UNIVERSITÉ DE STRASBOURG

UNIVERSITE DE STRASBOURG

Ecole Doctorale des Sciences de la Vie et de la Santé

THESE

Présentée en vue de l'obtention du grade de

Docteur de l'Université de Strasbourg

Discipline : Sciences du Vivant

Spécialité : Biologie Moléculaire

par

Bertrand BECKERT

GROUP I LIKE RIBOZYMES:

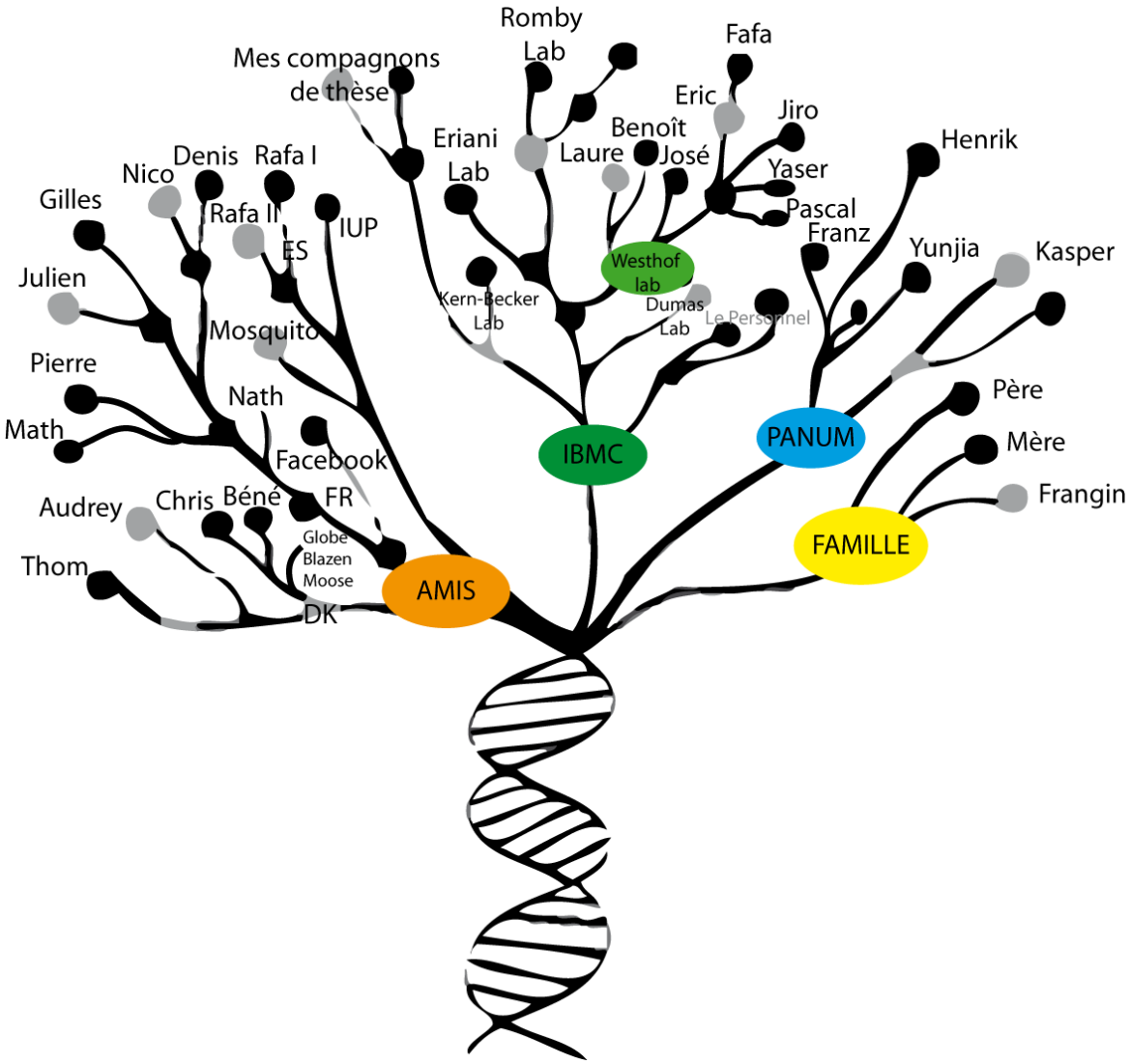
STRUCTURE, FUNCTION AND EVOLUTION

Soutenue le 15 Octobre 2010 devant la commission d'examen :

Dr. Benoît MASQUIDA	Directeur de thèse
Pr. Henrik NIELSEN	Directeur de thèse
Dr. Bruno SARGUEIL	Rapporteur externe
Dr. François MICHEL	Rapporteur externe
Dr. Philippe DUMAS	Examinateur
Pr. Eric WESTHOF	Examinateur

UPR 9002 du CNRS, Institut de Biologie Moléculaire et Cellulaire, Strasbourg

Remerciements/Acknoledgements



Je voudrais tout d'abord commencer par remercier tous les membres de mon jury qui ont accepté de juger mon travail de thèse : Bruno Sargueil, François Michel, Philippe Dumas, Eric Westhof.

Pour l'UPR9002 version Française :

Je voudrais tout d'abord commencer par remercier la personne qui m'a, si je puis me permettre, aiguillé vers la modélisation et vers l'IBMC : Fabrice Jossinet.

Je remercie Eric Westhof qui m'accueillit dans son équipe. Je remercie également mon premier directeur de thèse Benoît Masquida qui a eu la patience m'encadrer (pas facile d'être Boss tout les jours…..). Je lui souhaite bon courage pour la suite des épreuves qui se profilent à l'horizon….

Un grand merci également à tous les membres de l'équipe d'Eric : Fabrice, Pascal, Yaser, José, Jiro, Valérie et bien entendu Laure. Je tiens dans cette équipe à remercier trois personnes qui m'ont beaucoup aidé durant cette thèse, Fabrice, Jiro et Laure.

Merci beaucoup à tous les membres de l'UPR9002 qui m'ont aidé que ce soit pour un conseil ou pour emprunter du matériel. Je tiens tout d'abord à remercier l'équipe Eriani (Gilbert, Franck, Laure (encore)) pour les discussions scientifiques et sans oublier les vendredis... Je tiens aussi à remercier l'équipe Kern/Becker. Finalement, je tiens à remercier tout particulièrement Pascale Romby ainsi que son équipe pour les discussions mais aussi pour l'aide que Pascale a pu m'apporter tout au long de cette thèse même en étant à l'étranger.

Je remercie également le club Med de l'IBMC pour l'organisation de sessions de Ski dans les Alpes Suisses (JC, Guillaume, Thom)… super WE les gars, bravo et superbe ambiance (merci également à Marco !)… De même je tiens à remercier mes compagnons d'écriture/ de thèse (Clem, Lucile, José, Marie, Laurence, Tania, Yaser, Marc, Mathieu, Nizar, Gaëlle)

Plus personnellement je tiens à remercier mes amis (Julien, Nath, Math, Pierre,) et tout particulièrement Gilles… Je tiens également à remercier ceux que j'ai laissé au Danemark (Audrey, Thom, Chris, Béné), merci pour ces deux années à CPH ! De ce fait comment ne pas remercier mes collocs de strass Rafa I et Rafa II (la connexion espagnol !) sans oublier Marie !

Merci encore à toute les personnes que j'ai croisé pendant ma thèse et qui ont contribué à me faire évoluer ! Merci à vous !

For the Panum (I can't write it in Danish…. Undskyld)

I would like to express my deep and sincere thanks and gratitude to Professor Henrik Nielsen. I had a really good time to work with him and thus I really thank him for encouraging me from the beginning to work on the project and also for his academic guidance throughout the thesis. I would like also to thank him for both scientific discussion and social support during my time in his lab. It has been also a real pleasure to be invited at Henrik's place and to see the Danish way of life. Thank you very much.

I would like also to thank members from Nielsen lab. In this way I really thank Yunjia, Kasper my benchmates for many scientific and not scientific talks and for good times that we spent in CPH event if it was not easy every day but I really enjoy the time that I spent with you! I also thank Franz for the technical assistance and also his social support during my time in the lab.

Finally I would like to thank the people from the Panum (18.2.2), Thomas, Neel, Karin, Marianne, Lyse, Quiang… In this way I really thank Karin and Neel for helping me ordering material and also the thousand tips that I have borrowed. I also thank everybody for the organization of the two "Julefrokost" that I have done! It was such a good Christmas diner/party!

Many thanks to everybody.

--- Bertrand Beckert

CONTENTS

INTRODUCTION

Research carried out during the last three decades has challenged the central dogma of molecular biology stating that genetic information is transfered forward from DNA to RNA and ultimately to proteins without back-coding possibility (Crick, 1970). This oversimplified view relegated RNA to a sequence carrier (messenger RNA) or to architectural scaffoldings in large RNPs such as the ribosome (ribosomal and transfer RNAs) in which there was no doubt among the scientific community of the 70's that the peptidyl-transferase activity was carried out by ribosomal proteins.

In this context, it is not surprising that the discovery of catalytic RNAs or ribozymes like group I introns and the RNase P represented a major breakthrough and led their discoverers to be awarded the 1989 Nobel prize in chemistry. Additional ribozymes were then spotted, notably in plant viruses and viroids where they participate in the replication of the RNA genome (Flores et al., 2001; Flores et al., 2009) and recently in humans (Salehi-Ashtiani et al., 2006; Luptak and Szostak, 2008). The ribosome was characterised as a ribozyme since no ribosomal protein could be detected around the peptidyl transferase centre (PTC) in the various crystal structures revealing its architecture in various biologically relevant conformations (Rodnina, 2008; Noller et al., 2001; Yusupov et al., 2001; Yusupova et al., 1991). Moreover, specific RNA chemical groups of residues from the PTC and of the tRNA could be identified as necessary for the peptidyl transfer (Moroder et al., 2009; Graber et al., 2010; Clementi et al., 2010; Clementi and Polacek, 2010). The 21st century has also started by witnessing the central roles played by RNAs in cellular regulatory processes in all kingdoms of life. In bacteria, untranslated regions from mRNAs called riboswitches interact with metabolites to control the expression of the downstream gene (Tucker and Breaker, 2005). Interestingly some of these riboswitches can be ribozymes like the glmS (Winkler et al., 2004; Lim et al., 2006; Link and Breaker, 2008) or fused to ribozymes (Lee et al., 2010). In eukaryotes, interfering RNAs derived from microRNAs (miRNA) or small interfering RNAs (siRNA) have been shown to form an RNA-protein complex that triggers a nuclease-

mediated degradation mechanism of a complementary mRNA target ultimately leading to the temporary knock-down of the encoded gene or group of genes (Fire, 2007; Mello, 2007). And these mechanisms seem to represent only the emerged part of the iceberg according to recent transcriptomic analysis that show that close to 100% of genomes is usually transcribed (Amaral et al., 2008; Mattick, 2009).

Except in the case of the PTC, the chemistry of natural ribozymes is based on the transesterification reaction. Such a reaction usually corresponds to a type 2 nucleophilic substitution (SN_2) where the nucleophilic group, provided either by the RNA itself or by an activated water molecule, attacks a phosphodiester bond. Depending on the nature and location of the attacking group, the products of the reaction can be quite diverse. Several cases summarised in Table 1 have been described. The chemically simplest reaction corresponds to the hydrolysis of the RNA chain by a water molecule. This situation is typical of the reaction catalysed by RNase P which leads to process the 5' end of the tRNA precursors by accurately cleaving at the position between the tRNA core and the 5' leader sequence (Guerrier-Takada et al., 1983; Guerrier-Takada and Altman, 1984; Maquez et al., 2008). The two products of the reaction consist of one strand starting with a 5'-phosphate and a second one ending with a 3'-hydroxyl (Figure 1). When increasing the complexity, self-cleaving ribozymes come next. This family includes ribozymes from viruses and viroides such as the hammerhead (Forster and Symons, 1987; Lambert and Burke, 2008; Scott, 2008), hairpin (Lilley, 2008b), hepatitis delta virus (Wu et al., 1989; Koo et al., 2008) and Neurospora crassa VS ribozymes (Saville and Collins, 1990; Lilley, 2008b). In these ribozymes, the cleavage is mediated by the activated 2'-hydroxyl group of the residue tethered to the scissile phosphate group. The reaction generates a 5' cleavage product with a 2'-3'-cyclic phosphodiester end and a 3' cleavage product with a 5'-hydroxyl end (Figure 1). These ribozymes are able to perform the reaction opposite to the cleavage since the 2'-3'-cyclic phosphodiester is activated and tends to evolve by hydrolysis to generate either a 3' or a 2' phosphate end (Figure 1). Depending on self-cleaving ribozymes, the ligation reaction can be even more efficient than the cleavage reaction.

Ribozyme	*Sequenced Sample*	*Size (nt)*	*Activity*	*Reaction performed*	*Crystal structure*	*Organism*
Group I intron	>10000	>=200	Self-splicing	Nucleotidyl transfer	Yes	B/E
Group II intron	>700	>=500	Self-splicing	Nucleotidyl transfer	Yes	B/E
Group I Like intron	30	~200	Self-cleavage	Nucleotidyl transfer	No	E
Hammerhead	>15	~40	Self-cleavage	Transesterification 2'-O	Yes	E/B/A
Haipin	>3	~70	Self-cleavage	Transesterification 2'-O	Yes	E/B/A
HDV/CPEB3	>5	~90	Self-cleavage	Transesterification 2'-O	Yes	E/
Varkud Satellite	1	~160	Self-cleavage	Transesterification 2'-O	No	E
glmS	>20		Self-cleavage	Transesterification 2'-O	Yes	B
RNase P*	>500	300	Processing	Hydrolysis	Yes	E/B/A
Ribosome*	>10000	>3000	Protein synthesis	Peptidyl transfer	Yes	E/B/A

Table 1

Classification of natural ribozyme according to their main catalytic strategy.

This table is adapted from (Doudna and Cech, 2002; Doudna and Lorsch, 2005)). The number of sequenced examples is a snapshot as of 2010 and is greatly influenced by the new sequencing strategy, technology and database upkeep and may grow exponentially with the rise of the deep sequencing technology. However it gives a rough idea of the ribozyme abundance. (*) The ribosome and the RNase P are both ribonucleoprotein enzyme. Interestingly the RNase P has been shown to possess a relevant catalytic activity in absence of the protein, while no activity has yet been detected with the protein-free ribosome and the large-subunit rRNA alone.

The most complex class of ribozymes, self-cleaving introns like group I and group II ribozymes are genetic elements that interrupt functional genes. They reach one step beyond in terms of complexity since they need to carry out successively two transesterification reactions in order to splice out and to ligate the exons (group I intron reviewed in (Cech, 1990)), group II intron reviewed in (Pyle, 2008; Pyle, 2010; Toor et al., 2010)). Their main pathway consists in two successive transesterification steps. The first one results in the cleavage of the bond between the 5' exon and the intron. The nucleophilic attack is mediated by the 3'-hydroxyl group of an external nucleotide (αG) in the case of group I introns, and by the 2'-hydroxyl group of an internal A residue in the case of group II introns (Figure 1). Consequently, αG is tethered to the 5' end of the group I intron whereas the branching reaction taking place in the group II intron results in the formation of a lariat (Figure 1). The second step of the reaction is more similar in both ribozymes. The 3'-hydroxyl group of the 5' exon attacks the phosphodiester bond linking the intron to the 3' exon promoting splicing as well as exon ligation. It is worth to note that the second catalytic step relies on fine-tuned dynamic

reshaping of the catalytic site in order to bring to the right places the participating chemical groups. It is worth to note that group I ribozymes can also follow alternative pathways that lead to the formation of full length circles either before or after splicing (Nielsen et al., 2003). Truncated circles on the 5' side can also be formed after splicing (Tanner and Cech, 1996; Nielsen and Johansen, 2009).

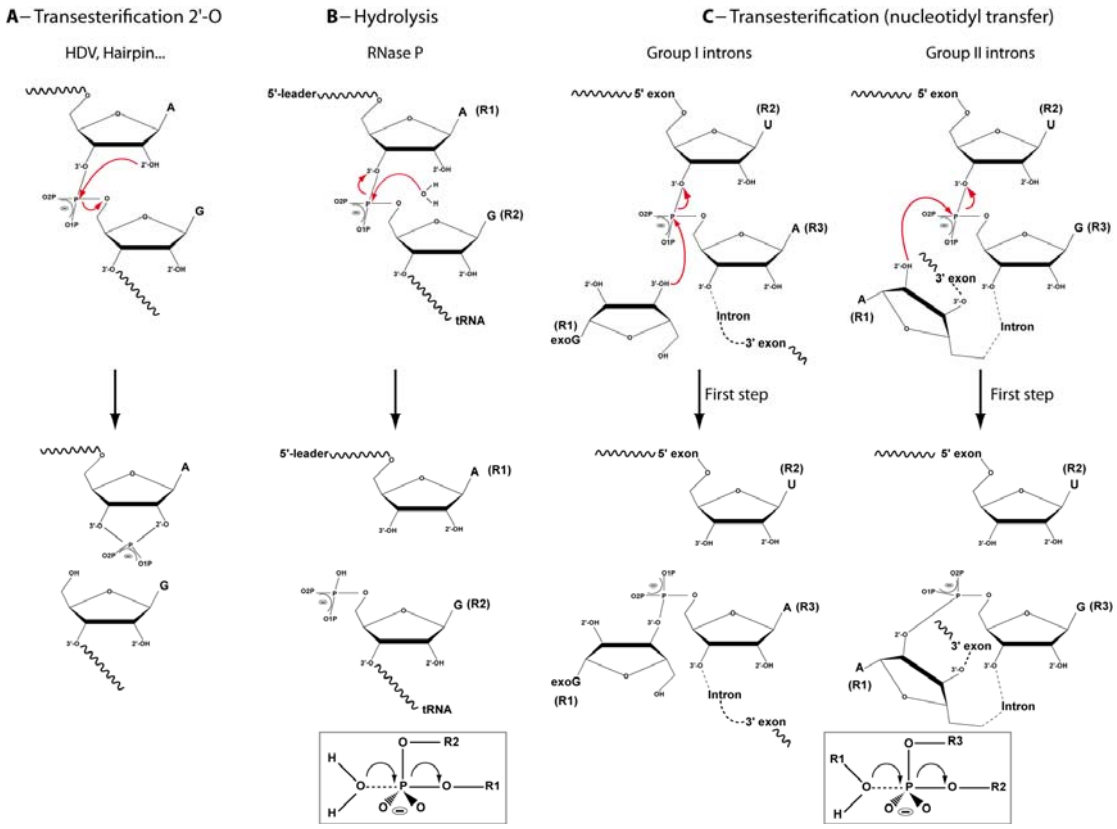


Figure 1
Four types of reactions catalyzed by natural ribozymes.

(A) Self-cleavage reaction of small ribozymes. The reaction involves an attack by the 2' hydroxyl of the nucleotide on the phosphorus atom of the nucleotide. The reaction generates a 5' cleavage product with a 2'-3'-cyclic phosphodiester end and a 3' cleavage product with a 5'-hydroxyl end. (B) Hydrolytic reaction carries out by the RNase P to maturate tRNA molecules. The reaction generates two products which consist of one strand starting with a 5'-phosphate and a second one ending with a 3'-hydroxyl. (C) First step of transesterification reaction catalysed by group I and group II intron.

The diversity of the products results from the architectural diversity of RNA catalytic sites. RNA catalytic sites should be programmed to activate the nucleophile and to facilitate the release of the leaving groups. Since activation often relies on acido-basic catalysis in most self-cleaving ribozymes, catalytic sites should be built so as to control the pKa of critical

chemical groups (Nakano et al., 2000; Bevilacqua, 2008; Luptak et al., 2001). Moreover, group I and II ribozymes orchestrate catalysis by activating the nucleophiles using magnesium ions (reviewed in (Hougland et al., 2006)). These should be bound precisely in space and time to correctly move through the two consecutive catalytic steps. Intuitively, one can realise that it is not easy to build up complex catalytic sites by just tethering the residues directly involved in the reaction. The structure of the catalytic site and ultimately the way it will act on the substrates and products result from the overall structure of the ribozyme. However there is no obvious relationship between the degree of complexity of the reaction and the size of the ribozyme. Self-cleavage ribozymes are usually quite short ranging from 50 to 100 residues for the hammerhead and the VS ribozyme, respectively. Group I and group II introns have between 200 to 400 residues and RNase P with an apparently simple hydrolytic reaction to perform is also about 400 residues long (see Table 1). Thus, the size of a ribozyme also depends from functions that should be distinct from catalysis.

A consequence of the fact that group I introns carry out splicing is that they adopt a particular secondary and tertiary structure (Figure 2). Their catalytic core is composed of short helices (Pi, i=1 to 10) precisely connected by single-stranded junctions (Ji/j) (Cech et al., 1994) (Figure 2). The detailed structure of group I introns will be described in chapter I section 3. Nonetheless, a brief description of some features helps understanding how splicing constraints result in a precise architectural organisation of each RNA structural element with respect to one another. In all group I introns P7 harbours the G cofactor binding site initiating the first step of the splicing pathway (Michel et al., 1990). The internal loop between P4 and P5 serves as a platform for docking the P1 substrate. Consequently P1 is sandwiched between P4/P5 and P7. It is worth to note that P1 and P10 are formed transiently prior to the first and the second catalytic steps, respectively which adds a molecular dynamics dimension to the system. In brief, the intricate network of interactions that constrains very strongly the whole RNA mainly results from the selection pressure for splicing (Golden, 2008). However some parts of the introns do not seem sensitive to them. Sequence analysis show that specific sites appear to concentrate insertion events (Figure 2) either under the form of structural variation of RNA peripheral domains or by integrating exogenic RNA like ORFs (Open Reading Frames) (see chapter I).

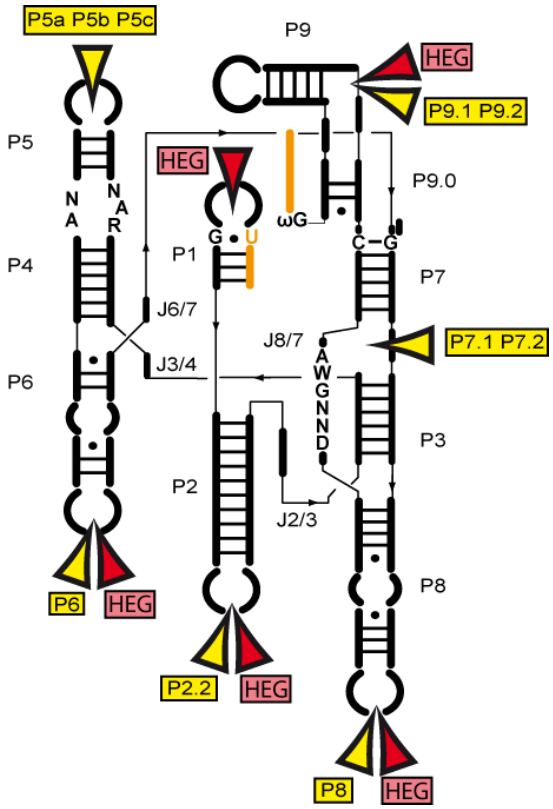


Figure 2
Secondary structure of group I intron catalytic core

Classical secondary structure representation of group I ribozymes according to (Cech et al., 1994) adapted from (Adams et al., 2004a). Invariant residues are represented (P1 G•U, P7 G=C) (R=A/G, W=A/U, D=A/G/U and N= A/C/G/U) according to (Michel and Westhof, 1990).The yellow arrows represent the insertion sites of peripheral domains. The red arrows represent the insertion site of Open reading Frames Encoding for Homing Endonuclease.

Structural variations of RNA peripheral domains have been extensively studied and have led to the phylogenetic classification of group I introns in 13 subgroups (Michel and Westhof, 1990). However the impact of these differences on catalysis and folding of only some group I introns phyla has been systematically studied (Woodson, 1992; Lehnert et al., 1996; Schroeder et al., 2004; Chauhan and Woodson, 2008; Chauhan et al., 2009). Sometimes introns are accompanied by ORFs encoding maturases, endonucleases or other genes (see chapter I). The relationship between group I introns and their encoded ORFs has been much less studied. Many group I introns have insertions corresponding to a homing endonuclease gene (HEG). The homing mechanism is thought to mainly account for dissemination of introns (Haugen et al., 2005a; Haugen et al., 2005b). In such cases, introns can be considered as selfish elements that remain neutral to the host since splicing leaves the sequence at the insertion site unchanged. Nonetheless sequencing data show how complex some group I

introns are and raise the fundamental questions of what are the evolutionary causes and the functional consequences of those situations for both the introns and their hosts.

Group I introns are distributed from bacteria to lower eukaryotes with a main incidence in fungi and an apparent absence from Archaea (Haugen et al., 2005a). Some of those are highly complex like the obligatory intron found in the gene of subunit 5 of the mitochondrial NADH dehydrogenase (ND5) from hexacorals (Figure 3, (Beagley et al., 1996; van Oppen et al., 2002)). Phylogenetic studies (Medina et al., 2006) show that in the course of evolution from Octocorallians to Corallimorpharians, the P8 element of this intron has gradually integrated more and more mitochondrial genes including other NADH dehydrogenase subunit genes, ribosomal RNAs and the cytochrome oxidase I (COI) which contains a HE-encoding group I intron. The observation that the ND5-717 intron is maintained throughout the phylogenetic tree prompts to ask wether it participates in regulating the expression of the NADH dehydrogenase which has a key role in respiratory control. Complex intronic context can also be observed in bacteria. In the human pathogen *Clostridium Difficile*, an intron (Cd*ISt*1) is inserted in the *tcdA* enterotoxin gene (Hasselmayer et al., 2004a; Hasselmayer et al., 2004b). This intron is fused with an insertion element (IS) encoding a degenerated transposase (TlpA) and a second functional one (TlpB). The association of an IS-element to an intron (IStron) represents an example of molecular symbiot but strikingly, it is always inserted within a protein coding genes. IStrons seem to have a leaky 3' splice site which may induce protein diversity by modifying the length and sequence of the 5' end of the second exon. In the case of the Cd*ISt*1 the IStron may well help the toxin gain structural diversity in order to vary its effect on the host. Mutants may be selected if they confer a selective advantage to the bacterium. Interestingly, the degenerated IS-element has been fused to the P9 region of the intron where insertions do not prevent splicing of the ribozyme.

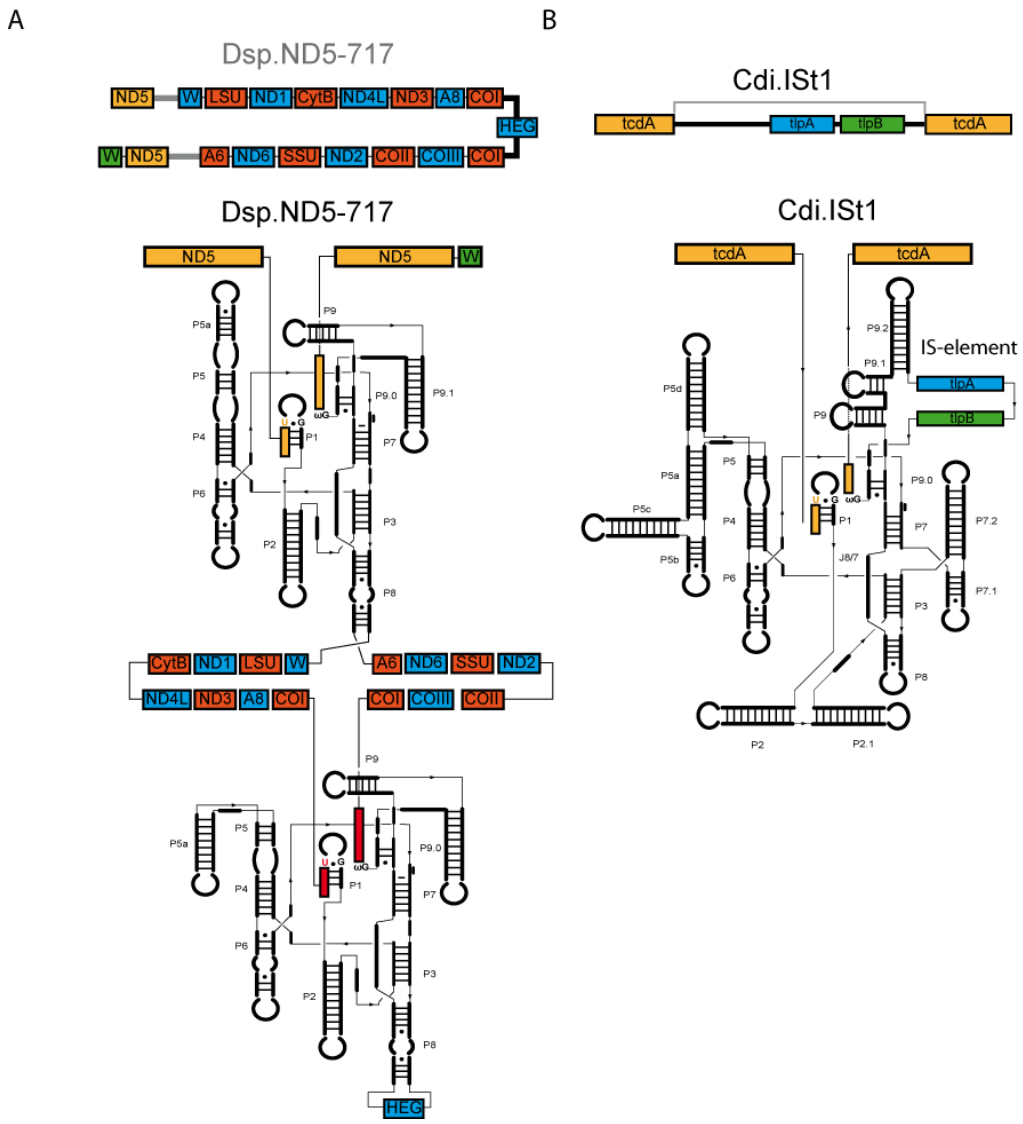


Figure 3
Highly complex group I introns
(A) Schematic secondary diagram representation of the ND5-717 intron obligatory in hexacoral mitochondrial genome. The 18.807 nt Discosoma presented herein contains 15 additional mitochondrial genes (boxed blue and red) embedded within the structure of the intron. Only the tRNATrp gene (W) is freestanding in the mitochondrial genome. The cytochrome oxidase I (COI) gene is interrupted by a HEG-containing group I intron. (figure from (Nielsen and Johansen, 2009)). (B)) Schematic secondary diagram representation of IStron Cd*I*St1 from ***C. difficile*** (Hasselmayer et al., 2004a; Hasselmayer et al., 2004b).

These examples chosen among many others suggest that introns are not only selfish transposable elements but can also benefit their host. Studying such complex introns in their biological context should allow for better understanding their relationship and very importantly the molecular mechanism by which such a relationship has appeared, is maintained and works. The ambition of this thesis is to illustrate these concepts by studying a

complex intron found in the small ribosomal subunit (SSU) gene of the myxomycete *Didymium iridis* and the amoebo-flagellates *Naegleria*. This intron is made from a HEG-containing group I intron (GIR2) and another ribozyme (GIR1) inserted right before the HEG coding sequence (Figure 4). Although GIR1 strongly resembles a group I intron (Einvik et al., 1998c), Nielsen et al (2005) have demonstrated that it instead carries out a branching reaction leading to the formation of a tiny lariat at the 5' end of the homing endonuclease (HE) mRNA. This work had considerable impact on our view of this twin-intron. Actually, since GIR1 is fully transcribed before GIR2, early cleavage would prevent correct splicing of the whole intron together with the ligation of ribosomal exons leading to depletion of active ribosomes in the cell (Vader et al., 2002). To prevent this situation to occur, a control mechanism should exist that couples the activities of the two ribozymes. Analysis of the pre-rRNA transcripts of *D. iridis* cells under starvation-induced encystment have shown the accumulation of a 7.5 kb RNA corresponding to the cleavage product of GIR1 extending through the HEG up to the end of the spliced large subunit (LSU) rRNA (Vader et al., 2002). These results demonstrate the coupling between the two ribozymes and motivated us to investigate the intron at different levels. It was first needed to better understand the molecular basis of the GIR1 branching reaction (**Paper I** (Beckert et al., 2008)). We deduced from this work and other data that the flanking sequences of GIR1 may have a role on its catalytic activity. Thus, we have started studying the effect of the flanking sequences on the structure of the catalytic core and the shape of GIR1 (**Paper III**) and on the release mechanism of the branching product from the ribozyme (**Paper II**). In parallel, we have designed constructs for crystallisation studies which are still under process. Finally, we turned towards GIR1 ribozymes from Naegleria in order to explore whether the strategy used by the twin-intron of these organisms to couple the activities of both ribozymes was identical or different provided the RNA regions at the GIR1 interface are different in *D. iridis* and in Naegleria (Supp Results in chapter V).

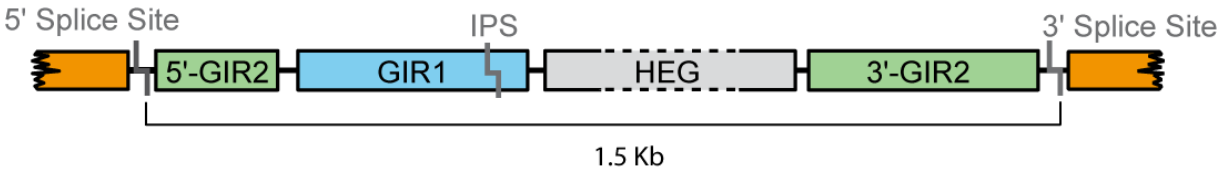


Figure 4
Linear representation of the twin-ribozyme intron with the internal processing site (IPS) of the GIR1 ribozyme.

This corpus of data gives the following manuscript which is organised as follows: the introduction compiles the structural knowledge on group I introns in order to deliver the keys to understand the mechanism of the twin-intron from *D. iridis* and Naegleria. The results are then presented under the form of publications or preprints. Additional data that have obtained at the end of the PhD period are presented in the Supplementary Results section. Finally, a general conclusion and perspectives section follows summarising the milestones of the work and suggesting further ideas that still need to be experimentally addressed.

CHAPTER I: GROUP I INTRONS

1. Function of group I intron:

Most group I introns are able to perform two different reaction pathways. Those two different pathways recognized within the group I introns are “the self-splicing pathway” and “the circularization pathway”. Even through the two pathways are somewhat similar in their individual steps, they lead to the formation of different products (Nielsen et al., 2003).

1.1. The self-splicing pathway:

During the self-splicing pathway, group I introns catalyse their own excision from the primary transcript by two coupled transesterification reactions followed by the release of the ligated flanking exons RNA strands (reviewed in (Cech, 1990)). Many group I introns are able to promote their self-splicing activity in the absence of proteins. However, some others retain some poor self-splicing activity *in vitro* or only at high non-physiological Mg^{2+} concentration. They require protein co-factors or maturases in order to be catalytically active (Garriga and Lambowitz, 1986; Wallweber et al., 1997). Regardless of the maturase dependency to perform their splicing reaction, the chemical mechanism is the same (Figure 5). The splicing pathway of group I introns has been well characterized for the Tth.L1925 intron from the large subunit rRNA gene of the ciliate *Tetrahymena thermophila* by biochemical and biophysical methods. Remarkably it has been found to be representative for the self-splicing pathway of all the group I introns (reviewed in (Cech, 1990)).

The ribozyme first reaches its active conformation (Figure 5-1). During the folding process, the 5’ exon base-pairs with the internal guide sequence (IGS) in order to form the P1 hairpin. P1 is then positioned in the vicinity of the catalytic core of the ribozyme adjacent to the G-binding site (Figure 5-2). Given that the 5’ splice site is recognized and the binding of one exogenous guanosine (exoG) or one of its phosphorylated forms (GMP, GDP, GTP) to the G-binding site occurs, the first step of the splicing is initiated. The 3’OH group of the exoG acts as a nucleophile and attacks the phosphorous atom at the 5’ splice site. This results in breaking the bond between the 5’ exon and the intronic part (Figure 5-3). Then a new

covalent bond between the exoG and the first nucleotide of the intron is formed. Following the first step (Figure 5-4), conformational rearrangements occur in which the exoG exits the G-binding site and is replaced by the conserved guanosine residue at the 3'end of the intron (ωG) and finally the P10 domain is folded. The second transesterification reaction can then occur. The newly formed 3'OH group of the 5'exon then attacks the 3'splice site and simultanously completes the splicing reaction by forming the phosphodiester bond between the exons (Figure 5-5). The ligated exons are then released from the intron (Figure 5-6) (reviewed in (Cech, 1990)). This ability to self-splice has the potential to render the presence of group I introns genetically neutral to their host.

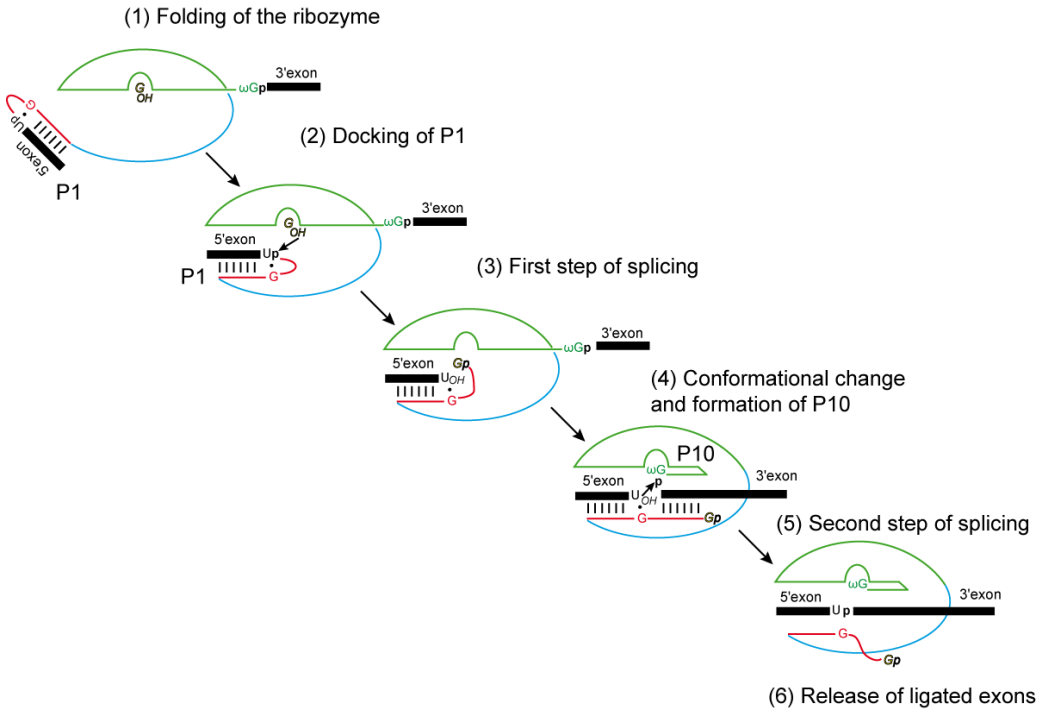


Figure 5
Self-splicing pathway of group I introns scheme.

In order to accomplish the first splicing reaction step, group I ribozymes must first bind the exoG in the G-binding site. However, during this step a competition between the binding of either the exoG or the ωG in the G-binding site may exist (Rangan et al., 2003). On the opposite, the second splicing step is only restricted to the binding of ωG. Thus, the relative affinities for the two guanosines seem to change during the splicing pathway thereby driving the reaction to completion (Golden and Cech, 1996; Zarrinkar and Sullenger, 1998). The selection of which of the two guanosines that should bind to the G-binding site is provided by sequestration of ωG during the first splicing step. After the first transesterification reaction,

conformation changes release ωG which can dock to the G-binding site (Rangan et al., 2004). In this way, the ability of a group I intron to perform its self-splicing pathway relies on the ribozyme capacity to adopt a particular 3D conformation. However, in addition to the self-splicing pathway, group I introns are able to perform an alternative pathway, the circularization pathway. What are the characteristics of the pathway? And what are the differences and the similarities between these two pathways?

1.2. The Circularization pathway:

Group I introns have been shown to be able to form three types of circular RNAs: "truncated intron circle"; "full-length circles" and "longer than full-length circles". The first type of circular group I introns are truncated intron circles (Figure 6). They result from the circularization of the spliced out intron. In this particular case, after the self-splicing reaction, the intron terminal residue (ωG) is still docked in the G-binding site. The ωG makes an attack at an internal phosphodiester bond within the intron 5'end. It leads to the formation of a circular RNA and the release of a short RNA (Tanner and Cech, 1996).

The second type of circular group I introns are full-length circles (FLC) that are specific of the circularization pathway (Haugen et al., 2004b; Nielsen et al., 2003; Zaug et al., 1983) (Figure 6). Like the self-splicing pathway, the circularization pathway is a two step reaction: a hydrolytic cleavage directly followed by a transesterification reaction. This alternative pathway to the self-splicing one is initiated by the docking of the ωG in the G-binding site. This situation is likely to mimic the second step of the self-splicing pathway with the difference in that there is no free 3'OH group on the 5'exon. The first reaction, the hydrolysis reaction is performed by a water molecule that attacks the 3' splice site. Following the 3'splice site hydrolysis, the 3'OH of ωG performs a nucleophilic attack at the 5'splice site resulting in the formation of a covalent full length circle (FLC). The formation of FLC results in unligated exons and non-functional gene products (Nielsen et al., 2003).

Finally the third type of circular group I intron that has been recently described in an *in vitro* study of the splicing of the *Anabaena* $tRNA^{Leu}$ intron (Vicens and Cech, 2009) is longer than the FLC (Figure 6). In this particular intron, the formation of the circles results from the circularization of the spliced out intron. Unlike the case of the formation of a

truncated circle in which ωG makes an attack at an internal phosphodiester bond within the intron near the 5'end, ωG attacks the triphosphate of the GTP that was coupled to the intron 5'end after the first step of splicing. Thus, pyrophosphate is released and the circles incorporate the guanosine co-factor.

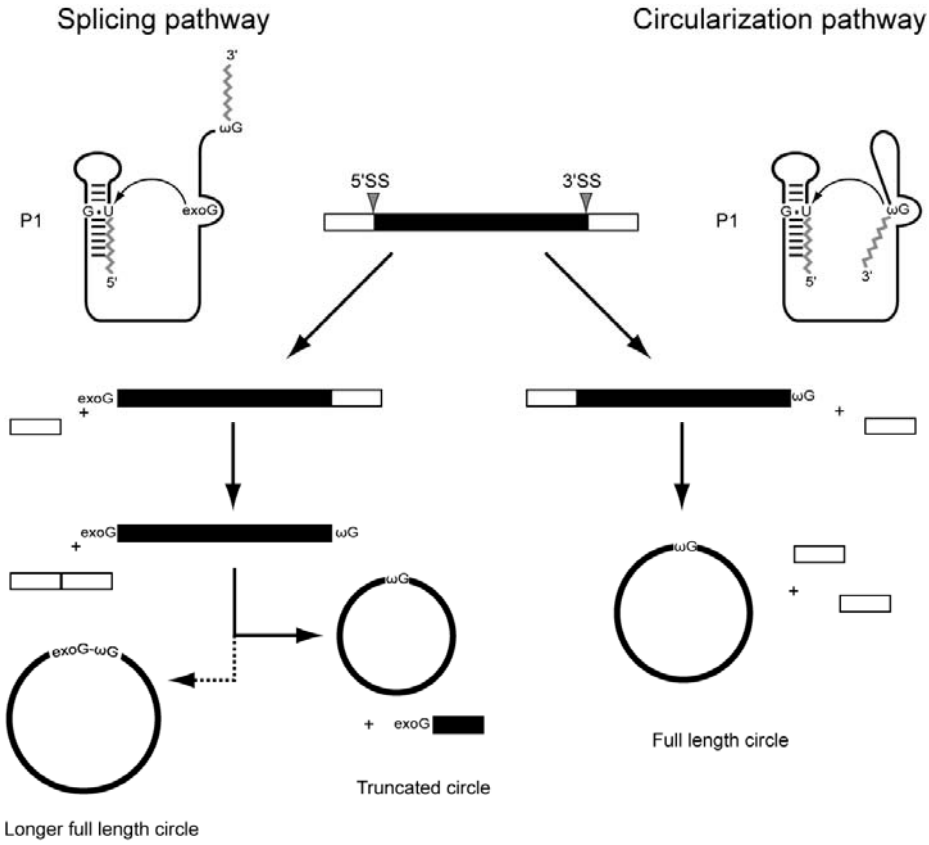


Figure 6
Schematic representation of the group I introns' processing pathways.
(Left): The splicing pathway. This pathway is initiated by the binding of the exoG at the G-binding site within the intron core. The nucleophilic attack by the 3'OH of the exoG at the 5'splice site results in the breaking of the bond between the 5' exon and the intronic part and the covalent binding of the exoG to the intron. Then the 5'exon is positioned to attack the 3' splice site. The second transesterification reaction leads to the ligation of the 5' and 3' exons and the release of the linear intron. Further processing leads either to the formation of truncated circle or either to the formation of circle longer than the full-length (Vicens and Cech, 2009). (Right) The full length circle formation pathway. The circularization pathway starts with the binding of ωG to the G-binding site. The ribozyme then catalyzes a hydrolytic reaction at the 3' splice site. Subsequently a transesterification reaction mediated by the attack of ωG at the 5' splice site results in the formation of the full length circle and two free non-ligated exons. (figure adapted from (Nielsen et al., 2003)).

Most group I introns are found inserted in rRNA genes from all three phylogenetic kingdoms. However they seem to have a sporadic distribution. It points out to a complex evolutionary past (Belfort and Perlman, 1995b). Phylogenetic studies and mapping of group I

introns insertion sites in the rRNA have revealed that the group I introns are present in more than 50 distinct integration sites.

2. Group I intron mobility:

A common spreading mechanism of group I introns is vertical inheritance. Once inserted into a genome an intron can be stably maintained within its host gene over long periods of time. The $tRNA^{Leu}$ group I intron is in this view a remarkable example. It has been retained in the plastid DNA after primary endosymbiosis of the cyanobacterium into a primordial eukaryote (Kuhsel et al., 1990; Xu et al., 1990). Therefore, $tRNA^{Leu}$ introns can be used to infer host phylogeny in the same manner as the commonly used ribosomal DNA (Simon *et al.*, 2003). The phylogenetic observations that highly similar introns are often inserted in the same site of several different species, support that group I introns have been lateraly transferred. This strongly indicates that introns are not passive genetic entities but rather dynamic mobile elements that can both be lost or gained by horizontal transfer. Thus, vertical inheritance and horizontal transfer are the two mechanisms that account for the particular distribution of group I introns (Figure 7). Intron spreading can be mediated by homing endonucleases (Figure 7 A) often found inserted into the peripheral regions of group I introns. Mobility may be also achieved by the ribozyme itself, presumably by reversal of the splicing reaction (Figure 7 B).

