Ribonucleoprotein Complexes: Biochemical and Computational Structure Analysis of RNA and Protein Constituents

Celia W.G. van Gelder

Ribonucleoprotein Complexes: Biochemical and Computational Structure Analysis of RNA and Protein Constituents

Ribonucleoprotein Complexes: Biochemical and Computational Structure Analysis of RNA and Protein Constituents

een wetenschappelijke proeve
op het gebied van de Natuurwetenschappen

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op
dinsdag 16 januari 1996,
des namiddags te 3.30 uur precies

door

Celia Wilhelmina Geertruida van Gelder

geboren op 26 juni 1965
te Renkum

Promotor: Prof. Dr. W.J. van Venrooij

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Gelder, Celia Wilhelmina Geertruida van

Ribonucleoprotein complexes: biochemical and computational
structure analysis of RNA and protein constituents /
Celia Wilhelmina Geertruida van Gelder. - [S.l. : s.n.]
Thesis Katholieke Universiteit Nijmegen. - With ref. - With
summary in Dutch.
ISBN: 90-9008866-0
Subject headings: ribonucleoprotein complexes / RNA ; structure
analysis / protein ; structure analysis

Dit proefschrift is gedrukt door Johan "De Hommel" Wiegers.

The investigations described in this thesis were carried out at the Department of Biochemistry, Faculty of Science, University of Nijmegen, The Netherlands.

"Il y a plus d'une façon de regarder les choses"

Voor mijn ouders

Voor Ronald

CONTENTS

DANKWOORD

Als je als "computerchemicus" op een biochemisch lab komt werken, gaat er een hele nieuwe wereld voor je open. "Let maar op, wij maken wel een biochemicus van je", werd er gezegd. En nu dit proefschrift voltooid is, wil ik op deze plek iedereen bedanken die mij heeft geholpen die onbesuisde uitspraak waar te maken.

Alleerst wil ik mijn promotor Walther van Venrooij bedanken, die met mij in het diepe gesprongen is, en mij altijd het vertrouwen heeft gegeven dat we samen de overkant zouden halen. José Thijssen en Eric Jansen hebben mij geholpen met mijn eerste schreden aan de labtafel, en samen met José heb ik heel wat uurtjes boven de lichtbak gezeten om onbegrijpelijke bandjes te interpreteren.

Ger Pruijn, Wilbert Boelens en Ron Peek wil ik bedanken voor hun meedenken tijdens verschillende stadia van het promotieonderzoek. Erik Klaassen en Rob Schouten hebben een bijdrage geleverd aan respektievelijk Hoofdstuk 3 en het homologie-modeling onderzoek van LDH, dat niet in dit proefschrift vermeld staat. Sander Teunissen wil ik bedanken voor zijn bijdragen aan Hoofdstuk 4 en 5, en voor de vele, vele discussies met name het afgelopen jaar. De VRT is een gezellige en inspirerende werkgroep, waar ik met veel plezier heb gewerkt, en alhoewel ik hier niet iedereen kan noemen, hebben alle VRT-leden aan deze sfeer bijgedragen.

Els van Genne wil ik natuurlijk bedanken voor allerlei praktische ondersteuning, maar met name voor de oprechte belangstelling in het wel en wee op zowel het persoonlijke als het "promotie" vlak. Hanneke van Boxtel was mijn carpoolpartner, en op de ritjes van 's-Hertogenbosch naar Nijmegen v.v. deelden wij veel meer dan alleen de brandstofkosten.

Van het CAOS/CAMM Center hebben Jan Noordik, Frank Leusen en Jack Leunissen, met hun grote kennis van CAMM en CAMMSA veel bijgedragen aan met name Hoofdstuk 2. De andere leden van het CAOS/CAMM Center wil ik bedanken voor het feit dat ik er als ex-collega altijd heel welkom was.

I would like to thank Dr. Iain Mattaj and Dr. Sam Gunderson (EMBL) for their collaborations concerning the U1A mRNA research, Dr. Alain Krol (IBMC, Strasbourg) for his advice concerning the chemical modification experiments, and Prof. Eric Westhof (IBMC, Strasbourg) for his collaboration on the RNA modeling studies.

Johan ("de Hommel") en Wil Wiegers hebben gezorgd voor de fantastische verschijningsvorm van dit proefschrift, dat dus ook met recht een familiestuk te noemen is!

Mijn vader en moeder wil ik bedanken omdat ze me hebben aangemoedigd om door te studeren en me al die jaren hebben gesteund in iets wat ik natuurlijk wel heel graag wilde, maar wat voor hen slechts abracadabra was.

Ronald, zonder jou had dit proefschrift er vast heel anders uitgezien. In de letterlijke zin natuurlijk omdat je me erg veel geholpen hebt met figuren maken, lay-outen, etc. Maar veel belangrijker zijn al die keren dat je me moed insprak en me leerde te relativeren als ik het met mijn bekende zwartgalligheid weer eens niet zag zitten. Dank je wel, schat!

CHAPTER 1

General Introduction

List of abbreviations

3'UTR	3' untranslated region
Ala	Alanine
bp	base pairs
CD	Circular Dichroism
CMCT	1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide methane-p-toluene sulfonate
CStF	Cleavage Stimulation Factor
DEPC	Diethylpyrocarbonate
DMS	Dimethylsulfate
ENU	Ethylnitrosourea
Gly	Glycine
Gln	Glutamine
hnRNP	Heterogeneous Nuclear Ribonucleoprotein
MD	Molecular Dynamics
MM	Molecular Mechanics
mRNA	messenger RNA
NMR	Nuclear Magnetic Resonance Spectroscopy
Phe	Phenylalanine
PAB II	Poly(A) binding protein II
PABP	Poly(A) binding protein
RNP1	Conserved octamer sequence in RNP motif
RNP2	Conserved hexapeptide sequence in RNP motif
RNPs	Ribonucleoprotein particles
RNP motif	RNA binding domain, found in many RNA binding proteins
RNP80 motif	alternative name for RNP motif
RRE	Rev Response Element
rRNA	Ribosomal RNA
RRM	RNA Recognition Motif; alternative name for RNP motif
RT	Reverse Transcriptase
SF2/ASF	Splicing factor 2/Alternative splicing factor
SLE	Systemic lupus erythematosus
snRNA	small nuclear RNA
ssDNA	single-stranded DNA
ssRNA	single-stranded RNA
snRNP	small nuclear ribonucleoprotein particle

TAR RNA	Transactivation response element
Tyr	Tyrosine
U2AF	U2 auxilliary factor

General Introduction

INTRODUCTION

The control of eukaryotic gene expression involves several steps in which specific sequences in pre-mRNA transcripts, as well as in small RNA molecules, are recognized by RNA-binding proteins, in this way forming ribonucleoprotein complexes (RNPs). These RNA-binding proteins mediate interactions in transcription, pre-mRNA processing (capping, splicing and 3'-end formation), regulation of translation and the stability of mRNA (Burd and Dreyfuss, 1994). Furthermore, these RNP complexes are common targets for autoimmune responses, especially in individuals with systemic lupus erythematosus (SLE) (Van Venrooij and Maini, 1994).

In this thesis the structural features of protein and RNA components of two different RNA-protein complexes are described. The first is the complex between the U1A protein and its own mRNA. The second one is in fact a group of RNA-protein complexes, namely the cytoplasmic Ro RNP particles. The results are based on both experimental and computational approaches (modeling) and mainly focus on the structure of the RNA components of these RNA-protein complexes. Some results concerning structural features of the protein components are described as well. In this chapter, an introduction on RNA structure and its determination, and on RNA-binding proteins is given.

RNA STRUCTURE

Since the three-dimensional structure formed by RNA molecules is crucial to their biological function, knowledge of RNA structure is essential. Folded RNA molecules are stabilized by a variety of interactions, the most prevalent of which are stacking and hydrogen bonding between bases (Saenger, 1984). An RNA chain can fold back upon itself to form hydrogen-bonds between bases. Most commonly Watson-Crick base pairs (bp) between A-U and G-C, which involve two and three hydrogen bonds, respectively, are formed. G-U pairs, containing two hydrogen bonds, also occur in RNA, and are approximately as stable as A-U pairs (Chastain and Tinoco, 1991). The interactions found in a three-dimensional RNA structure can be divided in secondary and tertiary interactions.

Secondary structure elements

Secondary interactions mostly involve duplex and loop regions and can be divided into different types (Chastain and Tinoco, 1991), depicted in Figure 1.

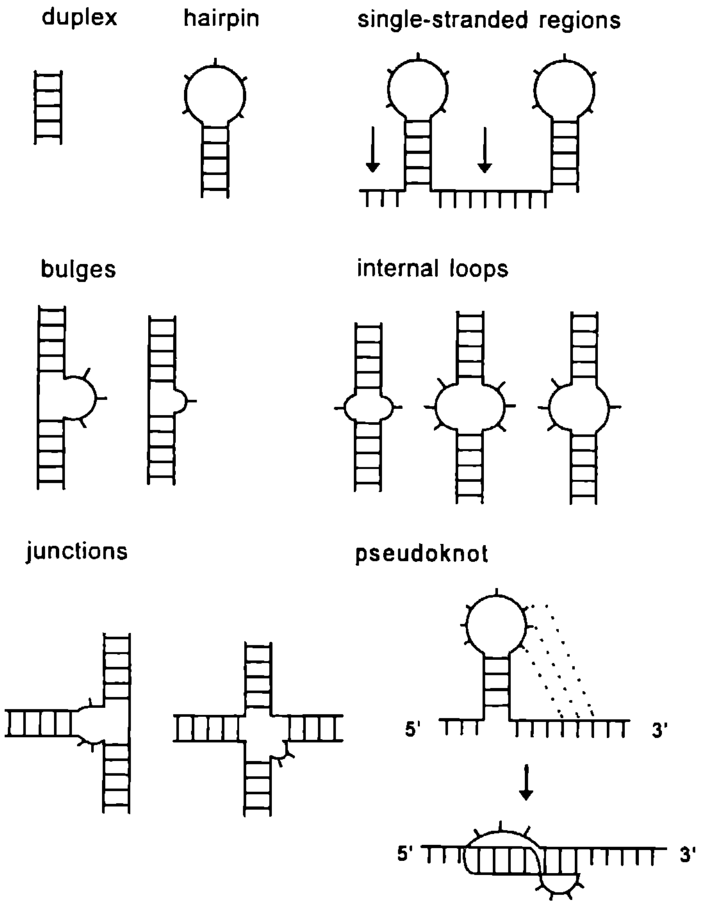


Figure 1. Secondary structure elements occurring in RNA. Also included is the definition of an RNA pseudoknot.

Helix/duplex. Uninterrupted base pairs in RNA can form a right-handed double helix. This helix has A-form geometry as opposed to the B-form of DNA duplexes. There are 11 base pairs per turn, the minor groove is wide and shallow while the major groove is narrow and deep. The sugars have the 3'-endo conformation and the base pairs are tilted with respect to the helix axis and displaced from it by about 4 Å (Saenger, 1984).

Single-stranded regions. Unpaired nucleotides form single-stranded regions and in the absence of secondary and tertiary interactions to constrain them they are assumed to be roughly ordered by base stacking in a helical geometry (Chastain and Tinoco, 1991).

Hairpins/stemloops. Hairpins are the most predominant elements of RNA secondary structure (Varani, 1995). A hairpin consists of a duplex bridged by a loop of unpaired nucleotides. The smallest loop possible is thought to be three nucleotides; DNA and RNA loops containing 4 or 5 nucleotides are most stable (Chastain and Tinoco, 1991; Hilbers *et al.*, 1994; Varani, 1995). In *E.coli* 16S rRNA 50 % of all loops contain 4 unpaired bases, and about 70% of these tetraloops contain the loop sequences GNRA, UNCG or CUUG (N=A,C,G,U; R=A,G) (Gutell, 1993). These hairpins form unusually stable tetraloop conformations, and NMR studies of UUCG (Varani *et al.*, 1991) and GAAA (Heus and Pardi, 1991) hairpins showed that the conformation of the sugar-phosphate backbone throughout the loop is very different from the A-form geometry. Hairpin loops can be actively involved in the tertiary structure (as is seen in tRNA), they often are important sites for specific RNA-protein interactions (see below) and they can be nucleation sites for RNA folding (Varani, 1995).

Bulge loops or bulges. Bulges are defined as unpaired nucleotides on one strand of a double-stranded region. Bulged nucleotides can be either looped out or stacked into the helix, creating a bend in the double helix. They affect the structure of the surrounding duplex for several base pairs and can open the major groove. NMR data of single-base bulges in DNA and RNA have shown that purines tend to stack between adjacent pairs, while pyrimidines are frequently excluded from the helix (Tang and Draper, 1990; Van den Hoogen, 1988).

Internal loops or bubbles. A mismatch is formed by two opposed nucleotides that cannot form a Watson-Crick base pair (Saenger, 1984). For example, GA mismatches occur frequently in rRNA (Gutell, 1993). Internal loops can be formed when the helix is interrupted by nucleotides on both strands that are not Watson-Crick - or GU-paired. The loops can be open or can be closed by the formation of non-Watson-Crick hydrogen bonds (non-canonical base pairs) (Santalucia *et al.*, 1991). Symmetric or asymmetric loops can be formed depending on whether an equal or an unequal number of nucleotides is on opposing strands, respectively (Chastain and Tinoco, 1991). In RNA it is known that asymmetric loops destabilize a helix more than symmetric loops (Peritz *et al.*, 1991). Most naturally occurring internal loops are rich in purines, especially adenosines (Chastain and Tinoco, 1991; Peritz *et al.*, 1991).

Junctions (three-way, four-stem). Junctions, or multibranched loops, contain three or more double helical regions with a variable number of unpaired nucleotides where the helical regions meet each other. Examples include tRNA, which contains a four-way junction, and 5S RNA and the hammerhead ribozyme, both with a three-way junction. Junction regions are important because helical regions can stack coaxially at these regions to form longer helical regions, a phenomenon which contributes to the structural stability of nucleic acid tertiary folds.

Tertiary interactions

Tertiary interactions in RNA bring together nucleotides in regions which are not close to each other in the primary or secondary structure. They also govern the characteristic three-dimensional fold of an RNA molecule, as is for example seen in tRNA. Several types of tertiary interactions have been identified so far:

Loop-loop interactions. These are found in tRNA, RNase P and 16S and 23S rRNA. Tertiary base pairing is also found in RNA pseudoknots, which involve intramolecular base pairing of bases in a hairpin loop with bases outside (but adjacent to) the stem of the loop to form a second stem and loop region (reviewed in Pleij, 1994) (see Figure 1). The second stem can be stacked upon the first to form a quasi-continuous coaxial helix. Many RNA pseudoknots exist, for example the phylogenetically proven pseudoknots in group I and II introns and 16S rRNA (Jaeger *et al.*, 1993). In ribosomal RNA they have a function in translation (Gutell, 1993). In certain plant viruses the 3'UTRs of the mRNAs contain a pseudoknot, a so-called tRNA-like structure, which mimics the structure and function (it can be aminoacylated) of tRNA (Mans *et al.*, 1991).

Single-strand/helix interactions. One example is the intercalation of bases into a helix, like G57 in tRNA (between G18 and G19). Another example is a base triple which occurs when a Watson-Crick base pair has an interaction with a third nucleotide. This can occur in the major or the minor groove and can be formed by one or two hydrogen bonds or stacking. For example, tRNA contains three base triples, while in 5S rRNA one base triple is proposed (Brunel *et al.*, 1991). In the recently determined X-ray structure of a hammerhead ribozyme (Pley *et al.*, 1994) an intermolecular interaction is found between a GAAA tetraloop and a duplex from a different molecule in the asymmetric unit.

Helix/helix interactions. Helix-helix contacts can be formed between the grooves of different helices when RNA molecules fold into compact tertiary structures. The 2'OH group appears to play an important role in stabilizing helix-helix contacts, as is seen in the crystal structure of an RNA duplex (Chastain and Tinoco, 1991).

RNA STRUCTURE DETERMINATION

The first step toward predicting the three-dimensional structure of an RNA molecule is to predict its secondary structure. This secondary structure can usually be established to a large extent without considering tertiary interactions when it is assumed that the interactions between secondary structure motifs will be weaker than the interactions within these secondary structure motifs. This assumption is known to be true for tRNA (Jaeger *et al.*, 1990). Two major approaches exist for the determination of RNA secondary structure, namely comparative sequence analysis and prediction of thermodynamic stability (Jaeger *et al.*, 1993). The combination of both methods is often most useful.

Phylogenetic comparison or comparative sequence analysis

In comparative sequence analysis (reviewed in Gutell, 1993) RNA sequences with identical function in different organisms are compared. The goal is to find structural features which have been conserved during evolution and can be formed by all sequences (Fox and Woese, 1975). Most often the nucleotide sequence of conserved helices will differ but changes in base composition at one side of a helix will be compensated for by matching changes in base composition at the opposite side of the same helix. A helix is usually considered to exist if at least two of such compensating base changes can be demonstrated (Chastain and Tinoco, 1991).

The method critically depends on the choice of the sequences, which must be sufficiently different but not so much different that homologous residues cannot be aligned with confidence. Another limitation is that phylogeny cannot provide information about conserved regions and therefore might predict fewer helices than actually exist. Phylogeny is considered a very strong method for RNA secondary structure prediction and examples of predicted RNA secondary structures include 5S rRNA (Fox and Woese, 1975), 16S rRNA (Noller and Woese, 1981), Group I introns (Michel and Westhof, 1990), U snRNAs (Guthrie, 1988) and the RNA moiety of RNase P (Pace *et al.*, 1989).

For the prediction of tertiary interactions phylogeny has also been used for example in the case of RNase P, 16S rRNA, and the base triples in tRNA (Jaeger *et al.*, 1993). In many cases, the nature of the covarying bases can indicate the geometry of the base pair, thereby providing very valuable spatial constraints (Gautheret and Cedergren, 1993). For tertiary interactions, more sophisticated covariance analysis algorithms are used which correlate positions regardless of the type of pairing between nucleotides and independent of the surrounding structure (Gutell, 1993).

Thermodynamic stability

In the second approach for determining RNA secondary structure computer algorithms are used to predict Gibbs free energies (ΔG°) for the formation of particular RNA secondary structures (Jaeger *et al.*, 1993). In contrast to phylogenetic comparison, predictions are already possible when only one sequence is available. The secondary structure of lowest free energy is thought to dominate at equilibrium. However, structures with similar free energies (suboptimal structures) may exist in dynamic equilibrium and have to be considered as well (Jaeger *et al.*, 1990).

In the prediction algorithms, each base pair and each stacking interaction contributes an empirically determined free energy to the total free energy of the RNA. Stems will contribute negative (favorable) free energy while loops and other single-stranded nucleotides are assumed to destabilize the folded molecule and thus contribute positive free energy (Turner and Sugimoto, 1988).

The most common method used to predict RNA secondary structures involves recursive (or dynamic) algorithms (reviewed in Turner and Sugimoto, 1988). Dynamic algorithms first find the lowest free energy secondary structure for all pentanucleotides, then for all hexanucleotides, and so on, until the final fragment encompasses the entire sequence. Computation of a new subfragment is performed by using the results from computations on smaller subfragments and the execution time is proportional to N^3 - N^4 (with N the number of nucleotides). The well-known Zuker program MFOLD predicts 70% of the phylogenetically deduced helices correctly and the suboptimal structures which are predicted within 10% of the lowest free energy contain roughly 90% of phylogenetically known helices (Jaeger *et al.*, 1989). The program also is able to force specific regions of the molecule to be either single-stranded or double-stranded if such information is available, for example from chemical modification data, and this greatly increases the significance of the results.

RNA folding programs make several simplifying assumptions. The first one is that the stability of a structural element in an RNA molecule is dependent only on the identity of adjacent base pairs (nearest neighbor model) (Turner and Sugimoto, 1988). The rationale behind this is that the major interactions in RNA, stacking and hydrogen bonding, are short-range. Secondly, tertiary interactions will be weaker than secondary interactions. Thus, it is assumed that the sum of the free energies of component secondary structures is a reasonable approximation of total free energy. Finally, knots are not considered in most secondary structure prediction algorithms.

The thermodynamic parameters (free energies) for secondary structure motifs are obtained by either varying the parameters until known RNA structures are predicted (Turner and Sugimoto, 1988) or by deriving them using absorbance-versus-temperature melting curves for small RNA molecules containing one or more structural motifs. Parameters are known for all combinations of adjacent base pairs involving Watson-Crick or G-U base pairs (Freier *et al.*, 1986; He *et al.*, 1991), for unpaired terminal nucleotides (dangling ends) and terminal mismatches (Freier *et al.*, 1986), for internal and hairpin loops (Jaeger *et al.*, 1993; Antao and Tinoco, 1992; Jaeger *et al.*, 1989). Junctions are implemented as well; their energy depends on the number of stems and the number of unpaired nucleotides within the junctions (Chastain and Tinoco, 1991).

The assessment of the significance of a folded structure is difficult. If folding is performed with varying parameters or with successively overlapping pieces of the RNA sequence, motifs that appear in most or all of the structures may represent the more significant local structures. Another approach for assessing significance of locally optimal secondary structures uses a Monte-Carlo method which will not be further discussed here (Abrahams *et al.*, 1990; Le *et al.*, 1988).

Several other approaches have been proposed for the prediction of RNA secondary structure. Combinatorial algorithms (Turner and Sugimoto, 1988; Jaeger *et al.*, 1993) first develop a list of all helices that can be formed, and then determine the combination of these helices that gives the lowest free energy. The advantage is that they can include knotted structures and non-nearest-neighbor interactions. However, because the number of possible helix combinations grows exponentially the algorithm is only applicable for sequences up to 200 nucleotides (Turner and Sugimoto, 1988). Other algorithms (Abrahams *et al.*, 1990; Martinez, 1984) consider the formation of secondary structure as a stepwise process, in which intermediate structures evolve into the native one by subsequent addition of stems. This approach is meant to simulate the folding process, and assumes that RNA folding proceeds from "nucleation" centers which are focal points for local RNA folding. Finally, a combination of phylogenetic and thermodynamic methods is postulated, in which optimal and suboptimal secondary structures are predicted by energy minimization and structural comparison of these secondary structures is used to find conserved structures (Konings, 1989; Le and Zuker, 1991).

The use of thermodynamic methods in predicting tertiary interactions is hampered by the fact that rules for forming tertiary interactions have not been established and that the free energies of most tertiary structures have not been determined. Despite these difficulties, one algorithm has been proposed which can predict pseudoknots (Abrahams *et al.*, 1990; Van Batenburg *et al.*, 1995).

Chemical and enzymatic reactivity

The secondary structure of an RNA molecule can be established experimentally by using a variety of chemical and enzymatic probes that distinguish between base paired and single-stranded nucleotides in the RNA (reviewed in Ehresmann *et al.*, 1987; Krol and Carbon, 1989; Knapp, 1989). Each reagent has a distinct specificity; so the larger the number of probes applied, the more accurate the derived structure of the folded RNA will be.

Ribonucleases that cleave the phosphodiester bond of a nucleotide in a single-stranded configuration are RNase A, T1, U2, T2 and nuclease S1 (see Table I). RNase V1 can be used to detect double-stranded or stacked regions. The Watson-Crick positions of the RNA bases and the N7 atoms of the purines can be modified with base specific chemicals (see Table I and Figure 2). All the Watson-Crick positions of the atoms are unreactive when involved in base pairing, except for a G-U base pair, in which N2-G is accessible. For example, dimethylsulfate (DMS) modifies the N1 atom in adenosine, the N3 atom in cytosine and the N7 atom in guanosine.

The strategy to probe the RNA secondary structure is to bring the RNA under certain conditions in which it can be subjected to limited RNase hydrolysis or chemical modification. These conditions include native conditions (presence of magnesium and monovalent cations), semi-denaturing conditions (presence of EDTA) and denaturing conditions (high temperature, presence of EDTA). Tertiary interactions are generally less stable than Watson-Crick interactions and are expected to melt under semi-denaturing conditions (Krol and Carbon, 1989). Semi-denaturing conditions also provide information about the stability of the different helical domains in an RNA molecule. The cleavages or modifications are

TABLE I. ENZYMATIC AND CHEMICAL PROBES FOR RNA STRUCTURE DETERMINATION

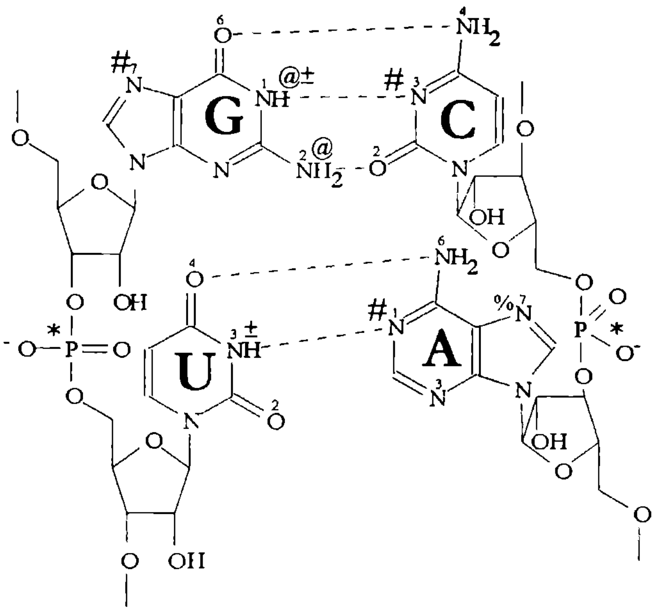
Probes	MW	Specificity	Cleavage	Phosphate	Comments
Enzymatic probes					
RNase A	13,700	unpaired U or C	Yes	3' phosphate	Some preference for CpA and UpA
RNase T1	11,000	unpaired G	Yes	3' phosphate	
RNase U2	12,490	unpaired A>G	Yes	3' phosphate	Low pH optimum
RNase CL3	16,800	unpaired C>>A>U	Yes	3' phosphate	
RNase T2	36,000	unpaired N	Yes	3' phosphate	Ap>Np
Nuclease S1	32,000	unpaired N	Yes	5' phosphate	Low pH optimum; requires Zn^{2+}
RNase V1	15,900	paired or stacked N	Yes	5' phosphate	Requires divalent cations
Chemical probes					
Dimethylsulfate (DMS)	126	N1-A	No		N7-G>N1-A>N3-C
		N3-C, N7-G	Yes		DMS also modifies cysteine
Diethylpyrocarbonate (DEPC)	174	N7-A	Yes		DEPC also modifies imidazole ring of histidine
CMCT [a]	424	N3-U, N1-G	No		N3-U>N1-G
kethoxal [b]	148	N1-G, N2-G	No		Kethoxal also modifies guanidino group of arginine
Ethylnitrosourea (ENU)	117	phosphates	Yes		Leaves ethylated phosphate on 3' oxygen
MPE-Fe(II) [c]	780	paired N	Yes		
Fe(II)EDTA	NA	exposed riboses	Yes		Hydroxyl radicals generated in presence of O_2 or H_2O_2

[a] 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide methane-p-toluene sulfonate

[b] ß-etoxy-α-ketobutyraldehyde

[c] methidiumpropyl-EDTA-Fe^{2+}

NA = Not Applicable



(#) DMS, (%) DEPC, (±) CMCT, (@) kethoxal, (*) ENU

Figure 2. Nucleotide bases and their modifying reagents. Dotted lines indicate the hydrogen bonds which are present when the bases are paired The indicated reagents can modify their target atoms only if they are accessible, that is if they are not hydrogen bonded or stacked.

introduced at a level of on the average less than one hit per molecule (single-hit conditions). Control incubations, in which the reagent is omitted, are performed in parallel. After a certain incubation time, the products of the reaction are detected using one of two possible detection methods: direct detection when end-labeled RNA is used or detection by primer extension.

End-labeled RNA can be used with enzymatic probing and with some chemical probes that induce cleavage of the RNA. The reaction products are analyzed on a denaturing polyacrylamide-urea gel. The size of the cleaved products, which is determined by running a sequence reaction on the same gel, indicates the cleavage site and in this way provides structure information. Dephosphorylated RNAs can be 5'-end-labeled using [γ-^{32}P]ATP and T4 polynucleotide kinase while for 3'-end-labeling [^{32}P]pCp and T4 RNA ligase are used (Ehresmann *et al.*, 1987). An advantage of direct detection is that only picomole amounts of RNA are needed. Disadvantages are that it can only be applied to small RNA molecules (< 200 nucleotides) due to the resolution of a sequencing gel, and that chemical probes which do not induce cleavage cannot be used.

In primer extension analysis, the unlabeled RNA is first subjected to enzymatic orchemical attack. After stopping the reaction, the RNA is hybridized with a 5'-end-labeled oligodeoxyribonucleotide complementary to a chosen sequence in the RNA. Using reverse transcriptase (RT) the primer is elongated until a modified nucleotide causes the RT enzyme to stop at the nucleotide immediately 3' to the modification. The reverse transcriptase products are then analyzed on a denaturing gel. This method is very useful for longer RNA molecules, because the start of the RT reaction, i.e. the complementary sequence of the primer, can be varied. A disadvantage of the method is the fact that no information about the very 3'- end of the RNA molecule can be obtained. Other disadvantages are that pauses of RT (RT-stops) are found which reflect spontaneous pyrimidine-purine breaks (Krol and Carbon, 1989; Kwakman *et al.*, 1990) and the tendency of RT to pause at particular structural elements in the RNA (Kwakman *et al.*, 1990).

Another means to study RNA structure and RNA-protein interactions is by using chemical nucleases (Huber, 1993), i.e. metal complexes that cleave nucleic acids with little or no dependency on the identity of the attached base. For example, Fe(II)-EDTA is a versatile probe of RNA tertiary structure (Latham and Cech, 1989). In solution, Fe(II)-EDTA complexes generate hydroxyl radicals in the presence of hydrogen peroxide or molecular oxygen. Hydroxyl radicals attack solvent-exposed riboses to cause strand scission of the RNA and in this way discriminate between solvent-accessible and solvent-inaccessible regions. The small size of the hydroxyl radical, and its uniform reactivity make it an excellent probe.

Information about tertiary structure is obtained by use of crosslinking to reveal the proximity of parts of the RNA widely separated in the sequence. Although crosslinking results cannot be directly interpreted in terms of secondary or tertiary interactions, they do provide distance constraints of great value for RNA modeling. The localization of the crosslink can be identified by partial hydrolysis of the RNA and identification of the reacting nucleotides. UV light can induce nucleic acid - nucleic acid photocrosslinks by forming cyclobutane bridges between bases that are in direct contact (Hubbard and Hearst, 1991a). Psoralen can intercalate into helical regions of DNA and RNA and, upon irridation, can covalently crosslink pyrimidines across the helix (Jaeger *et al.*, 1993). Usually psoralen crosslinks are found in helical structures but other base stacking geometries can also be cross-linked and may point to tertiary interactions (Hubbard and Hearst, 1991a). Another RNA crosslinking reagent is *bis*-(2-chloroethyl)-methylamine ("nitrogen mustard") (Hubbard and Hearst, 1991b).

Optical Spectroscopy

Absorbance-versus-melting curves can be used to measure the melting temperature Tm and therefore the stability of RNA molecules (Turner and Sugimoto, 1988).

Circular dichroism (CD) is the difference in extinction for right and left circularly polarized light (Jaeger *et al.*, 1993). For nucleic acids the CD spectrum is mainly dependent on the sequence and stacking geometry of the bases and is often used as a qualitative measure of conformation (Jaeger *et al.*, 1993).

Information about the arrangement of secondary structure elements in three dimensions can also be obtained from fluorescence energy transfer experiments (Chastain and Tinoco, 1991). It has been used for example to study the conformation of a four-way junction in DNA (Chastain and Tinoco, 1991). When aromatic amino acids such as tryptophan, tyrosine and phenylalanine are involved in RNA-binding, it has been possible to detect binding by measuring the reduction in fluorescence (quenching) (Keene and Query, 1991).

X-ray crystallography

The experimental method providing the highest resolution in structure analysis is X-ray crystallography, which, however, suffers from the requirement of large amounts of highly purified material (milligrams of RNA) (Jaeger *et al.*, 1993). Unfortunately, only a few RNAs and RNA-protein complexes have yielded crystals able to diffract at high resolution. X-ray structures for several tRNAs and for two RNA duplexes have been published. Recently, also the structure of the hammerhead ribozyme was determined (Pley *et al.*, 1994). In case of RNA-protein complexes, several complexes of tRNA and its cognate synthetase have been described, as well as the complex of the RNP motif of U1A stemloop II of U1 snRNA.

NMR

Nuclear magnetic resonance (NMR) spectroscopy provides a means for determining the three-dimensional structure and conformational properties of nucleic acids in solution (reviewed in Van de Ven and Hilbers, 1988; Wijmenga *et al.*, 1993). NMR techniques measure distances with through-space interactions (nuclear Overhauser effect or NOE) and dihedral angles using through-bond interactions (J-coupling). NMR experiments are currently limited to oligoribonucleotides of about 40-50 nucleotides (Varani and Tinoco, 1991; Chastain and Tinoco, 1991), but the use of multi-dimensional heteronuclear NMR will allow NMR studies on larger RNA molecules (Chastain and Tinoco, 1991; Jaeger *et al.*, 1993). In NMR experiments, millimolar amounts of very pure RNA are needed, but these quantities can now be synthesized efficiently by either enzymatic or chemical methods (Varani and Tinoco, 1991).

The structure of several RNA oligonucleotides containing sequences from functional or structural domains of larger RNAs have been determined by NMR. Two very stable hairpins with tetraloops UUCG (Varani *et al.*, 1991) and GAAA (Heus and Pardi, 1991) have been determined. Structures of helix I (White *et al.*, 1992) and loop E (Wimberly *et al.*, 1993) of 5S rRNA, of a pseudoknot and of the bulged TAR RNA have been determined at medium-resolution (Jaeger *et al.*, 1993).

RNA-BINDING PROTEINS: THE RNP MOTIF

The most widely found and best-characterized RNA binding motif is called the RNP motif (Burd and Dreyfuss, 1994; Birney *et al.*, 1993) (alternative names: RRM (Keene and Query, 1991) and RNP80 motif (Scherly *et al.*, 1989)). The RNP motif is a domain of about 90 amino acids present in one or more copies in proteins that bind, for example, pre-mRNA, mRNA, pre-ribosomal RNA or snRNA. The RNP family of proteins functions at several levels in RNA metabolism, including pre-mRNA transcription, splicing, and possibly, stability and transport. Some members are involved in tissue-specific and in developmentally regulated gene expression.

Animal, plant, fungal and bacterial cells contain RNP motif proteins in nearly all organelles in which RNA is present, suggesting that it is an ancient protein structure with an important function (Burd and Dreyfuss, 1994). Despite the strong homology between them they contain unique properties of recognition that allow them to distinguish between RNAs of diverse sequence and secondary structure. Table II gives an overview of some RNP proteins and their substrates. Some RNP proteins recognize single-stranded RNA (ssRNA) (for example hnRNP A1 and La) while others recognize RNA secondary structure elements (for example U1A and Ro60). Furthermore, members of this family span a spectrum of binding affinities (Kenan *et al.*, 1991; Burd and Dreyfuss, 1994) ranging from high affinity (Kd= 10^{11} - 10^{-8} M in U2AF and U1A) to low affinity (Kd=10^{-7} - 10^{-6} M in the major hnRNP proteins and UP1).

TABLE II. EXAMPLES OF RNP PROTEINS

Protein	# RNP domains	Substrate and Comments	Reference
hnRNP proteins			
hnRNP A1/UP1	2	ssRNA and ssDNA, sequences at 3' end intron, Gly-rich C-terminus	(1,2)
hnRNP C1/C2	1	strong preference for poly(U), acidic C-terminus	(1,3,4)
hnRNP E/UP2	1	may prefer homopolymeric RNA	(2)
snRNP proteins			
U1-70K	1	U1 RNA stemloop 1, charged C-terminus (RS, RE and RD repeats)	(1)
U1-A	2	U1 RNA stemloop 2, U1A mRNA	(5,6)
U2-B''	2	U2 RNA stemloop 4, may bridge to unknown RNA	(1)
Splicing factors			
U2AF^{65}	3	branchpoint-polypyrimidine tract at 3' end of intron, contains SR domain	(7)
SF2/ASF	2	splicing factor, also alternative splicing factor, can switch 5' splice site selection, contains SR domain	(7,8)
PTBP	3	polypyrimidine tract-binding protein, poorly conserved RNP1 and RNP2 elements	(2)
sex-lethal (sxl)	2	transformer (tra) pre-mRNA, near U2AF binding site	(7)
transformer-2 (tra-2)	1	affects alternative splicing of double-sex transcript, binds six 13-nt repeats, contains SR domain	(2,9)
Translation			
PABP	4	3'-poly(A) tails of processed mRNA, Pro-rich C-terminus, translational control	(2,4)
eIF 4B	1	binds near 5'-cap of mRNA, role in translation	(2)
Other			
La	1	3' oligo(U) of RNA polymerase III transcripts, role in transcription termination	(10)
Ro60	1	5' and 3' base-paired termini of hY RNAs, unknown function	(10)
nucleolin	4	binding sequence unknown, pre-rRNA processing, contains RGGF repeats	(1,2)
CStF	1	crosslinks to U-rich sequence downsream of polyA signal, necessary for cleavage during polyadenylation	(11,12)
PAB II	1	binds to growing poly(A) tail in polyadenylation, stimulates PAP, Pro and Gly-rich C-terminus	(11,13)
chloroplast proteins	2	bind ssDNA, function unknown	(2,9)
elav	3	binding site unknown, neuron-specific, needed for visual system development	(2)
Atypical			
E coli rho	1	ssDNA and UTRs of mRNA (cytosine-rich), transcription termination, contains ATP binding domain	(1,2,14)

1 Keene,J D and Query,C C (1991) Prog Nuc Acid Res Mol Biol , **41**, 179 202 2 Kenan,D J , Query,C C and Keene,J D (1991) TIBS, **16**, 214 220 3 Mattaj,I W (1993) Cell, **73**, 837 840 4 Swanson,M S , Nakagawa,T Y , Levan,K and Dreyfuss,G (1987) Mol Cell Biol , **7**, 1731 1739 5 Scherly,D , Boelens,W , Van Venrooij,W J , Dathan,N A , Hamm,J and Mattaj,I W (1989) EMBO J , **8**, 4163-4170 6 Boelens,W C , Jansen,E J R , Van Venrooij,W J , Stripecke,R , Mattaj,I W and Gunderson,S I (1993) Cell, **72**, 881 892 7 Burd,C G and Dreyfuss,G (1994) Science, **265**, 615-621 8 Zuo,P and Manley,J L (1994) Proc Natl Acad Sci U S A , **91**, 3363 3367 9 Fukamikobayashi,K , Tomoda,S and Go,M (1993) FEBS Lett , **335**, 289 293 10 Pruijn,G J M , Slobbe,R L and Van Venrooij,W J (1991) Nucl Acids Res , **19**, 5173 5180 11 Wahle,E and Keller,W (1992) Annu Rev Biochem , **61**, 419 440 12 Manley,J L and Proudfoot,N J (1994) Gene Develop, **8**, 259 264 13 Wahle,E , Lustig,A , Jeno,P and Maurer,P (1993) J Biol Chem , **268**, 2937-2945 14 Modrak,D and Richardson,J P (1994) Biochem , **33**, 8292-8299

The most conserved feature of the RNP motif is an octamer sequence RNP1. A less conserved hexamer sequence - RNP2 - is located approximately 30 amino acids amino terminal to RNP1. The RNP1 and RNP2 sequences contain aromatic and basic residues but there are many other conserved positions in the RNP motif that contain in particular Phe, Gly or Ala (Keene and Query, 1991), which constitute the hydrophobic core of the protein (Fukamikobayashi *et al.*, 1993).

Based on secondary structure predictions the fold of the RNP motif was predicted to be ßαßßαß (Ghetti *et al.*, 1989) (α=alpha-helix, ß=beta-sheet), and

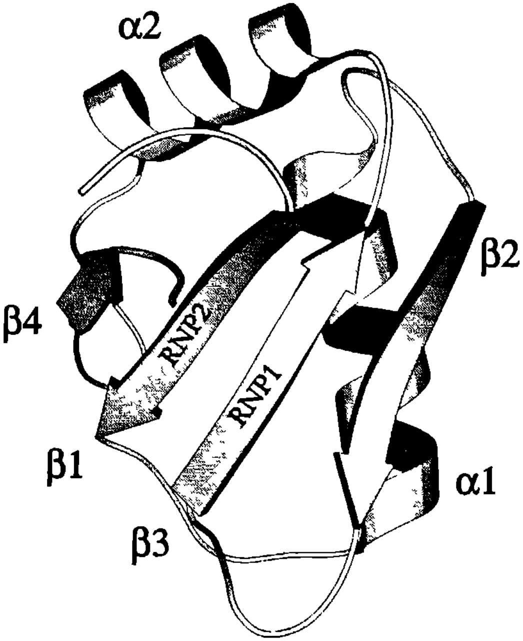


Figure 3. The ßαßßαß structure of the RNP motif. Indicated are the numbers of the ß-strands and α-helices. The RNP1 and RNP2 sequences are the so-called consensus sequences of 8 and 6 amino acids, respectively.

structural analyses by NMR spectroscopy and X-ray crystallography showed that for the U1A protein (Nagai *et al.*, 1990; Hoffman *et al.*, 1991), the hnRNP C protein (Wittekind *et al.*, 1992), and the *Drosophila* sxl protein (Lee *et al.*, 1994) this was indeed the case. The structure shows a four-stranded antiparallel ß-sheet, flanked on one side by two α-helices (see Figure 3). The conserved RNP1 and

RNP2 segments are located in the two central ß-strands, ß1 and ß3, respectively. Distinctive features are several solvent-exposed aromatics (Phe and Tyr) implicated in stacking interactions with nucleic acid bases.

RNA-binding studies indicate that three distinct structural elements contact the RNA: the ß-sheet, the loops, and the N- and C-terminal regions of the RNP motif (Burd and Dreyfuss, 1994). UV-crosslinking experiments with radioactive oligo dT and the hnRNP A1 protein showed that two Phe residues, one in RNP1 and the other in RNP2, could be crosslinked to the substrate, while in U1A protein a Tyr residue in RNP1 could be crosslinked to the second stem-loop of U1 snRNA (Stump and Hall, 1995). Protein mutagenesis experiments with U1A protein showed that U1 snRNA loop II binds to the surface of the four-stranded ß-sheet, as well as to loops at one edge of the sheet (Nagai *et al.*, 1990). In the U1A protein, RNP1 is preceded by a stretch of amino acids, which form a loop and were shown to determine the RNA-binding specificity of the domain (Scherly *et al.*, 1990). However, in the hnRNP C protein there is essentially no ß2-ß3 loop, but instead a tight turn is present (Wittekind *et al.*, 1992).

For two RNA-protein complexes detailed structural information is available. NMR experiments were performed on the complex between hnRNP C protein and rU_8 (Görlach *et al.*, 1992) while for the complex of the N-terminal RNP motif of U1A and stemloop II of U1 snRNA NMR data (Howe *et al.*, 1994; Hall, 1994) and also X-ray data (Oubridge *et al.*, 1994) are available. In both complexes, the structure of the RNP motif when bound to RNA is nearly identical to the unbound structure. Amino acids in the ß-sheet and in the N- and C-termini are involved in RNA-binding while the α-helices are largely unaffected. Bound RNA remains relatively exposed and potentially accessible for interaction with other RNA sequences or RNA-binding proteins.

When RNP protein sequences are analyzed there is a strong conservation of residues aligning with the hydrophobic core positions of U1A. It is thus believed that all RNP proteins share a common fold and a similar protein-RNA interface, and that non-conserved residues contribute additional contacts for sequence-specific RNA recognition (Kenan *et al.*, 1991). Most of the RNP proteins need sequences flanking the RNP motif for RNA-binding, suggesting that the motif alone may not contain sufficient information to function as a sequence-specific RNA-binding domain. For example, U1A needs 6 amino acids C-terminal of the RNP motif (Scherly *et al.*, 1989; Lutz-Freyermuth *et al.*, 1990) while the minimal segment of U1-70K required for RNA-binding is 111 amino acids (Query *et al.*, 1989). However, in case of the Ro60 protein, no mutations are allowed at the N- or C-terminus illustrating the diverse nature of the RNA binding domains concerning long-range intramolecular interactions that are involved (Pruijn *et al.*, 1991).

Many of the proteins with multiple RNP motifs (hnRNP A1, PABP, U2AF) require contiguous RNP motifs for wild-type RNA-binding specificity (Burd and Dreyfuss, 1994). The highly conserved organization of their RNP motifs implies that each motif has its own unique functional role. Indeed, each of the four domains of PABP has its own RNA-binding capacity and specificity (Fukamikobayashi *et al.*, 1993). It has been postulated that the presence of multiple RNP motifs in a protein may allow bridging between two different RNA molecules and it was shown that the U1A protein, which contains 2 RNP motifs, may bind simultaneously to the U1 snRNA and to the 3' UTR of mRNA sequences (Lutz and Alwine, 1994; but see also Lu and Hall, 1995). Proteins without discernible RNP motifs may contain analogous RNA-binding surfaces. Some ribosomal proteins possess a tertiary structure similar to the RNP fold (Hoffman *et al.*, 1991), suggesting either an evolutionary relationship or a convergent RNA-binding strategy. Furthermore, an RNP1 sequence which forms the central strand of a three-stranded ß-sheet was found in the bacterial nucleic acid-binding cold shock protein (Csp) (Schindelin *et al.*, 1993).

Many RNP proteins are composed of conserved RNP motifs linked to divergent auxiliary domains characterized by monotonous repetitions of distinctive amino acids (Biamonti and Riva, 1994). Such a modular structure can account for a multiplicity of interactions and it is intriguing that they are often situated at the extremities of their respective proteins. Several auxiliary domains have been identified (see also Table II). A glycine-rich domain is found (e.g., in basic hnRNP proteins and in nucleolin), which contains closely spaced RGG repeats (R=Arginine, G=Glycine), interspersed with other, often basic or aromatic amino acids. Another auxiliary domain identified is the SR domain found in splicing factors (SF2, SC35, U2AF), and in the splicing regulators *tra* and *tra-2* of *Drosophila* (Keene and Query, 1991). Auxiliary domains may be important functional constituents of the RNP proteins since various functions have been ascribed to them, including non-specific RNA-binding, annealing activity, interaction with other proteins and determinants of intracellular localization (Biamonti and Riva, 1994).

RNA-PROTEIN INTERACTIONS

General considerations

Structural, biochemical and molecular-genetic studies have established two important determinants of sequence specificity in protein-nucleic acid interactions. The first one is direct hydrogen bonding and van der Waals interactions of protein side chain and main chain atoms with nucleic acid bases. The second source of sequence specificity is provided by the sequence-dependent bendability of nucleic acids. Binding may induce conformational changes in both proteins and nucleic acids. For example, a significant distortion of the tRNA structure is observed in the X-ray structure of *E. coli* glutaminyl-tRNA synthetase complexed with tRNA(Gln) (Rould *et al.*, 1989).

In B-DNA the major groove is wide enough to accommodate an α-helix and antiparallel ß-strands but the major groove of regular A-form RNA is too narrow to allow insertion of protein secondary structure elements (Steitz, 1990). One might therefore expect proteins to discriminate between RNA sequences via interactions in the minor groove. In the complex of tRNA synthetase with its tRNA two sequence-specific contacts in the minor groove of tRNA were found (Rould *et al.*, 1989). However, there are fewer hydrogen bonding possibilities presented in the RNA minor groove (as compared with the major groove) that allow discrimination between the two base pairs and their two orientations. Fortunately, the major groove in RNA mostly is accessible in the neighbourhood of bulges, loops and non-Watson-Crick base pairs, which allows many opportunities for specific recognition. In fact, most of the protein binding sites characterized in RNA are loop regions: hairpins, bulges and internal loops, many of which undergo (gross) conformational changes upon protein binding. Hairpins form binding sites of several snRNP proteins to their cognate snRNAs. A purine bulge was shown to be involved in the binding of bacteriophage R17 coat protein to its RNA (Witherell *et al.*, 1990). HIV tat protein binds specifically to a 3-nucleotide bulge in the TAR RNA stem-loop (Harper and Logsdon, 1991). Internal loops form the binding sites of U1A on U1A mRNA (Van Gelder *et al.*, 1993) and of Rev, a regulatory RNA-binding protein that facilitates the export of unspliced HIV pre-mRNAs, on the Rev Response Element (RRE) (Burd and Dreyfuss, 1994).

