. C. W. A. Pleij, .*Trends Biochem. Sci.* **15,** 143 (1990).

167. R. M. W. Mans and C. W. A. Pleij, *in* "Nucleic Acids and Molecular Biology" (F. Eckstein and D. M. J. Lilley, eds.), p. 250. Springer-Verlag, Berlin and New York, 1993.

168. K. Rietveld, R. Van Poelgeest, C. W. A. Pleij, J. M. Van Boom and L. Bosh, *NARes* **10,** 1929 (1982).

169. C. W. A. Pleij, K. Rietveld and L. Bosh, *NARES* **13,** 1717 (1985).

170. K. Rietveld, K. Linschooten, C. W. A. Pleij and L. Bosh, *EMBO J.* **3,** 2613 (1984).

171. R. L. Joshi, S. Joshi, F. Chapeville and A. L. Haenni, *EMBO J.* **2,** 1123 (1983).

172. A. Van Belkum, J. P. Abrahams, C. W. A. Pleij and L. Bosh, *NARes* **13,** 7673 (1985).

173. S. Stern, B. Wieser and H. F. Noller, *JMB* **204,** 447 (1988).

174. C. R. Woese and R. R. Guttel, *PNAS* **86,** 3119–3122 (1989).

175. D. S. McPheeter, G. D. Stormo and L. Gold, *JMB* **201,** 517 (1988).

176. C. K. Tang and D. E. Drapper, *Cell* **57,** 531 (1989).

177. T. Jacks, H. D. Madhani, F. R. Masiarz and H. E. Varmus, *Cell* **55,** 447 (1988).

178. I. Brierley, P. Digard and S. C. Inglis, *Cell* **57,** 537 (1989).

179. R. W. Davies, R. B. Wawring, J. A. Ray, T. A. Brown and C. Scazzocchio, *Nature* **300,** 719 (1982).

180. J. D. Puglisi, J. R. Wyatt and I. Tinoco, Jr., *Nature* **331,** 283 (1988).

181. J. R. Wyatt, J. D. Puglisi and I. Tinoco, Jr., *JMB* **214,** 455 (1990).

182. J. R. Wyatt and I. Tinoco, Jr., *in* "RNA World" (R. F. Gesteland and J. F. Atkins, eds.), p. 465. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1993.

183. J. D. Puglisi, J. R. Wyatt and I. Tinoco, Jr., *JMB* **214,** 437 (1990).

184. H. S. Chan and K. A. Dill, *J. Chem. Phys.* **90,** 492 (1989).

185. J.-M. Neefs and R. De Wachter, *NARes* **18,** 5695 (1990).

186. J.-H. Chen, S.-Y. Le and J. V. Maizel, *CABIOS* **8,** 243 (1992).

187. W. Fontana, D. A. M. Konnings, P. F. Stadler and P. Schuster, *Biopolymers* **33,** 1389 (1993).

188. W. Fontana, P. F. Stadler, P. Tarazona, E. D. Weinberger and P. Schuseter, *Phys. Rev. E* **47,** 2083 (1993).

189. P. Schuster, *Artificial Life* **1,** 39 (1994).

190. M. A. Huynen, D. A. M. Konnings and P. Hogeweg, *J. Theor. Biol.* **165,** 251 (1993).

191. A. M. Gutin, A. Yu. Grosberg and E. I. Shakhnovich, *J. Phys. A: Math. Gen.* **26,** 1037 (1993).

192. O. B. Ptitsyn, *Mol. Biol.* **18,** 574 (1984) (in Russian).

193. H. Moine, P. Romby, M. Springer, M. Grunberg-Manago, J.-P. Ebel, B. Ehresmann and C. Ehresmann, *JMB* **216,** 299 (1990).