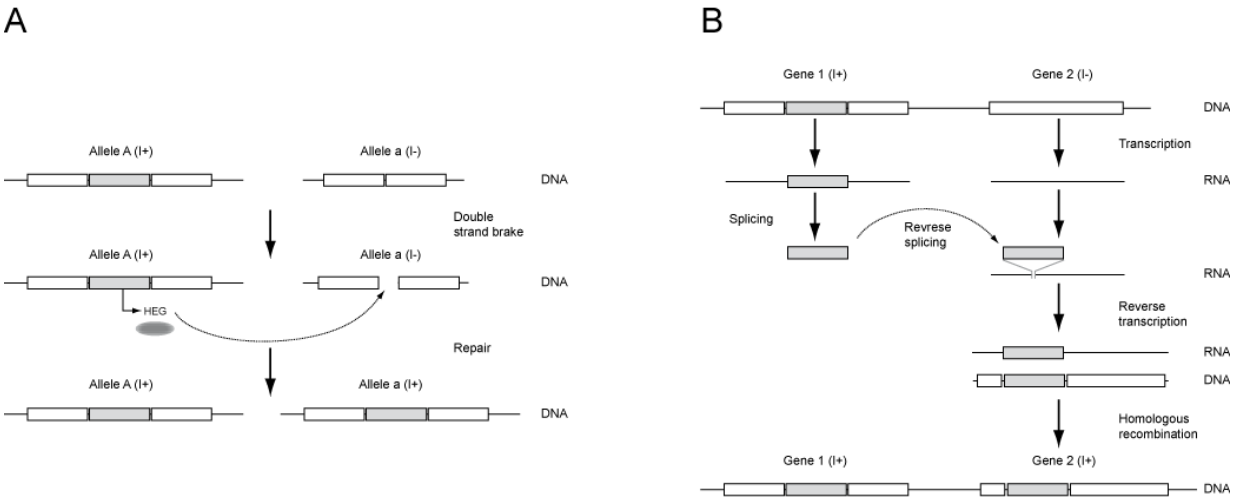


Figure 7
The two mobility mechanisms proposed for group I introns.
(A) HEG-mediated transfer via double break repair mechanism. The endonuclease is expressed from an intron I+, recognizes a specific sequence in the allele I- and then creates a

double stranded break. The damaged allele is then repaired using as template the allele containing the intron. (B) Model for ribozyme-mediated transposition. Two different genes (Gene 1 and Gene 2) contain an intron I+ and an intron-less respectively. After transcription the intron I+ is processed and released from the RNA precursor. Subsequently the intron reverse splice into the intron-less RNA. A DNA homolog is then created by reverse transcription of the new RNA template intron I+. By homologous recombination of the flanking sequences the intron is integrated into the genome.

2.1. Homing endonuclease mediated mobility at the DNA-level:

Homing endonuclease genes (HEGs) are mobile genetic element. They are widely distributed in fungi, protists, bacteria and viruses. They are found as freestanding genes or embedded in the peripheral regions of group I introns. The homing endonucleases have been classified into four different families characterized according to their conserved structural motifs: LAGLIDADG, GIY-YIG, H-N-H and His-Cys box (Belfort and Roberts, 1997; Chevalier and Stoddard, 2001). They have no known host function and are rather thought to exemplify selfish parasitic elements (Goddard and Burt, 1999a). The homing endonucleases found in group I introns promote intron homing in a process where the intron moves from an intron-containing allele to an intron-less allele by the double-strand-break-repair (DSBR) pathway (Belfort and Perlman, 1995a). The site specific transfer is initiated by recognition of an asymmetric target sequence of 14-40 bp with a tolerance to single site mutation by the homing endonuclease (Galburt and Stoddard, 2002). The endonuclease makes a specific double-strand break in the intron-less allele. The break is subsequently repaired by gene conversion from the intron containing allele (reviewed in (Lambowitz and Belfort, 1993)).

In nature, group I introns and associated homing endonuclease genes appear to undergo an evolutionary cycle of gain and loss currently known as the Goddard-Burt cyclical model (Goddard and Burt, 1999b) (Figure 8). This model describes the life cycle of the intron and its associated homing endonuclease gene. When all the alleles in a population are occupied, the HEG has no longer a biological function and then a degeneration process is initiated. The HEG accumulates mutations, becomes truncated, non-functional and finally lost from the group I intron by sporadic deletion. Unable to spread, the intron is then lost and in this way, leaves the host allele free for recurrent new insertions or invasions to restart the cycle. Consequently, similar introns with different HEG states (functional/mutated/degenerated) can be found in the same position in the intron of different

genome species. Introns do, however, manage to "escape" from the cycle by inserting into a new genetic position.

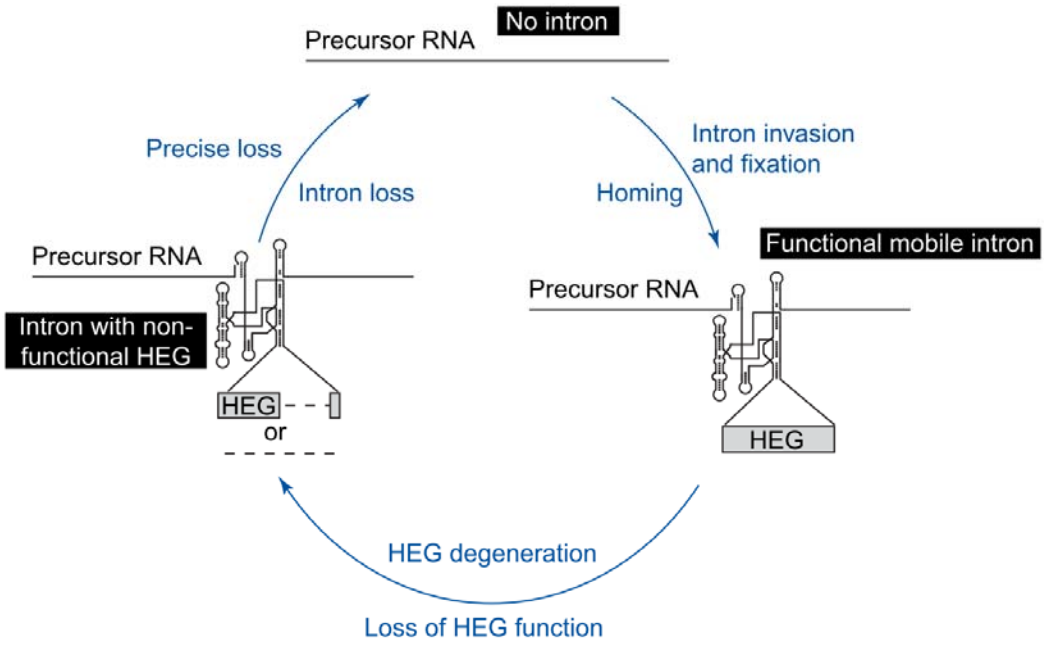


Figure 8
The Goddard-Burt cyclical model (Goddard and Burt, 1999b).
The evolutionary model of group I intron gain and loss is divided into different steps. 1) Intron invasion and fixation: A mobile intron with a functional HEG invades an intron-less population and becomes fixed in all the homologous insertion sites via the homing process. 2) HEG degeneration and loss of function: because all the homologous already possess an intron and are protected from the HEnase activity, the HEG does not need to be active decreasing selection pressure toward it function. The HEG becomes redundant, mutated and eventually lost. However some introns may escape from the cycle and gain a new function. Depending on the selection pressure those new RNA elements can be conserved or be transient (Nielsen and Johansen, 2009). 3) Intron precise loss: Finally the intron is lost leaving the host allele free and ready for a new invasion.

Homing endonucleases encoded by group I intron have been shown to bind to their cognate intron RNAs. They have been demonstrated to have a maturase activity, i.e. promoting the splicing of the intron (Belfort, 2003; Longo et al., 2005) (See paragraph 5: Group I intron looking for protein partner). Thus, both mobility and splicing of some group I introns rely on the expression of their embedded protein-coding genes. How protein-coding genes embedded in nuclear ribosomal DNA can be expressed and also regulated?

2.2. Mobility at the RNA level:

Homing endonucleases found in group I introns promote intron homing by the DSBR pathway. However, this particular pathway is highly allele specific and does not favour transposition of introns to other loci. Moreover, the particular wide phylogenetic distribution of group I introns and the fact that only about 30% of group I introns encode a protein with a putative endonuclease activity (reviewed in (Guhan and Muniyappa, 2003)) support the idea that other processes may contribute to the site-specific insertion or transfer of group I introns. Interestingly, transesterification reactions catalysed by the group I ribozymes are reversible. Thus, a mechanism based on reverse splicing directly followed by reverse transcription of the recombinant RNA and integration into the genomic DNA could contribute to intron mobility (Figure 9). By this process, the intron mobility could result either in homing, if the intron integrates into a homologous site or an intronless allele, or transposition, if the intron integrates into a heterologous site.

The mechanism responsible for group I intron heterologous site invasion is mainly based on the reverse-splicing ability of group I introns shown by *in vitro* studies done with the *Tetrahymena* intron (Woodson and Cech, 1989). This reverse-splicing mechanism depends on the group I ribozyme ability to reform a P1 helix by recognition of a 4-6 nt target sequence complementary to the intron encoded internal guide sequence (IGS). However *in vitro* integration experiments between the linear *Tetrahymena* intron and the 23S *E. coli* rRNA have shown that there was a large degree of flexibility in the selection of the 5' sequence for the formation of the stable P1 (Roman and Woodson, 1995). Thus, the integration site differed significantly from the original 5' exon and the frequency of the reverse splicing site was modulated by the structure of the rRNA (Roman and Woodson, 1995). Nevertheless, reverse splicing was shown to occur only when the P1 helix was reformed and if it could be docked into the intron core then the forward splicing was promoted (Woodson and Cech, 1989). *In vivo* studies based on the expression of the *Tetrahymena* intron in *E. coli* have confirmed these *in vitro* observations. The *Tetrahymena* intron is able to reverse-splice into sites homologous to the splice junction of the *Tetrahymena* organism (Roman and Woodson, 1998). Interestingly, the ribozyme has also been shown to integrate into several non-homologous sites but less efficiently (Roman and Woodson, 1998; Roman et al., 1999). Additional experiments also showed that a strong P10 was required for the choice of the

integration site and in some cases it could even enhance reverse-splicing reactivity (Roman et al., 1999).

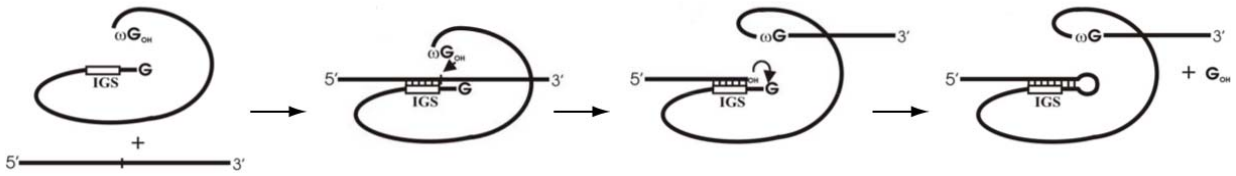


Figure 9
Model of reverse splicing process leading to intron integration at the RNA level.
The reverse splicing is initiated by recognition of a substrate that possesses a sequence complementary to the IGS. The 3'OH group of the ωG attacks the novel 3' splice site in the substrate. Then the 3'OH group of the novel 5' exon attacks the phosphodiester bond between the intron and the exoG attached to its 5' end. This results in the integration of the intron in the RNA molecule and also the release of the exoG that was previously incorporated in the spliced intron.

It is worth to note that, depending on the intron, the reverse-splicing efficiency and integration site diversity could vary. *In vitro* and I*n vivo* reverse-splicing studies of the twin-ribozyme intron containing the DiGIR2 ribozyme promoting the formation of FLC intron (Nielsen et al., 2003) (**Paper IV**), revealed that this particular ribozyme reverse-splices into its phylogenetically conserved insertion site (i.e. S956) in the SSU rRNA of *E. coli* and yeast. Further analysis of the entire *E. coli* SSU rRNA confirmed that intron integration was exclusively restricted at the site S956 (Birgisdottir and Johansen, 2005). Remarkably, DiGIR2 possesses a high affinity for its own integration site. This is to put in contrast with the previous observation done with the *Tetrahymena termophila* ribozyme which partially reverse splices into its own insertion site but also targets several novel rRNA sites (Roman and Woodson, 1998).

Although the mechanism by which the group I introns invade heterologous genetic sites and then integrate in a DNA gene remains elusive, the reverse-splicing remains the most plausible model. While the homing model remains highly efficient but highly targeted with the recognition sequence recognition of 15-45 nt, the reverse-splicing seems to be less targeted with its sequence requirement of 4-6 nt and also less efficient because of the need of two additional steps (i.e. reverse-transcription and then recombination). Reverse-splicing between a linear group I ribozyme and a ligated exon or rRNA has been shown to take place *in vitro* (Roman and Woodson, 1995; Woodson and Cech, 1989) and also *in vivo* (Roman and

Woodson, 1998; Roman et al., 1999). Interestingly, the circular form of group I introns could also be a serious candidate for intron mobility by reverse splicing.

2.3. The FLC intron, a possible role in the group I intron mobility?

During *in vitro* processing by some group I introns, the different reactions catalyzed were shown to promote accumulation of different intermediates (Johansen and Vogt, 1994; Decatur et al., 1995; Nielsen et al., 2003; Haugen et al., 2004a). Interestingly, in the case of the DiGIR2 group I ribozyme, the formation of FLC introns and the circularization pathway products were detectable both by northern blot analysis (Vader et al., 1999; Vader et al., 2002) and qRT-PCR (**see Paper IV**). Moreover, the qRT-PCR revealed that these FLC introns were accumulating in the cell in response to external factors (**Paper IV**). These observations support the idea that the FLC is a biologically relevant molecule and also reflect that RNA circularity increases the resistance against degradation (Harland and Misher, 1988; Chan et al., 1988). Then, it was speculated that the presence of those FLCs may have several biological roles: (1) they could act as intermediates in the expression of the intron-encoded homing endonuclease by stabilizing the messenger RNA, in other words they could be regarded as expression vectors (Nielsen et al., 2003); (2) they could also be involved in translocation from the nucleus to the cytoplasm prior to translation (Haugen et al., 2002); (3) finally, they could be involved in the group I intron mobility (**Paper IV**).

Why only the FLC introns can be involved in the mobility? First, these FLC are in fact constituted by the entire intron sequence in comparison with the truncated circle. Thus, the junction sequence created between the 5' end and the 3' end of the intron in the FLC can be virtually used to recognize the integration site in the target RNA (Nielsen and Johansen, 2009). Second, the model proposed for FLC integration is based on the energetic advantage that could drive the integration reaction by breaking three phosphodiester bond (circle opening, removal of exoG and splitting of the ligated exon) and formation of two phosphodiester bonds (establishment of the 5' and the 3' splice site) (Nielsen and Johansen, 2009). The circle integration mechanism is currently under investigation but the role for the FLC in intron mobility and horizontal transfers has been already suggested (Haugen et al., 2002; Nielsen et al., 2003) and shown to integrate ligated RNA exons *in vitro* (Birgisdottir, 2005) but with an unknown mechanism currently under investigation.

In RNA, function is dictated by the structure. The ability of a group I intron to perform all its reactions thus relies on the ribozyme capacity to reach its active conformation.

3. Structure of group I intron:

Group I introns vary significantly not only in size but also in there self-splicing abilities. Group I introns have been categorized into 5 main groups (IA-IE) with a total of 13 subclasses by comparative sequence analysis (Lehnert et al., 1996; Li and Zhang, 2005; Michel, 1990).

3.1. Secondary structure representation improvement according to the first 3D model:

Group I ribozymes are highly structured and their structure has been extensively characterized using biochemical methods: chemical and enzymatic structure probing (Inoue and Cech, 1985), mutagenesis, chemical footprinting (Latham and Cech, 1989), covalent cross-linking (Downs and Cech, 1990a; Wang and Cech, 1992; Wang et al., 1993). They all retain a central core with a conserved secondary structure composed of a series of base-paired helices (P) numbered P1 through P9 (Michel *et al.*, 1982). The paired helices are separated by loops (L) or connected together by single-stranded junctions (J), named according to the two helices that are linked (Figure 10 A and B). The amount of data and the alignments of group I intron sequences available in the 80's led to the building of the first three dimensional model of the group I intron core sufficiently detailed to guide knowledge-based experiments (Michel, 1990) (Fig 3 C).

Figure legend: Group I intron secondary structures and the first 3D model
Improvements of the group I intron secondary structure schema drawing and the first 3D model of group I intron. (A) Classical representation of the group I secondary structure (adapted from Burk et al. 1987). Paired segments (P1-P9) are indicated. For clarity of the figure only the universally conserved residues are included. Exons are shown as thick black lines. (B) The modern secondary structure representation of group I intron core (adapted from Cech et al. 1994). Domains and helices are represented more accurately according to their organization within the intron. Principal domains are separated by dotted lines. (C) The first 3D model of group I intron done by Michel and Westhof based on sequence alignment (from Michel & Westhof 1990).

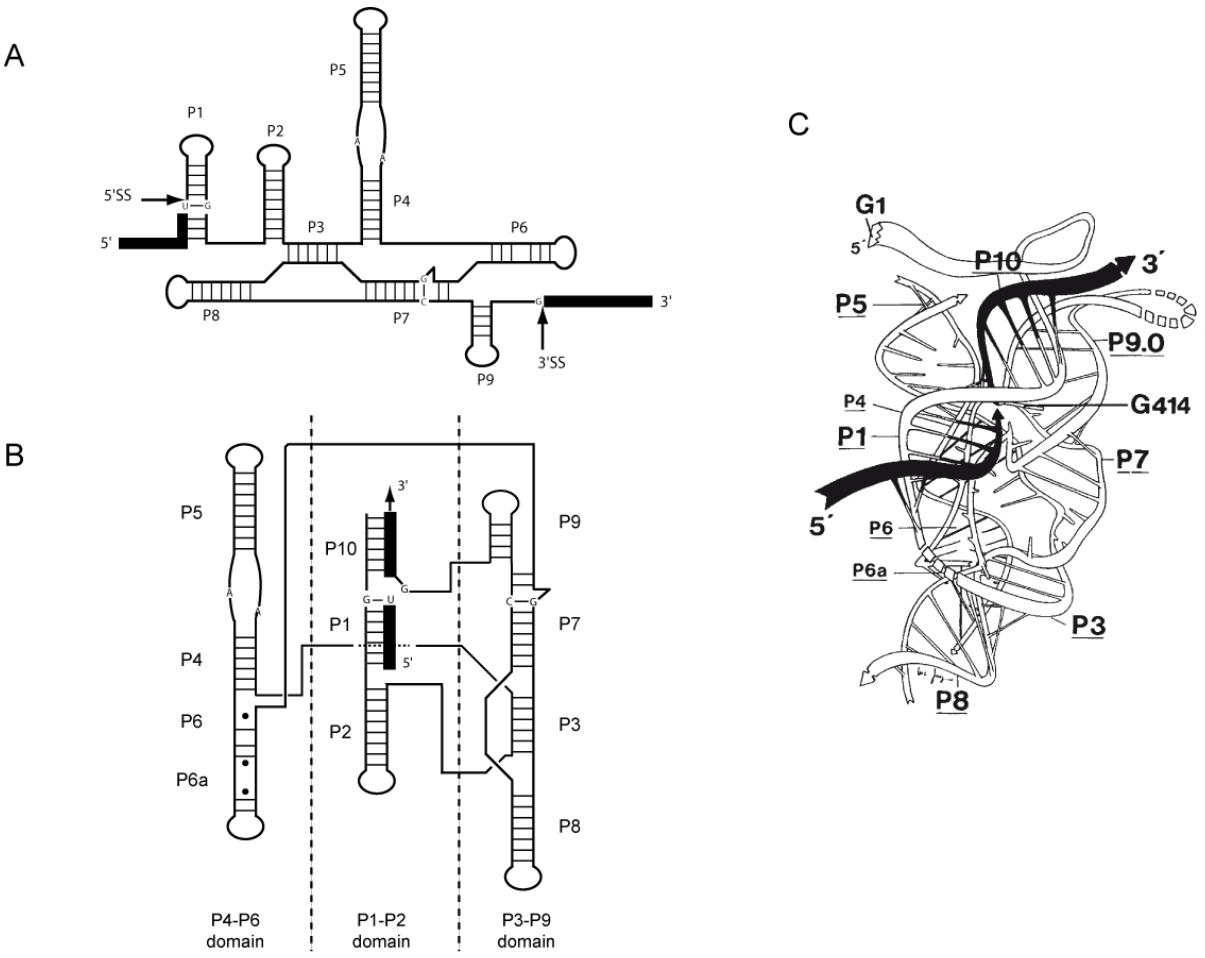


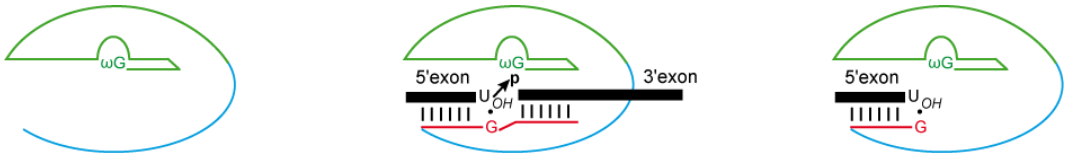
Figure 10
Group I intron secondary structures and the first 3D model

As deduced from biochemical data in combination with structure modelling, the secondary structure of the phylogenetically conserved core of the ribozyme can be divided into three helical domains: P1-P2 (substrate domain), P4-P6 (stabilization domain) and P3-P9 (catalytic domain). The conserved regions of the P1-P2 and P4-P6 domains consist of 2 and 3 helices respectively that are nearly coaxially stacked (Figure 10 B and C). The P3-P9 domain contains the pseudoknot P3/P7 conserved in all group I introns and is interrupted by the P4-P6 domain. The P4-P6 domain serves as a scaffolding domain. Tertiary contacts between P5 and the P9 loop (L9) of the catalytic domain and also between P3 and P6 allow the P4-P6 domain to wrap around the P3-P9 domain (Figure 10 C). This creates a cleft at the domain interface into which can dock the P1 hairpin harbouring the 5' splice site (Kim and Cech, 1987; Michel, 1990) (Figure 10 C).These structural elements represent the catalytic core of the intron and are conserved in the vast majority of the self-splicing group I introns.

Although group I ribozymes are highly structured and the secondary structure of their core is highly conserved, the comparison of their primary sequence reveals very few conserved positions (Michel, 1990). Conserved residues are directly involved in the recognition and positioning of the substrate. These conserved nucleotides spread over the three different domains like in the substrate domain with the G•U wobble base pair from P1; in the stabilization domain with the internal bulge J4/5 located in the P4-P6 domain and finally in the catalytic domain with both ωG and the G-C base pair at the G-binding site in P7. As an example the A residues from the J4/5 have been shown to be important for the correct 5' splice site selection. ωG participates directly in the reactions catalysed by the ribozyme and the G-C base pair plays an important role in selection and the coordination of the exoG and the ωG residues that are bound alternatively to the G-binding site (Michel, 1990).

3.2. The advent of crystallographic structures:

Generally crystal structures of enzyme or ribozyme represent a major breakthrough in our understanding of their activity. Twenty-two years after the discovery of the first ribozyme, crystal structures of different group I introns from three distinct types of exons (rRNA, mRNA and tRNA) and trapped in three distinct chemical states (Figure 11) have revealed at the atomic level their catalytic strategy. Three different group I intron structures were solved: the *Tetrahymena* intron (IC1) (Golden et al., 1998; Guo et al., 2004), the $tRNA^{Ile}$ intron from *Azoarcus* (IC3) (Adams et al., 2004a; Adams et al., 2004b) and the orf142-I2 intron form Twort bacteriophage (IA2) (Golden et al., 2005).



Tetrahymena crystal structure | *Azoarcus* crystal structure | Twort crystal structure

Figure 11
Schematic representation of the three distinct chemical states in which the ribozymes were crystallized.

3.2.1. Overview of the atomic level architecture of group I introns:

3.2.1.1. Global overview of the 3D architecture:

The three different group I intron structures belong to three different group I intron families (IC3 ,IC1, IA2), they show a global organisation and a catalytic core structures remarkably similar (Dror *et al.*, 2005) even if they are locked in different states of the self-splicing reaction pathway (Figure 12 A). Interestingly these crystal structures globally confirm the previous 3D models (Lehnert et al., 1996; Michel, 1990; Rangan et al., 2004). As an example, the overall *rsmd* of 3.85 Å between the *Azoarcus* crystal structure and its model, (Rangan et al., 2004) shows that the architectural elements were correctly predicted and positioned with respect to one another. Moreover, models can provide a good basis for designing experiments (Adams et al., 2004a). Deduced from biochemical data in combination with 3D modelling and confirmed by crystal structures, the structure of group I introns is able to form by the assembly of the three main domains: (P4-P6), (P3-P9), (P1-P2) (Figure 12 A and B). Interestingly, the 3D structures confirm that the conserved nucleotides found in the alignment to be localized mostly in the P4, P6 and P7 helices (Michel, 1990) are clustered in and around the active site (Figure 12 C).

The striking point of those three different crystal structures comes from the G-binding site located in P7 which is occupied by the 3'-terminal ωG. Even if the ribozymes are locked in different states of the self-splicing reaction pathway, the structure of the G-binding site is remarkably conserved (i.e. the root mean square deviation of the sugar-phosphate is below 1 Å Figure 12 C). The crystallographic structures highlight the interaction mode of the ωG in the G-binding site by making hydrogen-bonding interactions to the deep groove of the universally conserved G-C pair of P7 helix which is consistent with the previous observations (Michel et al., 1989). Furthermore they reveal the structural context in which the base-triple interaction (ωG-G=C) takes place which consists in a sandwich of base-triple interactions comprising residues from the P7 helix and the J6/7 junction. Given that ωG is stacked and stabilized by those three other base-triple interactions it also provides a specific recognition of the guanine base, while leaving the sugar moiety of ωG accessible for the catalytic reaction.

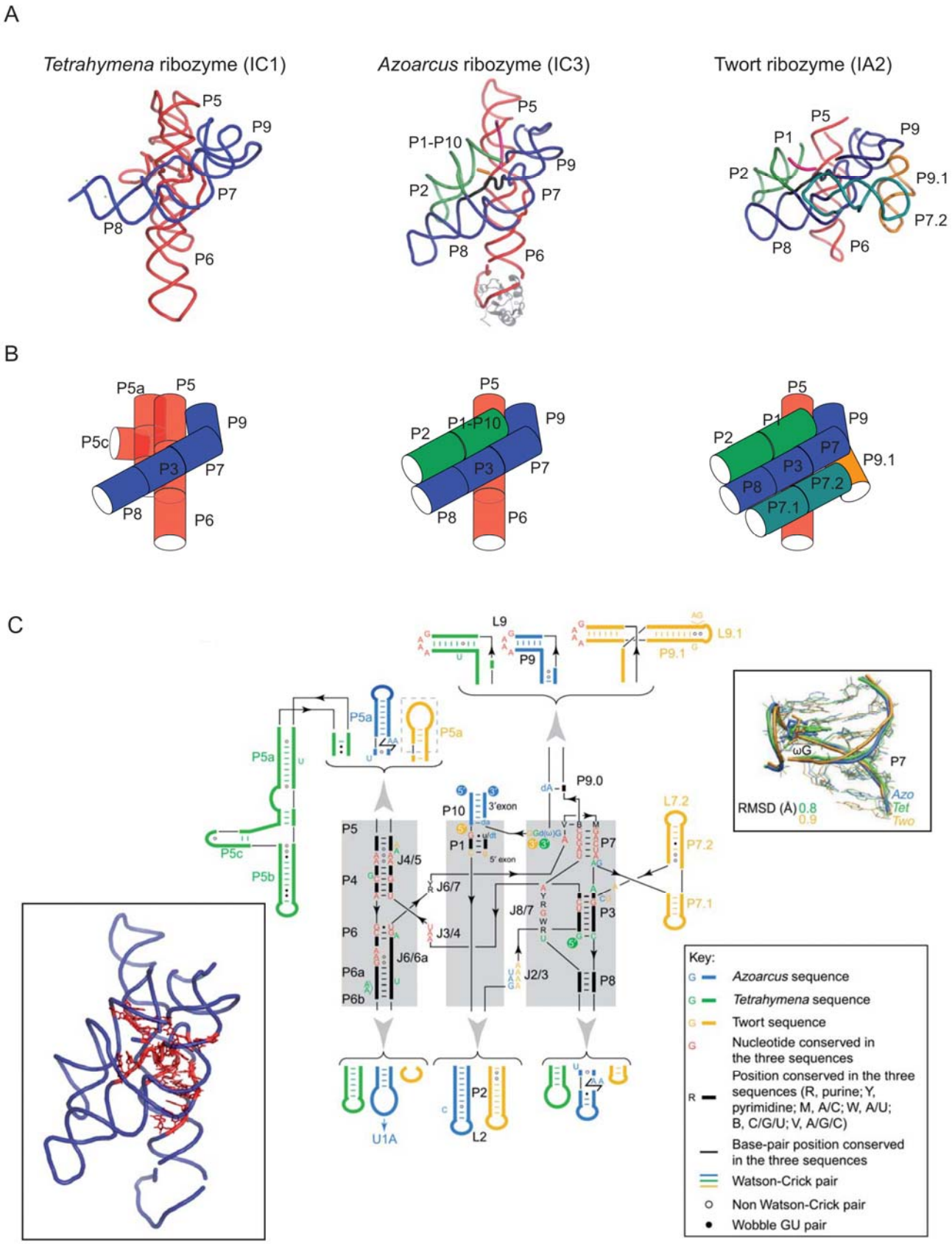


Figure 12
The common catalytic core organization of group I ribozyme.

The common catalytic core organization of group I ribozyme. (A) Crystal structures of the *Tetrahymena*, *Azoarcus* and Twort ribozyme. The three domain organization is highlighted by colouring in red the P4-P6 domain, in blue the P3-P9 domain and in green the P10-P1-P2 domain. (B) Schematic representation of the three group I introns 3D structure. Those schematic representations show the relative positioning of the domains previously shown by the coloured ribbon 3D structure. They also demonstrate the global architecture conservation. (C) Overlay of secondary structure diagrams emphasizing conserved structural elements (black) inside a conserved core (formed by the three shaded boxes that define, from left to right, the P4–P6, P1–P2 and P3–P9 domains). P1–P10 elements, important junctions and loops, and the intron-specific peripheral domains are indicated. Coloured 5' and 3' symbols locate the different ends of the three constructs. The broken box around the P5a element of Twort specifies that this region is disordered in the final model. Lower- and upper-case characters stand for exon and intron sequences, respectively. Nucleotides conserved in the three sequences are shown in red: these residues on the 3D structure of Azoarcus are shown in the inset. In the other inset is shown the superimposition of the P7 and the G-binding site in the *Azoarcus* (blue), *Tetrahymena* (green) and Twort (orange) structure. ωG is shown in balls and sticks. Values in green and orange indicate the root mean square deviations (RMSDs) for the *Tetrahymena/Azoarcus* and Twort/*Azoarcus* comparisons, respectively (Vicens and Cech, 2006).

3.2.1.2. Importance of junctions:

The crystal structures have also highlighted the importance of junctions that form specific structures essential for both folding and substrate recognition. J3/4 and J6/7 were predicted from early mutagenesis studies (Michel et al., 1990) to form base-triples at the P4-P6 domain interface. Crystal structures confirm these predictions and show that these junctions are involved in the formation of base triples and they facilitate contacts between the P4-P6 domain and the P3-P9 domain (Figure 13 A). It also has been revealed, from the *Azoarcus* crystal structure, the importance of the J6/7 in the stabilization and the formation of base-triple interactions with residues of P7. In this structure the backbone of the last two nucleotides from this junction has been found to serve as ligand for an active site metal ion (Adams et al., 2004a). They also provide, in combination with the other features present in P7, a specific recognition of the guanine base (exoG and ωG). This structure explain why the length, but not the sequence of J6/7, is absolutely conserved (Damberger and Gutell, 1994).

Two other critical junctions (J8/7 and J4/5) have been shown notably by these structures to be directly involved in the substrate recognition. The first junction J8/7, highly conserved among the group I introns (7 nucleotides in the IC1 introns and 6 nucleotides in the other introns) with a particular order of conserved purines and pyrimidines, is involved in

bringing together the P3-P7 domain that contains the G-binding site and the P1 helix (Figure 13 B). The second junction J4/5 that forms a tandem of sheared A◦A pairs, is directly involved in the recognition of the G•U pair on the 5'splice site. By making A-minor type interactions (see Annexes Review: Exploring RNA structure by integrative molecular modelling) this junction stabilizes and recognizes the G from the G•U wobble pair. Each junction lines on one side of the domain, providing complementary surface for interdomain packing.

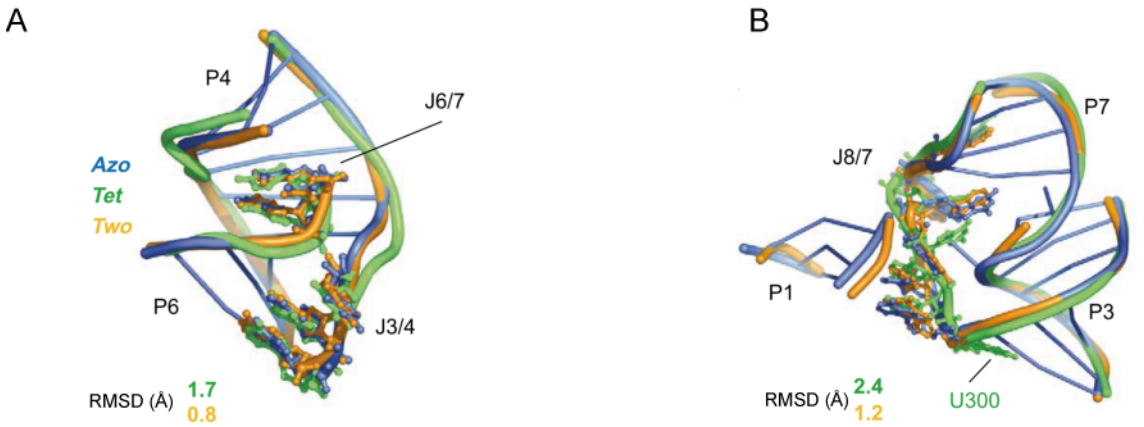


Figure 13
Superimposition of the conserved junctions.
(A) J6/7 junction interacts between the P4 and P6 helices. (B) The J8/7 junction brings together helices P1, P3 and P7. Those two examples illustrate interdomain packing mediated by the junctions. (Azo: *Azoarcus* ribozyme; Tet: *Tetrahymena* ribozyme; Two: Twort ribozyme)

3.2.1.3. Tetraloop docking interaction:

Generally group I intron helical stems are often capped by one of the three classes of tetraloop sequences: CUYG, UNCG and GNRA (where, N = any nucleotide, R = Purine, Y = Pyrimidine). As an example, a survey of 230 selected group I intron from various origin (nuclear, mitochondrial and chloroplastic), in which the domain P9 and P2 are capped by a loop, reveals that P9 is capped at 82.2% by a GNRA loop while P2 is only capped at 47.4% by a GNRA loop (Prathiba and Malathi, 2008). The three crystal structures emphasize the importance of the L9 loop. By making long range interactions with the P4-P6 domain, L9 clamps the P3-P9 and P4-P6 domain together. The GNRA loops known to stabilize hairpin/helix, participate in long range tertiary contacts by interacting within the shallow groove of helices (Figure 14). Phylogenetic, experimental data and crystallographic structures have shown that GNRA tetraloops have their specific receptors (Costa and Michel, 1995;

Costa and Michel, 1997; Ikawa et al., 1999; Ikawa et al., 2001; Jaeger et al., 1994). Thus, the L9 GAAA tetraloop found in the crystal structures interact with its conserved 11-nucleotide internal loop motif (Costa and Michel, 1995).

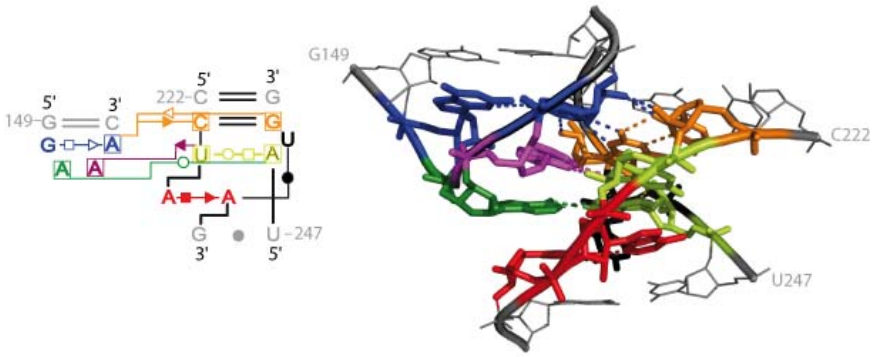


Figure 14
Tetraloop receptor

3.2.1.4. The P3 pseudoknot belt:

An important structural feature that characterises group I introns is the pseudoknot P3/P7. Interestingly, the *Azoarcus* crystal structure shows a novel feature: "a pseudoknot belt". This pseudoknot belt is composed of a single stretch of ~25 nucleotides included and roams through all three intron's domains, wrapping the whole intron in its equatorial region (Adams et al., 2004b) (Figure 15). This particular structure is consistent with the previous kinetic investigations of ribozyme folding that indicated that P3 is the last helix to form (Pan and Woodson, 1998; Rangan et al., 2003; Sclavi et al., 1998; Zarrinkar and Williamson, 1994). In a similar manner in the Twort and *Tetrahymena* other crystal structures, this pseudoknot belt brings P1/P2 into the active cleft where the P1 helix and its 5'-exon could adopt a docked conformation (Adams et al., 2004b). Despite the differences in the set of peripheral domains from each intron, the belt remains on the outside of the molecule.

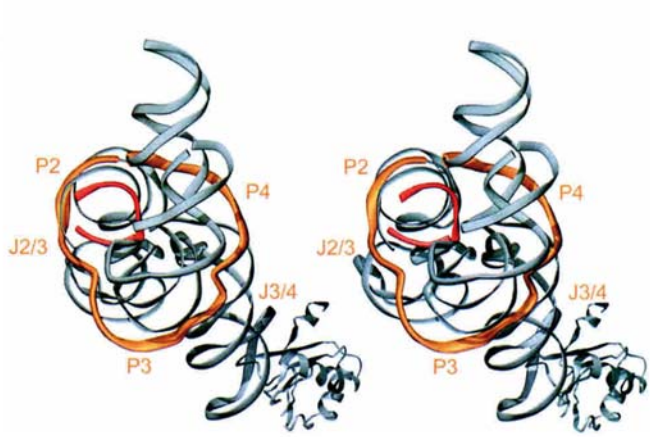


Figure 15
Cross-eyed stereographic view of the P3 pseudoknot belt.
The P3 pseudoknot belt reaches around the circumference of the *Azoarcus* intron's midpoint. The belt is shown in orange and involved ~25 residues (A32 to A59) included within the three domain. The 5' and 3' exon are shown in red. (Adams et al., 2004b).

3.2.1.5. Role of metal ions in the catalytic site:

Comparison with protein enzyme active sites in combination with biochemical experiments and modelling approaches have led several groups to suggest that the presence of metal ions in the active site contributes to the self-splicing reaction mechanism (reviewed in (Hougland et al., 2006)). Early biochemical experiments provided evidence for three metal ions in the catalytic site during the first step of the self-splicing pathway (reviewed in (Hougland et al., 2006)). *In vitro* experiments based on metal ion rescue of modified substrates have led to uncover a first metal ion (M_A) (Piccirilli et al., 1993) and then a second (M_B) (Weinstein et al., 1997). In this model the metal ions M_A and M_B have each a specific role. M_A (bound to 3'OH group of the U of the G•U substrate) serves as nucleophile activator while M_B (bound to the 3'OH group of the exoG) serves as a leaving group stabilizer (Weinstein et al., 1997). Subsequent experiments also based on metal ion rescue of modified substrates have led to uncover a third metal ion (M_C) that contacts the 2'OH group of exoG during the first step of splicing (Shan et al., 1999; Shan et al., 2001; Shan and Herschlag, 1999) (Figure 16 A). Interestingly, it has been also demonstrated that the two metal ions M_B and M_A exchange their role during the second step of splicing with M_B now bound to the 3' oxygen atom of the ωG and M_A still bound to the same uridine (Steitz and Steitz, 1993).

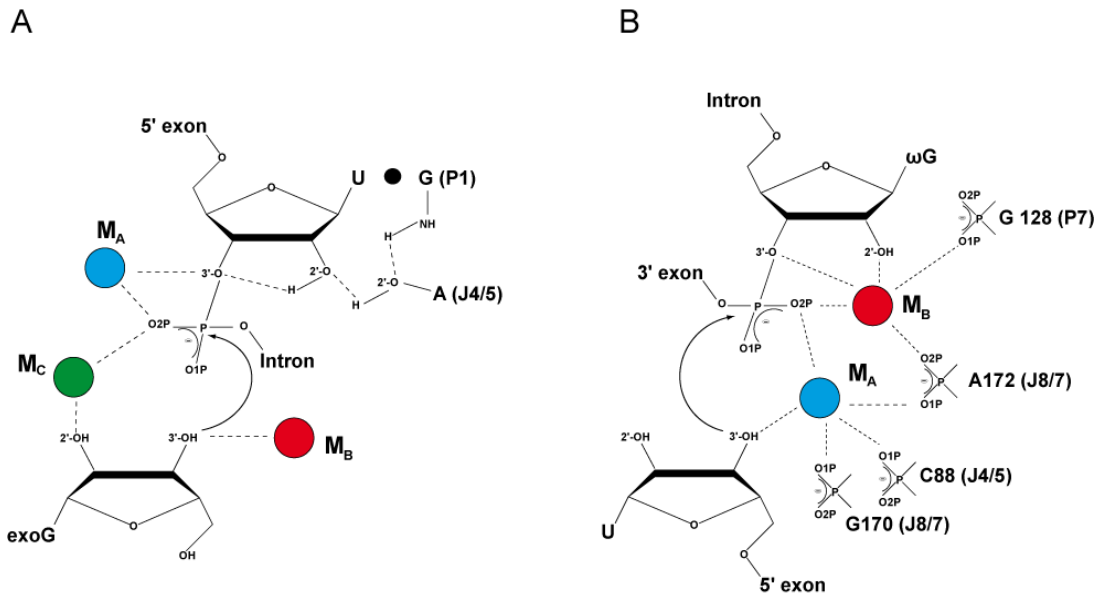


Figure 16
Metal ions at the active site.
(A) The three metal ion model. The magnesium ions in the core are proposed to contribute to the catalytic property of the intron in three ways: 1) By positioning the substrates with respect to each other 2) By deprotonating 3'-oxygen of the attacking nucleophile (ExoG, 5'exon or ωG). 3) By stabilizing a negative charge both in the transition state and on the leaving group. (B) Representation of the metal ion in the active site from the Azoarcus ribozyme crystal structure.

The high resolution crystal structure of the *Azoarcus* ribozyme trapped in the splicing intermediate prior to the exon ligation reaction has enabled to observe metal ions in the active site (Adams et al., 2004a; Adams et al., 2004b). Interestingly, this crystal structure supports a "two metal ion model" previously proposed (see (Steitz and Steitz, 1993)). The crystal structure furthermore demonstrates the binding mode but also the activation and stabilization of the various group involved in binding the two metal ions identified in the active site. Thus, it seems that only two metals ions are required for the catalysis to occur in the second step of splicing, with M_A that acts as a nucleophile activator and M_B as a leaving group stabilizer. The two metal ions coordinate the hydroxyl and phosphate oxygen group from the exon intron respectively as previously predicted by biochemical experiments. Moreover, the metal-metal distance of 3.9 Å is in good agreement with the "two metal ion model" catalytic center (Steitz and Steitz, 1993) (Figure 16 B).

The three different crystal structures show differences in the number and specific contacts of the metal ions bound to the active site. This heterogeneity highlights the inherent difficulty in assigning catalytic metal-ion-binding sites using structural approaches. It is also

possible that those variabilities in the binding of metal ions are the result from the crystallisation condition but also from the state in which the molecules were trapped. Thus, the crystal structures do not rule out the presence of the third metal ion at the active site.

3.3. The peripheral elements:

The classification of group I introns into 13 structural subgroups is based on subgroup-specific peripheral elements (Lehnert et al., 1996; Li and Zhang, 2005; Michel, 1990). Generally these elements are inserted away from the catalytic core in the loop regions (P1, P2, P5, P6, P8 and P9) and in some particular cases in the core (P7 in the Twort ribozyme). Two kinds of peripheral elements exist. The first ones are those longer than 500 nt and contain open reading frames (ORF) encoding endonucleases that promote intron mobility (Haugen et al., 2005a) and should have no impact in the stabilization of the ribozyme core. The second type of peripheral elements (numbered P11-P17) are those that are involved in the stabilization of the ribozyme core by establishing long-range tertiary interactions. Through long-range loop-helix or loop-loop interactions these peripheral elements frame the ribozyme and tie it together (Golden et al., 2005; Lehnert et al., 1996) (Figure 17). Moreover, the progressive deletion of these peripheral elements can result in the ribozyme loss of activity (Cech, 1990) or in a shift from its self-splicing pathway to the circularization pathway (Haugen et al., 2004a). However, some group I introns have lost their peripheral domains but the high G+C content and the highly structured 5' and 3'exons, like in the case of the *Azoarcus* ribozyme, may account for the high activity of the ribozyme even at high temperature (Tanner and Cech, 1996).

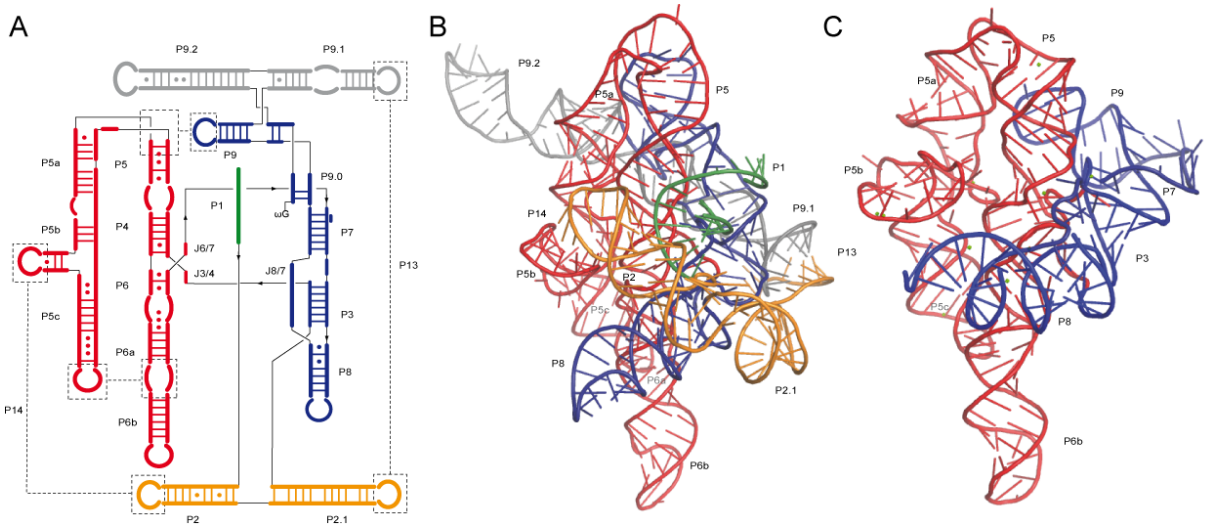


Figure 17

Overview of the peripheral domain in the *Tetrahymena* Tth.L1925 group I ribozyme.
(A) Secondary schematic drawing of the *Tetrahymena* Tth.L1925 group I ribozyme with filled in red the stabilization domain, in blue the catalytic domain in orange the P2-P2.1 peripheral domain, in cyan the P9.1-P9.2 peripheral domain and in green the P1 domain. The tertiary interactions are represented by dashed boxes and dashed lines. Long-range interactions that mainly involve loop-loop interactions are numbered P13 and P14. (B) 3D model of the *Tetrahymena* Tth.L1925 group I ribozyme with its peripheral domains (Lehnert et al., 1996). (C) Crystal structure of the *Tetrahymena* Tth.L1925 (Guo et al., 2004). The comparison between the 3D model and the crystal structure shows how the peripheral elements interact away from the catalytic core in order to frame the ribozyme.