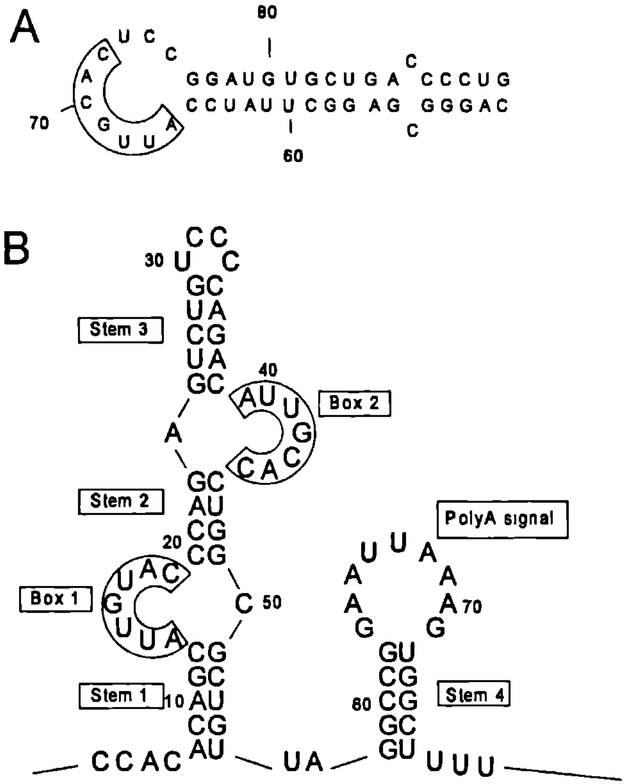


Figure 4. (A) The second stemloop of human U1 snRNA. The boxed sequence is important for U1A protein binding. (B) The secondary structure of the conserved region of the 3' UTR of the human U1A mRNA The boxed regions, which are essential for U1A protein binding, show similarity to the single-stranded U1 snRNA sequence in stem-loop II (Data taken from Van Gelder *et al.*, 1993).

The U1A - U1A mRNA complex

The removal of introns from the pre-messenger RNA, i.e. RNA splicing, is an important process in which several small ribonucleoprotein particles (snRNPs) participate (Sharp, 1994). One of them, U1 snRNP, interacts with the pre-mRNA by a mechanism that includes base pairing between the 5' end of U1 snRNA and the 5' splice site. U1 snRNP contains at least eight common proteins (B', B, D1, D2, D3, E, F and G), which also occur in other U snRNPs, as well as three U1 specific proteins named U1-70K, U1C and U1A (Lührmann *et al.*, 1990). The U1A protein binds directly to the second stemloop of U1 snRNA (Scherly *et al.*, 1989; Lutz-Freyermuth *et al.*, 1990), but its function in splicing is unknown yet. Roles for the U1A protein (Lutz and Alwine, 1994) and for the U1 snRNP (Wassarman and Steitz, 1993) have been suggested in the coupling of splicing and polyadenylation and in the coupling of polyadenylation and translation (Proudfoot, 1994).

The U1A protein contains two RNP motifs, of which the N-terminal copy is responsible for binding to U1 snRNA (Scherly *et al.*, 1989; Lutz-Freyermuth *et al.*, 1990). The structure of the RNP motif (Nagai *et al.*, 1990; Hoffman *et al.*, 1991) and of its complex with U1 snRNA is known (Oubridge *et al.*, 1994; Howe *et al.*, 1994; Hall, 1994) and has been discussed above. The loop of the second hairpin of human U1 snRNA contains 10 nucleotides (see Figure 4A). It has been shown that the first seven of them, which are highly conserved between species, are critically important for U1A protein binding, although the structural context of this sequence affects binding affinity (Scherly *et al.*, 1989; Scherly *et al.*, 1990; Tsai *et al.*, 1991).

In the 3' UTR of vertebrate U1A pre-mRNA there is a conserved region (Boelens *et al.*, 1993) which contains two stretches of seven nucleotides (called Boxes 1 and 2) similar to those of the second stemloop of U1 snRNA. These Box sequences are located close to the polyadenylation signal (see Figure 4B). It has been demonstrated that two U1A proteins can bind to these Box regions (Boelens *et al.*, 1993; Van Gelder *et al.*, 1993) and *in vitro* and *in vivo* experiments showed that excess U1A protein specifically inhibits polyadenylation of its own pre-mRNA (Boelens *et al.*, 1993). The mechanism of this regulation involving pre-mRNA binding and inhibition of polyadenylation has been further elucidated by *in vitro* studies. The inhibition of polyadenylation was shown to depend on a specific interaction of U1A protein with mammalian poly(A) polymerase in which the C-termini of both proteins might be involved (Gunderson *et al.*, 1994).

The Ro RNPs

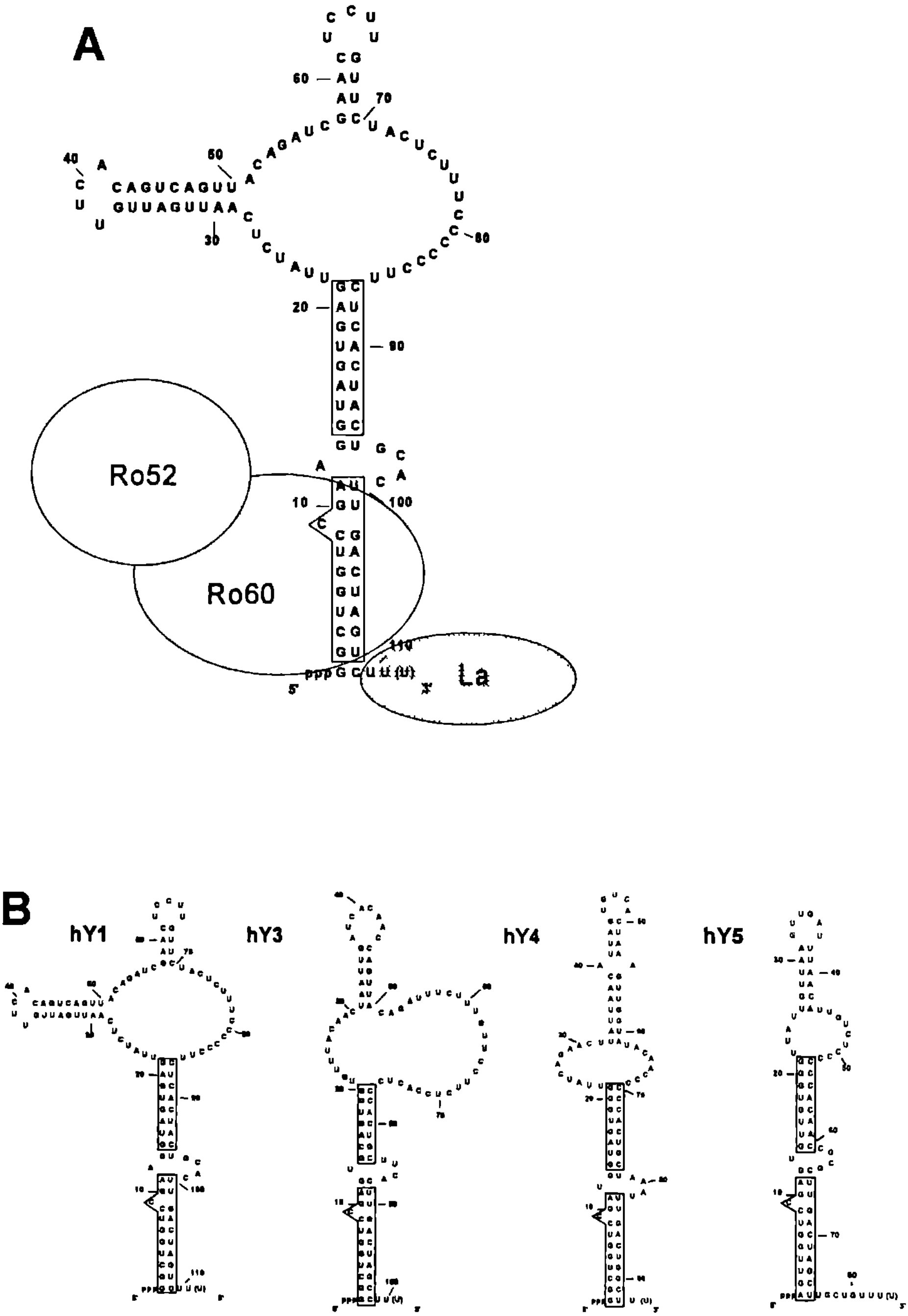
The Y RNAs (or Ro RNAs) are small cytoplasmic RNAs which are components of the Ro (SS-A) ribonucleoprotein complexes in eukaryotes (for a review see Van Venrooij *et al.*, 1993). The Ro RNPs are recognized frequently by antibodies present in sera of patients with autoimmune diseases like Sjögren's syndrome or SLE. Despite their relative abundance (~ 1-5 x 10^5 copies/cell) and evolutionary conservation no function has as yet been ascribed to these complexes. Several functions in processes such as mRNA stability, mRNA localization or translation have been suggested (reviewed in Pruijn *et al.*, 1990; Van Venrooij *et al.*, 1993).

The Ro RNPs consist of one Y RNA molecule and at least three proteins, Ro60, Ro52 and La (see Figure 5A). Within a cell, distinct subpopulations of the Ro RNPs with characteristic physicochemical properties can be distinguished and differences between cells within a species have also been observed (Pruijn *et al.*, 1990).

In human cells four Y RNAs have been identified, called hY1, hY3, hY4 and hY5 RNA (hY2 appeared to be a degradation product of hY1), ranging in length from 84 to 112 nucleotides, while in other species two to four Y RNAs were found (Pruijn *et al.*, 1993). The secondary structures of the hY RNAs show many similarities and are characterized by base pairing of the 5'- and 3'- termini (see Figure 5B). The stem structure formed in this way is the binding site for the Ro60 protein, and contains a bulged C-residue which is very important for protein binding (Pruijn *et al.*, 1991).

Ro60 is the most common Ro protein (see Figure 5C) and contains an RNP motif. The human protein also contains a zinc finger structure, but this motif is not conserved in the *Xenopus* Ro60 protein. Deletion mutagenesis showed that in both Ro60 and La, the RNP motif alone is not sufficient for the association with hY RNAs (Pruijn *et al.*, 1991), but that substantial parts of the proteins flanking the RNP motif are needed as well.

The La (or SS-B) protein is a 47 kDa ubiquitous phosphoprotein which functions in RNA polymerase III transcription termination and is localized predominantly in the nucleus (Hendrick *et al.*, 1981). It is (transiently) associated with RNA polymerase III transcripts, including the Y RNAs, adenovirus VA RNAs, Epstein-Barr virus EBER RNAs, and precursor forms of tRNA and 5S rRNA. The common sequence motif present in these RNAs is the 3'-oligouridine stretch and this is also the site of interaction with the La protein (Stefano, 1984; Pruijn *et al.*, 1991). The interaction of La with most of the RNA polymerase III products is lost upon maturation of the transcripts. However, mature Y RNAs still contain a complete La binding site and a stable association with La has been demonstrated (Boire and Craft, 1990). Furthermore, most, if not all, hY RNA molecules in cultured cells appear to be associated with La (Peek *et al.*, 1993). Besides a N-terminal RNP motif, a second RNP motif has recently been identified in the La protein (Birney *et al.*, 1993; see Figure 5C). Furthermore, La contains three so-called PEST regions and a conserved ATP binding site, also found in ATP-


A
Ro52
Ro60
La
pppGCUU(U)
5'
3'
10
20
30
40
50
60
70
80
90
100
110
B
hY1
hY3
hY4
hY5

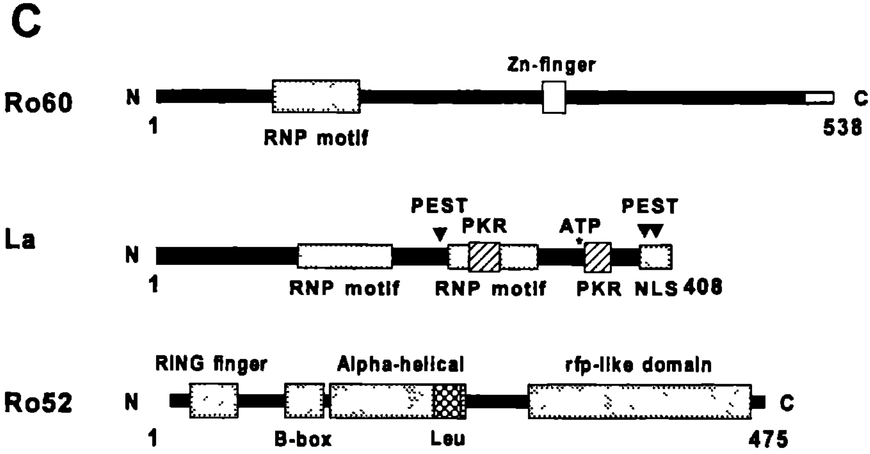


Figure 5. (A) (left) Schematic drawing of hY1 RNP Proteins La, Ro60 and Ro52 are indicated (Data taken from Van Venrooij *et al*, 1993) (B) (left) The secondary structures of the human Y RNAs. (Data taken from Van Gelder *et al.*, 1994b) (C) (above) Schematic overview of the functional domains contained in proteins La, Ro60 and Ro52 PEST region rich in Proline (P), Glutamic Acid (E), Serine (S) and Threonine (T) NLS Nuclear localization signal. PKR regions which show homology with the dsRNA dependent protein kinase PKR rfp-like region which shows homology with human transforming protein *rfp*. B-box Cys/His rich domain Leu Leucine zipper. (Modified from Van Venrooij *et al.*, 1993).

dependent DNA and RNA helicases (reviewed by Van Venrooij *et al.*, 1993). In addition to the 3'-oligouridine stretch, La may have some affinity for (an)other RNA structure(s) since La binding to RNAs lacking a 3'-oligouridine stretch has been observed as well (Van Venrooij *et al.*, 1993). Recently it was shown that La can also bind and unwind dsRNA substrates (Xiao *et al.*, 1994).

The Ro52 protein (52 kD) contains a zinc finger-like motif, called the RING finger (Freemont *et al.*, 1991), and a central leucine zipper domain (Chan *et al.*, 1991; Itoh *et al.*, 1991). In contrast to the well-conserved La and Ro60 proteins, Ro52 can be detected imunologically in primate cells only (Slobbe *et al.*, 1991). No direct interactions between Ro52 and the Ro RNAs could be identified, but the presence of Ro60 appears to be required for the Ro52 protein to bind to Ro RNPs, presumably via protein-protein interactions (Pruijn *et al.*, 1991; Slobbe *et al.*, 1992).

RNA MODELING

The limited number of RNA structures determined by X-ray crystallography and NMR spectroscopy compels the use of theoretical methods to obtain information on RNA conformation. The goal of these methods is to produce models consistent with all available experimental data, and although such structures are approximations, they provide valuable information for the design and interpretation of experiments.

RNA structure prediction is difficult because the flexibility of RNA is very large. There is extensive rotational freedom around seven intra- and internucleotide bonds per nucleotide and interactions between bases, phosphates, sugars, and solvent add even more complexity. Observations obtained by chemical modification, crosslinking and footprinting experiments can lead to constraints to restrict possible regions of the molecule in space. Mutational analyses can be useful in assessing the importance of specific residues and base pairs in the function of RNAs, although care must be taken in the interpretation of the results. Detailed analyses of RNA structure and function is possible by the substitution of specific functional groups in bases, sugars or phosphates. For example, involvement of phosphate oxygens can be monitored using phosphothioate analogs (Gautheret and Cedergren, 1993). One essential criterion for judging the validity of an RNA structure model is its generalization to RNAs belonging to the same class through biological evolution. All these RNA molecules should be able to form the same general fold in which insertions and deletions must be accommodated.

Several approaches of RNA modeling have been described (reviewed in Gautheret and Cedergren, 1993). All of them use interactive graphics programs, such as SYBYL or Quanta/CHARMm, in one or more stages of the building process, for example for visualizing the structure built or for energy minimization during the procedure.

Interactive modeling

In interactive modeling (reviewed by Westhof, 1993), a valid RNA secondary structure, obtained from phylogenetic and/or probing data, is replaced by computer-generated structural elements, which are often taken from known RNA structures. Interactive graphics modeling is then used to dock the subunits manually and in this way a starting conformation can be generated, that agrees with known structural features of RNA and with all available experimental data.

The docking of the substructures into the whole structure is very often open to numerous possibilities, especially when the links are single-stranded regions. Therefore, the generation of this initial structure is a crucial step that defines most of the interactions. After this building process the structure can be energy-minimized and successive cycles of loop modeling and docking of secondary elements can be tried until all available three-dimensional interactions are optimally dealt with.

This interactive modeling approach has been used to construct structural models of 16S rRNAs (Brimacombe *et al.*, 1988; Stern *et al.*, 1988), 5S rRNAs (Brunel *et al.*, 1991), tRNA (Dock-Bregeon *et al.*, 1989), U1 snRNA (Krol *et al.*, 1990), the *Tetrahymena* group I intron (Michel and Westhof, 1990), M1 RNA (the catalytic RNA subunit of ribonuclease P) (Westhof and Altman, 1994) and the hepatitis delta virus ribozyme (Tanner *et al.*, 1994).

Rules used in RNA modeling are based mostly on observations of available X-ray and NMR structures and can be summarized as follows (Gautheret and Cedergren, 1993; Malhotra *et al.*, 1994).

- Stacking and hydrogen bonding are the main determinants for RNA structure (Gautheret and Cedergren, 1993).
- Double-stranded regions are modeled as regular A-form RNA helices with the bases in the *anti* conformation and the riboses in the 3'-endo conformation (Gautheret and Cedergren, 1993). Sequence-dependent distortions of the A-helix are generally ignored during model building. Duplexes which are separated by less than 3 single-stranded nucleotides are assumed to stack colinearly (Kim and Cech, 1987).
- In building single-stranded regions, energy parameters are useful to predict stacking disruption (Gautheret and Cedergren, 1993).
- Base mismatches and internal loops are constructed by maintaining the integrity of the double helix while optimizing base pairing and stacking inside the loop. Non-Watson-Crick base pairs are allowed at the junction of two helices (Kim and Cech, 1987).
- Bulges are placed either inside or outside the helix, depending on the experimental information and on stacking energy parameters. Often, single bulged nucleotides are stacked into the helix (Benedetti and Morosetti, 1991; Kim and Cech, 1987).
- No general rules are as yet available for hairpin loop modeling. RNA loops are characterized by extensive stacking and extension of the A-form of the helix into the loop (Malhotra *et al.*, 1994). Available information concerning known

structures such as tRNA hairpin loops and RNA tetraloops may guide the modeling process.

- Multibranched loops cannot be modeled without considerable supplemental information on possible interactions between and among structural elements. If a base can stack on either of the two helices the stacking with the most favorable ΔG is chosen (Jaeger *et al.*, 1989).

Computational techniques

Molecular Mechanics and Dynamics techniques have also been used in RNA modeling. In Molecular Mechanics (MM) the potential energy of a molecule is described as a function (the force-field) of its atomic coordinates, and is the sum of the energy contribution of structural features such as bond lengths, bond angles, nonbonded interactions etc. (reviewed in Burkert and Allinger, 1982). Minimization of this function will lead to a low-energy structure but considering the numerous local energy minima of an RNA molecule, it is likely that only a local minimum is found, rather than the global energy minimum. Examples of RNA structures built via this method include histone mRNA loops (Gabb *et al.*, 1992), tetraloops (Kajava and Ruterjans, 1993) and the Rev Response Element (Le *et al.*, 1994).

In Molecular Dynamics (MD) both the potential and kinetic energy of a molecule is calculated and in this way a part of the conformational space of the molecule can be sampled and energy minima over a larger range of conformations can be identified. During MD studies the ends of helices are often constrained, to avoid disrupture of the helix (Nilsson *et al.*, 1990; Fritsch and Westhof, 1991). Examples of RNA structures built via this method are the helices of 5S RNA (Kim and Marshall, 1992) and the T4 self-splicing *nrdB* intron (Nilsson *et al.*, 1990).

There is a high computational cost for explicit consideration of solvent molecules and counterions in energy calculations. Both calculations with explicit solvent (Hausheer *et al.*, 1990) and without solvent (Kim and Marshall, 1992) have been performed. In the latter case the screening effect of counterions and solvent can be modeled implicitly in two ways. A distance (r) dependent dielectric constant (ϵ) can be used for the calculation of electrostatic interactions between atoms and examples include $\epsilon = r$ (Nilsson *et al.*, 1990), $\epsilon = 4r$ (Veal and Wilson, 1991; Brahms *et al.*, 1992) and a sigmoidal distance dependent function (Brahms *et al.*, 1992). An alternative method is to use partially neutralized phosphates because it is known experimentally that nucleic acid polymers maintain a net partial charge per phosphate of $\sim -0.2e$ (Veal and Wilson, 1991).

Two systems that position helical elements instead of atoms have been described. Malhotra and coworkers described a modified MM approach in which nucleotides are replaced by pseudoatoms (reviewed in Malhotra *et al.*, 1994). In this method a random construction mode produces widely varying conformers that are adjusted and evaluated by molecular mechanics techniques. In this way structures for 16S and 23S rRNAs and RNase P were built (Malhotra *et al.*, 1994; Harris *et al.*, 1994; Malhotra and Harvey, 1994). The second pseudoatom method treats the RNA molecule as a set of double-stranded helices linked by flexible single-strands of variable length. Tertiary distance constraints derived experimentally or by phylogeny are used to fold the molecule and distance geometry, developed primarily to solve NMR structures, is used for this purpose (Hubbard and Hearst, 1991b). Models for tRNA and 16S rRNA were built using this method (Hubbard and Hearst, 1991b; Hubbard and Hearst, 1991a).

Finally, a 'constraint satisfaction' algorithm was published, that automates the structure-building procedure (Major *et al.*, 1991; Gautheret and Cedergren, 1993; Gautheret *et al.*, 1993; Major *et al.*, 1993). A unique search procedure quickly yields a family of structures all satisfying a predetermined set of three-dimensional constraints in a given discrete space. These structures can then be refined by techniques such as energy minimization.

PROTEIN MODELING

Determination of the three-dimensional structure of a protein is a major step towards the elucidation of its biological function. Although the number of protein structures determined by X-ray and NMR methods is increasing steadily, the total number of known three-dimensional structures is still several orders of magnitude lower than the number of proteins for which the sequence is known. Therefore there is much interest in the prediction of protein structures and for this computer modeling is an essential tool able to complement experimental methods.

Molecular mechanics and dynamics simulations have many applications (reviewed in Karplus and Petsko, 1990; Van Gunsteren and Mark, 1992; Van Gunsteren *et al.*, 1994) in the study of the conformation and flexibility of proteins and in the modeling of protein structures or protein-ligand complexes. One of the most successful methods is homology modeling, in which a three-dimensional model of the target protein is constructed from its amino acid sequence and the known X-ray or NMR structure of a homologous protein (reviewed in Johnson *et al.*, 1994).

Furthermore, the MD method is used as a refinement technique in determining X-ray or NMR structures. MD calculations are also used to estimate the relative binding free energies of two related ligand molecules to an enzyme, or of an enzyme and mutant enzyme to a specific substrate (Reynolds *et al.*, 1992). This technique is referred to as the free energy perturbation method and is based on thermodynamic cycles.

OUTLINE OF THIS THESIS

The aim of the work described in this thesis was to integrate both experimental and theoretical approaches in order to gain insight in structural aspects of the RNA and protein components of two different RNA-protein complexes.

Chapter 2 describes a Molecular Dynamics approach used for the generation of complete protein coordinates from its Cα coordinates (Van Gelder *et al.*, 1994a). This study was inspired by an attempt to build two RNP proteins, U1A and La, by homology modeling using a template structure. For these proteins, only the Cα coordinates of a template structure were available in the Brookhaven Protein Databank (Bernstein *et al.*, 1977). Our study shows that extensive MD calculations are promising for capturing details of the native protein conformation. They are generally applicable in protein structure prediction when limited coordinate information is available. The resulting protein structures can be used (within limits) with confidence to study the general structure of the protein involved, or as a basis for further model building of homologous protein structures.

All available secondary structures for the hY RNAs were deduced from low-energy structure predictions (with minor adaptations in some cases). We therefore investigated the conformation of human hY1 and hY5 RNA using both chemical and enzymatic structure probing, while for hY3 and hY4 RNA some preliminary enzymatic probing was performed. The results, presented in Chapter 3, show that both for hY1 and for hY5 RNA the secondary structure largely corresponds to the structures predicted by sequence alignment and computerized energy-minimization. However, some important deviations were observed, the most important of which is a yet unidentified tertiary interaction in hY1 RNA, involving the pyrimidine-rich region (Van Gelder *et al.*, 1994b).

We have investigated the human U1A protein - U1A pre-mRNA complex and the relationship between its secondary structure and function in inhibition of polyadenylation *in vitro* (Chapter 4; Van Gelder *et al.*, 1993). The secondary structure of the conserved region of the 3'UTR of U1A mRNA was determined by a combination of theoretical predictions, phylogenetic sequence alignment, enzymatic structure probing and analyses of structure and function of mutant mRNAs. It was shown that the integrity of a large part of this structure is required for both high affinity binding to U1A protein and specific inhibition of polyadenylation *in vitro*.

After this, detailed chemical probing of the U1A mRNA was performed, as well as footprinting experiments on the U1A-U1A mRNA complex. Additionally, we propose a possible tertiary structure model for this RNA-protein complex. These results are described in Chapter 5.

In Chapter 6, a general discussion related to the work described in this thesis is presented.

ACKNOWLEDGEMENTS

I would like to thank Dr Ger Pruijn for critical reading of this chapter.

REFERENCES

Abrahams, J.P , van den Berg, M , van Batenburg, E and Pleij, C W.A Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.* 18.3035-3044, 1990.

Antao, V.P. and Tinoco, I ,Jr. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops *Nucl. Acids Res.* 20 819-824, 1992

Benedetti, G. and Morosetti, S. Three-dimensional folding of Tetrahymena thermophila rRNA IVS sequence· a proposal *J. Biomol Struct. Dyn.* 8 1045-1055, 1991

Bernstein, F C , Koetzle, T F., Williams, E J B , Meyer, E F Jr , Kennard, O , Shimanouchi, T. and Tasumi, M. The protein data bank. A computer based archival file for molecular structures *J. Mol. Biol.* 112.535-542, 1977

Biamonti, G. and Riva, S New insights into the auxiliary domains of eukaryotic RNA binding proteins. *FEBS Lett.* 340·1-8, 1994.

Birney, E , Kumar, S and Krainer, A R. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucl. Acids Res.* 21 5803-5816, 1993.

Boelens, W.C., Jansen, E J R , Van Venrooij, W J , Stripecke, R., Mattaj, I W. and Gunderson, S.I. The human U1 snRNP-specific U1A protein inhibits polyadenylation of its own pre-messenger RNA. *Cell* 72.881-892, 1993.

Boire, G. and Craft, J. Human Ro ribonucleoprotein particles: characterization of native structure and stable association with the La polypeptide. *J. Clin. Invest.* 85 1182-1190, 1990

Brahms, S., Fritsch, V , Brahms, J G and Westhof, E. Investigations on the dynamics structures of Adenine- and Thymine-containing DNA. *J. Mol Biol.* 223 455-476, 1992.

Brimacombe, R., Atmadja, J., Stiege, W and Schuler, D A detailed model of the three-dimensional structure of Escherichia coli 16S ribosomal RNA in situ in the 30S subunit *J. Mol. Biol.* 199 115-136, 1988.

Brunel, C , Romby, P., Westhof, E., Ehresmann, C. and Ehresmann, B. Three-dimensional model of Escherichia coli ribosomal 5 S RNA as deduced from structure probing in solution and computer modeling *J Mol Biol* 221 293-308, 1991

Burd, C G and Dreyfuss, G Conserved structures and diversity of functions of RNA-binding proteins *Science* 265 615-621, 1994

Burkert, U and Allinger, N L *Molecular Mechanics*, Washington American Chemical Society, 1982

Chan, E K , Hamel, J C , Buyon, J P and Tan, E M Molecular definition and sequence motifs of the 52 kD component of human SS-A/Ro autoantigen *J Clin Invest* 87 68 76, 1991

Chastain, M and Tinoco, I , Jr Structural elements in RNA *Prog Nucleic Acid Res Mol Biol* 41 131 177, 1991

Dock Bregeon, A C , Westhof, E , Giege, R and Moras, D Solution structure of a tRNA with a large variable region yeast tRNA(ser) *J Mol Biol* 206 707-722, 1989

Ehresmann, C , Baudin, F , Mougel, M , Romby, P , Ebel, J-P and Ehresmann, B Probing the structure of RNAs in solution *Nucl Acids Res* 15 9109-9129, 1987

Fox, G E and Woese, C R 5S RNA secondary structure *Nature* 256 505-507, 1975

Freemont, P S , Hanson, I M and Trowsdale, J A Novel Cysteine-Rich Sequence Motif *Cell* 64 483-484, 1991

Freier, S M , Kierzek, R , Jaeger, J A , Sugimoto, N , Caruthers, M H , Neilson, T and Turner,D H Improved free-energy parameters for predictions of RNA duplex stability *Proc Natl Acad Sci, USA* 83 9373 9377, 1986

Fritsch, V and Westhof, E 3 Center hydrogen bonds in DNA - Molecular Dynamics of Poly(dA) Poly(dT) *J Am Chem Soc* 113 8271-8277, 1991

Fukamikobayashi, K , Tomoda, S and Go, M Evolutionary clustering and functional similarity of RNA binding proteins *FEBS Lett* 335 289 293, 1993

Gabb, H A , Harris, M E , Niranjan, B P , Marzluff, W F and Harvey, C Molecular Modeling to predict the structural and biological effects of mutations in a highly conserved histone mRNA loop sequence *J Biomol Struct Dyn* 9 1119-1130, 1992

Gautheret, D , Major, F and Cedergren, R Modeling the 3 dimensional structure of RNA using discrete nucleotide conformational sets *J Mol Biol* 229 1049-1064, 1993

Gautheret, D and Cedergren, R Modeling the 3 dimensional structure of RNA *FASEB J* 7 97 105, 1993

Ghetti, A , Padovani, C , Di Cesare, G and Morandi, C Secondary structure prediction for RNA binding domain in RNP proteins identifies $\beta\alpha\beta$ as the main structural motif *FEBS Lett* 257 373-376, 1989

Gorlach, M , Wittekind, M , Beckman, R A , Mueller, L and Dreyfuss, G Interaction of the RNA-binding domain of the hnRNP C proteins with RNA *EMBO J* 11 3289 3295, 1992

Gunderson, S I , Beyer, K , Martin, G , Keller, W , Boelens, W C and Mattaj, I W The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A)polymerase *Cell* 76 531-541, 1994

Gutell, R R Comparative studies of RNA - Inferring higher-order structure from patterns of sequence variation *Curr Opin Struct Biol* 3 313-322, 1993

Guthrie, C Spliceosomal snRNAs *Annu Rev Genet* 22 387-419, 1988

Hall, K B Interaction of RNA hairpins with the human U1A N-terminal RNA binding domain *Biochem* 33 10076-10088, 1994

Harper, J W and Logsdon, N J Refolded HIV 1 tat protein protects both bulge and loop nucleotides in TAR RNA from ribonucleolytic cleavage *Biochem* 30 8060-8066, 1991

Harris, M E , Nolan, J M , Malhotra, A , Brown, J W , Harvey, S C and Pace, N R Use of photoaffinity crosslinking and molecular modeling to analyze the global architecture of ribonuclease P RNA *EMBO J* 13 3953-3963, 1994

Hausheer, F H , Singh, U C , Palmer, T C and Saxe, J D Dynamic properties and electrostatic potential surface of neutral DNA heteropolymers *J Am Chem Soc* 112 9468-9474, 1990

He, L , Kierzek, R , Santalucia, J , Walter, A E and Turner, D H Nearest neighbor parameters for

G U mismatches - 5'GU3'/3'UG5' is destabilizing in the contexts CGUG GUGC,UGUA AUGU, and AGUU UUGU but stabilizing in GGUC CUGG *Biochem.* 30 11124-11132, 1991

Hendrick, J P , Wolin, S L , Rinke, J , Lerner, M R and Steitz, J A Ro small cytoplasmic ribonucleoproteins are a subclass of La ribonucleoproteins Further characterization of the Ro and La small ribonucleoprotein particles from uninfected mammalian cells *Mol Cell Biol.* 1 1138-1149, 1981

Heus, H A and Pardi, A Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops *Science* 253 191-194, 1991

Hilbers, C W , Heus, H A , Van Dongen, M J P , and Wijmenga, S S The hairpin elements of nucleic acid structure DNA and RNA folding In *Nucleic Acids and Molecular Biology*, 8, pp 55-104, 1994

Hoffman, D W , Query, C C , Golden, B L , White, S W and Keene, J D RNA binding domain of the A-protein component of the U1 small nuclear ribonucleoprotein analyzed by NMR spectroscopy is structurally similar to ribosomal proteins *Proc Natl. Acad Sci., USA* 88 2495-2499, 1991

Howe, P W A , Nagai, K , Neuhaus, D and Varani, G NMR studies of U1 snRNA recognition by the N-terminal RNP domain of the human U1A protein *EMBO J* 13 3873-3881, 1994

Hubbard, J M and Hearst, J E Predicting the three-dimensional folding of transfer RNA with a computer modeling protocol *Biochem.* 30 5458-5465, 1991a

Hubbard, J M and Hearst, J E Computer modeling 16S ribosomal RNA *J Mol Biol.* 221 889-907, 1991b

Huber, P W Chemical nucleases - Their use in studying RNA structure and RNA protein interactions *FASEB J* 7 1367-1375, 1993

Itoh, K , Itoh, Y and Frank, M B Protein heterogeneity in the human Ro/SSA ribonucleoproteins The 52 and 60-kD Ro/SSA autoantigens are encoded by separate genes *J Clin. Invest* 87 177-186, 1991

Jaeger, J A , Turner, D H and Zuker, M Improved predictions of secondary structures for RNA *Proc Natl Acad Sci., USA* 86 7706-7710, 1989

Jaeger, J A , Turner, D H and Zuker, M Predicting optimal and suboptimal secondary structure for RNA *Methods Enzymol* 183 281-303, 1990

Jaeger, J A , Santalucia, J and Tinoco, I Determination of RNA structure and thermodynamics *Annu. Rev Biochem* 62 255-287, 1993

Johnson, M S , Srinivasan, N , Sowdhamini, R and Blundell, T L Knowledge-based protein modeling *Crit. Rev Biochem Molec Biol.* 29 1-68, 1994

Kajava, A and Ruterjans, H Molecular modelling of the 3-D structure of RNA tetraloops with different nucleotide sequences *Nucl Acids Res* 21 4556-4562, 1993

Karplus, M and Petsko, G A Molecular dynamics simulations in biology *Nature* 347 631-639, 1990

Keene, J D and Query, C C Nuclear RNA-binding proteins *Prog Nucleic Acid Res Mol Biol.* 41 179-202, 1991

Kenan, D J , Query, C C and Keene, J D RNA recognition - Towards identifying determinants of specificity *TIBS* 16 214-220, 1991

Kim, J H and Marshall, A G Structural investigation of helices II, III, and IV of B megaterium 5S ribosomal RNA by Molecular Dynamics calculations *Biopol* 32 1263 1270, 1992

Kim, S H and Cech, T R Three-dimensional model of the active site of the self-splicing rRNA precursor of Tetrahymena *Proc Natl Acad Sci., USA* 84 8788-8792, 1987

Knapp, G Enzymatic approaches to probing of RNA secondary and tertiary structure *Methods Enzymol* 180 192-212, 1989

Konings, D A M *Pattern analysis of RNA secondary structure, Phd Thesis*, Utrecht 1989

Krol, A , Westhof, E , Bach, M , Luhrmann, R, Ebel, J P and Carbon, P Solution structure of human U1 snRNA Derivation of a possible three-dimensional model *Nucl. Acids Res* **18** 3803-3811, 1990

Krol, A and Carbon, P A guide for probing native small nuclear RNA and ribonucleoprotein structures *Methods Enzymol* **180** 212-227, 1989

Kwakman, J H , Konings, D A , Hogeweg, P , Pel, H J and Grivell, L A Structural analysis of a group II intron by chemical modifications and minimal energy calculations *J Biomol. Struct Dyn.* **8** 413-430, 1990

Latham, J A and Cech, T R Define the inside and outside of a catalytic RNA molecule *Science* **245** 276-282, 1989

Le, S Y , Chen, J H , Currey, K M and Maizel, J V A program for predicting significant RNA secondary structures *Comp Appl Biosci.* **4** 153 159, 1988

Le, S Y , Pattabiraman, N and Maizel, J V RNA tertiary structure of the HIV RRE domain II containing non Watson Crick base pairs GG and GA Molecular modeling studies *Nucl Acids Res* **22** 3966-3976, 1994

Le, S Y and Zuker, M Predicting common foldings of homologous RNAs *J Biomol. Struct. Dyn.* **8** 1027-1044, 1991

Lee, A L , Kanaar, R , Rio, D C and Wemmer, D E Resonance assignments and solution structure of the second RNA binding domain of sex lethal determined by multidimensional heteronuclear magnetic resonance *Biochem* **33** 13775 13786, 1994

Lu, J R and Hall, K B An RBD that does not bind RNA NMR secondary structure determination and biochemical properties of the C terminal RNA binding domain from the human U1A protein *J Mol Biol* **247** 739 752, 1995

Lutz, C S and Alwine, J C Direct interaction of the U1 snRNP-A protein with the upstream efficiency element of the SV40 late polyadenylation signal *Gene Dev* **8** 576-586, 1994

Lutz Freyermuth, C , Query, C C and Keene, J D Quantitative determination that one of two potential RNA-binding domains of the A protein component of the U1 small nuclear ribonucleoprotein complex binds with high affinity to stem loop II of U1 RNA *Proc Natl Acad Sci., USA* **87** 6393 6397, 1990

Luhrmann, R, Kastner, B and Bach, M Structure of spliceosomal snRNPs and their role in pre-mRNA splicing *Biochim Biophys Acta* **1087** 265 292, 1990

Major, F , Turcotte, M , Gautheret, D , Lapalme, G , Fillion, E and Cedergren, R The combination of symbolic and numerical computation for three-dimensional modeling of RNA *Science* **253** 1255-1260, 1991

Major, F , Gautheret, D and Cedergren, R Reproducing the 3-dimensional structure of a transfer RNA molecule from structural constraints *Proc Natl. Acad Sci., USA* **90** 9408 9412, 1993

Malhotra, A , Tan, R K Z and Harvey, S C Modeling large RNAs and ribonucleoprotein particles using molecular mechanics techniques *Biophys J* **66** 1777-1795, 1994

Malhotra, A and Harvey, S C A quantitative model of the Escherichia coli 16S RNA in the 30S ribosomal subunit *J Mol Biol* **240** 308-340, 1994

Mans, R M W , Pleij, C W A and Bosch, L Transfer RNA like structures - Structure, function and evolutionary significance *Eur J Biochem.* **201** 303-324, 1991

Martinez, H M An RNA folding rule *Nucl Acids Res* **12** 323-334, 1984

Michel, F and Westhof, E Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis *J Mol Biol* **216** 585 610, 1990

Nagai, K , Oubridge, C , Jessen, T H , Li, J and Evans, P R Crystal structure of the RNA binding domain of the U1 small nuclear ribonucleoprotein-A *Nature* **348** 515-520, 1990

Nilsson, L , Ahgren Stalhandske, A , Sjogren, A S , Hahne, S and Sjoberg, B M Three-dimensional model and molecular dynamics simulation of the active site of the self-splicing intervening sequence of the bacteriophage T4 nrdB messenger RNA *Biochem.* **29** 10317-10322, 1990

Noller, H F and Woese, C R Secondary structure of 16S ribosomal RNA *Science* **212** 403-411, 1981

Oubridge, C, Ito, H, Evans, P.R., Teo, C.H. and Nagai, K Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin *Nature* 372:432-438, 1994

Pace, N.R., Smith, D.K., Olsen, G.J and James, B D. Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA - a review. *Gene* 82:65-75, 1989.

Peek, R, Pruijn, G J.M., van der Kemp, A.J.W and Van Venrooij, W.J. Subcellular distribution of Ro ribonucleoprotein complexes and their constituents. *J. Cell. Sci.* 106.929-935, 1993.

Peritz, A.E., Kierzek, R., Sugimoto, N. and Turner, D.H. Thermodynamic study of internal loops in oligoribonucleotides: symmetric loops are more stable than asymmetric loops. *Biochem.* 30:6428-6436, 1991.

Pleij, C W.A. RNA pseudoknots. *Curr. Opin. Struct. Biol.* 4·337-344, 1994.

Pley, H.W., Flaherty, K.M. and McKay, D B. Three-dimensional structure of a hammerhead ribozyme. *Nature*, 372:68-74, 1994.

Proudfoot, N.J. Chasing your own poly(A) tail. *Curr. Biol.* 4:359-361, 1994.

Pruijn, G J M., Slobbe, R.L. and Van Venrooij, W J. Structure and function of La and Ro RNPs. *Mol. Biol. Rep.* 14.43-48, 1990.

Pruijn, G.J.M., Slobbe, R L. and Van Venrooij, W.J. Analysis of protein - RNA interactions within Ro ribonucleoprotein complexes. *Nucl. Acids Res.* 19 5173-5180, 1991.

Pruijn, G J.M., Wingens, P.A.E.T M, Peters, S.L.M., Thijssen, J P H. and Van Venrooij, W.J. Ro RNP associated Y RNAs are highly conserved among mammals. *Biochim. Biophys. Acta* 1216.395-401, 1993.

Query, C.C., Bentley, R.C. and Keene, J.D A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. *Cell* 57:89-101, 1989.

Reynolds, C.A., King, P M and Richards, W.G. Free energy calculations in molecular biophysics. *Mol. Phys.* 76·251-275, 1992.

Rould, M.A., Perona, J J, Soll, D. and Steitz, T.A. Structure of E coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 Å resolution. *Science* 246 1135-1142, 1989.

Saenger, W. *Principles of nucleic acid structure*, Berlin:Springer-Verlag, 1984

Santalucia, J., Kierzek, R. and Turner, D.H. Stabilities of consecutive A C, C C, G G, U C, and U.U mismatches in RNA internal loops - Evidence for stable hydrogen-bonded U U and C.C+pairs. *Biochem.* 30 8242-8251, 1991.

Scherly, D., Boelens, W., Van Venrooij, W J., Dathan, N.A., Hamm, J. and Mattaj, I.W. Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA. *EMBO J.* 8.4163-4170, 1989.

Scherly, D., Boelens, W., Dathan, N A., Van Venrooij, W.J. and Mattaj, I.W. Major determinants of the specificity of interaction between small nuclear ribonucleoproteins U1A and U2B" and their cognate RNAs. *Nature* 345:502-506, 1990.

Schindelin, H., Marahiel, M.A. and Heinemann, U. Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-Shock protein. *Nature* 364 164-168, 1993.

Sharp, P.A. Split genes and RNA splicing. *Cell* 77 805-815, 1994.

Slobbe, R.L., Pruijn, G.J.M., Damen, W G.M., van der Kemp, A J.W. and Van Venrooij, W J. The detection and occurrence of the 60 and 52 kDa Ro (SS-A) antigens and of autoantibodies against these proteins. *Clin. Exp. Immunol.* 86:99-105, 1991.

Slobbe, R L, Pluk, W., Van Venrooij, W.J. and Pruijn, G.J.M. Ro ribonucleoprotein assembly in vitro. Identification of RNA-protein and protein-protein interactions. *J. Mol. Biol.* 227:361-366, 1992.

Stefano, J.E. Purified lupus antigen La recognizes an oligouridylate stretch common to the 3'termini of RNA polymerase III transcripts. *Cell* 36:145-154, 1984.

Steitz, T.A Structural studies of protein-nucleic acid interaction: the sources of sequence-specific binding. *Quart.Rev.Biophys.* 23·205-280, 1990.

Stern, S., Weiser, B and Noller, H.F. Model for the three-dimensional folding of 16S Ribosomal

RNA. *J. Mol. Biol.* 204.447-481, 1988.

Stump, W.T. and Hall, K B. Crosslinking of an iodo-uridine-RNA hairpin to a single site on the human U1A N-terminal RNA binding domain *RNA* 1 55-63, 1995.

Tang, R. and Draper, D.E Bulge loops used to measure the helical twist of RNA in solution *Biochem.* 29·5232-5237, 1990.

Tanner, N.K., Schaff, S., Thill, G., Petit-Koskas, E., Crain-Denoyelle, A and Westhof, E. A three-dimensional model of hepatitis delta virus ribozyme based on biochemical and mutational analyses. *Curr Biol* 4 488-498, 1994.

Tsai, D E., Harper, D S. and Keene, J D U1-snRNP-A protein selects a ten nucleotide consensus sequence from a degenerate RNA pool presented in various structural contexts *Nucl. Acids Res.* 18·4931-4936, 1991.

Turner, D.H. and Sugimoto, N. RNA structure prediction. *Ann. Rev. Biophys. Biophys. Chem.* 17 167-192, 1988.

Van Batenburg, F.H D., Gultyaev, A P. and Pleij, C.W A. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.*, 174 269-280, 1995

Van den Hoogen, Y.T, Van Beuzekom, A.A., De Vroom, E , Van der Marel, G.A., Van Boom, J.H and Altona, C , Bulge-out structures in the single-stranded trimer AUA and in the duplex (CUGGUGCGG)(CCGCCCAG) A model building and NMR study. *Nucl Acids. Res,* 16 5013, 1988

Van de Ven, F J M , and Hilbers, C.W. Nucleic acids and nuclear magnetic resonance. *Eur. J. Biochem.*, 178·1-38, 1988

Van Gelder, C.W.G., Gunderson, S I , Jansen, E J.R., Boelens, W.C., Polycarpou-Schwarz, M , Mattaj, I.W. and Van Venrooij, W J A complex secondary structure in U1A pre-messenger RNA that binds two molecules of U1A protein is required for regulation of polyadenylation. *EMBO J.* 12 5191-5200, 1993

Van Gelder, C W.G., Leusen, F J J , Leunissen, J A M. and Noordik, J H. A Molecular Dynamics approach for the generation of complete protein structures from limited coordinate data *Protein-Struct. Funct. Genet.* 18 174-185, 1994a.

Van Gelder, C W G., Thijssen, J P H M , Klaassen, E.C.J , Sturchler, C., Krol, A , Van Venrooij,W J and Pruijn, G.J M. Common structural features of the Ro RNP associated hY1 and hY5 RNAs *Nucl. Acids Res* 22·2498-2506, 1994b.

Van Gunsteren, W F , Luque, F J , Timms, D. and Torda, A.E Molecular mechanics in biology: From structure to function, taking account of solvation. *Annu. Rev. Biophys. Biomol. Struct.* 23 847-863, 1994.

Van Gunsteren, W F. and Mark, A.E. On the interpretation of biochemical data by molecular dynamics computer simulation. *Eur J Biochem.* 204·947-961, 1992.

Van Venrooij, W J , Slobbe, R L. and Pruijn, G J.M. Structure and function of La and Ro RNPs. *Mol. Biol. Rep.* 18·113-119, 1993.

Van Venrooij, W J. and Maini, R N. *Manual of biological markers of Disease, Part B: Autoantigens.* Kluwer, 1994.

Varani, G., Cheong, C and Tinoco, I.,Jr. Structure of an unusually stable RNA hairpin. *Biochem.* 30 3280-3289, 1991.

Varani, G. Stable nucleic acid hairpins *Annu. Rev. Biophys Biomol. Struct.* 24 379-404, 1995

Varani, G. and Tinoco, I.,Jr RNA structure and NMR spectroscopy *Q. Rev. Biophys.* 24·479-532, 1991.

Veal, J M. and Wilson, W D. Modeling of nucleic acid complexes with cationic ligands: a specialized molecular mechanics force field and its application. *J. Biomol. Struct. Dyn.* 8 1119-1145, 1991.

Wassarman, K M. and Steitz, J.A. Association with terminal exons in pre-mRNAs: a new role for the U1 snRNP? *Genes Dev.* 7.647-659, 1993.

Westhof, E. Modelling the three-dimensional structure of ribonucleic acids. *J. Mol. Struct.*

286:203-210, 1993.

Westhof, E. and Altman, S. Three-dimensional working model of M1 RNA, the catalytic RNA subunit of ribonuclease P from Escherichia coli. *Proc. Natl. Acad. Sci., USA* 91:5133-5137, 1994.

White, S.A., Nilges, M., Huang, A., Brunger, A T and Moore, P.B. NMR analysis of helix I from the 5S RNA of Escherichia coli. *Biochem.* 31.1610-1621, 1992.

Wijmenga, S.S., Mooren, M.W., and Hilbers, C.W., NMR of nucleic acids; from spectrum to structure. In: *NMR of macromolecules. A practical approach*, 1993. Ed. Roberts, G.C K.

