JOURNAL OF VIROLOGY, July 1992, p. 4144–4153
0022-538X/92/074144-10$02.00/0

Vol. 66, No. 7

The Human Immunodeficiency Virus Type 1 Packaging Signal and Major Splice Donor Region Have a Conserved Stable Secondary Structure

GEOFFREY P. HARRISON AND ANDREW M. L. LEVER*

University of Cambridge Department of Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, United Kingdom

Received 10 February 1992/Accepted 14 April 1992

Interaction of *cis*-acting RNA sequences with nucleocapsid proteins is one of the critical events leading to retroviral genomic RNA packaging. We have derived a potentially stable secondary structure for the packaging signal region of human immunodeficiency virus strain IIIB, using a combination of biochemical analysis and computer modelling. This region encompasses the major splice donor (SD), which is found in a highly structured conserved stem-loop. Comparison with other published human immunodeficiency virus type 1 sequences shows almost absolute nucleotide conservation in base-paired regions required to maintain this structure. Presently and previously described packaging-defective mutants would disrupt the structure. The structure depends on base pairing between nucleotide sequences 5′ of the major SD which are common to both genomic and subgenomic RNAs and sequences 3′ of SD which are unique to the unspliced RNA. This may explain how in retroviruses such as Rous sarcoma virus, mutations in regions common to genomic and subgenomic RNA might prevent packaging of the unspliced mRNA by disrupting a signal structure which can exist only in the genomic RNA species.

Packaging of genomic retroviral RNA into virion particles involves specific selection of two full-length capped polyadenylated viral mRNAs from a large number of cellular mRNAs of identical structure, many of which will be of similar size. Since the full-length viral mRNA is packaged almost exclusively, attention has centered on the region 3′ to the major splice donor (SD), which is unique to this transcript. Study of spontaneously arising and engineered retroviral mutants has revealed the presence of *cis*-acting packaging signals predominantly at the 5′ end of the genomic RNA. In spleen necrosis virus (56), Moloney murine leukemia virus (Mo-MuLV) (34), Rous sarcoma virus (RSV) (26), and human immunodeficiency virus type 1 (HIV-1) (1, 10, 29), deletions between the SD and the *gag* initiation codon were shown to cause a packaging deficiency. In RSV, however (40), packaging signals were identified 5′ of the SD and in regions flanking the v-*src* gene (47). In Mo-MuLV (4), *cis*-acting sequences extending into the *gag* gene were shown to further enhance packaging. A model system for the selection process is that the packaging region (ψ) adopts a stable secondary structure which is recognized by *gag* proteins and allows for selective transport and encapsidation of this RNA into the assembling virion capsid. Such a structure has been recently proposed for Mo-MuLV (2). Dimerization of the RNA would not appear necessary for this selectivity, since freshly budded virions initially contain monomeric RNA (7, 9).

We have used the following three experimental methods to model the RNA secondary structure of the 5′ leader sequence of HIV-1, involving the SD and ψ regions: (i) biochemical analysis, (ii) free-energy minimization, and (iii) nucleotide sequence comparison of cloned, sequenced isolates, which is analogous to phylogenetic comparison.

Biochemical probes. Biochemical and enzymatic structure-specific probes have been used to investigate tRNA precursors (50), tRNAs (15, 37), apolipoprotein II mRNA (44), rRNAs (23, 32, 38, 52, 54, 55, 58, 59), poliovirus (46), and pre-mRNAs (51). These generally involve in vitro transcription of RNA and annealing to form intrachain base pairing. This is followed by modification of the resultant structure with sequence- and structure-specific probes which disrupt cDNA synthesis from a downstream primer. The size of the cDNA corresponds to the point of structural modification.

Free-energy minimization. Free-energy minimization can provide only an outline working model that must be compared with both phylogenetic and biochemical data. Free-energy parameters are known to an accuracy of only 10% at best, and in most cases a structure that is about 80% correct is predicted in this way (61). Small changes in energy parameters used for these algorithms can result in large changes in predictions.

Phylogenetic analysis. Phylogenetic analysis (comparing potential secondary structures in naturally occurring variants and related species and corresponding second-site reversions) has been instrumental in the elucidation of the structures of rRNAs (21, 57), tRNAs (48), class I and II introns (8, 11, 36), small nuclear RNAs (45), and catalytic RNAs (25). Packaging of genomic RNA probably entails common mechanisms, and it is likely that similar RNA secondary and tertiary structural motifs are involved in different lentiviruses.

MATERIALS AND METHODS

Biochemical analysis. Biochemical analysis was carried out by the general procedure of Stern et al. (49; see also references 38, 44, 46, 50, and 55). The *Hin*dIII fragment of HIV-1 strain IIIB (16) from bases 541 to 1086 was cloned into the *Hin*dIII site of the expression vector Bluescript KS II (Stratagene), and RNA was transcribed in vitro from the T3 promoter. In vitro-transcribed RNA was extracted twice

* Corresponding author.

with phenol-chloroform and precipitated by addition of 0.1 volume of 3 M sodium acetate (pH 6.0) and 2.5 volumes of ethanol and treated with 1 μg of RNase-free DNase I per 100 μg of nucleic acid at 0°C in DNase I buffer, for 30 min. RNA was then extracted twice with phenol-chloroform, precipitated by addition of 0.1 volume of 3 M sodium acetate (pH 6.0) and 2.5 volumes of ethanol, and then reannealed by heating to 75°C for 3 min and cooling over a period of between 15 and 45 min to room temperature in reannealing buffer (30 mM Tris-HCl [pH 7.8], 20 mM $MgCl_2$, 300 mM KCl). The following RNA modifications were made, with approximately 2 μg of RNA per reaction mixture: RNase V_1 (Pharmacia) (0 to 1 U) at 0°C for 30 min; RNase T_1 (Boehringer, Lewes, England) (0 to 50 U) at 0°C for 30 min; dimethyl sulfate (DMS) (BDH, Poole, England) (0.05 to 0.3%) at 20°C for 10 min; and kethoxal (United States Biochemical; 10 to 35 mg/ml in 20% ethanol) at 0.1 volume at 20°C for 10 min.

Photoreaction with hydroxymethylpsoralen (psoralen) (Sigma) was with 0.1 to 2.5 μg/μg of RNA irradiated at between 10^{14} and 10^{16} photons per s at a wavelength of 370 nm for 10 min at room temperature.

RNase V_1. RNase V_1 cuts internucleotide bonds in helical regions, leaving a 5′ phosphate. RNase V_1 is not base specific but requires a minimum of 4 to 6 nucleotides that are in helical conformation (1 or 2 bases on either side of the target), whether base paired or single stranded and stacked. Sequences can be sensitive to both V_1 nuclease and a single strand-specific nuclease or chemical reagent (20, 27). This can be the case because a secondary structure is unstable or because the RNA has alternate conformations in equilibrium. RNase V_1 can recognize some single-stranded sites because single-stranded polynucleotides can transiently adopt helical conformation.

Psoralen. Psoralen is a planar, aromatic compound that intercalates double-stranded regions of nucleic acids. Irradiation of psoralen-nucleic acid complexes with long-wave UV light results in the formation of covalent psoralen adducts normally with pyrimidines and principally with uridines. Where there are neighboring pyrimidines facing on a paired strand, a diadduct can form so that the two strands are covalently cross-linked through a psoralen molecule (32, 51).

Kethoxal. Kethoxal (2-keto-3-ethoxybutyraldehyde) reacts with unpaired guanines at N-1 and N-2, causing reverse transcriptase to terminate or pause. If a guanine residue is not modified by this chemical, it may be involved in secondary or tertiary interaction.