The ability of a group I intron to perform all its catalytic reactions relies on the ribozyme capacity to adopt particular secondary and tertiary structures but also on the folding path of molecule can reach its active conformation. In the next section we will have an overview of the group I intron folding.

4. Group I intron folding:

The Levinthal's paradox exposes that the time needed for a protein to randomly search through all their possible conformations in order to find the one with the lowest free energy is much more longer than the lifetime of the universe (Levinthal C., 1969). This paradox can be applied to the RNA folding issue. In other words, this folding issue can be reframed in: how do RNA molecules to fold into a unique structure without searching through all possible conformations available to them on a physiologically relevant time scale?

4.1. The *in vitro* hierarchical model for RNA and group I introns:

Early work on tRNA and group I introns led to a hierarchical model for the RNA folding. In this model, the regular secondary elements such as helices, loops, bulges and junctions are produced by nearest neighbour interactions which are mainly base stacking and base pairing (Brion and Westhof, 1997). Those local secondary structures were shown to be stable under a wide range of conditions (i.e. pH jump and salt concentration) (Tinoco, Jr. and Bustamante, 1999). Interestingly they were also shown to fold *in vitro* in a time scale of 10-100 µs and to be thermodynamically stable. The subsequent collapse of these secondary structure elements together generates the tertiary structure which is stabilized by interhelical tertiary interaction and specific coordination of metal ion (Treiber and Williamson, 2001; Woodson, 2000a; Woodson, 2000b; Woodson, 2005). In this model the secondary structure forms first quickly and is followed by a slower establishment of the tertiary interactions.

Due to its particular chemical nature, the RNA molecules are highly negatively charged which works against folding into a compact structure. The formation of the tertiary structure induces an increase of the local negative charge density. Thus, the tertiary structure is strongly coupled to the electrostatic environment. Tertiary structure is also more sensitive than the secondary structure to the size, valence and concentrations of counterions (Draper, 2004; Draper et al., 2005; Woodson, 2005). Those ions promote folding by reducing the electrostatic repulsion between RNA phosphate groups. The close relationship between structural hierarchy, stability, and electrostatics have been shown by many folding studies of several group I introns and is illustrated by the equilibrium folding pathway of the *Azoarcus* group I ribozyme dependency to Mg^{2+} (Rangan et al., 2003). Interestingly, the studies of the *Azoarcus* Mg^{2+} dependence folding pathway have led to a model with the presence of two macroscopic transitions (Figure 18) (Rangan et al., 2003). At low ionic strength, the ribozyme adopts some secondary structures (U). The sub-millimolar Mg^{2+} (~0.2 mM) neutralizes the negative phosphate charges and induces the assembly of core helices. It allows compaction of the core and results in appearance of ordered intermediates (I_C). Additional Mg^{2+} (~2 mM) induces formation of the native tertiary structure (N), which also correlates with the fact that this ribozyme shows a burst of reactivity. Interestingly, the folding studies of the *Azoarcus* group I ribozyme have also led to the identification of a stably folded inactive form of the ribozyme (I_F). Furthermore, hydroxyl radical footprinting experiments performed onto those compact intermediates (I_C and I_F) have demonstrated that the interior of the RNA remains

accessible to the solvent (Das et al., 2003; Rangan et al., 2003). This suggests that the tertiary interactions taking place in the I states are dynamic.

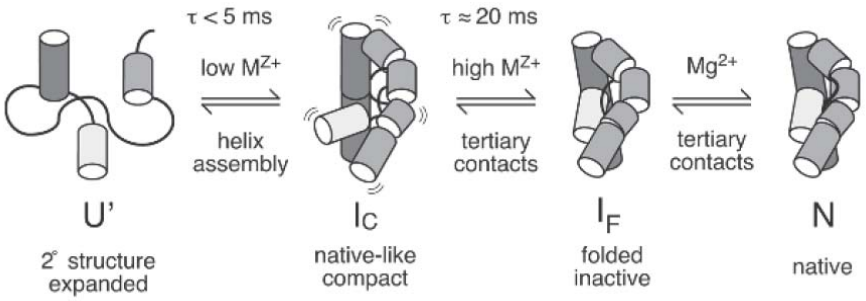


Figure 18
Hierarchical folding of the *Azoarcus* ribozyme.
The Mg2+ dependent folding pathway involves at least two macroscopic transitions (I->N). (U) The unfolded state contains most of the secondary structure (helices, loops and junction). Low Mg2+ concentration (~0.2 mM) induces arrangement and cooperative organization of secondary structural elements to compact intermediates: native-like intermediates (I_C). Those compact intermediates are folded in an open dynamic conformation. In higher Mg2+ concentration the tertiary interactions start to stabilize the ribozyme structure (I_F). In above 2 mM Mg2+ the tertiary structure becomes stably folded and catalytic activity ensues (N). The overall folding time of the *Azoarcus* ribozyme from its secondary structure to its native catalytic tertiary structure has been estimated to <50 ms at 37°C (Rangan et al., 2003).

4.2. Folding intermediates, dynamics and misfolding:

Azoarcus ribozyme folding studies have shown that the ribozyme collapses quickly to native-like intermediates that transform to native structure in <50 ms (Chauhan et al., 2005; Rangan et al., 2003). In contrast, the investigations of the *Tetrahymena* folding pathway have revealed that the ribozyme requires longer time to fold in its native structure in comparison to the *Azoarcus* ribozyme. Early studies based on hydroxyl radical footprinting, base chemical modification and UV crosslinking experiments have shown that the P4-P6 domain can fold faster and at lower Mg^{2+} concentration than the P3-P9 domain (Downs and Cech, 1990b; Zarrinkar and Williamson, 1994). These results are consistent with the previous observation that the P4-P6 domain can fold independently of the rest of the ribozyme (Celander and Cech, 1991; Downs and Cech, 1990b; Murphy and Cech, 1993). By folding independently in less than 1 sec the P4-P6 domain has been proposed to assemble first and then to provide a scaffold for the assembly of the rest of the ribozyme (Zarrinkar and Williamson, 1996) but it is not required for the assembly of the P3-P9 domain. The folding rate of the entire ribozyme

has been then deduced by the use of classical biochemical experiments and has been determined to be above 1 min or longer (about 1-2 sec for the P4-P6 domain and 1 min or longer for the P3-P9 (Downs and Cech, 1990b; Zarrinkar and Williamson, 1994). Interestingly, variations of folding times observed for these two ribozymes and also between the different domains in the *Tetrahymena* ribozyme mainly depend on how closely the intermediates resemble the native structure, and thus how much the initial structures must reorganize before reaching the native conformation. Thereby, the *Azoarcus* ribozyme collapses to native-like intermediates that are very close to the native ribozyme. It is thus interesting to identify the elements involved in the rate-limiting folding step in the *Tetrahymena* ribozyme.

The folding time variation between the P4-P6 and the P3-P9 domain observed in the *Tetrahymena* ribozyme is due to the presence of metastable folding intermediates. Interestingly, those metastable intermediates were first observed on non-denaturing polyacrylamide gels (Emerick and Woodson, 1994; Pan and Woodson, 1998) and slowly disappear with increasing temperatures, Mg^{2+} concentration or action of osmolytes that can stabilize or destabilize both the secondary and tertiary structure (Lambert and Draper, 2007). These observations suggest the simultaneous refolding of many different intermediates (Emerick and Woodson, 1994). Several studies based on hydroxyl radical footprinting, chemical base modifications and site-directed mutagenesis showed that the P3/P7 pseudoknot is replaced by a non-native base pairing (altP3), which stabilizes those populated folding intermediates (Figure 19). This altP3 is formed by non-native base pairing between the 3' strand of P3 and J8/7. Interestingly, peripheral domains (i.e. P14, P13) stabilize altP3, increasing the stability and the lifetime of the misfolded conformer. Furthermore, it has been shown that the *Tetrahymena* ribozyme promotes mispairing of other helices within the pre-rRNA, like for example the P1 helix, to form P(-1) (Woodson and Cech, 1991; Woodson, 1992; Emerick and Woodson, 1993) (Figure 19). P-1 is a co-transcriptionally favored hairpin located in the 5'-exon that prevents the formation of the active intron with P1 harboring the 5'-splice site (Emerick and Woodson, 1993). Remarkably, the formation of P(-1) induces misfolding of the catalytic core and promotes the formation of altP3 (Pan and Woodson, 1998). The formation of P-1 results in decreases of the splicing activity. Thus, altP3 and P(-1) must unfold before the ribozyme can have a chance to refold into its native structure (Pan et al., 1997; Pan and Woodson, 1998).

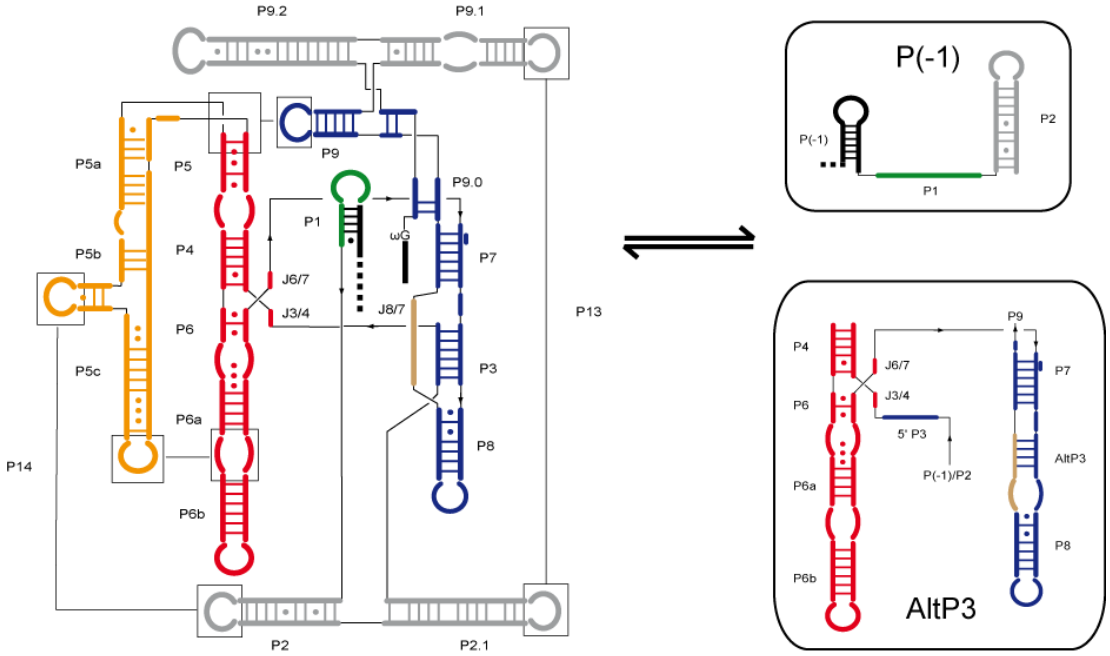


Figure 19
Schematic secondary structure a well characterized form of misfolded *Tetrahymena* ribozyme.

4.3. The flanking sequence context and the co-transcriptional folding:

Group I introns and more generally ribozymes are often embedded in larger RNA molecules. Thus the characterization of the ribozyme's catalytic core starts with the problem of delimiting the functional unit since large RNA molecules are impractical to study *in vitro*. However, flanking sequences or in other words their sequence context were shown to be required for efficient self-splicing activity of the ribozyme (Woodson, 1992). They were also shown to have a significant influence on the folding rates of the ribozyme (Woodson, 1992; Emerick and Woodson, 1993; Cao and Woodson, 1998).

The folding of RNA helices is 2-3 orders of magnitude faster than the rate of transcription (Cruz and Westhof, 2009). This allows base-base recognition to take place as soon as the emerging strand of RNA has reached sufficient length to promote folding. Consequently, the folding occurs during the transcription and is dictated by the sequential nature of RNA synthesis. However, the sequential formation of RNA interactions during the transcription can bias the folding pathway and ultimately determine the functional state of a transcript. Furthermore, the folding pathway can also be greatly affected by the transcription

rate of different RNA polymerases. As an example, the T7 polymerase usually used in *in vitro* experiments elongates at 200-400 nt/sec, versus 10-35 nt/sec for the *E. coli* RNA polymerase (Uptain et al., 1997). Thus, the transcription speed can drastically affect the propensity of group I intron to fold properly or misfold (Heilman-Miller and Woodson, 2003; Jackson et al., 2006). Interestingly, RNA polymerases have been shown to pause during their transcription process. This ability to pause during the transcription plays an important role in folding by avoiding the formation of undesirable structures as the nascent RNA emerges from the RNA polymerase. In the paused complexes, the nascent RNA forms labile structures by temporary sequestrating sequences involved in the formation of non-native helices and that will form the native structure more efficiently later in the transcription process (Wong et al., 2007).

5. Group I intron looking for protein partner:

The group I introns catalytic activity relies on the ribozyme ability to reach the 3D structure corresponding to active conformation. A recent survey on the ability of group I intron to fulfil their self-splicing activity has shown a correlation between a 35% GC content criteria and the self-splicing activity (Vicens et al., 2008). This 35% GC content criteria is generally in good agreement with self-splicing efficiency reported in the literature. However, some group I introns have GC content lower than the 35% GC content criteria. These group I introns have lost their ability to fulfill their self-splicing pathway. They require the presence of co factor protein(s) to carry out splicing. Thus, group I introns can recruit several types of protein cofactors: tRNA synthetases, DNA homing endonucleases, homologs of DNA junction resolvases, homologs of RNA helicases and proteins with no known function (Lambowitz and Caprara, 1999).

5.1. The co-factor protein CYT-18:

Several examples of protein cofactor recruited by group I introns to promote splicing are found in nature. The best studied example is the *Neurospora crassa* mitochondrial tyrosyl-tRNA synthetase (mt TyrRS or CYT-18), encoded by the *cyt-18* gene. *in vitro* experiments have shown that no splicing activity was detected at physiological Mg^{2+} concentration in the absence of the protein but splicing activity was detected when the purified CYT-18 protein

was added (Garriga and Lambowitz, 1986; Wallweber et al., 1997). Interestingly, the introns spliced by the help of the CYT-18 have little sequence similarity, indicating that the protein recognizes conserved structural features of the ribozyme (Guo and Lambowitz, 1992). By using RNA footprinting experiments, the protein binding mode and the ribozyme's region targeted have been identified. The CYT-18 protection sites are mainly clustered in the P4-P6 domain around the junction between P4 and P6. Additional protected sites were found in the region of the P3-P9 domain including P7 and P9 (Caprara et al., 1996a; Caprara et al., 2001) (Figure 20 B). Further experiments have shown that CYT-18 interacts intensively with the stabilization domain and promotes in this way its assembly by helping to establish the correct geometry around the P4-P6 helical junction (Chen et al., 2000). Moreover, CYT-18 contacts in the same time the P3-P9 domain. By forming this interaction between the two domains, the protein helps stabilizing and orienting of those two domains to form the ribozyme's active site (Caprara et al., 1996a; Caprara et al., 1996b). The recently solved crystal structure of the Twort ribozyme co-crystallized with the CYT-18 protein without the C-terminal domain is consistent with previous experimental observations (pdb:2RKJ at 4.5 Å (Paukstelis et al., 2008)). The structure shows that the protein binds along one face of the coaxially stacked P4-P6 helices but it also reveals what are the protein features involved in binding the P4-P6 domain (Figure 20 A and B). The protein particular structure creates several anchoring points that clamp the P4-P6 domain, mimicking P5abc extension of the *Tetrahymena* ribozyme, stabilizing the correct conformation of this domain (Paukstelis et al., 2008) (Figure 20 C and D). Interestingly, CYT-18 also stabilizes the junctions between the stabilization domain and the catalytic domain (J3/4 and J6/7) but also the L9-P5 tetraloop-receptor (Paukstelis et al., 2008) (Figure 20 A and C). In conclusion the structure shows a unique structural adaptation of the protein that is related to a co-evolution of the protein in relation with the introns. By interacting with the folded ribozyme the protein induces stabilization of key tertiary interactions. In fact, the protein core has additional structural adaptations, including basic amino-acid substitutions relative to non-splicing bacterial TyRSs that contribute to group I intron binding (Paukstelis et al., 2005). However, CYT-18 protein is not the only example of proteins that help group I intron splicing.

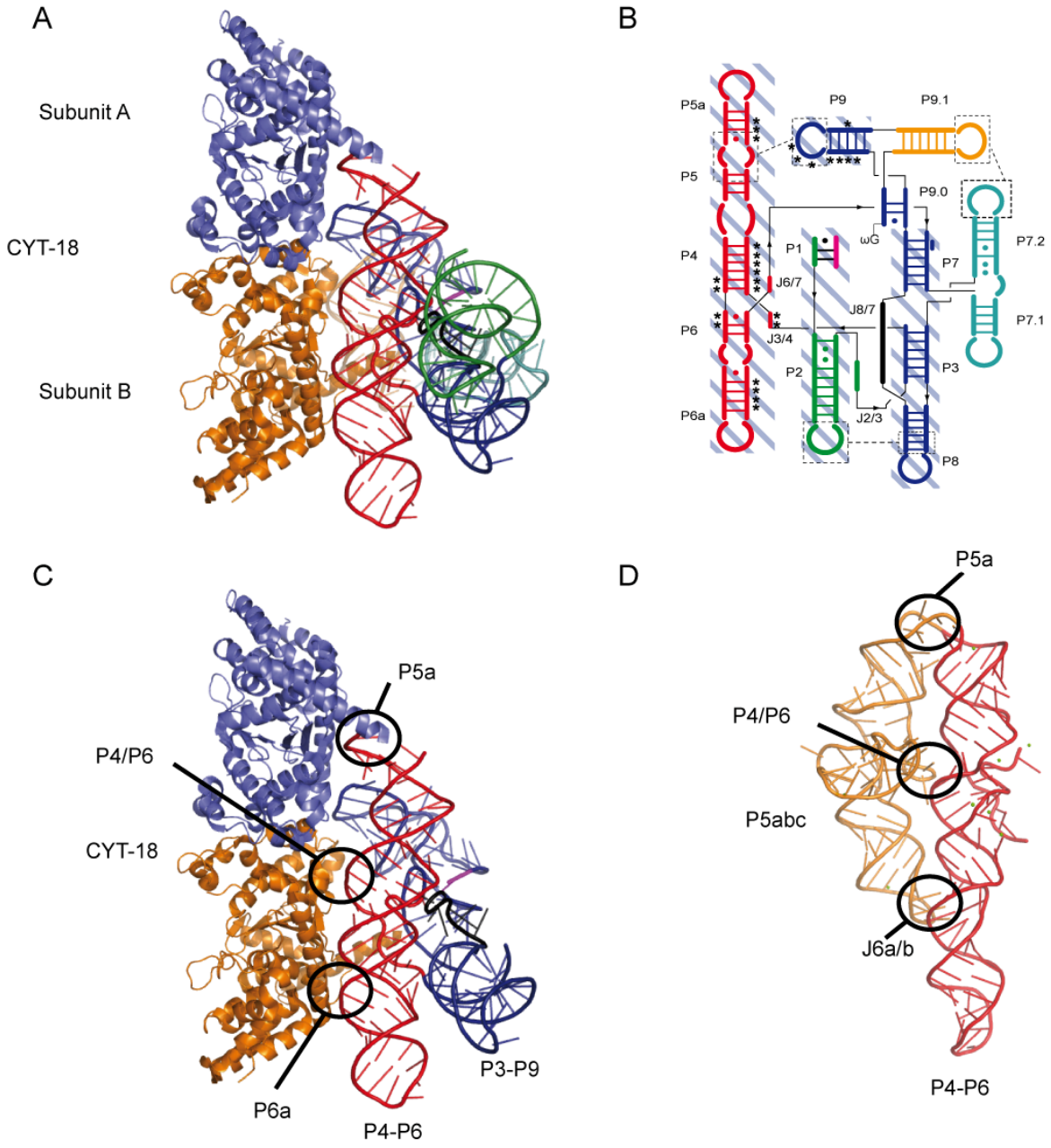


Figure 20
Crystal structure of the Twort orfI142-I2 group I intron bound to CYT-18

Crystal structure of the Twort orfI142-I2 group I intron bound to CYT-18/Δ424-669 and comparison with the P5abc peripheral domain of the *T. thermophila* LSU intron. (A) Ribbon diagram of the CYT-18/Δ424-669 bond to the Twort ribozyme (Paukstelis et al., 2008). The subunits A and B are coloured in blue and orange respectively. The CYT-18/Δ424-669 interacts mainly with the P5-P6 domain of the ribozyme but also with the P9 domain. (B) Schematic secondary structure of the Twort ribozyme. The (*) represents the residues protected by the binding of the protein. The boxed regions correspond to the phosphodiester-backbone positions protected by the full-length CYT-18 protein in *N. crassa* ND1 intron (Caprara et al., 1996b; Caprara et al., 2001). The protections in P2, P3, P8, which are not seen in the crystal structure, are attributable to the C-terminal domain of the CYT-18 protein, wich is absent from the structure. (C) Orthogonal ribbon diagrams of CYT-18/Δ424-669 bound to Twort P4-P6 (in red) and P3-P9 domain (in blue). (D) The corresponding view of the *T. thermophila* LSU intron crystal structure (Guo et al., 2004), with P5abc extension in orange.

Previous work showed that CYT-18 could replace P5abc peripheral domain and then promotes the splicing at physiological Mg^{2+} concentration (Mohr et al., 1994). This comparison between the two system shows that both the CYT-18 and the P5abc bind the length of the P4-P6 domain with contacts at the P5, the P4-P6 junction and the distal region of P6 (P6a for CYT-18 and J6a/b for for P5abc), enabling them to stabilize the blackbone conformation on both sides of the P4-P6 junction.

5.2. Other protein co-factors:

Other proteins have been found to assist group I introns splicing (Table 2). Unlike CYT-18, these protein splicing cofactors mainly work in concert with mitochondrial mt intron-encoded maturase proteins. For example, the yeast mt Leucyl-tRNA synthetase (mt LeuRS or Nam2), encoded by the nuclear gene Nam2, together with the intron-encoded maturase has been shown to assist splicing of the bI4 and aI4α group I introns of the mtDNA (Rho and Martinis, 2000). The bI4 intron forms a ternary complex by binding each of its protein splicing partners (Rho and Martinis, 2000).

The yeast bI3 group I intron is also an instructive example of an RNA that has become dependent on proteins to fold and function. This intron requires the binding of 6 proteins: two dimers of the Mrs1 protein (Bassi et al., 2002; Bassi and Weeks, 2003) and the binding of its bI3 maturase. Each protein seems to have a special role. The Mrs1 proteins that mainly bind to the tetraloop receptor (Duncan and Weeks, 2010), have been shown to induce large conformation rearrangements in both the secondary structure and tertiary structures (Duncan and Weeks, 2008). The bI3 maturase that binds to the P4-P6 domain (Duncan and Weeks, 2010), has been shown to promote long-rang tertiary interactions (Duncan and Weeks, 2008).

In attempting to reach catalytic activity, the aI5β group I intron is in fact the one that wins the gold medal regarding the diversity of required splicing factors. This group I intron requires at least six different proteins, including Mrs1, Pet54, Mss116, Mss18 and Suv3 (Turk and Caprara, 2010). However, the unusual number of required cofactors is not yet clear. Interestingly, studies of those group I intron protein cofactor complexes help deduce evolutionary models in which a functional ribozyme became dependant on proteins, coevolved with them and recruited multiple and maybe new proteins to maintain, enhance or regulate its splicing activity.

Protein	Protein type	Group I intron	Intron cellular localization	Protein Intron binding site	Proteins bind per intron
TyrRS (CYT-18)	tRNA synthetase	bI3	mitochondria	P4-P6	2
LeuRS (Nam2)	tRNA synthetase	bI4 and aI4α		P4-P6	2
Cbp2	Unknown	bI5		5' end of the ribozyme	1 + maturase
Pet54	Translational activator	aI5β		5' end of the ribozyme	1 + maturase
Mss18	RNA binding protein	aI5β		5' end of the ribozyme	1 + maturase
Mrs1	DNA junction resolvases homolog	aI5β		Tetraloop receptor	4 + maturase
Suv3	RNA Helicase	aI5β		Unknown	1+Mrs1, Pet54, Mss116, Mss18
Intron-encoded maturase	DNA homing endonucleases	various	various	various	various

Table 2
Non-exhaustive table of the co-factor proteins with their introns.

5.3. The role of proteins in *in vivo* folding of group I introns:

It has been proposed that RNA chaperones were involved in assisting RNA in achieving a unique native conformation by resolving non-native conformation. The *Tetrahymena* ribozyme has been shown to self-splice 10-50 times more efficiently *in vivo* than *in vitro* (Brehm and Cech, 1983; Zhang et al., 1995). This is mainly due to the fact that important features of the intracellular environment cannot reliably be reproduced *in vitro*. These are: the full flanking sequence context, vectorial folding during the transcription, the RNA polymerase rate and pause, the ion homeostasis and finally presence of protein-assisted folding.

There are two classes of proteins that are not necessarily mutually exclusive (Herschlag, 1995; Lorsch, 2002; Schroeder et al., 2004) and help manage RNA folding: specific RNA binding proteins and RNA chaperones. The proteins that bind to a specific RNA, recognize a defined 3D structure, thereby stabilizing it through tight binding as previously presented with the CYT-18 study case. The second class of proteins is the RNA chaperones that can resolve the non-native conformations without recognizing specific structures or sequences. They rescue RNAs that are trapped in unproductive folding states by

helping the RNA to get out of the kinetic traps that are known to sprinkle the RNA folding pathway.

For example, StpA was first isolated as a suppressor of a splicing-deficient group I intron mutant of the phage T4 *td* gene in an *in vitro* assay. This protein is homologous to the nucleoid-associated protein possessing the ability to bind bent DNA and is thought to be a global transcriptional regulator (Zhang et al., 1996). The influence of StpA was monitored by *in vivo* chemical structure probing (Waldsich et al., 2002a; Waldsich et al., 2002b). The experiments reveal that StpA enhances the accessibility of the bases participating in tertiary-structure elements, which indicates that it loosens the intron structure (Waldsich et al., 2002a). In this way, StpA has a mechanism of action opposite to CYT-18 that favors the compactness of the tertiary structure. Thus, StpA resolves the tertiary fold of the intron, giving the molecule another chance to fold into the native splicing-competent structure (Waldsich et al., 2002a). Other examples of group I intron RNA-chaperones were found. They highlight the close relation between those two classes of protein. CYT-18 that binds to the P4-P6 domain, does not complete folding of the ribozyme, but further recruits the RNA helicase CYT-19, a dead-box protein, which works together with CYT-18 protein to promote the group I intron splicing (Mohr et al., 2002; Lorsch, 2002). More recently, the CYT-19 protein was shown to unfold both the native and the misfolded forms of the *Tetrahymena* group I intron with a particular preference for less stable structures lacking tertiary interactions allowing the RNAs to explore a wider range of conformations (Bhaskaran and Russell, 2007).

6. Summary of group I intron structure and function:

Group I introns have been found embedded within essential genes and in various organisms that range from bacteriophages to plants and fungi. Most of them have been shown to retain self-splicing activity. However, this self-spicing ability sometimes relies on the presence of protein co-factors (i.e: maturase or splicing protein co-factors) that either help the ribozyme to reaches its active conformation or stabilize the ribozyme in its catalytic conformation. Regardless of dependency the maturase or protein co-factor to perform their splicing reaction, the chemical mechanism is the same. The self-splicing pathway has the potential to render the presence of group I intron genetically neutral to the host. On the opposite, group I introns have been also shown to promote the formation of full-length circle

introns. The formation of FLC results in un-ligated exons and non-functional gene products. Thus, group I introns may have gained a regulatory role that can impair gene expression.

In the circularization and the self-splicing pathways, the catalytic ability of the group I introns relies on the folding into a complex 2D and 3D structure. Thus, group I introns are also good model systems for folding studies, in part because they contain a well-folded and phylogenetically conserved core.

Finally, Group I introns are not only catalytic RNAs, but also mobile genetic elements. The success of these introns as mobile elements relies on their self-splicing and circularization abilities which enable them to propagate by inserting into host genes. It also relies on the presence of intron-encoded homing endonuclease gene. However, group I introns and associated homing endonuclease genes appear to undergo an evolutionary cycle of gain and loss currently known as the Goddard-Burt cyclical model (Goddard and Burt, 1999b) (Figure 21). Interestingly, some introns may escape from the cycle (Figure 21) and thus it may result in the acquisition of new biological functions, far beyond intron removal and intron mobility, selected under given evolutionary pressure (Nielsen and Johansen, 2009). The growing amount of deep-sequencing data may greatly influence the discovery of new complex group I system/organization. The diversity of the group I intron organizations can be illustrated by the exhaustive list of examples presented in the next section.

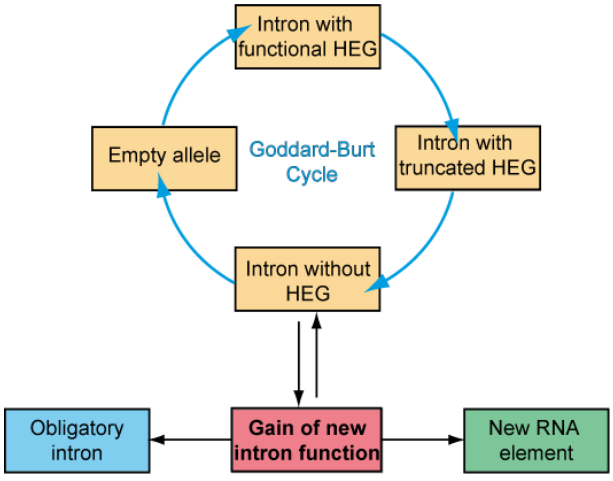


Figure 21
The Nielsen-Johansen group I intron "escape" routes from Goddard-Burt cycle (Nielsen and Johansen, 2009).
Introns may escape from the Goddard-Burt cycle and subsequently gain a new function by either becoming an obligatory host-essential intron or becoming a new RNA element with functions different from classical self-splicing.

CHAPTER II: THE TWIN-RIBOZYME INTRON

In this part, I describe briefly the twin-ribozyme discovery, the organization and all the elements that define this complex group I intron organization. I briefly describe the GIR2 ribozyme that belongs to a classical group I intron. Then, I focus on the GIR1 ribozyme from the myxomycete *Didymium iridis*, describing its function and its secondary structure. This part ends with a presentation of the organism where this singularity is found and the processing pathways of the twin-ribozyme intron in its biological context.

1. Discovery, distribution and structural organization of the twin-ribozyme intron:

1.1. Discovery and distribution of the twin-ribozyme intron:

Numerous group I introns were discovered and characterized from studies of nuclear rDNA genes in fungi. Interestingly, insertion sites of group I introns that interrupt both the SSU and the LSU, were shown to be phylogenetically conserved (Cannone et al., 2002; Jackson et al., 2002). Furthermore, group I introns remain an interesting model to understand the mechanisms and the pathways of horizontal gene transfer due to their widespread distribution among fungi rDNA genes. The screening of fungi rDNA gene with mapping of group I intron insertion sites in combination with group I intron phylogenetic characterization has led to surprising discoveries. The SSU rRNA of the myxomycete *Didymium iridis* was found to harbour one of the most complex group I intron organization (Johansen and Vogt, 1994). This new group I intron category harbours a classical group I intron in a peripheral domain of which are embedded both a ribozyme and an open reading frame encoding a homing endonuclease. Further studies led to the definition of a twintron subgroup (or twin-ribozyme intron). Finally this new subgroup was shown to be present in several species of *Naegleria* amoeba flagellates (Johansen and Vogt, 1994; Einvik et al., 1997; Einvik et al., 1998a; Johansen et al., 2002; Wikmark et al., 2006) and more recently in *Heterolobosea sp* (EMBL/GenBank DQ388519 unpublished results).

All natural variants of twin-ribozyme introns known today (Johansen et al., 2002; Wikmark et al., 2006), are inserted into conserved regions of the nuclear SSU rRNA host

gene. However, they present some differences in distribution and inheritance. The twin-ribozyme introns from *Naegleria* species were found to be all inserted at position 516 according to rDNA group I introns nomenclature system based on the *E. coli* numbering system (Johansen and Haugen, 2001) (Figure 22 B): the intron is coded on 3 letters corresponding to the organism name followed by capital letter for the small and large ribosome subunit and the insertion position according to the *E. coli* numbering. The Nae.S516 is restricted to the *Naegleria* genus with a widespread but sporadic distribution that includes 21 insertions among 70 strains analysed (Wikmark et al., 2006). In comparison with Nae.S516, the twin-ribozyme intron found in the myxomycete *Didymium iridis* was found to be inserted at position 916 and hence named Dir.S916. Strikingly, the Dir.S916 intron seems to be restricted to the *Didymium iridis* Panama 2 isolate.

Analysis of the twin-ribozyme intron distribution among those species have highlighted different inheritance patterns of the intron. The *Naegleria* intron was gained by the *Naegleria* phylum and has been strictly vertically inherited by strains that harbour it. However, due to evolutionary pressure, the intron was subject to mutations and subsequently to deletion that might explain the loss of the twin-ribozyme intron from approximately 70% of *Naegleria* isolates (Goddard and Burt, 1999b). In *D. iridis*, the process of intron inheritance is different from *Naegleria*. In the *Didymium* phylum, the twin-ribozyme intron is unique to the *Didymium iridis* Panama 2 isolate. This argues in favour of a recent gain by horizontal transfer. Interestingly, Dir.S956 has been shown recently to be a mobile element both at the RNA level through a reverse splicing event (Birgisdottir and Johansen, 2005) and at DNA level due to the presence of intron-encoded homing endonuclease (Johansen et al., 1997b). These phylogenetic differences add to structural differences between the twin-introns from the two phyla that could indicate a distinct origin.

1.2. Global structural organization of the twin-ribozyme intron:

The twin-ribozyme intron consists of an unusual ribozyme (GIR1) and a homing endonuclease gene (HEG) both embedded in a peripheral domain of a self-splicing group I ribozyme called GIR2 (subgroup IE) (Johansen and Vogt, 1994) (Figure 22). In the case of Dir.S956, GIR1 is inserted in domain P2. GIR1 is directly followed by an open reading frame (ORF) encoding the homing endonuclease I-*Dir*I, which is a member of the His-Cys box

family (Johansen et al., 1993). Moreover, a small spliceosomal intron of 51nt, situated in the ORF, seems to be a particular feature of I-*Dir*I HEG. A comparison between Dir.S956 and Nae.S516 reveals a general conservation of twin-ribozyme intron organization. However, some local differences can be pinpointed. In Nae.S516, GIR1 is inserted in domain P6 of GIR2 which belongs to subgroup IC1 (Figure 22).

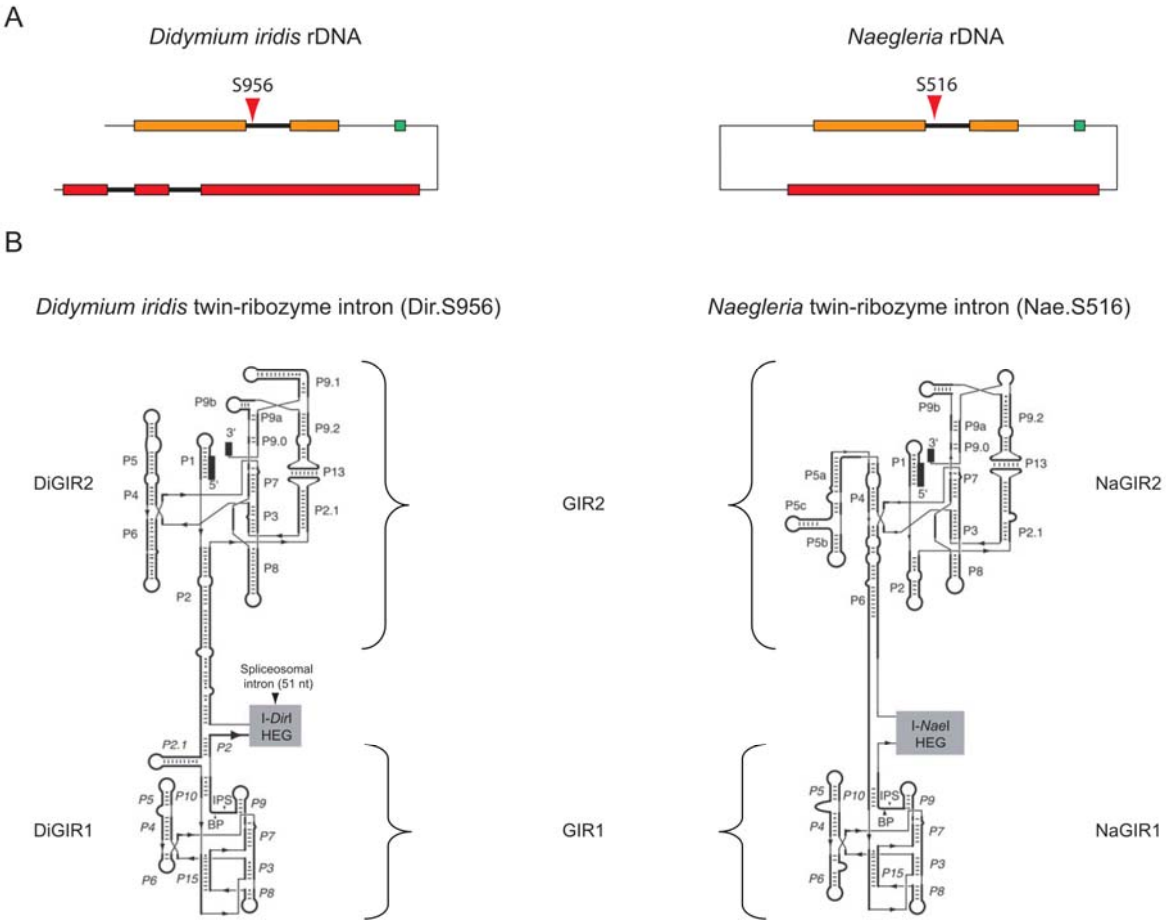


Figure 22
DNA and RNA organization of two twin-ribozyme introns.
(A) The Didymium iridis twin-ribozyme intron (Dir.S956). (B) The Naegleria twin-ribozyme intron (Nae.S516). In both cases, the GIR2 ribozyme is a splicing ribozyme involved in processing of the pre-rRNA. GIR1 is involved in processing and release of HE mRNA.

2. <u>The DiGIR2 ribozyme from the Dir.S956 twin-ribozyme intron:</u>

The DiGIR2 ribozyme from Dir.S956 belongs to the group IE intron subgroup (Figure 23). This subgroup is characterized by specific peripheral elements. P2 and P9 domains harbour characteristic IE subgroup extensions: the P2.1 domain; an elongated P9.0; P9.1 and P9.2 peripheral elements. Interestingly, long-range tertiary interactions between L2.1 and L9.2 that create P13, were also shown to take place and are exclusively conserved among IE

subgroup (Li and Zhang, 2005). Finally, sequence alignments of the ribozyme's core with both presence of GG in J3/4 and C residue in bulge next to the universal conserved G-C base pair in P7 have allowed to identify the DiGIR2 ribozyme as a ribozyme belonging to the group IE intron subgroup without ambiguousness (Suh et al., 1999).

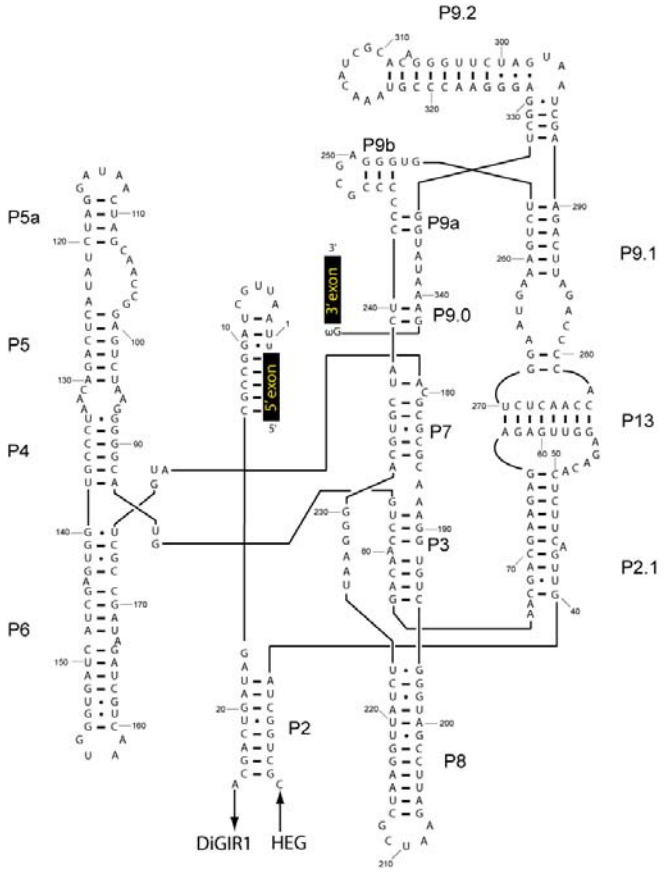


Figure 23
Secondary structure of the DiGIR2 group I ribozyme.

DiGIR2 carries out self-splicing *in vitro* (Johansen and Vogt, 1994; Decatur et al., 1995) without the help of any protein co-factor. In addition to self-splicing, DiGIR2 also produces circular RNAs. During *in vitro* processing of Dir.S956, different catalytic reactions catalyzed by DiGIR2 were shown to promote accumulation of intermediates, resulting in the circularization pathway (Johansen and Vogt, 1994; Decatur et al., 1995; Nielsen et al., 2003; Haugen et al., 2004a). Moreover, formation of full-length circles and circularization pathway products were detectable both by northern blot analysis (Vader et al., 1999; Vader et al., 2002) and QRT-PCR (**Paper IV**). Finally, reverse splicing of Dir.S956 has also been investigated. It reveals that it only depends on the DiGIR2 splicing ribozyme (Birgisdottir and Johansen, 2005). Thus, all these observations show that DiGIR2 is a functional group I

ribozyme despite the presence of both another ribozyme and an open reading frame that are embedded in its P2l domain.

3. The group-I-like ribozyme: GIR1

3.1. The DiGIR1 reactions, the good, the bad, the ugly:

In vitro studies concluded first that the DiGIR1 cleavage reaction was independent from added GTP. Secondly three different reactions were characterized (Figure 24 A). The natural reaction is the branching reaction (Figure 24 A-1) at the Branching Point (BP) (The good). This reaction leads to the formation of a tiny lariat cap where the first and the third nucleotides are joined by a 2',5' phosphodiester bond (Nielsen et al., 2005). The DiGIR1 branching reaction is initiated by a nucleophilic attack involving the 2'OH of U232 (Nielsen et al., 2005) Figure 24 B). The chemistry of the reaction was demonstrated by deoxy-substitutions in the substrate next to the catalytic core at positions corresponding to C230, A231, U232, and C233 (Figure 24 B). Two deoxy-substitutions had no impact on the branching reaction rates (i.e. C230 and C233). The other two deoxy-substitutions were shown to abolish or decrease the branching cleavage rate. First the U232 deoxy-substitution was found to completely inhibit the branching reaction. This implies that the 2'OH of the U232 is required for the branching reaction. Secondly, a strong effect of the deoxy-substitution at A231 was detected. It was then speculated that the A231 may have a structural role in the catalytic core (**see Paper I and review 1**).

The other reactions observed *in vitro* are parasitic (Figure 24 A). The reaction opposite to branching, referred as the ligation reaction is very efficient (Figure 24 A-2) (The ugly). This ligation reaction can result in the complete masking of the branching reaction in DiGIR1 length variants longer than 166 nucleotides upstream of the IPS (Nielsen et al., 2005; Nielsen et al., 2009) (see paragraph 3.2.2). Finally, DiGIR1 catalyses hydrolytic cleavage at the IPS site (Figure 24 A-3) (The bad). This reaction is observed with the full length intron and several length variants (Johansen and Vogt, 1994; Decatur et al., 1995; Einvik et al., 2000; Nielsen et al., 2009). This reaction is irreversible and is considered as an *in vitro* artefact resulting from a failure in the correct folding of the catalytic core.

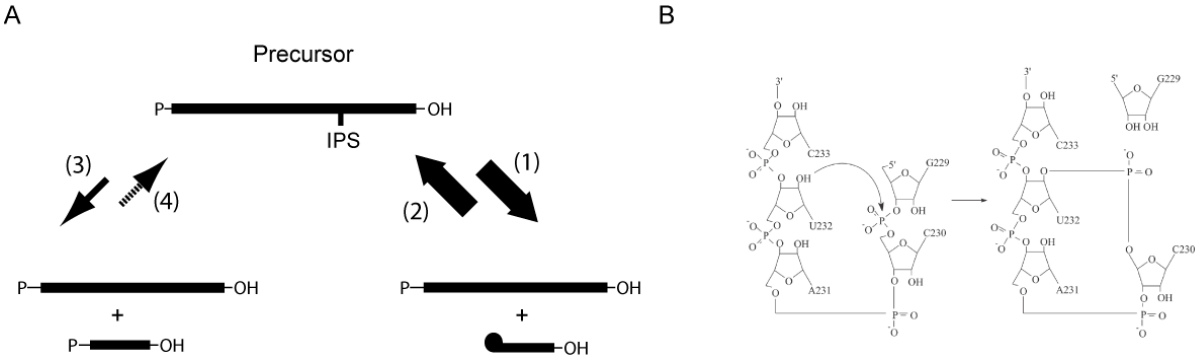


Figure 24

Reaction catalyzed by the DiGIR1 ribozyme and scheme of the branching reaction.
(A) Reaction catalyzed by the GIR1 branching cleavage ribozyme. The main activity (1) is branching. However, the branching reaction is highly reversible and can even be masked by the ligation reaction (2). A hydrolytic cleavage reaction (3) is less pronounced and only observed *in vitro*. (4) Hypothetical reaction that has not been observed. (B) Detail of the branching reaction catalyzed by the GIR1 ribozyme. U232 makes a nucleophilic attack at the IPS site. This results in the formation of a tiny lariat cap where the first and the third nucleotide are joined by a 2',5' phosphodiester bond (Nielsen et al., 2005).