Wimberly, B , Varani, G and Tinoco, I. The conformation of loop E of eukaryotic 5S ribosomal RNA. *Biochem.* 32 1078-1087, 1993.

Witherell, G.W., Wu, H N. and Uhlenbeck, O.C. Cooperative binding of R17 coat protein to RNA. *Biochem.* 29:11051-11057, 1990

Wittekind, M., Gorlach, M., Friedrichs, M., Dreyfuss, G. and Mueller, L. 1H, 13C, and 15N NMR assignments and global folding pattern of the RNA-binding domain of the human hnRNP C proteins. *Biochem.* 31.6254-6265, 1992.

Xiao, Q.R., Sharp, T.V., Jeffrey, I.W., James, M C., Pruijn, G.J.M., Van Venrooij, W.J. and Clemens, M J. The La antigen inhibits the activation of the interferon-inducible protein kinase PKR by sequestering and unwinding double-stranded RNA. *Nucl. Acids Res.* 22:2512-2518, 1994.

CHAPTER 2

A Molecular Dynamics Approach for the Generation of Complete Protein Structures From Limited Coordinate Data

PROTEINS: Structure, Function and Genetics 18:174-185 (1994)

A Molecular Dynamics Approach for the Generation of Complete Protein Structures From Limited Coordinate Data

Celia W.G. van Gelder,[1] Frank J.J. Leusen,[2] Jack A.M. Leunissen,[2] and Jan H. Noordik[2]

[1]Department of Biochemistry and [2]CAOS/CAMM Center, Faculty of Science, University of Nijmegen, 6500 HB Nijmegen, The Netherlands

ABSTRACT

Generation of full protein coordinates from limited information, e.g., the Cα coordinates, is an important step in protein homology modeling and structure determination, and molecular dynamics (MD) simulations may prove to be important in this task. We describe a new method, in which the protein backbone is built quickly in a rather crude way and then refined by minimization techniques. Subsequently, the side chains are positioned using extensive MD calculations. The method is tested on two proteins, and results compared to proteins constructed using two other MD-based methods. In the first method, we supplemented an existing backbone building method with a new procedure to add side chains. The second one largely consists of available methodology. The constructed proteins are compared to the corresponding X-ray structures, which became available during this study, and they are in good agreement (backbone RMS values of 0.5-0.7 Å, and all-atom RMS values of 1.5-1.9 Å). This comparative study indicates that extensive MD simulations are able, to some extent, to generate details of the native protein structure, and may contribute to the development of a standardized methodology to predict reliably (parts of) protein structures when only partial coordinate data are available.

Key words: computer modeling, protein structure prediction, α-carbons, structure evaluation, molecular dynamics

Abbreviations CSB-MD, crude structure building followed by MD refining, CG, conjugate gradients, MD, molecular dynamics, RMS, root mean squares deviation, SD, steepest descents, SP, spare parts method, SP-MD, combined spare parts and molecular dynamics method

INTRODUCTION

Determination of the three-dimensional structure of a protein is a major step in the elucidation of its biological function. Although the number of protein structures determined by X-ray and NMR methods is increasing steadily, the total number of known three-dimensional structures is still several orders of magnitude lower than the number of proteins for which the sequence is known. For this reason, much effort is being made to develop computational methodologies for the prediction of protein conformation. These methods will lead to an accumulation of our knowledge of protein structure, and the ultimate goal would be to generate a protein structure on the basis of its sequence only.

However, the problem which has to be solved first is to predict reliably a protein structure when only limited coordinate information is available. One example of this problem is seen in homology modeling, where the known tertiary structure of a protein is used as a template to predict the structure of an homologous (and preferably functionally related) protein.[1] In most cases, the backbone coordinates are taken from the template, as are the side chains of identical amino acids in both proteins. Insertions (including loops) and deletions in the backbone, however, must be predicted, as must the side chains of nonidentical amino acids. In addition, crystallographers and NMR spectroscopists might find such predictive methodologies useful to create an approximate structure in the early stages of the structure determination.

A good test problem for methods which can extend an incomplete protein coordinate set consists of the prediction of complete protein structures from only Cα coordinates. This approach has in fact been used in many studies, including this one. In our case it was inspired by an attempt to build two proteins by homology. For both of these proteins, only the Cα coordinates of a template structure were available in the Brookhaven Protein Databank.[2]

Several approaches to generate backbone and/or side chain coordinates from Cα coordinates have been described. For the generation of backbone coordinates, one promising method, the "spare parts" (SP) method,[3-5] uses fragments from known protein structures to build a polyalanine backbone which fits the known X-ray Cα positions (within a preset RMS limit*). It is based upon the emerging idea, that protein structures contain several "supersecondary" folding motifs or domains,

*The RMS value is the root mean squares deviation in atomic positions after optimal superimposition of two structures.

which may represent independent building blocks from which complete protein structures can be constructed.[6,7] In test cases,[3,8-10] good results (backbone RMS values between 0.35 and 0.6 Å) were obtained. Correa[11] has developed a method to generate a complete protein structure from its Cα trace, which does not need a priori knowledge of protein structure. The method is solely based on the known Cα positions, the topology of the 20 amino acids, and the flat nature of the peptide bond. In this method, long MD calculations (at high temperature) are used, which could hamper its applicability. The advantages, however, are that little protein expertise is needed and that the procedure is rather straightforward. Backbones constructed by this method showed RMS values of 0.2-0.5 Å. Recently, Bruccoleri[12] described a directed conformational search to generate backbone coordinates, which resulted in structures with backbone RMS values of 0.5-0.99 Å.

When the backbone coordinates are known, the side chain atoms can be built. Several approaches also exist for this task. The approach of Reid and Thornton[9] mainly consists of carefully and manually adjusting the χ torsion angles, after these were initially set at the preferred values, taken from the distribution of χ angles in known protein structures.[13-15] The method, tested on the protein flavodoxin,[9] yielded an RMS value of 2.4 Å for all the side chain atoms and an RMS of 1.7 Å for all the nonhydrogen atoms in the protein, as compared to the X-ray structure. Correa's MD method, already mentioned above, is also able to construct protein side chain coordinates and test cases showed RMS values of 1.3-1.7 Å for all the heavy atoms.[11] In the MaxSprout[8] program, Holm and Sander implemented a Monte Carlo procedure to optimize the side chain conformations. Their method has been tested on several proteins, resulting in average RMS values of 2.2 Å for all side chain atoms.[8] Recently, an automatic segment matching protocol has been described which uses information from a database of known protein structures to position the side chain atoms.[5] The resulting structures showed a side chain RMS of on average 1.87 Å.

As has been mentioned above, MD calculations seem promising in the prediction of protein structure, both for backbone and side chain atoms.[11] We have explored the MD approach further, evaluating the ability of extensive MD calculations to capture details of the native protein conformation, and to what level of precision.

We present a method to build a complete protein structure from partial coordinate information, i.e., the Cα coordinates. In this CSB-MD method (crude structure building followed by MD refining), the protein backbone is generated in a fast and rather crude way; known protein structures are not needed, unlike the

SP method. Then, extensive MD calculations are applied to position the side chain atoms. Details of the CSB-MD approach will be discussed in the Methods section and results will be compared to the results of two other methods in which MD calculations play a major role. We supplemented the first of these, the existing (SP) method to build backbones, by a MD procedure to add side chains (combined SP-MD approach). For the second one we used the MD method of Correa for which we suggest (and have applied) some minor modifications.

All three methods were tested on two proteins, from our current research, for which only the Cα coordinates were available. The first protein, yeast enolase, is a globular protein of 436 amino acids, which catalyzes the dehydration of 2-phosphoglycerate to phosphoenolpyruvate in the glycolytic pathway. The structure of yeast enolase has been solved at 2.25 Å resolution.[16] The second protein is the RNA binding domain of the A protein, which is part of the U1 small nuclear ribonucleoprotein particle (U1 snRNP). The U1 snRNP particle plays an important role in the removal of introns from pre-messenger RNA, the process known as splicing.[17] A number of proteins which can bind RNA contain one or more copies of a conserved motif of about 80 amino acids, the so called RNP-80 motif.[18] The N-terminal part of the A protein which binds to U1 snRNA contains an RNP-80 motif. Part of the protein (amino acids 1 to 95) was crystallized and the structure has been determined at a resolution of 2.8 Å.[19] Recently, full coordinate sets for both enolase and the RNP-80 motif have become available, thus allowing us to evaluate the precision of our constructed model structures. Some other validation criteria which can be used in the absence of complete coordinate sets are also discussed.

METHODS

Atomic coordinates

Cα coordinates of enolase were taken from the Brookhaven Protein Data Bank, PDB code 2ENL. The protein contains 436 amino acids arranged in a N-terminal domain (ca. 140 amino acids) and a main 8-fold barrel domain which contains the unusual topology $\beta\beta\alpha\alpha(\beta\alpha)_6$.[16] Recently, complete sets of coordinates have become available; one of them (4ENL) was used for the evaluation of the constructed protein structures. Cα coordinates of amino acids 6-90 from the RNP-80 motif of the U1 A protein were kindly provided by K. Nagai (MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK). Amino acids 1-5 and

91-95 are poorly ordered in the electron density map and are not used in this study. The structure contains a four-stranded antiparallel ß-sheet and two α-helices arranged in the order ßαßßαß in the primary structure. The four ß strands lie in a plane with the two helices on the same side of the sheet.

Model building

Backbone building

In our new CSB-MD method the GROMOS force field[20] was used in all computations (see Table I for computational details). The procedure starts with a very crude backbone, which was created by positioning all intermittent backbone atoms (C, N) on one-third and two-third of the distance between $C\alpha_i$ and $C\alpha_{i+1}$. Then carbonyl oxygen atoms and amide hydrogens were placed at idealized bond distances and with ω torsions of 180°, followed by a small random shift of all N and C atoms to avoid undefined $C\alpha_i$-C-N-$C\alpha_{i+1}$ backbone dihedrals. The resulting crude polyglycine chain was subjected to energy minimization using steepest descents (SD) to relieve the strain in the initial backbone, keeping all $C\alpha$ atoms fixed to their X-ray coordinates.

In our combined SP-MD approach, used for comparison with the CSB-MD method, the standard "construct backbone" option, as implemented in SYBYL,[21] was used to generate the protein backbone, from fragments of a protein database. The constructed backbone was then minimized briefly with SD ($C\alpha$ atoms fixed).

The third backbone building method applied comparatively was the one described by Correa,[11] but with several specific adjustments. Using the CHARMm force field,[22] we built the backbone chain using only Ala, Gly and Pro residues (Gly and Pro to account for greater and lesser flexibility, respectively, in the chain). Residues were built sequentially, and after each amino acid addition a short SD minimization was performed. During backbone building, a harmonic constraint was imposed on all $C\alpha$ atoms and dihedral constraints were set on the ω torsion angles to keep the peptide bonds in the trans orientation (*cis* orientation for both *cis*-prolines in enolase). The resulting protein backbone was refined with MD calculations at 1000 K.

Side chain building

In the CSB-MD method, we add all side chain atoms both simultaneously and in an extended conformation to the constructed backbone. Direct optimization of the

TABLE I. Overview of the Three Building Procedures

	CSB-MD	SP-MD*	Correa (adapted)†
General settings			
Nonbonded cutoff distance	8 Å	9 Å	10 Å
Dielectric constant	ε=1	ε=r	ε=r
Charges	Not scaled	Not scaled	Not scaled (RNP-80); scaled (RNP-80 and enolase)
Force field	GROMOS, united atom; version 37D4	AMBER; united atom; SYBYL, version 5.41	CHARMm 21 3 (enolase united atom; RNP-80 all atom)
Backbone construction	Build backbone in a "brute force" way (see text) SD, Cα fixed	SYBYL "construct backbone" RMS threshold 0 5 Å database-size: 215 proteins SD, Cα fixed	Build backbone gradually SD, Cα harmonic constraints (120 0 kcal/mol) MD 1000 K (enolase 60 psec; RNP-80 100 psec); Cα harmonic constraints (10 0 kcal/mol)
Side chain construction	Add extended side chains SD with increasing nonbonded cutoff distance SD, Cα harmonic constraints (2.0 kcal/mol) MD calculations	Add extended side chains Conformational search χ (delta=30 deg) SD, backbone fixed SD, unconstrained MD calculations	MD calculations
MD settings			
SHAKE[29]	Bonds to hydrogen (RNP-80); all bonds (enolase)	Bonds to hydrogen	Bonds to hydrogen
Timestep	1 fsec	1 fsec	1 fsec
Constraints	Cα harmonic constraints (2 5 kcal/mol) Omega harmonic constraints (4 0 kcal/mol)	Cα fixed	Cα harmonic constraints (10.0 kcal/mol) Omega harmonic constraints (5 0 kcal/mol)

MD	RNP-80 (psec)	Enolase (psec)	Temp. (K)	RNP-80 (psec)	Enolase (psec)	Temp. (K)		RNP-80 (psec)	Enolase (psec)	Temp. (K)
Heating	3	5	0→800	2	2	0→300	Heating	3	3	0→800
Simulation	47	35	800	28	15	300	Simulation γ-level	30	20	800
Cooling	7 5	8 2	800→0	4	4	300→0	Simulation δ-level	30	20	800
							Simulation ε level	30	20	800
							Simulation total	30	30	1000
							Cooling	8	8	1000→0

	CSB-MD	SP-MD*	Correa (adapted)†
Final structure	SD, Cα harmonic constraints (2 5 kcal/mol) CG, constrained SD, unconstrained	SD, unconstrained	SD, unconstrained

*The SP-MD approach uses a standard SP method to create the protein backbone and a new MD method to create the side chains.

†The adaptations to the original method of Correa are described in the text

resulting structure failed, because of obvious very short nonbonded interactions but this problem was easily overcome by SD minimization with a gradually increasing nonbonded cutoff distance, ranging from 0.01 Å to 8 Å (Cα atoms fixed). During further minimization the positional constraint was replaced by a harmonic constraint on the Cα crystal structure coordinates. Subsequently, MD simulation at 800 K was performed, with harmonic constraints on the Cα positions. Long MD calculations are necessary because the building process is driven by the gradual formation of hydrogen bonds which takes considerable time. Since the α carbons are harmonically restrained, the temperature must be high in order to allow conformational changes (such as flips of peptide units). Dihedral constraints were applied to the ω torsions; for enolase, however, GROMOS could not cope with 435 dihedral constraints, and the parameter set had to be adapted to contain a higher than usual force constant (12 instead of 8 kcal/mol) for this type of torsion angle. After cooling to 0 K, the structure was subjected to constrained SD and conjugate gradient (CG) minimizations, followed by an unconstrained minimization step until convergence (see also Table I for computational details). The whole building procedure was monitored using the Quanta molecular modeling package.

In our combined SP-MD method, applied for comparison to the CSB-MD method, the existing SP method to build the backbone is supplemented with a new MD procedure to add side chains. As in the CSB-MD method, we added the side chains in their fully extended conformation. A conformational search with an increment of 30° was performed for all the χ torsion angles, in combination with a quick SD routine, to relieve initial bumps in the structures. The structure was then subjected to a short SD minimization (backbone atoms were kept fixed) to relieve further close contacts. For enolase, which contains two *cis*-prolines, some additional manual building had to be done because the generated backbone contained all ω torsions in the *trans* orientation. Subsequently, the complete structure was minimized briefly with SD (no constraints) to generate a starting conformation for the MD simulations at 300 K. Due to software limitations of the SYBYL package, it was not possible to harmonically constrain the Cα positions of the constructed backbone to the X-ray Cα coordinates and, therefore, the Cα coordinates had to be fixed during all the MD calculations. After cooling to 0 K, the structures were finally subjected to SD minimization without constraints on any atoms, until convergence.

With the adapted Correa method, the second method which we applied for comparison, the side chains were built by sequential addition of (respectively) the

γ-, δ- , ϵ-, and ζ-atoms. During this sequential building of levels of atoms, the side chains of the aromatic rings of Tyr and Phe were added in one step, to avoid undesirable effects of partially built rings. After addition of each level of side chain atoms, the resulting structure was refined with MD at 800 K. To achieve the final structure, all constraints were removed and the cooled structure (0 K) was minimized to convergence with the SD minimizer. For the globular enolase structure, charges on side chain atoms were scaled in the following way: charges of atoms between 0 an 12 Å from the center of the molecule were not scaled; charges of atoms between 12 and 18 Å were scaled by a factor 0.7; and charges of atoms between 18 and 40 Å from the center were scaled by a factor of 0.3. This downscaling of charges near the surface of the protein is a way, in addition to the distance-dependent dielectric, to mimic solvent screening effects.[11,23] Two model structures were built for RNP-80. The first one was built using unscaled atomic charges, because this structure deviates too much from an ideal, globular shape; moreover, since it is only part of a larger protein, not every part of the current structure will necessarily be part of the surface of the native protein structure. The second structure was built using charges scaled down by a factor two, which made it possible to evaluate the effect of the charges on the excessive backfolding of side chains on the backbone of the protein, a phenomenon that was found in an earlier study.[11]

RESULTS AND DISCUSSION

Protein Structure Building

The crude polyglycine backbones generated with the CSB-MD method showed Cα RMS values of 0.50 Å for RNP-80 and 0.55 Å for enolase (abbreviated to 0.50/0.55 Å in the following) as compared to their respective crystal structure coordinates. After addition of the side chains and energy minimization with increasing nonbonded cutoff distance, the resulting Cα RMS was 0.74/0.73 Å. During the MD steps Cα RMS values were 0.8 Å during heating, 0.6-0.7 Å during simulation, and 0.4 Å during cooling. The difference between Cα RMS values during the initial stages of protein building with the CSB-MD method vs. the adapted Correa method (see below) is remarkable. In the CSB-MD approach, the initial backbone is allowed to deviate considerably from the X-ray Cα coordinates, and the structure is pulled gradually towards these coordinates during the MD simulation. In the Correa approach, however, the initial backbone fits very well to the Cα X-ray coordinates, and during the modeling work the structure is allowed to deviate from the X-ray coordinates to find an optimal compromise between RMS-fit and protein structure. Some manual adjustments of the CSB-MD structures were necessary during the high temperature MD simulation, as it turned out that the GROMOS force field is considerably more sensitive than the CHARMm force field to these high temperatures. After unconstrained minimization of the cooled CSB-MD structures a Cα RMS of 0.39 Å was achieved for both enolase and the RNP-80 motif.

In our combined SP-MD approach, applied for comparison to the CSB-MD method, the generated polyalanine backbone in the RNP-80 motif was built from 18 fragments with an average length of 7.3 amino acids; the backbone of enolase was generated from 93 fragments with an average length of 7.6 amino acids. The Cα RMS values were 0.38/0.36 Å. The side chains were placed and, after an initial conformational scan, the structures were subjected to MD calculations (Cα positions fixed). After cooling down, unconstrained SD minimizations until convergence resulted in final structures with Cα RMS values of 0.35/0.42 Å.

In our adapted Correa method, the first building step, in which Ala, Gly and Pro residues were added sequentially, yielded structures with Cα RMS values of 0.05/0.04 Å. These low RMS values resulted from the very large force constant of the harmonic Cα constraints (120 kcal/mol). The subsequent MD simulation (Cα constraint constants reduced to 10 kcal/mol), resulted in backbone structures with Cα RMS values of 0.17/0.14 Å. In the next step, all γ-level side chain atoms were added to the structures, which were then subjected to MD simulation at 800 K. Then the δ-, ϵ- and ζ- atoms were added, respectively, and MD calculations were performed after each step. RMS values for the Cα carbons were 0.35-0.4/0.35 Å during these steps of the building procedures. SD minimization was performed on the final cooled structures, after removal of all constraints, resulting in final structures with RMS values of 0.30/0.20 Å for the Cα atoms with respect to the X-ray Cα coordinates.

Evaluation of the Constructed Protein Structures

Because X-ray coordinates for both of our test proteins became available during this study, these were used as a reference to judge the constructed model structures, thereby providing an implicit comparison of the construction methods. The quality of the backbone structures will be evaluated according to the following criteria:

1. the deviations in the φ and ψ torsion angles;
2. the percentage and quality of backbone hydrogen bonds;
3. the peptide flips that occur, and, of course;
4. the RMS deviation from the X-ray structures.

The quality of the side chain conformations will be evaluated by

1. the deviations in side chain torsion angles; and
2. the RMS values when compared to the X-ray structures.

Because complete X-ray data to test the validity of constructed protein structures are not always available (and were not available when we started this study), some other criteria to judge model building quality will be discussed for the RNP-80 model structure which we built by the adapted Correa method. Because it has been shown that criteria like total energy or total surface accessible area are not good discriminatory factors to distinguish between correctly folded and misfolded structures,[24] we have used, among others, relative surface accessibility of side chain atoms and the known distributions of side chain torsion angles in high resolution proteins.

Quality of Backbone Conformation

Deviations in backbone torsion angles

The backbone conformation of the protein models constructed is generally in good agreement with the X-ray structures (average deviations in the φ and ψ torsion angles of 15° to 20° with the adapted Correa and the combined SP-MD methods and up to 25° with the CSB-MD method). The deviations of the φ and ψ torsion angles of the RNP-80 structure constructed with our CSB-MD method are shown in Figure 1. Although some large local deviations can be observed, these are nearly always located between or at the end of secondary structure elements. Furthermore, a correlation between the magnitude of deviations in $\psi(i)$ and $\varphi(i+1)$ is almost always present. If both deviations are large and of opposite sign, this has no influence on the direction of the backbone and the Cß atoms.[3,9] Regarding the prediction of positive φ-values, which are rare except for glycines and, to a lesser extent, for asparagines, we found that with the adapted Correa, SP-MD and CSB-MD methods respectively 78, 87, and 74% of the residues with positive φ-values were predicted correctly. The percentage amino acids incorrectly built with $\varphi > 0$ is

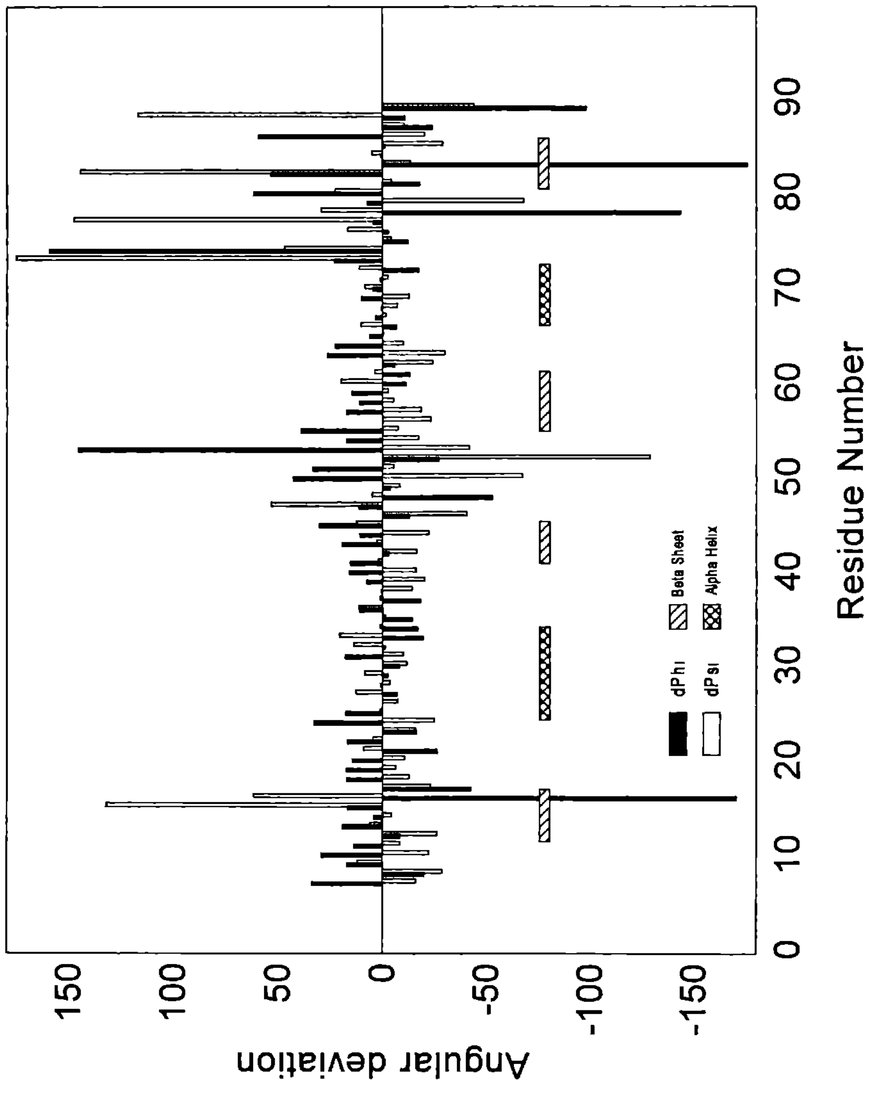


Figure 1. Deviations in backbone torsion angles φ and ψ between the RNP-80 structure generated by our CSB-MD method and its X-ray structure The secondary structure elements are also shown.

ca. 2% in all three methods.

Backbone hydrogen bonding

In native proteins the percentage of residues which is involved in main chain hydrogen bonding is on the average larger than 50%.[8,25] Our modeled structures showed 56 to 71% main chain hydrogen bonds, whereas the X-ray structures of RNP-80 and enolase showed 57 and 76%, respectively. Considering the backbone hydrogen bonds in α-helices and ß-sheets, our CSB-MD method shows results similar to the two methods used for comparison, while the adapted Correa method shows a somewhat better backbone hydrogen bonding pattern in turn regions than the two other methods. In all three methods, the created backbones contained most of the main chain hydrogen bonds in the α-helices and ß-sheets. During the MD calculations, the additional hydrogen bonds of the α-helices and ß-sheets were formed, as well as the majority of hydrogen bonds in the turn regions. The RNP-80 structure that was built by using scaled charges (see Methods) showed lower Cα RMS values but in this structure not all of the main chain hydrogen bonds were formed correctly.

Peptide flips

A peptide flip is a badly oriented peptide unit present in the constructed structure, in comparison to the X-ray structure, and occurs when the angle between the X-ray carbonyl oxygen atom, the X-ray carbonyl carbon, and the model carbonyl oxygen atoms is larger than 90°.[8] In general, the number of flips reported in the literature for similar studies is less than 5%,[3,8] and in most cases, peptide flips do not occur in regular α- and ß- regions, but rather in turn and coil regions of a protein. In particular the SP-MD method of backbone generation is expected to be sensitive to flips, because junctions between fragments can easily generate a peptide flip, but we did not observe such a sensitivity. From our modeled structures, only the CSB-MD structures showed a percentage peptide flips of 6-8% while the other two methods yielded structures with on average 4.4% peptide flips, the majority of which occurs inturn and coil regions.

RMS values

Several RMS values of the generated structures, as compared to their X-ray structures, are given in Table IIA, while Table IIIA shows results of other recent model building studies. Both the CSB-MD method and the two comparatively applied methods generate structures with acceptable RMS values for the backbone atoms and all three methods compare well to the values given by other authors. When discussing the magnitude of RMS values, it should be born in mind that for independently determined high resolution structures (< 2 Å) of the same protein in a different crystallographic environment (e.g., two unique molecules in an asymmetric unit, or the same proteins crystallized from different solvent conditions) generally RMS values of 0.3-0.6 Å are found for heavy atoms in the secondary structure elements.[26,27] Even higher values are reported for all heavy atoms.

The extremely low Cα RMS values reported by Correa[11] and Wendoloski[10] are remarkable. It is questionable if it is correct to try to achieve such a low value regarding the low resolution of the X-ray data involved. In our application of the Correa method we therefore decreased the Cα harmonic constraint force constant during the initial building steps (from 120 kcal/mol to 10 kcal/mol), aiming at the often occurring situation in which structures of which only Cα coordinates are known, are usually only poorly refined.

Our application of Correa's method shows, on average, the lowest RMS values while the CSB-MD method shows the largest, although the differences are small. Furthermore, the distribution of RMS values (data not shown), and the deviations in backbone torsion angles (see Figure 1) are clearly related to the presence of secondary structure elements; the backbone RMS values are significantly lower for residues in secondary structure motifs (α-helices and ß-sheets) than in other regions. This is to be expected, since hydrogen bonding patterns dominate these secondary structure elements and lower the conformational freedom of their atoms. As a consequence, the α-helices and ß-sheets are built more accurately than the less geometrically confined areas. Residues in the core of the protein also have restricted conformational freedom, compared to residues closer to the surface of the protein. Indeed, most structures show a somewhat lower Cα RMS value for residues in the core. The relatively high RMS values of the backbone oxygen atoms can be explained by their longer distance to the main chain, as it takes only a minor shift in backbone atoms to move the oxygen atoms considerably.

Quality of Side Chain Conformation

RMS values

The RMS deviations of the constructed side chains are given in Table IIB and are compared with results from other recent model building studies in Table IIIA.

TABLE II. RMS Values of Constructed Protein Structures as Compared to Their X-Ray Coordinates

	RNP-80[*]			Enolase[†]		
	CSB-MD	SP-MD	Correa[‡]	CSB-MD	SP-MD	Correa[‡]
A. Backbone conformation						
Cα RMS (Å)	0 40	0.35	0.30	0.42/0.39[§]	0 42/0 37[§]	0.27/0.20[§]
Cα RMS (secondary structure[**]) (Å)	0.30	0 34	0.26	0.37	0.33	0.20
Cα RMS (core[††]) (Å)	0.38	0.31	0.28	0 42	0.42	0 24
Backbone RMS (Å)	0 70	0.53	0.49	0.64	0.64	0.50
Backbone RMS (secondary structure) (Å)	0.44	0.51	0.32	0.49	0.41	0.30
Backbone oxygen RMS (Å)	1.11	0.83	0.75	1.04	0.96	0 84
B. Side chain conformation						
RMS of side chains (Å)	2.52	2.52	2.53	2.49	2.41	2.14
RMS of side chains (core) (Å)	2.22 (1.64 excl. Y86[¶])	1.34	1.24	2.47	2.10	1.53

[*] In the RNP-80 X-ray structure[19] several side chains on the surface of the molecule were placed arbitrarily, because the structure was disordered in those parts of the protein; these residues are not taken into account in the calculation of the RMS values.

[†] The model structures of enolase are compared to the protein 4ENL, which shows a Cα RMS value of 0.23 Å when compared to 2ENL (used to build the enolase structures) and which also shows much lower thermal parameters Thus in the case of enolase the modeled structure is compared to the atomic coordinates of a further refined structure.

[‡] The adaptations to the original method of Correa are described in the text.

[§] Cα RMS values in comparison to 2ENL.

[**] Secondary structure residues are defined as amino acids located in α-helices or in ß-sheets.

[††] Core residues are defined as those residues with a side chain solvent accessible surface of less then 15%, relative to the tripeptide Gly-X-Gly, corresponding to ca. 40-50 % of the residues.

[¶] Tyrosine 86 is particularly ill-placed in the RNP-80 CSB-MD structure and has a large influence on the total RMS; therefore also the RMS value without tyrosine 86 is given.

The CSB-MD method generates structures with side chain RMS values (> 2 Å) comparable to those of the two methods we applied comparatively. In particular, side chains at the surface of the proteins appear difficult to predict correctly and often deviate rather far from the X-ray structure, as was found in an earlier study.[11] Our side chain RMS values are slightly larger than those obtained in most other studies (see Table IIIA), despite the fact that we performed MD calculations on extended side chains (CSB-MD and SP-MD methods), and on gradually growing side chains (adapted Correa method), while Reid and Thornton used a careful building scheme[9] and Holm and Sander used a rotamer library and Monte Carlo procedure.[8]

TABLE III. Comparison to Other Studies*

A. RMS values

	Cα RMS (Å)	Backbone RMS (Å)	Side chain RMS (Å)	Side chain RMS (Å) (core residues)	All RMS (Å)
Correa[11]	0.02	0.19	NR	NR	1.29
	0.03	0.49	NR	NR	1.68
	0.3	0.41	NR	NR	1.64
Claessens et al.[3]	ca. 0.4	0.58	NA	NA	NA
Holm and Sander[8]	0.1-0.2	0.4-0.6	2.21	1.56	1.57
Reid and Thornton[9]	NR	0.57	2.41	NR	1.73
Tuffery et al.[13]	NA	NA	1.69	1.54	NA
Wendoloski and Salemme[10]	0.04	0.35	2.05	NR	1.41
Levitt[5]	NR	0.42	1.78	NR	1.26
Bassolino-Klimas and Bruccoleri[12]	0.30-0.87	0.5-0.99	NA	NA	NA
This study					
CSB-MD	0.40	0.70	2.52	2.22 (1.64 excl. Y86)	1.86
	0.42	0.64	2.49	2.47	1.76
SP-MD	0.35	0.53	2.52	1.34	1.83
	0.44	0.64	2.41	2.10	1.72
Correa (adapted)	0.30	0.49	2.53	1.24	1.83
	0.27	0.50	2.14	1.53	1.51

B. Deviations in side chain torsion angles

	χ1 ± 20° (%)	χ1 ± 30° (%)	χ1 ± 40° (%)	χ1 ± 60° (%)	Δ(χ1) (deg)
Correa[11]	NR	NR	NR	62	75
Holm and Sander[8]	44	54 (core 67)	NR	NR	NR
Reid and Thornton[9]	40	NR	NR	NR	58
Wendoloski and Salemme[10]	NR	NR	59	NR	NR
Levitt[5]	NR	72	NR	NR	NR
This study					
CSB-MD	37	50	58	67	52
	44	49	54	56	57
SP-MD	45	51	63	68	48
	49	54	57	59	52
Correa (adapted)	42	49	53	59	56
	53	63	67	70	44

*In most studies, several proteins were built and this table shows the average RMS values and the average χ1 deviations. NR, not reported; NA, not applicable (Claessens et al.[3] and Tuffery et al.[13] constructed only backbone and side chain conformation, respectively).

In most cases the side chain RMS values of core residues are significantly lower although for some structures (e.g., the enolase CSB-MD structure) the RMS values in the core are comparable to the total side chain RMS, which was also found for some of the proteins tested in an earlier study.[8] For the core residues, only the Correa and SP-MD methods show side chain RMS values which are similar to earlier studies, while the CSB-MD method performs worse.

When comparing the three methods used in this study, our adapted Correa method gives the best results, although the total side chain RMS of the RNP-80 motif is equal to that of the other methods. This may be due to the fact that, in a small protein like RNP-80, relatively many residues are on the surface of the molecule. Another reason could be the fact that in enolase charges towards the surface are scaled, which may yield better side chain conformations. In the RNP-80 structure built with scaled charges, the side chain positioning was in fact better than in the RNP-80 built with the unscaled procedure, but there the backbone conformation was worse (data not shown). Our comparatively applied SP-MD calculations shows reasonable side chain RMS values but performs less well than the Correa method, especially in case of the enolase structure. Our CSB-MD method results in side chains with rather large side chain deviations which are evenly distributed over core and surface residues. In both our SP-MD and CSB-MD computations, MD calculations were performed on side chains added in extended fashion. The differences between the two methods consist of a quick conformational search before the MD simulation in the SP-MD method, the application of a different temperature for the MD calculations, and the use of a different force field.

One might expect that for some amino acids, correct predicton of side chain conformation is more difficult than for others. For example, charged and aromatic amino acids are known to be very difficult to calculate correctly.[11,13] Indeed, we find in the CSB-MD enolase structure, for example, that the highest deviations occur in the Arg, Lys, His, Trp and Tyr residues, i.e., amino acids with large and/or flexible side chains (data not shown).

Figure 2 shows the RNP-80 structure generated by our application of Correa's method compared to the X-ray structure. Residues which are present in the RNP1 and RNP2 regions (which are responsible for the interaction with U1 RNA) are shown and agree very well with the X-ray data.

Deviations in side chain torsion angles

Deviations in the $\chi 1$ angles, with reference to X-ray coordinates are given in Table IIIB. This table also includes some data from other recent model building studies. The average deviations in the $\chi 1$ angles are comparable in all three methods tested and show similar levels of accuracy to other studies. The CSB-MD

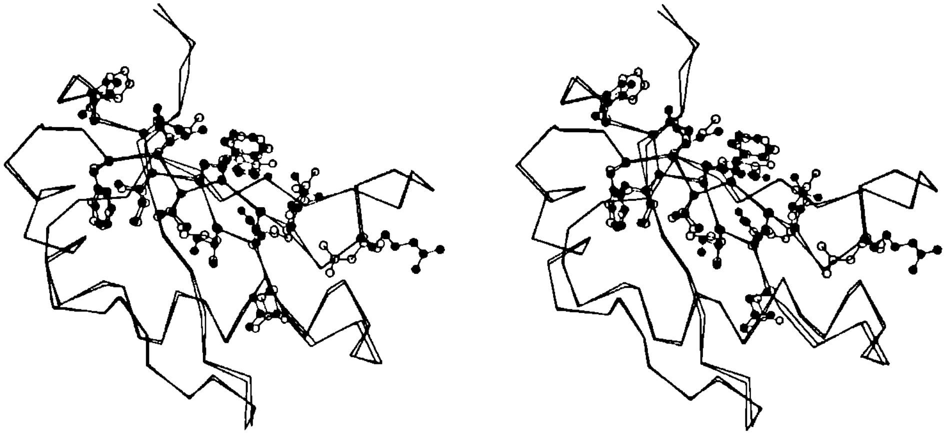


Figure 2. Superposition of the X-ray structure and the structure of RNP-80 created with the adapted Correa method. Side chains on the RNA-binding surface of the RNP-80 motif are shown. Filled circles correspond to the X-ray structure while open circles correspond to the RNP-80 model structure.

method performs almost as well as the two comparatively applied calculations, predicting on average 56% of the $\chi 1$ angles correctly within 40°. The distribution of the deviations in $\chi 1$ angles for the enolase structure, constructed with our CSB-MD and SP-MD methods, is shown in Figure 3. A similar distribution was found in the Correa method and in a earlier building study.[8] However, the CSB-MD and SP-MD structures show a regular distribution of misplaced side chains throughout the protein, whereas the adapted Correa method performs better, most misplaced side chains occurring in turn or coil regions.

Structure Validation in the Absence of Complete X-ray Data

Prior to the availability of the full enolase and RNP-80 protein X-ray structures, we had already assessed the reliability of the RNP-80 motif constructed in our application of the Correa method. As a first criterion, we used the accessible surface area of side chain atoms to a 1.4 Å spherical probe (equivalent to the radius of a water molecule).[9,28] The relative surface accessibility is given as the ratio between the solvent accessible surface of a side chain of amino acid X in the model structure and the solvent accessible surface of a side chain X in the tripeptide Gly-X-Gly (φ=-139, ψ= 135, $\chi 1$=120).[9] The distribution of these values in high

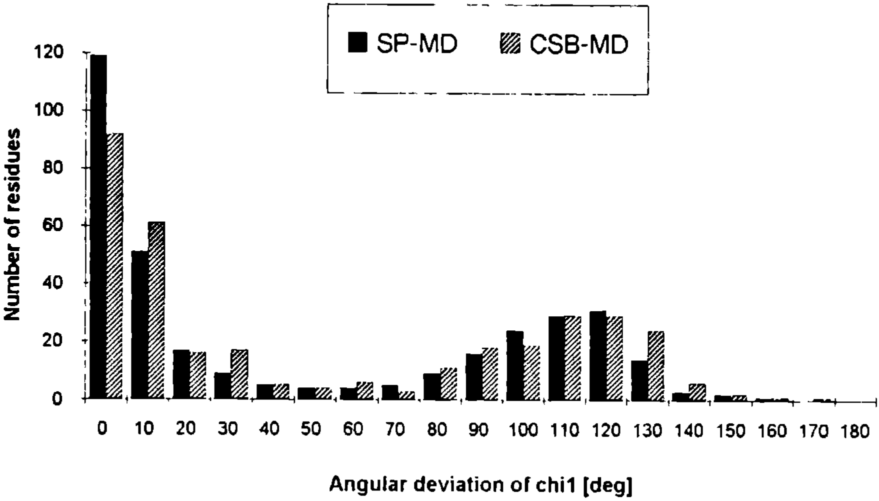


Figure 3. Angular deviation of the side chain torsion angle $\chi 1$ in the enolase structures computed by the CSB-MD and SP-MD methods.

resolution protein structures was studied by Reid and Thornton.[9] With the aid of their data amino acids can be identified which may have adopted an unusual conformation. If the side chain conformation of an amino acid shows a relative surface accessibility value which occurs in less than 7% of the side chain conformations of that same amino acid in high resolution proteins, it might be badly placed. In our RNP-80 structure, built with unscaled charges, we could locate 15 residues which were possibly ill-placed. A majority (10) of these residues is located at the end of secondary structure elements or in turn and coil regions.

Subsequently, we compared side chain torsion angles to the statistical distribution of side chains in known protein structures, containing 106 rotamers of the 19 nonglycine amino acids.[13] Our model structure showed that 8 of 85 amino acids are in an unusual conformation, of which 6 are present in turn or coil regions.

Finally, as a last criterion for the validity of a predicted structure, we examined the placement of the side chains of the polar amino acids, looking for side chains which point into the core of the protein even though they are not hydrogen bonded.[9] In our RNP-80 structure, 7 polar side chains of amino acids were not hydrogen bonded, but they all point into the solvent and not into the core of the protein.

When the atomic coordinates of RNP-80 became available, we checked whether or not we could find a correlation between amino acids predicted to be unusual according to the above criteria and amino acids for which there was a large deviation in the χ angles between the modeled and the X-ray structure. Indeed, about 70% of the $\chi 1$ angles with a deviation of more than 90° from the X-ray structure were detected by one or more of the criteria mentioned above, showing that these criteria can, to a certain extent, assist to estimate the validity of a model structure in the absence of complete X-ray data.

SUMMARY AND CONCLUSIONS

Our study shows that a molecular dynamics approach to generate full protein model structures from only the Cα coordinates yields reasonable structures.

To construct a protein backbone, the CSB-MD and two other MD based procedures applied for comparison all produced structures of comparable quality. Our CSB-MD and combined SP-MD methods both yield good backbone structures very quickly. The SP-MD method uses information from known protein structures whereas in the CSB-MD method the backbone atoms are initially placed without any prior knowledge. The adapted Correa method performs best, suggesting that building the backbone one amino acid at a time, followed by stepwise minimization and by long MD calculations yields the best structures. If sufficient computer resources are available, this might be preferred, but if computer time is limited, either the CSB-MD or the SP-MD method can be used to generate a reliable backbone quickly.

To position the side chains, the results of the combined SP-MD method show that MD calculations (at 300 K) on side chains, initially placed in an extended conformation, followed by a quick conformational search to relieve initial bumps, give reasonable results, which are comparable to other studies. In the CSB-MD method, side chains are also added in extended fashion, but no initial conformational search is performed and the MD calculations are done at high temperatures (800 K). GROMOS appeared to be very sensitive to these high temperatures, and it is not yet clear whether it is the high temperature or the absence of an initial conformational search of the side chains that causes side chain conformations of lesser quality. MD calculations on extended side chains at high temperature should be avoided until it is clear whether this high temperature causes the bad positioning of side chains in the CSB-MD method, or whether the GROMOS force field can not cope with these high temperatures. Again, the adapted Correa method performs best, suggesting that the MD calculations are most valuable when used on a gradually growing structure, in which every level of newly added atoms is given time to acommodate. This method is recommended for the construction of side chains, when computer time is not limited.

The CSB-MD method is rather time consuming, as are the two methods applied for comparison. For the small RNP-80 structure, the CSB-MD method took 80 hr of CPU time on a Convex C120, while the SP-MD method consumed 95 hr of CPU time and the adapted Correa method about 250 hr of CPU time (both on a Silicon Graphics Iris 4D/70GT). However, using each method on currently available hardware, computation times would be drastically reduced. An advantage of the CSB-MD method and the two methods run for comparison is that they all apply standard molecular modeling sofware packages and need no special databases, special computational routines, or even expert protein structure knowledge to extend a limited coordinate set to a complete protein structure.

It is evident that the CSB-MD method, as well as the two other ones, can be used in other, more complicated, modeling problems. For example, structure prediction in homology modeling could rely on framework structures to be refined by MD techniques. Our calculations were on low-resolution Cα coordinates but in all cases where an approximate outline of the protein backbone is available, this backbone could be extended to a full coordinate set.

In conclusion, our study shows that extensive MD calculations are promising in capturing, to some extent, details of the native protein conformation. These MD-based methods will be generally applicable in protein structure prediction and the resulting protein structures can be used (within limits) with confidence to study the general structure of the protein involved, or as a basis for further model building of homologous protein structures.

ACKNOWLEDGMENTS

The authors thank Dr. K. Nagai for making available the coordinates of the RNP-80 structure and Dr. Lebioda for making available the coordinates of the enolase structure. The authors thank Dr. Hilbert Bruins Slot for assistance in the use of the SYBYL package. Use of the services and facilities of the Dutch CAOS CAMM Center, under Grants SON 326-052 and STW NCH99.1751, is gratefully acknowledged.

REFERENCES

1. Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M. Knowledge based prediction of protein structures and the design of novel molecules. Nature 326:347-352, 1987.
2. Bernstein, F.C., Koetzle, T.F., Williams, E.J.B., Meyer, G.F.Jr., Kennard, O., Shimanouchi, T., Tasumi, ,M. The protein data bank: a computer based archival file for molecular structures. J. Mol. Biol. 112:535-542, 1977.
3. Claessens, M., Van Cutsem, E., Lasters, I., Wodak, S. Modelling the polypeptide backbone with 'spare parts' from known protein structures. Prot. Eng. 2:335-345, 1989.
4. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5:819-822, 1986.
5. Levitt, M. Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 226:507-533, 1992.
6. Richardson, J.S., Richardson, D.C. In: "Prediction of Protein Structure and the Principles of Protein Conformation." Fasman, G.D. (ed.). New York: Plenum Press, 1989.
7. Taylor, W.R., Thornton, J.M. Prediction of super secondary structure in proteins. Nature 301:540-542, 1983.
8. Holm, L., Sander, C. Database algorithm for generating protein backbone and side-chain co-ordinates from a Cα trace. Application to model building and detection of co-ordinate errors. J. Mol. Biol. 218:183-194, 1991.
9. Reid, L.S., Thornton, J.M. Rebuilding flavodoxin from Cα coordinates: A test study. Proteins 5:170-182, 1989.
10. Wendoloski, J.J., Salemme, F.R. PROBIT: A statistical approach to modeling proteins from partial coordinate data using substructure libraries. J. Mol. Graph. 10:124-126, 1992.
11. Correa, P.E. The building of protein structures from α-carbon coordinates. Proteins 7:366-377, 1990.
12. Bassolino-Klimas, D., Bruccoleri, R.E. Application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates. Proteins 14:465-474, 1992.
13. Tuffery, P., Etchebest, C., Hazout, S., Lavery, R. A new approach to the rapid determination of protein side chain conformation. J. Biomol. Struct. Dyn. 8:1267-1289, 1991.
14. Ponder, J.W., Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775-791, 1987.
15. McGregor, M.J., Islam, S.A., Sternberg, M.J.E. Analysis of the relationship between side chain conformation and secondary structure in globular proteins. J. Mol. Biol. 198:295-310, 1987.
16. Stec, B., Lebioda, L. Refined structure of Yeast Apo-enolase at 2.25 Å resolution. J. Mol. Biol. 211:235-248, 1990.
17. Lührmann, R, Kastner, B., Bach, M. Structure of spliceosomal snRNPs and their role in pre-mRNA splicing. Biochim. Biophys. Acta 1087:265-292, 1990.
18. Scherly, D., Boelens, W., Van Venrooij, W.J., Dathan, N.A., Hamm, J., Mattaj, I.W. Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA. EMBO J. 8:4163-4170, 1989.
19. Nagai, K., Oubridge, C., Jessen, T.H., Li, J., Evans, P.R. Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein-A. Nature 348:515-520, 1990.
20. Van Gunsteren, W.F. and Berendsen, H.J.C.. "Groningen Molecular Simulation (GROMOS), Library Manual." Groningen: BIOMOS, 1987.
21. SYBYL program, St. Louis, Missouri: TRIPOS Associates, Inc., 1991.
22. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M.