RNase T_1. RNase T_1 cleaves internucleotide bonds 3′ of unpaired guanine residues (leaving a 3′ phosphate). The enzyme may be limited in cleaving some unpaired G residues by steric hindrance.

DMS. DMS methylates unpaired adenine at N-1 and unpaired cytosine at N-3, causing reverse transcriptase to terminate or pause (it also methylates unpaired Gs at N-7, but this modification does not stop reverse transcriptase). Residues protected from DMS modification are probably involved in secondary or tertiary interactions.

Modified RNA was extracted twice with phenol-chloroform and precipitated by addition of 0.1 volume of 3 M sodium acetate (pH 6.0) and 2.5 volumes of ethanol. Synthetic oligonucleotide primers (1 μl of 100 mM) were annealed to 1 μg of modified RNA in 10 μl of renaturation buffer (40 mM Tris-HCl [pH 8.3], 240 mM KCl, 4 mM dithiothreitol) by heating to 75°C for 3 min and then allowing to cool over 15 min to room temperature.

Extension analyses from these primers were carried out with 1 U of avian myeloblastosis virus reverse transcriptase (Northumbrian Biologicals Ltd., Cramlington, England) per 5 μl of hybridization mix at 55°C for 30 min in a solution containing 50 mM Tris-HCl (pH 8.3), 6 mM $MgCl_2$, 40 mM KCl with 1 mM each dCTP, dTTP, and dGTP, and 1 μl of [α-^{32}P]dATP (ICN-Flow; 3,000 Ci/mM). cDNAs were precipitated under ethanol, washed with 70% ethanol, dried, and dissolved in 1× Tris-EDTA, and 1/10 volume of formamide dye mix was added. The samples were heated to 90°C for loading onto 6% polyacrylamide–7 M urea gels. Pauses or stops give rise to bands corresponding to the length of DNA from the 5′ end of the primer to the nucleotide immediately preceding the modified position. Identification of the modified base was facilitated by electrophoresing samples against dideoxy sequencing ladders (42) made from the same sequence with the same primer. For this purpose, the *Hin*dIII fragment of HIV-1 strain IIIB from bases 541 to 1086 was cloned into the *Hin*dIII site of M13mp18. On other occasions, another known sequence ladder was employed to determine the sizes of bands. Enzymatic cleavage causes unique termination bands because of the fragmentation of the template for primer extension. Unmodified RNA provided a completely reproducible pattern of stops (44) (referred to as invariable bands in Fig. 1 to 5), from which the positions of new stops could be ascertained.

Chemical modifications by DMS, kethoxal, and psoralen cause reverse transcriptase to terminate or pause as it moves along the template. Bands in the lanes of the untreated RNA arise from nicks in the RNA template, from strong secondary-structure features, or from sequence-dependent termination of cDNAs. Pauses or stops in reverse transcriptase caused by base modification give rise to bands corresponding to the length of DNA from the 5′ end of the primer to the nucleotide immediately preceding the modified position.

The relative quantity of truncated cDNA in a given band is proportional to the reactivity or digestibility of the nucleotide responsible for the band.

The apparent site of an enzymatic cleavage is displaced 5′ but never 3′ to the actual cleavage-modification site (44). This is because the nucleotide responsible for a stop itself cannot base pair and the corresponding cDNA fragment extends up to but does not include the complement to the modified position.

Free-energy minimization. Free-energy minimization studies were carried out by folding windows of 100 to 150 bases at 10-bp intervals from bases 630 to 840 (Los Alamos AIDS data base numbering) (39). The programs employed were FOLD and SQUIGGLES (University of Wisconsin Genetics Computer Group) running on a VAX (13) and RNA Secondary Structure Predictor (version 1.22; 1987) (62), all of which work on the basis of free-energy minimization.

Phylogenetic comparisons. Phylogenetic comparisons were made between the potential secondary structures of this region in the 18 different HIV-1 isolates and simian immunodeficiency virus strain CPZ (SIV_{CPZ}) (39).

RESULTS

Biochemical analysis. The results of only a sample of individual biochemical analyses are shown in Fig. 1 to 5. All modifications were mapped in at least two independent experiments, and the majority were mapped three or more times.

(i) Primer binding site to base 677. In our study, DMS (which methylates unpaired Cs and As) modified all of the As between 644 and 669 (Fig. 4), which are unpaired in the

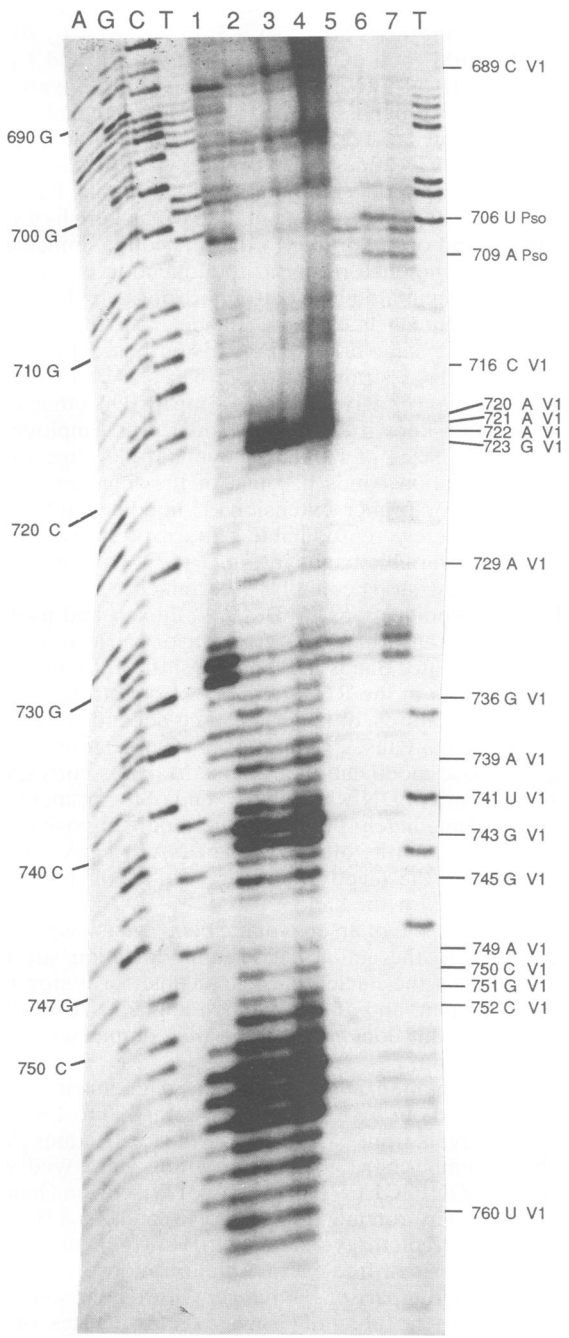

FIG. 1. Autoradiograph of a gel showing comparison of reverse transcribed RNA templates. The sequence of the primer used in all lanes was 5′-ATCTCTCTCCTTCTAGCC-3′ (nucleotides 790 to 773). The dideoxy sequencing ladder was generated from the same sequence. The T track was duplicated next to lane 7. Shown are unmodified RNA (lane 1), RNA digested with 0.25 U of RNase V_1 (lanes 2 to 4), and RNA irradiated at 350 nm with 10 μg of psoralen per ml (lane 5), with 20 μg of psoralen per ml (lane 6), and with 40 μg of psoralen per ml (lane 7). The sequence is numbered on the left side. Stops specific to modified RNA are labelled on the right. V1, RNase V_1 modification; Pso, psoralen modification.