Experimental studies have revealed that the three reactions can be separated. In this way, the branching reaction has been isolated from the reverse reaction by adding an osmolyte (urea) that (1) inhibits the ligation reaction and (2) decreases the contribution from the hydrolytic cleavage. Among GIR1s, the branching cleavage rates vary considerably (see chapter V supplementary results). Interestingly in DiGIR1, the flanking 5' and 3' sequences were shown to affect the rate of cleavage by branching (Nielsen et al., 2005; Nielsen et al., 2009; Nielsen et al., 2008) (see paragraph 3.2.2).

3.2. Structural organization of the GIR1 ribozyme:

3.2.1. Secondary structure of the DiGIR1 ribozyme:

Comparative sequence analysis of DiGIR1 and NaGIR1, together with DiGIR1 structure probing and molecular modelling have revealed the structural organization of the DiGIR1 ribozyme (Einvik et al., 1998c) (Figure 25). The secondary structure was found to be closely related to the group I intron structure. The secondary structure can be divided, like in the classical group I intron, into three different helical domains: P10-P15 (substrate domain), P4-P6 (stabilization domain) and P3-P9 (catalytic domain). Moreover, parallel structure probing of *Naegleria andersoni* (NaeGIR1) ribozyme have corroborated the overall global base-pairing scheme (Jabri et al., 1997). Despite the fact that the secondary structure was

found to be closely related to group I intron secondary structure, GIR1 presents some specific structural characteristics that define it as a distinct class of self-cleaving ribozyme (the group-I-like ribozyme). In this way, most notable features are: (1) the lack of P1; (2) the lack of the J8/7 junction; (3) the presence of a novel P3/P15 pseudoknot; (4) the presence of the three-way junction between P3, P8 and P15; (5) the presence of two new junctions J15/7 and J9/10 and finally (6) the unusual structure of J4/5 (Figure 25 part highlighted in yellow). Although the secondary structures of GIR1 and group I introns are closely related, the differences should impact the topology and resulting 3D structure. Moreover, the specificities of the GIR1 secondary structure may be responsible for the branching reaction. Thus, by taking advantage of recent group I intron crystal structures in combination with the recent finding of the GIR1 branching activity of this ribozyme, the 3D structure of the DiGIR1 ribozyme has been re-modelled and gave new insight into both structure requirements for the branching reaction and evolution of group I like ribozymes (**see Paper I and review 1**).

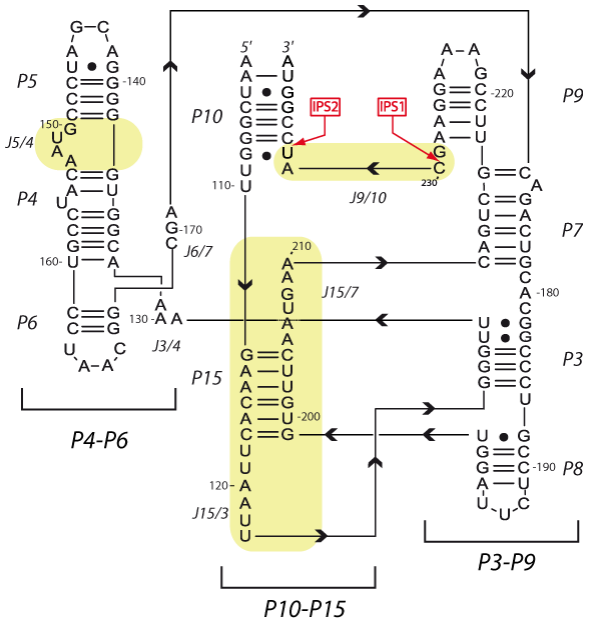


Figure 25
Secondary structure diagram of the DiGIR1 ribozyme

Secondary structure diagram of the DiGIR1 ribozyme according to (Einvik et al., 1998c). The regions highlighted in yellow are the 2D structure parts that differ from the classical group I intron secondary structure.

3.2.2. Role of the flanking sequences/peripheral domains:

From a biochemical point of view, characterization of a ribozyme often starts by delimitating the functional unit. In the case of twin-ribozyme introns, it comes down to define which RNA elements enter in the structure of DiGIR2, DiGIR1 and finally of the HEG messenger. However, according to the twin-ribozyme intron secondary structure (Figure 22), the way to isolate DiGIR1 from the other two components DiGIR2 and HEG seems straightforward. In this way and in order to define the minimal version of DiGIR1, linker regions between DiGIR1, DiGIR2 and HEG were shortened down by deletion. Different DiGIR1 length variants were constructed and named according to the number of nucleotides upstream and downstream the IPS. As an example, DiGIR1 157.22 length variant represents 157 nt located upstream and 22 nt downstream the IPS, respectively (Figure 26 A). Strikingly, some length variants were shown to fully reflect the full-length DiGIR1 cassette (i.e. 166.65 or 162.65) (Einvik et al., 2000). Some other length variants were shown to fully reflect the natural cleavage branching reaction (i.e. 157.22) while other length variants were shown to have a strong ligation activity masking the branching reaction under standard conditions (i.e 166.22) (Nielsen et al., 2005; Nielsen et al., 2009).

Two structural elements outside the GIR1 catalytic core were identified as being essential for branching (Figure 26 A). They form the so-called P2P2.1 peripheral domain. Interestingly, the P2.1 hairpin was shown to be critical for ribozyme catalysis. Mutations that destabilize P2.1 hairpin as well as hairpin shortening strongly affect catalytic activity while mutations in the L2.1 loop do not (Einvik et al., 2000). These observations indicate that the hairpin structure and length are important rather than individual bases in the loop. In the same line of evidence, the P2 has also been shown to be required for the ribozyme catalytic activity. Destabilization or shortening of P2 domain completely inhibits the ribozyme (i.e. shortest length variant for the branching activity: 157.22) (Einvik et al., 2000; Nielsen et al., 2005) (Figure 26 B).

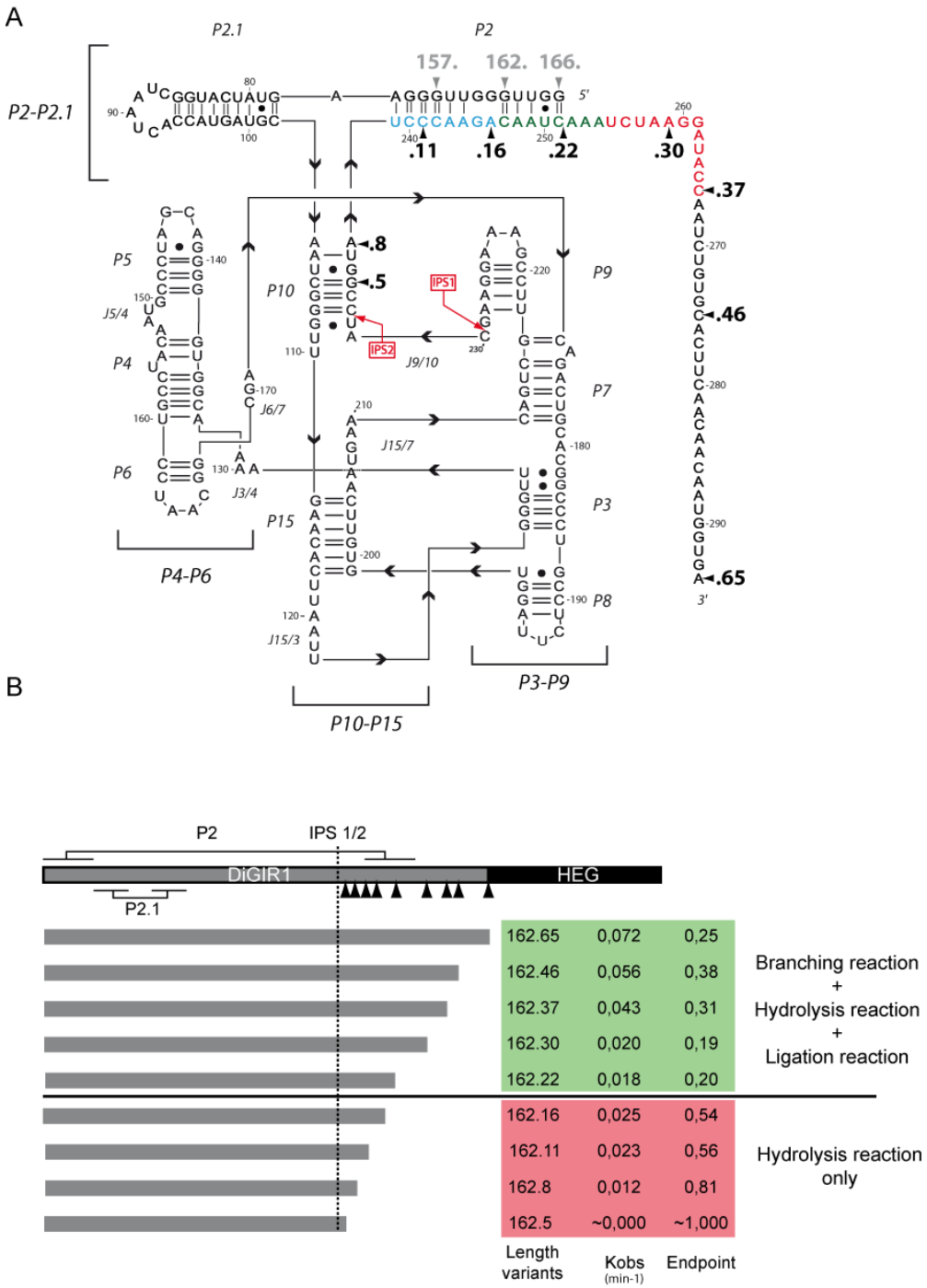


Figure 26
Secondary structure with the P2P2.1 domain.

Secondary structure with the P2P2.1 domain and schematic representation of various length variants. (A) Secondary structure of DiGIR1 according to (Einvik et al., 1998c), harbouring P2P2.1 peripheral domain. Different grey and black arrows represent 5' and 3' end length variants, respectively. (B) Linear representation of DiGIR1 with the paired region P2 and P2.1. The different 3'end length variants created during the deletion studies are also represented with their name.

DiGIR1 P2P2.1 peripheral domain has been shown to be necessary for ribozyme activity. I*n vitro* and *in vivo* studies of DiGIR1post-cleavage product have highlighted the

presence of a particular stem loop structure named HEG P1 located in the HE mRNA 5' UTR (Vader et al., 1999; Einvik et al., 2000). Furthermore, HEG P1 was recently shown to be involved in release mechanism of the HE mRNA from the ribozyme core (Nielsen et al., 2008; Nielsen et al., 2009) (**see Paper II**: interaction between L9 and an 11 nt receptor in HEG P1). Based on the observation that HEG P1 seems to be a stable stem loop structure, *in silico* predictions were carried out. Interestingly, it has been found that an alternative structure, containing the HEG P1 stem loop, can be formed when the 3' strand of P2 is longer than 46 nt (Figure 27). Whether this alternative structure can form *in vitro* or *in vivo* needs to be addressed as well as the role of this alternative structure.

In this way, the P2P2.1 peripheral domain and HEG P1 stem loop structure were recently demonstrated to play an important role in the ribozyme activity regulation (Figure 27) (**see Paper III**). It has been proposed that P2P2.1 folding activates the ribozyme, whereas the HEG P1 stem loop promotes destabilization of the catalytic core of the ribozyme by sequestering the 3' strands of both the P10 and P2 domains, (Figure 27) (**see Paper III**). In other words, the formation of HEG P1 turns "off" the ribozyme. Thus, the P2P2.1 domain, that is the DiGIR1 regulatory domain, has the potential to act as a conformational switch turning the ribozyme activity "on" and "off" (**see Paper III**).

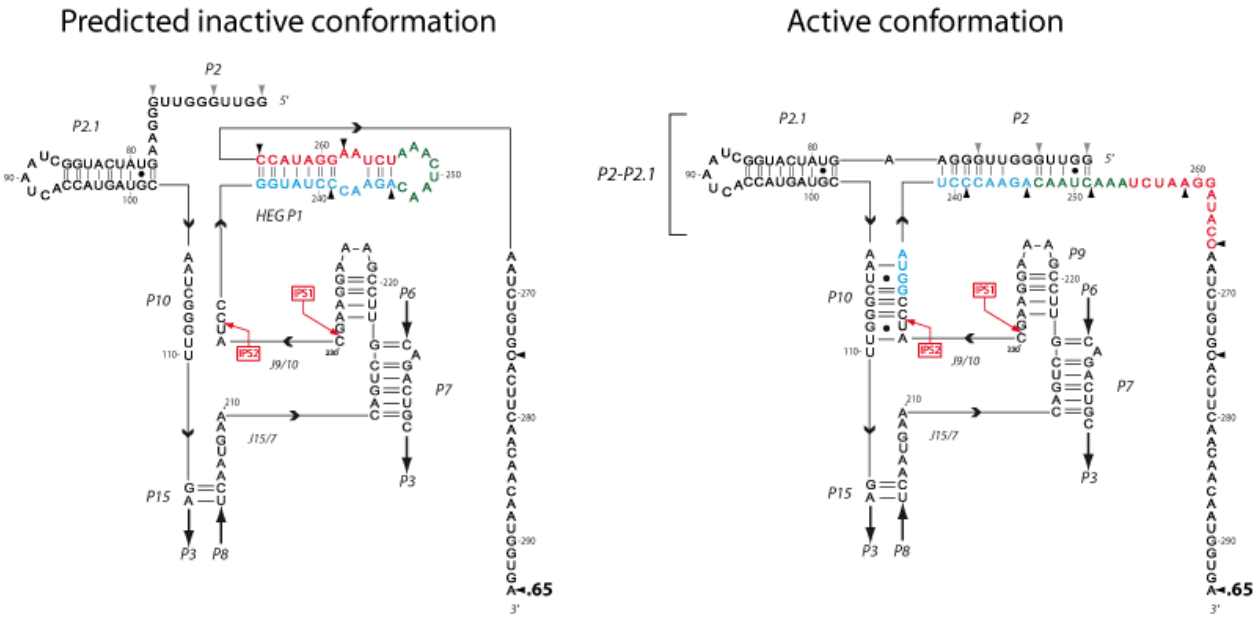


Figure 27
Representation of alternative conformations of the ribozyme's 3'end.

3.2.3. Similarities and differences between DiGIR1 and NaGIR1:

GIR1s were previously found in 21 among 70 *Naegleria* species according to (Wikmark et al., 2006). They present strong conservation of secondary structural features but also some differences between themselves and also in comparison with the DiGIR1 ribozyme (Figure 28) (Wikmark et al., 2006). In a recent survey, NaGIR1s were shown to promote branching at different rates *in vitro* (see Chapter V: Tang Y. unpublished result: Screening and characterization of GIR1 ribozymes from *Naegleria* genus). However, some structural features are universally conserved among GIR1 ribozymes (Figure 28 A). The unique catalytic core organization containing the P15 pseudoknot in combination with the three-way junction between the P3/P15 pseudoknot and P8, is uniformly conserved even at the sequence level (Figure 28 A). Interestingly, J3/4 or J9/10 are conserved at the nucleotide level. Finally, the sequence of P10 domain which consists of 6 bp including the G•U base pair at the IPS2 site, is uniformly conserved.

The comparison between secondary structures of various NaGIR1s and DiGIR1 also reveals some prominent differences within the ribozyme core. Those differences span all three domains. The L9 GAAA tetraloop that was shown to be important in the DiGIR1 release mechanism (**see Paper II**) is not present in NaGIR1s. They instead harbour a 7 to 11 nt loop (Figure 28 B). Stem P6 is at least 4 bp longer in NaGIR1s than in DiGIR1. More strikingly, the junction between P4 and P5 segments is a very variable part in both sequence and structure (Figure 28 B). In DiGIR1, the J5/4 four nucleotides junction was shown to play a key role in the recognition and stabilization of the P10 G•U base pair (**see Paper I**). In NaGIR1s, this junction is replaced by either an internal loop or insertion of helical segments. This can be illustrated by NclGIR1 that harbours an 11 nt single stranded J5/4 whereas some others NaGIR1s harbour a variable 3-6 bp or P5a stem insertion (Nca/Nit/Ngr/NloGIR1) (Figure 28 B). Finally, domains outside the catalytic core that were shown to be required for DiGIR1 branching activity and regulation are also different in NaGIR1s. DiGIR1 P2P2.1 domain is absent in the NaGIR1s. It has been replaced by a different P2 domain and a variable internal loop J2/10 (Figure 28). However, the P2 structure is poorly supported by comparative sequence analysis (Chapter V unpublished results). The J2/10 is an apparently unstructured variable internal loop as deduced from structure probing data (Jabri et al., 1997; Johansen et al., 2002).

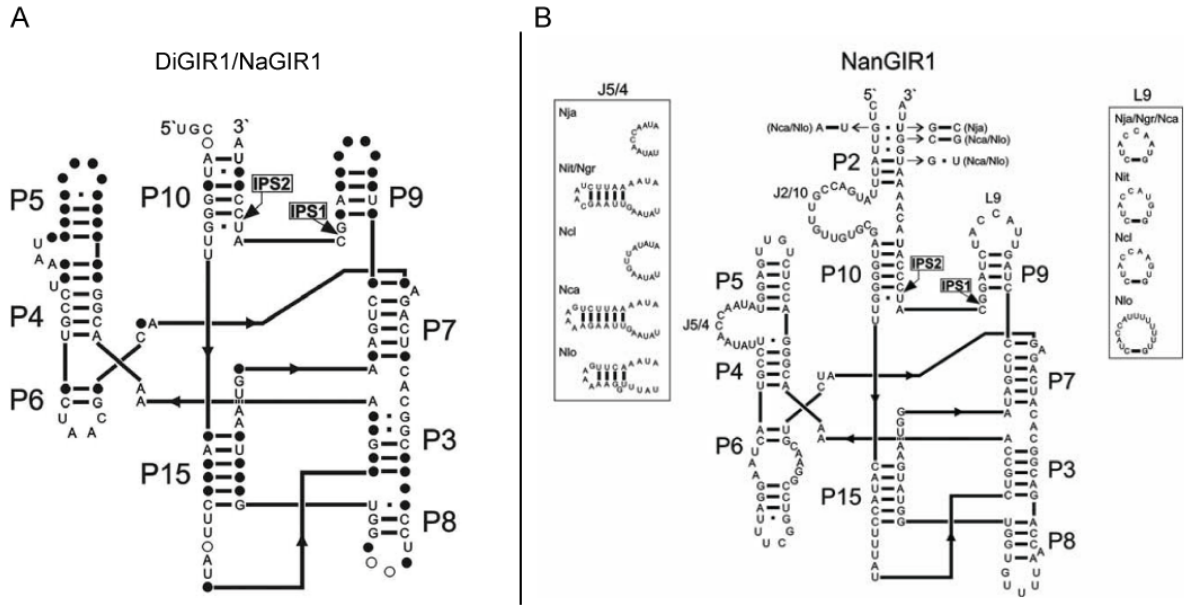


Figure 28
Secondary structure diagrams of the Naegleria GIR1 and similarities between DiGIR1 and NaGIR1.
(A) DiGIR1 and NanGIR1s consensus sequences. Nucleotide positions that are identical in pairwise comparisons are shown. Filled circles: non identical position; open circles, identical positions in some *Naegleria* sequences (from (Johansen et al., 2002)). (B) Secondary structure diagram of *Naegleria andersoni* GIR1. The secondary structure is drawn according to DiGIR1 secondary structure presented previously (Figure 25). Compensatory changes proposed in P2 among the various species are noted. The observed sequence variations in both the L9 loop and J5/4 junction are also represented in the box. (Nan: *N. andersoni*; Nja: *N. jamiesoni*; Nit: *N. italica*; Ngr: *N. gruberi*; Ncl: *N. Clarki*; Nca: *N. Carteri*; Nlo: *N. lovaniensis*) (from (Johansen et al., 2002)).

DiGIR1 has been intensively studied. Results obtained by mutation experiments and structure probing have revealed the ribozyme secondary structure. The recent understanding of the natural branching reaction in combination with previous structure probing and mutation experiments have led to suggest in of a new molecular model of the DiGIR1 ribozyme that supports the branching reaction (**see Paper I**). Flanking sequences were also shown to play a key role in both the ribozyme ability to perform its reaction (Einvik et al., 2000; Nielsen et al., 2005) and in the release of the product away from the core (Nielsen et al., 2009) (**see Paper II**). More recently, we have proposed that DiGIR1 P2P2.1 domain acts as a regulatory domain; adopting two mutual exclusive structures turning the ribozyme activity “on” and “off” (**see Paper III**). Thus, the knowledge accumulated on DiGIR1 allows us to better understand not only the structural requirements for the branching reaction but also its control. However, as presented previously, NaGIR1s show prominent differences with DiGIR1. Thus, these observations give rise to several questions about the NaGIR1s folding, the structure of the catalytic core, the role of peripheral domains and finally, the regulation. Given the fact

that first, P2.1 is not found in *Naegleria,* second, that domain P2 is slightly different from the counterpart in DiGIR1 P2 domain and third, that there is no such alternative structure predicted to disturb the catalytic core, then how is NaGIR1 branching activity regulated? In the same way, since the length of flanking sequences and sequence context (i.e.: formation of the P2 domain) were shown to be important in the DiGIR1 branching reaction, thereby is there a parallel that can be drawn with the NaGIR1? Finally, what is the impact of the sequence/domain insertions (i.e. J5/4) on NaGIR1 structure? (**see Chapter V**)

3.3. A complex rRNA processing pathway in the myxomycete *Didymium iridis*:

As mentioned previously, twin-ribozyme introns are found in the nuclear SSU rRNA gene of the myxomycete *Didymium iridis* and several *Naegleria* strains. Due to the presence of a cleavage ribozyme in the middle of a group I ribozyme, the rRNA processing is thereby more complex than the rRNA processing in the case of classical group I introns. Before starting describing the various points that characterize the rRNA processing in the *D. iridis,* let me digress for a moment by introducing the organism in which the twin-ribozyme intron is found.

3.3.1. The myxomycete *Didymium iridis* life cycle:

The myxomycete *Didymium iridis* is a phagotroph, living on the forest floor and feeding mainly on soil bacteria. *D. iridis* has, like other slime moulds, a complicated life cycle which mainly depends on environmental conditions. The life cycle can be roughly divided into a microscopic haploid stage and a macroscopic diploid stage (Figure 29). The diploid stage is initiated by sexual fusion of two haploid cells. The fusion generates a giant cell, the plasmodium. Interestingly the plasmodium grows into a large cell (diameter range is about one cm or more) which contains more than 10^8 synchronouzed dividing diploid nuclei in a common cytoplasm. This plasmodium then differentiates and undergoes meiosis to create a large number of haploid spores. The spores are then released and spread by the wind. If the spores land in a favourable environment they will germinate and produce subsequent haploid flagellate or amoeba cells. Then the haploid cells populate their new habitat by vegetative reproduction (Einvik et al., 1998a).

During the vegetative haploid state the myxomycetes can differentiate in three haploid forms that can be easily recognized: amoeba, flagellate and microcyst (Figure 29). Interestingly all three forms are interchangeable and mainly depend on the environmental condition. As an example, if the environment is aqueous the slime mould will carry flagella in order to "swim" but if the surroundings are drier it will become amoeba. However it happens that the environmental conditions become unsuitable to sustain basic amoeba/flagellate forms. Then both forms can differentiate into a hard-shelled highly resistant microcyst waiting for favourable conditions to excyst (Figure 29) (Einvik et al., 1998a).

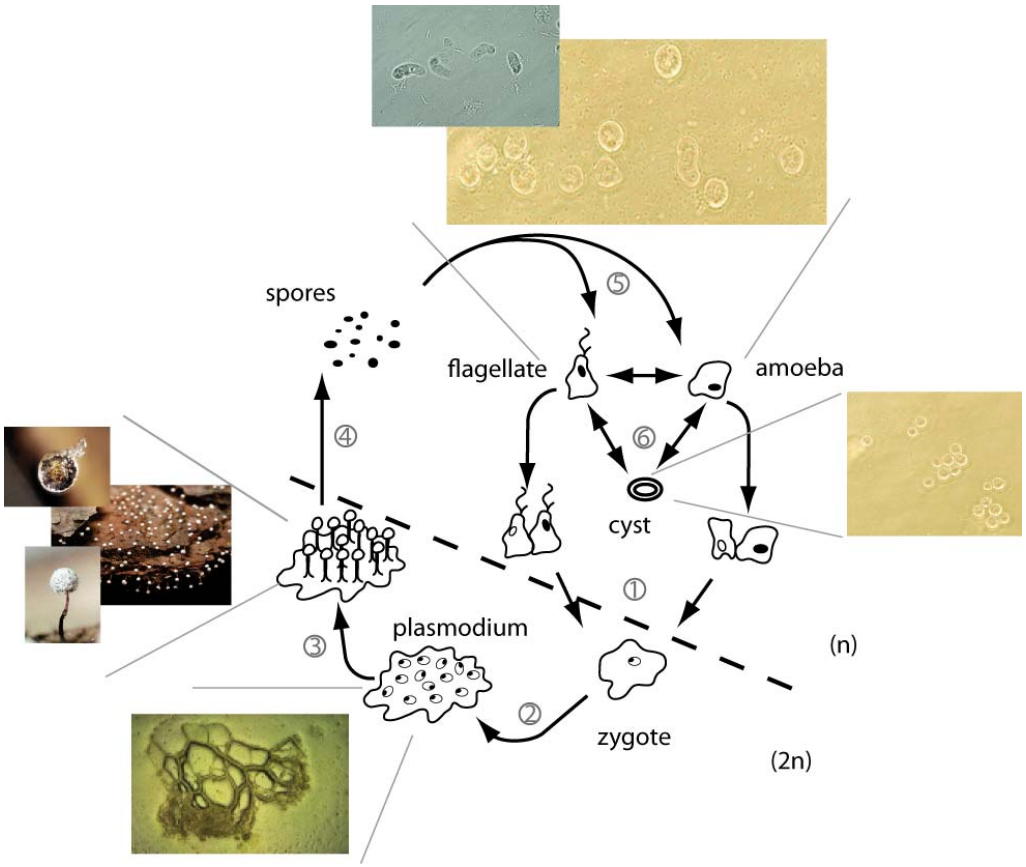


Figure 29
Life cycle of the myxomycete Didymium iridis

The myxomycete life cycle alternates between a haploid macroscopic stage (top part) and a diploid macroscopic stage (bottom part). Step 1: sexual fusion of amoeba or flagellates carrying different mating-type alleles (black and white nucleus). Step 2: development of the zygote into a multinucleate plasmodium. Step 3: irreversible development into sporangia. Step 4: production of haploid spores by meiosis. Spores are released and from them new cells will germinate. Step 5: from the germination of spores, new haploid cells grow as flagellates, amoebae. Depending on environmental growing conditions, flagellate or amoebae forms can differentiate into resting microcyst cells waiting for right environmental conditions to differentiate into the other forms (step 6).

3.3.2. rRNA processing pathway in the biological context:

From a combination of *in vitro* studies and *in vivo* observations, three different processing pathways of the Dir.S956 intron were mapped according to cellular conditions (i.e.: only in the haploid life stage of the myxomycete previously exposed).

The first pathway relies on DiGIR2 self-splicing activity (Figure 30). During exponential cell growth (i.e.: flagellate/amoeba forms), DiGIR2 catalyses its own excision from the primary transcript, resulting in ligation of exons and free standing intronic part. Subsequently, DiGIR1 cleaves and releases the HE mRNA (Vader et al., 1999) with a lariat cap in place of the conventional m^7G cap. The HE mRNA 3'end is then formed by cleavage at a site referred to as IPS3. Following that, the mRNA is polyadenylated at a polyA signal. The short spliceosomal intron (I51) is then spliced out and the mature mRNA is exported to the cytoplasm (Vader et al., 1999). It is worth to note that DiGIR1 requires completion of the DiGIR2 self-splicing pathway before the branching reaction can occur. Thus, the branching activity of the DiGIR1 ribozyme is regulated *in vivo*.

The second pathway results in the formation of full-length circle (FLC) introns and un-ligated exons (Nielsen et al., 2003). This pathway relies only on DiGIR2 ability to produce FLC as previously presented in Chapter I and Chapter II. In this pathway, DiGIR1 is not active and the circle re-opening seems to be required for its activation. Interestingly, this circularization pathway is responsive to cellular conditions (**see Paper IV**).

Finally, the third pathway is induced by starvation-induced encystment (Vader et al., 2002). During encystment, the pre-rRNA is processed into a 7.5 kb RNA product that accumulates within the cell. Interestingly, the processing is accomplished by DiGIR1 without prior DiGIR2 activity (Vader et al., 2002). This pathway results in rRNA production regulation by down-regulating the rRNA expression. Although the biological function of the 7.5 kb RNA is unknown, it has been speculated that it can be stored as a precursor that will allow fast expression of the HE when conditions favourable for rRNA expression will be restored (Nielsen et al., 2008).

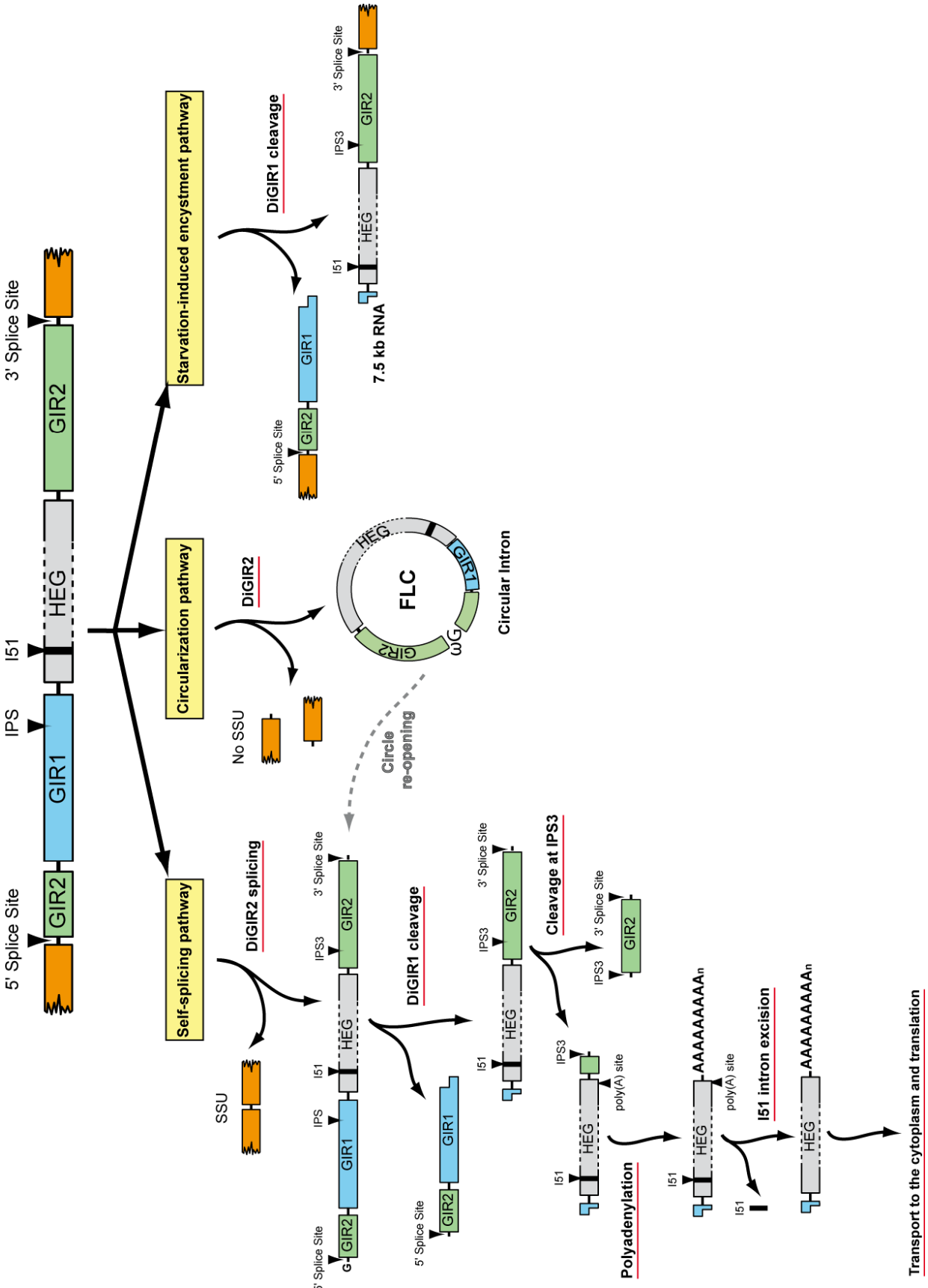


Figure 30
The three different processing pathways of the twin-ribozyme intron.

CHAPTER III: SUMMARY OF ARTICLES

<u>Paper I:</u> B. Beckert, H. Nielsen, C. Einvik, S.D. Johansen, E. Westhof, B. Masquida (2008). "Molecular modelling of the GIR1 branching ribozyme gives new insight into evolution of structurally related ribozymes". EMBO J.

In this study, we have rationalized the detailed structural and functional analysis of the DiGIR1 ribozyme into a new three-dimensional model. As previously presented, the Dir.S916 twin-ribozyme intron from the myxomycete *D. iridis* is composed of a branching ribozyme (DiGIR1) followed by a homing endonuclease (HE) encoding sequence. Both of them are embedded in a peripheral domain of a group I splicing ribozyme (DiGIR2). DiGIR1, by a unique catalytic reaction, catalyzes the formation of a tiny lariat with a 3 nt loop, which caps the HE mRNA. Interestingly DiGIR1, even with its unusual small size mainly due to the lack of large peripheral extension often found in group I ribozymes, is structurally related to group I ribozymes and more specifically to the IC3 eubacterial subgroup. Structure probing and phylogenetic studies have revealed the lack of a P1 segment playing the role of the substrate and conserved in all other known group I ribozymes. However, a P10 base paired segment as well as a novel pseudoknot P3/P15 were proposed to take place in the core region. Thus, these particular observations raise the question of why GIR1 carries out a branching reaction despite its resemblance with group I ribozyme.

In order to understand the structural requirements for the branching activity of DiGIR1, the ribozyme core was modelled. This was done by taking into account (1) structural probing data in combination with new mutagenesis data and (2) the recent crystal structure of the *Azoarcus* group I ribozyme. The three-dimensional model presented in this article highlights the similarities between DiGIR1 and the *Azoarcus* group I splicing ribozyme. In this way, the highly compact DiGIR1 structure resembles the core of the group I splicing-ribozyme. The P4-P6 and P3-P9 domain are oriented as two juxtaposed, elongated, coaxially stacked helixes. This helical arrangement is roughly similar to that of the *Azoarcus* structure. Interestingly, the P10 base-paired region and the extended P15 suggested pseudoknot were proposed to be positioned in the cleft like it is the case in group I introns with P1/P2.

The comparison between DiGIR1 and the *Azoarcus* ribozyme also reveals some differences that span the DiGIR1 ribozyme core model in all three domains. The first of them is the extended P10 substrate domain containing a G•U pair distinct from the nucleophilic residue that docks onto the catalytic core in classical group I ribozyme. It has been proposed that the group I intron conserved J8/7 junction was reduced in DiGIR1 to 3 nt due to its fusion with nucleotides of the previous P2 3'strand, thus forming the characteristic P15 helix. The new junction J15/7 composed from remaining J8/7 nucleotides is complemented by residues from the new J9/10 junction that docks in the catalytic core forming a pre-lariat fold. Fourth, the J4/5 conserved from the classical group I intron has been replaced by the J5/4 junction which recognized the substrate domain. Finally, J15/3 organizes the three-way junction between P15, P3 and P8 resulting in a side-by-side parallel orientation of P3-P9 to P10-P15 domain, which mimics the relative orientation of P1/P2 against the P3-P9 domain found in the *Azoarcu*s ribozyme. These differences in the catalytic core organization result in acquisition of a new reaction mechanism that is the branching activity of DiGIR1, rather than sequential hydrolytic cleavage as it was previously suggested.

Review: H. Nielsen, B. Beckert, B. Masquida and S. D. Johansen (2008). "The GIR1 branching ribozyme". In Ribozymes and RNA catalysis, Lilley DMJ and Eckstein F, eds. (London: The Royal Society of Chemistry), pp. 229-252.

In this book chapter, we describe the GIR1 ribozyme, focussing on GIR1 from the myxomycete D. iridis. The different features that characterize this ribozyme are described in detail from the biochemical to the structural point of view.

Paper II: Á. B. Birgisdottir, H. Nielsen, B. Beckert, B. Masquida, S. D. Johansen (2010). "Intermolecular interaction between a branching ribozyme and associated homing endonuclease mRNA". Submitted to Biological Chemistry.

The DiGIR1 ribozyme promotes the maturation of its associated homing endonuclease mRNA by a unique branching reaction that leads to the formation of a tiny 3-nt lariat cap. Upon its release, the 5' end of the mRNA has been shown to form an alternative stem-loop structure termed HEG P1. In this study, we have focused on the release mechanism of the mRNA messenger from DiGIR1 after the branching reaction. Interestingly, the release mechanism involves an intermolecular interaction between the L9 GAAA tetraloop of the

DiGIR1 ribozyme and a GNRA tetraloop receptor-like motif found in the HEG P1 stem-loop structure.

In order to better understand and characterize the intermolecular interaction between the L9 tetraloop and its receptor-like motif within the HEG P1 stem-loop structure, bimolecular gel-shift assays based on a composite *Tetrahymena* ribozyme have been performed. The *in vitro* association between L9 and the HEG-P1 receptor-like motif has been also characterized by secondary structure probing in concert with molecular modelling.

As a result, a new 11 motif receptor (UCUAAG-CAAGA) has been characterized and represents a new and specific GAAA tetraloop-receptor in RNA-RNA interaction. Interestingly, this interaction taking place between the cleaved mRNA product and the ribozyme has revealed its biological role which seems to promote the post-cleavage release of the lariat-capped mRNA. Thus, the receptor in HEG P1 adapts to the position of P9 and the L9 loop pulls the lariat out of the catalytic pocket and thereby contributes to the release of the mRNA. This finding adds to our general understanding of how protein-coding genes embedded in group I introns can be expressed and also controlled by ribozymes.

Paper III: B. Beckert, M. Marquardt Hedegaard, B. Masquida, H. Nielsen (2010). ”Identification of an on/off switch in the DiGIR1 ribozyme” (manuscript)

Group I ribozymes by their self-splicing ability are involved in internal reorganization of RNA molecules, by catalyzing their own removal from transcript precursors and promoting the ligation of the two exons. For these ribozymes and due to their particular localization in rRNA, timing of folding and thus catalysis is of particular importance because it affects dramatically the function of the molecule within which they reside. Interestingly, group I ribozymes have been found to rapidly fold in their active conformation. Thus, DiGIR1 is in this way particularly interesting mainly due to (1) its characteristic cleavage branching activity; (2) its particular location within a peripheral domain of the DiGIR2 ribozyme that is inserted in the rRNA SSU and (3) because its structure has been shown to be closely related to group I introns. Consequently, harbouring this cleavage branching ribozyme may have some consequences on rRNA processing. Thus, the formation of SSU rRNA depends on the ability to keep GIR1 inactive until splicing and exon ligation has taken place. Therefore, the ribozyme has to fold initially into an inactive conformation to avoid untimely cleavage of the

ribosomal RNA precursor. It has been proposed that the peripheral domain of DiGIR1 (the P2P2.1 domain) may act as a regulatory domain. In the inactive conformation, part of the regulatory domain has been hypothesized to adopt a stem-loop structure (HEG P1) that prevents the formation of a stable active site. Conversely, in the active conformation, HEG P1 is replaced by helices (P2 and P10) that may allow the organization of a three-way junction (P2-P2.1-P10).

In order to investigate the inactive and active conformations, as well as the folding of the DiGIR1 ribozyme, a combination of non denaturing polyacrylamide gel electrophoresis of various DiGIR1 length variants (WT and mutant as well) and twin-intron ribozyme chemical and enzymatic probing have been used. Finally, the conformation of the P2 P2.1 P10 three-way junction conferring activity to DiGIR1 has been deduced by combining biochemical data with structure modelling and theoretical accessibility calculations from the previously modelled DiGIR1 ribozyme core.

As a result, native gel analysis has revealed that folding of the DiGIR1 ribozyme was intrinsically linked to both the Mg^{2+} concentration and the 3' length end variant extension. Interestingly, the global picture araising from native gel assays is that folding of the DiGIR1 ribozyme *in vitro* is a partitioning between two alternative conformations mediated by formation of a hairpin at the 3´end of the ribozyme (HEG P1). The inactive conformation has been found to be favoured at low Mg^{2+} concentration and even co-exists with the active conformation at physiological and higher Mg^{2+} concentrations. Interestingly, structure probing has revealed the presence of the inactive conformation characterized by the presence of the HEG P1 stem-loop structure concomitantly with the high accessibility of all internal key junctions. Finally, the active conformation of DiGIR1 has been deduced by a combination of structure probing and structure modelling strategy. Fe-EDTA structure probing of DiGIR1 has made it possible to monitor the melting of the HEG P1 stem loop structure and the P2P2.1P10 three-way junction folding. This results in the docking of P2.1 onto the core. Consequently, the key junctions are protected. As a summary, the regulatory domain acts as an on/off switch that orchestrates the activities of the branching and splicing ribozymes in the twin-ribozyme intron in several different processing pathways and thus regulates the interplay between the intron and the host. Thus, P2 formation imposes architectural constraints on P2 P2.1 P10 3-way junction that lead to conformation triggering activity.

<u>Paper IV:</u> K. L. Andersen, B. Beckert, B. Masquida, M. Andreassen, S. D. Johansen , H. Nielsen (2010). "Accumulation of stable full-length circular group I introns during heat-shock". RNA submitted, accepted upon revision.

The DiGIR2 splicing ribozyme, found in the twin-ribozyme intron organization, belongs to subgroup IE of group I introns. Furthermore, it has been characterized to possess a strong circularization pathway that can be favoured by playing on three different parameters: the concentration of the guanosine cofactor, the deletion of peripheral elements and the antisense oligonucleotides targeting the 3' end of the intron. In this study, we have focused on the ability of DiGIR2 group I intron to form full-length circle (FLC) via its circularization pathway. Both the structure of FLC and its linear counterpart (linear intron LIVS) and *in vivo* conditions that may result in up-regulating the circularization pathway in the cell have been investigated.

In this paper, we have first investigated the copy number of FLC present in the myxomycete *D. iridis* under different growth conditions (normal, heat-shock and cold-shock conditions). During this first part, qRT-PCR has been used to determine the copy number of FLC in the cell. Interestingly, FLC nucleolar or cytosolic localization has also been investigated. We have then focused on the structure of both FLC resulting from the circularization pathway and LIVS intron resulting from the self-splicing pathway. By using enzymatic and chemical structure probing method, the secondary structures of both LIVS and FLC have been deduced. Finally, structure probing results of these two different forms have been rationalized in two detailed three-dimensional molecular models.

As a result, FLCs have been found to be predominantly, if not exclusively, nuclear with approximately 70 copies/cell during exponential growth. Interestingly, external factors (i.e. cellular stress by heat shock) have been shown to up-regulate the formation of FLC with more than 500 copies/cell. These results support the notion that circular form of group I introns is a biologically relevant molecule. Interestingly, we can assume from these results that FLC may be instrumental in group I intron mobility. Comparison of the FLC and LIVS structure probing data in combination with their respective structural model highlights the structural differences found within the catalytic core. Remarkably, theses differences have been found to be mainly clustered both next to the active site (P7) and in joining segments (J6/7, J8/7, J5/4). These results emphasize that a large proportion of FLC molecules harbour a

relaxed active site with a disassembled G-binding pocket. Moreover, the FLC P1-P2 domain is structurally perturbed. The circularized junction together with P1 show high accessibilities with an internal guide sequence exposed to the solvent. These FLC structural characteristics observed by both structure probing and by modelling are consistent with group I intron mobility and the potential ability of circular intron to reverse integrate into their cognate insertion site. Finally, the molecular model suggestes that the P9 extension could play a role in promoting circularization. In this way, by providing alternative stabilization of P7, the P9 domain seems to orchestrate the balance between splicing and circularization.

CHAPTER IV: ARTICLES

ARTICLE I:

Molecular modelling of the GIR1 branching ribozyme gives new insight into evolution of structurally related ribozymes

B. Beckert, H. Nielsen, C. Einvik, S.D. Johansen, E. Westhof, B. Masquida.

The EMBO Journal (2008) 27, 667–678 | © 2008 European Molecular Biology Organization | Some Rights Reserved 0261-4189/08
www.embojournal.org

THE EMBO JOURNAL

EMBO open

Molecular modelling of the GIR1 branching ribozyme gives new insight into evolution of structurally related ribozymes

Bertrand Beckert[1], Henrik Nielsen[2,3,*], Christer Einvik[4], Steinar D Johansen[3], Eric Westhof[1] and Benoît Masquida[1,*]

[1]Architecture et Réactivité de l'ARN, Université Louis Pasteur de Strasbourg, IBMC, CNRS, Strasbourg, France, [2]Department of Cellular and Molecular Medicine, The Panum Institute, University of Copenhagen, Copenhagen, Denmark, [3]Department of Molecular Biotechnology, Institute of Medical Biology, University of Tromsø, Tromsø, Norway and [4]Department of Pediatrics, University Hospital of North Norway, Tromsø, Norway

Twin-ribozyme introns contain a branching ribozyme (GIR1) followed by a homing endonuclease (HE) encoding sequence embedded in a peripheral domain of a group I splicing ribozyme (GIR2). GIR1 catalyses the formation of a lariat with 3 nt in the loop, which caps the HE mRNA. GIR1 is structurally related to group I ribozymes raising the question about how two closely related ribozymes can carry out very different reactions. Modelling of GIR1 based on new biochemical and mutational data shows an extended substrate domain containing a GoU pair distinct from the nucleophilic residue that dock onto a catalytic core showing a different topology from that of group I ribozymes. The differences include a core J8/7 region that has been reduced and is complemented by residues from the pre-lariat fold. These findings provide the basis for an evolutionary mechanism that accounts for the change from group I splicing ribozyme to the branching GIR1 architecture. Such an evolutionary mechanism can be applied to other large RNAs such as the ribonuclease P.