CHARMM A program for macromolecular energy, minimization and dynamics calculations J Comp Chem 4 187-217, 1983

23 Harvey, S C Treatment of electrostatic effects in macromolecular modelling Proteins 5 78-92, 1989

24 Novotny, J, Rashin, A A, Bruccoleri, R E Criteria that discriminate between native proteins and incorrectly folded models Proteins 4 19-30, 1988

25 Kabsch, W, Sander, C Dictionary of protein secondary structure Pattern recognition of hydrogen-bonded and geometrical features Biopolymers 22 2577-2637, 1983

26 Avbelj, F, Moult, J, Kitson, D H, James, M N G, Hagler, A T Molecular dynamics study of the structure and dynamics of a protein molecule in a crystalline ionic environment, *streptomyces griseus* Protease A Biochemistry 29 8658 8676, 1990

27 Dauber-Osguthorpe, P, Roberts, V A, Osguthorpe, D J, Wolff, J, Genest, M, Hagler, A T Structure and energetics of ligand binding to proteins *Escherichia Coli* dihydrofolate reductase-trimethoprim, a drug-receptor system Proteins 4 31-47, 1988

28 Lee, B, Richards, F M The interpretation of protein structures Estimation of static accesibility J Mol Biol 55 379-400, 1971

29 Van Gunsteren, W F, Berendsen, J H C Algorithms for macro molecular dynamics and constraint dynamics Mol Phys 34 1311-1327, 1977

CHAPTER 3

Common structural features of the Ro RNP associated hY1 and hY5 RNAs

Nucleic Acids Research 22:2498-2506 (1994)

Common structural features of the Ro RNP associated hY1 and hY5 RNAs

Celia W.G. van Gelder, José P.H.M. Thijssen, Erik C.J. Klaassen, Christine Sturchler[1], Alain Krol[1], Walther J. van Venrooij and Ger J.M. Pruijn

Department of Biochemistry, University of Nijmegen, PO Box 9101, 6500 HB Nijmegen, The Netherlands and [1]UPR du CNRS: Structure des Macromolécules Biologiques et Mécanismes de Reconnaissance, IBMC, 15 Rue René Descartes, 67084 Strasbourg Cedex, France

ABSTRACT

The secondary structures of human hY1 and hY5 RNAs were determined using both chemical modification techniques and enzymatic structure probing. The results indicate that both for hY1 and for hY5 RNA the secondary structure largely corresponds to the structure predicted by sequence alignment and computerized energy-minimization. However, some important deviations were observed. In the case of hY1 RNA, two regions forming a predicted helix appeared to be single-stranded. Furthermore, the pyrimidine-rich region of hY1 RNA appeared to be very resistant to reagents under native conditions, although it was accessible to chemical reagents under semi-denaturing conditions. This may point to yet unidentified tertiary interactions for this region of hY1 RNA. In the case of hY5 RNA, two neighbouring internal loops in the predicted structure appeared to form one large internal loop.

INTRODUCTION

Ro ribonucleoprotein particles (Ro RNPs) are present in the cytoplasm of eukaryotic cells (1, 2). They consist of one RNA molecule (called Y RNA) and three common proteins, called Ro60, Ro52 and La. The function of these RNPs is not yet known.

In human cells four Y RNAs, called hY1, hY3, hY4 and hY5 RNA (ranging in length from 84 to 112 nucleotides) have been identified, while in other species two to four Y RNAs were found (3, 4). Recently, four distinct Y RNAs from *Xenopus laevis* have been identified and sequenced (5), three of which appeared to be related to hY3, hY4 and hY5 RNA, respectively. The fourth *Xenopus* Y RNA (called xYα) did not appear to be a homologue of a human Y RNA.

The hY RNAs do not contain modified nucleotides (6, 7) and their sequences show mutual homology, especially in the 5' and 3' parts. They also exhibit similarity at the secondary structure level. The predicted base-pairing between the 5-' and 3'-regions of the molecule yields a conserved stem structure interrupted by a bulged residue and an internal loop. Furthermore, all hY RNAs contain a pyrimidine-rich region which varies in size between the different hY RNAs. Ribonuclease protection experiments showed that the Ro proteins bind to the lower part of the conserved stem (8) and studies with RNA mutants clearly demonstrated the importance of the bulged nucleotide in this region for Ro60 binding (9). The La protein binds to the 3'-oligouridine stretch present in all hY RNAs (9, 10).

All secondary structures published for the hY RNAs originated from low-energy structure predictions (with minor adaptations in some cases) (3, 7, 8, 11, 12). Only in one case (7) limited nuclease S1 digestion data for hY5 RNA were used. We therefore decided to investigate the conformation of hY1 and hY5 RNA in more detail, while for hY3 and hY4 RNA some preliminary structure probing experiments were performed. Several RNases were used to establish single-stranded regions in the RNA, while RNase V1 was used to locate double-stranded or stacked regions. Furthermore, chemical modifications with DMS, CMCT and kethoxal were carried out to probe the Watson-Crick positions of all four bases.

MATERIALS AND METHODS

Sequence alignment and secondary structure prediction

The alignment of the hY RNA sequences was made with the program CLUSTAL, which is part of the Wisconsin Package V 7.0 (13), and adjusted manually. The programs FOLD and MFOLD (14, 15) were used to generate optimal and suboptimal foldings.

Preparation of hY RNAs

In vitro transcription by T7 RNA polymerase was carried out as described (16). The hY RNAs were cloned into the *Eco*RI and *Hind*III sites of pGEM-3Zf(+) (9) resulting in hY transcripts with 10 additional nucleotides (nts) at the 5' end (GGGCGAAUUC) and 5 nucleotides at the 3' end (AAGCU) derived from the vector. For hY1 RNA synthesis an additional construct was made in which the transcription start site was positioned exactly at the first nucleotide of the hY1 RNA encoding sequence and in which a *Dra*I site was positioned at the sequence corresponding to the 3' end of hY1 RNA. *Dra*I linearization followed by T7 RNA polymerase transcription of this construct resulted in the synthesis of hY1 RNA lacking additional nucleotides.

5-' and 3-' end-labeling

For 5'-end-labeling the RNAs were dephosphorylated at their 5'-ends and then labeled using [γ-^{32}P]ATP and T4 polynucleotide kinase (Boehringer) as described previously (17). For 3'-end-labeling, a 20 μl reaction, containing 20 pmol RNA, 40 μCi [^{32}P]pCp (specific activity ~3000 Ci/mmol), 1 mM ATP, 50 mM Tris-HCl pH 7.8, 10 mM $MgCl_2$, 10 mM ß-mercaptoethanol, 40 U RNAsin, and 9 U of T4 RNA ligase, was incubated overnight at 4°C. The labeled RNAs were run over a Sephadex G-50 coarse spin column and were further purified by electrophoresis in a 10% polyacrylamide/urea gel. The full-length RNA products were excised from the gel and eluted overnight at 4 °C in a buffer containing 0.5 M NH_4Ac (pH 6.5), 10 mM $MgCl_2$ and 0.1 % SDS (18). The labeled RNAs were precipitated with ethanol and dissolved in water. In case of 5'-end-labeling of oligodeoxyribonucleotides 10 pmoles of each oligonucleotide were incubated with 15 pmoles [γ-^{32}P]ATP (specific activity ~3000 Ci/mmol) and 5 U of T4 polynucleotide kinase for 45 min at 37°C in a 50 μl reaction containing 50 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 5 mM DTE and 1 mM EDTA. The labeled oligonucleotides were run on a Sephadex G-25 coarse spincolumn, after which they were precipitated with ethanol and dissolved in water.

Enzymatic structure probing

All enzymatic probing experiments were performed under native (N) and denaturing (D) conditions and were repeated at least three times to obtain consistent data, while preliminary data under semi-denaturing (SD) conditions were obtained. The amount of enzyme added was optimized to obtain single hit conditions. 5'-end-labeled RNA (2 x 10^4 cpm) was supplemented with 4 μg of total yeast RNA as carrier. Digestions with RNase T1 (0.1 U), RNase A (1x10^{-6} U), RNase T2 (0.1 U), RNase U2 (0.1 U), or RNase V1 (0.08 U) were performed at room temperature for 10 minutes (N and SD), or at 50°C for 5 minutes (D). Buffer N contained 10 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$ and 50 mM KCl; Buffer SD contained 10 mM Tris-HCl pH 7.5, 50 mM KCl and 1 mM EDTA. Buffer D contained 7 M urea, 1 mM EDTA and 25 mM sodiumacetate. Nuclease S1 reactions at pH 7.5 were performed with buffer N supplemented with 2.5 mM $ZnCl_2$, and nuclease S1 reactions at pH 4.5 were performed in a buffer containing 50 mM sodiumacetate, pH 4.5, 10 mM $MgCl_2$, 50 mM KCl and 1 mM $ZnCl_2$.

Chemical modification

All chemical modification experiments were performed at least three times to obtain consistent data. Concentrations of chemicals were optimized to obtain single hit conditions. Chemical modifications were performed on 3'-end-labeled, 5'-end-labeled, and on unlabeled RNA. The RNAs were modified under native, semi-denaturing and denaturing conditions. Modification reactions were essentially carried out as described (18). In the primer-extension method, chemical modifications were performed using 0.3-0.5 μg unlabeled RNA, while in the reactions with end-labeled RNAs 3 x 10^4 cpm was used.

Buffers: Buffer 1: 200 mM HEPES pH 8.0, 10 mM $MgCl_2$, 50 mM KCl. Buffer II: 200 mM HEPES pH 8.0, 1 mM EDTA. Buffer III: 50 mM Na-borate pH 8.0, 10 mM $MgCl_2$, 50 mM KCl. Buffer IV: 50 mM Na-borate pH=8.0, 1 mM EDTA. Buffer V: 80 mM cacodylate pH 7.0, 100 mM KCl, 10 mM $MgCl_2$. Buffer VI: 80

mM cacodylate pH 7.0, 1 mM EDTA.

Dimethylsulfate (DMS) treatment: 0.5 - 2 μl DMS was added to the sample in 200 μl of Buffer I (native conditions) or Buffer II (semi-denaturing conditions); incubation 5 min at 30°C. Under denaturing conditions 0.5-2.0 μl DMS was used in 300 μl Buffer II; incubation 1 min at 90°C. Reactions were stopped by ethanol precipitation with 10 μg carrier tRNA. For modification of N3-C using end-labeled hY1 RNA different amounts of DMS were added to $2x10^4$ cpm of hY1 RNA in 200 μl of Buffer I (N) of Buffer II (SD, D); incubation for 5 min at 30°C, after which the reaction was stopped by ethanol precipitation. After this, a hydrazine-aniline treatment was carried out (18) to produce strand scission at the site of the modification.

CMCT treatment: A 42 mg/ml solution of CMCT (1-cyclohexyl-3-(2-morpholino ethyl)-carbodiimide metho-p-toluene sulfonate; Merck) was used. Under N and SD conditions, 50 μl CMCT was added to the sample in 150 μl Buffer III (N) or IV (SD) for a number of different incubation times at 30°C. Under D conditions, 5-25 μl CMCT was added to the sample in 150 μl of Buffer IV; incubation 1 min 90°C. Reactions were stopped by ethanol precipitation.

Kethoxal treatment: A solution of kethoxal (20 mg/ml in 20% ethanol) was used and 1-5 μl of this solution was added to the sample in 50 μl Buffer V (N) or VI (SD); incubation 10 min at 30°C; under D conditions 0.5-2 μl kethoxal was added to the sample in 50 μl Buffer VI and incubated for 1 min at 90°C.

Primer extension analysis

Primer extension was carried out essentially as described (18). Oligodeoxyribonucleotide primers 5'- CTAAGCTTAAAAGACTAGTCAAGTG-CAGT-3' and 5'-CTAAGCTTAAAACACGAAGCTAGTCAA-3', complementary to nucleotides 93-112 and 66-84 in hY1 and hY5 RNA, respectively, were 5'-end-labeled. Annealing was performed by dissolving the modified RNA template in 2 μl H_2O containing 10 μg tRNA and $5x10^4$ cpm of labeled primer, heating at 90°C for 1 min, incubating on ice for 1 min and returning to room temperature for 10 min. Extensions were achieved by adding 3 μl of a reverse transcription mix containing one unit of AMV reverse transcriptase (Boehringer) in 5 mM Tris-HCl, pH 8.0, 7 mM $MgCl_2$, 50 mM KCl, 5 mM DTT, 170 μM dNTPs and incubation at

37°C for 45 min. Reactions were stopped by adding 20 μl stopbuffer (50 mM Tris-HCl pH 8.3, 75 mM EDTA, 0.5% SDS). The RNA was hydrolyzed by adding 3 μl 3M KOH, followed by incubation at 90°C (3 min) and 37°C (1 hr). Then 6 μl concentrated acetic acid was added and the DNA fragments were ethanol precipitated.

Sequencing ladders of unmodified RNA were prepared by adding a dideoxynucleotide:deoxynucleotide mix (in a 1:10 ratio) to four different reverse transcriptase reactions. For obtaining a RNase T1 ladder a reaction was performed as described under enzymatic probing, followed by primer extension.

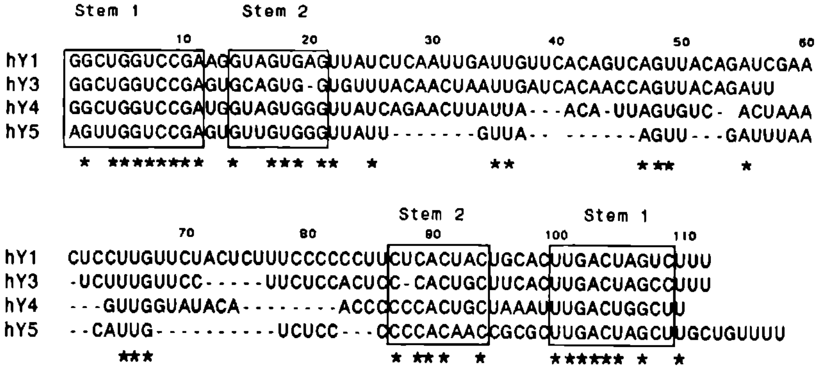
Reverse transcripts were analyzed on 10% denaturing polyacrylamide gels.

RESULTS

Sequence alignment and secondary structure prediction

Secondary structures for the hY RNAs were predicted with the programs FOLD and MFOLD (which also computes suboptimal foldings) (14, 15). Furthermore, a sequence alignment of the human Y RNAs was performed which is shown in Figure 1. Combination of the predicted structures and the sequence alignment resulted in structural models for the hY RNAs, which were subsequently tested experimentally.

Figure 4 (see page 94) shows the secondary structure models for hY1 and hY5 RNA resulting from our studies and also the proposed nomenclature for the different stems and loops. The most prominent feature of the structures obtained is the sequence conservation of the 5' and 3' terminal regions of the hY RNAs, which are proposed to base-pair and form the characteristic stem of the hY RNAs (stems 1 and 2 in Figure 4), with the bulged cytidine at the 9th position from the 5' end (C9; a C8-bulge is equally favourable from an energetic point of view; see below). Conservation also exists at the secondary structure level, showing covariation of paired residues, such that the conserved stems 1 and 2 can be formed in all hY RNAs. In this conserved part of the structure an asymmetrical internal loop with 1 nucleotide on the 5' part and 4 nucleotides on the 3' part is possible in all four hY RNAs.



```
     Stem 1        Stem 2
             10        20        30        40        50        60
hY1  GGCUGGUCCGAAGGUAGUGAGUUAUCUCAAUUGAUUGUUCACAGUCAGUUACAGAUCGAA
hY3  GGCUGGUCCGAGUGCAGUG-GUGUUUACAACUAAUUGAUCACAACCAGUUACAGAUU
hY4  GGCUGGUCCGAUGGUAGUGGGUUAUCAGAACUUAUUA---ACA-UUAGUGUC- ACUAAA
hY5  AGUUGGUCCGAGUGUUGUGGGUUAUU-------GUUA-- ------AGUU---GAUUUAA
      * ********  *  *** **  *         **          ***     *

                                Stem 2        Stem 1
             70        80        90       100       110
hY1  CUCCUUGUUCUACUCUUUCCCCCCUUCUCACUACUGCACUUGACUAGUCUUU
hY3  -UCUUUGUUCC-----UUCUCCACUCC-CACUGCUUCACUUGACUAGCCUUU
hY4  ---GUUGGUAUACA--------ACCCCCCACUGCUAAAUUUGACUGGCUU
hY5  --CAUUG----------UCUCC---CCCCACAACCGCGCUUGACUAGCUUGCUGUUUU
         ***                   * ***  *     ****** *  *
```

Figure 1. Sequence alignment of the human hY RNAs. Conserved nucleotides are indicated with asterisks, while the sequences forming stem 1 and 2, are outlined The indicated numbering is from hY1 RNA.

Structure probing strategy

For *in vitro* transcription of hY1 and hY5 RNA constructs were used (see Materials and Methods) which resulted in some additional nucleotides at the 5'- and 3'-ends of the molecules. These nucleotides, however, did not influence the predicted secondary structure, nor eliminated the binding of the Ro and La proteins (9). For hY1 RNA only, a second construct was prepared which allows the production of hY1 RNA without any additional nucleotides. HY1 RNA from both constructs behaved similarly in the enzymatic probing experiments and only the hY1 RNA without extensions was used in the chemical probing experiments.

Two RNAs, hY1 and hY5 RNA, were extensively probed both by enzymes and by chemical reagents. The enzymatic probing was performed with several RNases. Enzymes that cleave RNA when it is single-stranded are RNases A, T1, U2, T2 and nuclease S1. RNase T1 cleaves GpN bonds, RNase U2 ApN bonds, RNase A (Py)pN bonds while RNase T2 and Nuclease S1 do not exhibit known sequence specificities (19). However, for RNase U2, the sequence specificity is limited at physiological pH (19, 20), and we did not succeed in obtaining satisfactory cleavage with this RNase under native conditions. RNase V1 was used to detect double-stranded or stacked regions.

In the chemical probing experiments, the Watson-Crick positions of the bases in hY1 and hY5 RNA were modified with three base-specific chemicals, DMS, CMCT and kethoxal. DMS modifies the N1 atom in adenosines (N1-A) and (more slowly) N3-C, CMCT reacts with N3-U and (more slowly) with N1-G, and kethoxal reacts with N1-G and N2-G (both N1 and N2 are required for the kethoxal reaction). All these positions are unreactive when base-pairing involving the atoms at Watson-Crick positions occurs, except in the case of a G-U base-pair, where N2-G is accessible.

Both the enzymatic and chemical probing experiments were performed under native conditions (N) (in the presence of magnesium), semi-denaturing conditions (SD) (in the presence of EDTA) and denaturing conditions (D) (high temperature, in the presence of EDTA). Tertiary interactions are generally less stable than Watson-Crick interactions and are expected to melt under semi-denaturing conditions (18). Semi-denaturing conditions also give information about the stability of the different helical domains in an RNA molecule.

Chemically modified nucleotides were detected by primer extension analysis: a modified nucleotide causes reverse transcriptase (RT) to stop at the nucleotide immediately 3' to the modification and the reverse transcriptase products are subsequently analyzed on a denaturing polyacrylamide gel. Furthermore, the N3 positions of cytosines in hY1 RNA were also probed with DMS using end-labeled RNA, in this way allowing direct detection of the modifications.

Control incubations, in which the reagent was omitted, were always performed in parallel to detect spontaneous pyrimidine-purine breaks - which easily occur in RNA (18, 21), and, in the case of the primer extension method, to detect stops of RT. RT-stops reflect the tendency of RT to stop or pause at particular structural elements in the RNA (21).

The structure of hY1 RNA

Figures 2A through 2D show examples of the enzymatic and chemical probing results for hY1 RNA, while Figure 4A summarizes the results of several independent probing experiments.

Stems 1 and 2 and internal loop 1. The formation of the (conserved) stems 1 and 2 is clearly substantiated by RNase V1 cleavages (see Figure 2A) and by the chemical probing data, although the region around nucleotide 88 to 90 may be somewhat less stable. The bulged C9 was moderately cleaved by single-strand-specific RNase

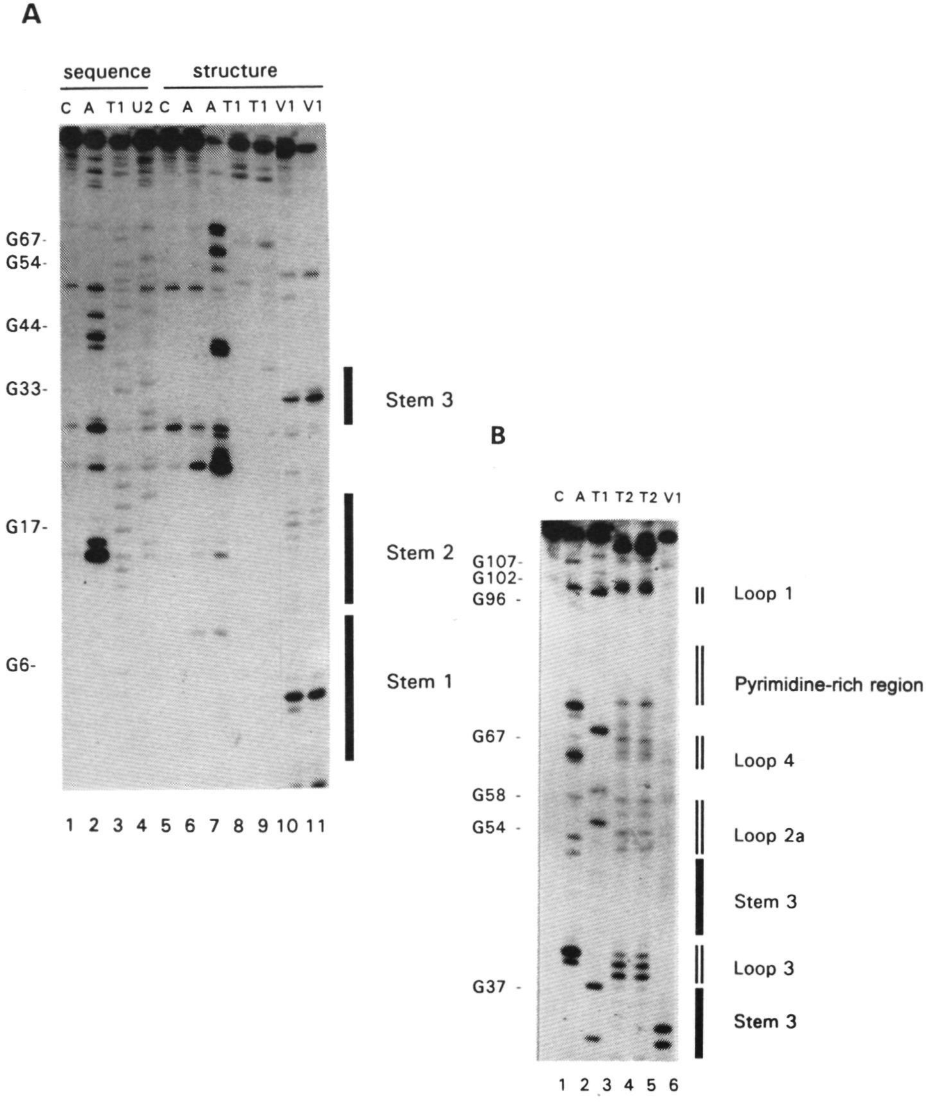


Figure 2. Structure probing of hY1 RNA. **(A)** Enzymatic probing of hY1 RNA. C: control lanes, in which no enzyme was added. Denaturing conditions: lanes 1 to 4. Lane 2: RNase A (6 x 10^{-6} U). Lane 3: RNase T1 (0.13 U). Lane 4: RNase U2 (0.08 U). Native conditions: lanes 5 to 11. Lanes 6 and 7: RNase A (1.5 x 10^{-5} and 4.5 x 10^{-5} U). Lanes 8 and 9: RNase T1 (0.13 and 0.5 U). Lanes 10 and 11: RNase V1 (0.08 and 0.3 U). Note that products of RNase V1 digestion run one base more slowly than those in the other lanes due to the absence of a 3' phosphate (33,34). **(B)** Enzymatic probing of hY1 RNA. Native conditions: lanes 1 to 6. Lane 2: RNase A (6 x 10^{-6} U). Lane 3: RNase T1 (0.5 U). Lanes 4 and 5: RNase T2 (0.1 and 0.05 U). Lane 6: RNase V1 (0.08 U). *(Figure continued on next page).*

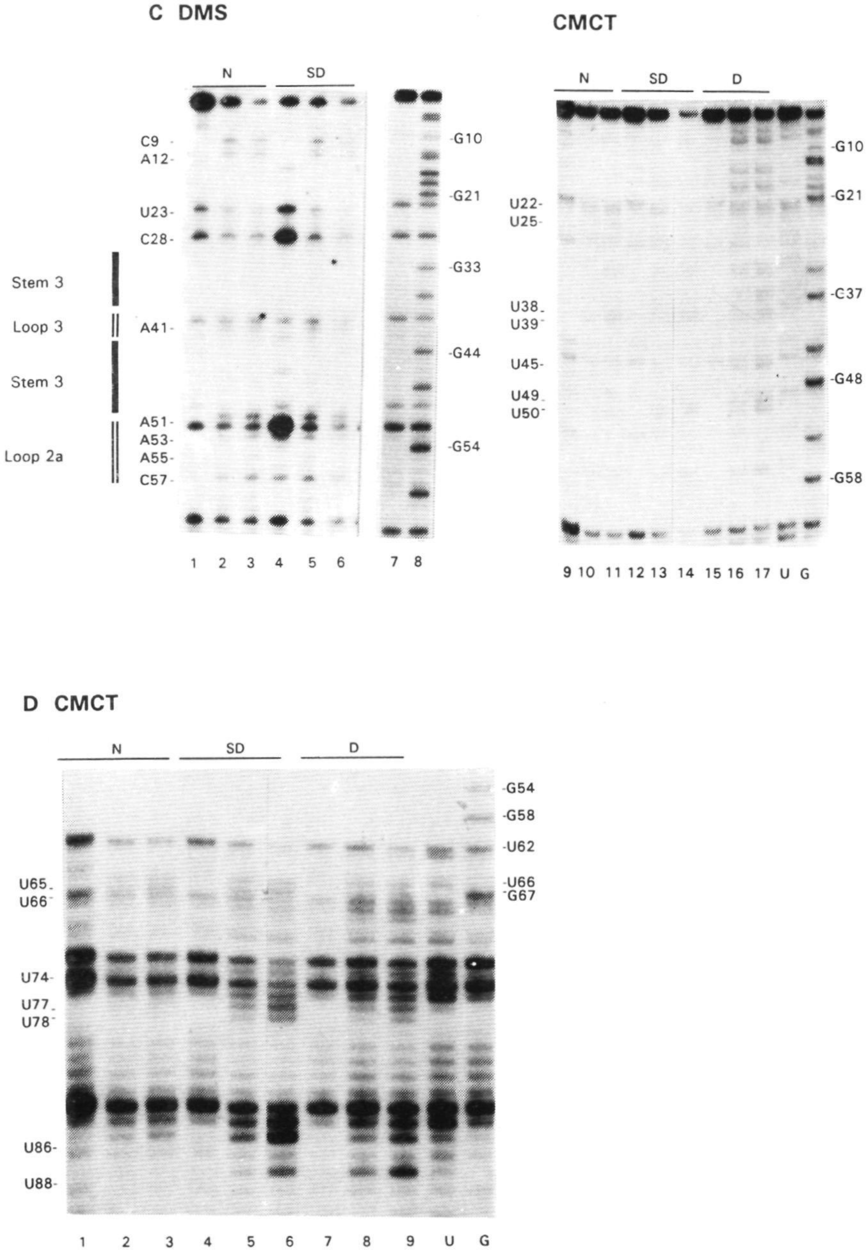

C DMS
N
SD
C9
A12
U23
C28
Stem 3
Loop 3
A41
Stem 3
A51
A53
A55
Loop 2a
C57
G10
G21
G33
G44
G54
1 2 3 4 5 6 7 8
CMCT
N
SD
D
U22
U25
U38
U39
U45
U49
U50
G10
G21
C37
G48
G58
9 10 11 12 13 14 15 16 17 U G
D CMCT
N
SD
D
G54
G58
U62
U66
G67
U65
U66
U74
U77
U78
U86
U88
1 2 3 4 5 6 7 8 9 U G

Figure 2 (continued) (C) Chemical probmg of hY1 RNA wnth DMS and CMCT DMS modificatıons are shown ın lanes 1 to 6 and CMCT modıficatıons ın lanes 9 to 17 Samples ın lanes 1, 4, 7, 9, 12, 15 are control ıncubatıons where reagent was omıtted. The reactıon condıtıons are ındıcated above the figures· N (natıve condıtıons), SD (semı-denaturıng condıtıons), D (denaturıng condıtıons). Lanes 2 and 5: 5 μl DMS ıncubated for 15 mınutes Lanes 3 and 6: 10 μl DMS ıncubated for 15 mınutes. Lanes 10, 11, 13 and 14: 50 μl CMCT ıncubated for 20, 30, 5 and 10 mınutes, respectıvely. Lanes 16 and 17: 10 and 25 μl CMCT, ıncubated for 1 mınute Lane 8. kethoxal modıficatıon under D condıtıons to generate a sequence ladder. U,G: dıdeoxy sequencıng lanes, ındıcated nucleotıdes are converted ınto hY1 RNA sequence (D) Chemıcal probıng of hY1 RNA wıth CMCT. Samples ın lanes 1, 4 and 7 are control ıncubatıons where reagent was omıtted. Lanes 2, 3, 5 and 6. 50 μl CMCT ıncubated for 20, 30, 5 and 10 mınutes, respectıvely. Lanes 8 and 9. 10 and 25 μl CMCT, ıncubated for 1 mınute. U,G dıdeoxy sequencıng lanes

A, while also some RNase V1 cleavage was found in this region. Both by primer extension (Figure 2C) and by using end-labeled hY1 RNA (data not shown) the N3-C position of C9 was shown to be accessible to DMS under native conditions. Moderate C8 modification was only detected using end-labeled RNA (data not shown), which suggests an equilibrium between C9-G102 and C8-G102 pairing, with C8-G102 pairıng in the majority of the molecules. A12, located in internal loop 1, and A11, located in an A-U pair bordering this loop, both show an accessible N1 atom (see Figure 2C). Nucleotides 96-99 were shown to be single-stranded, consistent with their localization in internal loop 1 (see Figure 2B).

The 3' oligo-U stretch, the binding site of the La protein, was found single-stranded since RNase T2 efficiently trimmed the full-length hY1 RNA to a length corresponding to 109 nucleotides (Figure 2B). Because we used 5'-end-labeled RNA in this experiment the RNA must have lost the oligo-U stretch at the 3' end confirming that the oligo-U stretch of hY1 RNA is single-stranded.

Stem-loops 3 and 4. The formation of stem 3 was confirmed by both enzymatic (Figure 2A and 2B) and chemical probing (Figure 2C), although this stem appeared to be breathing. It contains mainly A-U and G-U pairs, and thus is less stable than G-C rich stems. Indeed, several bases in this region were reactive under semi-denaturing conditions and some even under native conditions. Stem 4 appeared to be even less stable than stem 3, with sometimes weak RNase V1 cleavages at nucleotides 58 and 59. The chemical probing results also indicate that stem 4 exists, but is relatively unstable.

Loop 3 and loop 4 were clearly single-stranded, as shown by the absence of RNase V1 cleavages, the presence of RNase A and T2 cleavages (Figure 2A and 2B) and by the chemical probing data (Figures 2C and 2D).

Internal loop 2. Loop 2a (see Figure 4A), which was in a former secondary structure model predicted to form a stem (base-pairing of nts 24-27 with nts 53-56), was efficiently cleaved by the single-strand specific enzymes (Figure 2B). Furthermore, the chemical modification data show that almost all bases were fully reactive at their Watson-Crick positions under native conditions (Figure 2C). Therefore, it can be concluded that this part of the structure is single-stranded under the conditions used for probing and is part of a large internal loop (loop 2), as is shown in Figure 4A. In some experiments, weak RNase V1 cleavage was found between nt G21 and U22. These cleavages are probably due to base stacking in this region of the loop.

Remarkably, the large pyrimidine-rich region in hY1 RNA (loop 2b) could not be cleaved at all by the enzymes (see Figures 2A and 2B), not even under denaturing conditions (urea and 50°C), although in some experiments this region did show a smear under semi-denaturing conditions (data not shown). Reactions with nuclease S1 at low (pH=4.5) and neutral pH (pH=7.5), did not result in cleavages in the pyrimidine-rich region (data not shown). In the chemical modification reactions with DMS and CMCT, most of the nucleotides in the pyrimidine-rich loop could be modified under SD conditions, but not under native conditions. In Figure 2D this is shown for the CMCT reaction. The behaviour of the pyrimidine-rich region may point to the existence of long-range interactions in the tertiary structure of the molecule (see Discussion). Many RT-stops in this pyrimidine-rich region were reproducibly found, but DMS modifications on end-labeled hY1 RNA (data not shown), allowing the probing of N3-C positions, facilitated the deduction of information on these cytosines.

The structure of hY5 RNA

Figures 3A through 3D show examples of the enzymatic and chemical probing results for hY5 RNA, while Figure 4B summarizes the results of several independent experiments.

Stems 1 and 2 and internal loop 1. The formation of the conserved stem 1 was confirmed both by the chemical and enzymatic probing data. Strong RNase V1 cuts were only found near residues C9 and G10 (Figure 3A), probably indicating that C9 is stacked in the helix, whereas weak RNase V1 cleavages were found in other parts of stem 1. As in hY1 RNA, a weak RNase A cleavage was found at C9. Because the primer extension method showed an RT-stop at C9, no definite

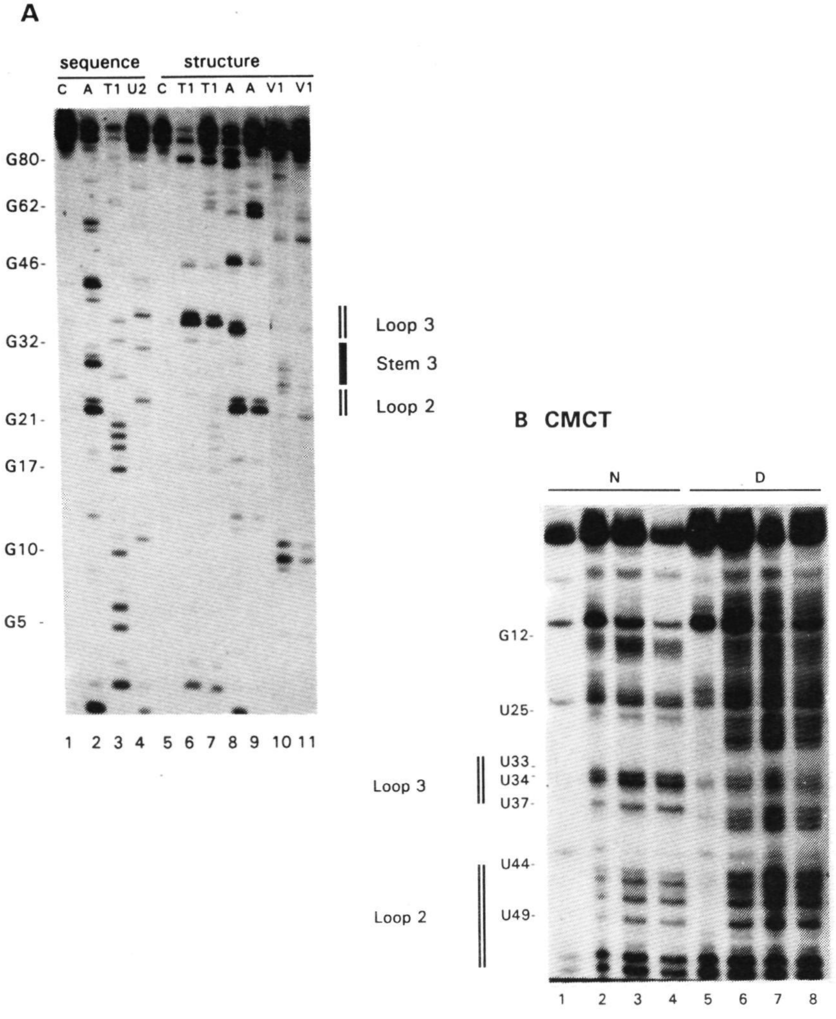


Figure 3. Structure probing of hY5 RNA. **(A)** Enzymatic probing of hY5 RNA. C: control lane, in which no enzyme was added. Denaturing conditions: lanes 1 to 4. Lane 2: RNase A (6 x 10^{-6} U). Lane 3: RNase T1 (0.13 U). Lane 4: RNase U2 (0.08 U). Native conditions: lanes 5 to 11. Lanes 6 and 7: RNase T1 (0.15 and 0.05 U). Lanes 8 and 9: RNase A (5 x 10^{-6} and 2 x 10^{-6} U). Lanes 10 and 11: RNase V1 (0.06 U). **(B)** Chemical probing of hY5 RNA with CMCT. Lanes 1 and 5: control lanes. The reaction conditions are indicated above the figures. Lanes 2, 3 and 4: 50 μl CMCT incubated for 10, 20 and 30 minutes, respectively. Lanes 6, 7 and 8: 10, 25 and 40 μl CMCT incubated for 1 minute. *(Figure continued on next page).*

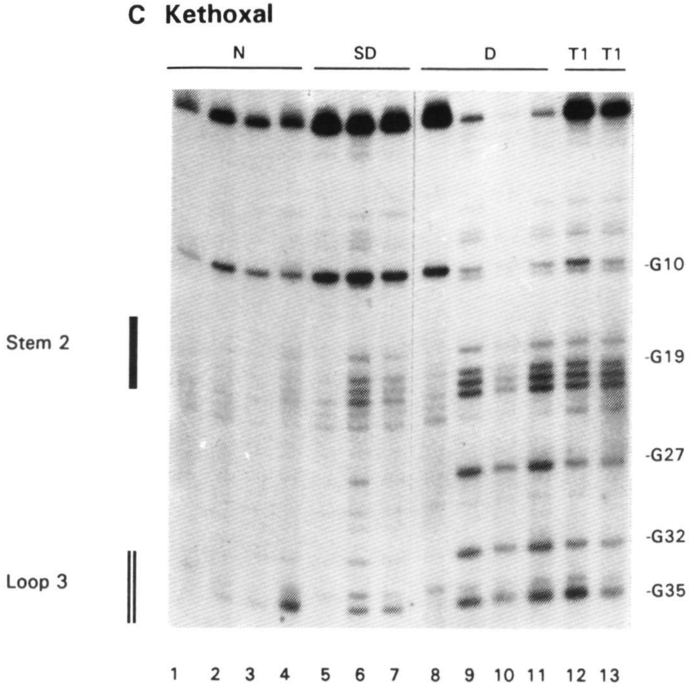


Figure 3 (Continued) (C) Chemical probing of hY5 RNA with kethoxal. Lanes 1, 5, 8: control lanes. The reaction conditions are indicated above the figures. Lanes 2, 3 and 4: 1, 2 and 5 μl kethoxal, incubated for 10 minutes. Lanes 6 and 7: 1 and 2 μl kethoxal incubated for 10 minutes. Lanes 9, 10 and 11: 0.5, 1 and 2 μl kethoxal, incubated for 1 minute. Lanes 12 and 13: RNase T1 ladder. *(Figure continued on next page).*

conclusion can be drawn whether C9 is looping out or is stacked inside the helix. The reactivity of the G's in stem 1 under SD conditions was only seen in the kethoxal and not in CMCT experiments (Figure 3C; see also Discussion).

Stem 2 contained RNase V1 cleavages, confirming its double-stranded nature, although the lower part of this stem appeared to be breathing. U13 is accessible to CMCT (Figure 3B), which is comparable to the situation in hY1 RNA where A12 is accessible to DMS. G12 and G14 are accessible to CMCT under native conditions (see Figure 3B) and the base-pairing of G10 through G12 with C65 through U67 appeared to be of intermediate stability. In comparison with the low-energy secondary structure predictions, our probing data indicated that loop 1 is located at a different position (nucleotides 13 and 61-64 rather than nucleotides 16

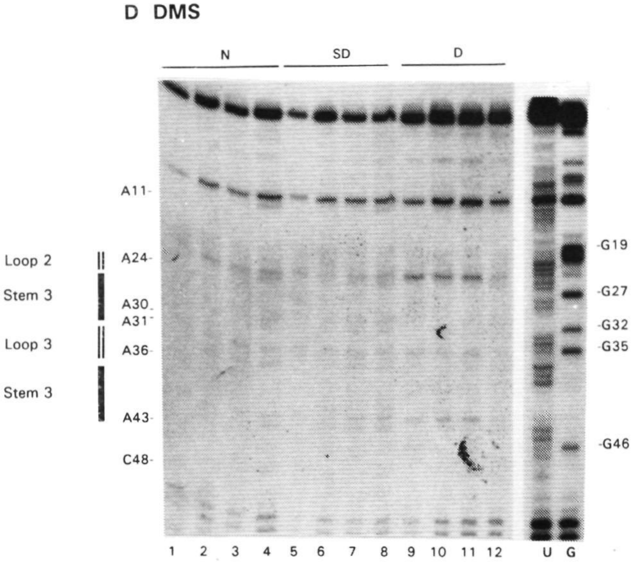


Figure 3 (Continued) (D) Chemical probing of hY5 RNA with DMS. Lanes 1, 5, 9: control lanes. The reaction conditions are indicated above the figures. Lanes 2, 3 and 4: 0.5, 1 and 2 µl DMS incubated for 10 minutes. Lanes 6, 7 and 8: 0.5, 1 and 2 µl DMS incubated for 10 minutes. Lanes 10, 11 and 12: 0.5, 1 and 2 µl DMS incubated for 1 minute.

and 58-61), which also is in better agreement with the alignment data (Figure 1).

Regarding the 3' oligo-U stretch, the binding site of the La protein, always two bands were seen in the enzymatic probing experiments; one with a length corresponding to the full-length hY5 RNA, the other containing some additional nucleotides (see Figure 3A) as described in the Material and Methods section. This probably indicates that the additional nucleotides are not stable and are probably removed during the incubation, even in the absence of any enzyme. Several phosphodiester bonds in the 3' oligo-U stretch are cleaved by single-strand-specific RNases corroborating the notion that it is single-stranded.

Stem-loop 3. The hairpin formed by stem 3 and loop 3 was shown to exist in solution by both enzymatic and chemical probing. However, the stem consists almost completely of A-U base-pairs and appears to be not very stable. Most of the adenosines are moderately reactive with DMS under native conditions, while the complementary uridines were reactive to CMCT only under semi-denaturing and denaturing conditions (Figures 3B and 3D). This difference is probably related to the fact that the N1-A is modified by the relatively small DMS reagent (M_r=126), while the N3-U on the opposite strand cannot be modified by the more bulky CMCT reagent (M_r=423). A similar phenomenon has been observed frequently in other RNAs (18, 21-23).

The nucleotides of loop 3 are accessible under native conditions for both single-strand-specific enzymes and chemical reagents (Figures 3A through 3D). A differential reactivity was observed for G32 and G35 which were modified by kethoxal but not by CMCT (see Discussion).

Internal loop 2. In the predicted secondary structure in this region, two base-pairs exist, U23-G46 and A24-U45, which separate the asymmetrical internal loop 2 into a smaller internal loop and a mismatch U25-U44. However, both the enzymatic and chemical probing data show that only one 'large' internal loop is present, of which nearly all the nucleotides are accessible at their Watson-Crick positions. See for example the CMCT probing results in Figure 3B. In contrast to the large pyrimidine-rich region in hY1 RNA, the pyrimidine-rich region in hY5 RNA is fully accessible. Some RNase V1 cleavages were detected at nucleotides 43-45, probably due to stacking tertiary interactions of these bases (see Discussion).

DISCUSSION

The utilization of structure-specific probes allowed us to map the hY RNA conformation in detail and provided experimental evidence for the secondary structures of hY1 and hY5 RNA. The secondary structures obtained largely correspond to the structures predicted by combining free-energy minimizations and alignment data, although some important deviations could be observed.

Secondary Structure

The conserved stems 1 and 2 do exist, both in hY1 and hY5 RNA. The internal loop 1 containing one nucleotide in the 5' part and four nucleotides in the 3' part is present in both RNAs at the position predicted after combining alignment data and low-energy predictions. U13 in hY5 RNA and A12 in hY1 RNA are located outside the helix and the bordering nucleotides (G12 and G14 in hY5 RNA and A11 in hY1 RNA) showed enhanced reactivities, which agrees with their location at the end of helical regions bordering an internal loop.

The bulged nucleotide C9 in both hY1 and hY5 RNA showed some RNase V1 reactivity, which might result from stacking of the cytosine in the helix, but also some RNase A cleavage was observed. In the chemical modifications the N3 position of C9 could be shown to be modified by DMS in hY1 RNA. Taken together, the behaviour of C9 indicates that the base is probably looping out in the majority of the molecules, while in a small percentage of the molecules base-pairing of C9-G102 occurs and C8 is bulging out. Similar equilibria, although not involving G-C pairs, have been observed in other RNAs as well (24, 25). Previous experiments have demonstrated that the identity of the base at position 9 is important for the recognition by Ro60 (9). Possibly Ro60 selects one of the alternative structures in this region and stabilizes its conformation during binding.

Our results also suggest that the base-pairs which separate C9 and the internal loop 1 are not very stable. This may point to the existence of an equilibrium between two structures, one in which there is a bulge and an internal loop (loop 1) and a second structure in which there is a larger internal loop. Both alternatives are also possible when thermodynamic data are considered.

The La-binding site, the 3' oligo-U stretch, is single-stranded, as was expected because the La protein can bind to both Y RNA constructs (9).

A pyrimidine-rich region of different length is present in internal loop 2 of both hY1 and hY5 RNA. In hY1 RNA the structure of this region was difficult to assess. Attempts to demonstrate that the loop 2b region in hY1 RNA is single-stranded by hybridization with an antisense oligonucleotide followed by RNase H cleavage were not successful (data not shown). However, the pyrimidine-rich region was accessible to modifying reagents under SD conditions, which indicates that it is single-stranded under these conditions but not under native conditions (see also below).

In contrast, in hY5 RNA, the pyrimidine-rich region is clearly single-stranded under native conditions. A differential reactivity with kethoxal on one hand and CMCT on the other hand was found for several guanosines in hY5 RNA. For example, G46 was reactive with kethoxal but not with CMCT. This may be explained by stacking of these bases, but alternative explanations could be that CMCT reacts more slowly with G's than kethoxal or the fact that CMCT is larger than kethoxal and therefore cannot approach the nucleotide. Similar differences in reactivity of a single base with CMCT and kethoxal have been observed before (26, 27).

Stem 3 and 4 in hY1 RNA and stem 3 in hY5 RNA are formed but in all cases these stems are relatively unstable and are breathing.

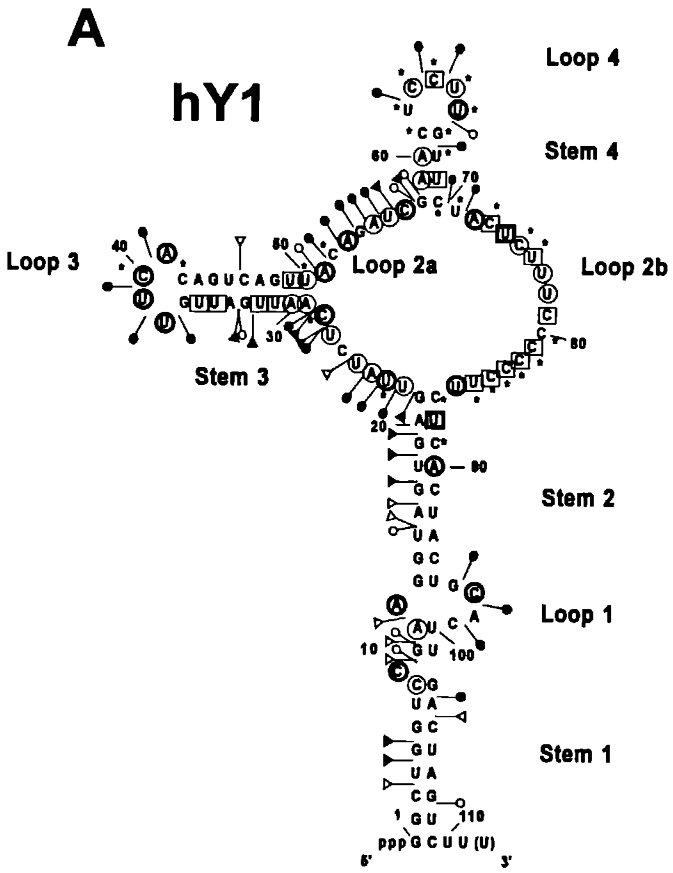
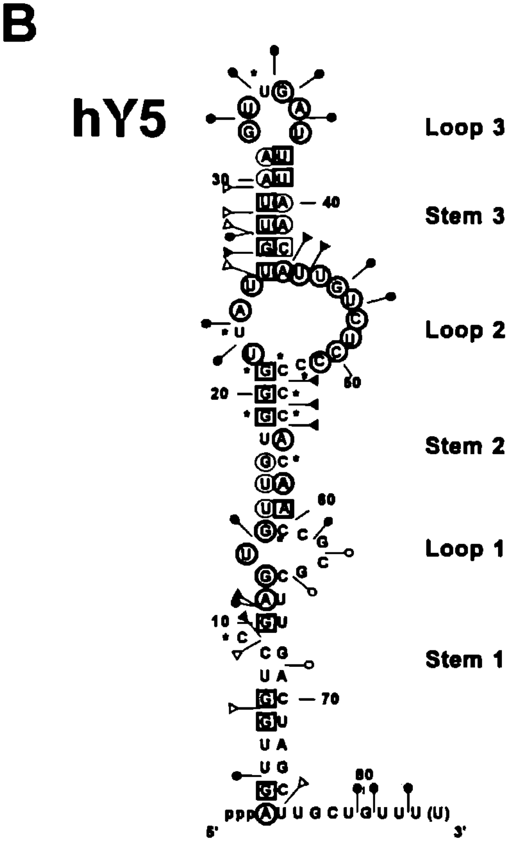


Figure 4. Summary of the structure probmg of hY1 and hY5 RNA (A) Reactivities of nucleotides in hY1 RNA towards chemical and enzymatic probes Reactivity towards chemical probes. Reactivity under native conditions is indicated by bold (strong reactivity) and light (moderate reactivity) circles around the nucleotides Nucleotides which are unreactive under native conditions but reactive under semi-denaturing conditions are indicated by bold (strong reactivity) and light (moderate reactivity) squares around the nucleotide Asterisks indicate natural stops of reverse transcriptase. Reactivity towards single-strand-specific RNases (T1, A, T2) under native conditions is indicated by small solid circle (strong reactivity) and small open circle (moderate reactivity). RNase V1 cleavage is shown with solid (strong reactivity) and open (moderate reactivity) triangles, respectively (B) (opposite page) Reactivities of nucleotides in hY5 RNA towards chemical and enzymatic probes. Symbols are identical to those in (A).