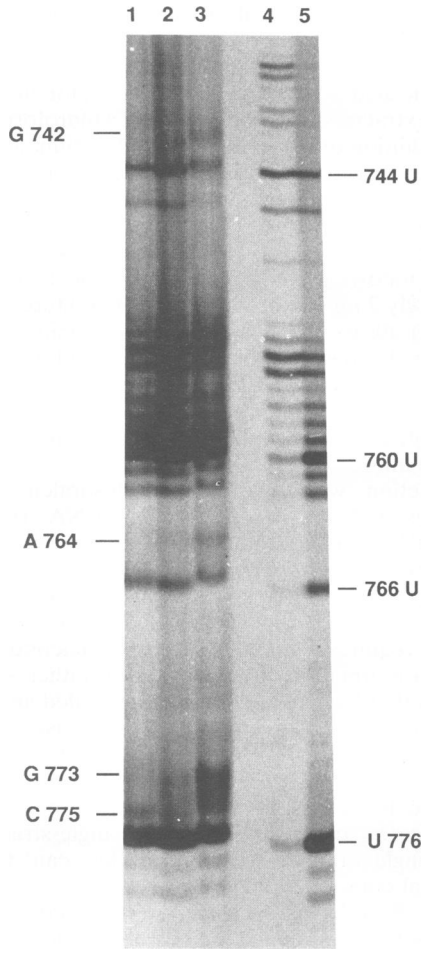

FIG. 2. Autoradiograph of a gel showing comparison of reverse transcribed RNA templates. The sequence of the primer used in all lanes was 5′-CTAATTCTCCCCCGC-3′ (nucleotides 828 to 814). No dideoxy sequencing ladder was generated with these cDNAs. Nucleotide positions were identified from the invariant pattern of stops. Shown are RNA digested with 0.017 U of RNase V_1 (lane 1), with 0.10 U of RNase V_1 (lane 2), and with 0.25 U of RNase V_1 (lane 3) and unmodified RNA (lanes 4 and 5). Stops specific to modified RNA are labelled on the left. The sequence is numbered on the right.

computed model (Fig. 6). RNase T_1 (which cleaves unpaired Gs) digested all of the Gs in the region 654 to 667G (Fig. 4).

(ii) Stem I. At the base of stem I, RNase V_1 (which cleaves within or near helical structures) digested at bases 681U and 787A (data not shown), which face one another in both of the computer predictions. Also, there was a V_1 site at nucleotide 676A (data not shown), despite the lack of predicted base pairing. It has previously been reported (33, 50) that V_1 can reach out beyond stem regions and make cuts in adjacent unpaired regions. DMS and kethoxal (the latter methylates unpaired Gs) reacted with bases in stem I and RNase T_1 digested at 786 (data not shown). We therefore conclude that stem I is labile. The stem may exist transiently or in equilibrium with the unpaired state. Despite attempts to produce alternative computer models in which this region is single stranded, stem I and stem II were invariably predicted to be base paired.

(iii) Stem II. In stem II, an RNase V_1 at 729A (Fig. 1) faces a site of psoralen photoreaction at 696U (psoralen forms

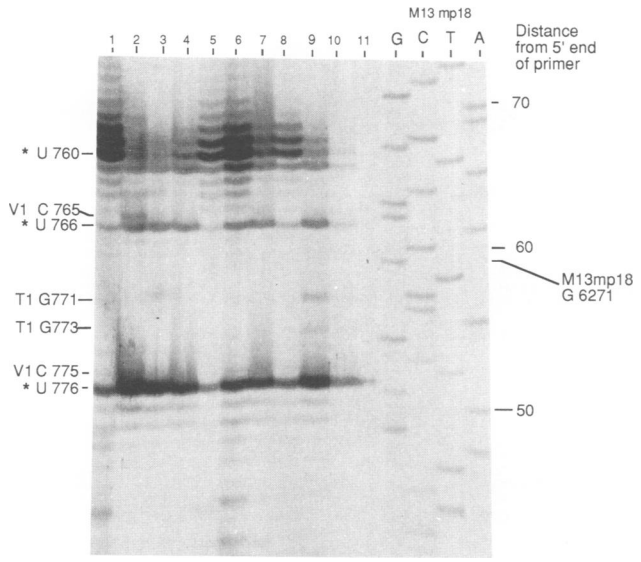

FIG. 3. Autoradiograph of a gel showing comparison of reverse transcribed RNA templates. The sequence of the primer used in all lanes was 5′-CTAATTCTCCCCCGC-3′ (nucleotides 828 to 814). The dideoxy sequencing ladder was generated from M13mp18. Nucleotide positions were determined from this and from the invariant pattern of stops. Shown are unmodified RNA (lanes 1 and 5) and RNA digested with 0.017 U of RNase V_1 (lanes 2 and 6), with 0.10 U of RNase V_1 (lanes 3 and 7), with 0.25 U of RNase V_1 (lanes 4 and 8), with 3 U of RNase T_1 (lane 9), with 20 U of RNase T_1 (lane 10), and with 50 U of RNase T_1 (lane 11). Stops specific to modified RNA are labelled on the left (V_1 or T_1), as are invariant stops (*) in unmodified samples. The sequence is numbered on the right.

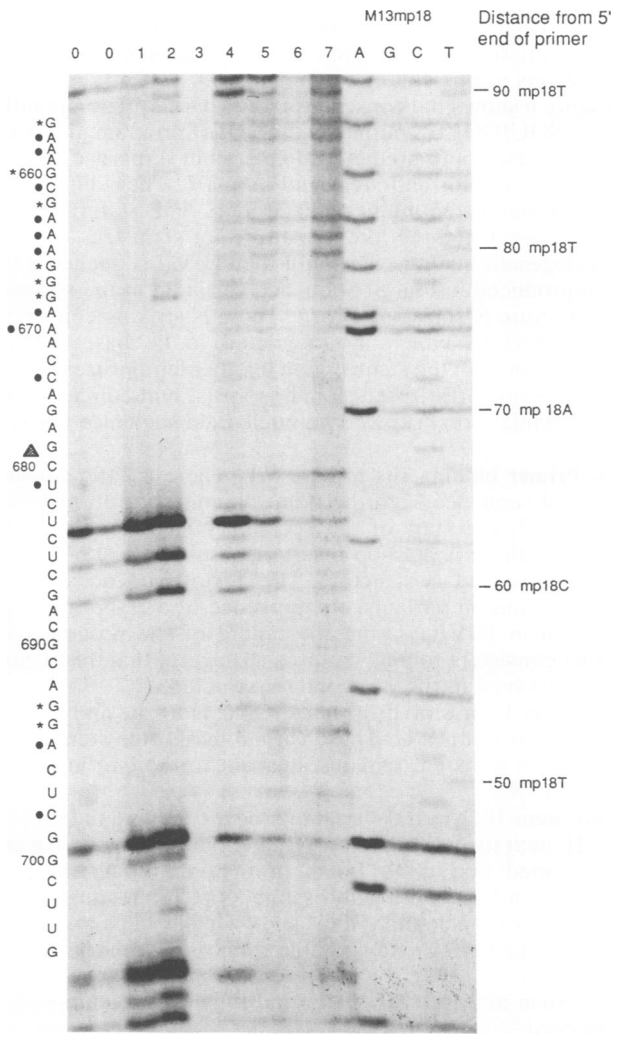

FIG. 4. Autoradiograph of gel showing comparison of reverse transcribed modified RNA templates. The primer used for cDNA synthesis was 5′-ACCAGTCGCCGCCCC-3′ (nucleotides 744 to 730). The dideoxy sequencing ladder was generated from M13mp18. The lengths of the cDNAs were determined from the sequencing ladder and by comparison with invariant stops. Shown are untreated RNA (lane 0), RNA digested with 5 U of RNase T_1 (lane 1) and with 30 U of RNase T_1 (lane 2), blank (lane 3), and RNA modified with 0.1% DMS (lane 4), with 0.2% DMS (lane 5), with 0.3% DMS (lane 6), and with 3.0% DMS (lane 7). Stops specific to modified RNA are labelled on the left. *, RNase T_1; ●, DMS; ▲, bases 677 to 679 absent.

adducts in double-stranded regions), and there are very strong V_1 sites at 720C to 723G (Fig. 1). There was a psoralen modification on the opposite strand at 701C (data not shown).