The EMBO Journal (2008) **27,** 667–678. doi:10.1038/emboj.2008.4; Published online 24 January 2008
Subject Categories: RNA
Keywords: GIR1 branching ribozyme; group I intron; RNA evolution; 3D modelling of RNA

Introduction

The list of naturally occurring ribozymes comprises a few that are fundamental for cellular life (the ribosome, RNase P, and possibly the spliceosome), two types of splicing ribozymes that are abundant in organellar and microbial genomes (within group I and group II introns), and a number of cleavage ribozymes with a sporadic occurrence in viroids, plant satellite RNAs, bacteria and, more recently, within the human genome (hammerhead, hairpin, VS, HDV, glmS, and the CPEB3 ribozymes). Apart from the ribosome, all naturally occurring ribozymes catalyse phosphor transfer reactions (Ditzler *et al*, 2007; Scott, 2007; Serganov and Patel, 2007). A recent addition to the list is the GIR1 branching ribozyme. This ribozyme (Figure 1) catalyses cleavage of the RNA chain by transesterification resulting in the formation of a 2′,5′ phosphodiester bond between the first and the third nucleotide of the 3′-cleavage product. The downstream cleavage product is an mRNA encoding a homing endonuclease (HE) that is thereby capped with a lariat containing 3 nt in the loop (Nielsen *et al*, 2005). Both GIR1 and the downstream HE mRNA are inserted into a peripheral domain of a regular splicing ribozyme (GIR2) making up the characteristic configuration of a twin-ribozyme intron. Such introns have so far only been found in the SSU rDNA genes of a unique isolate of *Didymium iridis* and in several *Naegleria* strains where it has been vertically inherited from a common ancestor (Johansen *et al*, 2002; Wikmark *et al*, 2006; Nielsen *et al*, 2008). The biological function of GIR1 appears to be in the formation of the 5′ end of the HE mRNA during processing from the spliced out intron and the resulting lariat cap seems to contribute by increasing the half-life of the HE mRNA (Vader *et al*, 1999; Nielsen *et al*, 2005), thus conferring an evolutionary advantage to the HE.

One of the interesting features of GIR1 is that the sequence and the secondary structure are very similar to that of eubacterial group IC3 introns at the second step of splicing (Figure 2), suggesting an evolutionary relationship with this specific subgroup of splicing ribozymes (Johansen *et al*, 2002). The secondary structure of GIR1 displays paired segments numbered P3–P10, similar to what is known in group I introns (Figure 2). The paired segments are generally shorter than those observed in group I introns consistent with the fact that the shortest form of DiGIR1 shown to catalyse branching *in vitro* is only 179 nt (Nielsen *et al*, 2005). Both ribozymes are organized as a compact bundle of three helical stacks (Figure 2; domains P3–P9, P4–P6, and P10–P2 (group I ribozyme) or P10–P15 (GIR1)). Group I intron classification is based on structural variation of peripheral elements organized around a very well-conserved catalytic core (Michel and Westhof, 1990). In contrast, the main distinctive features of GIR1 towards group I ribozymes occur within the catalytic core. Several characteristic single-stranded junctions tether the helices of a catalytic core containing a double pseudoknot in a way that leads to significant topological modifications

*Corresponding authors. B Masquida or H Nielsen, Architecture et Réactivité de l'ARN, Université Louis Pasteur de Strasbourg, IBMC, CNRS, 15 rue René Descartes, Strasbourg 67084, France. Tel.: +333 88 41 70 45; Fax: +333 88 60 22 18; E-mail: b.masquida@ibmc.u-strasbg.fr or Hamra@imbg.ku.dk

Received: 10 August 2007; accepted: 4 January 2008; published online 24 January 2008

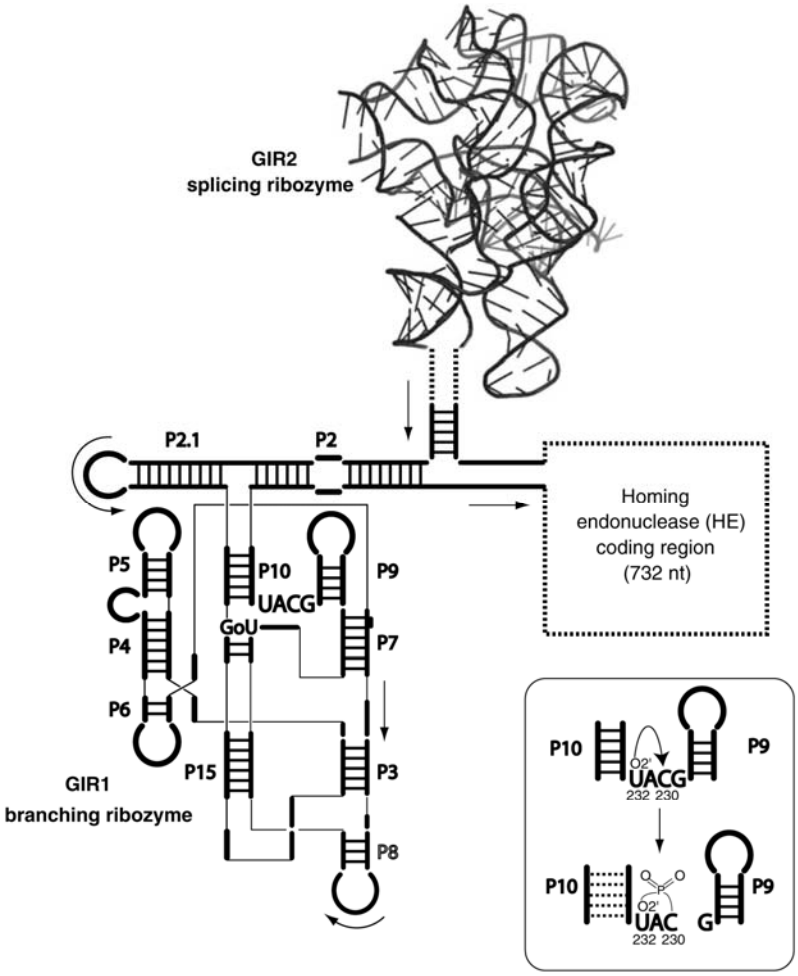


Figure 1 The RNA transcript from the Dir.S956-1 locus is composed of a regular group I splicing ribozyme (GIR2) represented as a 3D ribbon generated from Guo *et al* (2004) containing an ORF encoding an homing endonuclease (HE, dashed line) preceded by a branching ribozyme (GIR1, schematic secondary structure). The branching reaction consists of the O2′-hydroxyl group from U232 attacking the phosphate group from C230 leading to the formation of a 3-nt lariat (box). The branching reaction releases the cleaved ribozyme from the 5′ end of the HE mRNA. In addition to stems P3–P15, DiGIR1 harbours a specific extension, namely P2/P2.1 (Einvik *et al*, 2000), which could not be included unambiguously into the present model.

(Figure 2). GIR1 harbours a substrate domain different from the canonical P1 and P2 elements. However, the biochemical and mutational data presented in this study support the idea that they are replaced by a distinctive and unique 9-bp P15 stem starting with a GoU pair that should be able to dock onto the catalytic core in a way similar to that observed for group I introns. The close resemblance of GIR1 to a splicing ribozyme in an unrelated group of organisms and the structural organization of twin-ribozyme introns may be related to the propagation of group I introns by horizontal transfer. Group I introns are considered as mobile elements due to their sporadic occurrence in a wide variety of organisms, including protists, fungal mitochondria, bacteria, and phages (Haugen *et al*, 2005). Many lines of evidence point to reverse splicing and homing as mechanisms by which group I introns can transfer horizontally (Goddard *et al*, 2001; Bhattacharya *et al*, 2005; Haugen *et al*, 2005). The homing mechanism is well documented and appears to be particularly relevant to GIR1 because its activity is intimately related to the expression of a HE mRNA (H Nielsen, in preparation).

The intriguing observation that GIR1 and the group I splicing ribozymes are structurally related, yet carry out different reactions (splicing versus branching) prompted us to revise our previous structural model of GIR1. This model (Einvik *et al*, 1998b; Johansen *et al*, 2002) was based on structure probing and mutational studies. It predated the discovery of the branching reaction (Nielsen *et al*, 2005) and could not account for this reaction. The model presented in this study is based on new mutational data and furthermore benefits from the recent crystal structures of various group I ribozymes (Guo *et al*, 2004; Adams *et al*, 2004a, b; Golden *et al*, 2005) in the sense that the GIR1 regions organized identically in group I introns could be modelled more accurately. In our new model, residues that are key to the branching reaction lie within a pocket formed at the interface of P10, P15, P7, and J5/4. All the distinctive features

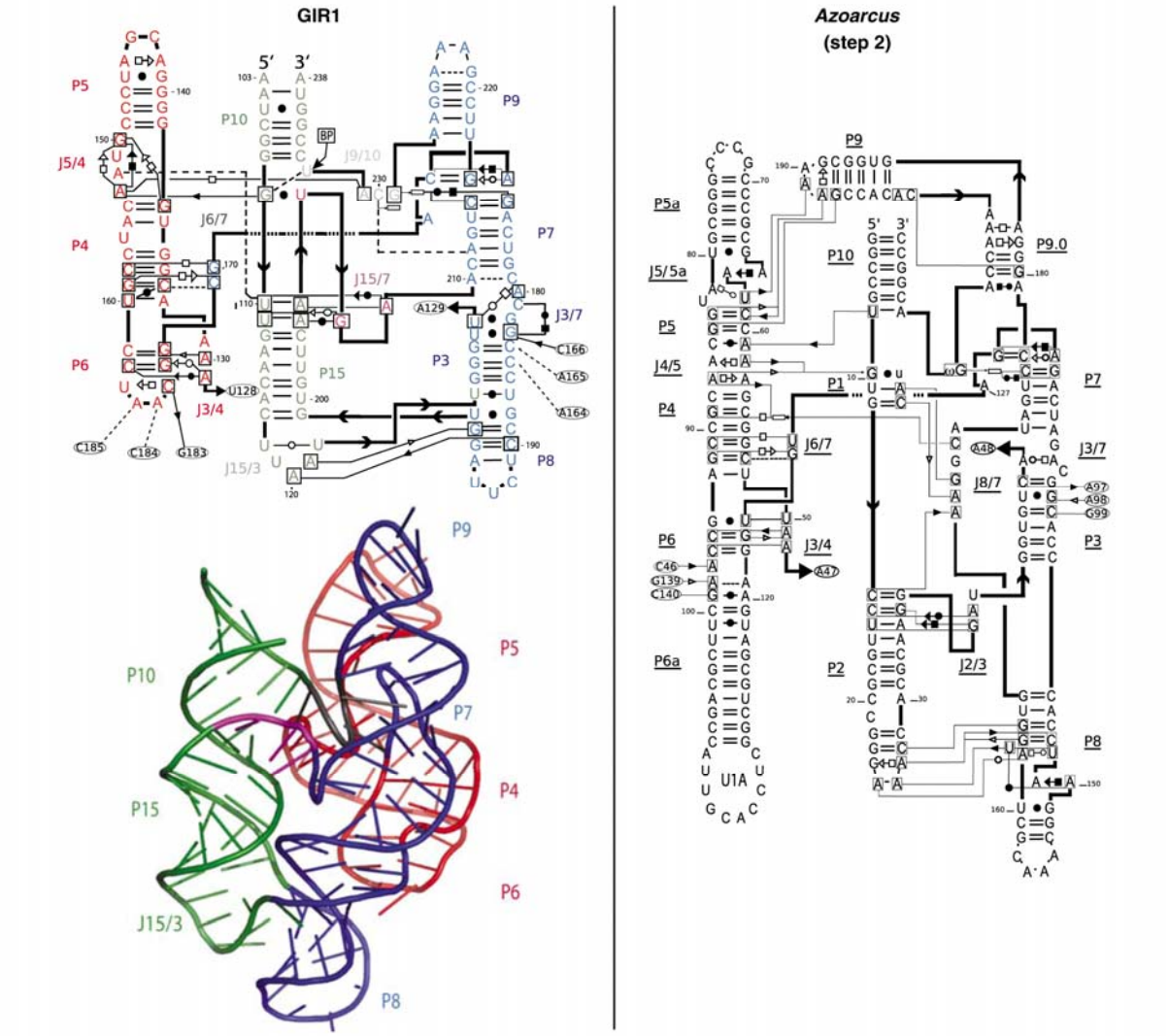


Figure 2 Overall representation of *Didymium* GIR1 secondary and 3D structures (left panel) and *Azoarcus* group I tRNA^{Ile} intron (*Azo*) secondary structure in (pre-) step 2 state (right panel). The secondary structure corresponding to the crystal structure model of Adams *et al* (2004a) harbours a U1A protein receptor motif. The non-canonical interactions are displayed on both secondary structures using the formalism elaborated by Leontis and Westhof (2001). The three helical domains of GIR1 (P10–P15 corresponding to *Azo* P10–P2, P4–P6, and P3–P9) are organized as a compact bundle at the centre of which lies the junction harbouring the residues involved in catalysis (J9/10, black ribbon). The overall architecture is stabilized by contacts recurrently observed in group I ribozymes, J15/3–P8 corresponding to L2–P8 in *Azo*, and L6–P3 corresponding to J6/6a–P8 in *Azo*. The contact between L9 and P5 is not observed in GIR1 despite the presence of the tetraloop sequence 5′ GAAA in L9. Topological differences are due to the presence of the double pseudoknot involving P3, P7, and P15 in GIR1.

of GIR1 concentrate in this pocket and result in a topology very different from what is observed in group I ribozyme crystal structure models. The structure of the critical J8/7 segment of group I introns is dramatically changed and has been partly replaced by residues belonging to the GIR1 lariat fold J9/10. Other key features are the detachment of the nucleophile from the GoU pair at the catalytic site and a structural alteration of the GoU pair receptor. Taken together, these structural differences account for the different chemical reaction catalysed by GIR1. Comparison of the models of the *Azoarcus* tRNA^{Ile} intron at the second step of splicing and GIR1 suggests a relatively simple model for the conversion of the topology of one ribozyme to the other based on strand mispairing. Similar scenarios can apply to other RNAs, for example, RNase P, and could constitute a general way of viewing the evolution of RNA molecules.

Results

In this section, the structure model of the *Didymium* GIR1 (DiGIR1) ribozyme is extensively compared with the *Azoarcus* group I ribozyme (*Azo*) crystal structure (Adams *et al*, 2004a). Hence, secondary structure elements and nucleotides corresponding to *Azo* are underlined throughout the text. The secondary structure of DiGIR1 and the similar ribozyme from *Naegleria* (NaGIR1) is generally supported by enzymatic and chemical probing (Einvik *et al*, 1998a; Jabri *et al*, 1997; Jabri and Cech, 1998). Furthermore, the *Naegleria*

structure is supported by covariations observed in most of the helical stems when NaGIR1 from different strains are compared (Johansen *et al*, 2002; Wikmark *et al*, 2006). NaGIR1 performs a branching reaction similar to that of DiGIR1 (H Nielsen, unpublished data) supporting the notion that the two GIR1 ribozymes adopt similar secondary structures. DiGIR1 harbours an additional domain P2/P2.1 (Einvik *et al*, 2000) not found in NaGIR1. This domain is excluded from the model because it is currently impossible to discriminate between several different models.

Extension of the P15 stem

Modelling of GIR1 is facilitated by the presence of a double pseudoknot at the core. In addition to the P3–P7 pseudoknot also found in all group I ribozymes (Michel and Westhof, 1990), a second pseudoknot, P3–P15, is found as a characteristic feature of GIR1 (Einvik *et al*, 1998b). P15 arises from base-pairing interactions between the 5′ strand of P2 with residues that could be derived from the 3′ strand of P8 and from J8/7, while the 3′ strand of P2 has been shortened and now makes up the J15/3 segment (Figure 3). Thus, one can visualize P15 as replacing the shallow/minor groove interactions taking place between J8/7 and P2, which are conserved in group I ribozymes (Strauss-Soukup and Strobel, 2000; Soukup *et al*, 2002). Inspired by the comparison with *Azo*, we now propose an extension of P15 involving residues 205–207. Residues A205 and A206 appear to be equivalent to A residues in J8/7 responsible for recognition of the P1–P2 substrate (Figure 2). J8/7 is a highly conserved joining segment in group I ribozymes that is part of the active site and makes contacts with all of the three principal domains of the group I ribozyme. During the first and second steps of splicing, the two conserved adenosines at the 5′ end of J8/7 are involved in recognition of the P1–P2 interface (Pyle *et al*, 1992; Tanner *et al*, 1997; Strauss-Soukup and Strobel, 2000; Adams *et al*, 2004a). In the original model of GIR1, a P1 was not included but could arise from a 3-bp extension of P15 resulting from the base complementarity between residues A205–U207 from J15/7 and U111–G109 separating P10 from P15, respectively. The existence of these three base pairs could originate from the interaction between a P1 having lost its 5′ exon making the residues from the internal guide sequence (IGS) prone to base pair with J8/7. Indeed, the crystal structure of the *Tetrahymena* group I ribozyme (Guo *et al*, 2004) shows that residues in J8/7 are directed towards the solvent when the substrate domain P1–P2 is absent. It is therefore likely that some J8/7 residues could form Watson–Crick interactions with a substrate domain containing unpaired nucleotides as it occurs when the IGS is separated from the 5′ exon. To confirm this possibility, disruptive and restoring mutations of the central base pair U110–A206 were tested by kinetic cleavage analyses (Supplementary Figure S1). *In vitro*, GIR1 catalyses (i) a forward branching reaction in equilibrium with (ii) a very efficient reversed reaction, and (iii) an inefficient hydrolytic cleavage reaction (Nielsen *et al*, 2005; Nielsen and Johansen, 2007). The outcome of the reaction can be analysed by primer extension with stop signals at the branch nucleotide or at the cleavage site representing branching and hydrolysis, respectively. All disruptive mutations resulted in reduced cleavage rates. The double mutations that restored base pairing (U110A–A206U and U110C–A206G) performed branching at a rate comparable to that of wild type. The possibility for nucleotides A205 and A206 to engage in base pairing with U110 and U111 additionally suggests that G109 base pairs with U207 to form a continuous helical stack at the junction between P10 and P15.

The GoU pair at the P10–P15 interface

The secondary structure of DiGIR1 allows for two different possibilities of forming a GoU pair at the catalytic site in analogy with the GoU pair in P1. In both cases, G109 is involved but the pairing partner could either be U207 or the branch nucleotide U232, as in the original model (Einvik

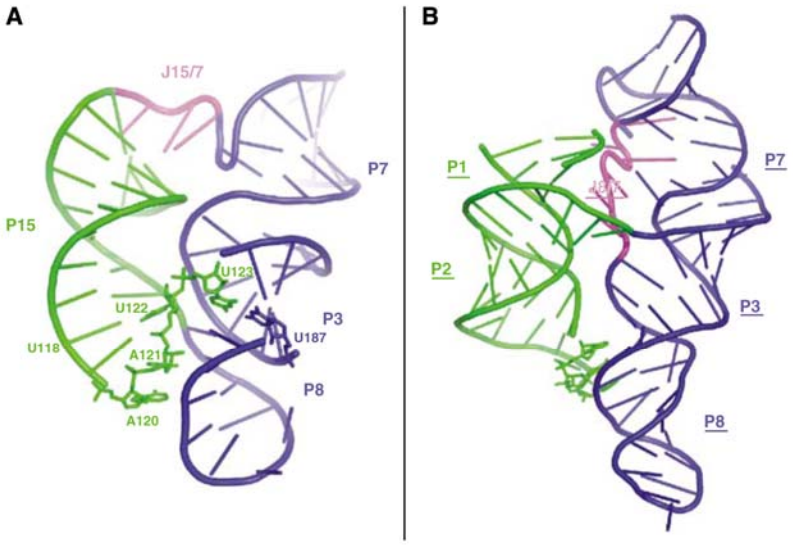


Figure 3 Comparison between the double-pseudoknot fold of GIR1 (**A**) and the corresponding region in *Azo* (**B**). (A) J15/7 forces P15 to be placed along P3 with which it forms a pseudoknot. Hence, the position of P15 corresponds to the position of P2 in *Azo* (B). In addition, this fold is confirmed by the conformation of the 3WJ since J15/3 (green stick residues) makes A-minor interactions in the shallow groove of P8, thus functionally replacing the interaction of L2 with the latter.

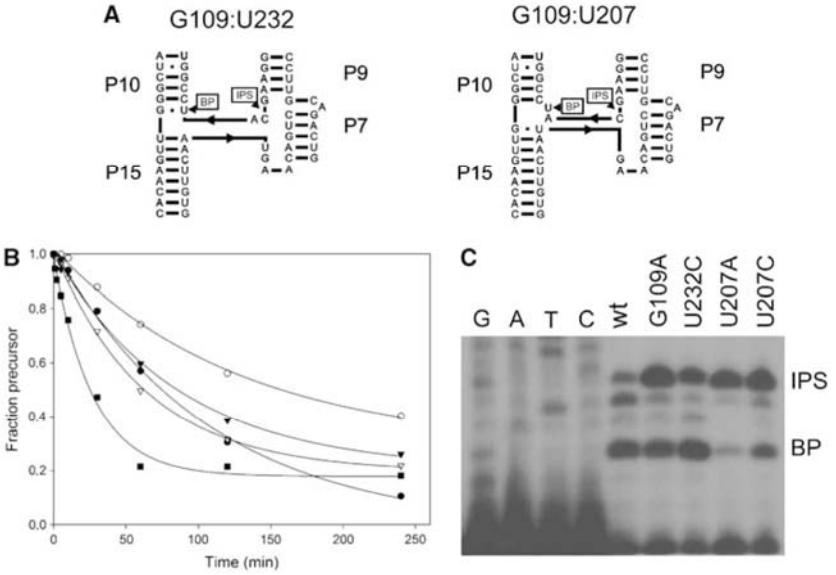


Figure 4 Mutational analysis of the GoU pair at the active site. (**A**) Two putative base-pairing schemes involve G109. In the left panel, G109 is base paired to the branch point (BP) nucleotide U232. In the right panel, this nucleotide is extra-helical and G109 is base paired to U207. (**B**) Kinetic cleavage analysis of mutations of the nucleotides involved. All of the mutations result in a decrease in the cleavage rate and mutations of the two potential base-pairing partners have a similar effect on the cleavage rate ((●) G109A, (○) U207A, (▼) U207C, (▽) U232C, and (■) wild-type GIR1). (**C**) Primer extension analysis of end point samples from the cleavage analysis. A primer extension stop signal at the BP indicates cleavage by branching and a stop at the internal processing site (IPS) indicates cleavage by hydrolysis. Mutations of G109 and U207 result in accumulation of more of the hydrolysis than the branching product in contrast to mutation of U232.

et al, 1998b) (Figure 4A). To distinguish between these two possibilities, the effect of mutations of the involved nucleotides on the cleavage rate (Figure 4B) and on the type of reaction (branching versus hydrolytic cleavage; Figure 4C) was assessed. The wild type is characterized by predominant cleavage by branching with only a small fraction (<10%) of stop signal, indicating hydrolytic cleavage after a 4 h incubation. Mutations that disrupt the catalytic pocket would be expected to affect both the branching and the hydrolysis rates. Such an activity loss has been previously observed following mutations of the ωG nucleotide (G229; Johansen *et al*, 2002) and disruption of the ωG-binding site in P7 (G174C; Decatur *et al*, 1995). Conversely, mutations that affect the positioning of the U232 relative to the ωG without disrupting the catalytic pocket would be expected to shift the reaction from branching to hydrolysis. The mutation G109A maintains the base-pairing potential with U232 or U207. The effect of the mutation is a moderate reduction in cleavage rate (Johansen *et al*, 2002). However, the reaction results in more hydrolysis than branching products. The mutations U232C and U207C similarly maintain the ability of these residues to base pair with G109. The effect on the cleavage rate is a comparable reduction in cleavage rate to that of G109A. However, cleavage in the U232C mutant results in more branching than hydrolysis product as in wild type, whereas cleavage in the U207C mutant results in more hydrolysis than branching product. The disruptive mutant U207A displays an even more reduced cleavage rate and cleaves almost exclusively by hydrolysis. The accumulation of more hydrolysis than branching product as seen in G109A and U207 mutants is an unusual phenotype as judged from our analysis of over 50 GIR1 mutants. Furthermore, mutants U232A and U232G that would likely disrupt base pairing involving this nucleotide cleave more by branching than by hydrolysis similar to U232C (H Nielsen, in preparation). These observations are in favour of base pairing of G109oU207 instead of G109oU232 as in the original model (Einvik *et al*, 1998b). In this way, the critical GoU pair at the active site belongs to a P1-like helix (the extended P15) as in splicing group I ribozymes and not to P10 as in the original model (Einvik *et al*, 1998b). A further implication is that U207 forming the GoU pair does not provide the nucleophile for the branching reaction as it occurs in group I ribozymes. Rather, U232 lies in the shallow/minor groove of the G109oU207 pair, where it potentially interacts with the amino group of G109 to drive the branching reaction (Figure 5A). These mutational data furthermore validate the 3-bp extension of P15, which contributes significantly to the re-design of the catalytic core by forming a continuous helical stack between P10 and P15.

Recognition of the substrate domain P10–P15 by J5/4

In group I ribozymes, the GoU pair in P1 is recognized by a wobble receptor located at the interface of P4 and P5 (Michel and Westhof, 1990; Wang and Cech, 1992; Strobel and Cech, 1994; Strauss-Soukup and Strobel, 2000). When viewed in secondary structure diagrams, the structure of this interface is a 3-nt symmetrical internal loop. In DiGIR1, the interface between P4 and P5 is asymmetrical with a 4-nt junction, 5′-GUAA, as J5/4 and no intervening nucleotides at the 5′ strand (Supplementary Figure S2A). Furthermore, J5/4 is one of the most variable segments as deduced from the *Naegleria* GIR1 sequence alignment albeit with conserved features (Wikmark *et al*, 2006). To assess the importance of J5/4 in DiGIR1, systematic mutational analysis of J5/4 residues was

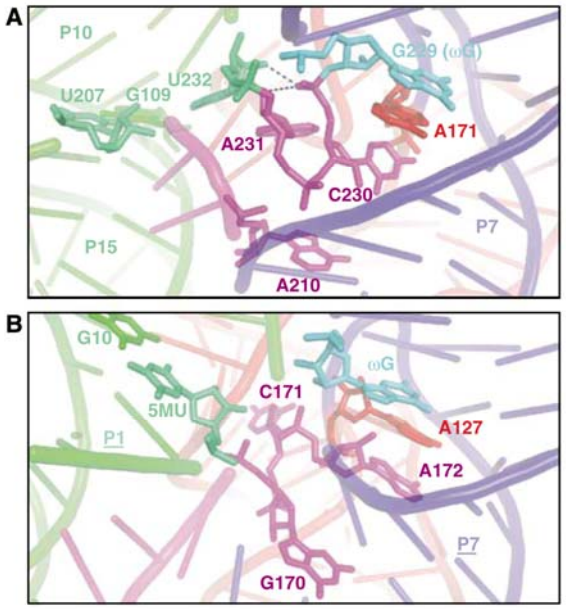


Figure 5 The catalytic pocket of the GIR1 ribozyme (**A**) spans from the ωG (G229, cyan residue) binding pocket to the nucleophile U232 (green residue); the GoU pair in the substrate domain lies at the P10–P15 interface (left side). G229 interacts in the deep groove side of the second base pair of P7 (blue ribbons). The lariat fold residues C230 and A231 interact with P7 and J5/4, respectively. The oxygen atoms of the scissile phosphate group are in contact with the 2′-hydroxyl groups of A231 and U232 (dashed lines). The 2′-hydroxyl and phosphate groups in the vicinity of the scissile phosphate potentially provide ligands for the catalytic magnesium ions. (**B**) In the same orientation and using the same colouring scheme (ωG in cyan, nucleophilic 5-methyl uracil in green (5MU)), the *Azoarcus* ribozyme shows that the GIR1 lariat fold residues functionally replace nucleotides from J8/7. The lariat fold together with J15/7 thus constitute a composite J8/7.

performed. Major alterations of the structure, such as deletion of J5/4, substitution of J5/4 with 5′-UUCG, or deletion of the bulged U156 all resulted in a complete loss of activity (data not shown). Substitution of the individual nucleotides resulted in decreased cleavage rates in all cases (Supplementary Figure S2B). The effect of mutating G150, U151, and A152 was moderate but the effect of the A153G mutant was dramatic pointing to this nucleotide as a key nucleotide for reactivity. Taken together, these results demonstrate an important function of J5/4 in GIR1 consistent with a preserved role of this structure in GoU recognition at the active site.

Alterations in the catalytic core do not affect the overall structure

The double pseudoknot provides a high level of constraint that guarantees confident model building of this region. The three stems P15, P3, and P8 together form a three-way junction (3WJ) already constrained in the P3–P15 pseudoknot. In the present model, the extended P15 is docked along P3 and adopts a parallel orientation with the co-axial stack occurring between stems P3 and P8. This conformation is promoted by the presence of the fairly long J15/3 stretch that forms a loop capping P15 and is able to interact in the shallow groove of P8 (Figure 3). A recent survey of 3WJ structures shows that J15/3 is part of a kind of 3WJ that occurs at 10 ribosomal RNA locations and in several other RNA crystal structures (Lescoute and Westhof, 2006). Moreover, the above-mentioned survey shows that, when present in the longest loop, adenine residues are instrumental in stabilizing the junction architecture through the formation of A-minor interactions in the narrow groove of the facing stem, hence mimicking the GNRA/tetraloop receptor inter-domain interactions between P2 and P8 observed in group I introns crystal structures (see below).

Consequently, the single strands connecting P15 to the neighbouring helices can be considered as characteristic features distinguishing GIR1 from group I ribozymes. The constraints due to the 3WJ and to the double pseudoknot result in P15 occupying the same place as P2 (Figure 3). Furthermore, P15 directly stacks onto P10 by taking advantage of the 3-bp extension of P15 that was not considered in the original model (Einvik *et al*, 1998b). Hence, P10 and P7 adopt a relative position to the 3-bp P15 extension similar to what is observed for stem P1 in the crystal structures of group I introns (Adams *et al*, 2004b; Golden *et al*, 2005). This conformation is also supported by the fact that it leads to the formation of a pocket where all the structural elements necessary to form the catalytic site, namely ωG, U232, and the G109oU207 pair from P15 are gathered, a condition not satisfied by other tested models of 3WJ.

P8 and P9 were then directly connected to the double pseudoknot to form the catalytic domain. The connections between the P3–P9 catalytic core and the P4–P6 domain of GIR1 are similar to what is observed in *Azo* crystal structure (Adams *et al*, 2004a). In other words, J3/4 and J6/7 are modelled so as to weave the same contacts as those observed in *Azo* with the shallow groove of P6 and the narrow groove of P4, respectively (Figure 2). The last residue from J6/7 (A171) plays the same role as in *Azo* by providing stacking continuity between G229 (ωG) and the closest residue from J9/10 (C230), which corresponds to the last residue of J8/7 (A172) in *Azo* (Figure 5B). Since the P4–P6 domain is connected to the core as in group I introns, P6 consequently resides in the vicinity of P3, and the P4–P5 interface is able to contact the P10–P15 interface. Regarding the P7–P9 interface, a very discrete difference occurs. P7 is tethered to P9 without the intervening A residue frequently observed in group I ribozymes. This observation is important because J7/9 has been proposed to sequester ωG during the first step of splicing (Rangan *et al*, 2004), a condition not necessary in GIR1.

Apart from tertiary interactions specifically found in the core of the ribozyme, the group I intron architecture is stabilized by three sets of tertiary interactions (Figure 2). The first two interlock elements P2 and P6 with P8 and P3, respectively, and the third one allows L9 to contact P5 (Jaeger *et al*, 1996). The structural homology between group I introns and GIR1 ribozymes would plead for the existence of similar inter-domain interactions. However, these interactions are necessarily affected by the fact that the secondary structure elements from GIR1 are different or shorter than in group I introns.

The double pseudoknot corresponds to a motif swap for the known interaction between P2 and P8 (Michel and Westhof, 1990; Salvo and Belfort, 1992; Costa and Michel, 1995). The resulting model suggests that J15/3 replaces the tetraloop located at the tip of P2 and interacts in the shallow groove of P8 (Figure 3). As in the *Tetrahymena* ribozyme

(Guo *et al*, 2004), the loop receptor in GIR1 P8 consists of two consecutive G = C pairs instead of an 11-nt receptor motif as observed in *Azo* (Adams *et al*, 2004b). J15/3 loops over itself to enter the 5′ strand of P3. U123 interacts with U187 and the resulting base pair intercalates at the P3–P8 interface. U122 base pairs with U118 and provides stacking continuity between P15 and J15/3. A120 and A121 form A-minor interactions in the minor groove of P8. This conformation is supported by mutants of U residues to C that affect the structure of the 3WJ (data not shown).

In *Didymium*, P6a does not exist and J6/6a thus becomes the tetraloop L6 capping P6. Hence, L6 is perfectly located to form the recurrent interaction between J6/6a and P3 (Waldsich *et al*, 2002; Adams *et al*, 2004a; Golden *et al*, 2005) using A-minor interactions (Doherty *et al*, 2001; Nissen *et al*, 2001) between A residues from L6 and two consecutive G = C pairs from P3 in the shallow/minor groove. Moreover, this tertiary contact is supported by mutational analysis of the two consecutive A residues from L6 (data not shown) and by the secondary structure of the NaGIR1 which displays a P6 element longer than in DiGIR1, albeit interrupted by an A-rich internal loop that could presumably function as J6/6a (Einvik *et al*, 1998b). It is noteworthy that the two tertiary interactions described above are also supported by chemical probing experiments showing that A residues important for the described contacts are protected from DMS and DEPC (Einvik *et al*, 1998b).

In contrast to the previous interactions, the characteristic interaction formed in group I between L9 and P5 is not conserved in GIR1 ribozymes (Figure 2). In DiGIR1, P9 is short (4 bp) and does not contain any hinge point that allows it to bend towards P5. However, its 5′-GAAA tetraloop could interact with a receptor embedded in the P2/P2.1 extension (Einvik *et al*, 2000; Nielsen *et al*, 2005) that was not included in the model (Figure 1).

Organization of the catalytic core

The next step consisted in understanding the architecture of the catalytic core in this unforeseen structural context. Around the catalytic region, the only fully identical feature shared by group I and GIR1 ribozymes resides in the organization of the ωG (G229) binding pocket (Figures 2 and 5). The Watson–Crick edge of G229 H-bonds with the Hoogsteen edge of the second G = C pair of P7, and is stabilized by stacking interactions between the first C = G pair of P7 and A171 from J6/7.

On the opposite side of the ribozyme, the J5/4 junction on which the substrate domain docks is organized quite differently in GIR1 compared to *Azo*. The dramatic loss of activity in the A153G mutant is consistent with its protection from chemical modification by DMS (Einvik *et al*, 1998b), and justifies orienting this key residue towards the core of the ribozyme. To achieve this, the two 5′ nucleotides from J5/4 (G150 and U151) are placed so as to lie in the deep/major groove of P4 to improve the stacking continuity between P4 and P5. A kink performed around the phosphate group of A152 allows A153 to loop back into P5 ejecting A152 and A153 towards P10–P15. Thus, J5/4 becomes part of the catalytic pocket and shields P4 (Supplementary Figure S3). In such a situation, A153 can bind the GoU base pair from P15 as does A58 in J4/5 (Adams *et al*, 2004b). Moreover, A152 interacts with U110 to provide a tandem of A-minor interactions. These A-minor interactions account for the observed loss of activity in the A153G mutant since G residues are rarely observed in contact with G = C base pairs in this motif (Doherty *et al*, 2001; Nissen *et al*, 2001). We propose that the interaction between P1 and J4/5 is replaced by A-minor tandem interactions involving the G109oU207 and U110–A206 pairs resulting from the extension of P15 with A153 and A152 from J5/4, respectively.

The lariat residues C230 and A231 replace important residues from the J8/7 junction in group I ribozyme

To suggest a relevant position for the residues involved in the lariat, a best-fitting lariat model with a 3-nt loop, obtained by an NMR study of an A2′-pG branched RNA (Agback *et al*, 1993) was accommodated in the catalytic pocket between P7, P5, and P10. The NMR models of these lariat RNAs provide a starting model from which several structural features can be characterized. The short length of the lariat loop forces the ribose–phosphate backbone to form the inner ring of the loop while ejecting the base moieties on the outside. As a result, the base rings cannot stack together but occupy distinct volumes in which they could interact with other chemical moieties. A lariat harbouring the DiGIR1 5′-CAU sequence while keeping the conformation of the RNA studied by NMR (Agback *et al*, 1993) was docked in the active site in a search for the best orientation. In the course of the refinement, the lariat was debranched to model a conformation corresponding to the pre-cleavage state. The ring formed by the lariat is short and tight with a kink around the phosphate group of A231 forcing base moieties of residues C230 and A231 to point towards structural elements forming the catalytic pocket with which they can interact (Figure 4A).

Interestingly, the lariat residues could be placed within the pocket left free following the relocation of residues 207–209 from J15/7 that extend P15. In this position, A231 and C230 take over the role of residues C171 and A172 from J8/7 in their ability to interact towards P4 and P7, respectively. C230 stacks with A171 (J6/7) strengthening the deep groove 4-nt stack including G229, and taking the place occupied by the 3′ A residue from J8/7 in *Azo* (Figure 5). A231 points towards J5/4 as a consequence of the lariat sharp turn and places this nucleotide at hydrogen bonding distance of A153. Although a base-pairing interaction is implicated, a geometry explaining the deep effect of the A231G mutant could not be clearly deduced based on chemical footprinting data (H Nielsen, in preparation). In the course of the catalytic formation of the GIR1 U2′-pC lariat, C230 and U232 are covalently attached following the nucleophilic attack of the O2′ group of U232 onto the phosphorus atom of C230. It is thus reasonable to place these chemical groups at H-bonding distance (2.8 Å) by taking advantage of the closest oxygen atom of the phosphate group. In the conformation proposed, the 2′-hydroxyl group of A231 interacts with the oxygen atom of the phosphate group of C230, which is not already in contact with the O2′ atom of U232 (Figure 4A). A deoxy substitution scan experiment (Nielsen *et al*, 2005) pointing out the important role of the 2′-hydroxyl group of A231 comes to support the proposed architecture since nucleotides at the 3′ end of J8/7 are involved in coordinating the catalytic magnesium ions using phosphate oxygen atoms and/or hydroxyl groups. Hence, the model strongly suggests that magnesium ions are relocated in J9/10 in the vicinity of C230 and A231.

J15/7 and J9/10 complement each other to stabilize the ribozyme catalytic core by forming a composite J8/7 junction that coordinates magnesium ions, and places the nucleophilic U residue in close vicinity of the targeted phosphate group.

Discussion

Building a new structural model for the GIR1 ribozyme was prompted by the recent finding that the ribozyme catalyses the formation of a short lariat containing a 3-nt loop by transesterification (Nielsen *et al*, 2005). The new model is consistent with previous biochemical and mutational data and incorporates new mutational data presented in this study. The model shows that the group I ribozyme substrate stems P1 and P2 are replaced in GIR1 by a distinctive and unique 9-bp P15 stem starting with a GoU pair. The modelling strategy relied on the existence of a double pseudoknot involving stems P3 and P7 on the one hand (a mandatory feature of group I intron catalytic core structure), and stems P3 and P15 on the other hand (Figure 3). The proposed architecture of this highly constrained double pseudoknot is consistent with the conformation of the 3WJ additionally encompassing P8 (Lescoute and Westhof, 2006). Apart from the characteristic P15, the secondary structure of GIR1 is similar enough to canonical group I introns to unambiguously claim their phylogenetic relationship. Surprisingly, the elements distinguishing GIR1 from the group I splicing ribozymes lie within the usually very well-conserved catalytic core (Michel and Westhof, 1990). The different topology results in a core that despite the marked similarity in base-pairing scheme between GIR1 and the group I ribozymes at the second step does not carry out splicing. Rather, the position of the nucleophile is shifted from the last base pair in P1 to the interface between the analogous P15 extension and P10, thereby allowing for the branching reaction. Thus, the function of carrying the nucleophile is detached from the GoU pair and the catalytic reaction occurs in *cis* rather than in *trans*.

The topological differences between the catalytic cores of group I and GIR1 ribozymes resulting from the presence of the double pseudoknot heavily impacts the architecture of the catalytic pocket. GIR1 harbours a G-binding pocket in P7 identical to the pocket observed in *Azoarcus* pre-tRNA^{Ile} intron. The extended substrate helix P15 (analogous to P1–P2) is recognized by the protruding J5/4 using two consecutive A-minor interactions (Doherty *et al*, 2001; Nissen *et al*, 2001) instead of the base-pair tandem formed between the sugar and Hoogsteen edges (Leontis and Westhof, 2001) of A residues in J4/5. Residues C230 and A231 in the loop of the lariat fold replace key residues from *Azo* J8/7 (A172 and C171) in their ability to interact with P7 and P4, respectively. Since in *Azo*, C171 and A172 are involved in binding the two magnesium ions that are required for catalysis (Adams *et al*, 2004b), it is tempting to suggest that residues constituting the lariat fold provide some of the ligands for binding the active site metal ions.

Shortening of J8/7 may account for the appearance of the branching reaction

The most dramatic feature of the topological change in GIR1 is that the joining segment that connects to the 5′ strand of P7 comes from P15 (J15/7) rather than from P8 (J8/7) and that it has been shortened down to 3 nt as a consequence of the extension of P15. Hence, J15/7 is stretched and adopts a very different path compared to J8/7 in *Azo*. As a consequence, the two 5′ residues are excluded from residing inside the pocket and are located towards the outer shell of the molecule and the branching residue U232 is allowed to dock in the shallow groove of P15 and be accommodated into the catalytic pocket. The 3′ residue of J15/7 (A210) functionally replaces the fourth nucleotide of J8/7 (G170) in *Azo* by interacting with J7/3. Hence, the space left free by the absence of the two 3′ nucleotides of J8/7 can be occupied by the two first residues (C230 and A231) from the lariat fold. These nucleotides are followed by the branching U232, with a 3′-hydroxyl group already tethered to the downstream RNA chain. U232 is positioned similarly with respect to the cleavage site as the 3′ U of the 5′ exon in *Azo*. The absence of a free 3′-hydroxyl group in an environment prone to bind magnesium ions and generate nucleophilic oxygen atoms from hydroxyl groups may have driven the O2′ of U232 to attack the facing phosphate group of C230 and form the 3-nt lariat characteristic of GIR1 ribozymes.

GIR1 topology may have arisen by drift of the 3′ strand of P2, the 5′ strand of P3, and J8/7 sequences of the ancestor intron

The three most notable features of GIR1 are that (i) it has so far only been found in the setting of twin-ribozyme introns, (ii) it closely resembles the eubacterial IC3 introns, and (iii) despite this close similarity, GIR1 catalyses a branching reaction rather than splicing. Although it is generally difficult to demonstrate any evolutionary path, only a few discrete events would be required to account for the emergence of the GIR1 branching ribozyme from group I introns. For the emergence of the twin-ribozyme configuration, it is plausible that a bacterial intron invaded a group I intron containing a HEG insertion. Myxomycetes are rich in nuclear rDNA introns with a relatively large proportion containing HEGs and the possibility of an invading bacterial intron is supported by the recent observation of a sister intron to Dir.S956-1 in the myxomycete *Diderma* (SD Johansen, unpublished data). The *Diderma* intron is located at the exact same rDNA position as the *Didymium* intron and has an almost identical group I splicing ribozyme with a very similar HEG inserted into the P2 segment. However, the *Diderma* intron lacks a GIR1 ribozyme and thus may represent the pre-existing receptor intron in the model. The configuration of an intron within an intron is reminiscent of the case of group II/III twintrons (Copertino and Hallick, 1993) and renders the invading intron in a situation with no stringent requirement to preserve the splicing activity. This could have set the stage for the subsequent transition of the invading intron into a branching ribozyme.

In the present study, we have shown that the difference between the *Azoarcus* intron representing the IC3 group I introns and the GIR1 branching ribozyme consists mainly of topological changes in the core. At the sequence level, we noticed that a single transposition event of the 5′-GUGUUC stretch from the 3′ strand of P15 of the wild-type GIR1 to A120 in J15/3 would restore the topology and the base-pairing scheme found in *Azo* (Supplementary Figure S4). Further, at the 3D level, exchanges of phosphate bonds at positions that come in fairly close distance on the *Azo* crystal structure would result in the GIR1 topology (Figures 2 and 6A). From a mechanistic point of view, a model based

on sequence drift and strand exchange promoted by the absence of a selection pressure for splicing can be envisaged. In this scenario, a gradual sequence change leads to sequence similarities between the 5′ strand of P3, the 3′ strand of P2 and J8/7 resulting in a strand pairing switch within the core (Figure 6B). This evolutionary pathway via alternative pairing is similar to the mutational drift experimentally demonstrated between the HδV ribozyme and the artificial class III ligase (Schultes and Bartel, 2000). In the evolutionary model of GIR1, misfolding has been promoted by the loss of the 5′ exon that has driven the J8/7 from GIR1 to form a pseudo P1 that expanded to form P15. It is noteworthy that this process gives rise to the double pseudoknot (P3–P7 and P3–P15) that contributes to the energetic stabilization of the core of the ribozyme. The higher stability conferred to the ribozyme core by the appearance of the double pseudoknot may have been important to allow the peripheral domains to evolve with only minor implications on the core structure explaining why they are reduced to short appendices. In a context with low selection pressure towards splicing, since GIR1 was already embedded in a self-splicing intron, this process relied on sequence similarities between the segments of the ribozyme involved in mispairing (P1, P2, P3, and J8/7). Misfolding of a group I ribozyme around J8/7 and P3 has been experimentally observed (Pan and Woodson, 1998). The misfolding generating GIR1 may have been positively selected by allowing the branching reaction to occur, which conferred a selective advantage by increasing the half-life of the HE mRNA (H Nielsen, in preparation; Johansen *et al*, 2007). The topological shuffle described here fully accounts for the topological

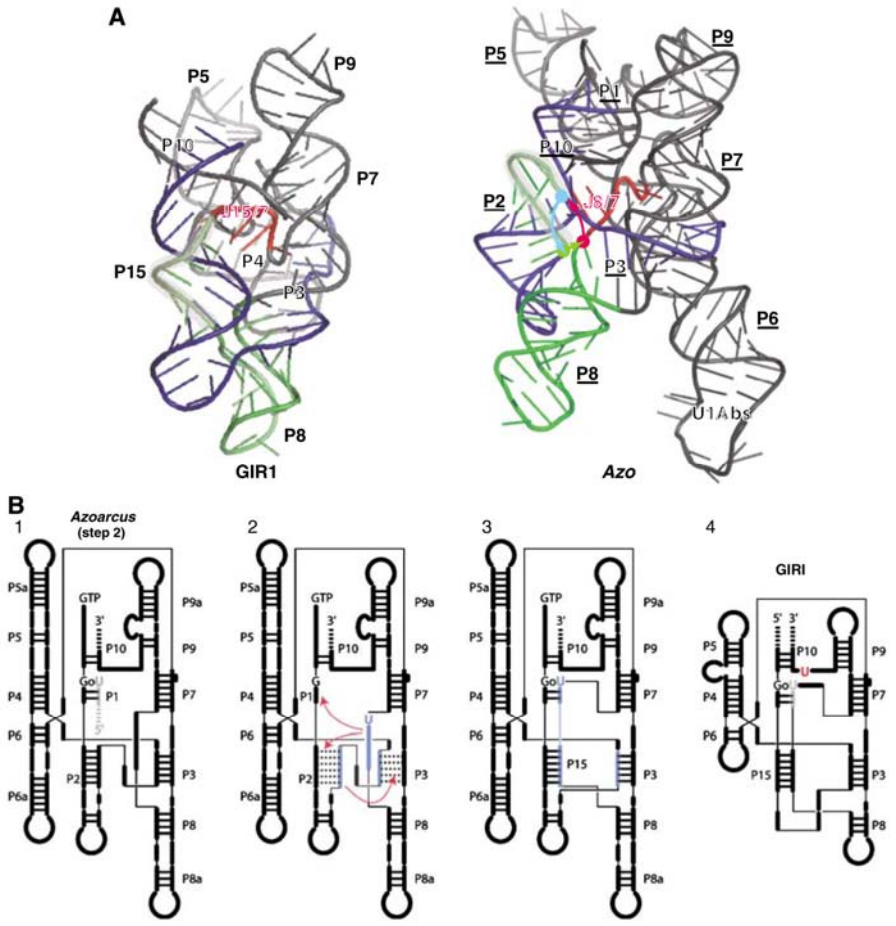


Figure 6 The different topologies of GIR1 and group I ribozymes. (**A**) At the 3D level, shuffling the backbone in regions where RNA backbones are in close vicinity (coloured arrows on the right panel displaying *Azo* crystal structure) leads to changing the topology of the *Azo* ribozyme to the GIR1 ribozyme. The first shuffling point intervenes at A27 in P2, the second involves G37 in P2, and the third one takes place at G166 in P8. Once *Azo* is cleaved at these positions, the intron can be religated. C28 is attached to G38 to connect the loop of P2 to P3 becoming J15/3. G37 is then attached to A167 to make J15/7, and finally G166 is linked to C28 to complete the shuffling. Equivalent backbone portions are coloured identically. Note the conserved position of the outlined segment which corresponds either to the 3′ strand of P15 in GIR1 or to the 3′ strand of P2 in *Azo* (U1Abs: U1A protein binding site). (**B**) A model for the evolution of a group I intron into the GIR1 ribozyme. (1) The *Azoarcus* group I ribozyme at a stage prior to the second catalytic step undertakes mutations in J8/7, P2, and P3 that lead to partial alternative pairing in folding. (2) The loss of the 5′ exon favours misfolded species by drifting of neighbouring sequence stretches altering the overall secondary structure. (3) The GIR1 fold is selected due to the appearance of a new chemical reaction allowing the formation of the lariat and conferring an increased half-life to the homing endonuclease mRNA. (4) The gain in energetic stabilization due to the presence of the new pseudoknot P3–P15 allows for a shortening of some peripheral elements leading to the final version of GIR1 ribozyme that is always shorter than group I ribozymes.