Comparison with other Y RNA structures

In Figure 5, the secondary structures of all four hY RNAs are shown. For hY3 and hY4 RNAs we also performed some preliminary enzymatic probing experiments (unpublished data), and the data obtained support the structures depicted in Figure 5. However, in both hY3 and hY4 RNA the pyrimidine-rich regions, which form the 3' part of an internal loop, could not be efficiently probed by enzymes, analogous to the situation in hY1 RNA. The 5' part of the internal loop showed both single-stranded behaviour and reactivity with RNase V1, which is similar to what was observed for hY1 RNA.

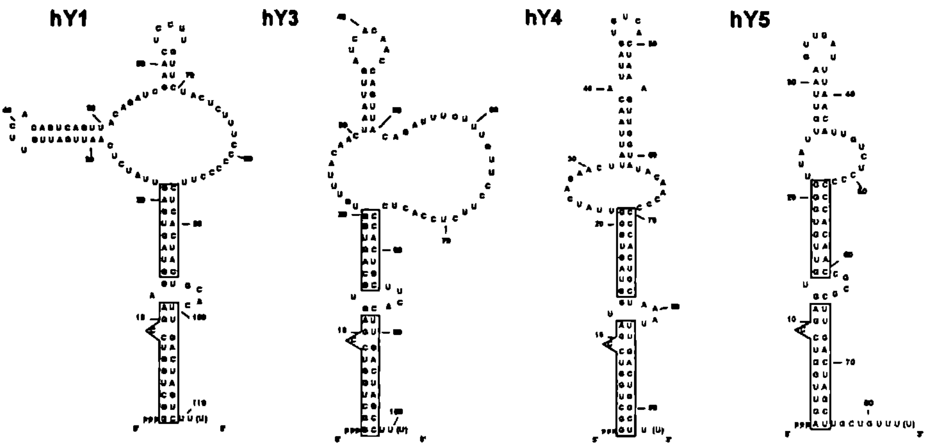


Figure 5. Secondary structure models for the human Y RNAs obtained by combining predicted structures and the obtained probing results.

Recently, *Xenopus laevis* Y RNAs were isolated and sequenced (5). *Xenopus laevis* cells contain four distinct Y RNAs, three of which (xY3, xY4 and xY5 RNA) show sequence similarity to hY3, hY4 and hY5 RNA, respectively. Surprisingly, no homologue for hY1 RNA was found. Instead another Y RNA (called xYα) was characterized, which does not appear to represent the *Xenopus laevis* counterpart of one of the hY RNAs. All the Xenopus Y RNAs are predicted to form a secondary structure similar to that of the hY RNAs (Figure 5). In all xY RNAs stem 1, including a bulged C9 residue, can be formed. In the conserved stem 2, there is a mismatch in xY4 and a bulged residue in xYα, while the other two xY RNAs apparently contain a 'perfect' stem 2. All xY RNAs also contain a pyrimidine-rich region, predicted to be located in an internal loop, as has been found for the hY RNAs.

Possible tertiary interactions

Bases that are reactive with chemical reagents under semi-denaturing conditions but protected under native conditions are most likely involved in tertiary interactions. An alternative explanation is that these nucleotides are stacked under native conditions. Nucleotides in interhelical regions (like internal loops) in RNAs are known to show all kinds of behaviour in structure probing experiments, ranging from high reactivity to complete protection (28). Examples of this type of behaviour are found in the adenovirus-associated RNAs (VA RNAs) and Epstein-Barr virus RNAs (EBER RNAs) (29, 30). These RNAs show secondary structure similarity to the hY RNAs, and can also bind the La protein. EBER-2 RNA contains an internal loop of 20 nucleotides (22), most of which are accessible for modification under native conditions. However, in the case of VA I RNA, several nucleotides in a large internal loop, called the central domain, are resistant to single-strand-specific RNases. This region, which is important for the recognition of VA RNA by the double-stranded RNA dependent protein kinase PKR (29), shows some RNase V1 cleavages, and is believed to contain an alternative structure in which tertiary interactions play a significant role. Another example is mouse RNase MRP RNA, which contains a large single-stranded loop of which only one third of the nucleotides are accessible at their Watson-Crick positions under native conditions (31).

In hY1 RNA, the behaviour of the nucleotides in the pyrimidine-rich region suggests that either tertiary interactions under native conditions block accessibility or that this region under native conditions adopts a strange, yet unidentified, structure, in which Watson-Crick positions are inaccessible. Furthermore, RNase V1 cleavage in the 5' part of loop 2 was observed which may also be explained by tertiary interactions in this region or alternatively by base stacking. In conclusion, our results suggest that an intrinsic tertiary structure involving loop 2 is present in hY1 RNA. It is possible that this tertiary folding is determined by non-Watson-Crick interactions. Examples of such interactions have been described in 5S rRNA, and in the Rev Responsive Element (RRE) of HIV-1 RNAs (32), where in both cases these unusual regions form the binding site of a protein.

In contrast to the situation in hY1 RNA, the pyrimidine-rich region in hY5 RNA appeared to be fully accessible at the Watson-Crick positions. Under native conditions some RNase V1 cleavages were found (nt 43-45), suggesting that to some extent base-pairing or stacking occurs. Base-pairing of U23 and A24 with G46 and U45, respectively, as was the case in the predicted secondary structure for hY5 RNA, would explain these cleavages and thus might be present temporarily or in a subset of the molecules.

Nearly all the human and *Xenopus* Y RNAs have limited base-pairing possibilities (ranging from 2 to 4 base-pairs, mostly G-U and A-U) between the 5' part and 3' part of the large internal loop. It is possible that these potential interactions, which may be stabilized by protein binding, are present when the Y RNAs exert their yet unknown function, while at another time the internal loop is open.

Knowledge about the structure of the Y RNAs is an important step towards understanding more about the function of the Y RNAs and Ro RNPs. Structure probing experiments performed on the Ro RNP particles are required to determine if the RNA structure changes when the RNA is bound by protein and to determine exactly which nucleotides are involved in protein binding

ACKNOWLEDGEMENTS

The research was supported by the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organization for Scientific Research (NWO) The research of Dr Pruijn has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences Use of the services and facilities of the Dutch CAOS/CAMM Center, under grant numbers SON 326-052 and STW NCH99 1751, is gratefully acknowledged

REFERENCES

1 Van Venrooij, W J , Slobbe, R L and Pruijn, G J M (1993) *Mol Biol Rep*, 18, 113 119

2 Peek, R , Pruijn, G J M , van der Kemp, A J W and Van Venrooij, W J (1993) *J Cell Sci.*, 106, 929 935

3 Pruijn, G J M , Slobbe, R L and Van Venrooij, W J (1990) *Mol Biol Rep*, 14, 43 48

4 Pruijn, G J M , Wingens, P A E T M , Peters, S L M , Thijssen, J P H and Van Venrooij, W J (1993) *Biochim Biophys Acta*, 1216, 395-401

5 O'Brien, C A , Margelot, K and Wolin, S L (1993) *Proc Natl Acad Sci. U S A*, 90, 7250-7254

6 Hendrick, J P , Wolin, S L , Rinke, J , Lerner, M R and Steitz, J A (1981) *Mol. Cell. Biol.*, 1, 1138 1149

7 Kato, N , Hoshino, H and Harada, F (1982) *Biochem Biophys Res. Commun.*, 108, 363 370

8 Wolin, S L and Steitz, J A (1984) *Proc Natl. Acad Sci. U S A*, 81, 1996-2000

9 Pruijn, G J M , Slobbe, R L and Van Venrooij, W J (1991) *Nucl. Acids Res*, 19, 5173-5180

10 Stefano, J E (1984) *Cell*, 36, 145 154

11 Wolin, S L and Steitz, J A (1983) *Cell*, 32, 735-744

12 O'Brien, C A and Harley, J B (1990) *EMBO J*, 9, 3683 3689

13 Devereux, J , Haeberli, P and Smithies, O (1984) *Nucl. Acids Res.*, 12, 387-395

14 Zuker, M , Jaeger, J A and Turner, D H (1991) *Nucl. Acids Res*, 19, 2707-2714

15 Jaeger, J A , Turner, D H and Zuker, M (1990) *Methods Enzymol*, 183, 281-303

16 Scherly, D , Boelens, W , Van Venrooij, W J , Dathan, N A , Hamm, J and Mattaj, I W (1989) *EMBO J*, 8, 4163 4170

17 Van Gelder, C W G , Gunderson, S I , Jansen, E J R , Boelens, W C , Polycarpou-Schwarz, M , Mattaj, I W and Van Venrooij, W J (1993) *EMBO J*, 12, 5191-5200

18 Krol, A and Carbon, P (1989) *Methods Enzymol.*, 180, 212-227

19 Knapp, G (1989) *Methods Enzymol*, 180, 192-212

20 Ehresmann, C , Baudin, F , Mougel, M , Romby, P , Ebel, J-P and Ehresmann, B (1987) *Nucl Acids Res*, 15, 9109-9129

21 Kwakman, J H , Konings, D A , Hogeweg, P , Pel, H J and Grivell, L A (1990) *J Biomol*

Struct. Dyn., 8, 413-430.

22. Glickman, J.N., Howe, J.G. and Steitz, J.A. (1988) *J. Virol.*, 62, 902-911.
23. Krol, A., Westhof, E., Bach, M., Lührmann, R, Ebel, J.P. and Carbon, P. (1990) *Nucl. Acids Res.*, 18, 3803-3811.
24. Moine, H., Romby, P., Springer, M., Grunberg-Manago, M., Ebel, J-P., Ehresmann, C. and Ehresmann, B. (1988) *Proc. Natl. Acad. Sci. U. S. A.*, 85, 7892-7896.
25. Leal de Stevenson, I., Romby, P., Baudin, F., Brunel, C., Westhof, E., Ehresmann, C., Ehresmann, B. and Romaniuk, P.J. (1991) *J. Mol. Biol.*, 219, 243-255.
26. Brunel, C., Romby, P., Westhof, E., Ehresmann, C. and Ehresmann, B. (1991) *J. Mol. Biol.*, 221, 293-308.
27. Sturchler, C., Westhof, E., Carbon, P. and Krol, A. (1993) *Nucl. Acids Res.*, 21, 1073-1079.
28. Baudin, F., Ehresmann, C., Romby, P., Mougel, M., Colin, J., Lempereur, L., Bachellerie, J-P., Ebel, J-P. and Ehresmann, B. (1987) *Biochimie*, 69, 1081-1096.
29. Peery, T., Mellits, K.H. and Mathews, M.B. (1993) *J. Virol.*, 67, 3534-3543.
30. Clemens, M.J. (1993) *Mol. Biol. Rep.*, 17, 81-92.
31. Topper, J.N. and Clayton, D.A. (1990) *J. Biol. Chem.*, 265, 13254-13262.
32. Ellington, A.D. (1993) *Current Biology*, 3, 375-377.
33. Miller, W.A. and Silver, S.L. (1991) *Nucl. Acids Res.*, 19, 5313-5320.
34. Sinnett, D., Richer, C., Deragon, J.M. and Labuda, D. (1991) *J. Biol. Chem.*, 266, 8675-8678.

CHAPTER 4

A complex secondary structure in U1A pre-mRNA that binds two molecules of U1A protein is required for regulation of polyadenylation

EMBO Journal 12: 5191-5200 (1993)

A complex secondary structure in U1A pre-mRNA that binds two molecules of U1A protein is required for regulation of polyadenylation

Celia W.G. van Gelder, Samuel I. Gunderson,[1] Eric J.R. Jansen, Wilbert C. Boelens,[1] Maria Polycarpou-Schwarz,[1] Iain W. Mattaj[1] and Walther J. van Venrooij

University of Nijmegen, Department of Biochemistry, PO BOX 9101, 6500 HB Nijmegen, The Netherlands and [1]European Molecular Biology Laboratory, Gene Expression Programme, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

Communicated by I.W. Mattaj

The human U1A protein-U1A pre-mRNA complex and the relationship between its structure and function in inhibition of polyadenylation *in vitro* were investigated. Two molecules of U1A protein were shown to bind to a conserved region in the 3' untranslated region of U1A pre-mRNA. The secondary structure of this region was determined by a combination of theoretical prediction, phylogenetic sequence alignment, enzymatic structure probing and molecular genetics. The U1A binding sites form (part of) a complex secondary structure which is significantly different from the binding site of U1A protein on U1 snRNA. Studies with mutant pre-mRNAs showed that the integrity of much of this structure is required for both high affinity binding to U1A protein and specific inhibition of polyadenylation *in vitro*. In particular, binding of a single molecule of U1A protein to U1A pre-mRNA is not sufficient to produce efficient inhibition of polyadenylation.

Key words: polyadenylation/RNA-protein interaction/RNA structure/U1 snRNP/ U1A protein

INTRODUCTION

The removal of introns from pre-messenger RNA, known as splicing, is an important process in which several small ribonucleoprotein particles (snRNPs) participate. One of them, U1 snRNP, interacts with the pre-mRNA by a mechanism that includes pairing between bases at the 5' end of U1 snRNA and sequences located at the 5' splice site. U1 snRNPs contain at least eight proteins (B', B, D1, D2, D3, E, F and G), which also occur in other U snRNPs, and three U1-specific proteins named 70K, U1C and U1A (Lührmann *et al.*, 1990). The U1A protein binds directly to the second stem-loop of U1 snRNA (Scherly *et al.*, 1989; Lutz-Freyermuth *et al.*, 1990). The protein contains two RNP motifs, of which the N-terminal copy is responsible for binding to U1 snRNA (Scherly *et al.*, 1989; Lutz-Freyermuth *et al.*, 1990; Nagai *et al.*, 1990; Jessen *et al.*, 1991; Hall and Stump, 1992). The structure of this domain of the U1A protein has been determined by X-ray crystallography and NMR studies (Nagai *et al.*, 1990; Hoffman *et al.*, 1991) and consists of a four-stranded antiparallel ß-sheet with two α-helices lying on the same side of the sheet.

The loop of the hairpin to which U1A binds has the sequence AUUGCACUCC. It has been shown that the first seven nucleotides, AUUGCAC, which are highly conserved between U1 snRNAs from various species, are critical for specific U1A protein binding, while the structural context of this sequence affects binding affinity (Scherly *et al.*, 1989, 1990; Bentley and Keene, 1991; Tsai *et al.*, 1991). If the loop sequence of stem-loop II of U1 snRNA is present in the absence of a stable stem, the affinity for the U1A protein drops (Scherly *et al.*, 1989; Tsai *et al.*, 1991). Quantitative mobility shift assays of the loop sequence of stem-loop II, present either in a linear structural context or in a hairpin structure with a loop larger than that found in U1 snRNA, showed an ~100-fold reduction in binding affinity for the U1A protein relative to the wild type stem-loop II (Tsai *et al.*, 1991). RNase protection experiments on U1 snRNP particles showed that both the loop sequence and ~5 bp of the stem are protected by the U1A protein (Bach *et al.*, 1990). Bound U1A protein also protects several 5' stem phosphates, as well as some loop phosphates, against ethylation by ethylnitrosourea (Jessen *et al.*, 1991).

Recently it has been shown that the 3' untranslated region (UTR) of the U1A pre-mRNA contains a region which has been conserved between vertebrate species (Boelens *et al.*, 1993). This region contains two stretches of seven nucleotides, one of which is identical to the seven nucleotides of the U1 snRNA loop mentioned above, while the other is the same in six out of seven positions. These sequences will be referred to as Box 1 and Box 2 respectively in this paper, with Box 1 being the more 5' of the two. Boxes 1 and 2 are located in close proximity to the cleavage and polyadenylation signal. The distance between the two boxes is conserved, as is the distance from Box 2 to the polyadenylation signal.

It was demonstrated that binding of U1A protein to this region of the U1A pre-mRNA, which depends upon these U1 snRNA-like sequences, causes inhibition of polyadenylation of the U1A pre-mRNA (Boelens *et al.*, 1993). Although it was not determined how many molecules of U1A protein were bound to each pre-mRNA, the number was shown to be greater than one.

In the experiments reported here, the structure of the U1A-binding region of the pre-mRNA was investigated by a variety of techniques and the number of protein molecules bound was determined. Further, the structural characteristics of the U1A protein - U1A pre-mRNA complex were examined in relation to its function in inhibition of polyadenylation. The U1 snRNA-like sequences are shown to form parts of two asymmetric internal loops present in a complex secondary structure. This part of the U1A pre-mRNA is compared with stem-loop II of U1 snRNA, the other RNA structure to which U1A protein is known to bind specifically.

RESULTS

Two molecules of U1A protein bind to the pre-mRNA

It was previously shown (Boelens *et al.*, 1993) that more than one molecule of U1A protein is able to bind to each U1A pre-mRNA. To determine the exact number of bound protein molecules, we adapted an assay often used to examine DNA-protein complexes (Hope and Struhl, 1987). Two differently sized U1A protein derivatives that bind to U1 snRNA with similar affinity (Lutz-Freyermuth *et al.*, 1990; Nagai *et al.*, 1990) were produced in, and purified from, *Escherichia coli.* These were full-length U1A (Awt) protein and a fragment of U1A containing the N-terminal 101 amino acids (A101). The two proteins were allowed to bind to a region of the U1A pre-mRNA (the Ag fragment) shown to be necessary and sufficient for U1A binding (Boelens *et al.*, 1993) and the resultant complexes analyzed by native gel electrophoresis. The Ag fragment contains the human U1A pre-mRNA sequences shown in Figure 2A plus 33 nt of 3' flanking sequence from the U1A gene and 8 nt of 5' flanking sequence derived from the cloning vector.

The position of unbound Ag RNA after native gel electrophoresis is shown in lane 1 of Figure 1. Addition of either A101 or Awt protein (represented by empty and filled squares respectively) results in the appearance of two retarded complexes (lanes 2 and 4) suggestive of binding of either one or two proteins to the RNA. The differential requirement for Awt and A101 protein in complex formation was probably due to the fact that much of the Awt protein in this particular preparation was not competent in RNA binding, since other preparations of Awt exhibited greater RNA binding capacity (data not shown). No additional intermediate complexes were seen when less protein was added (data not shown) while increasing the amount of either protein resulted in disappearance of both free RNA and the lower of the two RNA-protein complexes with a concomitant increase in the upper complex (lanes 3 and 5). Next, the two U1A derivatives were mixed before RNA binding (lanes 6 and 7). In addition to the two previously

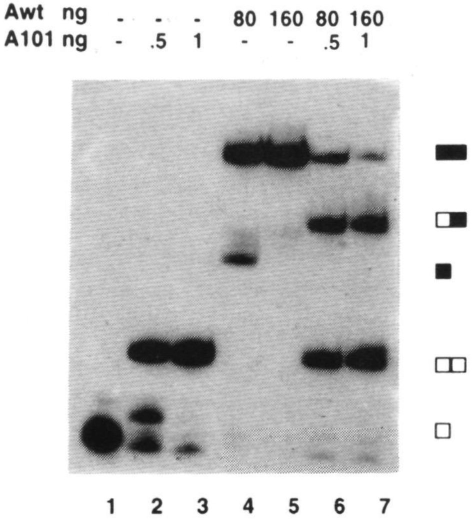


Figure 1. Two molecules of U1A protein bind to each pre-mRNA. ^{32}P-labelled Agwt RNA was incubated without U1A protein (lane 1), with A101 protein (lanes 2-3), with U1Awt protein (lanes 4-5) or with a mixture of A101 and U1Awt (lanes 6-7). The amount of protein added is indicated above the lanes. The boxes on the right represent the protein components of the complexes. Filled boxes are U1Awt and empty boxes A101.

detected, slowly migrating complexes (cf. lanes 3 and 5) a single additional complex of intermediate mobility was seen. The lack of additional intermediate complexes indicates that two, and not more, molecules of U1A bind to each RNA.

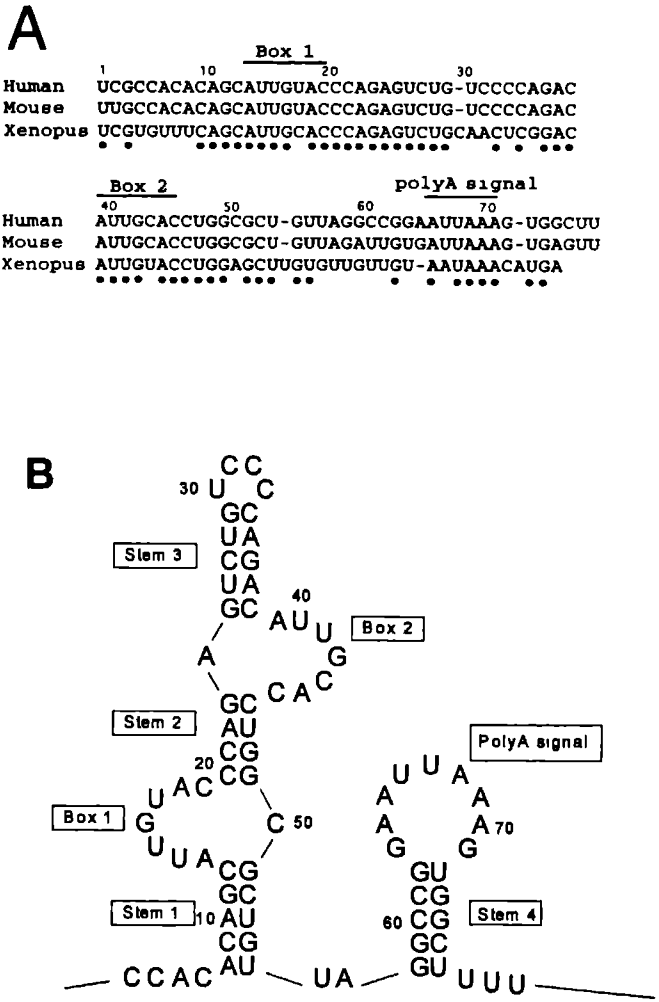
Sequence alignment and structure prediction

While the entire 3' UTRs of human and mouse U1A mRNAs are very similar (79% identical), the only region of high conservation of both with the *Xenopus* U1A mRNA sequence starts ~55 nt upstream of the A(A/U)UAAA cleavage and polyadenylation signals (Figure 2A) (Sillekens *et al.*, 1987; Scherly *et al.*, 1991; M. Bennett and J. Craft, personal communication). The entire region encompassing Boxes 1 and 2 and the cleavage and polyadenylation signal (Figure 2A) is 73% identical between human and *Xenopus* and 93% identical between human and mouse. The spacing between Box 2 and the polyadenylation signal is also identical for the three sequences. This localized sequence conservation and the fact that the Ag fragment has an affinity for U1A protein indistinguishable from that of the entire U1A pre-mRNA (see below), led us to expect that the Ag fragment would fold similarly either alone or in the context of the complete pre-mRNA.

Optimal and suboptimal foldings were calculated for the complete human and *Xenopus* U1A pre-mRNAs, for the 3' UTR sequences of the cDNAs and for segments of these 3' UTRs, using the FOLD and MFOLD programs (Jaeger *et al.*, 1990; Zuker *et al.*, 1991). In the majority of the predicted low-energy structures the Box 1 and 2 sequences are partially or completely single-stranded and are separated by a phylogenetically conserved stem-loop. Several possible structures exist for the sequences flanking the boxes, especially for the region which contains the cleavage and polyadenylation signal.

Secondary structure models were derived by combining phylogenetic and free energy data. The version in Figure 2B is that for the human U1A mRNA. The model consists of two distinct parts. The 5' part contains three stems (numbered 1, 2 and 3), separated by two asymmetric internal loops containing the Box 1 and 2 sequences. A single unpaired nucleotide is present on the strand opposite each box sequence.

Comparison with the *Xenopus* U1A mRNA sequence provides support for the 5' part of the model between A10 and U53 (Figure 3B; the conserved region lies between the large arrows). All non-conserved nucleotides in this region are at



```
A
                        Box 1
         1        10         20        30
Human    UCGCCACACAGCAUUGUACCCAGAGUCUG-UCCCCAGAC
Mouse    UUGCCACACAGCAUUGUACCCAGAGUCUG-UCCCCAGAC
Xenopus  UCGUGUUUCAGCAUUGCACCCAGAGUCUGCAACUCGGAC
         • •     •••••••• ••••••••••••   • • •••

          Box 2                     polyA signal
         40        50        60        70
Human    AUUGCACCUGGCGCU-GUUAGGCCGGAAUUAAAG-UGGCUU
Mouse    AUUGCACCUGGCGCU-GUUAGAUUGUGAUUAAAG-UGAGUU
Xenopus  AUUGUACCUGGAGCUUGUGUUGUUGU-AAUAAACAUGA
         •••• •••••• ••• ••      •  • ••••  ••
```

Figure 2. (A) Sequence alignment of the conserved part of the 3'UTR sequences of human, mouse and *Xenopus laevis* U1A pre-mRNAs Nucleotides that are identical in all three sequences are marked and the Box 1, Box 2 and polyadenylation sequences are indicated No sequences 3' to those shown are available from either U1A cDNAs or the U1A gene of *Xenopus* (B) Proposed secondary structure of the 3'UTR of the human U1A pre-mRNA The Box 1 and 2 sequences, the cleavage and polyadenylation signal and stems 1, 2, 3 and 4 are indicated unpaired positions with the exception of the A35 to G change, but this difference replaces an A-U pair with G-U. The extra nucleotide that is inserted in the *Xenopus* sequence is located in the terminal loop. Outside of this region phylogenetic support for the model is weak. Stem 1 in the *Xenopus* sequence is only 3 bp long. Further, although it is possible to draw hairpin structures in which the cleavage and polyadenylation signals are in loops, these are not well conserved with respect to the human structure in either the *Xenopus* or mouse U1A pre-mRNAs. One interesting aspect of the *Xenopus* sequence is that the non-conserved nucleotides in the two boxes are reversed in position (Figure 3B).

Enzymatic structure probing

To test the proposed structure, RNase digestions of 5' end-labelled Ag fragments were carried out under native and denaturing conditions using RNases A, U2, T1, T2 and V1. A typical example of the results is shown in Figure 3A, while Figure 3B summarizes the results of several independent experiments. It can be seen that the central three nucleotides of the Box 1 and Box 2 sequences [nt 15-17 (UGU) in Box 1 and nt 41-43 (UGC) in Box 2] are cleaved efficiently by the enzymes T1, A and T2, which are known to cut 3' of nucleotides present in single-stranded regions. In some experiments RNase T2 also appears to cut between other nucleotides in Boxes 1 and 2, but these cleavages were less reproducible. The terminal loop (nt 30-33) was almost never cut under native conditions, suggesting that its structure is very compact.

RNase V1 cuts, which indicate double-stranded or stacked bases, were clearly seen in the regions of stems 2 and 3 and, less reproducibly, in stem 1. In the latter stem, some positions were also cut by RNases A and T1. Therefore stem 1, if it exists, does not seem to be very stable under these conditions. RNase V1 cuts were found in the 5' part of Box 2, which could point to some base stacking. The bulged nucleotides A24 and C50 were never cut under native conditions and are therefore probably located inside the helix. The polyadenylation signal is clearly single-stranded (RNase T2 cuts), flanked by double-stranded regions (RNase V1 cuts). The region between the 5' and 3' parts of the structure was efficiently cleaved by RNase T2, although in a few experiments weak V1 cuts were also found.

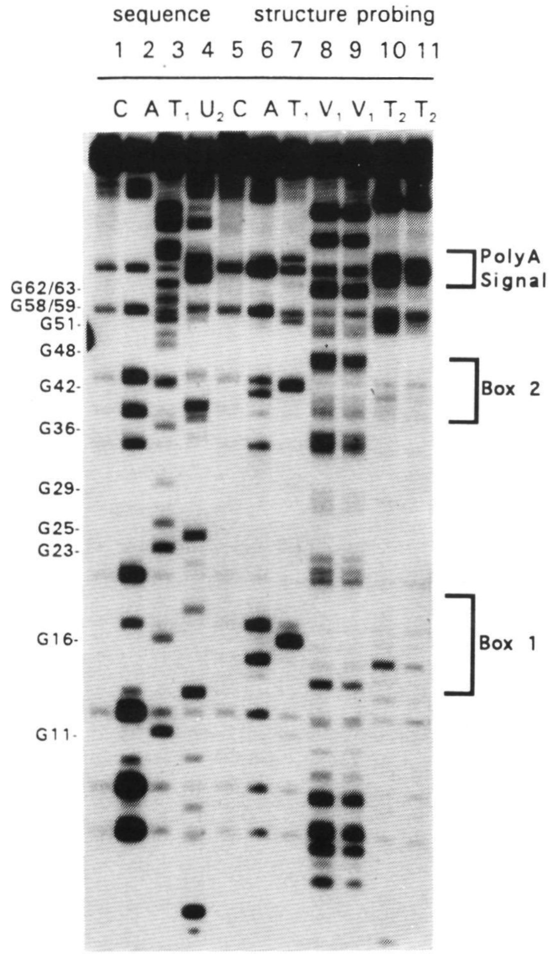


Figure 3. **(A)** Enzymatic digestions of the Ag RNA under denaturing (sequence) and native (structure probing) conditions. RNA samples were treated as described in Materials and methods. The samples in lanes 1 and 5 are control reactions, to which no enzyme was added. Lanes 2-4 contain reactions under denaturing conditions (the enzymes used are indicated) while lanes 6-11 designate reactions under native conditions (two concentrations were used for RNases V1 and T2). The positions of guanosines cleaved by RNase T1 under denaturing conditions are indicated on the left. *(Figure continued on next page).*

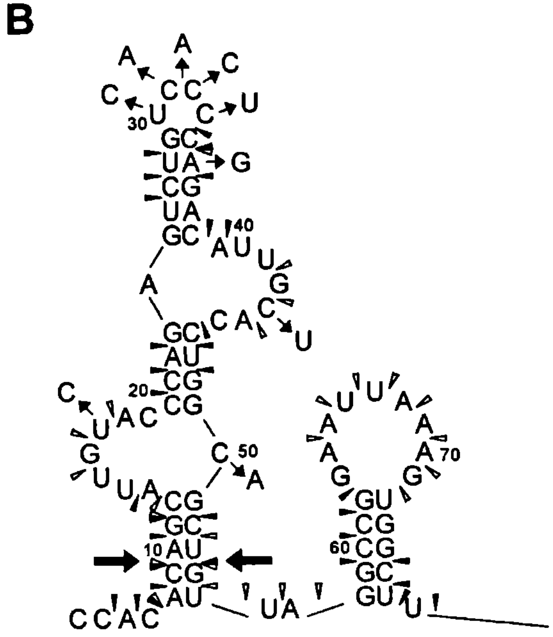


Figure 3 (continued). (B) The secondary structure of the human Ag RNA sequence. Consensus data from several independent experiments (only strong cuts) are shown. RNase V1 cleavage is indicated by closed triangles, RNase A/T1/T2 cleavage with open triangles. For the most conserved part of the structure (nt 10-53, indicated with large arrows), the nucleotide changes in the corresponding *Xenopus* RNA are indicated (small arrows).

RNA mutants

To test the 5' part of the structure more thoroughly and to obtain more information on the less conserved regions we next constructed mutants in the human Ag fragment. Single mutants (called 1A, 1B, 2A, 2B and 3A) were designed to disrupt each of the three 5' stem structures by mutating individual strands of each putative helix (see Table I for mutant sequences). In the double mutants (1AB, 2AB and 3AB), which were designed to maintain complementarity and the putative structure, the sequences of both strands of each stem were interchanged. Further, mutant 3CTD was constructed, in which stem 3 and the terminal loop were replaced by CGGCGCUUCGGCGCCG. This sequence is predicted to form a stem composed of six GC base pairs with a highly stable tetraloop (Tuerk *et al.*,

Table I. Binding assays of 3' UTR mutants

A. Stem mutants

	Sequences	Direct/indirect assay (150 mM)	Direct/indirect assay (500 mM)	Inhibition of polyadenylation
Agwt		+	+	+
3A	${}_{25}$GUCUG GUCUG$_{34}$	+	–	–
3AB	${}_{25}$CAGAG GUCUC$_{34}$	+	+	+
3CTD	see text	N.D.	+	+
2A	${}_{20}$CCAG CCAG$_{46}$	+	–	–
2B	${}_{20}$GGUC GGUC$_{46}$	+	–	–
2AB	${}_{20}$GGUC CCAG$_{46}$	+	+	+
1A	${}_{8}$ACAGC UCAGC$_{51}$	+	–	–
1B	${}_{8}$AGUCG UGUCG$_{51}$	+	+	+
1AB	${}_{8}$AGUCG UCAGC$_{51}$	+	+	+
4A	${}_{58}$GAGAG UCGGU$_{72}$	+*	N.D.	+
4B	${}_{58}$GGCCG UUCU$_{72}$	+*	N.D.	+
4AB	${}_{58}$GAGAG UUCUU$_{72}$	+*	N.D.	+

B. Loop mutants

	Sequences	Direct assay (150 mM)	Direct assay (500 mM)	Indirect assay (150 or 500 mM)	Inhibition of polyadenylation
Agwt		+	+	+	+
ΔB2	${}_{39}$GGAUCCC$_{45}$	+	–	–	–
ΔB1	${}_{13}$GGAUCCC$_{19}$	+	+	–	–
ΔB1/B2	both	–	–	–	–

N D., not determined

*Binding properties of stem 4 mutants were established by using bandshift assays.

1988). If the model were correct, this mutation should not disturb the U1A binding sites.

Enzymatic digestions, as described above, were performed on most of these mutant RNAs. The single mutants 2A and 3A clearly showed a distortion of the structure in the mutated region while the double mutants 2AB and 3AB had digestion patterns similar to the Agwt fragment (data not shown). The results with the stem 1 mutants were less easy interpretable; there was no clear difference between mutants 1B and 1AB, and, as with Agwt, both V1 and single-strand-specific enzymes cut in the stem 1 region of both these mutants.

Two single mutants and one double one were prepared in stem 4. In 4A and 4B the individual strands of the stem were mutated singly to disrupt the potential pairing and in 4AB the mutations were combined to restore pairing (see Table I). Nuclease digestion of mutant 4A suggested that this mutation disrupted stem 4. However, in the double mutant 4AB, in which stem 4 should reform, V1 cleavage was only partially restored (data not shown).

From these experiments we conclude that much of the proposed structure is likely to be correct although there is doubt about the existence, or at least the stability, of stem 1. The mutants could therefore be used to test the structural requirements for U1A protein binding and inhibition of polyadenylation. In addition to the mutants described above we also used the ΔB1, ΔB2 and ΔB1/2 mutants in which the sequences of Box 1 and 2 were altered individually or in combination (Boelens *et al.*, 1993; see Table IB).

Binding of U1A protein to the mutants

Two assays that can detect the binding of U1A protein to an RNA have been described previously (Boelens *et al.*, 1993). In the direct assay ^{35}S-labelled U1A protein is incubated with biotinylated RNA. Proteins that bind the RNA can be recovered via precipitation by Streptavidin-agarose and analysed by SDS-PAGE. In the indirect assay, which gives positive results only when at least two molecules of U1A protein are bound to each RNA, the ^{35}S-labelled U1A is precipitated via non-radioactive biotinylated U1A protein.

We tested the mutant AgRNAs in these assays. As reported (Boelens *et al.*, 1993), both ΔB1 and ΔB2 can still bind U1A in the direct assay in the presence of 150 mM KCl, while ΔB1/2 cannot (Figure 4A, left panel). However, if the KCl concentration is increased to 500 mM, binding to the ΔB2 mutant, which only retains Box 1 and thus an imperfect match to the U1 snRNA sequence, is undetectable (Figure 4A, right panel). Previously, it was shown that ΔB1 and ΔB2 bind maximally one molecule of U1A protein (Boelens *et al.*, 1993).

To characterize further the binding to these mutants, and to define better the reduction in affinity of the ΔB2 mutant, the dissociation constants (K_D) of their binding to U1A protein were determined. First, the K_D of the complex between U1 snRNA and the U1A protein was established. Under the conditions used (see Materials and methods) the K_D of this complex was 5(± 3) x 10^{11} M (Table II). This value is very similar to that (2x10^{11} M) determined by Hall and Stump (1992), who also assayed binding with a nitrocellulose filter binding assay, but used a much shorter RNA substrate and different buffer conditions. Both of these values are considerably (~10^3-fold) lower (i.e. indicative of tighter binding) than K_Ds determined for similar complexes measured by native gel electrophoresis (Lutz-Freyermuth *et al.*, 1990; Jessen *et al.*, 1991).

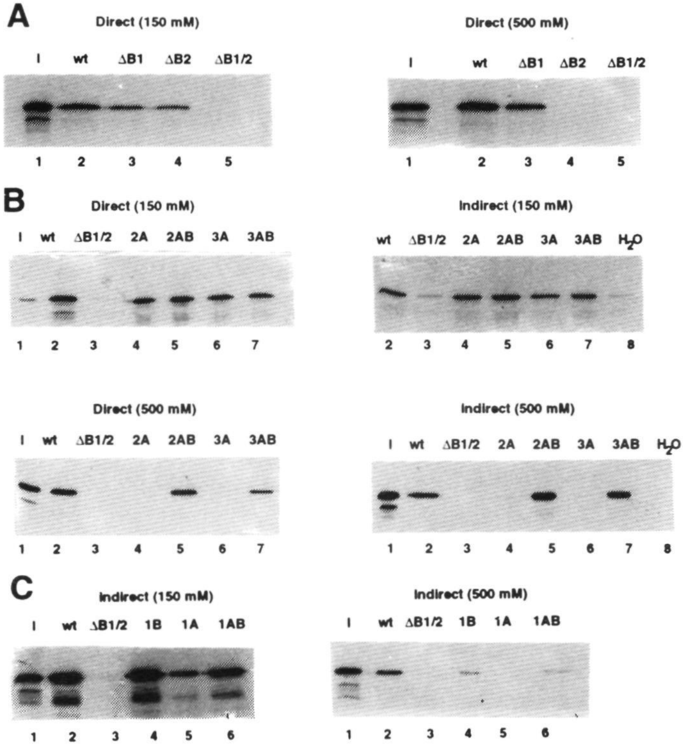


Figure 4. **(A)** Binding of ^{35}S-labelled U1A protein (lane 1, 10% of the input protein per assay) to various RNA substrates at 150 mM KCl (left panel) and 500 mM KCl (right panel). The RNAs used were Ag wt RNA (lane 2), mutant ΔB1 (lane 3), mutant ΔB2 (lane 4) and mutant ΔB1/B2 (lane 5). The mutants are from Boelens *et al.* (1993). **(B)** Binding assays for stem 2 and stem 3 mutants. Left panels: direct binding assay as described in panel A at either 150 mM (upper) or 500 mM (lower) KCl. Right panels: indirect binding assay. Precipitation of ^{35}S-labelled U1A protein via biotinylated U1A protein in the presence of various RNA substrates, as indicated above the lanes. **(C)** Binding assays for stem 1 mutants. Indirect binding assays carried out at 150 mM (left panel) or 500 mM (right panel) KCl.

Table II. Dissociation constants of various RNA-U1A protein complexes

RNA	K_D (M)	n
U1 RNA	5 (±3) x 10^{11}	4
U1A pre-mRNA	10 (±6) x 10^{-11}	4
Ag	6 x 10^{-11}	1
ΔB1	30 (±10) x 10^{-11}	3
ΔB2	800 (±100) x 10^{-11}	2

n, number of independent determinations.

In our assay the human U1A pre-mRNA-U1A protein complex has a K_D of 10(±6) x 10^{-11} M. Taking into account the measured variation in the K_D-values, U1 snRNA and the U1A pre-mRNA therefore exhibit comparable binding affinity. Note, however, that the K_D measured for the U1A pre-mRNA is complex since two U1A protein molecules bind to the pre-mRNA, and, in this assay, only one molecule has to be bound to score positive. In the single experiment carried out with the Ag fragment, the K_D was indistinguishable from those of either the pre-mRNA or U1 snRNA (Table II). The ΔB1 mutant showed an ~3-fold lower binding affinity [K_D=30(±10)x10^{11}M] than the wildtype (wt) pre-mRNA. In the case of the ΔB2 mutant, which only contains the imperfect Box 1 binding sequence, the binding affinity decreased by a factor of ~80 [K_D=800(±100)x10^{-11}M]. The affinity of the wt pre-mRNA is higher than the additive affinities of the two single-site mutants, indicating that there might be some cooperativity in the binding of the two U1A protein molecules. This conclusion was supported by electrophoretic mobility shift assays where, at protein concentrations at which low binding site saturation was achieved, the amount of U1A required to occupy both sites on an RNA was 2- to 4-fold greater than that required to occupy a single site (data not shown). The K_Ds of the two individual sites (Table II) would predict that, without cooperativity, ~30-fold more protein should be required.

Mutants affecting stems 1-3 of the structural model (Figure 2B) were next tested in the direct and indirect assays. Unexpectedly, mutants 2A, 2B and 3A, in which stems 2 or 3 were disrupted, could still bind U1A protein at 150 mM KCl in both the direct and indirect assays (Figure 4B, upper panels, lanes 4 and 6 and data not shown). Thus, disruption of either of the stems did not prevent interaction with U1A protein. When the assays were carried out at 500 mM KCl, however, it was evident that the affinity of the single mutants for U1A protein was reduced. Mutants 2A, 2B and 3A were incapable of interaction with even one molecule of U1A protein in these conditions (Figure 4B, lower panels, lanes 4 and 6 and data not shown). Restoration of stems 2 and 3 in the 2AB and 3AB double mutants restored U1A protein binding in both assays (Figure 4B, lanes 5 and 7). The 3CTD mutant, which contains a more stable terminal stem-loop, showed U1A protein binding comparable to that of wt U1A pre-mRNA (data not shown), providing further support for the presence of stem 3.

In case of the stem 1 mutants, a less clear-cut result was obtained. At high, but not at low, salt concentration one of the single mutants 1A, failed to bind U1A protein (Figure 4C, left and right panels, lane 5), suggesting that stem 1 might be needed for protein binding. Mutant 1A showed wt behaviour in the direct assay at 150 mM salt, but did not detectably bind U1A protein in this assay at 500 mM salt (data not shown). The other single mutant (1B), on the other hand, as well as the double mutant (1AB), both showed behaviour comparable to that of wt pre-mRNA (Figure 4C, lanes 2, 4 and 6). One explanation for this behaviour might be that some of the base positions mutated in mutant 1A (nt 51-54) are necessary for U1A protein binding in the absence of a stem structure. It is, however, also possible that the 1A mutation causes changes in the structure to occur in high salt and thus affects U1A protein binding in a less direct way.

Both the single and double mutants of stem 4 bound U1A protein like the wt RNA (data not shown), indicating that this part of the structure is not necessary for U1A protein binding. Taken together, these results support the structural data summarized earlier, since they indicate that the highly conserved and stable stems 2 and 3 are important for high affinity U1A binding. The less conserved and less stable stem 1 is not required for U1A binding, as is shown by mutant 1B, but the phenotype of the 1A mutant suggests that the stem might stabilize binding in some circumstances. Stem 4 is not needed for U1A protein binding.

Inhibition of polyadenylation

One functional consequence of U1A protein binding is inhibition of U1A pre-mRNA polyadenylation (Boelens *et al.*, 1993). The effects of the various mutations were therefore tested in an *in vitro* cleavage and polyadenylation assay. U1A protein addition to these assays results in specific inhibition of polyadenylation of the U1A wt substrate (Figure 5A, left panel). Considerably more U1A protein is required to inhibition polyadenylation of the double mutant ΔB1/B2, which cannot bind U1A specifically (Figure 5A, right panel). Polyadenylation of the two single mutants, ΔB1 and ΔB2 (Figure 5A, middle panels), is inhibited at a level of U1A protein only ~4-fold lower than that required for non-specific inhibition. This indicates that for efficient inhibition of the cleavage and polyadenylation reaction, it is crucial that two molecules of U1A protein can bind to the U1A pre-mRNA substrate.

The behaviour of the stem mutants in the polyadenylation inhibition assay closely mirrored their ability to bind U1A protein in 500 mM KCl in the binding assays described above. In the case of stems 2 and 3, the 2A and 3A single mutants, which are defective in U1A binding at high salt, behaved similarly to the ΔB1/2 double mutant (Figure 5B) while the 2AB and 3AB double mutants, in which stems 2 and 3 are restored, behaved similarly to the wt pre-mRNA. The 3CTD mutant also showed wt behaviour in polyadenylation inhibition (data not shown). The behaviour of the stem 1 mutants was also in agreement with the results of U1A protein binding at high salt. Mutants 1B and 1AB showed inhibition of polyadenylation comparable to the wt pre-mRNA (Figure 5C), while the 1A mutant showed no inhibition of polyadenylation.

As mentioned above, the existence of stem 4 in the human U1A pre-mRNA is supported by the nuclease digestion data but the stem has not been strongly conserved in evolution. To examine directly a possible functional role for this structure we tested the three mutants 4A, 4B and 4AB. All three mutants behaved similarly to the wt pre-mRNA in the polyadenylation inhibition assay (Figure 6). Thus, even if stem 4 does form, its existence does not seem to be important for the inhibition of polyadenylation by U1A protein.