RNase T_1 and DMS modified this region (data not shown); therefore, it may be unstable under some conditions. There was a psoralen modification at 709A (Fig. 1). It is unusual for psoralen to form adducts with purines, although this has been reported elsewhere (3). This adduct may represent a long-range interaction or pseudoknot. There could be base pairing across the loop IId, but there were no psoralen adducts detected elsewhere in the loop. There was, however, no biochemical evidence that the loop is single stranded.

(iv) Stem III. The helical nature of this region was very clear. RNase V_1 has also cut across the junction with stem II and along stem II (Fig. 1). RNase V_1 cuts at sharp bends in the backbone between two regions of helical conformation (27). At loop III, RNase V_1 has cut nucleotides predicted to be unpaired, which are adjacent to the ends of the loop, probably for the reasons given above. Some instability at the end of stem III is indicated by the RNase T_1 cleavage at 747G (Fig. 5) and the DMS modification at 750C (data not shown).

(v) Stem IV. RNase V_1 repeatedly cut in stem IV at positions 764A (data not shown), 765C (Fig. 2), 766U, 773G (Fig. 2), and 775C (Fig. 2). RNase T_1 digested the predicted nonpaired Gs at 771 and 773 (Fig. 3 and 5) in loop IV, and kethoxal reacted with all of the Gs (data not shown). Part a of stem IV appears to be labile, with vulnerability to chemical modification in the region in which non-Watson-Crick pairing was predicted (data not shown). All of the biochemical data are presented in Fig. 7.

Free-energy minimization. The free-energy minimization program predicted a similar complex stem-loop structure between the primer binding site and the *gag* initiation codon starting at base 677 and ending at base 789 (Fig. 6). Modelling of the region 5′ to base 677 produced inconsistent folding patterns with low stability, most probably reflecting the large natural sequence variation in this region. Surprisingly, no folding pattern could be obtained for the region downstream of base 789 (3′ to the *gag* initiation codon). The only predicted structure had a single small stem-loop with a free energy of −0.8 kcal (ca. −3.4 kJ)/mol, which can be consid-

ered nonsignificant. RNA secondary structure is known to inhibit initiation of translation and would probably be disadvantageous in this coding region (18, 22).

Figure 6 shows the consensus model based predominantly on the SQUIGGLES output of the FOLD program. The two programs used both predicted stem I, stem II (parts c and d), loops IIc and IId, and stem and loop IV. They differed at stem III and at the multibranched loop. The SQUIGGLES program predicted the folding pattern for stem III.

Phylogenetic analysis. The published HIV-1 sequences (39) are reproduced in Fig. 8. Sequence variation in the secondary structure is shown in Fig. 6. There is striking conservation of regions which are predicted to be base paired. Comparison with the same region in other lentiviruses (visna virus, caprine arthritis-encephalitis virus, and equine infectious anemia virus) showed no nucleotide sequence conservation.

(i) Primer binding site to base 677. There is little conservation of sequence 3′ to the primer binding site from bases 650 to 677, the start of the secondary structure of ψ. In HIV_{MAL}, there is a 13-bp insert at base 670, just 5′ to the predicted secondary structure. The lack of sequence conservation in natural variants, the presence of a 20-bp insert at base 670 in HIV_{MAL}, and the failure of the programs to predict consistent folding patterns all suggest that this region is not involved in the ψ secondary structure.

(ii) Stem I. In 8 of 19 isolates, 680C is an A, and in 2 of these isolates, it is a U. These variations shorten stem I, but the base of the stem remains adjacent to the *gag* initiation codon at 789.

(iii) Stem II. Natural variations occur only at the end of loop II, next to and within loop IId. Variations in loop IId are diagrammed in Fig. 9. In all of the variations, there is potential for base pairing across the loop. Interestingly, in 7 of 19 isolates, nucleotide 718 is an A which can pair with the U at 706, leading to a more stable stem. In all other isolates, 718 is a G.

(iv) Stem and loop III. Stem and loop III are completely conserved.

(v) Stem IV. Stem IV, part b, is completely conserved, but part a, near the multibranched loop, is completely disrupted in SIV_{CPZ} (Fig. 10). The only variation in loop IV is an A-to-U substitution in SIV_{CPZ} at base 772.

(vi) Multibranched loop. The greatest natural variability in the model is in the multibranched loop (Fig. 10), which will result in variations in the angles between stems II, III, and IV. The angle between the stems and the unpaired nucleotides in the loops may be crucial in RNA and protein interactions (5) such as those essential to virus assembly.

(vii) Downstream to base 789. Downstream to base 789, there is little sequence divergence, almost certainly because of constraints imposed by the *gag* coding functions.

None of the substitutions or insertions found in the phylogenetic comparison significantly disrupt the structure.

DISCUSSION

The three methods of secondary-structure analysis provide a remarkable level of consensus for the model presented. Deletion mutants of HIV-1 with a packaging-defective phenotype, reported by ourselves and others (1, 10, 29), have involved the 3′ half of the structure, between SD and the *gag* AUG. All of those described so far would effectively remove stem-loop IV and disrupt base pairing of stems I and III. The first described packaging mutant (HXBΔP1) had a 19-bp deletion from 753 to 773 which showed a level of

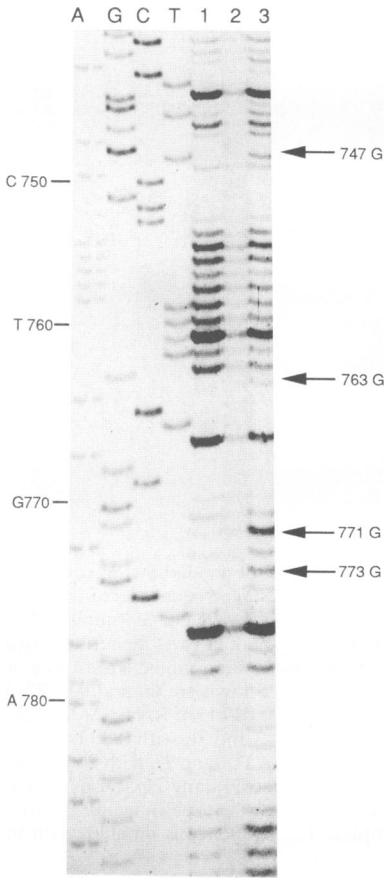

FIG. 5. Autoradiograph of a gel showing comparison of reverse transcribed RNA templates. The sequence of the primer used in all lanes was 5′-CTAATTCTCCCCCGC-3′ (nucleotides 828 to 814). The dideoxy sequencing ladder was generated from the same sequence. Lanes: 1, unmodified RNA; 2, RNA digested with 50 U of RNase T_1; 3, RNA digested with 10 U of RNase T_1. Stops specific to modified RNA are labelled on the right. The sequence is numbered on the left.

genomic packaging of about 2 to 5% of that of the wild-type virus. A more extensive deletion (HXBΔP2) from 750 to 785 packages itself at a level of about 50% of that of HXBΔP1 (data not shown), indicating that there is a direct relationship between the extent of disruption of this region and the decline in packaging efficiency. A further deletion totalling 103 bp in length and extending 5′ of SD gives a more profound packaging defect than one confined to 3′ (30), and, in RSV, deletions purely 5′ of SD give a packaging-defective phenotype (31, 47). With a secondary structure dependent on regions 5′ and 3′ of SD, it is clear why mutations in regions apparently common to genomic and subgenomic RNAs can produce a packaging-defective phenotype. Interaction of such RNA containing this secondary structure (or one with tertiary interactions based on it) with the *gag* precursor protein or the nucleocapsid protein is probably a critical early event in genomic RNA recognition.