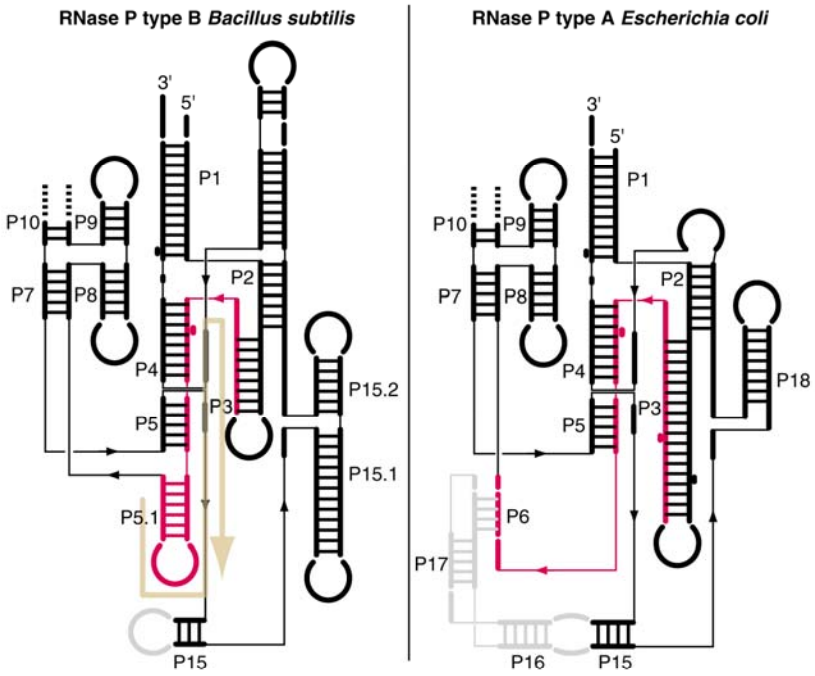


Figure 7 The drift model can be applied to the bacterial RNase P ribozymes of subtypes A and B. Starting from the type B ribozyme (left panel), pulling the 5′ end of the red strand, as suggested by the orange arrow, lengthens P3 as in type A ribozyme (right panel), and shortens P5.1. P5.1 does not fold anymore as a hairpin and finds a new pairing partner in its vicinity (region depicted in grey) leading to the formation of the P6 pseudoknot. This scheme is made possible by the existence of an extended L15 loop in type A ribozyme compared to the short L15 in type B ribozyme.

changes observed between the core of group I ribozymes and GIR1, the appearance of the double pseudoknotted structure with the extended P15, and the redefinition of the role of J8/7 (now J15/7).

The proposed mechanism for evolution of new RNA molecules may apply to other RNAs. The two main families of ribonuclease P ribozymes (Darr *et al*, 1992a, b) can be distinguished by secondary structure changes occurring in a single contiguous region: the path from family B to family A involves lengthening of stem P3, disruption of stem P5.1 with formation of a new pseudoknot P6 (Figure 7). In contrast, in the HδV/ligase case (Schultes and Bartel, 2000), all pairing stems are involved in strand exchange. Even though the above scenario leading to GIR1 evolution seems to be the most relevant, we cannot rule out that some unknown transposition events could have taken place at the RNA level with subsequent transfer to the DNA level by reverse transcription and integration into the genome or by the recently described RNA-directed DNA repair mechanism (Storici *et al*, 2007).

In conclusion, we have provided a model that correlates the branching activity of GIR1 with its topological difference compared to that of group I splicing ribozymes. We suggest also an evolutionary mechanism for the emergence of GIR1 based on the shuffling between functional motifs promoted by sequence shift and alternative pairings. Additional proofs will be needed and could be inspired from studies of other GIR1 ribozymes, such as those found in *Naegleria* (Einvik *et al*, 1997; Jabri *et al*, 1997; Johansen *et al*, 2002; Wikmark *et al*, 2006).

Materials and methods

In vitro mutagenesis

Extension of P15. Mutations at U110 were introduced by PCR using *Pfu* DNA polymerase of a wild-type GIR1 template (pDi162G1 Decatur *et al* (1995)) and oligos C377 (see Supplementary data for details) or C378 as the 5′-oligo and OP12 as the 3′-oligo. The PCR product was re-amplified to make templates for *in vitro* transcription (see below). Mutations of A206 were introduced by *in vitro* mutagenesis using the Quick Change site-directed mutagenesis kit (Stratagene) and oligos C405/C406 and C407/C408. To make double mutants, the U110 mutations were introduced into the A206 mutated templates.

GU pair. Construction of G109A and U232C were previously published (Johansen *et al*, 2002). Mutations at U207 were made as described above using oligos C477/C478 or C479/C480.

J5/4. Mutations were made in a wild-type GIR1 template as described above and oligos C415/C416, C424/C425, C270/C271, and C417.

Cleavage analyses and primer extension

Templates for *in vitro* transcription were made from wild-type and mutant templates by PCR using *Pfu* DNA polymerase (Fermentas) and oligos C287: 5′-AATTTAATACGACTCACTATAGGGTTGGGAAG TATCAT and C288: 5′-TCACCATGGTTGTTGAAGTGCACAGATTG. C287 carries a T7 RNA polymerase promoter. The run-off transcript from the PCR template includes 162 nt upstream and 65 nt downstream of the cleavage site. All templates were transcribed *in vitro* using T7 RNA polymerase (Fermentas) with trace amounts of [α-^{32}P]UTP. Cleavage analysis was performed as described in Einvik *et al* (2000). Briefly, radioactively labelled *in vitro* transcripts were renatured in 1 M KCl, 25 mM $MgCl_2$ at pH 5.5 for 10 min at 45°C. Then the reaction was jump started by increasing the pH to 7.5 by addition of Hepes-KOH. Time samples were withdrawn and run on

6% denaturing (urea) polyacrylamide gels. The gels were analysed on storage phosphor screens and the data fitted to a nonlinear first-order decay equation. The experiments shown are representative results of 3–5 independent experiments. Primer extension analysis was carried out as described (Einvik *et al*, 1998b) using end-labelled oligo C291: 5′-GATTGTCTTGGGAT. Sequencing ladders were made using the same primer and the plasmid pDi162G1 (Einvik *et al*, 1998b) as template. The reactions were analysed on 8% denaturing (urea) polyacrylamide gels.

Molecular modelling

Molecular modelling was performed as described in Masquida and Westhof (2005). The lariat model taken from Agback *et al* (1993) corresponds to the RNA lariat with a 3-nt loop in which all residues presents a C2′-endo conformation taken from http://www.boc.uu.se/boc14www/res_proj/final_struct/pictures/Welcome.html (file cGUAC_md25_A.pdb). The lariat was debranched to allow the phosphate group of the 5′ residue to be tethered to ωG. The sequence of J9/10 was applied to the lariat fold using the program fragment embedded in the manip software (Massire and Westhof, 1998). This program was also used to build in three dimensions (pdb file format) all GIR1 pieces similar to the corresponding group I intron regions. All the 3D elements were assembled interactively on a SGI Octane graphical workstation (IRIX64 v6.5, IP30) using the manip software. Each step of manual modelling was followed by several least-square refinement step (Westhof *et al*, 1985). The modelling/refinement cycles were iterated until a model satisfying all the constraints was obtained. Figures were prepared using the PYMOL program (DeLano WL, The PyMOL Molecular Graphics System (2002) http://www.pymol.org). Secondary structure diagrams in Figure 2 were directly generated from the PDB files using the program S2S (Jossinet and Westhof, 2005).

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (http://www.embojournal.org).

Acknowledgements

This project was supported the Danish Natural Science Research Council. BB was supported by CEE BAC RNA program (LSHG-CT-2005-018618) and the Lundbeck Foundation. We thank Pascale Romby for critical reading of the manuscript.

References

Adams PL, Stahley MR, Gill ML, Kosek AB, Wang J, Strobel SA (2004a) Crystal structure of a group I intron splicing intermediate. *RNA* **10:** 1867–1887

Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA (2004b) Crystal structure of a self-splicing group I intron with both exons. *Nature* **430:** 45–50

Agback P, Sandstrom A, Yamakage S, Sund C, Glemarec C, Chattopadhyaya J (1993) Solution structure of lariat RNA by 500 MHz NMR spectroscopy and molecular dynamics studies in water. *J Biochem Biophys Methods* **27:** 229–259

Bhattacharya D, Reeb V, Simon DM, Lutzoni F (2005) Phylogenetic analyses suggest reverse splicing spread of group I introns in fungal ribosomal DNA. *BMC Evol Biol* **5:** 68–78

Copertino DW, Hallick RB (1993) Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. *Trends Biochem Sci* **18:** 467

Costa M, Michel F (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J* **14:** 1276–1285

Darr SC, Brown JW, Pace NR (1992a) The varieties of ribonuclease P. *Trends Biochem Sci* **17:** 178–182

Darr SC, Zito K, Smith D, Pace NR (1992b) Contributions of phylogenetically variable structural elements to the function of the ribozyme ribonuclease P. *Biochemistry* **31:** 328–333

Decatur WA, Einvik C, Johansen S, Vogt VM (1995) Two group I ribozymes with different functions in a nuclear rDNA intron. *EMBO J* **14:** 4558–4568

Ditzler MA, Aleman EA, Rueda D, Walter NG (2007) Focus on function: single molecule RNA enzymology. *Biopolymers* **87:** 302–316

Doherty EA, Batey RT, Masquida B, Doudna JA (2001) A universal mode of helix packing in RNA. *Nat Struct Biol* **8:** 339–343

Einvik C, Decatur WA, Embley TM, Vogt VM, Johansen S (1997) *Naegleria* nucleolar introns contain two group I ribozymes with different functions in RNA splicing and processing. *RNA* **3:** 710–720

Einvik C, Elde M, Johansen S (1998a) Group I twintrons: genetic elements in myxomycete and schizopyrenid amoeboflagellate ribosomal DNAs. *J Biotechnol* **64:** 63–74

Einvik C, Nielsen H, Nour R, Johansen S (2000) Flanking sequences with an essential role in hydrolysis of a self-cleaving group I-like ribozyme. *Nucleic Acids Res* **28:** 2194–2200

Einvik C, Nielsen H, Westhof E, Michel F, Johansen S (1998b) Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites. *RNA* **4:** 530–541

Goddard MR, Greig D, Burt A (2001) Outcrossed sex allows a selfish gene to invade yeast populations. *Proc Biol Sci* **268:** 2537–2542

Golden BL, Kim H, Chase E (2005) Crystal structure of a phage Twort group I ribozyme–product complex. *Nat Struct Mol Biol* **12:** 82–89

Guo F, Gooding AR, Cech TR (2004) Structure of the *Tetrahymena* ribozyme; base triple sandwich and metal ion at the active site. *Mol Cell* **16:** 351–362

Haugen P, Simon DM, Bhattacharya D (2005) The natural history of group I introns. *Trends Genet* **21:** 111–119

Jabri E, Aigner S, Cech TR (1997) Kinetic and secondary structure analysis of *Naegleria andersoni* GIR1, a Group I ribozyme whose putative biological function is site-specific hydrolysis. *Biochemistry* **36:** 16345–16354

Jabri E, Cech TR (1998) *In vitro* selection of the *Naegleria* GIR1 ribozyme identifies three base changes that dramatically improve activity. *RNA* **4:** 1481–1492

Jaeger L, Michel F, Westhof E (1996) The structure of group I ribozymes. In *Nucleic Acids and Molecular Biology*, Eckstein F, Lilley DMJ (eds), Vol. 10, pp 33–51. Berlin: Springer Verlag

Johansen S, Einvik C, Nielsen H (2002) DiGIR1 and NaGIR1: naturally occurring group I-like ribozymes with unique core organization and evolved biological role. *Biochimie* **84:** 905–912

Johansen SD, Haugen P, Nielsen H (2007) Expression of protein-coding genes embedded in ribosomal DNA. *Biol Chem* **388:** 679–686

Jossinet F, Westhof E (2005) Sequence to structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* **21:** 3320–3321

Leontis NB, Westhof E (2001) Geometric nomenclature and classification of RNA base pairs. *RNA* **7:** 499–512

Lescoute A, Westhof E (2006) Topology of three-way junctions in folded RNAs. *RNA* **12:** 83–93

Masquida B, Westhof E (2005) Modeling the architecture of structured RNAs within a modular and hierarchical framework. In *Handbook of RNA Biochemistry*, Hartmann RK, Bindereif A, Schön A, Westhof E (eds), pp 536–545. Weinheim, Germany: Wiley VCH Verlag GmbH & Co.

Massire C, Westhof E (1998) MANIP: an interactive tool for modelling RNA. *J Mol Graph Model* **16:** 197–205, 255–197

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

Nielsen H, Beckert B, Masquida B, Johansen SD (2008) The GIR1 branching ribozyme. In *Ribozymes and RNA Catalysis*, Lilley DMJ, Eckstein F (eds), pp 229–252. London: Royal Society of Chemistry

Nielsen H, Johansen SD (2007) A new RNA branching activity: the GIR1 ribozyme. *Blood Cells Mol Dis* **38:** 102–109

Nielsen H, Westhof E, Johansen S (2005) An mRNA is capped by a 2′, 5′ lariat catalyzed by a group I-like ribozyme. *Science* **309:** 1584–1587

Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci USA* **98:** 4899–4903

Pan J, Woodson SA (1998) Folding intermediates of a self-splicing RNA: mispairing of the catalytic core. *J Mol Biol* **280:** 597–609

Pyle AM, Murphy FL, Cech TR (1992) RNA substrate binding site in the catalytic core of the *Tetrahymena* ribozyme. *Nature* **358:** 123–128

Rangan P, Masquida B, Westhof E, Woodson SA (2004) Architecture and folding mechanism of the *Azoarcus* group I pre-tRNA. *J Mol Biol* **339:** 41–51

Salvo JL, Belfort M (1992) The P2 element of the td intron is dispensable despite its normal role in splicing. *J Biol Chem* **267:** 2845–2848

Schultes EA, Bartel DP (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **289:** 448–452

Scott WG (2007) Ribozymes. *Curr Opin Struct Biol* **17:** 280–286

Serganov A, Patel DJ (2007) Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat Rev Genet* **8:** 776–790

Soukup JK, Minakawa N, Matsuda A, Strobel SA (2002) Identification of A-minor tertiary interactions within a bacterial group I intron active site by 3-deazaadenosine interference mapping. *Biochemistry* **41:** 10426–10438

Storici F, Bebenek K, Kunkel TA, Gordenin DA, Resnick MA (2007) RNA-templated DNA repair. *Nature* **447:** 338–341

Strauss-Soukup JK, Strobel SA (2000) A chemical phylogeny of group I introns based upon interference mapping of a bacterial ribozyme. *J Mol Biol* **302:** 339–358

Strobel SA, Cech TR (1994) Translocation of an RNA duplex on a ribozyme. *Nat Struct Biol* **1:** 13–17

Tanner MA, Anderson EM, Gutell RR, Cech TR (1997) Mutagenesis and comparative sequence analysis of a base triple joining the two domains of group I ribozymes. *RNA* **3:** 1037–1051

Vader A, Nielsen H, Johansen S (1999) *In vivo* expression of the nucleolar group I intron-encoded I-DirI homing endonuclease involves the removal of a spliceosomal intron. *EMBO J* **18:** 1003–1013

Waldsich C, Masquida B, Westhof E, Schroeder R (2002) Monitoring intermediate folding states of the td group I intron *in vivo*. *EMBO J* **21:** 5281–5291

Wang JF, Cech TR (1992) Tertiary structure of around the guanosine-binding site of *Tetrahymena* ribozyme. *Science* **256:** 526–529

Westhof E, Dumas P, Moras D (1985) Crystallographic refinement of yeast aspartic acid transfer RNA. *J Mol Biol* **184:** 119–145

Wikmark O-G, Einvik C, De Jonckheere J, Johansen S (2006) Short-term sequence evolution and vertical inheritance of the *Naegleria* twin-ribozyme group I intron. *BMC Evol Biol* **6:** 39–50

REVIEW I:

The GIR1 branching ribozyme.

H. Nielsen, B. Beckert, B. Masquida and S. D. Johansen.

In Ribozymes and RNA catalysis, Lilley DMJ and Eckstein F, eds. (London: The Royal Society of Chemistry), pp. 229-252.

CHAPTER 12

The GIR1 Branching Ribozyme

HENRIK NIELSEN,[a,b] BERTRAND BECKERT,[c] BENOIT MASQUIDA[c] AND STEINAR D. JOHANSEN[b]

[a] Department of Cellular and Molecular Medicine, The Panum Institute, University of Copenhagen, Copenhagen, DK-2200N, Denmark; [b] Department of Molecular Biotechnology, Institute of Medical Biology, University of Tromsø, Tromsø, Norway; [c] Institut de Biologie Moléculaire et Cellulaire, Centre National de le Recherche Scientifique, Université Pasteur, Strasbourg, France

12.1 Introduction

The GIR1 branching ribozyme is a *ca.* 179 nt ribozyme with structural resemblance to group I ribozymes.[1] It is found within a complex type of group I introns, termed the twin-ribozyme introns.[2] Rather than splicing, it catalyses a branching reaction in which the 2′OH of an internal residue is involved in a nucleophilic attack at a nearby phosphodiester bond.[3] As a result, the RNA is cleaved at an internal processing site (IPS), leaving a 3′OH and a downstream product with a tiny lariat at its 5′ end (Figure 12.1). The lariat has the first and the third nucleotide joined by a 2′,5′ phosphodiester bond and is referred to as "the lariat cap" because it caps an intron-encoded mRNA. The biological function of the GIR1 ribozyme thus appears to be in expression of an intron-encoded protein.

The GIR1 ribozyme was originally discovered during functional characterization of the *Didymium* twin-ribozyme intron. Combined deletion and *in vitro* self-splicing analyses revealed two distinct ribozyme domains within the intron.[2] Subsequently it was discovered that both ribozyme domains could be folded as group I ribozymes (GIRs), and named GIR1 and GIR2 according to the order of completion during transcription of the intron.[4] GIR2 was shown to be a conventional group IE splicing ribozyme and GIR1 a cleavage ribozyme acting at an intron internal site.

A description of the biology of GIR1 requires the introduction of a certain amount of nomenclature. Group I introns in ribosomal RNA (rRNA) genes are

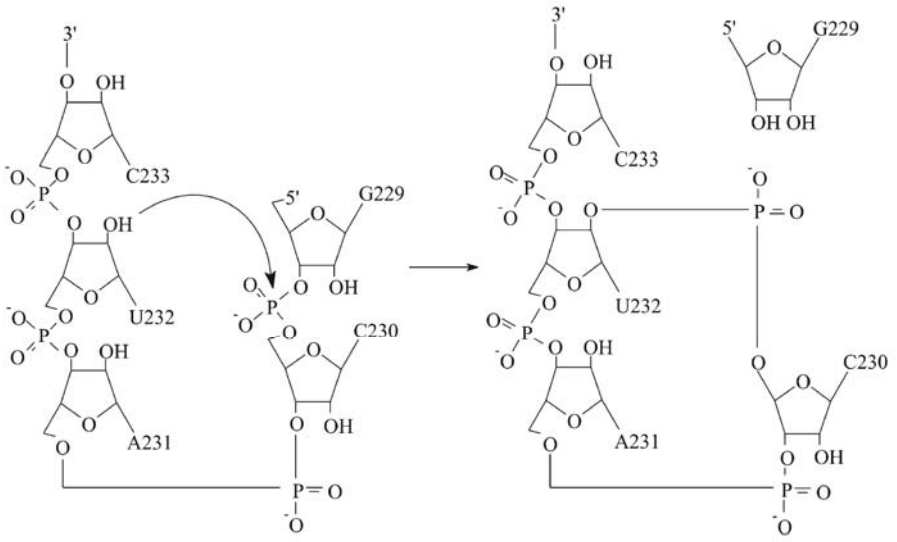


Figure 12.1 Branching reaction catalysed by GIR1. The 2′OH of the internal residue U232 makes a nucleophilic attack at the IPS. Bond lengths not drawn to scale. (The figure is reproduced from ref. 3 with permission.)

named according to Johansen and Haugen.[5] The name of host species and the insertion site within rRNA genes (*Escherichia coli* numbering) are reflected in the intron nomenclature. Dir.S956-1 is the twin-ribozyme intron inserted in the small rRNA gene of *Didymium iridis* at position 956. Since two different introns have been found in different isolates, this particular intron is numbered "-1". In a similar way, Nae.S516 is the twin-ribozyme intron from various species and isolates of *Naegleria*. Whenever a general feature of the branching ribozyme is described, it is referred to as "GIR1". When the description is specific for an individual ribozyme, or has only been investigated in one example, the species is indicated, *e.g.* "DiGIR1". Since several *Naegleria* ribozymes are known, "NaGIR1" refers to the *Naegleria* ribozyme in general and for example "NanGIR1" to the ribozyme from *Naegleria andersoni*. The classification of group I introns into subgroups follows the system by Michel and Westhof[6] based on differences in the structure of peripheral elements. DiGIR2 is a group IE,[7] and NaGIR2 a group IC ribozyme. Twin-ribozyme introns include a homing endonuclease gene (HEG). The transcript is referred to as the I-*Dir*I mRNA and the protein as the I-*Dir*I homing endonuclease following the rules for classification of these and similar enzymes.[8,9] Nucleotides in GIR1 are numbered according to their position within the twin-ribozyme intron. The minimal branching variant of DiGIR1 begins at pos. 73 and ends at pos. 251 in Dir.S956-1. The length variants are named according to how much sequence is included upstream, and downstream of the IPS. The minimal variant is thus named 157.22 to state that 157 nt upstream and 22 nt downstream of the IPS are included, respectively.

In this chapter, we describe the GIR1 ribozyme, focussing on GIR1 from the myxomycete *D. iridis* (DiGIR1). The emphasis is on describing GIR1 as a ribozyme for which a plausible hypothesis can be made as to how the structure was derived from group I introns and how this resulted in the gaining of a new reaction mechanism.

12.2 Distribution and Structural Organization of Twin-ribozyme Introns

The twin-ribozyme introns represent some of the most complex organized group I introns known and consist of a homing endonuclease gene (HEG) embedded in two functionally distinct catalytic RNA domains.[10] One of the catalytic RNAs is a conventional group I intron ribozyme (GIR2) responsible for intron splicing and reverse splicing, as well as intron RNA circularization. The other catalytic RNA domain is the group I-like ribozyme (GIR1) directly involved in homing endonuclease mRNA maturation.[3,4,11] Only two main natural variants of the twin-ribozyme group I introns are known,[1] and these include the Dir.S956-1 intron from the myxomycete *D. iridis* and the Nae.S516 from various species and isolates of *Naegleria* amoeboflagellates (see Table 12.1 below). Both introns are inserted into conserved regions of the nuclear small subunit rRNA (SSU rRNA) gene in their respective host organisms, and have a similar overall structural organization at the RNA level (Figure 12.2).

Several differences in distribution, inheritance, and structural organization are noted between the *Didymium* and *Naegleria* twin-ribozyme introns:

(i) The introns are located at different insertion sites in the SSU rRNA gene. Whereas the 1.4 kb *Didymium* intron is inserted after a U residue at position 956 (*E. coli* SSU rRNA numbering), the 1.2 kb *Naegleria* intron is located after residue 516. Both sites are frequently known to harbour group I introns in nuclear ribosomal DNA (rDNA) of eukaryotic microorganisms.

(ii) The pattern of intron distribution among host organisms is different. The Dir.S956-1 twin-ribozyme intron is unique to the Panama 2 isolate of *D. iridis*. In fact, very closely related *D. iridis* isolates either lack any group I intron at position 956 or harbour distantly related group I introns at this position.[12,13] The Nae.S516 intron, on the other hand, is restricted to the *Naegleria* genus, with a widespread but sporadic distribution that includes 21 of 70 strains analysed.[14]

(iii) The pattern of intron inheritance appears different. The *Didymium* intron is an example of an optional group I intron. The closest relatives known to the *Didymium* splicing ribozyme domain (DiGIR2) are found within different myxomycete genera, but not in a *Didymium* species, suggesting a recent gain by horizontal intron transfer. Interestingly, experimental support of Dir.S956-1 mobility has been obtained both at the RNA-level due to reverse splicing[15] and at the DNA-level due to

Table 12.1 GIR1 present in *Didymium* and *Naegleria* isolates.

Species/isolate	*GIR1 size (nt)*[a]	*Acc no*	*Ref.*
Didymium iridis Pan2-44	180	AJ938153	Johansen and Vogt 1994[2]
Naegleria andersoni A2	198	X78280	DeJonckheere 1994[46]
N. andersoni PPMFB-6	198	Z16417	DeJonckheere and Brown 1994[47]
N. carteri NG055	211	AM167878	Wikmark *et al.* 2006[14]
N. clarki RU30	197	AM167879	Wikmark *et al.* 2006[14]
N. clarki RU42	197	AM167880	Wikmark *et al.* 2006[14]
N. clarki Pd72Z/I	197	AF338417	Dykova *et al.* 2001[48]
N. clarki 4177/I	197	AF338418	Dykova *et al.* 2001[48]
N. clarki 4564/IV	197	AF338419	Dykova *et al.* 2001[48]
N. clarki 4709/I	197	AF338420	Dykova *et al.* 2001[48]
N. clarki Pd56Z/I	197	AF338422	Dykova *et al.* 2001[48]
N. clarki CB1S/I	197	DQ768725	Dykova *et al.* 2006[49]
N. gruberi CCAP1518-1D	208	X78278	DeJonckheere 1994[46]
N. italica AB-T-F3	209	U80249	Einvik *et al.* 1997[11]
N. jamiesoni T56E	197	U80250	Einvik *et al.* 1997[11]
N. martinezi NG872	219	AJ001399	DeJonckheere and Brown 1998[50]
N. philippinensis RJTM	209	AM167881	Wikmark *et al.* 2006[14]
N. pringsheimi 1D	208	AM167882	Wikmark *et al.* 2006[14]
Naegleria sp. CL/I	197	DQ768715	Dykova *et al.* 2006[49]
Naegleria sp. BCZ4/I	199	DQ768716	Dykova *et al.* 2006[49]
Naegleria sp. SUM3V/I	216	DQ768723	Dykova *et al.* 2006[49]
Naegleria sp. GP3/III	197	DQ768724	Dykova *et al.* 2006[49]
Naegleria sp. NG163	198	AM497929	S. Johansen and C. Einvik, unpublished
Naegleria sp. NG332	209	AM497930	S. Johansen and C. Einvik, unpublished
Naegleria sp. NG358	200	AM167883	Wikmark *et al.* 2006[14]
Naegleria sp. NG393	207	AM167884	Wikmark *et al.* 2006[14]
Naegleria sp. NG458	200	AM497931	S. Johansen and C. Einvik, unpublished
Naegleria sp. NG491	207	AM497932	S. Johansen and C. Einvik, unpublished
Naegleria sp. NG498	200	AM167885	Wikmark *et al.* 2006[14]
Naegleria sp. NG647	207	AM167886	Wikmark *et al.* 2006[14]

[a] The GIR1 sizes correspond to positions 1506–1685 of the *D. iridis* sequence (AJ938153), or positions 227–424 of the *N. andersoni* sequence (Z16417).

gene conversion initiated by the intron encoded I-*Dir*I homing endonuclease.[16] A very different inheritance pattern is noted among the *Naegleria* twin-ribozyme introns.[14] Here, the intron was apparently gained early in evolution of the *Naegleria* genus and co-evolved along with its host rDNA in a strict vertical inheritance pattern. Subsequently, the intron was lost by sporadic deletions in approximately 70% of the isolates.

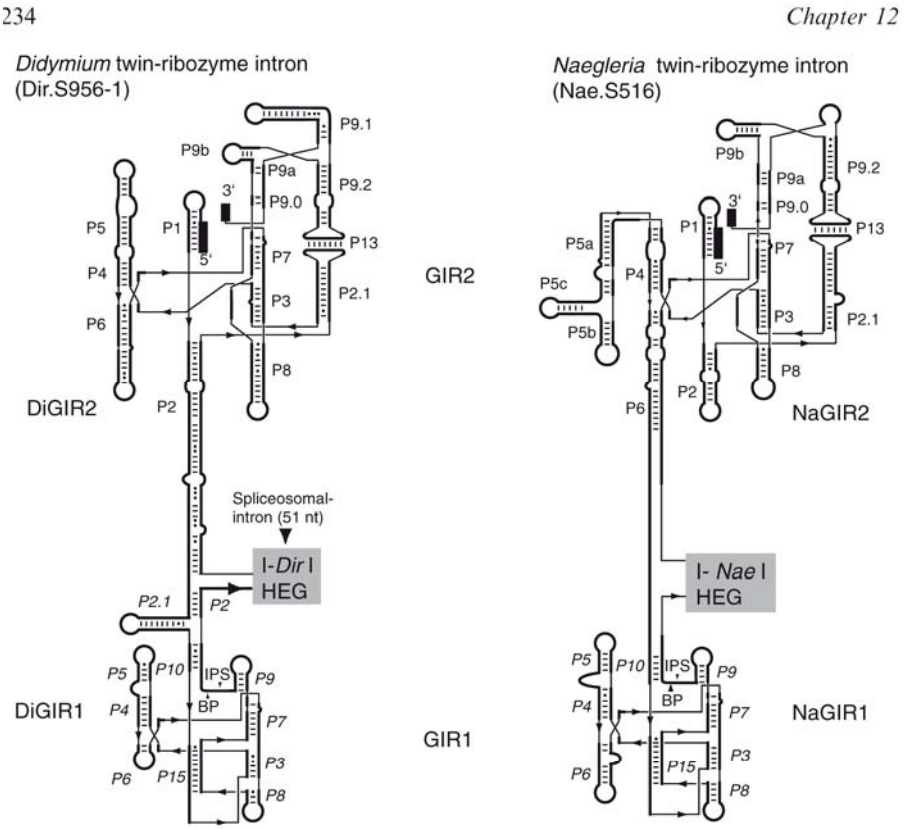


Figure 12.2 Structure diagrams showing the two known twin-ribozyme introns. GIR2 is the splicing ribozyme, GIR1 is the branching ribozyme, and HEG is a homing endonuclease gene. A spliceosomal intron is found in the I-DirI HEG in Didymium. BP: branch point. IPS: internal processing site. Exons are indicated by thick lines.

(iv) The splicing ribozyme domains (GIR2) of Dir.S956-1 and Nae.S516 represent two of the most distantly related nuclear group I intron subgroups known. Whereas the *Didymium* GIR2 has a typical IE intron fold, the *Naegleria* GIR2 represents the common IC1 intron subgroup (Figure 12.2). Structural differences between these two intron subgroups are observed in the organization of the P4-P6 region (Figure 12.2), as well as in sequence motifs flanking the P7 guanosine binding site.

(v) The GIR1-HEG insertions are located at different helical segments within the *Didymium* and *Naegleria* GIR2 ribozymes (Figure 12.2), namely P2 in DiGIR2 and P6 in NaGIR2.

(vi) The HEGs encode different homing endonucleases. Both the *Didymium* and *Naegleria* homing endonucleases (I-*Dir*I and I-*Nae*I, respectively) are members of the His-Cys box family, but they are distantly related in sequence and possess different target DNA specificities.[12,14,17] Interestingly, the I-*Dir*I HEG, but not the I-*Nae*I HEG, is interrupted by a

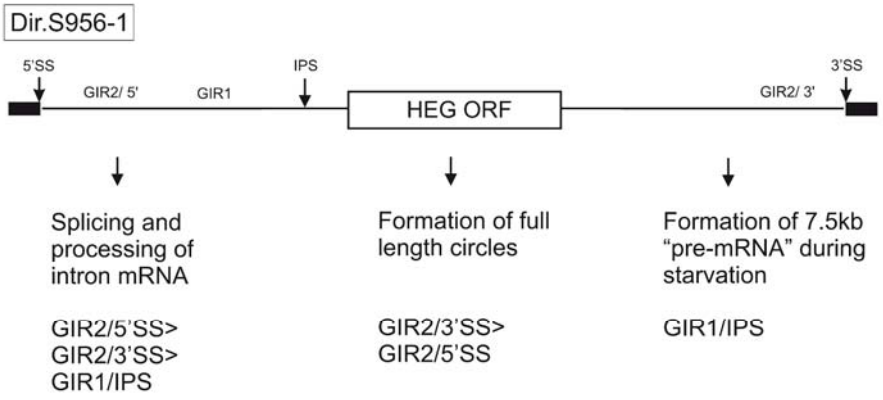


Figure 12.3 Ribozyme catalysed processing steps at the splice sites (SS) and the internal processing site (IPS) in three different processing pathways of the Dir.S956-1 intron. The order of the processing steps in each pathway is indicated. Black boxes: exons; Open box: The HEG open reading frame (ORF). Figure not drawn to scale.

small spliceosomal intron (Figure 12.2) that is removed during mRNA maturation in *D. iridis*.[18]

(vii) Finally, the *Didymium* and *Naegleria* GIR1 ribozymes have notable structural differences. Despite the fact that DiGIR1 and NaGIR1 have very similar secondary structure folds of the ribozyme core domain (Figure 12.3), important sequence differences are noted at flanking regions, in the P4-P6 domain as well as at several peripheral loop regions. However, comparisons of 29 natural *Naegleria* GIR1 variants (Table 12.1) confirm a high degree of conservation with only a very few variable nucleotide positions. Minor variations in size were noted at only four regions (J5/4, L6, L8, and L9).

12.3 Biological Context

Didymium iridis is a myxomycete that preys on other microorganisms and dead organic matter on the forest floor. It has a complex life-cycle that includes haploid amoebae, flagellates, and cysts, diploid amoebae, and a syncytial plasmodium that can differentiate into sporangia with haploid spores. Our studies have so far focused on the haploid life forms that are easily grown in liquid culture. Incidentally, these are the only life-cycle forms that have been observed in *Naegleria*. Obviously, given the complexity of both the organism and the twin-ribozyme intron itself, much can be learnt from studying the biological context of the GIR1 ribozyme. In this section we describe how our characterization of the molecular biology of the processing of the twin-ribozyme in *D. iridis* haploid life forms provides clues to understanding the structure and mechanism of the GIR1 ribozyme.

12.3.1 Three Processing Pathways of a Twin-ribozyme Intron

From a combination of *in vitro* and *in vivo* studies we have mapped three different processing pathways of the Dir.S956-1 intron that applies to different cellular conditions (Figure 12.3). The first pathway involves intron excision and exon ligation catalysed by GIR2, and subsequent processing of free intron to form the I-*Dir*I mRNA.[18] The formation of the I-*Dir*I mRNA involves the GIR1 activity and is detailed in the next section. The order of the activity of the ribozymes is (1) GIR2 at the 5′ splice site (SS), (2) GIR2 at the 3′SS and (3) GIR1. This pathway benefits both the host and the intron, and is the dominant pathway during the growth phase of haploid and flagellates.

The second pathway results in the formation of full-length circular introns and un-ligated exons. This is a general pathway in nuclear group I introns.[19] The order of reactivity is (1) hydrolytic cleavage at the 3′SS catalysed by GIR2 and (2) transesterification at the 5′SS catalysed by GIR2. GIR1 is not active in this pathway and circle re-opening by GIR2 seems to be required for activation of the GIR1 activity. This pathway benefits the intron at the expense of the host. The biological significance of the circularization pathway is not clear, but the circular introns are frequently suggested to be involved in the spreading of the intron. Interestingly, GIR2 adopts a particular conformation in the circles, and circularization is responsive to the cellular conditions (unpublished).

The third pathway is induced by starvation induced encystment.[20] *D. iridi*s undergoes frequent rounds of encystment and excystment in nature, depending on fluctuations in environmental conditions. During encystment, the rRNA precursor is processed into a 7.5 kb RNA product that accumulates within the cells to become the dominant intron containing molecular species. The processing is accomplished by the branching activity of GIR1 without prior GIR2 activity. The biological function is unknown but the current speculation is that the 7.5 kb RNA is stored as a precursor that will allow I-*Dir*I expression in the absence of rRNA expression. The pathway is at the expense of the host in the sense that functional rRNA is not being produced but it could also be viewed as a mechanism to down-regulate rRNA expression during starvation.

From the observation of the three different pathways and the order of activity of the ribozyme activities, it follows that the two ribozymes are regulated with respect to each other. It is currently not known if this regulation involves protein factors, RNA–RNA interactions within the twin-ribozyme intron, or both. The regulation, furthermore, implies that both ribozymes can fold into an inactive, yet biologically significant conformation.

12.3.2 Processing of the I-*Dir*I mRNA

Several examples are known of nuclear protein-coding genes embedded in rDNA. These genes are transcribed as an integral part of an RNA polymerase I transcript. Normally protein-coding genes are transcribed by RNA polymerase II and their expression is facilitated by co- and post-transcriptional modifications that are specific for RNA polymerase II transcription. This raises the

problem of how the protein-encoded genes within rDNA are brought on to the RNA polymerase II pathway A recent survey of the best described examples[21] concluded that several different strategies are being used and that those employed by the Dir.S956-1 intron in expression of the I-*Dir*I homing endonuclease rank among the most sophisticated.

The expression pathway is initiated by the splicing out of the intron by conventional group I intron splicing (Figure 12.4). Then, the 5′ end of the mRNA is formed by cleavage catalysed by the immediate upstream GIR1 ribozyme. The cleavage by branching provides the mRNA with a lariat cap as a substitution of the conventional m^7G cap. The 3′ end is then formed by cleavage at site referred to as IPS3. This poorly characterized cleavage reaction does not occur in isolated RNA and is thus believed to depend on a host factor.[18] Following this, the mRNA is further processed by cleavage and polyadenylation at a *bona fide* polyA signal.[18] Next, a short (51 nt) spliceosomal intron is spliced out.[18] The spliceosomal intron harbours the conventional splice signals but the splicing mechanism has not been studied. Similar short spliceosomal introns have been found in other homing endonuclease genes encoded within group I introns.[13] It is possible that the acquisition of the splicing and polyadenylation signals by the intron serves to recruit protein factors that help guide the mRNA onto the polymerase II pathway.

The lariat capped, spliced, and polyadenylated form of the mRNA is the only molecular species derived from the intron that is found in the cytoplasm.[18] The mRNA becomes associated with ribosomes[18] but the protein product has not been directly demonstrated. However, its activity has been shown to be present by enzymatic assays in cellular extracts,[22] and by demonstration of endonuclease intron homing in genetic crosses.[16] The lariat cap has indirectly been shown to be required for expression of the protein in yeast.[22] It substitutes for the conventional cap in protection of the mRNA against 5′ exonucleases and perhaps it even plays a role in recruitment of translation initiation factors.

12.3.3 Conformational Switching in GIR1

Group I introns are generally believed to fold directly into the active conformation *in vivo*, at least when the flanking exons are correctly folded. However, if GIR1 is folded in a similar way, this would result in cleavage of the ribosomal precursor and be detrimental to the cell. DiGIR1 and the I-*Dir*I HEG are inserted into P2 of DiGIR2. This positions DiGIR1 towards the 5′ end of the 1436 nt Dir.S956-1 intron. If GIR1 cleaves *in vivo* at a rate that is comparable to the *Tetrahymena* intron self-splicing ($t_{1/2}$ of *ca.* 2 s;[23]) there is ample time to fold into an active conformation and cleave before transcription and folding of GIR2 is completed. GIR1 will thus most likely fold initially into an inactive conformation. In fact, some key components of an inactive GIR1 conformation have been identified, primarily from structure probing experiments. One is a hairpin formed by nucleotides 235–266 immediately downstream of the IPS. The formation of this hairpin (HEG P1) is co-transcriptionally favoured and

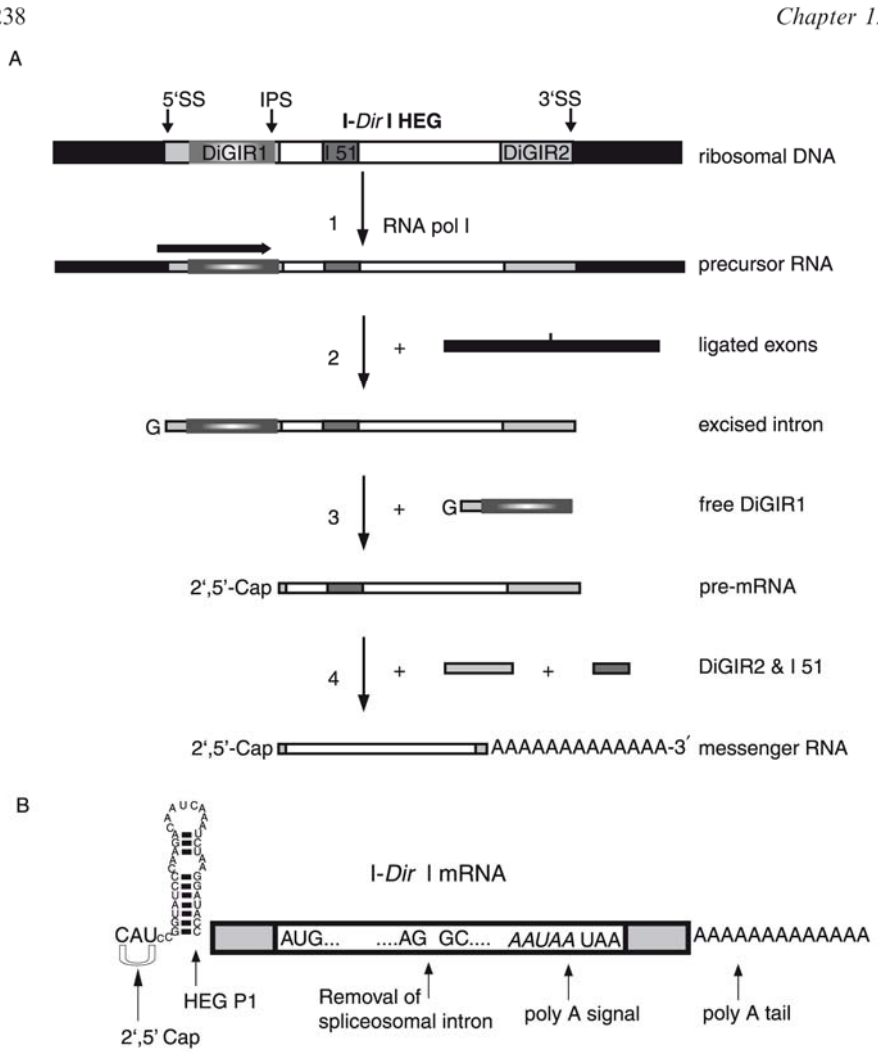


Figure 12.4 (*A*) Formation of the homing endonuclease I-DirI mRNA by processing of the Dir.S956-1 intron. The homing endonuclease is coded by the sense strand of the intron and is transcribed by RNA pol I as part of the rRNA precursor (1). The intron is spliced out and the exons are ligated by the GIR2 ribozyme (2). The spliced out intron is cleaved by the GIR1 branching ribozyme, leaving the free ribozyme and a downstream pre-mRNA equipped with a lariat cap (3). Finally, in a series of steps depending on host factors, the mRNA becomes polyadenylated, and a small spliceosomal intron is spliced out (4). (*B*) Detailed structure of the I-DirI mRNA, including the lariat cap and HEG P1 hairpin found in the short 5′-UTR. (Figure 12.4B is reproduced from ref. 3 with permission.)

precludes the formation of the catalytically active conformation (Figure 12.5). Incidentally, this hairpin is also found in the mature I-*Dir*I mRNA (Figure 12.4B). The conformational switching mechanism between P10-P2 of the catalytically active conformation and HEG P1 take place immediately

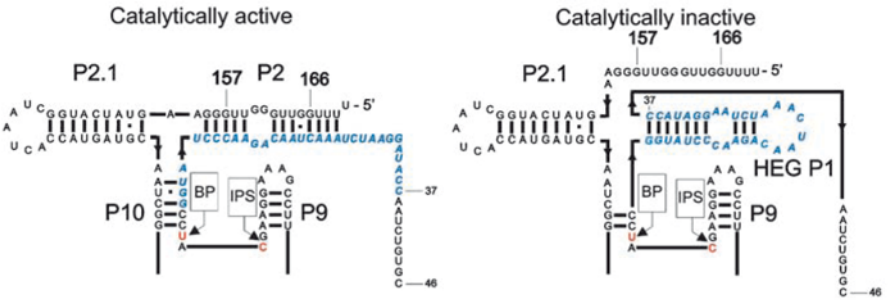


Figure 12.5 Alternative secondary structures of the part of DiGIR1 involved in conformational switching between catalytically active (including P2 and P10) and inactive (including HEG P1) conformations. The nucleotides involved in alternative pairings are in bold, blue italics. The nucleotides involved in formation of the 2′,5′ phosphodiester bond (C230 and U232) are in red. BP: branch point; IPS: internal processing site. The numbers indicate the distance in nt from the IPS.

following the branching reaction, where the formation of HEG P1 inhibits the reversal of the reaction and provides the I-*Dir*I mRNA with a pre-folded 5′-UTR.[24]

12.4 Biochemical Characterization

Characterization of a ribozyme often starts with the problem of delimitation of the functional unit. This is because many ribozymes are found as an integral part of a larger RNA molecule that is impractical to study. From looking at the structure diagrams of twin-ribozymes (Figure 12.2), isolation of a GIR1 functional unit appears to be straightforward. In the *Didymium* and *Naegleria* ribozymes, GIR1 is easily identified as a separate domain (Figure 12.2) presented on extended helices P2 and P6, respectively. This observation proved to be deceptive, mainly because of a misinterpretation of the activity of the ribozyme. The *Didymium* ribozyme was originally found to cleave at a single site (IPS1) by primer extension analysis of a construct carrying a deletion of most of the ORF and the 3′-part of GIR2.[2] Subsequently it has been shown that the full-length intron *in vitro* similarly is cleaved at a single site and with a relatively low efficiency.[24] Narrowing down by deletion studies to define a minimal version of GIR1 that could be studied in greater detail concluded that the 162.65 variant best reflected the functional unit.[10] This variant is cleaved at the same site as the full-length intron, but an additional stop site was mapped by primer extension analysis three nucleotides further downstream (IPS2). Incidentally, this is the only primer extension site observed in analysis of cellular RNA.[15,18,20] Subcloning of the *Naegleria* ribozyme (NanGIR1) led to a variant with 267 nt upstream and *ca.* 50 nt downstream of the IPS that showed the same cleavage pattern as the *Didymium* ribozyme.[11] This was further

narrowed down to 178.19 by Jabri *et al.*[25] who found this variant to cleave at a high rate compared to several other length variant but with cleavage only at IPS1, as in the case of the full-length intron. At this stage, the cleavage at the processing site (IPS1) was demonstrated to leave a 3′OH and a 5′ phosphate, as would be expected by hydrolytic cleavage.[25] The primer extension stop at IPS2 was not characterized, but it was inferred that this cleavage was hydrolytic as well.[10] In addition, it was concluded from a concatenation RT-PCR approach that the two cleavages occurred in an obligatory sequential manner.[10] The interpretation of the IPS2 primer extension stop proved to be incorrect (see below) and in this case it turned out that "less is more" in the sense that the full-length intron masks the fact that the reaction is actually highly reversible *in vitro*.