DISCUSSION

Structure of the 3' UTR of U1A (pre-)mRNA

The structure of the region of the U1A pre-mRNA responsible for binding to the U1A protein and thus for mediating autoregulatory inhibition of polyadenylation has been examined. Various lines of evidence suggest that the structure is complex. From top to bottom it starts with a tetraloop bounded by a stem of 5 bp (stem 3). Stem 3 is followed by an asymmetric internal loop, containing on one strand a 7 nt sequence required for U1A protein binding. On

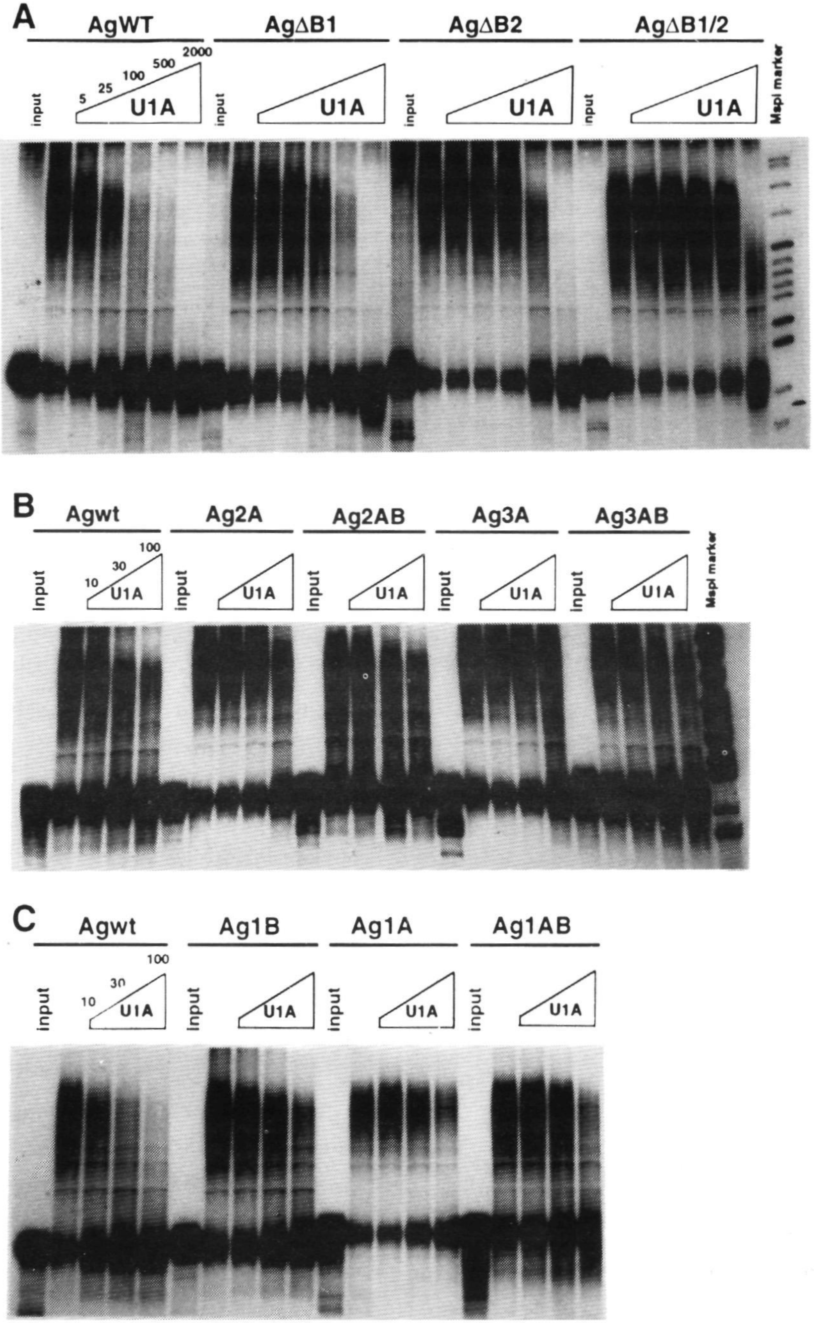

A
AgWT
AgΔB1
AgΔB2
AgΔB1/2
input
5
25
100
500
2000
U1A
input
U1A
input
U1A
input
U1A
MspI marker
B
Agwt
Ag2A
Ag2AB
Ag3A
Ag3AB
input
10
30
100
U1A
input
U1A
input
U1A
input
U1A
input
U1A
MspI marker
C
Agwt
Ag1B
Ag1A
Ag1AB
input
10
30
100
U1A
input
U1A
input
U1A
input
U1A

Figure 5 (opposite page). Effect of U1A protein on *in vitro* polyadenylation of Agwt RNA and the 3'UTR mutants. **(A)** Loop mutants. Recombinant, highly purified U1A protein was preincubated with the labelled RNA substrate for 5 min at room temperature. The 3' processing reaction was initiated by addition of the reaction buffers and nuclear extract. The labelled RNA assayed is indicated above each panel. The first lane of each panel is the input precursor RNA in the absence of nuclear extract or U1A protein. The second lane of each panel is polyadenylation in the absence of exogenously added U1A protein. The remaining lanes of each panel show the effect of addition of increasing amounts of exogenous U1A protein with the amounts indicated above each lane. The lane on the extreme right is a ^{32}P end-labelled *Msp*I digest of pBR322. **(B)** Stem 2 and 3 mutants. The type of labelled RNA used is indicated above each panel. The lanes of each panel are the same as described in panel A, except that the amounts of exogenously added U1A protein are different (ranging from 10 to 100 ng) as indicated above the panel. The lane on the extreme right is a ^{32}P end-labelled *Msp*I digest of pBR322. **(C)** Stem 1 mutants. The labelled RNA assayed is indicated above each panel. The lanes of each panel are the same as described in panel A, except that the amounts of exogenously added U1A protein are different (ranging from 10 to 100 ng) as indicated above the panel.

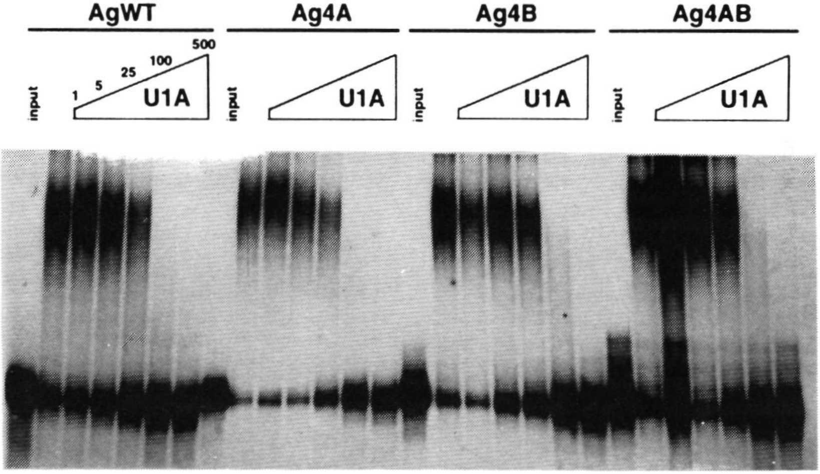


Figure 6. Effect of the U1A protein on the *in vitro* polyadenylation of the Agwt RNA and the stem 4 mutants. The labelled RNA assayed is indicated above each panel. The lanes of each panel are the same as described in Figure 5A, except that the amounts of exogenously added U1A protein are different (ranging from 1 to 500 ng) as indicated above the panel. The lane on the extreme right is a ^{32}P end-labelled *Msp*I digest of pBR322.

the other strand a single unpaired nucleotide is found, which probably stacks into the helix since it is inaccessible to nucleases. A second stem of four base pairs (stem 2) separates this internal loop from a second asymmetric internal loop, similar to the first, which may, or may not, be bounded by a further short helix (stem 1).

Apart from stem 1, all the secondary structure elements in the 5' part of the structure were shown to be required for optimal binding to U1A protein and for function in polyadenylation inhibition. In the case of stem 1, the evolutionary conservation of the potential to form at least a short stem at this position suggests that the stem, though metastable, may exist. The effects of mutations in this putative stem on U1A protein binding were diverse. The results obtained with mutant 1B, however, established that the potential to form stem 1 is not essential for U1A protein binding.

The second structural element in the conserved region of human U1A pre-mRNA is a stem-loop with the AUUAAA cleavage and polyadenylation signal forming most of the loop. This structural feature is unnecessary for U1A protein binding and for inhibition of polyadenylation and, in addition, is not well conserved in evolution. Thus, even if this part of the structure does form *in vivo* it is unlikely to have any relevance for autoregulation.

What might be the reason for the complexity of the proven part of the structure, the region to which U1A protein binds? First, the data presented indicate that efficient inhibition of polyadenylation is only possible when two molecules of U1A protein can bind to the pre-mRNA. Second, the binding studies show that the two protein molecules bind cooperatively. An attractive aspect of the structure from this point of view is that the two asymmetric internal loops are spaced approximately half a helical turn apart (if standard RNA geometry is applicable). Although it can be assumed that the internal loops will induce a distortion or a kink in the helix (Chastain and Tinoco, 1991), the U1A binding sites may therefore lie side by side on one face of the helix, favouring interaction between the two protein molecules during binding. Note, however, that we do not know whether the observed cooperativity of binding is due to protein-protein interaction or to changes induced in the pre-mRNA structure on binding the first molecule of U1A protein.

Comparison of two U1A binding sites

The K_Ds of the two physiologically relevant U1A protein-RNA complexes studied to date, those involving U1 snRNA and U1A pre-mRNA, are very similar and indicative of very high affinity binding. The tight binding to U1 snRNA is perhaps explicable since U1A protein in the free state would turn off its own production via autoregulation and, presumably, U1 snRNA without U1A might be non-functional (Hamm *et al.*, 1990; but see Liao *et al.*, 1993). There seems not to be an obvious rationale for such a strong interaction between U1A protein and its pre-mRNA.

Given the high affinity of both RNAs for U1A protein it is interesting to compare them. The U1A binding site on U1 snRNA is stem-loop II or B (Scherly *et al.*, 1989; Lutz-Freyermuth *et al.*, 1990). Parts of the 10 nt loop sequence and the presence of stable stem, but apparently not the detailed structure of the stem, are critical for tight binding (Scherly *et al.*, 1989, 1990; Lutz-Freyermuth *et al.*, 1990; Bentley and Keene, 1991; Jessen *et al.*, 1991; Tsai *et al.*, 1991; Hall and Stump, 1992). A model for the U1A-U1 snRNA interaction has been proposed (Jessen *et al.*, 1991) in which most of the protein-RNA contacts are with the phosphates of the RNA backbone and the loop sequence is proposed to be mainly required to generate the correct backbone conformation.

The structural context of the most U1 snRNA-like sequence in the U1A pre-mRNA (Box 2) as a 7 nt unpaired strand in an asymmetric loop sandwiched between two stems, would appear to be rather different from its context in U1 stem-loop II. Given the conformational flexibility of RNA it is premature to say that the structure of the two tight binding sites will be different, but further examination of the role of the single-stranded bases in protein binding as well as high resolution studies of the two RNAs to reveal similarities and differences in their structures would be particularly interesting areas of study.

Inhibition of polyadenylation

The major conclusions of this study with regard to polyadenylation inhibition are that structural changes in U1A pre-mRNA that result in either a reduction in affinity for U1A protein or in the loss of the capacity to bind two molecules of U1A protein alleviate the inhibitory effects of U1A protein on cleavage and polyadenylation reactions *in vitro*.

The requirement for two bound protein molecules for inhibition might be most easily compatible with a simple model in which U1A protein sterically hinders interaction of one of the multiple cleavage and polyadenylation factors (see Wahle and Keller, 1992 for a review) with the U1A pre-mRNA. However, more complex models involving specific interaction between U1A protein or a particular structure in U1A pre-mRNA induced by U1A binding and one or more of the processing factors are not ruled out. These possibilities can now be tested.

MATERIALS AND METHODS

Sequence alignment and secondary structure prediction

The alignment of the three U1A sequences was made with the program PILEUP, which is part of the University of Wisconsin GCG Package v 7 0 (Devereux *et al*, 1984), and was adjusted manually The programs FOLD and MFOLD (Zuker *et al.*, 1991) were used to generate optimal and suboptimal foldings of different regions of the three RNA sequences

Enzymatic structure probing

The Ag and mutant RNAs used in this study were dephosphorylated at their 5' ends and then radioactively labelled using [γ-^{32}P]ATP and T4 polynucleotide kinase according to Ehresmann *et al* (1987) The labelled RNAs were purified by electrophoresis on a 10% polyacrylamide-urea denaturing gel The full-length RNA products were cut out of the gel and eluted overnight at 4°C in a buffer containing 0 5 M NH_4Ac (pH 6 5), 10 mM $MgCl_2$ and 0 1 % SDS (Krol and Carbon, 1989) The RNA was precipitated with ethanol and resuspended in water

Labelled RNA (2-3x10^4 c p m) was supplemented with 4 μg of total yeast RNA as carrier Digestion with RNase T1 (0 01 U), T2 (0 005 U), U2 (0 2 U, only in buffer D), A (1x10^{-6} U) or V1 (0 06 U, only in buffer N) were performed at room temperature for 10 min in buffer N or at 50°C for 5 min in buffer D Buffer N (native conditions) contained 10 mM Tris pH=7 5, 10 mM $MgCl_2$ and 50 mM KCl Buffer D (denaturing conditions) contained 7 M urea, 1 mM EDTA and 25 mM sodium acetate

Preparation of mutants

The Ag sequence was inserted as an *Eco*RI-*Hind*III fragment into the pGEM-3z(+) vector Single-stranded DNA was produced with the helper phage M13 K07 and mutations were introduced using the oligonucleotide-directed mutagenesis kit from Amersham All mutants were checked by DNA sequencing

Binding and polyadenylation assays

RNA and biotinylated RNA transcription by T7 RNA polymerase, production of ^{35}S-labelled U1A protein in wheat germ extract, production of recombinant U1A protein from *E coli*, its biotinylation, the direct and indirect RNA-protein binding assays and *in vitro* polyadenylation reactions were all carried out as described by Boelens *et al* (1993) The nucleotide sequence of the Ag fragment of U1A extends from position 842 to position 951 in the sequence (Nelissen *et al*, 1991) and include 8 nt at the 5' end derived from the vector plasmid Since U1A protein loses polyadenylation inhibition activity when stored, the amount required to inhibit polyadenylation of the wt pre-mRNA was determined empirically for each experiment

For the electrophoretic mobility shift experiment, ^{32}P-labelled RNA was heated at 95°C for 3 min and quenched on ice for 1 min 2 x 10^4 c p m were added to the protein in a 10 μl reaction containing 10 mM Na-HEPES (pH 7 4), 50 mM KCl, 1 mM $MgCl_2$ and 200 ng of competitor tRNA at room temperature The reaction was immediately loaded on a 7% native acrylamide gel (60 1 acrylamide bisacrylamide), containing 10 mM Tris-borate pH 8 3, 1 mM EDTA and 0 1% Triton X-100 The gel was autoradiographed for 2-12 h at -80°C

Filter binding assay

To determine the dissociation constants for the interaction between U1A protein and RNA substrates a nitrocellulose filter binding assay was used A constant concentration of U1A protein in 10 μl buffer 1 containing 100 mM KCl, 2 mM MgCl$_2$, 20 mM HEPES-KOH pH 7.9, 5% glycerol, 0.5 mM DTE and 0.5 mg/ml BSA was mixed with 90 μl buffer 2 containing 10 mM Tris-HCl pH 7.5, 100 mM KCl, 2 mM MgCl$_2$, 0.1 mM EGTA, 0.5 mM DTE, 0.1 μg/μl tRNA and varying concentrations of ^{32}P-labelled RNA substrates. After equilibration at 20°C for 120 min, samples were filtered through pre-soaked Schleicher and Schuell BA85 0.45 μm nitrocellulose filters using a dot blot manifold (Schleicher and Schuell SRC96). The samples were subsequently washed twice with 200 μl buffer 2 without tRNA. The filters were dried and the amount of ^{32}P-labelled RNA bound to the filter was quantified by scintillation counting. The K$_D$s were determined by Scatchard plot analysis.

ACKNOWLEDGEMENTS

Use of the services and facilities of the Dutch CAOS CAMM Center, under grant numbers SON 326-052 and STW NCH99.1751, is gratefully acknowledged. We thank M. Bennett and J. Craft for making the mouse U1A cDNA sequence available prior to publication. The research was supported by the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organization for Scientific Research (NWO), EMBO, EMBL and a postdoctoral fellowship (to S.G.) from the Alexander von Humboldt Foundation.

REFERENCES

Bach, M., Krol, A. and Luhrmann, R (1990) *Nucleic Acids Res.*, **18**, 449-457.

Bentley, R.C. and Keene, J.D. (1991) *Mol. Cell. Biol.*, 11, 1829-1839.

Boelens, W.C., Jansen, E.J.R., Van Venrooij, W J., Stripecke, R., Mattaj, I W. and Gunderson, S.I. (1993) *Cell*, 72, 881-892.

Chastain, M. and Tinoco, I.,Jr (1991) *Prog. Nucleic Acid Res. Mol. Biol.*, 41, 131-177.

Devereux, J., Haeberli, P. and Smithies, O. (1984) *Nucleic Acids Res.*, 12, 387-395.

Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J-P. and Ehresmann, B. (1987) *Nucleic Acids Res.*, 15, 9109-9129.

Hall, K.B. and Stump, W.T. (1992) *Nucl. Acids Res.*, 20, 4283-4290.

Hamm, J., Dathan, N A , Scherly, D. and Mattaj, I.W. (1990) *EMBO J.*, **9**, 1237-1244.

Hoffman, D.W., Query, C.C., Golden, B.L., White, S.W. and Keene, J.D. (1991) *Proc. Natl Acad. Sci., USA*, **88**, 2495-2499.

Hope, I.A. and Struhl, K. (1987) *EMBO J.*, 6, 2781-2784.

Jaeger, J.A., Turner, D.H. and Zuker, M. (1990) *Methods Enzymol.*, **183**, 281-303.

Jessen, T.H., Oubridge, C., Teo, C.H., Pritchard, C. and Nagai, K. (1991) *EMBO J.*, 10, 3447-3456.

Krol, A. and Carbon, P. (1989) *Methods Enzymol.*, 180, 212-227.

Liao, X.C., Tang, J. and Rosbash, M (1993) *Genes Dev.*, 7, 419-428.

Luhrmann, R, Kastner, B. and Bach, M. (1990) *Biochim. Biophys. Acta*, 1087, 265-292.

Lutz-Freyermuth, C., Query, C.C. and Keene, J D. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 6393-6397.

Nagai, K., Oubridge, C., Jessen, T.H., Li, J. and Evans, P.R. (1990) *Nature*, **348**, 515-520.

Nelissen, R L.H., Sillekens, P.T.G., Beijer, R.P., Van Kessel, A.H.M.G. and Van Venrooij, W.J. (1991) *Gene*, 102, 189-196.

Scherly, D., Boelens, W., Van Venrooij, W.J., Dathan, N.A., Hamm, J. and Mattaj, I.W. (1989) *EMBO J.*, **8**, 4163-4170.

Scherly, D., Boelens, W., Dathan, N.A., Van Venrooij, W.J. and Mattaj, I.W. (1990) *Nature*, **345**, 502-506.

Scherly, D., Kambach, C., Boelens, W., Van Venrooij, W.J. and Mattaj, I.W. (1991) *J. Mol. Biol.*, **219**, 577-584

Sillekens, P.T., Habets, W.J., Beijer, R.P. and Van Venrooij, W.J. (1987) *EMBO J.*, **6**, 3841-3848.

Tsai, D.E., Harper, D S. and Keene, J.D. (1991) *Nucleic Acids Res.*, **18**, 4931-4936.

Tuerk, C. *et al.* (1988) *Proc. Natl Acad. Sci., USA*, **85**, 1364-1368.

Wahle, E. and Keller, W. (1992) *Annu. Rev. Biochem.*, **61**, 419-440.

Zuker, M , Jaeger, J.A. and Turner, D.H. (1991) *Nucleic Acids Res.*, **19**, 2707-2714.

CHAPTER 5

Chemical structure probing of the 3' UTR of U1A mRNA and footprinting analysis of its complex with U1A protein

Manuscript in preparation

Chemical structure probing of the 3' UTR of U1A mRNA and footprinting analysis of its complex with U1A protein

Celia W.G. van Gelder, Sander W.M. Teunissen and Walther J. van Venrooij
University of Nijmegen, Department of Biochemistry, PO Box 9101, 6500 HB Nijmegen, The Netherlands

ABSTRACT

The structure of the conserved region of the U1A pre-mRNA and its complex with U1A protein was investigated. The secondary structure of the U1A mRNA was determined using chemical modification techniques, while the RNA-protein complex was investigated by footprinting analyses using both ribonucleases and hydroxyl radicals.

The secondary structure of U1A mRNA deduced from the chemical probing largely corresponds to the structure predicted previously, which was based on enzymatic probing and analysis of structure and function of mutant mRNAs. However, some important additional information was obtained. All nucleotides in the conserved Box regions are fully accessible, as are the two unpaired nucleotides A24 and C50. Interestingly, the behavior of the two Box regions appears not to be completely identical, neither in the naked RNA nor in the RNA-protein complex. For the UCCC tetraloop, which could not be cleaved by RNases, chemical probing shows that three of the four bases in the loop are accessible.

Concerning the RNA-protein complex, the protection experiments show that the Box 1 and Box 2 regions are largely protected when the U1A protein is present. All stem regions in the 5' part of the structure seem protected against ribonucleases, while protection against the smaller hydroxyl probe is limited primarily to nucleotides in the Box regions. Interestingly, the nucleotides of the tetraloop become accessible to RNases in the RNA-protein complex. This result indicates that this loop undergoes a conformational change upon U1A protein binding. The 3' part of the structure, containing the polyadenylation signal in a hairpin, shows hardly any protection, a finding that agrees with the fact that U1A does not interfere with the binding of the cleavage polyadenylation specificity factor (CPSF) to the polyadenylation signal during polyadenylation.

INTRODUCTION

The removal of introns from the pre-messenger RNA, known as splicing, is an important process in which several small ribonucleoprotein particles (snRNPs) participate. One of them, U1 snRNP, contains a U1 snRNA molecule, at leasi eight Sm proteins also present in other U snRNPs, and three U1-specific proteins named U1-70K, U1C and U1A (1).

The U1A protein binds directly to the second stemloop of U1 snRNA (2, 3). The protein contains two RNP motifs, of which the N-terminal copy is responsible for binding to U1 snRNA (2-6). The structure of this RNA-binding domain of the U1A protein has been determined by X-ray crystallography and NMR studies (4, 7) and consists of a four stranded antiparallel ß-sheet with two α-helices both lying on the same side of the sheet. The loop of the second hairpin of human U1 snRNA contains 10 nucleotides. It has been shown that the first seven of them (with the highly conserved sequence AUUGCAC), are critical for U1A protein binding, although the structural context of this sequence affects binding affinity (2, 8-10). Recently, the complex between the N-terminal RNP motif of U1A and the second stemloop of U1 snRNA have been studied by NMR (11, 12) and cross-linking studies (13). The ß-sheet of U1A was shown to form the recognition surface and protein-RNA contacts mainly occur at the loop of the RNA hairpin. Furthermore, the crystal structure of this RNA-protein complex has been determined (14), revealing detailed information on the interaction of U1 snRNA with the U1A protein.

It has been shown that the 3' UTR of the U1A pre-mRNA contains a region which has been conserved among vertebrates (15). This region contains two stretches of seven nucleotides (called Boxes 1 and 2) having a sequence similar to that contained in the second, U1A binding stemloop of U1 snRNA and located in close proximity to the polyadenylation signal. It has been demonstrated that two human U1A proteins can bind to these two Box regions (15, 16) and *in vitro* and *in vivo* experiments have shown that the binding of two U1A proteins to this region specifically inhibits polyadenylation of the pre-mRNA (15). Thus, U1A protein regulates the production of its own mRNA via a mechanism that involves pre-mRNA binding and inhibition of polyadenylation. The mechanism of this regulation has been elucidated by *in vitro* studies where U1A protein was shown to inhibit both specific and nonspecific polyadenylation by poly(A) polymerase (PAP) (17). Furthermore, this inhibition was shown to depend on a specific interaction of U1A protein with mammalian PAP in which the C-termini of both proteins seem to be involved (17).

Recently the human U1A protein - U1A pre-mRNA complex and the relationship between its structure and function in inhibition of polyadenylation *in vitro* was investigated (16). The secondary structure of the conserved region of the 3'UTR (Ag RNA) was determined by a combination of theoretical predictions, phylogenetic sequence alignment, enzymatic structure probing and analysis of structure and function of mutant mRNAs (16). The structure shows both Box sequences as single stranded regions in two asymmetric internal loops which are flanked by two essential stem structures, and appears to be different from the binding site of U1A protein on U1 snRNA. The integrity of much of this structure is required for both high affinity binding to U1A protein and specific inhibition of polyadenylation *in vitro* (16).

Here a more detailed analysis of the U1A pre-mRNA and its complex with U1A protein is reported. Chemical probing was performed on the U1A mRNA both at room temperature and at zero degrees, which gave us a better understanding of some structural features which were not perfectly clear from the enzymatic probing experiments. The behavior of both the Watson-Crick positions and the N7 atoms of the purines was analyzed. Furthermore, the complex of U1A mRNA with U1A protein was studied by using ribonuclease and Fe(II)EDTA footprinting analyses.

MATERIALS AND METHODS

In vitro transcription of RNAs and purification of recombinant U1A protein

In vitro transcription by T7 RNA polymerase was carried out as described (16). The conserved region of the 3' UTR of the U1A mRNA is called Ag RNA and was cloned into the EcoRI and HindIII sites of pGEM-3Zf(+) resulting in Ag transcripts. The nucleotide sequence of the Ag fragment of U1A extends from VI-842 to VI-951 in the sequence (18) and includes 8 nucleotides at the 5' end derived from the vector plasmid. Production of recombinant U1A protein from *E.coli* was carried out as described (15).

5-' and 3-' end-labeling

For 5'-end-labeling the dephosphorylated RNAs or oligodeoxynucleotides were labeled using [γ-^{32}P]ATP and T4 polynucleotide kinase (Boehringer) as described previously (16). For 3'-end-labeling, the RNAs were labeled using [^{32}P]pCp and T4

RNA ligase as described (19). The labeled molecules were separated by electrophoresis in a 10% denaturing polyacrylamide/8 M urea gel. The full-length labeled products were excised and eluted from the gel (19), after which they were precipitated with ethanol and stored at -20°C. Just before use they were dissolved in water.

Chemical Modification

All chemical modification experiments were performed at least three times to obtain consistent data. Concentrations of chemicals were optimized to obtain 'single hit' conditions. Control incubations, in which the reagent was omitted, were always performed in parallel to detect spontaneous pyrimidine-purine breaks, which easily occur in RNA (20, 21), and, in the case of the primer extension method, to detect spontaneous stops of reverse transcriptase (RT). Chemical modifications were performed both on unlabeled RNA (0.3-0.5 μg) and on 3'-end-labeled (3 x 10^4 cpm), which was always renatured before use. The RNAs were modified under native conditions (N) (presence of magnesium), semi-denaturing conditions (SD) (presence of EDTA) and denaturing conditions (D) (high temperature, presence of EDTA). Native and semi-denaturing reactions were conducted both at 20°C and at 0°C. Modification reactions were essentially carried out as described (19). Chemically modified nucleotides were detected either by primer extension analysis or by using 3'-end-labeled RNA (in the case of N7-G, N7-A and N3-C).

Buffer solutions: Buffer I: 200 mM HEPES pH 8.0, 10 mM $MgCl_2$, 50 mM KCl. Buffer II: 200 mM HEPES pH 8.0, 1 mM EDTA. Buffer III: 50 mM Na-borate pH 8.0, 10 mM $MgCl_2$, 50 mM KCl. Buffer IV: 50 mM Na-borate pH 8.0, 1 mM EDTA.

DMS treatment: 0.5 - 2 μl DMS (dimethylsulfate) was added to the sample in 200 μl of Buffer I (native conditions) or Buffer II (semi-denaturing conditions) and incubated for 5 min at 20°C or 0°C. Under denaturing conditions, 0.5-2.0 μl DMS was used in 300 μl Buffer II and incubation was for 1 min at 90°C. Reactions were stopped by ethanol precipitation after addition of 10 μg carrier tRNA. The RNA was then subjected to primer extension to probe N3-C and N1-A positions. In case of probing the N7-G positions, the RNA pellets were resuspended in 300 mM sodium acetate and reprecipitated with ethanol. RNA pellets were dried and redissolved in 10 μl of 1 M Tris-HCl, pH 8.0, after which 10 μl of fresh 200 mM $NaBH_4$ was added. After an incubation on ice, in the dark for 20 min, 200 μl of cold 0.6 M HAc/NaAc (pH 4.5) was added, and the RNA was precipitated by adding 600 μl of cold ethanol. The pellets were resuspended in 5 μl of H_2O after which 20 μl of 9% aniline-acetate buffer, pH 4.5, was added, followed by an incubation for 15 min at 60°C in the dark. The RNA was again precipitated with ethanol and, either analysed on denaturing gels (in case of end-labeled RNA) or subjected to primer extension. For modification of N3-C using 3'-end-labeled RNA different amounts of DMS were added to $2x10^4$ cpm of Ag RNA in 200 μl of Buffer I (N) or Buffer II (SD, D). Incubations were for 5 min at 20°C or 0°C, after which the reaction was stopped by ethanol precipitation. Hydrazine-aniline treatment (20) was carried out to produce strand scission at the site of the modification.

CMCT treatment: A freshly prepared 42 mg/ml H_2O of CMCT (1-cyclohexyl-3-(2-morpholino ethyl)-carbodiimide metho-p-toluene sulfonate; Merck) was used. Under N and SD conditions, 50 μl CMCT was added to the sample in 150 μl Buffer III (N) or IV (SD) for several different incubation times at 20°C or 0°C. Under D conditions, 5-25 μl CMCT was added to the sample in 150 μl of Buffer IV; incubation was for 1 min at 90°C. Reactions were stopped by ethanol precipitation.

DEPC treatment: Ten to sixty μl DEPC (diethylpyrocarbonate) was added to the sample in 200 μl Buffer I (N) or II (SD); incubation was for 1 hr at 20°C. For D conditions, 3-10 μl DEPC was added to 200 μl Buffer II and incubated for 7 min at 90°C. Reactions were stopped by ethanol precipitation. An aniline step was performed to produce strand scission at the site of the modification (20).

Primer extension analysis

Primer extension was carried out essentially as described (19). Oligodeoxynucleotide primers 5'-GCTTAACAGCGCCAGG-3' and 5'-GATTGTGAAAAACCAAACCTC-3', complementary to nucleotides 45-60 and 81-101 in Ag RNA, respectively, were 5'-end-labeled. Annealing was performed by dissolving the modified RNA template in 2 μl H_2O containing 10 μg tRNA and $5x10^4$ cpm of labeled primer, heating it at 90°C for 1 min, followed by an incubation on ice for 1 min and returning to room temperature for 10 min. Extensions were achieved by adding 3 μl of a reverse transcription mix containing one unit of AMV reverse transcriptase (Boehringer) in 5 mM Tris-HCl (pH 8.0), 7 mM $MgCl_2$, 50 mM KCl, 5 mM DTT, 170 μM dNTPs and an incubation at 37°C for 45 min. Reactions were stopped by adding 20 μl stopbuffer (50 mM Tris-HCl pH 8.3, 75 mM EDTA, 0.5% SDS). The RNA was hydrolyzed by adding 3 μl of freshly prepared 3M KOH, followed by incubation at 90°C (3 min) and 37°C (1 hr). Then 6 μl concentrated acetic acid was added and the DNA fragments were ethanol precipitated. Reverse transcripts were analyzed on 10% denaturing polyacrylamide gels.

Enzymatic footprinting

All enzymatic footprinting experiments were repeated at least three times to obtain consistent data. In this type of experiment, renatured 5'-end-labeled Ag RNA (3 x 10^4 cpm, final concentration ~ 6 nM) was always used. Renaturation was achieved by heating the RNA at 65°C for 1-2 min followed by slow cooling to room temperature. Subsequently the specified amount of U1A wt protein (150 - 300 fold excess) or A101, containing the N-terminal 101 amino acids of U1A, was added (final volume: 20 μl). Buffer conditions were 10 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$ and 50 mM KCl. The complex was allowed to form for 30 min at room temperature after which the probing reactions were performed with RNase T1 (0.15 U; U1A wt only), RNase A (1x10^{-5} U; U1A wt only), RNase T2 (0.005 U), or RNase V1 (0.06 U) at room temperature for 10 min. The reactions were stopped by phenol extraction and the samples were analysed on a 10% denaturing gel. To establish cleavage positions in the Ag RNA, digestions were performed with RNase T1 and A, under denaturing conditions, at 50°C for 5 min in a buffer containing 7 M urea, 1 mM EDTA and 25 mM sodiumacetate.

Fe(II)EDTA footprinting

Cleavage reactions were carried out essentially as described by Darsillo *et al.* (22). In all experiments 5'-end-labeled Ag RNA (5 x 10^4 cpm) was used. The specified amount of U1A protein was added an the complex was allowed to form for 30 min at room temperature after which the probing reactions were performed at room temperature for 10 min (or for 30 min at 0°C). All reagents, freshly prepared, were placed on the rim of the tube and mixed subsequently by centrifugation. Final concentrations (final volume 10 μl) were 100 μM for Fe(II) ($Fe(NH_4)_2(SO_4)_2.6H_2O$) and EDTA, 1.5 mM for ascorbate and 0.0015-0.006% for H_2O_2. Buffer conditions were 20 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$ and 50 mM NaCl. Yeast tRNA (5 μg) was added as carrier. The reactions were stopped by adding thiourea (20 μl of a 0.1 M solution), which serves to quench the free-radical reaction, followed by a phenol extraction and ethanol precipitation. The samples were analysed on a 10%

denaturing gel. Fe(II)EDTA eliminates nucleoside moieties from the RNA to generate products with both 5' and 3'-phosphorylated termini. Consequently, the fragments produced migrate faster relative to corresponding fragments in the T1 lane (22, 23).

RESULTS

Structure probing of U1A mRNA

A previously proposed secondary structure of the conserved region of the 3' UTR of the U1A mRNA, called Ag RNA, is shown in Figure 1 and consists of two distinct parts which are separated by only two nucleotides, U56 and A57 (16). The 5' part, which has a very symmetric structure, contains three stems (numbered 1, 2, and 3), separated by two asymmetrical internal loops containing the Box 1 and 2 sequences, which are required for U1A binding. Two single unpaired nucleotides, A24 and C50, are present on the strand opposite Box 2 and Box 1, respectively. The 3' part of the structure is a stemloop with the AUUAAA polyadenylation signal occupying most of the loop.

Enzymatic structure probing experiments (16) clearly showed that the central three nucleotides in Box 1 and 2, and also the polyadenylation signal (loop 4) are single-stranded. The presence of the highly conserved stems 2 and 3, which are needed for U1A protein binding, was clearly established by RNase V1 cleavage and by analyses of structure and function of mutant mRNAs (16). However, the behavior of a few other parts of the structure was less easy interpretable. Stems 1 and 4 showed cleavage both by RNase V1 and by single-stranded-specific ribonucleases. This indicates that these two stems, which have not been strongly conserved in evolution, and which seem not important for either U1A protein binding or inhibition of polyadenylation by the U1A protein, are of weak stability or may not exist at all in solution (16). Furthermore, the tetraloop of stem 3 (nucleotides 30-33) was hardly cleaved by ribonucleases under native conditions, suggesting that its structure might be very compact. A similar behavior was found for the unpaired nucleotides A24 and C50. This could arise either from the fact that these two nucleotides are located inside the helix or from the fact that ribonucleases, because of their bulky size, are very sensitive to steric hindrance. In contrast, chemical probes, which are of small size, are not very sensitive to steric hindrance and therefore can provide more detailed insight in the mRNA structure at the atomic level.

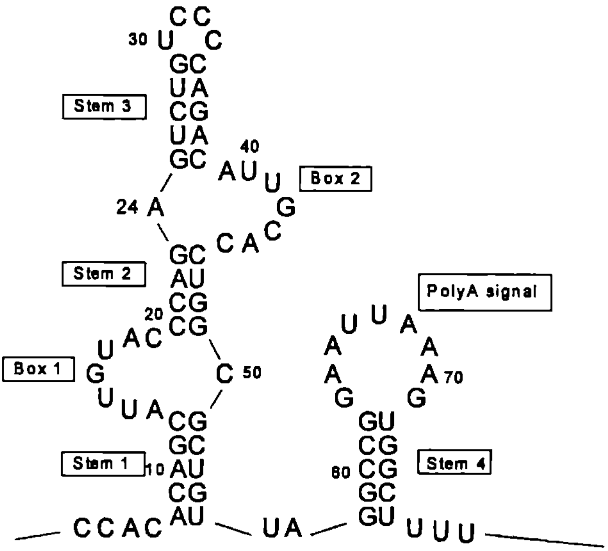


Figure 1. Secondary structure and nomenclature of the Ag RNA, the conserved region of the 3'UTR of the human U1A pre-mRNA (taken from (16)). The Box sequences, the polyadenylation signal and stems 1, 2, 3 and 4 are indicated.

The four bases were monitored at their Watson-Crick base-pairing positions by dimethylsulfate (DMS) at N1-A and N3-C and by carbodiimide (CMCT) at N3-U and N1-G. Position N7 of guanine and adenine residues was probed by DMS and diethylpyrocarbonate (DEPC), respectively. The experiments were performed under native conditions (N), semi-denaturing conditions (SD) and denaturing conditions (D). Tertiary interactions are generally less stable than Watson-Crick interactions and are expected to melt under semi-denaturing conditions (20). Experiments under such conditions will also give information about the stability of the different helical domains. Ag RNA was probed both at 20°C and at 0°C. The latter temperature was used to minimize the breathing in this relatively small RNA molecule, a phenomenon observed at 20°C (see below).

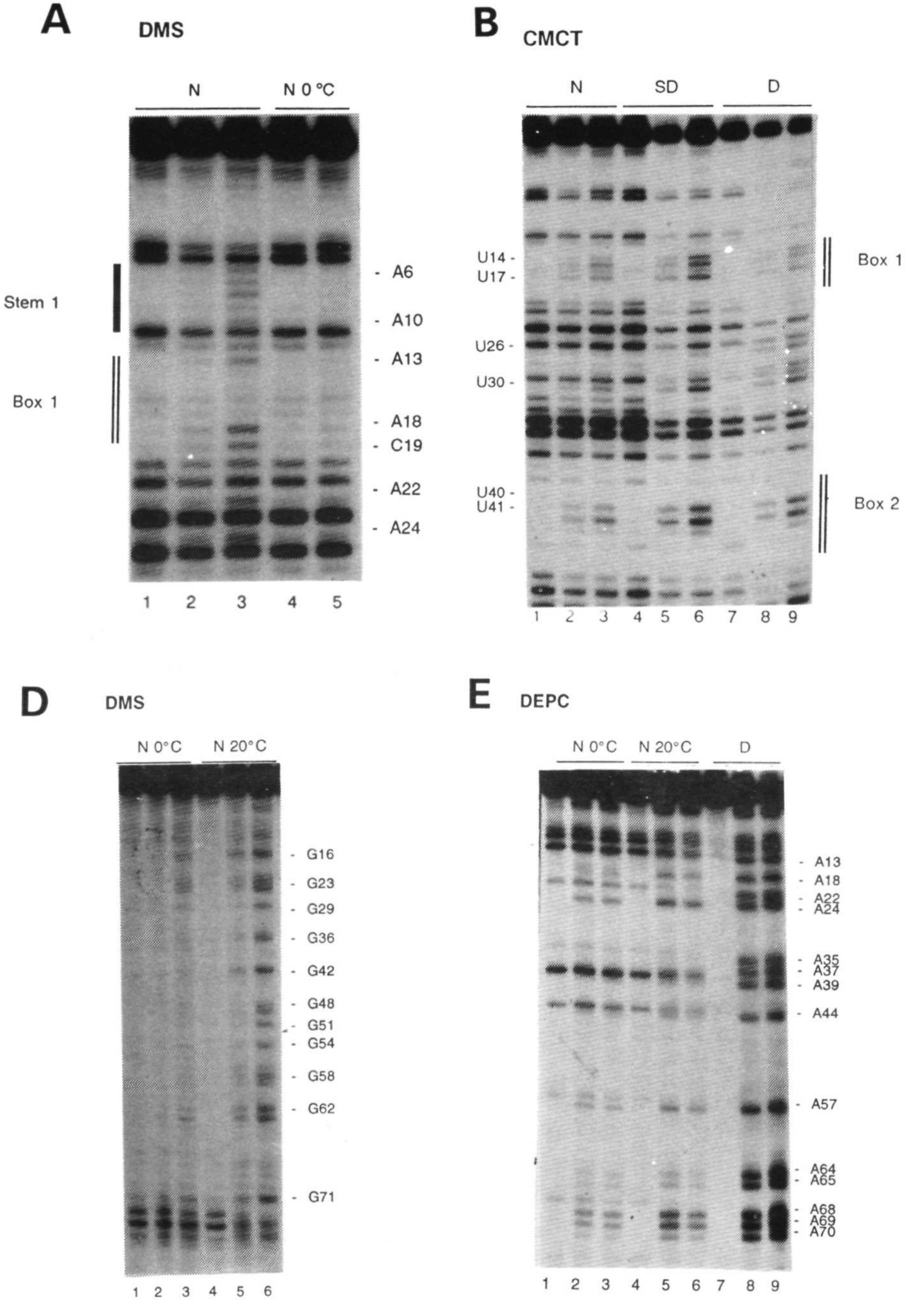
Figures 2A through 2E show examples of the chemical probing results for Ag RNA, while Figure 3 summarizes the results of several independent probing experiments both at 20°C (Figures 3A, 3B) and at 0°C (Figure 3C).

Stem regions

Stems 2 and 3. At 20°C the presence of stems 2 and 3 is clearly supported by the chemical modification data, since many nucleotides are only reactive under denaturing conditions. This is shown, for example, for nucleotides U26 and U28 in stem 3 in Figure 2B (lane 9) for the CMCT reaction. Their counterparts in stem 3, A37 and A35, respectively, are reactive with DMS (data not shown), as is A22 in stem 2 (see Figure 2A, lane 3). This difference in reactivity between adenosines and uridines in A-U base pairs has also been observed in helical regions of other RNAs (19, 20, 24) and is probably related to the fact that the N1-A can be modified by the relatively small DMS molecule (M_r=126), while the N3-U on the opposite strand cannot be modified by the more bulky CMCT reagent (M_r=423).

Concerning the N7 positions of the purines in stems 2 and 3, the guanosines are on the average more reactive towards DMS than the adenosines towards DEPC (see Figures 3B and 3C for a summary). This difference occurs because DEPC is larger than DMS, and in this way more sensitive to stacking (25, 26), and is, in our case most clearly visible in the zero degrees experiments. At this low temperature many N7-G positions in stems 2 and 3 are still accessible, although their reactivity is clearly reduced as compared to 20°C (Figure 2D, lanes 3 and 6), while N7 atoms of A35 and A37 are no longer available for modification (Figure 2E, lanes 3 and 6). Guanosines 23, 25, 49 and 51 are bordering the two internal loops where they are likely to be more accessible, a behavior also found in other RNA internal loops (19).

Stems 1 and 4. In agreement with the enzymatic probing (16), the chemical probing experiments show that stems 1 and 4 are of weak stability and are breathing at 20°C. In stem 1, many nucleotides are reactive at the Watson-Crick positions at this temperature (See Figure 2A, lane 3, for nucleotides 6-10), and the same is true for the N7-positions of G51 and G54 in stem 1 (see Figure 2D, lane 6). When we lower the temperature to 0°C, the Watson-Crick positions of nucleotides 6-10 can no longer be modified by DMS (Figure 2A, lane 5). Only U55 can still react with CMCT (data not shown), but this nucleotide is located at the end of stem 1, and thus is likely to be more reactive. Also the reactivity of the N7-atoms of the purines was diminished at 0°C. In stem 4 several nucleotides show reactivity at 20°C, both at their Watson-Crick and N7 -positions (see Figure 3 for a summary), but when the temperature is lowered the nucleotides can no longer be modified, or show much less reactivity. Taken together, these results suggest that stem 1 and 4


A
DMS
N
N 0 °C
Stem 1
Box 1
A6
A10
A13
A18
C19
A22
A24
1 2 3 4 5
B
CMCT
N
SD
D
U14
U17
U26
U30
U40
U41
Box 1
Box 2
1 2 3 4 5 6 7 8 9
D
DMS
N 0°C
N 20°C
G16
G23
G29
G36
G42
G48
G51
G54
G58
G62
G71
1 2 3 4 5 6
E
DEPC
N 0°C
N 20°C
D
A13
A18
A22
A24
A35
A37
A39
A44
A57
A64
A65
A68
A69
A70
1 2 3 4 5 6 7 8 9

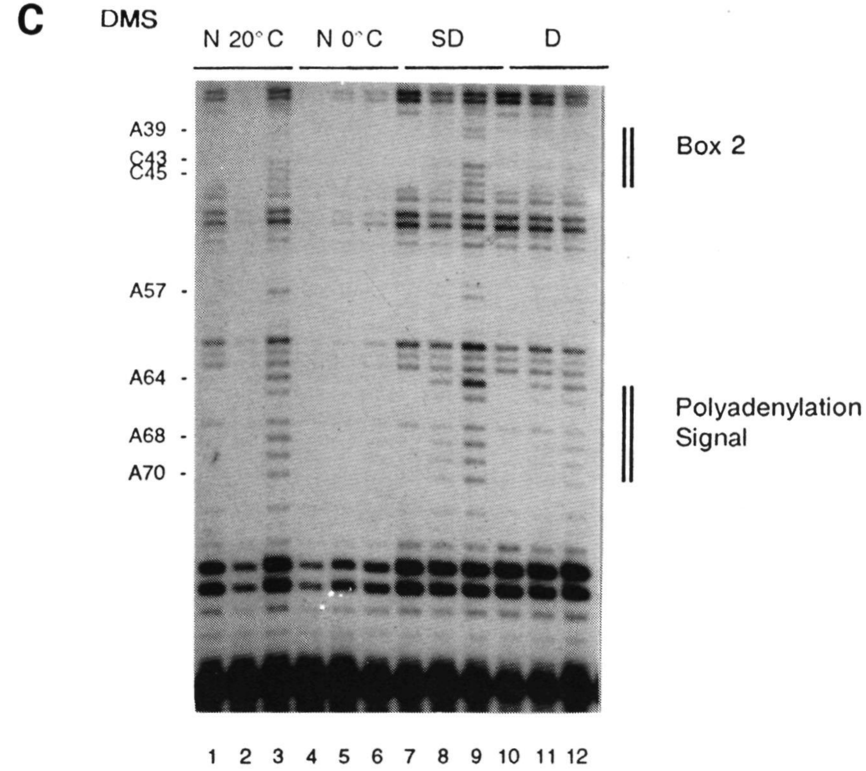


Figure 2. Structure probing of Ag RNA. **(A) (opposite page)** Chemical probing of Ag RNA with DMS at room temperature and at 0 °C. Detection of modifications was done by primer extension. Samples in lanes 1 and 4 are control incubations in which reagent was omitted. The reaction conditions are indicated: N (native conditions, 20°C) and N 0°C (native conditons, 0°C). Lane 2: 0.5 μl DMS incubated for 15 min. Lanes 3 and 5: 1.5 μl DMS incubated for 15 min. **(B) (opposite page)** Chemical probing of Ag RNA with CMCT. Detection of modifications was by primer extension. Samples in lanes 1, 4 and 7 are control incubations in which reagent was omitted. Lanes 2, 3, 5 and 6: 50 μl CMCT incubated at room temperature for 20, 30, 5 and 10 min, respectively. Lanes 8 and 9: 10 and 25 μl CMCT, incubated for 1 min at 90°C. **(C) (above)** Chemical probing of Ag RNA with DMS at room temperature and at 0°C. Detection of modifications was by primer extension. Samples in lanes 1, 4, 7, 10 are control incubations in which reagent was omitted. The reaction conditions are indicated: N (native conditions), SD (semi-denaturing conditions), D (denaturing conditions). Lanes 2 and 5: 0.5 μl DMS incubated for 7 min. Lanes 3 and 6: 1.5 μl DMS incubated for 7 min. Lanes 11 and 12: 0.5 and 1 μl DMS incubated for 30 s at 90°C. **(D) (opposite page)** Chemical probing of N7-G positions of 3'-end-labeled Ag RNA with DMS. The reaction conditions are indicated at the top of the figures. Samples in lanes 1, 4 are control incubations in which reagent was omitted. Lanes 2 and 5: 1 μl DMS incubated for 10 min. Lanes 3 and 6: 2 μl DMS incubated for 10 min. **(E) (opposite page)** Chemical probing of N7-A positions of 3'-end-labeled Ag RNA with DEPC. Samples in lanes 1, 4, 7 are control incubations in which reagent was omitted. The reaction conditions are indicated at the top of the figures: N (native conditions), D (denaturing conditions). Lanes 2 and 3: 10 and 20 μl DEPC incubated for 1 hr at 0°C. Lanes 5 and 6: 10 and 20 μl DEPC incubated for 50 min at room temperature. Lanes 8 and 9: 2 and 4 μl DEPC incubated for 4 min at 90°C.