The sequence variants noted are found in regions which appear unimportant for base pairing. This may mean that variation in these regions is tolerable and unimportant. Alternatively, the variants may be defining a further level of specificity of packaging between isolates. Detailed study involving attempts to cross package RNA into heterologous

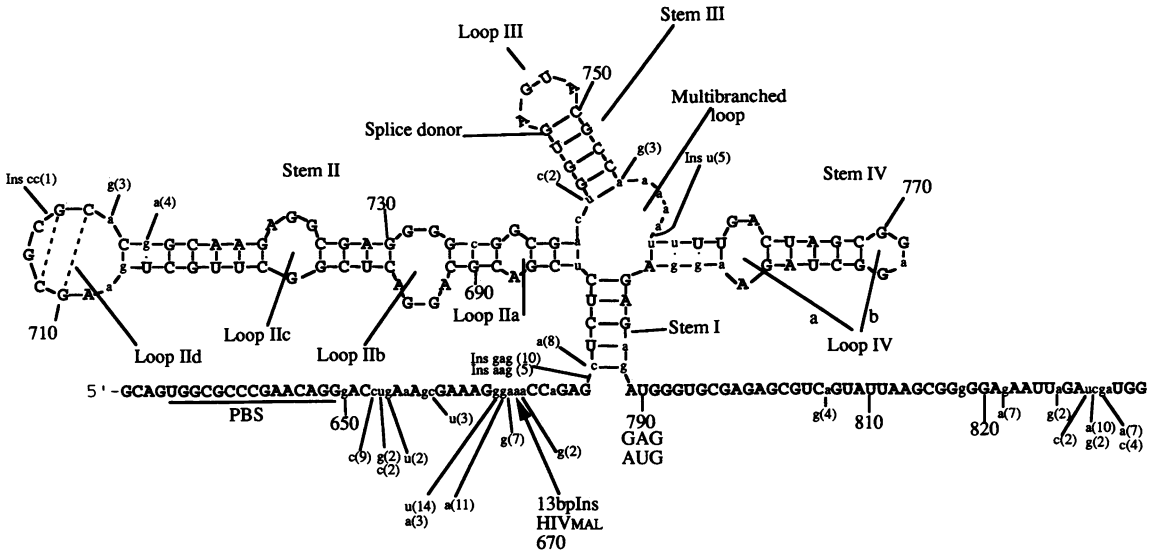

FIG. 6. RNA secondary-structure prediction, according to SQUIGGLES output from FOLD (University of Wisconsin Genetics Computer Group) running on a VAX, predicted a free-energy value of approximately −31.9 kcal (ca. −134 kJ). Residues in uppercase letters are absolutely conserved and those in lowercase letters are substituted in 1 of 19 homologous sequences. Larger insertions and more frequent substitutions are also in lowercase letters, with frequency among the 19 isolates in parentheses. PBS, primer binding site. Ins, insert.

virions would be necessary to answer this and could also address the question of whether different isolates showed variable degrees of efficiency of encapsidation due to nucleotide sequence variation.

In the vast majority of the structure, single- and double-stranded regions were clearly demarcated biochemically; however, some small regions in our model were modified with both single- and double-stranded probes. While we have termed these latter regions labile, it is impossible to be certain of how stable they may be in vivo in the presence of viral and cellular proteins. In vitro, it was noticeable that in labile areas, a prolonged annealing time (greater than 45 min) correlated with the appearance of double-stranded structures which sometimes were single stranded with annealing times of approximately 15 min. However, our studies make it clear that it is not tenable to infer that an RNA strand is or is not base paired by failure to modify with a specific probe. All structural motifs should be positively identified.

A similar study with Mo-MuLV, in which two somewhat different models were predicted from the biochemical data when different computer programs were used, has recently been reported (2). In that study, only single-strand-RNA-modifying agents were used, and failure to modify was interpreted as revealing the presence of RNA-RNA interactions, which we have found can be unreliable. In our study, despite the use of different computer programs, identical models were predicted for the complete structure, with the exception of stem III. The structure predicted by SQUIGGLES is consistent with the other two methods of modelling, and the large number of individual bases whose interactions we positively identified by chemical probing and the consistency of the sequence comparison lead us to suggest that the alternative computer-predicted structure is less likely.

This model of RNA secondary structure inevitably has limitations; it does not predict tertiary or quaternary RNA

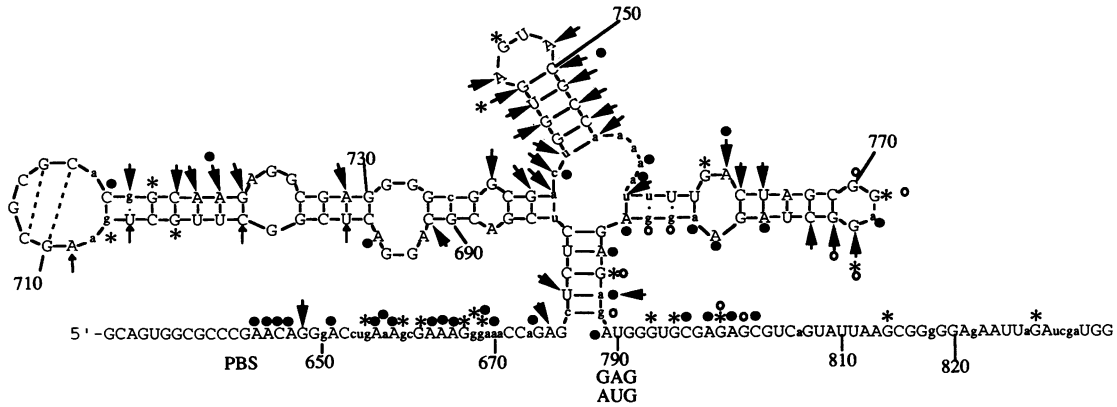

FIG. 7. Sites of enzymatic cleavage and positions of chemical modification. ►, RNase V_1; →, psoralen; *, RNase T_1; ●, DMS; ○, kethoxal; PBS, primer binding site.