12.4.1 GIR1 Catalyzes Three Different Reactions

DiGIR1 catalyses three different reactions. The natural reaction is the branching reaction (1A in Figure 12.6) in which a transesterification at the IPS results in the cleavage of the RNA with a 3′OH and a downstream lariat cap made by joining of the first and the third nucleotide by a 2′,5′ phosphodiester bond.[3] These products are the only products observed by analysis of cellular RNA.[18,20] *In vitro*, DiGIR1 catalyses the reverse reaction (1B), referred to as the ligation reaction. It is very efficient to the extent that the forward reaction is completely masked in reactions with full-length intron and length variants that include more than 166 nucleotides upstream of the IPS.[3,24] Finally, DiGIR1 catalyses hydrolytic cleavage at the IPS (2A) at a relatively low rate. This is the cleavage reaction observed with the full-length intron and several length variants.[24] The hydrolytic cleavage is irreversible and is considered an *in vitro* artefact resulting from a failure to present the branch nucleotide (BP) correctly for catalysis. NaGIR1 catalyses the same reactions as DiGIR1. Specifically, several length variants of NanGIR1, the smallest being 178.28, have been shown to catalyse branching (unpublished).

The three reactions can be experimentally separated. The branching reaction is isolated from the reverse reaction by cleavage in the presence of 2 M urea. This inhibits ligation completely and the contribution from hydrolytic cleavage at these conditions is negligible.[24] Ligation is studied under acidic conditions (*e.g.* pH 5.5) at which the forward reaction is inhibited.[24] Finally, the hydrolysis reaction is the only reaction observed at standard conditions with certain length

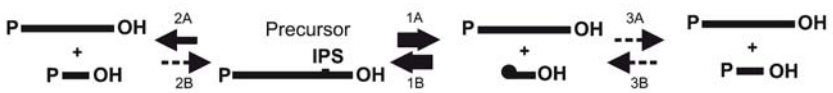


Figure 12.6 Reactions known to be catalysed by GIR1 as well as hypothetical reactions that have not been observed (dashed arrows). The main activity is the branching activity (1A) that is highly reversible *in vitro* (1B). A hydrolytic cleavage reaction (2A) is less pronounced and only observed *in vitro*.

variants. This is not a true isolation of the reaction because the major fraction of the molecules apparently is engaged in multiple rounds of branching and ligation reactions. The reaction rates vary considerably among the length variants. Analysed as described above, the 166.22 variant of DiGIR1 performs branching at a rate of $0.085\,min^{-1}$, ligation at a rate approaching $1\,min^{-1}$, and hydrolysis at $0.01\,min^{-1}$.[24] The branching rate is only one order of magnitude less than the cleavage rates of most optimized minimal cleavage ribozymes (*e.g.* 0.2–0.5 min^{-1} for the hairpin,[26] $1.0\,min^{-1}$ for the VS,[27] and 0.5–2.0 min^{-1} for the hammerhead ribozyme[28]).

12.4.2 Characterization of the Branching Reaction

The branching reaction is initiated by a nucleophilic attack involving the 2′OH of U232 at the phosphate of C230 (Figure 12.1). This was demonstrated in a trans-cleavage experiment using a ribozyme that was truncated at A222 in L9 (7 nt upstream of the IPS) and a substrate carrying the missing 7 nt followed by 22 nt downstream of the IPS.[3] Deoxy-substitutions were introduced in the substrate at positions corresponding to C230, A231, U232, and C233. Only deoxy-substitution of the 2′OH of U232 completely prevented the branching reaction. A strong effect of deoxy-substitution at A231 was ascribed to a structural effect (Section 12.5). Detailed characterization of the reaction mechanism is still in its initial phase.

The characteristic product of the branching reaction, the lariat cap, was first deduced from indirect analysis, such as primer extension, resistance towards degradation by enzymes and alkali. Subsequently, the lariat was sequenced by enzymatic degradation and analysis by thin-layer chromatography (TLC).[3] Curiously, the structure of the lariat cap was already known from the literature. Several small lariats, including a 4 nt lariat, were analysed by nuclear magnetic resonance (NMR) imaging by Agbäck *et al.*[29] in a study attempting to describe the lariat products from spliceosomal splicing. The 4 nt lariat was found to have an unusual structure with the lariat ring locked in a rigid South-type conformation.

12.4.3 Biochemistry of GIR1

Optimization of the GIR1 reaction is mainly due to the early work of Jabri *et al.* on the minimal version (178.19) of *Naegleria andersoni* GIR1 (NanGIR1).[25] Notably, this ribozyme was reported to exclusively cleave by hydrolysis at the IPS and thus the branching reaction may not have been optimized. The basic setup for GIR1 cleavage studies is to pre-incubate the *in vitro* transcribed RNA in cleavage buffer at pH 5.5 (10 mM cacodylate or 10 mM acetate) for 5–10 min and then start the reaction using a pH jump by addition of cleavage buffer containing Hepes-KOH at pH 7.5. This setup eliminates a lag phase required for folding of the RNA and permits kinetic analysis. Based on this, the optimal conditions for NanGIR1 were found to be 1 M KCl, 25 mM $MgCl_2$ at 45 °C.

$CaCl_2$ did not substitute for $MgCl_2$, but $MnCl_2$ could replace $MgCl_2$ and was even found to be more effective at lower concentrations. Cleavage in NaCl was less efficient than with KCl and LiCl inhibited the reaction. The polyamines spermine and spermidine could not replace the monovalent ion at concentrations used in group I introns. The activity of the NanGIR1 ribozyme under the above conditions increased linearly with temperature between 30 and 45 °C, and reduced again above 47 °C. The reaction rate was found to be first order in hydroxide ion concentration independent of the type of buffer in the range pH 4–8.5. *In vitro* selection was employed to select for NanGIR1 variants with a faster hydrolysis rate and less salt dependence. Variants were characterized that cleaved at 300-fold greater rates in 100 mM KCl.[30] A systematic analysis of the requirement for the *Didymium* ribozyme, and in particular in relation to the branching activity, has not been performed. Generally speaking, the conditions found by Jabri *et al*.[25] appear to apply to the *Didymium* ribozymes as well, but the mutations that relieved the NanGIR1 ribozyme of the high salt requirement did not have a similar effect in the DiGIR1 context.[1]

12.5 Modelling the Structure of GIR1

Most of the base-pairing scheme of GIR1 was based on the close similarity to group I introns (Chapter 10). In the first published 3D model,[10] known features of group I introns were supplemented with *in vitro* mutagenesis structure probing data. The most notable observations were the lack of P1, the presence of a novel P15 pseudoknot, and the unusual structure of J5/4. The P2-P2.1 domain was not included in the original model but was subsequently verified.[31] Parallel work with the *Naegleria* ribozyme corroborated the overall base-pairing scheme, except that P2.1 is absent in this species.[11,25]

The publication of an X-ray crystal structure at 3.1 Å resolution of the *Azoarcus* sp. $tRNA^{Ile}$ intron (*Azo*)[32,33] prompted a revision of the base-pairing scheme of DiGIR1. The *Azo* intron was crystallized in a version that represents the second step of splicing, the structure that most closely resembles DiGIR1. The similarity suggested an extended P15 corresponding to the P1-P2 stack in *Azo* at the second step of splicing. The next development was the unexpected finding that the reaction catalysed by GIR1 is a branching reaction rather than sequential hydrolytic cleavages.[3] The unique product of the reaction, a 4 nt lariat, had previously been characterized by NMR in Chattopadhyaya's group in an attempt to describe the structure of the lariat resulting from spliceosomal splicing.[29] With this at hand, we could construct a model[34] that incorporated features from recent X-ray crystal structures of group I introns, including the *Azo*,[32,33] *Tetrahymena thermophila* (*Tet*),[35] and the *Staphylococcus aureus* bacteriophage *Twort* (*Two*)[36] introns, together with the unique features of GIR1 like the P15 pseudoknot, the lariat fold and a type of three-way junction (3WJ) originally found in rRNA.[37] Two different base-pairing schemes were considered for the GU pair at the active site. Based on mutation analysis, G109:U207 was favoured over G109:U232 (Section 12.5.4). The P2-P2.1

domain presented an additional problem because the docking of this domain onto the remainder of the structure was not evident. The P2-P2.1 domain was thus excluded from the model.

12.5.1 Overall Structure

The model of the DiGIR1 is compact with three aligned helical stacks ($85 \times 52 \times 40$ Å without the P2-P2.1 domain). The P3-P7-P8-P9 domain (hereafter P3-P9 domain) and P10-P15 domains run in parallel and the P4-P5-P6 domain (hereafter P4-P6 domain) is slightly tilted with the L5 end pointing away from the rest (Figure 12.7). This helical arrangement is roughly similar to that of *Azo*.[33] The helical organization in DiGIR1 is mainly brought about by two structural features. First, the 3′ strand of P2 (G199-C204) has apparently been fused with nucleotides from J8/7 (A205-U207), thus forming the P15 pseudoknot. Since P3 is already embedded in the pseudoknot characteristic of group I introns, the resulting structure is a double pseudoknot. The formation of P15 shortens the canonical group I intron J8/7 from 7 (IC1 introns) or 6 nt (all other subgroups) to only 3 nt and in this way tightens the structure. Second, J15/3 organizes the three-way junction between P15, P3 and P8, resulting in a side-by-side parallel orientation of the P3-P9 and P10-P15 domains (Figure 12.8A).

12.5.2 Coaxially Stacked Helices

The P4-P6 domain in DiGIR1 is quite small. In the *Tet* intron this domain folds at an early stage and functions as a scaffold that facilitates the folding of the core.[38,39] This is unlikely to be the case in DiGIR1 and is consistent with the notion that DiGIR1 initially folds into an inactive conformation from which it is activated (Section 12.3.3). Substitution of L5 with a UUCG tetraloop has only a moderate effect on the cleavage kinetics, whereas substitution of the two As in L6 has a more pronounced effect (unpublished), probably due to an interaction with P3. Compared to group I introns, the interface between P5 and P4 is quite different in that J4/5 is missing. J5/4 and the architecture around it, including the bulged U156, are critical for the activity (Section 12.5.4). The *Naegleria* ribozyme differs considerably in the structure of the P4-P6 domain. First, P6 is extended and includes an internal loop that makes the tertiary interaction with P3 similar to what is known from the *Azo* intron. As a consequence, L6 is not involved in this tertiary interaction and the sequence is variable. Second, J5/4 is highly variable among *Naegleria* GIR1's, in many cases including an additional helix (P5.1) in some sequence variants.

The P3-P9 domain contains a P7 guanosine binding site that is highly conserved among all group I introns. Here, an exogenous guanosine factor (exoG) binds during the first step of splicing and the intron terminal guanosine (ωG) binds during the second step of splicing and during the first step of the circularization pathway. GIR1 P7 binds only G229, the equivalent of ωG.

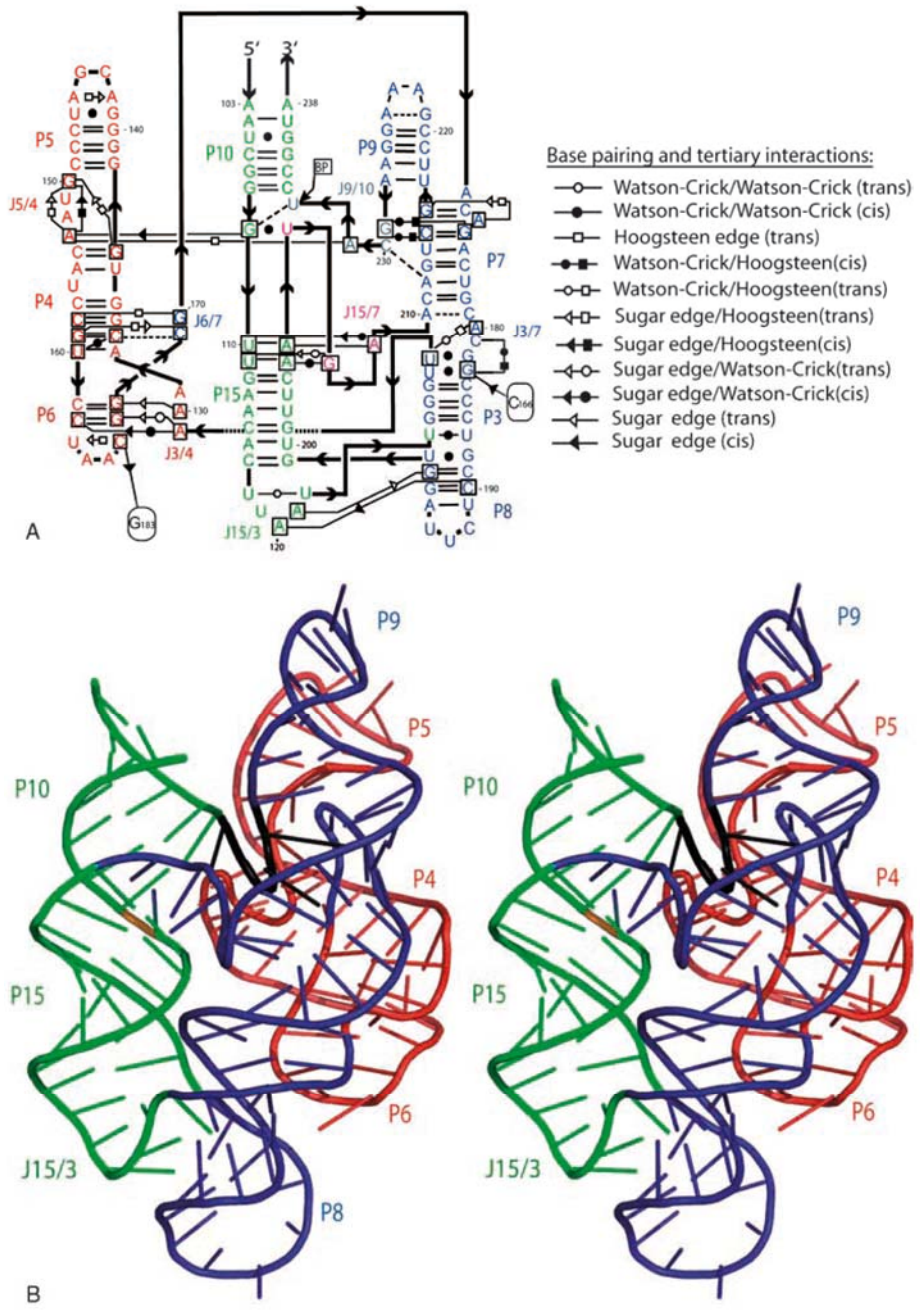


Figure 12.7 Model of the structure of the GIR1 ribozyme. (*A*) Secondary structure and proposed tertiary interactions of DiGIR1 showing Watson–Crick as well as non-Watson–Crick base-pairings. (*B*) 3D model of the overall structure of DiGIR1.[34] The paired stems (P) and one of the joining segments (J) are numbered.

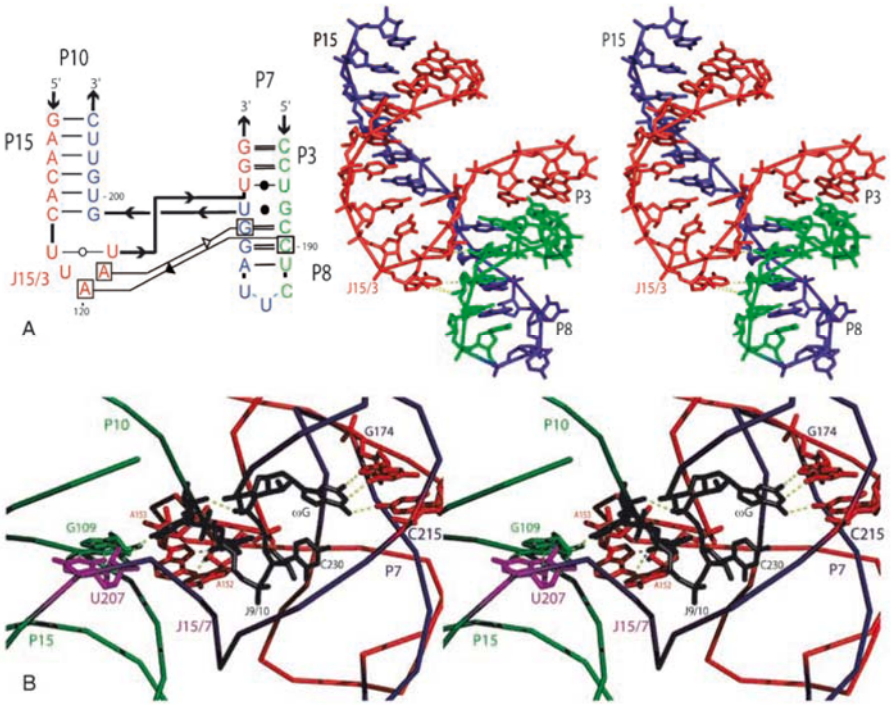


Figure 12.8 Details from the structural model of DiGIR1. (*A*) Base pairing in the three-way junction J15/3 and 3D model.[34] The boxed nucleotides are A120, A121, C190, and G197. (*B*) Modelled structure of the active site with the key nucleotides indicated. Colouring corresponds to that in Figure 12.7.

Addition of GTP to a GIR1 cleavage reaction has no effect (unpublished), suggesting that added guanosine is unable to compete with G229 for binding to P7. P9 stacks upon P7 without any intervening nucleotides. Both the length and the sequence of the GAAA tetraloop that caps P9 are important for activity (unpublished), but the tetraloop receptor remains elusive. Experiments and modelling have excluded a receptor in the P4-P6 domain. Rather, we favour an interaction within the P2-P2.1 domain. P2 is involved in a conformational switch (Section 12.3) that results in the formation of the alternative HEG P1, and a receptor for L9 has been suggested in this structure (unpublished). However, this does not rule out the existence of an alternative receptor in the catalytically active conformation. NaGIR1 differs from DiGIR1 in that L9 is a 7–13 nt loop. Part of the loop sequence is complementary to the sequence downstream P2″, an interaction that could very well be a functional replacement of the L9:HEG P1 interaction in DiGIR1.

The P10-P15 domain consists of a 5 base-pair P10 stacked directly upon P15. A 6 base-pair P15 was proposed in our original model, but the analogy to *Azo* suggested an extension by 3 base-pair that was subsequently supported by

mutagenesis analysis of the U110-A206 base-pair. The structure of this domain is very similar in DiGIR1 and NaGIR1. In fact, both P10 and P15 from *Naegleria* function in the DiGIR1 context[10] (unpublished). The P10-P15 stack is equivalent to the P10-P1-P2 stack formed at the second step of group I introns. Thus, the apparent absence of a P1 from the GIR1 ribozyme is related to the fact that GIR1 resembles a group I intron at the second, not the first, step (Section 12.6).

The P2-P2.1 top domain, not incorporated into the current model, forms a three-way junction with P10. P2 is the connection to the splicing group I intron component (GIR2) of the twin-ribozyme introns (Section 12.2). NaGIR1 has a loop of unknown structure in place of P2.1. One possible function of P2-P2.1 in DiGIR1 is to bridge the P4-P6 and P3-P9 domains and in this way stabilize the folding of the core in a manner similar to the direct interactions between these two domains known in group I introns.

12.5.3 Junctions and Tertiary Interactions Involving Peripheral Elements

The junctions between the P4-P6 and P9-P3 domains resemble those seen in group I introns. The interface between the two domains is modelled with the involvement of base-triple interactions between J3/4 and J6/7 in the shallow/minor groove of P6 and the deep/major groove of P4, respectively. J15/3 organizes a three-way junction of family C[37] (Figure 12.8A). U118 base-pairs with U122, and U123 base-pairs with U187 (P3). A120 and A121 form A-minor interactions in the minor groove of P8. Two other junctions, J15/7 and J9/10, are part of the active site and are discussed below.

Group I introns usually have peripheral structural elements that branch out from the core. These elements interact and contribute stabilization of the core but are few in GIR1, at least in the model that does not involve the P2-P2.1 domain. The two As in the L6 tetraloop make A-minor interactions in the shallow/minor groove of two consecutive G=C pairs in P3. In the *Naegleria* ribozyme the interaction is between As in J6/6a and P3, as observed recurrently in group I introns. As mentioned in Section 12.5.2, L9 is a candidate for an additional peripheral interaction, but an interaction partner has not been identified.

12.5.4 The Active Site

The key residues for the branching reaction interact in a pocket formed at the interface of P10, P7, J5/4, and J9/10 (Figure 12.8B). The architecture of the guanosine binding site is similar to that of group I introns (Chapter 10) and all substitutions of G229 are inactive.[1] A GU wobble pair is presented at the interface of P10 and P15 equivalent to the position at the interface of P10 and P1 at the second step in group I introns. In our original model, G109 forms a GU wobble base-pair with U232. Substitutions of U232 reduce cleavage by a

factor of 3–4 but are still permissive for the branching reaction.[34] In contrast, mutations of G109 and U207 both result in poor branching in addition to a 2–4-fold reduction in the cleavage rate.[34] The G109:U207 pairing was thus favoured in the model. The GU wobble in *Azo* is recognized by the wobble receptor motif in the symmetric J5/4 internal loop. Here, A58 and A87 belonging to two consecutive AA pairs are involved in an intricate hydrogen bonding network and recognize the ribose 2′OH and the exocyclic amine, respectively, of the G in the GU pair. In DiGIR1, J4/5 is lacking, and the recognition of the GU pair by J5/4 appears to be quite different. A231 of the lariat fold appears oriented towards A153 of J5/4 due to hydrogen bonding between the 2′OH and a phosphate oxygen at C230. These two adenosines are among the most critical residues in DiGIR1 and are proposed to form a *trans*-Hoogsteen-Hoogsteen pair. The pairing places A153 in a position to hydrogen bond to the N2 of G109. The nucleotide carrying the nucleophile (U232) lies in the shallow/minor groove of the GU pair and is close enough to hydrogen bond to the N2 amino group of G109. Perhaps this interaction can explain the preference for a U at this position. Mutational analysis in NaGIR1[30] indicates that GU recognition in this ribozyme could be slightly different, perhaps related to the variation in the structure of J5/4.

J8/7 is a highly conserved sequence among group I introns. In *Azo*, it pirouettes through the active site, making contacts to all three helical domains as well as participating in the binding of structural and catalytic metal ions.[33] The sequence of J8/7 is clearly recognized in GIR1 (4/6 nucleotides are conserved), but its organization appears dramatically altered. The first three nucleotides are involved in P15 base-pairings. Thus, the recognition of P1-P2 in *Azo* by the three 5′ nucleotides in J8/7 has been replaced by participation in the analogous stem structure by contributing base-paired residues. The third base-pair (involving U207) is a GU wobble base-pair that replaces the GU found in P1 in *Azo* at almost the same position. In *Azo*, the next three nucleotides make contacts with J7/3, P4, and P7, respectively.[33] However, in GIR1 the base-pairing of the 5′ part of J8/7 restricts the course of the backbone and the next two nucleotides (G208 and A209) contact the shallow/minor groove of P15. Then, the last nucleotide of J8/7 (A210) mimics the fourth nucleotide in *Azo* (G170) in the sense that it makes the same hydrogen bonding and stacking interaction with J7/3. This leaves GIR1 short of the two nucleotides (C171 and A172) that contact P4 and P7. One interesting possibility is that the two added nucleotides from the lariat fold (C230 and A231) might take on some of these functions. At least, C230 can be stacked under the guanosine binding site in P7 in the same way that A172 is stacked in *Azo*. In *Azo*, J8/7 is critical in the binding of both of the catalytic metal ions. A particularly well-studied aspect of the catalytic site of the *Azoarcus* intron is the coordination of catalytic metal ions.[40] The structural differences observed in GIR1 could influence metal ion binding but this has not yet been addressed.

In the active site, the nucleotide carrying the nucleophilic 2′OH (U232) is unpaired and occupies a pocket that will accommodate any base. This is consistent with *in vitro* mutagenesis results that show that branching can occur

with any nucleotide at this position (unpublished). In the *wt* sequence, a single hydrogen bond between U232 and A231 bond could help orient the 2′OH for nucleophilic attack. The distance U232 O2′–C230 O1P in the model is 2.80 Å, and in agreement with the proposed mechanism.

12.6 Phylogenetic Considerations

The similarities between GIR1 and *Azo* suggest a common ancestor or, alternatively, that one of the ribozymes was derived from the other. As previously noted, GIR1 may have originated from eubacterial group IC3 introns.[1] In this section we discuss the possibility that GIR1 originated by a misaligned reverse splicing event and that this was followed by structural transformations. Reversal of the splicing reaction is a well-established reaction suggested to be in part responsible for the observed phylogenetic distribution of group I introns.[41] G229 in GIR1 is equivalent to the ωG residue of a group I intron. The sequence that follows, 5′-CAU, is identical to the sequence of the 5′-exon immediately upstream of the *Azo* intron. In fact, this is the anticodon sequence of the $tRNA^{Ile}$ in which the *Azo* intron resides. The suggestion is that GIR1 originated by reverse splicing of a eubacterial tRNA-like intron into a tRNA-like molecule in a situation where the exons were misaligned by three nucleotides. This resulted in reverse splicing on the 5′ side of the anticodon instead of the 3′ side. As a consequence, the intron became linked at its 3′-end (G229) to three nucleotides of 5′-exon (5′-C230AU) followed by the 3′-exon (C233 and onwards). The inclusion of the three nucleotides of 5′-exon is postulated to have blocked any further steps in reverse splicing, thus trapping the intron in a conformation that is reminiscent of a group I intron prior to the second step of splicing. We now consider several predictions that follows from this evolutionary model.

The first prediction is that GIR1 has lost the features characteristic of the first step of splicing. Here, a P1 hairpin including the 5′ splice site is absent. Perhaps related to this, GIR1 is unable to bind the cofactor of the first step, GTP, well enough to compete with binding of the ωG analogue, G229. In *Azo*, an A residue links P7 and P9. This nucleotide has been hypothesized to sequester ωG in a G-A pair during the first step of splicing.[42] Consistent with the idea that such a function is redundant in GIR1, P7 and P9 are linked without intervening nucleotides.

The second prediction is a loss of conformational flexibility in GIR1 related to the conformational switch that occurs in *Azo* between the first and the second step of splicing. There are two important examples in GIR1 of replacement of tertiary interactions with base-pairing. This results in a loss in flexibility in the sense that the interaction is a one-step rather than a two-step interaction. The tethering of the (P10)-P1-P2 helical stack to the P3-P9 domain by an L2 tetraloop: P8 tetraloop receptor interaction in *Azo* is replaced by the lower 6 base-pairs of the P15 pseudoknot in GIR1. The docking of P1 into the core of a group I ribozyme is mediated by a minor groove interaction by the three

5′ nucleotides of J8/7. This interaction has to accommodate both the P1 hairpin at the first step, and the lower part of the hairpin sandwiched between P10 and P2 at the second step. This interaction is replaced by the three upper base-pairs in GIR1 P15.

The third prediction is that GIR1 has undergone a structural reduction while maintaining the essentials of the catalytic core. The IC3 introns present in eubacterial tRNA rank among the smallest group I introns known. GIR1, and in particular DiGIR1, is even smaller (Table 12.1). The integrations of P1, P2, and parts of J8/7 into the P15 pseudoknot are characteristics of the GIR1 ribozyme. In DiGIR1, the reduction in size of the P4-P6 domain is another prominent feature. The internal loop that is involved in a minor groove interaction with P3 is lost and the interaction is replaced by a similar interaction using nucleotides in L6. The reduction in size of the P4-P6 domain correlates with the loss of the function of this domain as a scaffolding domain during folding. The preservation of the core is mainly seen in the conservation of the P7 architecture and the maintenance of an interaction between J5/4 and a GU wobble base-pair. A final consequence of the proposed origin of GIR1 is that the flanking sequences could have retained tRNA features. In this view, P10 is regarded as a fusion with the anticodon stem and P2.1 as the D arm. Further similarities in the 3′ flanking sequences that depend on alternative folding are currently being analysed.

Perhaps the most stunning observation based on the model is that the topological differences between the GIR1 and the group I intron catalytic cores can be explained by relatively simple alterations at the sequence level. For example, a single transposition of six nucleotides (5′ GUGUUC) from P15″ of GIR1 to a position within J15/3 would shuffle some of the critical junctions within the core and alter the structural organization of J8/7 at the RNA level. The resulting topology would be that of a group I intron. Such a transposition event can be incorporated into the reverse splicing model or can it be envisaged without such an event.[34] Importantly, the feasibility of these evolutionary models can be experimentally addressed.

12.7 Concluding Remarks

The GIR1 branching ribozyme shows a clear structural resemblance to group I introns. In terms of the reaction catalysed, it is more reminiscent of the group II introns (Chapter 11) (or spliceosomal introns; Chapter 13). However, the catalytic core appears to constitute a unique fold and the product, although essentially being a branched RNA, appears to be unique in structure. For these reasons, we suggest that the GIR1 branching ribozyme is listed as an independent class of naturally occurring ribozymes, as originally suggested by Cech and Golden.[43] Besides being an addition to the shortlist of naturally occurring ribozymes, GIR1 is a new example of 2′,5′ phosphodiester bond-forming activity, of which around ten are known.[44] Compared with most other ribozymes, the characterization of the structure and reaction mechanism of GIR1 is

still in its initial phase. In contrast, a substantial amount of information of the biological context has been gathered. A particularly interesting aspect of GIR1 is that a plausible hypothesis can be made on its evolutionary history. There are very few examples of this, the most recent being the suggestion that the HDV ribozyme originated from the human transcriptome.[45] The proposed development of GIR1 is more radical because it involves a structural and a functional derivation. We suggest that two events, a misaligned reverse splicing, and a transposition of six nucleotides are sufficient to account for the derivation of the basic GIR1 structure from a group I intron.

References

1. S. Johansen, C. Einvik and H. Nielsen, DiGIR1 and NaGIR1: Naturally occurring group I-like ribozymes with unique core organization and evolved biological role, *Biochimie*, 2002, **84**, 905–912.
2. S. Johansen and V.M. Vogt, An intron in the nuclear ribosomal DNA of *Didymium iridis* codes for a group I ribozyme and a novel ribozyme that cooperate in self-splicing, *Cell*, 1994, **76**, 725–734.
3. H. Nielsen, E. Westhof and S. Johansen, An mRNA is capped by a 2′, 5′ lariat catalyzed by a group I-like ribozyme, *Science*, 2005, **309**, 1584–1587.
4. W.A. Decatur, C. Einvik, S. Johansen and V.M. Vogt, Two group I ribozymes with different functions in a nuclear rDNA intron, *EMBO J*, 1995, **14**, 4558–4568.
5. S. Johansen and P. Haugen, A new nomenclature of group I introns in ribosomal DNA, *RNA*, 2001, **7**, 935–936.
6. F. Michel and E. Westhof, Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis, *J. Mol. Biol.*, 1990, **216**, 585–610.
7. S.O. Suh, K.G. Jones and M. Blackwell, A group I intron in the nuclear small subunit rRNA gene of *Cryptendoxyla hypophloia*, an ascomycetous fungus: Evidence for a new major class of Group I introns, *J. Mol. Evol.*, 1999, **48**, 493–500.
8. M. Belfort and R.J. Roberts, Homing endonucleases: Keeping the house in order, *Nucleic Acids Res.*, 1997, **25**, 3379–3388.
9. R.J. Roberts, M. Belfort, T. Bestor, A.S. Bhagwat, T.A. Bickle, J. Bitinaite, R.M. Blumenthal, S.K. Degtyarev, D.T. Dryden and K. Dybvig *et al.*, A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes, *Nucleic Acids Res.*, 2003, **31**, 1805–1812.
10. C. Einvik, H. Nielsen, E. Westhof, F. Michel and S. Johansen, Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites, *RNA*, 1998, **4**, 530–541.
11. C. Einvik, W.A. Decatur, T.M. Embley, V.M. Vogt and S. Johansen, *Naegleria* nucleolar introns contain two group I ribozymes with different functions in RNA splicing and processing, *RNA*, 1997, **3**, 710–720.

12. P. Haugen, O.G. Wikmark, A. Vader, D.H. Coucheron, E. Sjottem and S.D. Johansen, The recent transfer of a homing endonuclease gene, *Nucleic Acids Res.*, 2005, **33**, 2734–2741.
13. S.D. Johansen, A. Vader, E. Sjøttem and H. Nielsen, *In vivo* expression of a group I intron HEG from the antisense strand of *Didymium* ribosomal DNA, *RNA Biol.*, 2007, **3**, 157–162.
14. O.G. Wikmark, C. Einvik, J.F. De Jonckheere and S.D. Johansen, Short-term sequence evolution and vertical inheritance of the *Naegleria* twin-ribozyme group I intron, *BMC. Evol. Biol.*, 2006, **6**, 39.
15. A.B. Birgisdottir and S. Johansen, Site-specific reverse splicing of a HEG-containing group I intron in ribosomal RNA, *Nucleic Acids Res.*, 2005, **33**, 2042–2051.
16. S. Johansen, M. Elde, A. Vader, P. Haugen, K. Haugli and F. Haugli, In vivo mobility of a group I twintron in nuclear ribosomal DNA of the myxomycete *Didymium iridis*, *Mol. Microbiol.*, 1997, **24**, 737–745.
17. M. Elde, N.P. Willassen and S. Johansen, Functional characterization of isoschizomeric His-Cys box homing endonucleases from *Naegleria*, *Eur. J. Biochem.*, 2000, **267**, 7257–7266.
18. A. Vader, H. Nielsen and S. Johansen, In vivo expression of the nucleolar group I intron-encoded I-*Dir*I homing endonuclease involves the removal of a spliceosomal intron, *EMBO J.*, 1999, **18**, 1003–1013.
19. H. Nielsen, T. Fiskaa, A.B. Birgisdottir, P. Haugen, C. Einvik and S. Johansen, The ability to form full-length intron RNA circles is a general property of nuclear group I introns, *RNA*, 2003, **9**, 1464–1475.
20. A. Vader, S. Johansen and H. Nielsen, The group I-like ribozyme DiGIR1 mediates alternative processing of pre-rRNA transcripts in *Didymium iridis*, *Eur. J. Biochem.*, 2002, **269**, 5804–5812.
21. S.D. Johansen, P. Haugen and H. Nielsen, Expression of protein-coding genes embedded in ribosomal DNA, *Biochem. J.*, 2007, in the press.
22. W.A. Decatur, S. Johansen and V.M. Vogt, Expression of the *Naegleria* intron endonuclease is dependent on a functional group I self-cleaving ribozyme, *RNA*, 2000, **6**, 616–627.
23. S.L. Brehm and T.R. Cech, Fate of an intervening sequence ribonucleic acid: Excision and cyclization of the *Tetrahymena* ribosomal ribonucleic acid intervening sequence in vivo, *Biochemistry*, 1983, **22**, 2390–2397.
24. H. Nielsen, C. Einvik, T.E. Lentz, M.M. Hedegaard and S.D. Johansen, A conformational switch in the DiGIR1 ribozyme involved in release and folding of the downstream I-*Dir*I mRNA, 2007, in preparation.
25. E. Jabri, S. Aigner and T.R. Cech, Kinetic and secondary structure analysis of *Naegleria andersoni* GIR1, a group I ribozyme whose putative biological function is site-specific hydrolysis, *Biochemistry*, 1997, **36**, 16345–16354.
26. M.J. Fedor, Structure and function of the hairpin ribozyme, *J. Mol. Biol.*, 2000, **297**, 269–291.
27. D.A. Lafontaine, T.J. Wilson, D.G. Norman and D.M. Lilley, The A730 loop is an important component of the active site of the VS ribozyme, *J. Mol. Biol.*, 2001, **312**, 663–674.

28. B. Clouet-d'Orval and O.C. Uhlenbeck, Hammerhead ribozymes with a faster cleavage rate, *Biochemistry*, 1997, **36**, 9087–9092.
29. P. Agbäck, A. Sandstrom, S. Yamakage, C. Sund, C. Glemarec and J. Chattopadhyaya, Solution structure of lariat RNA by 500 MHz NMR spectroscopy and molecular dynamics studies in water, *J. Biochem. Biophys. Methods*, 1993, **27**, 229–259.
30. E. Jabri and T.R. Cech, In vitro selection of the *Naegleria* GIR1 ribozyme identifies three base changes that dramatically improve activity, *RNA*, 1998, **4**, 1481–1492.
31. C. Einvik, H. Nielsen, R. Nour and S. Johansen, Flanking sequences with an essential role in hydrolysis of a self-cleaving group I-like ribozyme, *Nucleic Acids Res.*, 2000, **28**, 2194–2200.
32. P.L. Adams, M.R. Stahley, A.B. Kosek, J. Wang and S.A. Strobel, Crystal structure of a self-splicing group I intron with both exons, *Nature*, 2004, **430**, 45–50.
33. P.L. Adams, M.R. Stahley, M.L. Gill, A.B. Kosek, J. Wang and S.A. Strobel, Crystal structure of a group I intron splicing intermediate, *RNA*, 2004, **10**, 1867–1887.
34. B. Beckert, H. Nielsen, S.D. Johansen, E. Westhof and B. Masquida, Evolution of a group I intron to a branching ribozyme studied by molecular modeling. 2007, in preparation.
35. F. Guo, A.R. Gooding and T.R. Cech, Structure of the *Tetrahymena* ribozyme: Base triple sandwich and metal ion at the active site, *Mol. Cell*, 2004, **16**, 351–362.
36. B.L. Golden, H. Kim and E. Chase, Crystal structure of a phage *Twort* group I ribozyme-product complex, *Nat. Struct. Mol. Biol.*, 2005, **12**, 82–89.
37. A. Lescoute and E. Westhof, Topology of three-way junctions in folded RNAs, *RNA*, 2006, **12**, 83–93.
38. P.P. Zarrinkar and J.R. Williamson, Kinetic intermediates in RNA folding, *Science*, 1994, **265**, 918–924.
39. W.D. Downs and T.R. Cech, Kinetic pathway for folding of the *Tetrahymena* ribozyme revealed by three UV-inducible crosslinks, *RNA*, 1996, **2**, 718–732.
40. M.R. Stahley and S.A. Strobel, Structural evidence for a two-metal-ion mechanism of group I intron splicing, *Science*, 2005, **309**, 1587–1590.
41. D. Bhattacharya, V. Reeb, D.M. Simon and F. Lutzoni, Phylogenetic analyses suggest reverse splicing spread of group I introns in fungal ribosomal DNA, *BMC. Evol. Biol.*, 2005, **5**, 68.
42. P. Rangan, B. Masquida, E. Westhof and S.A. Woodson, Architecture and folding mechanism of the *Azoarcus* Group I Pre-tRNA, *J. Mol. Biol.*, 2004, **339**, 41–51.
43. T.R. Cech and B.L. Golden, Building a catalytic active site using only RNA, in: *The RNA World*, ed. R.F. Gesteland, T.R. Cech and J.F. Atkins, CSHL Press, Cold Spring, Harbor New York, 1999, pp. 321–349.
44. H. Nielsen and S.D. Johansen, A new RNA branching activity: The GIR1 ribozyme, *Blood Cells Mol. Dis.*, 2007, **38**, 102–109.

45. K. Salehi-Ashtiani, A. Luptak, A. Litovchick and J.W. Szostak, A genome-wide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene, *Science*, 2006, **313**, 1788–1792.
46. J.F. De Jonckheere, Evidence for the ancestral origin of group I introns in the SSUrDNA of *Naegleria* spp, *J. Eukaryot. Microbiol.*, 1994, **41**, 457–463.
47. J.F. De Jonckheere and S. Brown, Loss of the ORF in the SSUrDNA group I intron of one *Naegleria* lineage, *Nucleic Acids Res.*, 1994, **22**, 3925–3927.
48. I. Dykova, I. Kyselova, H. Peckova, M. Obornik and J. Lukes, Identity of Naegleria strains isolated from organs of freshwater fishes, *Dis. Aquat. Organ*, 2001, **46**, 115–121.
49. I. Dykova, H. Peckova, I. Fiala and H. Dvorakova, Fish-isolated *Naegleria* strains and their phylogeny inferred from ITS and SSU rDNA sequences, *Folia Parasitol. (Praha)*, 2006, **53**, 172–180.
50. J.F. De Jonckheere and S. Brown, Three different group I introns in the nuclear large subunit ribosomal DNA of the amoeboflagellate *Naegleria*, *Nucleic Acids Res.*, 1998, **26**, 456–461.

ARTICLE II:

Intermolecular interaction between a branching ribozyme and associated homing endonuclease mRNA

Á. B. Birgisdottir, H. Nielsen, B. Beckert, B. Masquida, S. D. Johansen.

Submitted Biological Chemistry.
Accepted upon revision

Intermolecular interaction between a branching ribozyme and associated homing endonuclease mRNA

Ása B. Birgisdottir [1], Henrik Nielsen [2], Bertrand Beckert [2,3], Benoît Masquida [3], Steinar D. Johansen [1 *]

[1] RNA and Transcriptomics Group, Faculty of Health Sciences, University of Tromsø, Tromsø, Norway, [2] Department of Cellular and Molecular Medicine, The Panum Institute, University of Copenhagen, Copenhagen, Denmark, [3] Architecture et Réactivité de l'ARN, Université de Strasbourg, IBMC, Centre National de le Recherche Scientifique, Strasbourg, France

Correspondence

Á.B. Birgisdottir, RNA and Transcriptomics Group, Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, N-9037 Tromsø, Norway

Fax: +47 77 64 53 50

Tel: + 47 77 64 46 23

E-mail: aasa.birna.birgisdottir@uit.no

Running title: Intermolecular interaction between ribozyme and mRNA

Abstract

The control of the folding of structurally complex RNA molecules like ribozymes or riboswitches often relies on the formation of specific tertiary interactions. Usually in the case of group I ribozymes, catalysis cannot be achieved without establishing critical tertiary contacts. In this study, we give experimental evidence that catalysis can also control folding of peripheral elements of a group I related ribozyme in order to control the fate of the cleavage products. The nuclear *Didymium* GIR1 branching ribozyme (DiGIR1) and the downstream homing endonuclease mRNA are found as a whole insertion in a group I intron splicing ribozyme. DiGIR1 generates a 2', 5' lariat cap at the 5' end of the homing endonuclease mRNA by catalysing a self-cleavage branching reaction at an internal processing site. Upon release, the 5' end of the mRNA forms a distinct hairpin structure termed HEG P1. The L9 GAAA tetraloop of DiGIR1 has the potential to interact with a GNRA tetraloop receptor-like motif found within HEG P1. Our biochemical data, in concert with molecular 3D modelling, provide experimental support for an intermolecular tetraloop-receptor interaction. We conclude that the 11-nt HEG P1 motif (UCUAAG-CAAGA) represents a new and specific GAAA tetraloop-receptor engaged in RNA-RNA interactions. The biological role of this interaction appears to be linked to the homing endonuclease expression by promoting post-cleavage release of the lariat capped mRNA. These findings add to our understanding of how protein-coding genes embedded in nuclear ribosomal DNA are expressed in eukaryotes and controlled by ribozymes.

Keywords: *Didymium* GIR1; gene expression; group I intron; ribozyme; RNA tertiary interaction

Introduction

GIR1 is the only known naturally occurring catalytic RNA with structurally resemblance to group I ribozymes, but with a biological function distinct from splicing (Cech and Golden, 1999; Johansen et al., 2002; Nielsen et al., 2005; Nielsen et al., 2008; Beckert et al., 2008). GIR1 ribozymes are found as small domains (180-220 nt) within twin-ribozyme group I introns including the Dir.S956-1 intron from the myxomycete *Didymium iridis* and the Nae.S516 introns from several species of the amoeba-flagellate *Naegleria* (reviewed in Nielsen et al., 2008; Nielsen and Johansen, 2009). The GIR1s are inserted into peripheral domains of group I splicing ribozymes (see Fig. 1A for the *Didymium* intron) and are always followed by homing endonuclease genes (HEGs) (Decatur et al., 1995; Einvik et al., 1997). The *Didymium* GIR1 (DiGIR1) has been reported to catalyse a self-cleavage branching reaction at an internal processing site (IPS, see Fig. 1B) resulting in the formation of a small 2', 5' lariat cap between the first and third nucleotide at the 5' end of the homing endonuclease (HE) mRNA (Nielsen et al., 2005). Furthermore, cleavage induces a conformational change in the ribozyme, resulting in release of the downstream I-*Dir*I HE mRNA containing both a lariat cap at the 5' end and a hairpin structure (HEG P1) in the 5' untranslated region (UTR) (Nielsen et al., 2009). Whereas experimental data support the involvement of DiGIR1 in HE expression (reviewed in Johansen et al., 2007), the role of HEG P1 in mRNA formation is currently unknown.