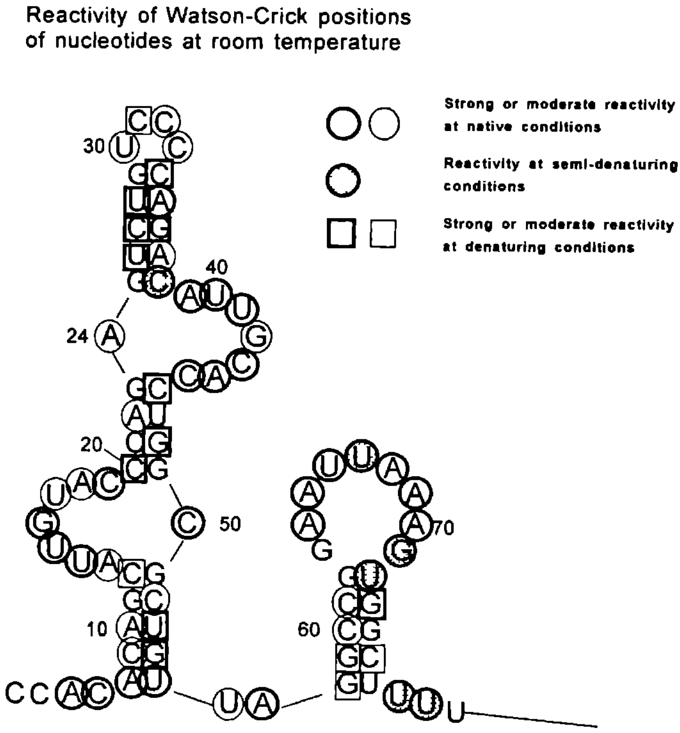

Reactivity of Watson-Crick positions
of nucleotides at room temperature
Strong or moderate reactivity at native conditions
Reactivity at semi-denaturing conditions
Strong or moderate reactivity at denaturing conditions
30
40
24
20
50
10
60
70
CC

Figure 3. Summary of the chemical structure probing of the 3' UTR of U1A mRNA (A) Reactivities of Watson-Crick positions of nucleotides in Ag RNA towards chemical probes at 20°C. Consensus data from several independent experiments using both primer extension and end-labeled detection are shown. Reactivity towards the chemical probes is indicated with symbols which are explained in the Figure. Nucleotides for which no reactivity is indicated show RT-stops in the primer extension reactions. *(Figure continued on next page).*

indeed can be formed, although they are of weak stability at 20°C.

Loop regions and linker region

Box 1 and Box 2 regions. All nucleotides in the Box 1 and 2 sequences are accessible at their Watson-Crick positions at 20°C (see Figure 3A for a summary of the data). Figure 2A (lane 3) shows accessibility to DMS of A13, A18 and C19 in Box 1 and Figure 2C (lane 3) of A39, C43, A44 and C45 in Box 2. Figure 2B (lane 3) shows the accessibility to CMCT of U14 to U17 in Box 1, and of U40 to G42 in Box 2. At 0°C a few nucleotides at the 5' part of Box 2 become inaccessible at their Watson-Crick positions (Figures 2C, compare lanes 3 and 6). In Box 1 the N1

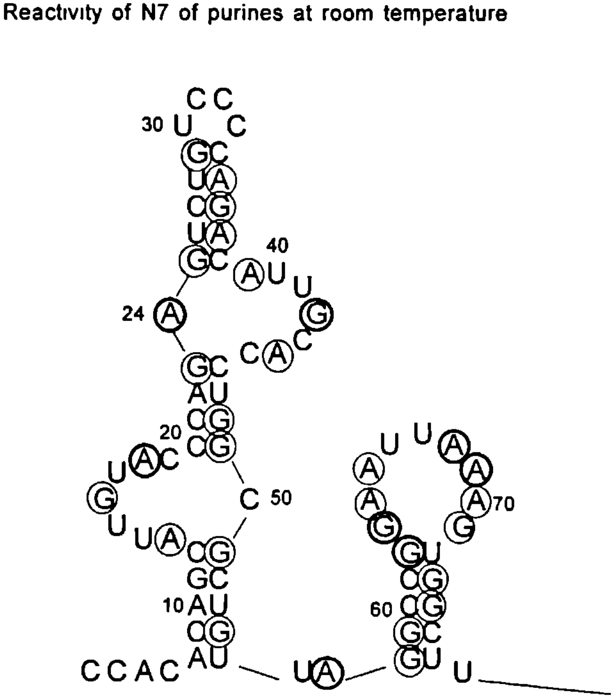


Figure 3 (continued). (B) Reactivities of N7-positions of purines in Ag RNA towards chemical probes at 20°C. Consensus data from several independent experiments using 3'-end-labeled U1A mRNA are shown. Symbols are identical to those used in Figure 3A. *(Figure continued on the next page).*

atom of A13 is no longer accessible while A18 and C19 show reduced accessibility (Figure 2A, lane 5). This behavior is probably due to stacking of the bases and this agrees with the RNase V1 cleavage found at the 5' parts of both Box sequences (16). It must be noted, however, that both Box sequences do not behave exactly the same at 0°C (see Discussion).

Although RNases were unable to cleave the unpaired nucleotides A24 and C50, our chemical probing results show that C50 is strongly reactive and that the N1 atom of A24 (Figure 2A, lane 3) is moderately reactive at 20°C. For C50 this had to be deduced from reactions with 3'-end-labeled RNA (data not shown) due to the occurrence of a natural stop of reverse transcriptase at C50 in the primer extension reactions. Note that in A24 also the N7 atom is also available for modification (Figure 2E, lane 6). At 0°C, N1 of A24 is no longer accessible (Figure 2A, lane 5), probably due to stacking, but the N7-atom of A24 still can be modified (Figure 2E,

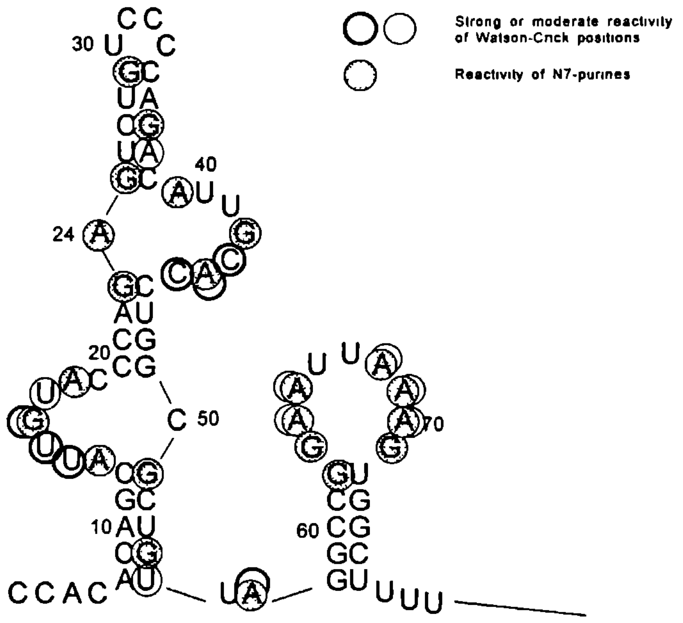

Reactivity of nucleotides at zero degrees
(Watson-Crick positions and N7-purines)

Figure 3 (continued). (C) Reactivities of Watson-Crick and N7-positions of nucleotides in Ag RNA towards chemical probes at 0°C. Consensus data from several independent experiments using both primer extension and end-labeled detection are shown. Reactivity towards the chemical probes is indicated with symbols which are explained in the Figure.

lane 3).

Tetraloop. Nucleotide U30 in the tetraloop is moderately reactive towards CMCT while the reactivity of cytosines 31-33 toward DMS is more difficult to evaluate due to the presence of RT-stops, especially at positions 33 and 34. By using 3'-end-labeled RNA, however, it was found that nucleotides 32 and 33 are moderately reactive at their N3 position, while N3 of C31 is only reactive at denaturing conditions (data not shown).

Loop 4 and linker region. In full agreement with the enzymatic probing, loop 4 is completely accessible at the Watson-Crick positions at native conditions (see Figure 2C, lane 3, for DMS results) with the exception of U67 which becomes accesible to CMCT only under SD conditions (data not shown). The N7 positions of the purines (see Figure 2D and 2E, lanes 6) are accessible and A68 through A70 seem more strongly modified than A64 and A65. At 0°C most of the nucleotides are still moderately accessible at both Watson-Crick and N7 positions, but the reactivity is clearly reduced as compared to 20°C (see Figures 2C lane 6, and 2D and 2E, lanes 3).

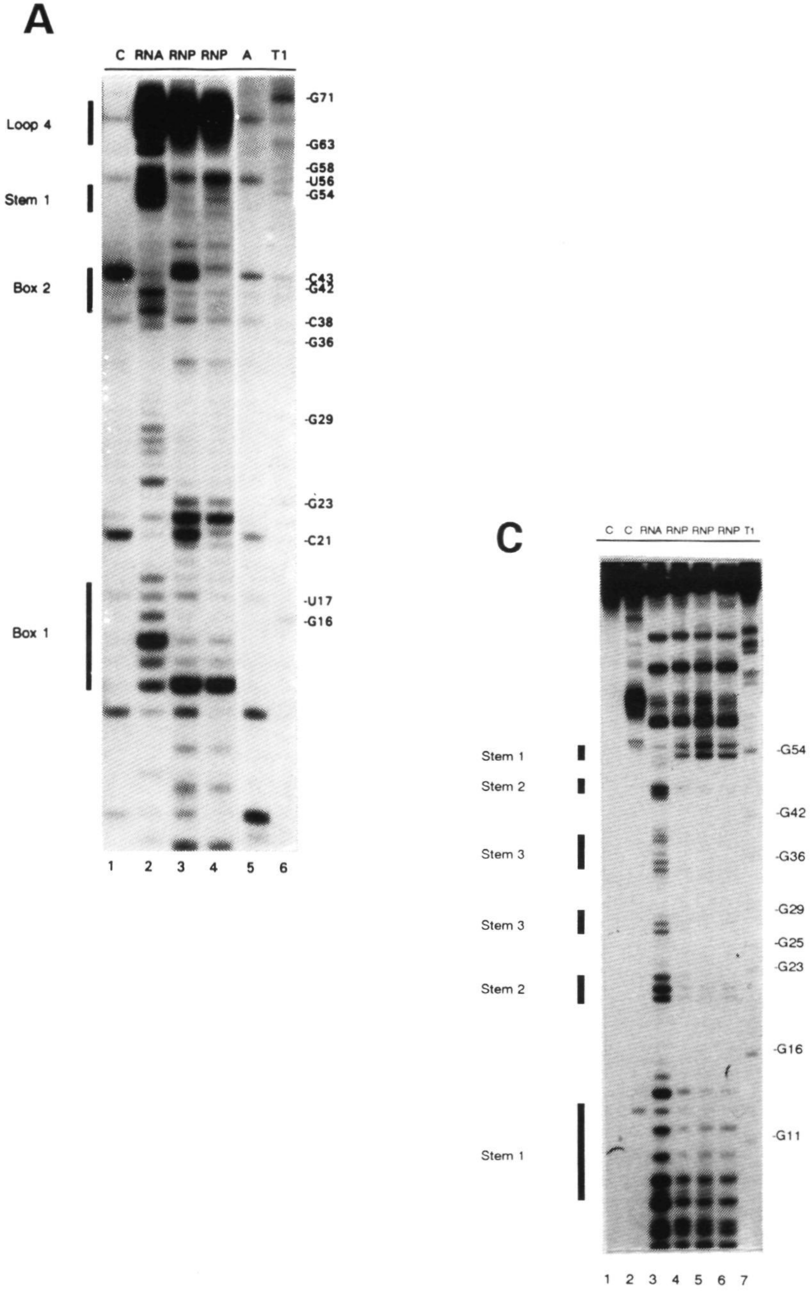
The linker region, which connects the 5' part with the 3' part of the structure, is formed by two nucleotides U56 and A57. Both nucleotides are fully accessible at room temperature (see Figure 2C, lane 3 for A57). At 0°C, U56 can no longer be modified but A57 is still accessible, both at N7 (Figure 2E, lane 3) and at N1 (see Figure 3C).

Analysis of the complex of U1A mRNA with U1A protein

To obtain information on the complex of U1A mRNA and U1A protein, footprinting experiments were performed using both various ribonucleases and Fe(II)EDTA. In these experiments 5'-end-labeled Ag RNA was incubated with an excess of U1A protein. The resulting RNP complexes were probed with RNases A, T1, T2, V1 or Fe(II)EDTA. Examples of RNase footprinting are shown in Figure 4A (RNase T2), Figure 4B (RNase A) and Figure 4C (RNase V1), while Figure 4D summarizes the results obtained by ribonuclease protection.

As might be expected, the Box 1 and 2 regions are almost completely protected by the U1A protein (compare lanes 2 and 3 in Figure 4A). The phosphodiester bond between C43 and A44 is a very sensitive spot in both RNA and RNP, which obscures clear interpretation of the protection pattern at that position. Such intrinsic fragility, especially for pyrimidine-adenosine bonds, is well known in RNA molecules (26). Nucleotide A13 in Box 1 becomes a hypersensitive site in the RNP complex (lane 3). The single-strand-specific RNases also cleave some nucleotides in the stem regions in the naked RNA, for instance nucleotides 26-28 in stem 3 and nucleotides in both stem halves of stem 1. These cleavages are absent or much weaker in the RNP complex (Figure 4A, compare lanes 2 and 3; see also Discussion).

The protection pattern obtained by RNase V1 in the presence of U1A wt protein (Figure 4C) shows that the stem regions in the 5' part of the RNA (stems 1, 2 and 3) become protected in the RNA-protein complex, while the 3' part remains unprotected (compare lanes 3 and 4). Footprinting experiments with RNase V1 and T2 in the presence of A101 (containing the N-terminal 101 amino acids of U1A) did not show significant differences in the protection patterns (data not


A
C
RNA
RNP
RNP
A
T1
Loop 4
Stem 1
Box 2
Box 1
-G71
-G63
-G58
-U56
-G54
-C43
-G42
-C38
-G36
-G29
-G23
-C21
-U17
-G16
1 2 3 4 5 6
C
C
C
RNA
RNP
RNP
RNP
T1
Stem 1
Stem 2
Stem 3
Stem 3
Stem 2
Stem 1
-G54
-G42
-G36
-G29
-G25
-G23
-G16
-G11
1 2 3 4 5 6 7

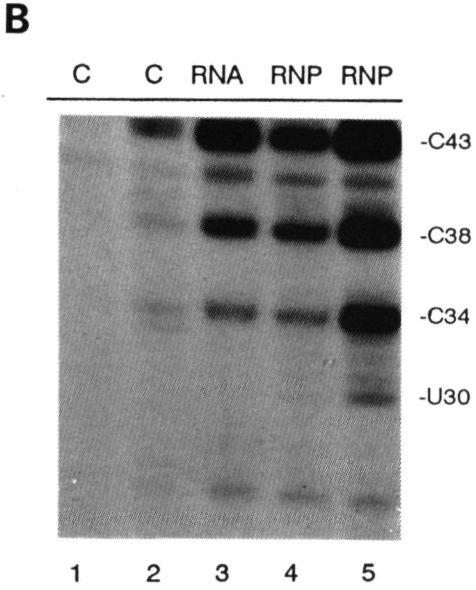
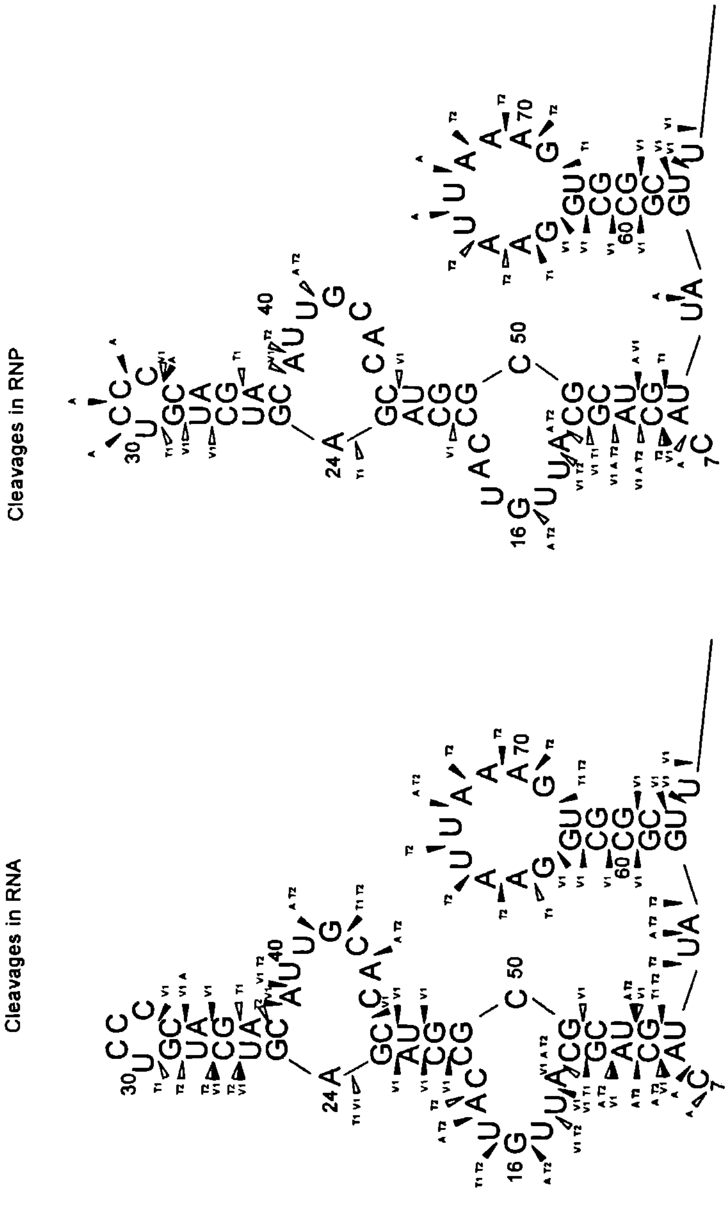


Figure 4. Enzymatic and chemical probing of the U1A - Ag RNA complex. **(A)** **(opposite page)** Enzymatic footprinting of the U1A-Ag RNA complex using 5'-end-labeled Ag RNA and single-strand-specific RNase T2: Lane 1: Control incubation where U1A protein and RNase T2 are omitted; Lane 2: RNA probed at room temperature for 10 min with RNase T2 (5 x 10^{-3} U); Lanes 3 and 4: RNA incubated with respectively 150 and 300 molar excess of U1A protein, probed with RNase T2 (5 x 10^{-3} U); Lanes 5 and 6: RNA probed under denaturing conditions with RNases A and T1, to obtain a sequence ladder for U/C and G, respectively. **(B)** **(above)** Enzymatic footprinting of the U1A-Ag RNA complex using 5'-end-labeled Ag RNA and single-strand-specific RNase A: The region around the tetraloop is shown. Lane 1: Control incubation in which U1A protein and RNase A are omitted; Lane 2: Control incubation in which both U1A mRNA and U1A protein (300-fold molar excess) are present but RNase A is omitted; Lane 3: RNA probed with RNase A (2 x 10^{-5} U) for 10 min at room temperature; Lanes 4 and 5: RNA with 150- and 300-fold excess of U1A protein, respectively, probed with RNase A as in lane 3. Note the accessibility of nucleotides U30, C31, C32 and C33. **(C)** **(opposite page)** Enzymatic footprinting of the U1A-Ag RNA complex using 5'-end-labeled Ag RNA and RNase V1. Lane 1: Control incubation in which U1A protein and RNase V1 are omitted; Lane 2: Control incubation in which both Ag RNA and U1A protein are present but RNase V1 is omitted; Lane 3: RNA probed at room temperature for 10 min with RNase V1 (0.06 U); Lanes 4, 5 and 6: RNA incubated with respectively 150, 300 and 500 molar excess of U1A protein, probed with RNase V1; Lane 7: RNA probed under denaturing conditions with RNase T1, to obtain a sequence ladder. *(Figure continued on next page).*



Cleavages in RNA

Cleavages in RNP

Figure 4 (continued, opposite page). (D) Summary of RNase data obtained for both the naked Ag RNA and the U1A-Ag RNA complex. On the left the digestion pattern of the RNA is shown (adapted from (16)), while on the right the digestion pattern of the U1A-Ag RNA complex is shown. Strong cleavages are indicated by solid arrows and weak cleavages by open arrows. The enzymes which do cut are indicated next to the arrows. In case of RNases V1 and T2, the data are both for U1A wt and for A101, while for experiments with RNases A and T1 only U1A wt was used. *(Figure continued on next page).*

shown). At the 3' side of stem 1 (nts 51-55) protection is found until nucleotide 53, while one V1 cleavage, between nucleotides 53-54, becomes stronger in the RNP as compared to the naked RNA.

Because nucleotides of the tetraloop in the naked RNA were not cleaved by RNases (see above and (16)) information about protection of this region was not expected to be obtained. Interestingly, however, the tetraloop becomes accessible to RNases in the RNP complex (see Figure 4B, compare lanes 3 and 5), indicating a structural change in this part of the RNA upon protein binding.

The 3' part of the structure (stem 4 and loop 4) does not show much protection (see Figure 4A), so this region appears to be accessible in the RNP complex. Only the 5' side of the polyadenylation signal (nucleotides 64-65) shows partial protection (compare lanes 2 and 4).

Next to using ribonucleases, RNA-protein interactions can also be analyzed by chemical nucleases (27), i.e. metal complexes that cleave nucleic acids with little or no dependence on the identity of the attached base. For example, Fe(II)EDTA complexes generate hydroxyl radicals in the presence of hydrogen peroxide or molecular oxygen. Hydroxyl radicals attack solvent-exposed riboses inducing strand scission of the RNA and in this way are able to discriminate between solvent-accessible and solvent-inaccessible (i.e. protected) regions. Furthermore, Fe(II)EDTA can be used to probe the conformation of naked RNA. It appears to act independently of the secondary structure, but can be used to analyze the tertiary folding of RNAs. In some cases, for example tRNA, ribose residues in the interior of an RNA molecule are protected from strand scission, while for example in 5S rRNA, only minor modulation in cleavage intensity along the molecule is found, indicating that this RNA possesses very little, if any, tertiary structure (22).

Results of Fe(II)EDTA probing of the U1A mRNA-U1A complex are summarized in Figure 4E while an example of a densitometer scan of the gel is shown in Figure 4F. The data obtained so far concern only nucleotides 10-50 and largely agree with the enzymatic protection data. The RNA lanes in our Fe(II)EDTA experiments in general show little modulation in intensity (data not shown), which indicates a lack of tertiary structure. This is in agreement with the

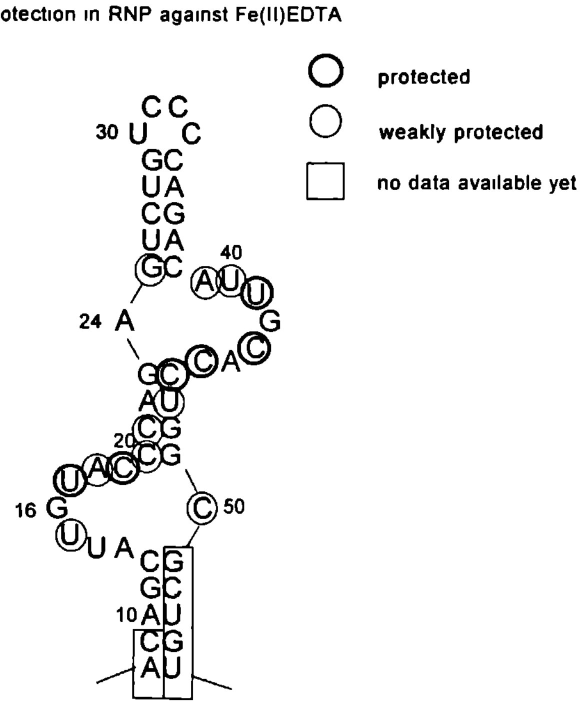


Figure 4 (continued). (E) Summary of Fe(II)EDTA footprinting of nucleotides 10-50 of the U1A-Ag RNA complex using 5'-end-labeled Ag RNA. The U1A wt protein is present in 150 fold excess and nucleotides which are protected against cleavage by hydroxyl radicals are shown in bold (strong protection) and light circles (weak protection). *(Figure continued on next page).*

results of the chemical probing of Ag RNA. In the RNA-protein complex, the majority of the nucleotides in Boxes 1 and 2 are protected against hydroxyl radicals. Nucleotides G16 and G42, which are positioned symmetrically in the structure, are not protected. This agrees with the enzymatic data, which show accessibility of the 5'phosphates of these two guanosines in the RNP complex (see Figure 4C). However, the behaviour of the loop nucleotides is not completely symmetrical. In Box 1, the nucleotides at the 5' side of the Box (A13 and U14) are more accessible to radicals than the corresponding nucleotides in Box 2 (A39 and U40). Some protection is found in stem 2 but stem 3 shows almost no protection as is the case with the 5' side of stem 1 (for the 3' side of stem 1 no data are available yet).

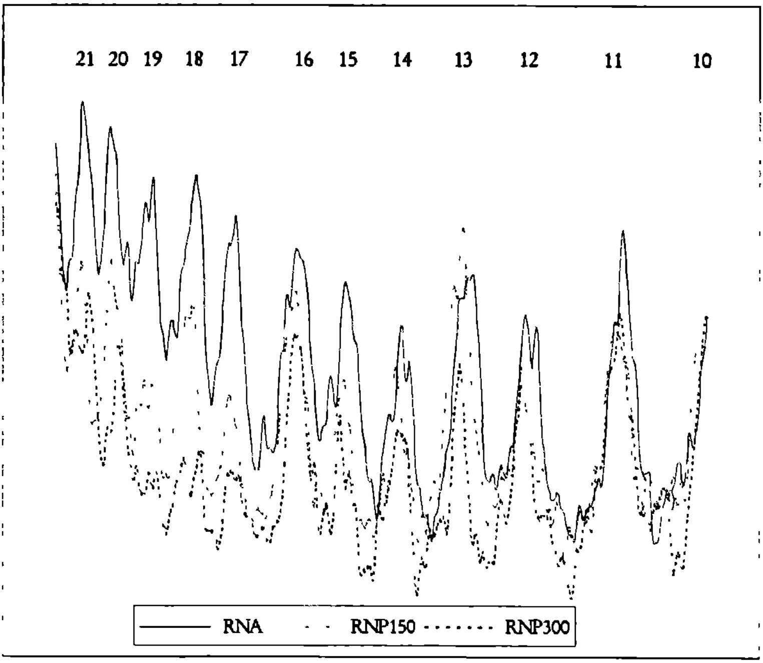


Figure 4 (continued). (F) Imaging densitometer scans of hydroxyl radical footprints of the U1A-Ag RNA complex. Data are shown for both 150 (RNP150) and 300-fold (RNP300) excess of U1A wt protein. The region of the Ag RNA shown is nucleotides 10-21. Box 1 is located between nucleotides 13 and 19.

DISCUSSION

Secondary structure of U1A mRNA

We probed the conserved region of the 3'UTR of U1A mRNA at nucleotide resolution by the utilization of structure-specific probes. The secondary structure obtained is in accord with the structure predicted previously which was based upon enzymatic probing and analysis of structure and function of mutant RNAs (16), but contains a number of additional features.

At 20°C, the highly conserved stems 2 and 3 are indeed present while stems 1 and 4 also exist but these are not very stable and probably breathing. At 0°C, all four stems clearly exist in the structure.

At 20°C all nucleotides in the Box 1 and 2 regions are fully accessible at both their Watson-Crick positions and at the N7-atoms of the purines. This behavior excludes the presence of tertiary interactions between these nucleotides and other parts of the RNA. At 0°C, several nucleotides in the Box regions are no longer accessible, probably because of stacking. Interestingly, the behavior of the two Box regions at 0°C is identical. Box 1 shows more reactivity of both Watson-Crick and N7-positions at its 5' end, while Box 2 shows most reactivity at the 3' end. A (somewhat) different structure of the two Boxes could be expected because the two sequences, although almost identical in sequence, have a different structural context in the U1A pre-mRNA and also differ in U1A binding capacity. Box 2 forms a much stronger (~30 fold) binding place for U1A protein than Box 1 (16).

The two unpaired nucleotides A24 and C50 are clearly accessible at 20°C. Whether the accessibility of A24 and C50 results from looping out of the helix or from the fact that the structure of the RNA is more open at the internal loops is not known. The fact that N7 of G25 also can be modified supports the possibility that A24 is not stacked in the helix.

Cleavage in the tetraloop by RNases was not observed (16), but chemical probing showed that in the RNA three of the four tetraloop nucleotides are moderately accessible at 20°C. N3 of C31 was only reactive at denaturing conditions, which could either indicate stacking or point to an involvement in a tertiary interaction under native conditions.

The chemical probing results clearly indicate the presence of stemloop 4. In the loop containing the AUUAAA polyadenylation signal, it can be seen that all adenosines are reactive at both the N1 and the N7 position, and this behavior persists at 0°C.

In conclusion, the probing studies provide a secondary structure for the Ag RNA as shown in Figure 3. Both the chemical probing and the Fe(II)EDTA results suggest that there are hardly any tertiary interactions present between different domains of the Ag RNA, which is reminiscent of the behavior of 5S rRNA towards chemical and enzymatic probes (28) and hydroxyl radicals (22).

The U1A mRNA - U1A protein complex

Footprinting experiments have been performed on the complex of U1A mRNA and the U1A protein. In the first set of experiments, several ribonucleases were used. Inhibition of reactivity at certain nucleotides can be inferred as a direct protection (and hence contact) of the RNA by the protein at that site. However, reduced reactivity can also be caused by conformational changes in the RNA chain brought about by the addition of the protein, and it is not easy or impossible to distinguish between these two modes of protection. Furthermore, since RNases are large molecules, steric hindrance may significantly enlarge the protected regions. For this latter reason, a probe with small size, the hydroxyl radical, has been used in a second set of experiments. Cleavage of RNA with Fe(II)EDTA appears relatively independent of the secondary structure, and its uniform reactivity makes it an excellent probe (22). However, the technique is tedious and not always as reproducible as one would like.

Both the ribonuclease and Fe(II)EDTA protection experiments show that the Box 1 and Box 2 regions, as might be expected, are largely protected when the U1A protein is present. Clearly, these sequences, which have been shown to be important for U1A binding to U1A mRNA (15), are in contact with the U1A protein. Only some nucleotides, located at the 5'-side of both Boxes, can be attacked by ribonucleases (A13, U15, A39 and U41) and hydroxyl radicals. Surprisingly, nucleotides G16 and G42, both localized at the center of a Box sequence, show no protection against the small hydroxyl radical. This lack of protection of G16 and G42 possibly reflects the fact that the loop turns sharply there, resulting in an exposed ribose.

All nucleotides in stems 1, 2 and 3 show complete or partial protection against ribonucleases in the presence of U1A protein, and also in the presence of A101, which contains only the N-terminal RNP motif of U1A. Around nucleotide 54 RNase V1 cleavage is enhanced when U1A protein is added. This might indicate that stem 1 becomes more stable as a result of U1A binding. Alternatively it could indicate a stacking of stems 1 and 4 onto each other. The reduction of cleavages by single-strand-specific RNases in stem 1 and 3 could indicate protection of these regions by the U1A protein, but could also be the result of a further stabilization of the double-stranded region upon binding of U1A protein. Footprinting of the complex of stem-loop II of U1 snRNA and U1A protein has been performed with both RNase V1 and ethylnitrosourea (5). However, in that case only one of the stem halves 'of stem-loop II appeared to be protected against the probes. This difference in behaviour of U1 snRNA (5) as compared to U1A mRNA (our results) can be due to the difference in size of the RNA substrates or to the presence of two rather bulky U1A proteins in the latter case, instead of one in case of U1 snRNA.

Interestingly, in the RNP complex the nucleotides in the tetraloop (loop 3) seem to become accessible. This may indicate that this loop undergoes a conformational change upon U1A protein binding. It appears that the loop opens up, with its nucleotides becoming available to the probes. Such behaviour is also found in bacteriophage R17 where a hairpin tetraloop structure is becoming more open upon R17 coat protein binding (29).

The 3' part of the structure is formed by stem-loop 4 and this part shows, as expected, no protection, except for some limited changes at the 5' side of the loop. This means that this region is accessible in the RNP complex, a finding which is in complete agreement with the finding that U1A does not interfere with the binding of the cleavage polyadenylation specificity factor (CPSF) to the polyadenylation signal during polyadenylation of the mRNA (17).

In conclusion, we have obtained detailed information concerning the structure of Ag RNA and its complex with U1A protein. This information allowed us to build a three-dimensional model for the conserved region of U1A mRNA and a possible tertiary structure model for this particular RNA-protein complex. Such a model will be discussed in the addendum of this chapter.

ACKNOWLEDGEMENTS

The authors want to thank Drs. Pascale Romby and Chantal Ehresmann (Institut de Biologie Moléculaire et Cellulaire, Strasbourg, France) for advice with the chemical probing reactions and Drs. Robert Walzcak and Herve Moine (Institut de Biologie Moléculaire et Cellulaire, Strasbourg, France) for advice concerning the Fe(II)EDTA footprinting experiments. The research was supported in part by the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organization for Scientific Research (NWO).

REFERENCES

1. Luhrmann, R, Kastner, B. and Bach, M. (1990) *Biochim. Biophys. Acta*, 1087, 265-292.
2. Scherly, D., Boelens, W., Van Venrooij, W.J., Dathan, N.A., Hamm, J. and Mattaj, I.W. (1989) *EMBO J.*, 8, 4163-4170.
3. Lutz-Freyermuth, C., Query, C C. and Keene, J.D. (1990) *Proc. Natl. Acad. Sci., USA*, 87, 6393-6397.
4. Nagai, K., Oubridge, C., Jessen, T.H., Li, J. and Evans, P.R. (1990) *Nature*, 348, 515-520.
5. Jessen, T H., Oubridge, C., Teo, C.H., Pritchard, C. and Nagai, K. (1991) *EMBO J.*, 10, 3447-3456.
6. Hall, K.B. and Stump, W.T. (1992) *Nucl. Acids Res.*, 20, 4283-4290.
7. Hoffman, D W., Query, C.C., Golden, B.L., White, S.W. and Keene, J.D. (1991) *Proc. Natl. Acad. Sci., USA*, 88, 2495-2499.
8. Scherly, D., Boelens, W., Dathan, N.A., Van Venrooij, W.J. and Mattaj, I.W. (1990) *Nature*, 345, 502-506.
9. Bentley, R.C. and Keene, J.D. (1991) *Mol. Cell. Biol.*, 11, 1829-1839.
10. Tsai, D E., Harper, D.S. and Keene, J.D. (1991) *Nucl. Acids Res.*, 18, 4931-4936.
11. Howe, P.W.A., Nagai, K., Neuhaus, D. and Varani, G. (1994) *EMBO J.*, 13, 3873-3881.

12 Hall, K B (1994) *Btochem* , 33, 10076-10088

13 Stump, W T and Hall, K B (1995) *RNA*, 1, 55-63

14 Oubndge, C , Ito, H , Evans, P R , Teo, C H and Nagaı, K (1994) *Nature*, 372, 432-438

15 Boelens, W C , Jansen, E J R , Van Venrooıj, W J , Stripecke, R., Mattaj, I W and Gunderson, S I (1993) *Cell*, 72, 881 892

16 Van Gelder, C W G , Gunderson, S I , Jansen, E J R., Boelens, W C , Polycarpou Schwarz, M , Mattaj, I W and Van Venrooıj, W J (1993) *EMBO J*, 12, 5191-5200

17 Gunderson, S I , Beyer, K , Martın, G , Keller, W , Boelens, W C and Mattaj, I W (1994) *Cell*, 76, 531-541

18 Nelıssen, R.L H , Sıllekens, P T G , Beıjer, R P , Van Kessel, A H M G and Van Venrooıj, W J (1991) *Gene*, 102, 189-196

19 Van Gelder, C W G , Thıjssen, J P H M , Klaassen, E C J , Sturchler, C , Krol, A , Van Venrooıj, W J and Pruıjn, G J M (1994) *Nucl Acıds Res* , 22, 2498-2506

20 Krol, A and Carbon, P (1989) *Methods Enzymol* , 180, 212-227

21 Kwakman, J H , Konıngs, D A , Hogeweg, P , Pel, H J and Grivell, L A (1990) *J Biomol. Struct. Dyn.*, 8, 413-430

22 Darsıllo, P and Huber, P W (1991) *J Biol. Chem.*, 266, 21075-21082

23 Celander, D W and Cech, T R (1990) *Biochem.*, 29, 1355 1361

24 Glıckman, J N , Howe, J G and Steıtz, J A (1988) *J Virol.*, 62, 902-911

25 Dock-Bregeon, A C , Westhof, E , Gıege, R and Moras, D (1989) *J Mol. Biol.*, 206, 707-722

26 Mougel, M , Eyermann, F , Westhof, E , Romby, P , Expert-Bezancon, A , Ebel, J-P , Ehresmann, B and Ehresmann, C (1987) *J Mol. Biol* , 198, 91 107

27 Huber, P W (1993) *FASEB J* , 7, 1367-1375

28 Chrıstıansen, J , Brown, R S , Sproat, B S and Garrett, R A (1987) *EMBO J* , 6, 453 460

29 Varanı, G (1995) *Annu. Rev Biophys Biomol. Struct.*, 24, 379-404

Addendum: Towards a three-dimensional model of the complex of Ag RNA with U1A protein

Celia W.G. van Gelder, Sander W.M. Teunissen and Walther J. van Venrooij

University of Nijmegen, Department of Biochemistry, PO Box 9101, 6500 HB Nijmegen, The Netherlands

ABSTRACT

With the help of the structure probing data of U1A mRNA and of its complex with U1A protein, we started with the construction of a three-dimensional model of this RNA-protein complex using computer modeling techniques. First, a model of the interaction of one of the U1A-binding regions of U1A mRNA with one RNP motif of U1A was created. This interaction is most probably similar as found in the complex of U1A protein with U1 snRNA stemloop II. We then postulated that the second U1A binding region of the mRNA is positioned in the same way on the N-terminal RNP motif of the second U1A protein.

After this, a possible orientation of the two U1A-binding sites in U1A mRNA, and therefore of the two U1A proteins, is discussed in the light of the currently available experimental data. At this stage of the modeling, the working model still is very crude. Further experiments are needed to test and refine it, and to determine the precise orientation of the two U1A proteins.

INTRODUCTION

In the two preceding chapters of this thesis, investigations concerning the structure of the conserved region of the U1A mRNA (Ag RNA) have been described. Furthermore, the complex of the U1A pre-mRNA was investigated by footprinting analyses using both ribonucleases and hydroxyl radicals. With this information an attempt can be made to postulate a three-dimensional model for Ag RNA and its complex with U1A protein.

The U1A protein contains two so-called RNP motifs (1), of which the N-terminal copy is responsible for binding to U1 snRNA (2, 3), and to U1A mRNA (4). The function of the C-terminal RNP motif of U1A is not known yet, but this domain does not appear to bind RNA (5). As described earlier in this thesis, the RNP motif contains a ßαßßαß fold, in which a ß-sheet formed by four antiparallel ß-strands is flanked at one side by two α-helices. The conserved RNP1 and RNP2 segments are located in the two central ß-strands (ß1 and ß3). In the U1A protein a Tyr residue in RNP2 could be crosslinked to a nucleotide in the second stemloop of U1 snRNA (6) and mutagenesis experiments showed that U1 snRNA stemloop II binds to the surface of the four-stranded ß-sheet, as well as to loops at one edge of the sheet (7). All these data agree with recent results of NMR experiments performed on the complex between hnRNP C protein and rU_8 (8, 9) and on the complex of the N-terminal RNP motif of U1A and stemloop II of U1 snRNA (10, 11). In the latter complex, the U1 snRNA hairpin loop, which is largely disorded in the absence of protein, becomes ordered upon protein binding (11). Most recently, a co-crystal structure has been described of the complex between the N-terminal RNP motif of U1A and an RNA substrate containing 21 nucleotides of stemloop II of U1 snRNA (12).

MATERIALS AND METHODS

Probing data of RNA and RNP

The chemical and enzymatic probing data for Ag RNA and for the U1A-Ag RNA complex are taken from Chapters 4 (13) and 5 of this thesis.

U1A protein coordinates

For U1A protein the X-ray structure of its N-terminal RNP motif was solved at 2.8 Å (7). Only the Cα coordinates of this structure are available in the Brookhaven Protein Databank (14). Full coordinates for U1A have been generated as described in Chapter 2 of this thesis (15).

RNA model building

The RNA secondary structure was divided into elementary motifs (helices, loops), which were assembled into a three-dimensional structure by using a computer graphics station and SYBYL software (16). The following principles were observed. First, the major interaction stabilizing RNA structure is base stacking, which is short-range and controlled by the nearest neighbors, followed by hydrogen bonding between complementary or non-canonical bases. Secondly, the sugar-phosphate backbone preferentially adopts right-handed helical conformations with the bases in the *anti* conformation and the ribose sugar in the C3'-*endo* pucker. Thirdly, the model must be consistent with the results obtained by enzymatic and chemical probing of the RNA and the RNP complex.

RESULTS AND DISCUSSION

Secondary structure of Ag RNA and protection data of the Ag RNA-U1A complex

Figure 1A shows the secondary structure of Ag RNA as deduced from our previous studies (Chapters 4 (13) and 5). The 5' part of the structure is involved in U1A protein binding, and the Box 1 and 2 sequences are the main determinants for this (4). In Figure 1B, the second stemloop of U1 snRNA is shown, which is the other substrate to which U1A protein binds. The first seven nucleotides of this loop, AUUGCAC, are critical for U1A protein binding (2, 17), while the three nucleotides at the 3' side of the loop are functioning as a kind of spacer and can be replaced by non-nucleotide linkers without disturbing complex formation (Dr. K. Hall, pers. communication). Furthermore, the loop sequence of 10 nucleotides has to be constrained by a stem of which the sequence does not seem important (5, 11). In Ag RNA, the Box 2 region, which contains a sequence identical to the U1A-binding region of U1 snRNA, can bind U1A protein with high affinity, while Box 1, which contains a sequence in which 6 out of 7 nucleotides are identical to the U1A binding sequence, shows a 30-fold lower affinity for U1A protein (13).

Nucleotides which are protected by ribonucleases or by hydroxyl radicals in the Ag RNA-U1A protein complex are shown in Figures 4D and 4E of Chapter 5. It was shown that the Box 1 and 2 regions are almost completely protected against ribonucleases by the U1A protein. It must be noted that the secondary structure of the 5' part of the conserved region of Ag RNA shows an approximate two-fold symmetry axis and this axis is depicted in Figure 1A. Furthermore, this symmetry is also found in the protection behaviour of some of the nucleotides. For example, the

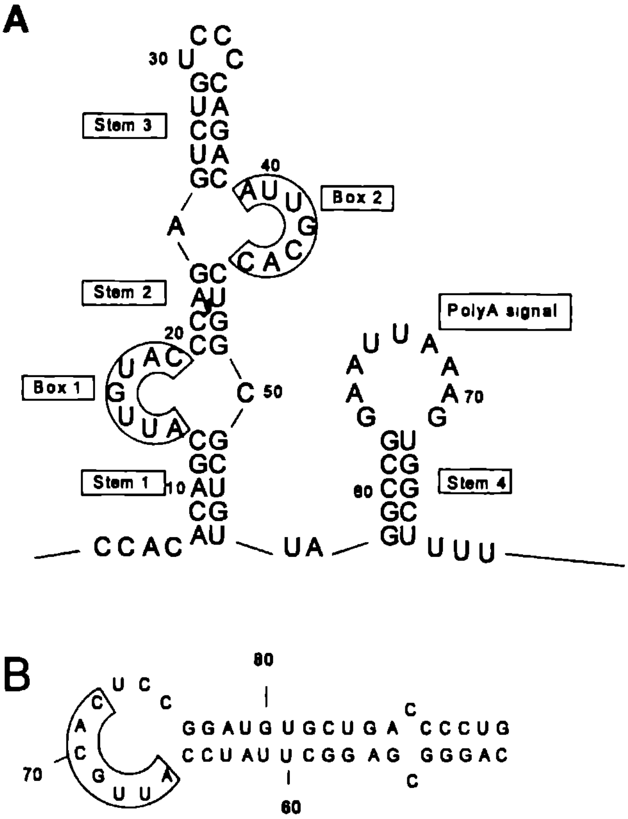


Figure 1. (A) The secondary structure of the conserved region of the 3' UTR of the human Ag RNA The boxed regions are important for U1A protein binding The nomenclature for the loop and stem regions is shown, as is the location of the two-fold symmetry axis (⧫) (see text). (B) The second stemloop of human U1 snRNA The boxed region is important for U1A binding. Numbering of nucleotides is according to wt U1 snRNA sequence

riboses of the symmetrically positioned G16 and G42 are not protected against hydroxyl radicals, and also stem 2 shows symmetrical behaviour toward hydroxyl radicals (Figure 4E of Chapter 5). The ribonuclease protection data of the RNP complex indicate that in stem 2 the only V1 cleavages are found between C20-C21 and C46-U47, which are positioned symmetrically in the RNA (Chapter 5, Figure 4D, right panel). In the Box regions RNases can moderately cleave A39 and U41, as well as A13 and U15, which are positioned symmetrically. In stems 1 and 3, both stem sides show protection against ribonucleases in the RNP complex, although the 5' side of stem 1 (nts 8-12) more than the corresponding side of stem 3 (nts 34-38).

Towards a three-dimensional model of the U1A-mRNA complex

Model building strategy

Taking into account the similarities between the two substrates of U1A protein, U1 snRNA and U1A mRNA, we started by creating one U1A-binding site of the Ag RNA similar to the U1 snRNA hairpin loop found in the X-ray structure of the complex of the N-terminal RNP motif of U1A protein with this U1 snRNA loop (12). The N-terminal RNP motif of the U1A protein was positioned on the partial Ag RNA structure. It is known that the structure of the RNP motif bound to RNA is nearly identical to the unbound protein structure (8, 10), so the reconstructed U1A structure (15), which is based on the protein crystal Cα-coordinates (7), can be used with confidence in the building of the RNA-protein complex. The second U1A-binding place of Ag RNA was assumed to be identical to the first, so the conformation of this first U1A-binding region of Ag RNA was copied. After this, possible relative orientations of the two protein binding sites in the mRNA were explored.

Complex of one U1A binding site with one U1A RNP motif

Our working model of the complex of an RNA molecule containing Box 2 and stemloop 3 with the N-terminal RNP motif of U1A is shown in Figure 2. The full-coordinate reconstruction of U1A (15) contains coordinates for amino acids 1-90. Amino acids 91 and 92 were added to the C-terminal end of the RNP motif and the structure of amino acid 90 was adjusted (it was disordered in (7)) to resemble the co-crystal (12).

Regarding the Ag RNA, we constructed stem 3 containing the tetraloop. At this stage of the modeling, the tetraloop was constructed in agreement with the N3-C modification data of the naked RNA (Figure 3A, Chapter 5). In the presence of U1A the structure of the loop will change to adopt a more exposed conformation because the bases are accessible to ribonucleases in the RNP complex but not in the RNA (see Chapter 5, Figure 4D). To position stem 3 on the RNP motif of U1A we used the Arg 52, Lys 20 and Lys 22 side chains. Base pair C38-G25 is thought to resemble the C65-G76 base pair in the U1 snRNA stem of stemloop II (Figure 1B), which has contact with Arg 52 in the co-crystal (12). Nucleotides 34-38 in Ag RNA are oriented towards the side chains of lysines 20 and 22. After this, the Box 2 nucleotides were

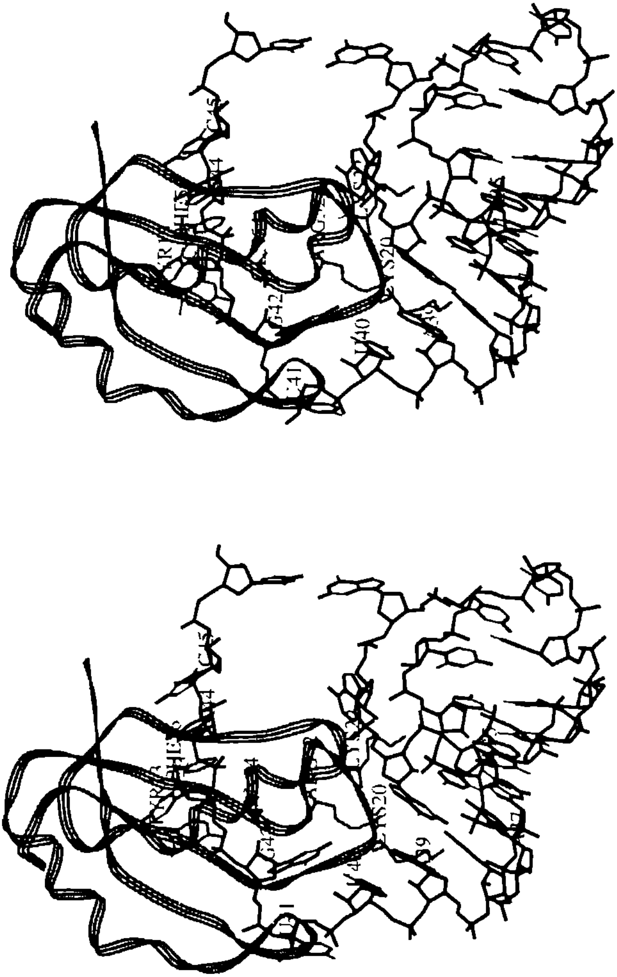


Figure 2. Model structure for one U1A binding region of Ag RNA with the N-terminal RNP motif of U1A. The Box 2 region is shown, as well as stem 3 and the tetraloop. For the N-terminal RNP motif of U1A only a ribbon representation is shown, as well as the side chains of some important amino acids (Arg 52, Tyr 13, Phe 56 and Gln 54). The nucleotides in Box 2 are labeled as well as some amino acid side chains. In stem 2, base pair G23-C46 is shown in black while the other 3 base pairs are shown in grey.

positioned. At the 5' side of Box 2, nucleotides A39 and U40 stack on the preceding stem 3. At the 3' side of Box 2, A44 and C45 are stacked on Phe 56. For nucleotides U41 and C43 a 2'-endo sugar pucker was chosen as starting conformation, in agreement with the pucker of the corresponding nucleotides in the co-crystal (12). G42 is located between the side chains of Asn 16 in $\beta 1$ and Gln 54 in $\beta 3$, and it was constructed with an exposed ribose, because this nucleotide is not protected against hydroxyl radicals in the RNP complex. C43 is stacked on the Tyr 13 side chain. This is in agreement with the co-crystal structure (12) and also with the fact that Tyr 13 can be crosslinked to the corresponding nucleotide - C70 - in U1 snRNA (6). Other examples of such contacts are the interactions of U40 with the side chain of Glu 19 and of U41 with the side chain atoms of Arg 83, Lys 80 and Asn 16. In fact, almost all stacking and hydrogen bonding contacts between the RNA and protein main chain and side chain atoms which have been described for U1 snRNA stemloop II can readily be made in our model. This further underscores that the tertiary reconstruction of U1A protein made by us previously (15) shows good agreement with the crystal structure (12).