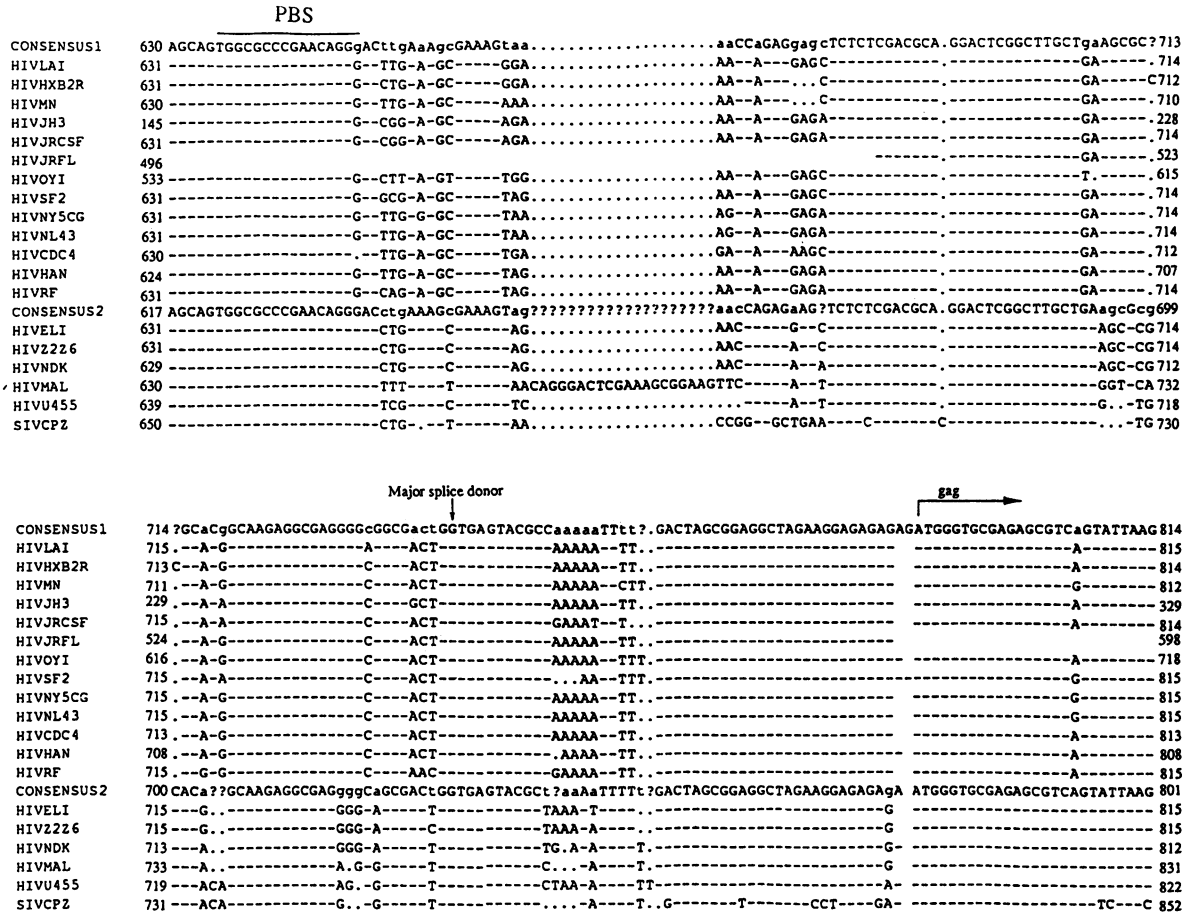

FIG. 8. Linear sequence analysis of 18 HIV-1 strains taken from the AIDS data base alongside the SIV_{CPZ} sequence. Regions of nucleotide identity (−) and difference are indicated. Regions of identity can be seen between those of the primer binding site (PBS) and the *gag* coding region, one of which holds the SD motif.

interactions and does not take account of the effects of RNA-protein interactions which must be involved in RNA encapsidation. Nevertheless, it provides a useful working model which raises interesting questions about the control of packaging, splicing, and translation in HIV-1.

Retroviral RNA packaging also involves dimerization of the genomic strand at a stage in HIV-1 that is still unknown (6, 41), although in RSV it apparently occurs after initial budding (7, 9). The dimer linkage site is postulated to be close to the *gag* initiation codon (35). It is conceivable that dimer formation might involve a strand swap between the labile palindromic stem I regions of two RNA molecules. RNA from this region shows dimer formation in sense and antisense (data not shown); however, these dimers are more stable than would be expected for Watson-Crick pairing.

The region described is also intimately involved in other RNA functions, including splicing, and the absolute conservation of the helical region containing the SD is very striking. The SD RNA is known to form a base pair with the U1 small nuclear ribonucleoprotein (60), and splice sites are known to be preferentially nuclease accessible (14). The easy accessibility to V_1 RNase in our studies is consistent with this.

The cloverleaf appearance of the model is also reminiscent of the secondary structure of tRNA, and it is notable that

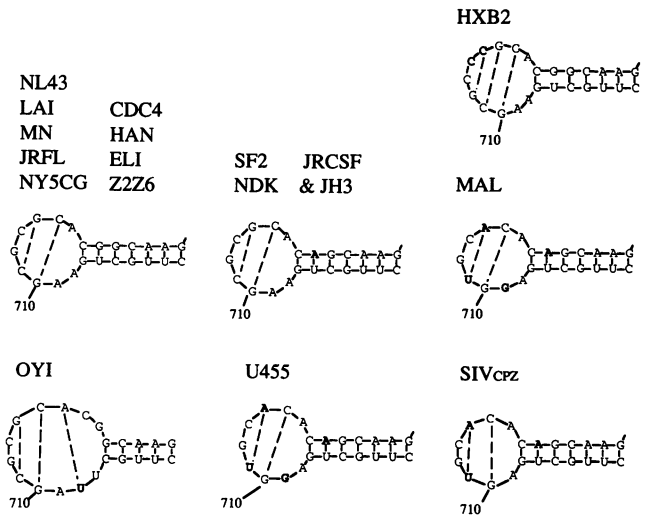

FIG. 9. RNA secondary-structure prediction for stem and loop IId in sequenced isolates. Base substitutions differing from those in HXB2 are shown in boldface type.

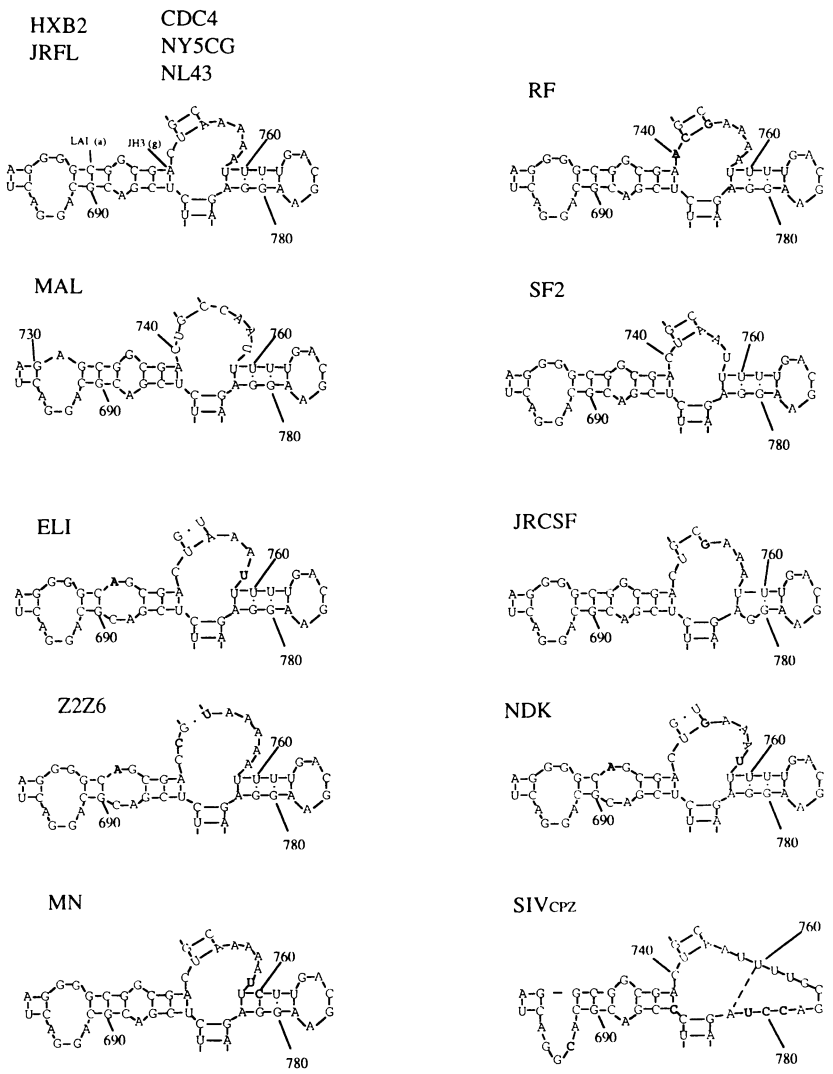

FIG. 10. RNA secondary structure predictions for the multibranched loop in sequenced isolates. Base substitutions which differ from bases in HXB2 are shown in boldface type.