Nucleotides C46, G23 and A24 in Ag RNA are thought to resemble nucleotides 73-75 in U1 snRNA. In the co-crystal, these 3 nucleotides do not contact the RNA (12). Concerning the conformation of A24, it was constructed as looping out with an accesssible N7-atom, in agreement with the chemical probing of the naked RNA (Chapter 5). However, the final positioning of A24 in the RNP can be made only when DMS data for the RNP complex providing information about the behaviour of the N1 and N7 atoms of A24, have been obtained.

Finally, the four base pairs of stem 2 were constructed. Because their position can not be determined at the moment, only base pair G23-C46 is colored black to indicate that they are at similar positions as U73 and C74 in U1 snRNA. The other 3 base pairs of stem 2 are colored grey in Figure 2 and their position will be the determining factor in the relative orientation of Box 1 and Box 2 in the Ag RNA (see also below).

To create a conformation for the second U1A-binding site of Ag RNA (Box 1 and stem 1), the conformation of the first U1A-binding site was copied (not shown). Stems 1 and 4 were constructed and stacked on each other. The linker between the two stems is only two nucleotides long and in RNA modeling, helices separated by less than three nucleotides are often assumed to stack (18). Furthermore, stacking could agree well with the enhanced RNase V1 cleavage as found in the RNP at the bottom of stem 1 (Chapter 5).

At this stage of the modeling we did not consider possible differences between the behaviour of the two Box regions in our protection experiments. However, a perfect identity can not be expected, given the differences in U1A binding of Box 1 and Box 2. C70 (in U1 snRNA (12)) and C43 (in Box 2) are stacked on Tyr 13. U17, the altered nucleotide in the U1A binding site of Box 1 can also be stacked on Tyr 13 but uracil is known to be a weaker "stacker" than cytosine (19). Three hydrogen bonds are found for positions N3 and N4-H of C70 in U1 snRNA with side chain and main chain atoms of U1A (12). Since in uracil the positions of the hydrogen bonding donors and acceptors is opposite as compared to cytosine (N3-C vs. N3H-U and N4H-C vs. O4-U), none of these three hydrogen bonds can be formed in Box 1 if the U17 adopts the exact same position as C43 in Box 2. This may explain the loss of binding affinity found for Box 1. In contrast to this, however, is the observation that when C⇔U mutations were made for all the nucleotides in the hairpin loop of a U1 snRNA-like substrate, no significant effect on U1A protein binding *in vitro* could be measured (22).

The 3' part of the structure (stem 4 and the polyadenylation loop) were built pointing away from the 5' part (data not shown), since hardly no protection by U1A was found, indicating that there is no contact with the N-terminal RNP motif of U1A.

Structure of U1A protein

In the crystal structure of the U1A protein dimer, two N-terminally located RNP motifs are present in the asymmetric unit and they are related to each other by a non-crystallographic dyad axis (Figure 3) (7). Many hydrophobic amino acids are found in this interface which suggests that the dimer is not an artefact of crystallization, but that the RNP domain can form such dimers with either itself or with an RNP domain in other proteins (7).

U1A is known to exist as a monomer in solution (20, 21) and to bind U1 snRNA as a monomer. However, two U1A proteins (and not more than two) bind specifically to the 3' UTR of U1A mRNA (13). A dimer of U1A could not be demonstrated in solution using crosslinking methods in the absence of RNA, but the two proteins can be crosslinked with dithio-bis(succinimidylpropionate) (DSP) when U1A mRNA is present (W. Boelens, unpublished results). The crosslink was found with two U1A wt proteins, and does not prove an orientation as in the crystal dimer, but does not exclude it either. If one looks at the interface between the N-terminal RNP motifs of the two U1A proteins in the dimer (7) it can be seen that most lysines are located near or at the interface.

NMR results obtained for the U1A-U1 snRNA interaction (10, 11) and for the hnRNP C- $r(U)_8$ interaction (8) showed that the α-helices of the RNP motif do not

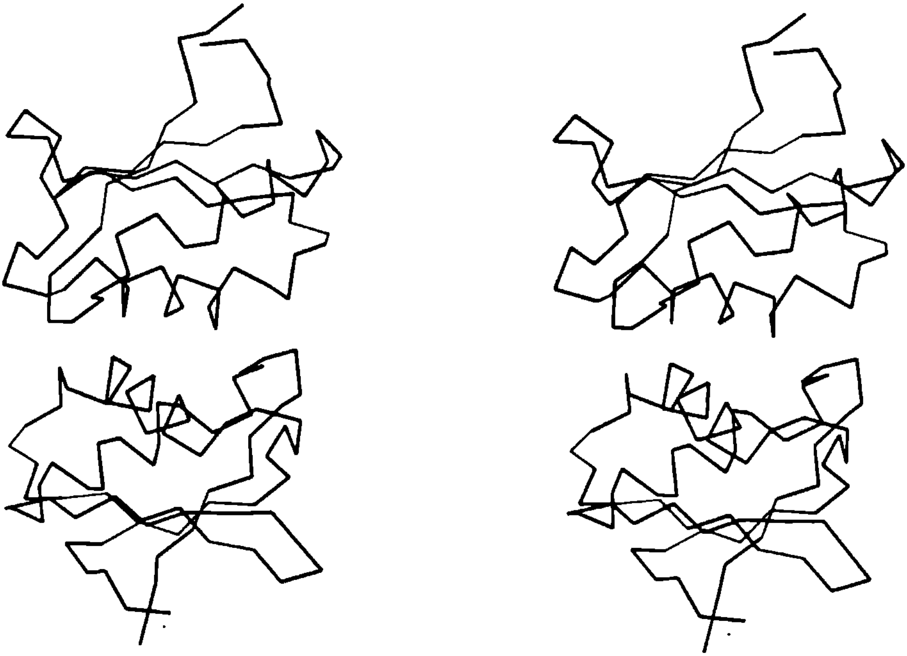


Figure 3. Stereopicture of the Cα-coordinates of the X-ray structure of the N-terminal RNP motif of the U1A protein dimer (7). The view is looking down the non-crystallographic dyad axis.

make contact with the RNA but are free in the structure and potentially available for protein-protein interactions. If one assumes that the same holds for U1A bound to U1A mRNA this would mean that the α-helices of the RNP motif are available for interaction with another protein domain, either from the same protein or from a different protein.

Considering the size of the U1A protein (282 amino acids) and the proximity of the two U1A binding sites on the RNA to each other, it can be expected that the two U1A proteins interact with each other. When the dissociation constants for the two single mutants (each missing one of the two Box sequences) are compared to that of wt Ag RNA, it seems apparent that there is some cooperativity between the two proteins (13). Furthermore, it appears that the U1A protein which binds to the weaker Box 1 needs sequences outside its RNP motif to bind the mRNA, which could mean that protein-protein interactions are necessary for the second U1A protein to bind (our unpublished data). Further support for cooperativity can be extracted from experiments with RNA mutants. Mutant ΔB2, in which only the weaker Box 1 is available can not bind U1A in the presence of 500 mM of salt, while if Box 2 is also present, two U1A proteins bind at these conditions (13). Further data on possible protein-protein interactions come from an U1A mRNA mutated in the Box 1 sequence, which still was able to (weakly) bind two U1A proteins (our unpublished data).

Recently, RNA mutants in which an increasing number of base pairs was added to stem 2, i.e. to increase the distance between Box 1 and Box 2, were tested in mobility shift assays (S. Gunderson, personal communication). When one base pair is added to stem 2, there is still a complex visible of Ag RNA with two U1A proteins, although already in much reduced amount as compared to wt Ag RNA. However, if two or more base pairs are added to stem 2, only the complex of Ag RNA with one U1A protein can be found, even at high U1A protein concentrations. This suggests that contact(s) between the two proteins is necessary for the second protein to bind to Box 1.

All these data point to an interaction between the two U1A proteins. However, it must be realized that the results described above were obtained in experiments with the full-length U1A protein, while the only structural information available for U1A protein is its N-terminal RNP motif. Two molecules of A101, containing only the N-terminal RNP motif, can bind to the U1A mRNA (13). Our ribonuclease protection data have been determined for U1A wt and for A101, while for Fe(II)EDTA footprinting only U1A wt was used so far (Chapter 5). In case of RNase V1, the protection experiments showed no difference between U1A wt and A101. Further studies are underway to determine the protection of A101- Ag RNA complexes for RNase T2 and Fe(II)EDTA. However, a problem with A101 is that two A101 domains do not bind strongly to the RNA and that the resulting complex is not a functional one (that is, it does not affect polyadenylation). In case of the complex of U1A with U1 snRNA not much difference in protection against RNase V1 and ethylnitrosourea (ENU) was found for U1A wt or A96 (amino acids 1-96 of U1A) (22).

All these data still leave several possibilities. The two U1A proteins interact via sequences outside the N-terminal motif or via their N-terminal RNP motifs or perhaps via both. In case of interacting RNP motifs, we think it most likely that the interaction between them occurs as is found in the X-ray structure (7).

"U1A dimer possibility"

We decided to start the modeling with the simplifying assumption that in the Ag

RNA - U1A complex the N-terminal RNP motifs of the two U1A proteins are positioned as seen in the X-ray structure of U1A (7), and to investigate this possibility in relation to the available experimental data.

Both RNP motifs of U1A were positioned as in the protein dimer (7), in which their relative orientation is determined by the non-crystallographic dyad axis. When the constructed Box 2 with stemloop 3, as well as Box 1 with stem 1 are positioned, each on one RNP motif, this twofold symmetry is conserved.

Nucleotides A24, C50 and the base pairs of stem 2 must then be positioned to connect the two RNA parts (not shown). Stem 2 contains only four base pairs, but should be able to bridge the required distance if one postulates a rather sharp bending of the RNA. Nucleotides A24 and C50 are single-stranded nucleotides, which can span a length of 6-7 Å (23). The stacking interactions of A44 with C45 will have to be released, and stem 3 and stem 1 will have to be rotated somewhat to make a connection between the two U1A-binding sites possible.

In such a structure of two RNP motifs with Ag RNA the major groove of stem 2 would be protected, while the shallow groove would not. However, the Fe(II)EDTA data of U1A wt indicate that the shallow groove is protected, but because this experiment has been performed with U1A wt this protection could be caused by other regions of the U1A protein. Therefore, the Fe(II)EDTA data can correspond with a positioning as described above. Fe(II)EDTA experiments for the complex of A101 with Ag RNA will give a more precise answer.

RNase V1, which recognizes the phosphates of nucleotides in a helical conformation (24), shows protection at both sides of stem 2, both in the presence of U1A wt or A101. This means that parts of the two RNP motifs must be positioned between the Box 1 and 2 regions, which indeed is the case in the postulated model. The protection against V1 can be explained by the model since the phosphates of stem 2 are not accessible from all directions.

Concerning stem 3 we hardly find any protection of the riboses. This is in agreement with the fact that the major groove is oriented towards the protein. Hydroxyl radicals are thought to attack C4' and/or C1' of the riboses (25), which are located in the shallow groove. The ribonuclease protection data found for both sides of stem 1 and 3 is not so easy to explain. It is found both with U1A wt and A101, which means that only the two RNP motifs are responsible for this protection. When a Box region is positioned on the RNP motif as is the U1 snRNA loop in the co-crystal (12), such protection is not conceivable. This suggests that most likely other parts of the Ag RNA are located in these areas, sterically hindering the ribonucleases to attack.

It is interesting to compare our model structure with the structure of the interaction between the second stemloop of U1 snRNA and U1A protein (12). The ribonuclease protection patterns for stems 1 and 3 are more extensive than those of the RNA stem in the U1A - U1 snRNA complex (22). In that case only one side of the stem (nucleotides 59-63 in Figure 1B), is protected against RNase V1. This discrepancy could be explained by taking into account the different size of the RNA substrate.

In U1 snRNA the phosphates of C59 to C64 in the stem are protected against ENU, indicating that they are oriented towards the protein (22). In our model structure, the corresponding phosphates of stem 3 are also oriented towards the protein.

In U1 snRNA, the three 3' nucleotides of the loop (nts 73 to 75) do not contact the RNP motif of U1A (12). In the Ag RNA- U1A model structure the corresponding bases of A24 and C50 are also accessible, but the base pairs G23-C46 and C20-G49 in stem 2 are located close to the U1A protein. This can be explained by the fact that their structural context is very different, since in U1 snRNA the corresponding nucleotides (74 and 75) are located in a single-stranded loop, while in Ag RNA they form a base pair and also are part of the stem linking the two U1A binding regions.

Concerning the AUUG(C/U)AC sequences, also a difference in protection behaviour is found between the Box regions on the one hand and the U1 snRNA loop on the other hand (22). Firstly, however, it must be noted that the ENU data found for the U1 snRNA loop (protection of C70 to G76) (22) do not seem to correspond with the U1 snRNA - U1A co-crystal structure, where all the phosphates of this sequence seem accessible (12). This indicates that the relation between the behaviour of nucleotides towards probing agents in solution and their position in a crystal structure is not so evident. This complicates also the interpretation of our Fe(II)EDTA data for the Box 1 and 2 regions. In particular we find better protection against hydroxyl radicals in the region around A39, U40 and C45, as compared to the corresponding nucleotides in U1 snRNA. This may indicate a different or tighter binding of these nucleotides to U1A protein.

Conclusions

We have attempted the building of a possible three-dimensional structure for the complex of Ag RNA with U1A. First, a working model of the complex containing one U1A protein binding site of Ag RNA and one U1A RNP motif was built. We then speculated about possible orientations of the two U1A-binding sites relative to each other. Only one possibility is discussed here, in which the symmetry present both in the U1A protein dimer crystal structure and in the Ag RNA, is maintained in the U1A-Ag RNA complex. However, the experimental data obtained so far are not sufficient to distinguish between this model and other possible models in which the two U1A proteins are not positioned as found in the crystal dimer.

Suggestions for further experiments would include crosslinking of two A101 proteins to try to obtain evidence for the dimer orientation, and DMS and kethoxal probing of the RNP to distinguish between major and minor groove accessibility of the nucleotides, in particular in the stem regions. The role of A24 and C50 can be tested by deleting or mutating them, and nucleotides in Box 1 and 2 can be mutated to test for important interactions, comparable to the studies performed on U1 snRNA (6, 22). Furthermore, mutants of U1A can be tested for RNA binding to further identify important amino acids.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Eric Westhof and Dr. Kathleen Hall for helpful discussions.

REFERENCES

1. Birney, E., Kumar, S. and Krainer, A.R. (1993) *Nucleic Acids Res.*, 21, 5803-5816.
2. Scherly, D., Boelens, W., Van Venrooij, W.J., Dathan, N.A., Hamm, J. and Mattaj, I.W. (1989) *EMBO J.*, 8, 4163-4170.
3. Lutz-Freyermuth, C., Query, C.C. and Keene, J.D. (1990) *Proc. Natl. Acad. Sci. U. S. A.*, 87, 6393-6397.
4. Boelens, W.C., Jansen, E.J.R., Van Venrooij, W.J., Stripecke, R., Mattaj, I.W. and Gunderson, S.I. (1993) *Cell*, 72, 881-892.
5. Lu, J.R. and Hall, K.B. (1995) *J Mol Biol*, 247, 739-752.
6. Stump, W.T. and Hall, K.B. (1995) *RNA*, 1, 55-63.
7. Nagai, K., Oubridge, C., Jessen, T.H., Li, J. and Evans, P.R. (1990) *Nature*, 348, 515-520.
8. Görlach, M., Wittekind, M., Beckman, R.A., Mueller, L. and Dreyfuss, G. (1992) *EMBO J.*, 11, 3289-3295.
9. Görlach, M., Burd, C.G. and Dreyfuss, G. (1994) *J Biol Chem*, 269, 23074-23078.
10. Howe, P.W.A., Nagai, K., Neuhaus, D. and Varani, G. (1994) *EMBO J*, 13, 3873-3881.
11. Hall, K.B. (1994) *Biochem.*, 33, 10076-10088.
12. Oubridge, C., Ito, H., Evans, P.R., Teo, C.H. and Nagai, K. (1994) *Nature*, 372, 432-438.
13. Van Gelder, C.W.G., Gunderson, S.I., Jansen, E.J.R., Boelens, W.C., Polycarpou-Schwarz, M., Mattaj, I.W. and Van Venrooij, W.J. (1993) *EMBO J.*, 12, 5191-5200.
14. Bernstein, F.C., Koetzle, T.F., Williams, E.J.B., Meyer, E.F.Jr., Kennard, O., Shimanouchi, T. and Tasumi, ,M. (1977) *J. Mol. Biol.*, 112, 535-542.
15. Van Gelder, C.W.G., Leusen, F.J.J., Leunissen, J.A.M. and Noordik, J.H. (1994) *Protein-Struct Funct Genet*, 18, 174-185.
16. SYBYL program. TRIPOS Associates, Inc., St. Louis, Missouri.
17. Scherly, D., Boelens, W., Dathan, N.A., Van Venrooij, W.J. and Mattaj, I.W. (1990) *Nature*, 345, 502-506.
18. Kim, S.H. and Cech, T.R. (1987) *Proc. Natl Acad. Sci. , USA*, 84, 8788-8792.
19. Cantor,C.R. and Schimmel,P.R. (1980) Biophysical Chemistry Part III, The behaviour of biological macromolecules. Freeman and Company, San Francisco.
20. Hoffman, D.W., Query, C.C., Golden, B.L., White, S.W. and Keene, J.D. (1991) *Proc. Natl Acad.*

Sci. , *USA*, 88, 2495-2499.
21. Hall, K.B. and Stump, W.T. (1992) *Nucl. Acids Res.*, 20, 4283-4290.
22. Jessen, T.H., Oubridge, C., Teo, C.H., Pritchard, C. and Nagai, K. (1991) *EMBO J.*, 10, 3447-3456.
23. Saenger, W. (1984) Principles of nucleic acid structure. Springer-Verlag, Berlin.
24. Lowman, H.B. and Draper, D.E. (1986) *J. Biol. Chem.*, 261, 5396-5403.
25. Huber, P.W. (1993) *FASEB Journal*, 7, 1367-1375.

CHAPTER 6

General Discussion

General Discussion

Introduction

The past decade has seen a rapid increase in our understanding of the role of RNA-protein complexes in biological processes such as translation, transcription, RNA processing and translocation of proteins. However, little is known about the details of sequence-specific recognition between RNA and protein components of these RNA-protein complexes. For this reason, much effort is being made to investigate their secondary and tertiary structures. In this thesis the structural features of RNA and protein components of two RNA-protein complexes have been described: the U1A-U1A mRNA complex and the Ro RNPs.

The U1A protein and the U1A-U1A mRNA complex

The U1A protein is a well-characterized protein, of which the structure and mode of U1 snRNA binding have been studied in great detail, while data on U1A mRNA binding are beginning to emerge. However, the function of the U1A protein in splicing has not yet been established, although it is known that recognition of the 5' splice site by U1 snRNP requires both U1 snRNA and U1-specific proteins (1). A possible role in the link-up between splicing and polyadenylation has been postulated for U1 snRNP, since anti-U1 snRNP antibodies specifically block cleavage and polyadenylation in nuclear extracts (2, 3), and immunoprecipitation of poly(A) polymerase results in specific co-precipitation of U1 snRNA, but not of other snRNAs (4). Furthermore, U1 snRNA can be crosslinked to pre-mRNAs in the region of the polyadenylation signal in a manner that is influenced by the presence of a 3' splice site on the RNAs (5). The finding that U1A protein can also bind to its own (pre-)mRNA strenghtened this idea of coupling between splicing and polyadenylation and both positive and negative regulating effects of U1A on polyadenylation have been found. U1A protein can inhibit polyadenylation of its pre-mRNA by binding to a specific region of the 3' UTR of this mRNA (6). However, U1A may also positively regulate polyadenylation efficiency by interacting with the upstream efficiency element of the SV40 late polyadenylation signal (7). This latter interaction has been proposed to occur via the second RNP motif of U1A. Unfortunately, these results could not be confirmed by Mattaj and Keller (personal communication) and by Lu and Hall (8).

From the above it is clear that despite the fact that much is known about the U1A protein, its precise function in mRNA processing still has to be discovered. The same holds for the specific recognition of the RNA substrates by both the N-terminal and the C-terminal RNP motifs of U1A. The studies described in this thesis have contributed significant information about the U1A mRNA as substrate for U1A. The structure of the conserved region of U1A mRNA has been thoroughly investigated, and the secondary structure has now been established ((9) and Chapter 5). Furthermore, a working model has been proposed for the structure of the U1A-U1A mRNA complex (addendum of Chapter 5). At this stage of the modeling, and with the limited amount of data available at present, the working model still is very crude. Further experimental studies are esssential for further refinement of the model. Mutations can be made in both the U1A mRNA and the U1A protein to identify important nucleotides and amino acids, respectively, and crosslinking techniques can be used to locate contact sites in the RNA-protein complex.

The Ro RNPs

Much less data are available on the Ro RNPs, as compared to the U RNPs. In fact, structural features of these complexes are just beginning to be unraveled. Although the Ro RNPs are conserved in a variety of vertebrate and invertebrate cells, and present in relatively abundant quantities, the function of these complexes in the cell is still unknown. However, possible function(s) in processes such as mRNA stability, mRNA localization or translation have been suggested (10, 11). Recently, it was discovered that the Ro60 protein could be involved in a novel quality control or discard pathway for 5S rRNA in the nucleus (12).

When the investigations on the structures of the Y RNAs were started, only the human sequences were known next to secondary structure predictions based on RNA folding algorithms. The secondary structure of the human hY RNAs was determined biochemically using chemical and enzymatic probing, and some interesting features were discovered (13). First, in hY1 RNA, the pyrimidine-rich region in the large internal loop appears to be involved in tertiary interactions (13). The behaviour of the N7-atoms of several adenosines in this loop also seem to point to stacking interactions or (tertiary) base pairing (our unpublished data). A second interesting feature is the finding that in nearly all Y RNA sequences base pairing between several nucleotides in a hairpin loop and nucleotides in the large internal loop appears possible (our unpublished data). Although the probing of hY1 and hY5 RNA at 30°C showed these regions to be fully single-stranded, it will be interesting to test whether these nucleotides become inaccessible at a lower temperature.

During our studies the Y RNA sequences of *Xenopus laevis* became available (14) and recently the Y3 and Y4 RNAs of iguana (15) as well as the single Y RNA of *C. elegans* (16) were characterized. The proposed structures of all these RNAs are in good agreement with the consensus secondary structures of the human Y RNAs proposed by us (13). With more Y RNA sequences becoming available, phylogenetic comparison can now be performed in search for conserved secondary and tertiary interactions. Such data will be very helpful in complementing the structure probing data.

Regarding the Ro60 and La proteins, not much structural information is available. In both proteins large regions outside the RNP motifs are needed to achieve RNA binding which hampers studies on the structures of these proteins and their interaction with RNA. In case of La, some information is available concerning the requirements of its RNA substrates. The recognition site on the protein can accommodate up to 4 uridylate groups with preference for a 3' OH-terminus (17). By using protein homology modeling with the U1A protein structure (Chapter 2) as a template, the RNP motif of the La protein was built (our unpublished data), and this model can be of use in future studies regarding the binding of La to the Y RNAs.

In principle, it is possible to build the RNP motif of Ro60 by protein homology modeling, but practically no structural information is available for the important regions flanking this domain. Several epitopes of Ro60 are known (18, 19), and since these regions probably are located on the outside of the protein, epitope data can be of some help. It is, however, obvious that more studies are needed to characterize the Ro60 protein and its RNA binding properties in more structural detail.

The Ro52 protein, the function of which is also not known, does not bind the Y RNAs directly, but presumably via protein-protein interactions (20, 21), since it contains zinc finger and leucine zipper motifs. Interestingly, Ro52 was shown recently to bind DNA, and striking similarities were found between the nucleic acid-binding motifs of Ro52 and a family of zinc finger proteins which bind DNA or regulate gene expression (22). These findings seem to indicate that Ro52 belongs structurally and functionally in this family and this interesting possibility certainly will initiate additional research leading to more (structural) information about Ro52.

A future line of Ro RNP research will be the probing of the RNA when present in the RNP particle (footprinting). Such an approach might identify protein binding sites and elements of the RNA which are not involved in protein binding. One interesting feature in this respect is the limited base pairing possibilities (ranging from 2 to 4 base-pairs) between the 5' part and 3' part of the large internal loop, present in nearly all known Y RNA sequences. It is possible that these potential interactions, possibly stabilized by protein(s), are only present when the Y RNAs exert their as yet unknown function.

Other options to study the structures of the Ro RNPs are RNA-RNA and RNA-protein crosslinking experiments and further use of RNA mutants. Interesting mutants could for instance include RNAs in which some of the cytosines in the pyrimidine-rich region of hY1 RNA are replaced by other nucleosides. It will be interesting to see at what conditions the folded loop turns into a "simple" single-stranded-region, that is, whether the presumed tertiary interactions will be broken. When nucleotides around the bulged C9 and interior loop 1 are systematically changed, further information concerning Ro60 binding might be obtained. Such studies would be even more interesting when the function of the Ro RNPs could be established, because then direct structure-function studies are feasible.

Considerations about RNA, protein and RNP structures

Methods of RNA secondary structure prediction have greatly improved in the last few years. Prediction of optimal and suboptimal structures and the determination of better en more free energy parameters, in particular for junctions and internal loops, have led to closer agreements with available models of RNAs, which were established independently by phylogenetic and experimental studies. There are no rules available yet for reliable prediction of tertiary interactions, like the one found in hY1 RNA. However, some progress has been made in the identification of tertiary motifs in RNA, deduced from sequence information (28, 29).

When dealing with RNA structures it is important to realize that a biological RNA molecule mostly does not form a single structure but instead, may have several alternative conformations in equilibrium. Furthermore, the lowest free energy structure and the biologically important structures are not necessarily the same; a conformational switch can occur between alternative configurations of an RNA during its functioning, as, for example, has been shown for the 7SL RNA molecule during the signal recognition particle cycle (23).

Finally, it should be realized that the structure of an RNA molecule present in an RNA-protein complex can be different from that of the naked RNA. For U1 snRNA it was found that the loop of hairpin II is flexible in the RNA but becomes more structured after binding of the U1A protein (24, 25). In case of the U1A mRNA, a change in the structural behaviour of the tetraloop is found in the presence of the U1A protein (Chapter 5). However, structural data obtained by experimental studies on RNAs in both free and protein-bound form often largely agree, indicating that global changes in the RNA backbone structure after association with protein mostly do not occur (26, 27).

The protein components of RNA-protein complexes are also not static entities. Their structure is flexible, although perhaps less so than that of the RNA components, and their structures may also change when association with the RNA takes place. In case of the RNP motif, however, it is known that the global protein structure does not change much during RNA binding (24, 25, 30).

All three protein constituents of the RNA-complexes described in this thesis, i.e. U1A, Ro60 and La, contain an RNP motif. However, the cognate RNA substrates differ considerably in sequence and in structure. In case of U1A the two substrates U1 snRNA and U1A mRNA, contain an RNA stem-loop and an internal loop, respectively, as binding sites for the protein. In case of Ro60 an RNA stem is the binding site, while for La a single-stranded oligo-U stretch appears sufficient. Another difference is that in case of U1A, the RNP motif can bind independently to RNA, while in case of both Ro60 and La sequences outside the RNP motif are necessary. In the latter two cases it is likely that these sequences stabilize the correct conformation of the RNP motif for RNA binding. However, their direct involvement in RNA binding in La and Ro60 cannot be excluded and needs further investigation.

The integration of experimental and theoretical approaches

High resolution structural techniques such as X-ray crystallography and NMR are presently not capable of handling systems of the size of most RNA-protein complexes, although progress is being made. For this reason, our understanding of the three-dimensional structure of RNAs is lagging behind that of other macromolecular systems. However, there is a wealth of low-resolution structural data available for several RNAs and RNPs. These include results from secondary (and sometimes tertiary) structure predictions based on phylogenetic studies, crosslinking and footprinting experiments, chemical accessibility, electron miscroscopy, mutational studies, etcetera, which all contribute valuable information useful for building tertiary structure models of RNA

With more data, both about protein and RNA components, becoming availabale, an attempt can be made to integrate them in a possible three-dimensional structure by using RNA and protein modeling methods. Structural models of several RNAs are now availabe (31-33) It must be realized, however, that the usefulness of such a modeling process does not reside in its current level of precision. The strength lies in the prediction of the global folding of the RNA, a 3D hypothesis destined to be subjected to experimental verification The model can be tested, for instance by using RNA mutants. Depending on the experimental results, the model will be adapted and tested again, and so on.

In conclusion, the integration of both experimental and theoretical tools for studying structural features of RNA and protein molecules and their interaction, will be very valuable in gaining insights into functionally important RNA structures and contributes to our understanding of experimentally observed phenomena concerning RNA molecules. Application and further development of such theoretical tools will be of considerable importance in future studies in molecular biology

The work described in this thesis has contributed to a better understanding of the secondary structures of the RNAs being studied and of the tertiary structure of the RNP motif Furthermore, in case of the U1A-mRNA complex a 3D-model has been obtained, which can now be tested and further refined In case of the Ro RNPs, the results obtained for the secondary structures of the hY RNAs have created possibilities for studies on possible tertiary interactions in the RNA and the RNA-protein complex. These studies will eventually lead to a three-dimensional structural model of the Ro RNPs

REFERENCES

1 Kohtz, J D , Jamison, S F , Will, C L , Zuo, P , Luhrmann, R, Garciablanco, M A and Manley, J L , *Nature*, 368 119 124, 1994

2 Moore, C L and Sharp, P A , *Cell*, 41 845-855, 1985

3 Hashimoto, C and Steitz, J A , *Cell*, 45 581-591, 1986

4 Raju, V S and Jacob, S T , *J Biol. Chem* , 263 11067-11070, 1988

5 Wassarman, K M and Steitz, J A , *Genes Dev* , 7 647-659, 1993

6 Boelens, W C , Jansen, E J R , Van Venrooij, W J , Stripecke, R , Mattaj, I W and Gunderson, S I , *Cell*, 72 881 892, 1993

7 Lutz, C S and Alwine, J C , *Genes Dev* , 8 576-586, 1994

8 Lu, J R and Hall, K B , *J Mol Biol* , 247 739-752, 1995

9 Van Gelder, C W G , Gunderson, S I , Jansen, E J R , Boelens, W C , Polycarpou-Schwarz, M , Mattaj, I W and Van Venrooij, W J , *EMBO J*, 12 5191-5200, 1993

10 Pruijn, G J M , Slobbe, R L and Van Venrooij, W J , *Mol Biol. Rep* , 14 43 48, 1990

11 Van Venrooıj, W J , Slobbe, R L and Pruıjn, G J M , *Mol. Bıol. Rep* , 18 113-119, 1993

12 O'Bnen, C A and Wolın, S L , *Genes Dev* , 8 2891-2903, 1994

13 Van Gelder, C W G , Thıjssen, J P H M , Klaassen, E C J , Sturchler, C , Krol, A , Van Venrooıj, W J and Pruıjn, G J M , *Nucl Acıds Res* , 22 2498 2506, 1994

14 O'Bnen, C A , Margelot, K and Wolın, S L , *Proc Natl Acad Scı., USA*, 90 , 7250-7254, 1993

15 Farns, A D , O'Brıen, C A and Harley, J B , *Gene*, 154 193 198, 1995

16 Van Horn, D J , Eısenberg, D , O'Bnen, C A and Wolın, S L , *RNA*, 1 293 303, 1995

17 Stefano, J E , *Cell*, 36 145-154, 1984

18 Huang, S C , Yu, H , Scofield, R H and Harley, J B , *Scand J Immunol.*, 41 220-228, 1995

19 Saıtta, M R , Arnett, F C and Keene, J D , *J Immunol* , 152 4192-4202, 1994

20 Pruıjn, G J M , Slobbe, R L and Van Venrooıj, W J , *Nucl. Acıds Res.*, 19 5173 5180, 1991

21 Slobbe, R L , Pluk, W , Van Venrooıj, W J and Pruıjn, G J M , *J Mol Bıol.*, 227 361 366, 1992

22 Frank, M B , Mccubbın, V R and Heldermon, C , *Bıochem J*, 305 359-362, 1995

23 Andreazzolı, M and Gerbı, S A , *EMBO J*, 10 767-777, 1991

24 Hall, K B , *Bıochem.*, 33 10076-10088, 1994

25 Oubrıdge, C , Ito, H , Evans, P R , Teo, C H and Nagaı, K , *Nature*, 372 432 438, 1994

26 Moazed, D , Stern, S and Noller, H F , *J Mol. Bıol.*, 187 399 416, 1986

27 Huber, P W , Blobe, G C and Hartmann, K M , *J Bıol. Chem* , 266 3278-3286, 1991

28 Costa, M and Mıchel, F , *EMBO J*, 14 1276-1285, 1995

29 Gautheret, D , Damberger, S H and Gutell, R R , *J Mol Bıol.*, 248 27-43, 1995

30 Howe, P W A , Nagaı, K , Neuhaus, D and Varanı, G , *EMBO J*, 13 3873 3881, 1994

31 Brunel, C , Romby, P , Westhof, E , Ehresmann, C and Ehresmann, B , *J Mol Bıol* , 221 293 308, 1991

32 Krol, A , Westhof, E , Bach, M , Luhrmann, R, Ebel, J P and Carbon, P , *Nucl Acıds Res* , 18 3803-3811, 1990

33 Westhof, E and Altman, S , *Proc Natl. Acad Scı., USA*, 91 5133-5137, 1994

Summary/Samenvatting

Summary

RNA-binding proteins as well as ribonucleoprotein complexes (RNPs) mediate interactions in pre-mRNA processing events (capping, splicing and polyadenylation), are involved in the regulation of translation and for the stability of mRNA. Furthermore, RNP complexes are common targets for autoimmune responses, especially in individuals with systemic lupus erythematosus (SLE).

Many RNA binding proteins contain a conserved RNA binding domain, the so-called RNP motif, which is present in one or more copies in proteins that bind pre-mRNA, mRNA, pre-ribosomal RNA or small nuclear RNAs (snRNAs). All these RNAs have their own unique structure, and since the three-dimensional structure formed by an RNA molecule contained in an RNP is crucial to its biological function, knowledge of such structures is essential for our understanding of the complex biochemical processes in which they participate.

In this thesis experimental and theoretical (computational chemistry) methods are combined to investigate structural aspects of the RNA and protein components of two different RNA-protein complexes. The first one is the complex between the U1A protein and its own mRNA. The other RNP complexes described are the cytoplasmic Ro RNP particles, particles consisting of one Y RNA and the proteins Ro60, La and Ro52.

Chapter 1 provides an introduction to RNA secondary and tertiary structure and describes methods that can be used to determine such structures. The RNP motif, present in the U1A, La and Ro60 proteins, is described as well, together with what is known about its structure and its interaction with RNA substrates. Some methods used in this thesis are also introduced in Chapter 1, among them an experimental approach in which a variety of chemical and enzymatic reagents is used to distinguish between base paired and single-stranded nucleotides in an RNA molecule. Furthermore, theoretical approaches concerning RNA secondary structure prediction and RNA and protein tertiary structure modeling are discussed.

Chapter 2 describes a Molecular Dynamics (MD) method which can be used for the generation of complete protein coordinates when only limited coordinate data, e.g. Cα coordinates, are available. This study was inspired by an attempt to build the structure of the RNP motifs of the U1A and La proteins by protein homology modeling, while only the Cα coordinates of a template structure were available in the Brookhaven Protein Databank. The study shows that extensive MD calculations are useful, to some extent, in capturing details of the native protein conformation and as such they appear to be generally applicable in protein

structure prediction. The resulting protein structures can be used (within limits) with confidence to study the general structure of the protein involved, or as a basis for further model building of homologous protein structures.

Chapter 3 describes structural studies on the Y RNAs, small cytoplasmic RNAs which are components of the Ro (SS-A) ribonucleoprotein complexes. The Ro RNPs are frequently recognized by autoantibodies present in autoimmune sera of patients with Sjögren's syndrome or SLE. Until recently, the secondary structures proposed for the hY RNAs originated from low-energy structure predictions only. We investigated the conformation of human hY1 and hY5 RNA, using both chemical and enzymatic structure probing. The results indicate that both for hY1 and hY5 RNA the secondary structure largely corresponds to the structure predicted by sequence alignment and RNA folding algorithms. However, some interesting deviations could be observed, one being an as yet unidentified tertiary interaction in hY1 RNA, involving the pyrimidine-rich region.

Chapter 4 concerns the U1A protein, a protein present in the U1 snRNP complex in which it is bound to the second stemloop of U1 snRNA. However, the U1A protein can also bind to a conserved region in the 3' UTR of its own pre-mRNA and in this way inhibits polyadenylation of this pre-mRNA. The secondary structure of the conserved region of the pre-mRNA able to bind the U1A protein has been determined by a combination of theoretical predictions, phylogenetic sequence alignment, enzymatic structure probing and analyses of structure and function of mutant mRNAs. The results show that the integrity of a large part of this structure is required for both high affinity binding of U1A and subsequent specific inhibition of polyadenylation *in vitro*.

Chapter 5 describes the chemical structure probing of the conserved region of U1A mRNA, which yielded structural information about the RNA at nucleotide resolution. Footprinting experiments on the U1A-U1A mRNA complex were performed as well. The experimental data obtained allowed us to propose a three-dimensional model for the conserved region of U1A mRNA, and for the complex between this mRNA and U1A protein. This model is discussed in the Addendum of Chapter 5.

In Chapter 6 a general discussion about the results described in this thesis is presented. Special attention has been paid to the powerful benefits of an integration of both experimental and theoretical methods to approach the analysis of RNA, protein and RNP structures.

Samenvatting

Zowel RNA-bindende eiwitten als RNA-eiwit complexen (RNP's) spelen een rol bij de processing (capping, splicing en polyadenylering) van boodschapper RNA (mRNA), bij de regulatie van translatie en bij de stabiliteit van mRNA. Ook zijn RNP complexen vaak het doelwit van autoimmuunreacties, met name bij patiënten met systemische lupus erythematodes (SLE).

Veel RNA-bindende eiwitten bevatten een geconserveerd RNA-bindings motief, het RNP-motief, dat in één of meer kopieën aanwezig is in eiwitten die pre-mRNA, mRNA, pre-ribosomaal RNA of snRNA (small nuclear RNA) binden. Al deze RNA's hebben hun eigen, unieke structuur, en aangezien de driedimensionale structuur van een RNA molecuul in een RNP-complex cruciaal is voor zijn biologische functie, is kennis over zulke structuren essentieel voor het doorgronden van de complexe biochemische processen waarin ze een rol spelen.

In dit proefschrift zijn zowel experimentele als theoretische ("computerchemie") methoden beschreven waarmee de structurele aspecten van de RNA- en eiwit-onderdelen van twee verschillende RNA-eiwit-complexen onderzocht zijn. Het eerste complex is dat tussen het U1A eiwit en zijn mRNA; het andere RNP complex is het cytoplasmatische Ro RNP, dat bestaat uit één Y RNA (of Ro RNA) en de eiwitten Ro60, La en Ro52.

Hoofdstuk 1 geeft een inleiding over verschillende aspecten van de secundaire en tertiaire structuur van RNA en beschrijft methoden die gebruikt kunnen worden om zulke structuren te bepalen. Het RNP-motief, dat aanwezig is in eiwitten als U1A, La en Ro60, wordt beschreven wat betreft zijn structuur en zijn interactie met RNA-substraten. Ook worden in Hoofdstuk 1 enkele technieken geïntroduceerd die gebruikt zijn in dit proefschrift, o.a. een experimentele methode waarmee, gebruikmakend van een scala aan chemische en enzymatische reagentia, enkelstrengs- en dubbelstrengs-gebieden in een RNA-molecuul onderscheiden kunnen worden. Tenslotte worden theoretische methoden beschreven om secundaire structuren van RNA en tertiaire structuren van zowel RNA- als eiwitmoleculen te voorspellen.

Hoofdstuk 2 beschrijft een Moleculaire Dynamica (MD) methode die gebruikt kan worden om de volledige coordinatenset te voorspellen voor een eiwit als men alleen de beschikking heeft over een onvolledige set, bijvoorbeeld de Cα-coordinaten. Deze studie kwam voort uit een poging om de structuur van het RNP-motief van de eiwitten U1A en La te voorspellen met behulp van eiwithomologie modelling, terwijl alleen de Cα-coordinaten van een voorbeeldstructuur beschikbaar waren. De studie laat zien dat uitgebreide MD-berekeningen gebruikt kunnen worden om, tot op zekere hoogte, details van eiwitconformatie weer te geven. De gevolgde methode lijkt algemeen toepasbaar bij het voorspellen van eiwitstructuren en de gegenereerde eiwitstructuren kunnen op hun beurt weer als basis dienen voor structuurvoorspelling van andere, homologe eiwitten.

Hoofdstuk 3 bevat een studie naar de structuur van de Y RNA's, kleine cytoplasmatische RNA's die voorkomen in Ro RNP complexen. De Ro RNP's worden vaak herkend door autoantistoffen van patiënten met Sjögren's syndroom of SLE. Tot voor kort waren de voorgestelde secundaire structuren van de hY RNA's slechts gebaseerd op computervoorspellingen. Wij hebben de conformatie van hY1 en hY5 RNA onderzocht met behulp van chemische en enzymatische methoden. De resultaten tonen aan dat zowel voor hY1 als voor hY5 RNA de secundaire structuren grotendeels overeenkomen met de door de computer voorspelde structuren. Niettemin werden ook een aantal interessante verschillen gevonden, waarvan de nog niet nader geïdentificeerde tertiaire interactie in het pyrimidine-rijke gebied van hY1 de belangrijkste is.

Hoofdstuk 4 gaat over het U1A eiwit, dat aanwezig is in het U1 snRNP complex waarin het gebonden is aan de tweede stamloop van U1 snRNA. Bovendien kan het U1A eiwit ook binden aan een geconserveerd gebied in de 3' UTR van zijn pre-mRNA. De secundaire structuur van dit gebied werd bepaald met behulp van theoretische voorspellingen en fylogenetische vergelijkingen, via enzymatische structuur-analyse en door bestudering van de structuur en functie van mutant mRNA's. De resultaten tonen aan dat een aanzienlijk deel van de structuur gevormd moet worden wil het U1A eiwit met hoge affiniteit binden en op deze manier polyadenylering kunnen remmen.

Hoofdstuk 5 beschrijft de chemische structuur-analyse van het U1A mRNA, een methode die zeer gedetailleerde structuur-informatie opleverde (resolutie op nucleotide-niveau). Er werden ook z.g. "footprinting" analyses uitgevoerd met het complex van U1A en zijn mRNA. De verkregen experimentele gegevens hebben geleid tot een driedimensionaal model voor het geconserveerde deel van U1A mRNA, en voor het complex met het U1A eiwit. Dit model is beschreven in het Addendum van Hoofdstuk 5.

Hoofdstuk 6 bevat een afsluitende discussie over de resultaten die beschreven staan in dit proefschrift. In deze discussie wordt de integratie van experimentele en theoretische methoden ter bepaling van RNA-, eiwit- en RNP-structuren nogmaals benadrukt.

List of Publications

Van Gelder, C.W.G., Gunderson, S.I., Jansen, E.J.R., Boelens, W.C., Polycarpouschwarz, M., Mattaj, I.W. and Van Venrooij, W.J. (1993) A complex secondary structure in U1A pre-messenger RNA that binds two molecules of U1A protein is required for regulation of polyadenylation. *EMBO Journal*, 12, 5191-5200.

Van Gelder, C.W.G., Leusen, F.J.J., Leunissen, J.A.M. and Noordik, J.H. (1994) A Molecular Dynamics approach for the generation of complete protein structures from limited coordinate data. *Proteins: Structure, Function, and Genetics*, **18**, 174-185.

Van Gelder, C.W.G., Thijssen, J.P.H.M., Klaassen, E.C.J., Sturchler, C., Krol, A., Van Venrooij, W.J. and Pruijn, G.J.M. (1994) Common structural features of the Ro RNP associated hY1 and hY5 RNAs. *Nucleic Acids Research*, 22, 2498-2506.

Van Venrooij, W.J. and Van Gelder, C.W.G. (1994) B cell epitopes on nuclear autoantigens. What can they tell us? *Arthritis and Rheumatism*, 37, 608-617.

Van Gelder, C.W.G., Teunissen, S.W.M. and Van Venrooij, W.J. (1995) Chemical structure probing of U1A mRNA and footprinting analysis of its complex with U1A protein *(manuscript in preparation)*.

Curriculum Vitae

Celia van Gelder is geboren op 26 juni 1965 in Renkum. In 1983 behaalde zij haar VWO diploma aan het Pax Christi College te Druten. Aansluitend studeerde zij scheikunde aan de Katholieke Universiteit Nijmegen (KUN) en het doctoraalexamen werd behaald in augustus 1988. Tijdens de studie werden een hoofdvak Analytische Chemie/Chemometrie (Prof. drs. G. Kateman) en een bijvak klinische chemie (Dr. R. Wevers, Afdeling Neurologie, Radboud Ziekenhuis, Nijmegen) gevolgd. Gedurende haar studie heeft zij tweemaal het derdejaars practicum "Computers in de Chemie" geassisteerd.

Van september 1988 tot augustus 1990 was zij aangesteld als toegevoegd onderzoeker bij het CAOS/CAMM Center. Haar taken bestonden uit het geven van cursussen, o.a. over Molecular Modeling en Sequentieanalyse van eiwitten en nucleïnezuren (CAMMSA), alsmede het opzetten van een afstudeervariant chemische informatiekunde voor scheikundestudenten van de KUN en het geven van het "Molecular Modeling" onderdeel van de cursus "Computers in de Chemie".

Van september 1990 tot augustus 1994 was zij werkzaam als junior onderzoeker bij de Afdeling Biochemie van de KUN (Prof. dr. W.J. van Venrooij). Gedurende deze periode werd het in dit proefschrift beschreven onderzoek verricht. Op de volgende congressen heeft zij haar onderzoek gepresenteerd: EMBO workshop "Structure and function of eukaryotic RNPs", te Arolla, Zwitserland in 1993; Wetenschappelijke vergadering SON-Nucleïnezuren, te Utrecht, 1993; Jaarlijkse SON-Nucleïnezuren bijeenkomst te Lunteren, 1993; EMBO Summerschool "RNA structure and function", te Spetsai, Griekenland, 1994; CAOS/CAMM Symposium "From Toy to Tool", te Nijmegen, 1995. Er werden verschillende werkbezoeken afgelegd, te weten aan Dr. K. Nagai (MRC, Cambridge), en aan Dr. A. Krol en Prof. E. Westhof (IBMC, Straatsburg). Naast dit onderzoek was zij verantwoordelijk voor alle software- en hardware ondersteuning van de medewerkers van de Afdeling Biochemie.

Van mei 1995 tot april 1996 is zij werkzaam op de Afdeling Biochemie en Voedselverwerking van het Agrotechnologisch Onderzoeks Instituut (ATO-DLO; Dr. ir. C. van Dijk), waar zij eiwitmodeling studies verricht naar structurele aspecten van verschillende voedselenzymen