splicing of the pre-tRNA intron occurs at a somewhat analogous position in stem-loop III. It is conceivable that retroviruses represent an intermediate stage in the splicing pattern between pre-tRNA and pre-mRNA or at least have features common to both.

Deletion mutants in the region between SD and the *gag* initiation codon have been noted to cause aberrant splicing (9a), and we have noted some reduction in spliced gene products with some mutations in this region (data not shown). However, this latter reduction could be an effect on translation, since RNA secondary or tertiary structures influence translation initiation and efficiency (12, 17–19, 22, 43, 53).

In prokaryotes, initiation codons tend to occur in regions devoid of RNA secondary structure, whereas internal AUGs are found to be preceded by regions of potential secondary structure (17), and it has been postulated that a loop-exposed AUG correlates with efficiency of expression (22, 24). Treatments that disrupt secondary structure tend to increase the ability of ribosomes to recognize correct sites (28). The presence of a secondary (or tertiary) structure just prior to the *gag* AUG may be central to the control and expression of both *gag* and *pol* proteins in HIV-1.

ACKNOWLEDGMENTS

We thank J. Almond for useful discussions.

This work was funded by the Medical Research Council AIDS-Directed Programme (United Kingdom).

REFERENCES

1. **Aldovini, A., and R. A. Young.** 1990. Mutations of RNA and protein sequences involved in human immunodeficiency virus type 1 packaging result in production of noninfectious virus. J. Virol. **64**:1920–1926.
2. **Alford, R. L., S. Honda, C. B. Lawrence, and J. W. Belmont.** 1991. RNA secondary structure analysis of the packaging signal for Moloney murine leukemia virus. Virology **18**:611–619.
3. **Bachellerie, J.-P., J. F. Thompson, M. Wegnez, and J. F. Hearst.** 1981. Identification of the modified nucleotide produced by

covalent photoaddition of hydroxymethyltrimethylpsoralen to RNA. Nucleic Acids Res. **9**:2207–2222.
4. **Bender, M. A., T. D. Palmer, R. E. Gelinas, and A. D. Miller.** 1987. Evidence that the packaging signal of Moloney murine leukemia virus extends into the *gag* region. J. Virol. **61**:1639–1646.
5. **Bhattacharyya, A., A. I. H. Murchie, and D. M. J. Lilley.** 1990. RNA bulges and the helical periodicity of double stranded RNA. Nature (London) **343**:484–487.
6. **Bieth, E., C. Gabus, and J.-L. Darlix.** 1990. A study of the dimer formation of Rous sarcoma virus RNA and of its effect on viral protein synthesis *in vitro*. Nucleic Acids Res. **18**:119–127.
7. **Canaani, E., K. V. D. Helm, and P. Duesberg.** 1973. Evidence for 30-40S RNA as precursor of the 60-70S RNA of Rous sarcoma virus. Proc. Natl. Acad. Sci. USA **70**:401–405.
8. **Cech, T. R., N. K. Tanner, I. Tinoco, B. R. Weir, M. Zuker, and P. S. Periman.** 1983. Secondary structure of the Tetrahymena ribosomal RNA intervening sequence. Proc. Natl. Acad. Sci. USA **80**:3903–3907.
9. **Cheung, K.-S., R. E. Smith, M. P. Stone, and W. K. Joklik.** 1972. Comparison of immature (rapid harvest) and mature Rous sarcoma virus particles. Virology **50**:851–864.
9a. **Clavel, F.** Personal communication.
10. **Clavel, F., and J. M. Orenstein.** 1990. A mutant of human immunodeficiency virus with reduced RNA packaging and abnormal particle morphology. J. Virol. **64**:5230–5234.
11. **Davies, R. W., R. B. Waring, J. A. Ray, T. A. Brown, and C. Scazzocchio.** 1982. Making ends meet: a model for RNA splicing in fungal mitochondria. Nature (London) **300**:719–724.
12. **De Smit, M., and J. van Duin.** 1990. Control of prokaryotic translational initiation by mRNA secondary structure, p. 1–35. *In* W. E. Cohn and K. Moldare (ed.), Progress in nucleic acid research and molecular biology, vol. 38. Academic Press Inc., San Diego, Calif.
13. **Devereux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence programs for the VAX. Nucleic Acids Res. **12**:387–395.
14. **Dodgson, J. B., and J. D. Engel.** 1983. The nucleotide sequence of the adult chicken alpha-globin genes. J. Biol. Chem. **258**: 4623-4629.
15. **Favorova, O. O., F. Fasioli, G. Keith, S. K. Vassilenko, and J.-P. Ebel.** 1981. Partial digestion of tRNA-aminoacyl tRNA synthetase complexes with cobra venom ribonuclease. Biochemistry **20**:1006–1011.
16. **Fisher, A. G., E. Collati, L. Ratner, R. C. Gallo, and F. Wong-Staal.** 1985. A molecular clone of HTLV-III with biological activity. Nature (London) **316**:262–265.
17. **Ganoza, M. C., E. C. Kofoid, P. Marliere, and B. G. Louis.** 1987. Potential secondary structure at translation initiation sites. Nucleic Acids Res. **15**:345–360.
18. **Gold, L.** 1988. Posttranscriptional regulatory mechanisms in Escherichia Coli. Annu. Rev. Biochem. **57**:199–233.
19. **Gualerzi, C. O., and C. L. Pon.** 1990. Initiation of mRNA translation in prokaryotes. Biochemistry **29**:5881–5889.
20. **Guerrier-Takada, C., and S. Altman.** 1984. Structure in solution of M1 RNA, the catalytic subunit of ribonuclease P from *Escherichia coli*. Biochemistry **23**:6327–6334.
21. **Gutell, R., and G. E. Fox.** 1988. A compilation of large subunit RNA sequences presented in a structural format. Nucleic Acids Res. **16**(Suppl.):r175.
22. **Hall, M. N., J. Gabay, M. Debarbouille, and M. Schwartz.** 1982. A role for mRNA secondary structure in the control of translation initiation. Nature (London) **295**:616–618.
23. **Inque, T., and T. R. Cech.** 1985. Secondary structure of the circular form of the Tetrahymena rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase. Proc. Natl. Acad. Sci. USA **82**:648–652.
24. **Iserentant, D., and W. Fiers.** 1980. Secondary structure of mRNA and efficiency of translation initiation. Gene **9**:1–12.
25. **James, B. D., G. J. Olsen, and N. R. Pace.** 1989. Phylogenetic comparative analysis of RNA secondary structure. Methods Enzymol. **180**:227–239.
26. **Katz, R. A., R. W. Terry, and A. M. Skalka.** 1986. A conserved *cis*-acting sequence in the 5′ leader of avian sarcoma virus RNA is required for packaging. J. Virol. **59**:163–167.
27. **Kean, J. M., and D. E. Draper.** 1985. Secondary structure of 345-base RNA fragment covering the S8/S15 protein binding domain of *Escherichia coli* 16S ribosomal RNA. Biochemistry **24**:5052–5061.
28. **Kozak, M., and D. Nathans.** 1972. Translation of the genome of a ribonucleic acid bacteriophage. Bacteriol. Rev. **36**:109–134.
29. **Lever, A. M. L., H. Gottlinger, W. Haseltine, and J. Sodroski.** 1989. Identification of a sequence required for efficient packaging of human immunodeficiency virus type 1 RNA into virions. J. Virol. **63**:4085–4087.
30. **Lever, A. M. L., J. H. Richardson, and G. P. Harrison.** 1991. Retroviral RNA packaging. Biochem. Soc. Trans. **19**:963–966.
31. **Linial, M., E. Medeiros, and W. S. Hayward.** 1978. An avian oncovirus mutant (SE21Q1b) deficient in genomic RNA biological and biochemical characterization. Cell **15**:1371–1381.
32. **Lipson, S. E., and J. E. Hearst.** 1988. Psoralen cross-linking of ribosomal RNA. Methods Enzymol. **164**:330–341.
33. **Lowman, H. B., and D. E. Draper.** 1986. On the recognition of helical RNA by cobra venom V_1 nuclease. J. Biol. Chem. **261**:5396–5403.
34. **Mann, R., R. C. Mulligan, and D. Baltimore.** 1983. Construction of a retroviral packaging mutant and its use to produce helper free defective retrovirus. Cell **33**:153–159.
35. **Marquet, R., F. Baudin, C. Gabus, J.-L. Darlix, M. Mougel, C. Ehresmann, and B. Ehresmann.** 1991. Dimerization of human immunodeficiency virus (type 1) RNA: stimulation by cations and possible mechanism. Nucleic Acids Res. **19**:2349–2357.
36. **Michel, F., and B. Dujon.** 1983. Conservation of RNA secondary structures in two intron families including mitochondrial-, chloroplast- and nuclear-encoded members. EMBO J. **2**:33–38.
37. **Moazed, D., and H. F. Noller.** 1986. Transfer RNA shields specific nucleotides in 16S ribosomal RNA from attack by chemical probes. Cell **47**:985–994.
38. **Moazed, D., S. Stern, and H. F. Noller.** 1986. Rapid chemical probing of conformation in 16S ribosomal RNA and 30S ribosomal subunits using primer extension. J. Mol. Biol. **187**:399–416.
39. **Myers, G., B. Korber, J. A. Berzofsky, R. F. Smith, and G. N. Pavlakis (ed.).** 1991. Human retroviruses and AIDS. Los Alamos National Laboratory, Los Alamos, N.Mex.
40. **Pugatsch, T., and D. W. Stacey.** 1983. Identification of a sequence likely to be required for avian retroviral packaging. Virology **128**:505–511.
41. **Roy, C., N. Tounekti, M. Mougel, J.-L. Darlix, C. Paoletti, C. Ehresmann, B. Ehresmann, and J. Paoletti.** 1990. An analytical study of the dimerization in *in vitro* generated RNA of Moloney murine leukaemia virus. Nucleic Acids Res. **18**:7287–7292.
42. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
43. **Schulz, V. P., and W. S. Reznikoff.** 1990. *In vitro* secondary structure analysis of mRNA from *lac Z* translation initiation mutants. J. Mol. Biol. **211**:427–445.
44. **Shelness, G. S., and D. L. Williams.** 1985. Secondary structure analysis of apolipoprotein II mRNA using enzymatic probes and reverse transcriptase. J. Biol. Chem. **260**:8637–8646.
45. **Siliciano, P. G., M. H. Jones, and C. Guthrie.** 1987. *Saccharomyces cerevisiae* has a U1-like small nuclear RNA with unexpected properties. Science **237**:1484–1487.
46. **Skinner, M. A., V. R. Racaniello, G. Dunn, J. Cooper, P. D. Minor, and J. W. Almond.** 1989. New model for the secondary structure of the 5′ non-coding RNA of polio virus is supported by biochemical and genetic data that also show that RNA secondary structure is important in neurovirulence. J. Mol. Biol. **207**:379–392.
47. **Sorge, J., W. Ricci, and S. H. Hughes.** 1983. *cis*-Acting RNA packaging locus in the 115-nucleotide direct repeat of Rous sarcoma virus. J. Virol. **48**:667–675.
48. **Sprinzl, M., T. Hartmann, F. Meissner, J. Moll, and T. Vorderwulbecke.** 1987. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. **15**(Suppl.):r53.

49. **Stern, S., D. Moazed, and H. F. Noller.** 1988. Analysis of RNA structure using chemical and enzymatic probing monitored by primer extension. Methods Enzymol. **164:**481–489.
50. **Swerdlow, H., and C. Guthrie.** 1984. Structure of intron-containing tRNA precursors. J. Biol. Chem. **259:**5197–5207.
51. **Teare, J., and P. L. Wollenzien.** 1990. The structure of a pre-mRNA molecule in solution determined with a site directed cross linking reagent. Nucleic Acids Res. **18:**855–864.
52. **Thompson, J. F., and J. E. Hearst.** 1983. Structure of *E. coli* 16S RNA elucidated by psoralen crosslinking. Cell **32:**1355–1365.
53. **van de Guchte, M., E. van der Lende, J. Kok, and G. Venema.** 1991. A possible contribution of mRNA secondary structure to translation initiation efficiency in Lactococcus lactis. FEMS Microbiol. Lett. **81:**201–208.
54. **Vassilenko, S. K., P. Carbon, J.-P. Ebel, and C. Ehresmann.** 1981. Topography of 16S RNA in 30S subunits and 70S ribosomes accessibility to cobra venom ribonuclease. J. Mol. Biol. **152:**699–721.
55. **Walker, T. A., K. D. Johnson, G. J. Olsen, M. A. Peters, and N. R. Pace.** 1982. Enzymatic and chemical structure mapping of mouse 28S ribosomal ribonucleic acid contacts in 5.8S ribosomal ribonucleic acid. Biochemistry **21:**2320–2329.
56. **Watanabe, S., and H. M. Temin.** 1982. Encapsidation sequences for spleen necrosis virus, an avian retrovirus, are between the 5′ long terminal repeat and the start of the *gag* gene. Proc. Natl. Acad. Sci. USA **79:**5986–5990.
57. **Woese, C. R., R. Gutell, R. Gupta, and H. F. Noller.** 1983. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. Microbiol. Rev. **47:**621–669.
58. **Wollenzien, P. L.** 1988. Psoralens as probes of nucleic acid structure and function, p. 51–66. *In* F. P. Gasparro (ed.), Psoralen DNA photobiology, vol. 2. CRC Press, Boca Raton, Fla.
59. **Youvan, D. C., and J. E. Hearst.** 1982. Sequencing psoralen photochemically reactive sites in *Escherichia coli* 16S rRNA. Anal. Biochem. **119:**86–89.
60. **Zhuang, Y., and A. M. Weiner.** 1986. A compensatory base change in U1 sn RNA suppresses a 5′ splice site mutation. Cell **46:**827–835.
61. **Zuker, M., J. A. Jaeger, and D. H. Turner.** 1991. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenic comparison. Nucleic Acids Res. **19:**2707–2714.
62. **Zuker, M., and P. Stiegler.** 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. **9:**133–148.