Biologically active large RNAs often contain hairpin structures capped by GNRA (N; any base; R; purine) tetraloops (Woese et al., 1990; Abramovitz and Pyle, 1997). Two other classes of tetraloop sequences CUYG (Y corresponds to a pyrimidine) and UNCG impart extra stability to a stem-loop structure (Tuerk et al., 1988; Antao et al., 1991; Molinaro and Tinoco, 1995). GNRA tetraloops, however, not only play a role in stabilization but also participate in long-range tertiary interactions in the shallow groove side of helical elements. Phylogenetic and experimental data have shown that GUGA, GUAA and GAAA have their specific receptors CU-AG, CC-GG and the 11-nt motif (CCUAAG-UAUGG), respectively (Jaeger et al., 1994; Costa and Michel, 1995; Costa and Michel, 1997). In addition, various

receptor motifs for the GAAA loop have been identified and include a 12 nt sequence (CCCUAAC-GAGGG), termed the IC3 motif (Ikawa et al., 1999). Folding of the *Tetrahymena* group I intron ribozyme to its compact shape is largely dependent on a GNRA-loop-receptor interaction (L5b GAAA loop-P6a 11-nt motif) in the P4/P6 folding domain (Cate et al., 1996; Guo et al., 2004).

The HEG P1 hairpin in the 5' UTR of I-*Dir*I mRNA (Vader et al., 1999) contains a potential GNRA-loop receptor motif resembling the 11-nt motif in P6a of the *Tetrahymena* ribozyme. The P9 hairpin in DiGIR1 is capped by a GAAA tetraloop that could be engaged in an interaction with the HEG P1 motif (Fig. 1B). In this study we demonstrate that the HEG P1 hairpin can function as a receptor for a GAAA loop, and that post-cleavage DiGIR1 and HEG P1 RNAs are able to interact *in vitro*.

Results

Structural features support the existence of a HEG P1 tetraloop-receptor.

The HEG P1 hairpin structure in 5' UTR of the released I-*Dir*I mRNA was first noted in (Vader et al., 1999). HEG P1 is formed by conformational switching following the DiGIR1 branching reaction (Nielsen et al., 2009) and is incompatible with the active conformation of DiGIR1 because several nucleotides are involved in the important P10-P2 stems essential for catalysis (Fig. 1B).

To gain experimental support for HEG P1, structure probing experiments were performed on post-cleavage HEG-containing RNA. In short, a DiGIR1 ribozyme construct containing 162 nt upstream and 65 nt downstream IPS was *in vitro* transcribed and subjected to cleavage and lariat formation. The released 65 nt downstream RNA, which contains the lariat, was gel-purified and further subjected to DMS, DEPC, or RNaseV1 cleavage. Experimental results (Fig. 1C) strongly support the existence of HEG P1 after cleavage of 162.65 RNA. RNase V1 which cleaves RNA at helical regions, gave a number of significant hits in the proximal base paired region. However the single stranded specific chemical probes DMS and DEPC reacted predominantly with residues within the internal and terminal loop regions. Close examination of the HEG P1 sequence revealed a putative GNRA tetraloop-receptor motif. This motif contains a sequence with the potential to form an AA platform (UCUAAG-CAAGA; Fig. 1B) resembling that of the well-described 11-nt GAAA tetraloop

receptor motif (CCUAAG-UAUGG) frequently noted in self-splicing introns (Costa and Michel, 1995).

The HEG P1 receptor motif binds to a GAAA tetraloop.

Next, we analyzed whether the HEG P1 motif could function as a specific GAAA tetraloop-receptor. Due to the complex involvement of the HEG P1 nucleotides in conformational switching (Nielsen et al., 2009) we used a gel mobility-shift assay based on the bimolecular *Tetrahymena* ribozyme system (Ikawa et al., 1999; Ikawa et al., 2001). The *Tetrahymena* ribozyme has a large P5 extension consisting of P5a, P5b and P5c regions (Cech et al., 1994), and it has been demonstrated that a separately prepared P5abc domain RNA added in *trans* can activate an inactive *Tetrahymena* ribozyme mutant (ΔP5abc) lacking the P5abc element (van der Horst et al., 1991). This is achieved by the formation of a stable RNA-RNA complex that functions as a catalytically active ribozyme (Fig. 2A). The stability of the bimolecular ribozyme complex mainly depends on a tertiary interaction between the L5b GAAA loop and the 11-nt receptor motif in the P6a region of ΔP5abc.

The *Tetrahymena* P6a 11-nt motif in the ΔP5abc intron was mutated (one position at a time) towards the HEG P1 motif, giving rise to 8 different constructs (Fig. 2B). All the different ΔP5abc constructs were incubated with an equal amount of uniformly ^{35}S-labelled P5abc containing the L5b GAAA tetraloop. P5abc is much longer than GIR1 P9 (9 vs 4 bp respectively). Consequently, the natural sequence of the P5b stem was used for the experiments to avoid engineering of P5abc that may have exhibited affected structural properties leading to experimental artefacts. Moreover, the stem supporting the L5b tetraloop is known to be very stable and the bases have not been observed to participate directly in the complex formation (Cate et al. (1996) Science, 273, 1696). The stability of P9 has been demonstrated elsewhere (Einvik et al. 1998 1st model) despite its short length of 4 base-pairs. The prefered C=G pair closing P9 readily before the loop is known to have a major effect on the correct folding of GNRA and UUCG tetraloops (Blose et al. (2009) JACS, 131, 8474). Complex formation was monitored by an RNA-RNA gel mobility-shift assay. The *Tetrahymena* P6a 11-nt motif was included as a control which is expected to roughly co-migrate with the complex since P5abc is necessary for the folding of the ribozyme. In the presence of 10 mM Mg^{2+}, complex formation was observed for all the receptor motifs tested (Fig. 2C). The results indicate that the 11-nt motif has the strongest affinity to the GAAA loop. Receptor mutant number 6 (Mut6) is a weak interacting partner whereas the HEG P1

motif and Mut5 show moderate complex formations. In order to verify that the observed retarded signals on the gels represented bimolecular complexes, we prepared two 3' end-extended versions of the ΔP5abc ribozyme bearing either the 11-nt receptor motif or the HEG P1 motif, respectively. Complex formation with P5abc should then result in retarded signals migrating slower in the gel than the originally identified signals. This was indeed the case (Fig. 2C) and confirms complex formation between P5abc and the ΔP5abc ribozyme with the different receptor motifs.

To assess the stability of the tetraloop-receptor interactions we performed the gel-mobility shift assay at 5 mM Mg^{2+}. In the bimolecular ribozyme system from *Tetrahymena*, the magnesium ion concentration required for the formation of a stable RNA-RNA ribozyme complex is inversely proportional to the binding affinity between the receptor and the loop (Naito et al., 1998; Ikawa et al., 1999; Ikawa et al., 2001). At 5 mM Mg^{2+} the binding affinity of all receptor motifs of the GAAA loop was reduced (data not shown). The 11-nt *Tetrahymena* motif had the strongest affinity but Mut4 and Mut6, as well as the HEG P1 receptor motif showed a significant reduction in the binding affinity, implicating that the complex formed with the HEG P1 motif is less stable than the complex formed with the 11-nt *Tetrahymena* motif. The affinity of the HEG P1 motif and the *Tetrahymena* 11-nt motif to a UUCG tetraloop was investigated by the same bimolecular approach. We observed practically no interactions for both tetraloop receptor motifs (Fig. 2C), demonstrating that the HEG P1 receptor motif is able, as the *Tetrahymena* 11-nt motif, to discriminate between a GAAA (GNRA loop) and a non-GNRA loop. In summary, the results show that the HEG P1 receptor motif can function as a GAAA loop receptor in the bimolecular system based on the *Tetrahymena* ribozyme. This receptor motif has lower affinity for the GAAA loop than the 11-nt receptor, but is able to discriminate between the GAAA loop and the non-GNRA loop UUCG.

Interaction between HEG P1 RNA and post-cleavage form of DiGIR1 ribozyme.

To examine the tertiary interaction between the HEG P1 and post-cleavage DiGIR1 P9 tetraloop, we performed gel-mobility shift analysis using uniformly ^{35}S-labelled HEG P1 RNA incubated with unlabelled DiGIR1 RNA. The DiGIR1 ribozyme was gel-purified (urea-PAGE) from an unlabelled cleavage reaction to obtain a post-cleavage form of the ribozyme. A 32 nucleotide HEG P1 hairpin was prepared separately by *in vitro* transcription (Fig. 3A). PAGE purified uniformly ^{35}S-labelled HEG P1 transcripts (0.5 μM) were incubated together with increasing concentrations of the prepared unlabelled DiGIR1 ribozyme bearing either

the wild-type (wt) L9 GAAA or a mutated tetraloop (L9 UUCG). At 10 mM Mg^{2+}, increasing amounts of complex formation between HEG P1 and DiGIR1 was observed on native gels with increasing concentrations of DiGIR1 L9 GAAA (Fig. 3B). In a parallel experiment we incubated unlabelled DiGIR1 with L9 GAAA or L9 UUCG with labelled HEG P1 transcripts of 56 nt (Fig. 3C). Similarly to the above experiment, gel mobility-shift analysis revealed increased complex formation with increased amounts of DiGIR1 L9 GAAA (Fig. 3D). These results demonstrate that the HEG P1 RNA is able to form a specific complex with a post-cleavage form of DiGIR1 RNA, an observation consistent with an interaction between the L9 GAAA in DiGIR1 and the receptor motif in HEG P1.

Molecular modelling of the post-cleaved DiGIR1-HEG P1 interaction.

A structural model representing the interaction between HEG P1 and the post-cleavage form of the DiGIR1 ribozyme (Fig. 4) was built based on our model of the DiGIR1 precursor (Beckert, 2008). Following cleavage, DiGIR1 undergoes secondary structure rearrangements due to the formation of HEG P1 that are characterized by the dissociation of strands forming stems P2 and P10 (Nielsen et al., 2009) (see Fig. 1). The tetraloop-receptor created by the formation of HEG P1 is able to interact with the L9 GAAA tetraloop. One of the aims was to propose a model supported by the mutational data that explains this interaction. Crystal structures of various tetraloops interacting with their 11-nt receptor motifs were compared (Cate et al., 1996; Guo et al., 2004; Adams et al., 2004). Superimposition of the tetraloops showed that the resulting positions and orientations of the receptors were preserved. A model of the GAAA-tetraloop receptor was built and refined according to the specific sequence of the HEG P1 motif. The architecture of HEG P1 is consistent with a functional 11-nt tetraloop receptor, but three important differences are noted. The proximal U-G pair (Fig. 2B) is replaced by a C-G pair, the bulging U is replaced by an A, and finally the G-C pair distal from the U-G pair is replaced by an A-U pair (Fig. 4A and B). However, these differences do not affect the key structural determinants that include the G-C pair immediately downstream the bulge, the *trans* A-U pair involving the Hoogsteen edge of the A as well as the Watson-Crick edge of the U, and the AA-platform. Contacts known to be essential in the interaction between GAAA and the well-studied 11-nt motif (CCUAAG-UAUGG) (Costa and Michel, 1995; Cate et al., 1996; Costa and Michel, 1997) are thus all present in the HEG P1 motif (UCUAAG-CAAGA) (see legend of Fig. 4 for details), but Watson-Crick base-pairs closing the receptor core motif have compensatory changes (see Fig. 2B). These substitutions, in concert with the bulging A residue located on

the side opposite to the AA platform, may explain the observed lower affinity of the HEG P1 receptor for the GAAA tetraloop (Fig. 2C).

Discussion

In the present study we have identified and investigated a putative intermolecular tertiary association between a GAAA tetraloop present in the P9 segment of the DiGIR1 branching ribozyme, and the HEG P1 receptor motif present in a separate I-*Dir*I mRNA. The results support that the putative receptor motif has affinity and specificity for a GAAA loop in a bimolecular system derived from the *Tetrahymena* ribozyme. Furthermore, separately prepared HEG P1 RNA and a post-cleavage DiGIR1 RNA are able to form complexes *in vitro*, demonstrated by a gel-mobility shift assay and further supported by molecular 3D modelling.

What is the biological role of the GAAA-receptor interaction? Recently, we noted that HEG P1 is formed as a result of RNA conformational switching following DiGIR1 cleavage (Nielsen et al., 2009). The DiGIR1 ribozyme branching reaction occurs by transesterification, leading to lariat capping of the downstream RNA (Nielsen et al., 2005). The lariat cap presumably protects the 5' end of I-*Dir*I mRNA from 5' →3' exonucleolytic degradation, and the conformational switch is proposed to be linked to the release of the 5' capped HE messenger from the DiGIR1 ribozyme core (Nielsen et al., 2009). Prior to branching, the HEG P1 nucleotides are involved in alternative base pairings with the 5' region of the DiGIR1 core (see Fig. 1B), which subsequently become destabilized during cleavage and conformational switching. The formation of the L9-HEG P1 interaction has major consequences on the fate of the lariat after catalysis. The rigidity of the helical stack containing stems P3, P7, P8 and P9 prevents the P9 hairpin to bend in order to interact with the receptor in a conformation that would let the lariat reside in the catalytic pocket. Rather, molecular modelling indicates that the receptor adapts to the position of P9 apparently to pull the lariat out of the catalytic pocket, thus contributing to the release of the mRNA.

Cleavage by the *Naegleria* GIR1 (NaGIR1) also forms a 2', 5' lariat cap at the 5' end of its associated HE mRNA (our unpublished results). After branching by NaGIR1, the released *Naegleria* mRNA contains a five nucleotide conserved sequence in its 5' end (Johansen et al., 2002), which is complementary to conserved sequences in the P9 loop of the NaGIR1 ribozyme. Similar to DiGIR1, the post-cleavage form of NaGIR1 appears associated

with the downstream mRNA, but by base pairing between sequences in L9 and the 5' end of the mRNA. Interestingly, *Didymium* and *Naegleria* systems appear to have evolved two different molecular strategies to achieve the same biological goal in homing endonuclease expression. This mechanism relies on intermolecular association between post-cleavage GIR1 RNA and mRNA. There are examples from other complex ribozymes of defined tertiary contacts mediated by two distinct molecules. In stabilizing the core of group I introns, an interaction between P9 and P5 is achieved by an L9b GNRA-P5 receptor interaction in most IC1 introns, but is replaced by base pairing (P17) in the group ID intron SdCob.1 (Lehnert et al., 1996). In some *Mycoplasma* species, the RNase P RNA either uses a tetraloop-receptor interaction or a pseudoknot to build the interaction between L9 and P1 (Massire et al., 1997)..

In summary, the expression of the I-*Dir*I homing endonuclease gene embedded in nuclear ribosomal DNA depends on a series of unusual events that include: 1) RNA polymerase I precursor transcription (Vader et al., 1999; Vader et al., 2002) ; 2) lariat capping of mRNA mediated by DiGIR1 branching (Nielsen et al., 2005), 3) post-cleavage release of the mRNA from the DiGIR1 ribozyme core by conformational switching and tetraloop-receptor interaction (Beckert et al., 2008; Nielsen et al. 2009; this work), 4) polyadenylation of the 3' end and removal of a 51-nt spliceosomal intron (Vader et al., 1999), and finally 5) nuclear export and polysome association of mature mRNA in the cytoplasm for subsequent translation (Vader et al., 1999). Our findings add to the general understanding of how protein-coding genes embedded in ribosomal DNA can be expressed in eukaryotes and controlled by ribozymes (Johansen et al., 2007).

Materials and Methods

Construction of plasmid templates for *in vitro* transcription

The pDiGIR1 L9UUCG mutant plasmid was constructed from the wild-type template pDi162G1 (Einvik et al., 1998) by site directed mutagenesis using the QuickChange™ Site-Directed Mutagenesis Kit from Stratagene. The L-21 form of the *Tetrahymena* group IC1 ribozyme (Zaug et al., 1988; Guo and Cech, 2002) lacking the first 5' 21 intron nucleotides was prepared through PCR amplification from a plasmid construct with the wild type

Tetrahymena ribozyme. The PCR product was then digested with *Eco*RI and *Hin*dIII and ligated into an *Eco*RI/*Hin*dIII cut pUC18 vector. Deletion PCR on the L-21 construct, performed as described in (Naito et al., 1998), resulted in a construct lacking the P5abc domain (ΔP5abc construct). This construct was used as a template in site-directed mutagenesis reactions to obtain the different P6a receptor mutants (Mut1, Mut2, Mut3 and Mut4). Mut3 was then used as template in site-directed mutagenesis reaction to construct Mut5, Mut6 and HEG P1. The wild-type *Tetrahymena* P5abc domain construct was designed as described in (Williams et al., 1992). The PCR product was ligated into the *Eco*RI/*Hin*dIII sites of the pUC18 vector. The P5abc domain with a UUCG loop in P5b was made by annealing partially overlapping primers containing the P5abc sequence with the T7 promoter sequence at the 5' end of the forward primer. The Klenow fragment (Amersham Pharmacia Biotech Inc) was then used to fill out ends. After phenol:chloroform extraction, the product was cloned into the *Eco*RI/*Hin*dIII sites of the pUC18 vector. The plasmid pHEG P1-56 contains a 56 nt sequence starting 5 nt upstream the DiGIR1 IPS, and includes HEG P1 as well as the sequence just preceding the AUG start codon of the I-*Dir*I HEG. The pHEG P1-56 was made by annealing partially overlapping oligonucleotide primers as described above for the P5abc L9UUCG variant. The product was phenol:chloroform extracted and ligated into the *Eco*RI/*Hin*dIII cut pUC18 vector.

***In vitro* transcription, ribozyme self-cleavage and RNA PAGE-purification**

In vitro, T7 RNA polymerase-based transcriptions (Stratagene) of the DiGIR1 and *Tetrahymena* ribozyme derived constructs were performed on linearized plasmids. The pHEG P1-56 plasmid template was linearized with *Nco*I. The pDiGIR1 L9wt (pDi162G1) and pDiGIR1 L9 UUCG were also linearized with *Nco*I to give DiGIR1-162.61 variants (162 nt upstream and 61 nt downstream of IPS). The L-21 ΔP5abc intron plasmids were linearized with *Sca*I (the intron thus lacks the last 5 nt) or *Pvu*II for the longer receptor constructs. The P5abc template plasmids were linearized with *Hin*dIII. The template for the HEG P1 32 nt transcript was prepared by annealing of a T7 promoter oligonucleotide with an oligonucleotide encoding the 32 nucleotides of HEG P1 coupled to a T7 promoter complementary sequence. The annealing was performed with equal molar amounts of the oligonucleotide primers (final concentration of 25 µM) in 10 mM Tris-HCl pH 8, incubated at 95ºC for 5 min and then gradually to 25ºC. Transcription by a T7 RNA polymerase (Stratagene) was then conducted using 2.5 µM of the annealed oligonucleotides. Uniformly labelled RNA was prepared using [α-^{35}S] CTP (10 µCi·µL^{-1}; Amersham Pharmacia Biotech)

during the transcription. Prior to PAGE-purification of the post-cleavage forms of DiGIR1 WT and DiGIR1 L9 UUCG ribozymes, the transcripts were subjected to self-cleavage conditions (40 mM Tris-HCl pH 7.5, 15 mM $MgCl_2$, 500 mM KCl, 2 mM Spermidine, 5 mM DTT) at 45ºC for 2 and 4 hours, respectively and then ethanol precipitated. The post-cleavage forms of DiGIR1 WT and DiGIR1 L9 UUCG, as well as transcripts from other constructs were excised from 8 M urea, 5% polyacrylamide gels and incubated in 400 µL elution buffer (300 mM NH_4Ac, 0.1% SDS, 10 mM Tris-HCl pH 8 and 2.5 mM EDTA pH 8) on a rotating wheel at 4ºC overnight.

Structure probing

The DMS and DEPC modification reactions, RNase V1 cleavage reaction, and subsequent primer extension analyses were done essentially according to (Christiansen et al., 1990). RNA for chemical probing was *in vitro* transcribed at 5°C over night, uniformly labeled using [α-^{32}P]UTP (10 µCi·µL^{-1}; Amersham Pharmacia Biotech) during transcription. Phenol-chloroform purified precursor-RNA was subjected to self-cleavage and chemically modified in modification buffer (70 mM Hepes-KOH (pH 7.8), 10 mM $MgCl_2$, 270 mM KCl, 1 mM DTT). Modified RNA species were PAGE-purified and submitted to primer extension analysis using the primer C301 (5'-TCACCATGGTTGTTGA-3'). RNA for V1 probing was uniformly labeled using [α-^{32}P]UTP (10 µCi·µL^{-1}; Amersham Pharmacia Biotech) during *in vitro* transcription at 37°C. Phenol-chloroform purified precursor-RNA was processed for 20 min at 45°C in modification buffer (70 mM Hepes-KOH pH 7.8, 10mM $MgCL_2$, 270 mM KCl) before being subjected to RNase V1. Reactions were terminated by adding NaOAc, phenol-chloroform extracted, and ethanol precipitated. Specificity of probes: Dimethyl sulphate -DMS (A > C), Diethyl pyrocarbonate - DEPC (unpaired A), and RNase V1 (Ambion) (double-stranded RNA).

Gel-mobility shift assay

Reactions of 15 µL combining 0.5 µM uniformly ^{35}S-labelled gel-purified P5abc or P5abc with P5bUUCG and 0.5 µM gel-purified unlabeled ΔP5abc or one of its variants (Mut1-Mut6; HEG P1 motif) were pre-incubated at 50ºC for 10 min in a buffer containing 50 mM Tris-HCl , pH 7.5, 10 mM $MgCl_2$ and 200 mM NH_4Cl. Before adding the P5abc part to the different ΔP5abc variants, the reactions were allowed to cool down to approximately 30ºC. After mixing, the reactions were incubated at 30ºC for 60 min to allow complex

formation. 0.5 µM uniformly ^{35}S-labelled PAGE-purified HEG P1 (32 nt) and 0.5 µM-3 µM unlabelled PAGE-purified post-cleavage forms of DiGIR1 were treated in the same way before gel electrophoresis of the interaction assays. After 60 min incubation, the reactions were immediately electrophoresed on 5% native polyacrylamide gels containing 32 mM Tris-HCl pH 7.5, 132 mM HEPES, 0.1 mM EDTA and 10 mM $MgCl_2$. The gels were run at 15 W in the same buffer at 4 ºC for 3 to 5 ½ hours. Dried gels were then subjected to autoradiography.

Molecular modelling

Molecular modelling was performed as described in Masquida and Westhof (Masquida and Westhof, 2005; Masquida et al., 2010), and based upon our recent pre-cleavage DiGIR1 model (Beckert et al., 2008). The sequence of the 11-nt receptor motif from GIR1 was applied to the receptor structure model from the P4-P6 crystal structure (Cate et al., 1996) using the program Fragment embedded in the RNA modelling software Manip (Massire and Westhof, 1998). Superimposition of the tetraloop-receptor structures was performed using lsqman (Kleywegt and Jones, 1994). The final model was refined using the nuclin/nuclsq least-square refinement program (Westhof et al., 1985) Figures were prepared using the PYMOL program (DeLano, 2002).

References

Abramovitz, D.L. and Pyle, A.M. (1997). Remarkable morphological variability of a common RNA folding motif: The GNRA tetraloop-receptor interaction. J. Mol. Biol. *266*, 493-506.

Adams, P.L., Stahley, M.R., Kosek, A.B., Wang, J., and Strobel, S.A. (2004). Crystal structure of a self-splicing group I intron with both exons. Nature *430*, 45-50.

Antao, V.P., Lai, S.Y., and Tinoco, I. (1991). A thermodynamic study of unusually stable RNA and DNA hairpins. Nucleic Acids Res. *19*, 5901-5905.

Beckert, B., Nielsen, H., Einvik, C., Johansen, S.D., Westhof, E., and Masquida, B. (2008). Molecular modeling of the GIR1 branching ribozyme gives new insight into evolution of structurally related ribozymes. EMBO J. *27*, 667-678.

Cate, J.H., Gooding, A.R., Podell, E., Zhou, K., Golden, B.L., Kundrot, C.E., Cech, T.R., and Doudna, J.A. (1996). Crystal structure of a group I ribozyme domain: principles of RNA packing. Science *273*, 1678-1685.

Cech, T.R., Damberger, S.H., and Gutell, R.R. (1994). Representation of the secondary and tertiary structure of group I introns. Nat. Struct. Biol. *1*, 273-280.

Cech, T.R., and Golden, B.L. (1999). Building an active site using only RNA. In: The RNA World, 2nd edition. R.F. Gesteland, T.R. Cech and J.F. Atkins, eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 321-349.

Christiansen, J., Egeberg, J., Larsen, N., and Garrett, R.A. (1990). Analysis of rRNA structures: Experimental and theoretical considerations. In: Ribozymes and Protein Synthesis - A practical approach. G. Spedding, ed. (Oxford, UK: Oxford University Press), pp 229-252.

Costa, M. and Michel, F. (1995). Frequent use of the same tertiary motif by self-folding RNAs. EMBO J. *14*, 1276-1285.

Costa, M. and Michel, F. (1997). Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution. EMBO J. *16*, 3289-3302.

Decatur, W.A., Einvik, C., Johansen, S., and Vogt, V.M. (1995). Two group I ribozymes with different functions in a nuclear rDNA intron. EMBO J. *14*, 4558-4568.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System. http://www.pymol.org.

Doherty, E.A., Batey, R.T., Masquida, B., and Doudna, J.A. (2001). A universal mode of helix packing in RNA. Nat. Struct. Biol. *8*, 339-343.

Einvik, C., Decatur, W.A., Embley, T.M., Vogt, V.M., and Johansen, S. (1997). *Naegleria* nucleolar introns contain two group I ribozymes with different functions in RNA splicing and processing. RNA *3*, 710-720.

Einvik, C., Nielsen, H., Westhof, E., Michel, F., and Johansen, S. (1998). Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites. RNA *4*, 530-541.

Guo, F. and Cech, T.R. (2002). Evolution of *Tetrahymena* ribozyme mutants with increased structural stability. Nat. Struct. Biol. *9*, 855-861.

Guo, F., Gooding, A.R., and Cech, T.R. (2004). Structure of the *Tetrahymena* ribozyme; base triple sandwich and metal ion at the active site. Mol. Cell. *16*, 351-362.

Ikawa, Y., Naito, D., Aono, N., Shiraishi, H., and Inoue, T. (1999). A conserved motif in group IC3 introns is a new class of GNRA receptor. Nucleic Acids Res. *27*, 1859-1865.

Ikawa, Y., Nohmi, K., Atsumi, S., Shiraishi, H., and Inoue, T. (2001). A comparative study on two GNRA-tetraloop receptors: 11-nt and IC3 motifs. J. Biochem. *130*, 251-255.

Jaeger, L., Michel, F., Westhof, E. (1994). Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. J. Mol. Biol. *236*, 1271-1276.

Johansen, S., Einvik, C., and Nielsen, H. (2002). DiGIR1 and NaGIR1: naturally occurring group I-like ribozymes with unique core organization and evolved biological role. Biochimie *84*, 905-912.

Johansen, S.D., Haugen, P., and Nielsen, H. (2007). Expression of protein-coding genes embedded in ribosomal DNA. Biol. Chem. *388*, 679-686.

Kleywegt, G.J. and Jones, T.A. (1994). A super position. CCP4/ESF-EACBM Newsletter Prot Crystal 9-14.

Lehnert, V., Jaeger, L., Michel, F. and Westhof, E. (1996). New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the Tetrahymena thermophila ribozyme. Chem. Biol. *3*, 993-1009.

Leontis, N.B. and Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. RNA *7*, 499-512.

Masquida, B. and Westhof, E. (2005). Modeling the architecture of structured RNAs within a modular and hierarchical framework. In: Handbook of RNA biochemistry. R.K. Hartmann, A. Bindereif , A. Schön, and E. Westhof, eds,(Weinheim, Germany: Wiley VCH Verlag Gmbh & Co), pp. 536-545.

Masquida, B., Beckert, B., and Jossinet, F. (2010). Exploring RNA structure by integrative molecular modelling. N. Biotechnol. *27*: in press, doi:10.1016/j.nbt.2010.02.022.

Massire, C., Jaeger, L., and Westhof, E. (1997). Phylogenetic evidence for a new tertiary interaction in bacterial RNase P RNA. RNA *3*, 553-556.

Massire, C. and Westhof, E. (1998). MANIP: an interactive tool for modelling RNA. J. Mol. Graph. Model *16*, 197-205.

Molinaro, M. and Tinoco, I. (1995). Use of ultra stable UNCG tetraloop hairpins to fold RNA structures: thermodynamic and spectroscopic applications. Nucleic Acids Res. *23*, 3056-3063.

Naito, Y., Shiraishi, H., and Inoue, T. (1998). P5abc of the *Tetrahymena* ribozyme consists of three functionally independent elements. RNA *4*, 837-846.

Nielsen, H., Westhof, E., and Johansen, S (2005). An mRNA is capped by a 2', 5' lariat catalyzed by a group I-like ribozyme. Science *309*: 1584-1587.

Nielsen, H., Beckert, B., Masquida, B., and Johansen, S.D. (2008). The GIR1 branching ribozyme. In: Ribozymes and RNA Catalysis. D.M. Lilley and F. Eckstein, eds. (Cambridge, UK: RSC Publishing), pp. 229-252.

Nielsen, H., Einvik, C., Lentz, T.E., Hedegaard, M.M., and Johansen, S.D. (2009). A conformational switch in the DiGIR1 ribozyme involved in release and folding of the downstream I-DirI mRNA. RNA *15*, 958-967.

Nielsen, H. and Johansen, S.D. (2009). Group I introns – moving in new directions. RNA Biol. *6*, 375-383.

Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B., and Steitz, T.A. (2001). RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. Proc. Natl. Acad. Sci. USA *98*, 4899-4903.

Tuerk, C., Gauss, P., Thermes, C., Groebe, D.R., Gayle, M., Guild, N., Stormo, G., d´Aubenton-Carafa, Y., Uhlenbeck, O.C., and Tinoco, I. Jr. (1988). CUUCGG hairpins: extraordinarily stable RNA secondary structures associated with various biochemical processes. Proc. Natl. Acad. Sci. USA *85*, 1364-1368.

Vader, A., Nielsen, H., and Johansen, S. (1999). In vivo expression of the nucleolar group I intron-encoded I-*Dir*I homing endonuclease involves the removal of a spliceosomal intron. EMBO J. *18*, 1003-1013

Vader, A., Johansen, S., and Nielsen, H. (2002). The group I-like ribozyme DiGIR1 mediates alternative processing of pre-rRNA transcripts in *Didymium iridis*. Eur. J. Biochem. *269*, 5804-5812.

van der Horst, G., Christian, A., and Inoue, T. (1991) Reconstitution of a group I intron self-splicing reaction with an activator RNA. Proc. Natl. Acad. Sci. USA *88*, 184-188.

Westhof, E., Dumas, P., Moras, D. (1985). Crystallographic refinement of yeast asparticacid transfer RNA. J. Mol. Biol. *184*, 119-145.

Williams, K.P., Fujimoto, D.N., and Inoue, T. (1992). A region of group I introns that contains universally conserved residues but is not essential for self-splicing. Proc. Natl. Acad. Sci. USA *89*, 10400-10404.

Woese, C.R., Winker, S., and Gutell, R.R. (1990). Architecture of ribosomal RNA: Constraints on the sequence of “tetra-loops”. Proc. Natl. Acad. Sci. USA *87*, 8467-8471.

Zaug, A.J., Grosshans, C.A., and Cech, T.R. (1988). Sequence-specific endoribonuclease activity of the *Tetrahymena* ribozyme: enhanced cleavage of certain oligonucleotide substrates that form mismatched ribozyme-substrate complexes. Biochemistry *27*, 8924-8931.

Figure legends

Figure 1. The DiGIR1 ribozyme. (**A**) Schematic view of the *Didymium* twin-ribozyme intron Dir.S956-1. The DiGIR1 branching ribozyme and the I-*Dir*I HEG constitute an insertion in the P2 segment of the splicing ribozyme DiGIR2. (**B**) Helix diagram of DiGIR1 before (pre-) and after (post-) cleavage. The internal processing site (IPS) and the branch point (BP) are indicated on the pre-cleaved form. The self-cleavage branching reaction that occurs at IPS induces a conformational change in DiGIR1, resulting in release of the I-*Dir*I mRNA with a lariat cap at the 5' end. The HEG P1 structure with a proposed GAAA tetraloop-receptor motif (the HEG P1 motif) is indicated. (**C**) Summary of structure probing analyses indicated on the HEG P1 helix diagram. Filled and open arrows/ arrowheads indicate modification DMS and DEPC, respectively. Black squares indicate RNaseV1 cleavages. Modifications are noted as strong and weak, indicated by the size of the symbols. The 2', 5' lariat cap and the I-DirI HEG are indicated at the 5' and 3' ends, respectively, of the HEG P1 RNA sequence.

Figure 2. Bimolecular system, derived from the *Tetrahymena* ribozyme, to test tetraloop-receptor interactions. (**A**) Schematic secondary structure of the bimolecular *Tetrahymena* ribozyme applied in the gel-shift assay (Ikawa et al., 1999). The L-21 variant of the *Tetrahymena* ribozyme was divided into two pieces; the ΔP5abc intron and the P5abc RNA domain. The P6a 11-nt GAAA-loop receptor motif in ΔP5abc and the L5b GAAA loop in P5abc are boxed. The two RNAs form a complex via tertiary interactions (marked with double-headed arrows) and the interaction between the L5b GAAA loop and the receptor motif in P6a (boxed) is responsible for the stability of the complex. (**B**) Sequence presentations of the 11-nt motif, wild-type (WT) 11-nt motif P6a receptor (left), the HEG P1 motif (right), and their chimeric motifs (Mut1-Mut6). The 11-nt motif in P6a of the ΔP5abc intron was gradually mutated towards the HEG P1 motif by site-directed mutagenesis. Mutated nucleotides are denoted in red. (**C**) RNA-RNA gel-mobility shift assay at 10 mM Mg^{2+}. The ^{35}S-labeled P5abc RNA (0.5 μM), with L5b as GAAA or UUCG, was mixed together with the unlabeled L-21 form of the ΔP5abc intron ribozyme (0.5 μM) containing the various P6a receptor motifs. The ^{35}S-labeled ΔP5abc ribozyme was used as a control in the absence of P5abc RNA (lanes labelled 'No P5abc'). 'L' (long) indicates a 3´end-extended

version of the ΔP5abc RNA with the 11-nt (WT) or the HEG P1 receptor motif due to linearization of the plasmid template further downstream the 3´end of the intron. The signals corresponding to a free labelled P5abc RNA domain and a ΔP5abc-P5abc RNA-RNA complex are indicated with arrows (right).

Figure 3. *In vitro* interaction between the post-cleavage DiGIR1 ribozyme and a separately prepared HEG P1 RNA. (**A**) Secondary structure depiction of the DiGIR1 post-cleavage form with the GAAA loop in P9 boxed. The HEG P1 hairpin structure at the 5' end of I-*Dir*I mRNA with the putative receptor motif boxed *(*right). (**B**) RNA-RNA gel-mobility shift assay at 10 mM Mg^{2+} with ^{35}S-labeled HEG P1 (0.5 μM) and increasing amount of the unlabeled post-cleavage form of DiGIR1, from 1:1 to 1:6 molar ratios for the L9 GAAA (WT) variant and 1:1 to 1:2 molar ratios for the L9 UUCG mutant. The mobility of the unbound hot HEG P1 and HEG P1 in complex with DiGIR1 is indicated by arrows. (**C**) Secondary structure of the 56 nt HEG P1 hairpin structure at the 5' end of I-*Dir*I mRNA (HEG P1 with tail) with the putative receptor motif boxed. **(D)** RNA-RNA gel-mobility shift assay at 10 mM Mg^{2+} with ^{35}S-labeled HEG P1 with tail (0.5 μM) and increasing amount of the unlabeled post-cleavage form of DiGIR1, from 1:1 to 1:2 molar ratios for the L9 GAAA (WT) and L9 UUCG variants (right). The mobility of the unbound hot HEG P1 with tail and HEG P1 in complex with DiGIR1 is indicated by arrows.

Figure 4. 3D structure model of the post-cleaved *Didymium* GIR1. (**A**) Secondary structure diagram of the cleaved form of the post-cleaved GIR1 showing the overall structural organization of the L9 loop bound to its HEG P1 receptor. Base-pair geometry is indicated using the Leontis-Westhof nomenclature (Leontis and Westhof, 2001; Masquida et al., 2010). The lariat is denoted by *. Coloured sequence segments indicate pairings formed in the pre-cleaved form of GIR1 that lead to the formation of P2 (slate blue) and P10 (green). (**B**) Structure model of the HEG P1 receptor motif bound to the L9 GAAA tetraloop. Colour code is with respect to A. The intermolecular contacts are as observed in (Cate et al., 1996). The Watson-Crick (WC) edge of the A and the sugar edge of the U forming the *trans* Hoogsteen-WC base-pair in the receptor (cyan) interact with the WC edge of the first A (upmost red) of the tetraloop and the sugar edge of the second to form an A-minor (type II) interaction (Doherty et al., 2001; Nissen et al., 2001), respectively. The intermolecular interaction between the two adenosines (cyan and red) is very specific and adopts a *trans*

WC-WC geometry. The third adenosine in the loop forms a type I A-minor interaction with the green G-C pair in the receptor. The first adenosine in the loop stacks on the AA-platform (yellow). (**C and D**) Stereoscopic views of the overall post-cleaved DiGIR1 ribozyme core bound to the corresponding I-*Dir*I mRNA by intermolecular tetraloop-receptor interaction. C is oriented according to the secondary structure diagram in A. HEG P1 is in the foreground and L9 in the background. D is oriented according to the 3D model in B. Atoms from ωG are represented by spheres and the lariat residues by sticks. The set of interactions between L9 and the HEG P1 receptor enforces the position of the lariat to become distal from the ribozyme core.

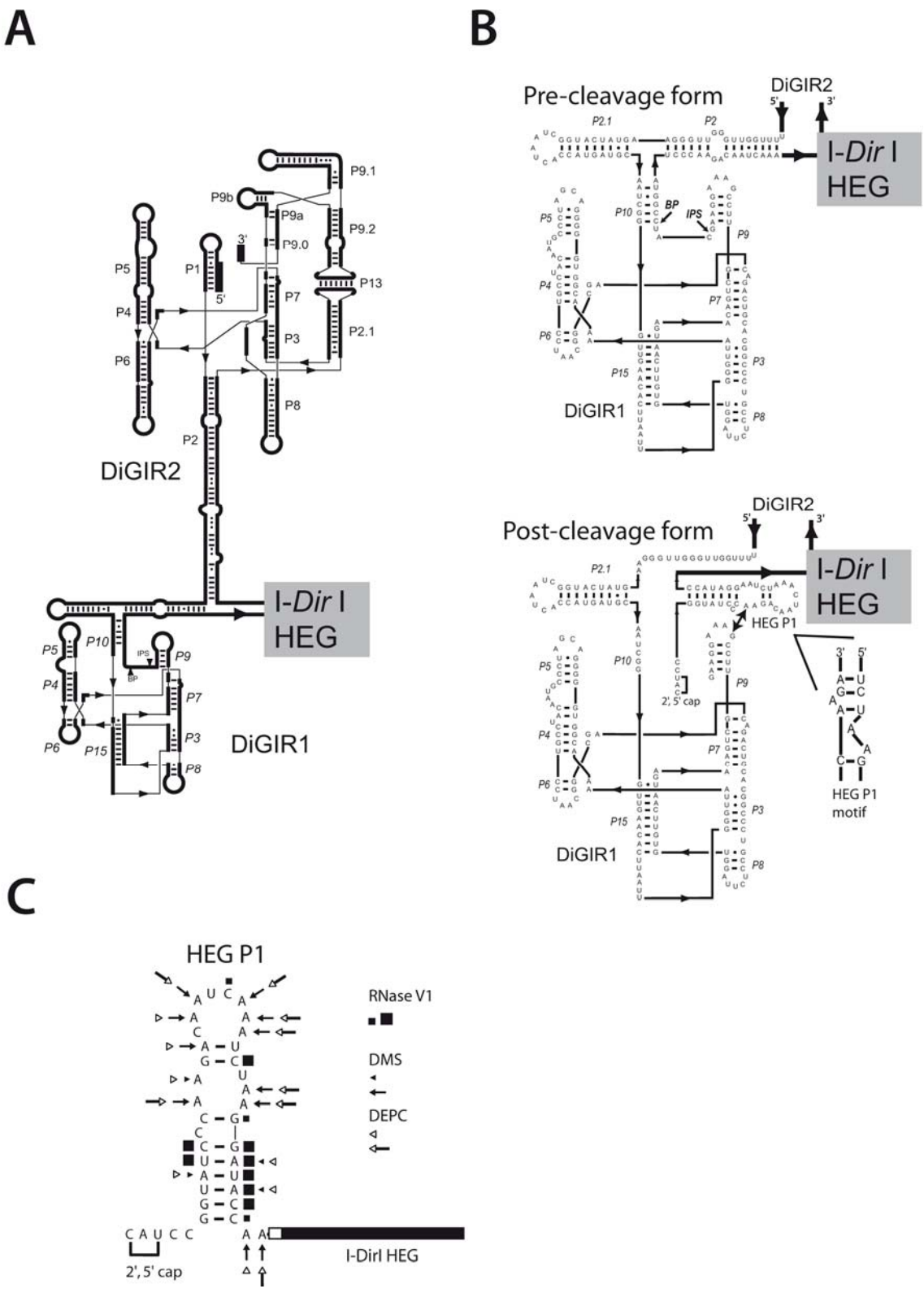

A
P9.1
P9b
P9a
P9.2
P9.0
3'
P5
P1
P13
5'
P7
P4
P3
P2.1
P6
P8
P2
DiGIR2
I-Dir I
HEG
P5
P10
IPS
P9
BP
P4
P7
P6
P15
P3
P8
DiGIR1
B
Pre-cleavage form
DiGIR2
5'
3'
P2.1
P2
I-Dir I
HEG
P5
P10
BP
IPS
P9
P4
P7
P6
P15
P3
DiGIR1
P8
Post-cleavage form
DiGIR2
5'
3'
I-Dir I
HEG
P2.1
HEG P1
P5
P10
2', 5' cap
P9
3'
5'
P4
P7
P6
P15
P3
DiGIR1
P8
HEG P1
motif
C
HEG P1
RNase V1
DMS
DEPC
C A U C C
2', 5' cap
I-DirI HEG

Figure 1

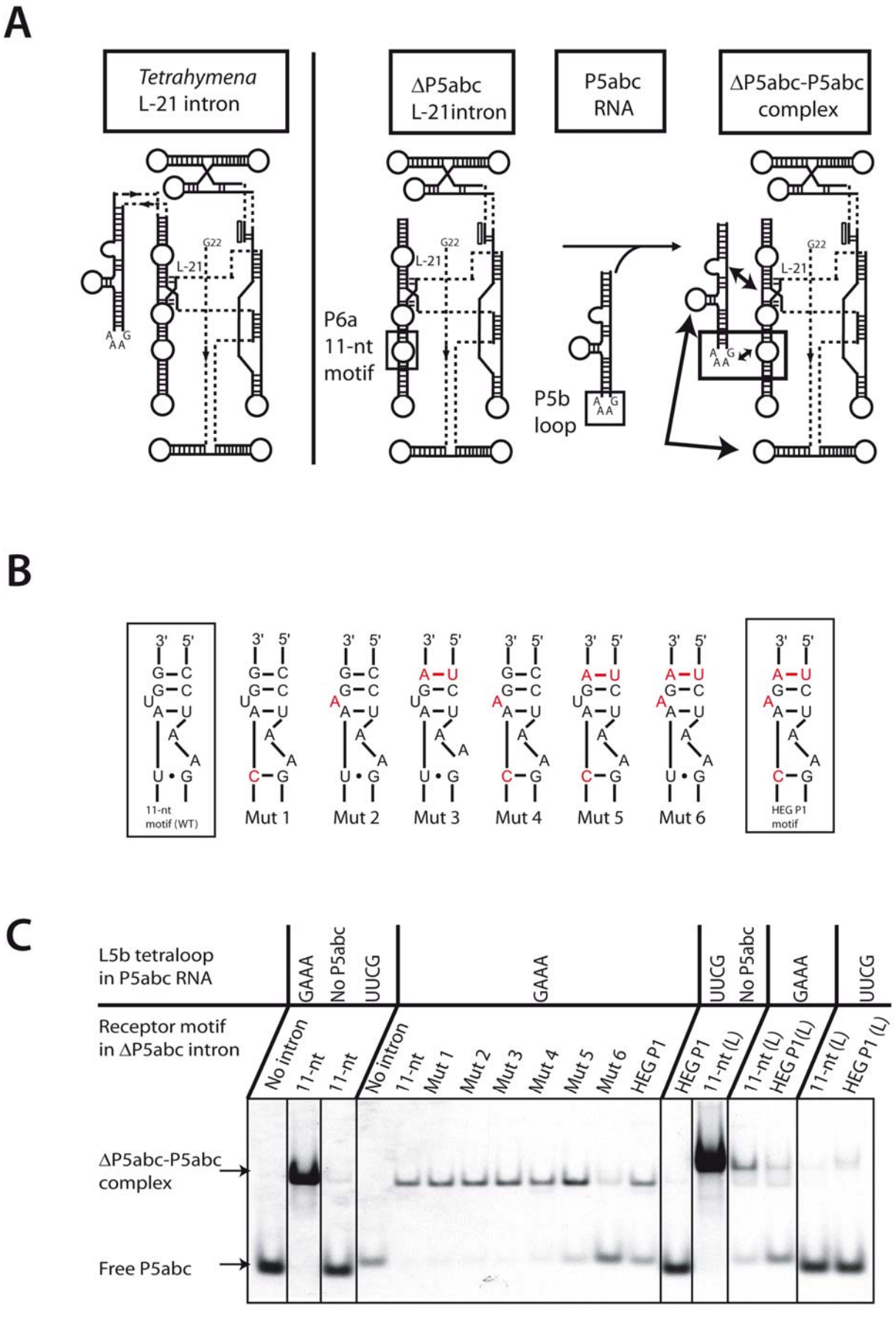

A
Tetrahymena L-21 intron
ΔP5abc L-21intron
P5abc RNA
ΔP5abc-P5abc complex
L-21
G22
P6a 11-nt motif
P5b loop
B
11-nt motif (WT)
Mut 1
Mut 2
Mut 3
Mut 4
Mut 5
Mut 6
HEG P1 motif
C
L5b tetraloop in P5abc RNA
GAAA
No P5abc
UUCG
Receptor motif in ΔP5abc intron
No intron
11-nt
Mut 1
Mut 2
Mut 3
Mut 4
Mut 5
Mut 6
HEG P1
11-nt (L)
HEG P1(L)
ΔP5abc-P5abc complex
Free P5abc

Figure 2

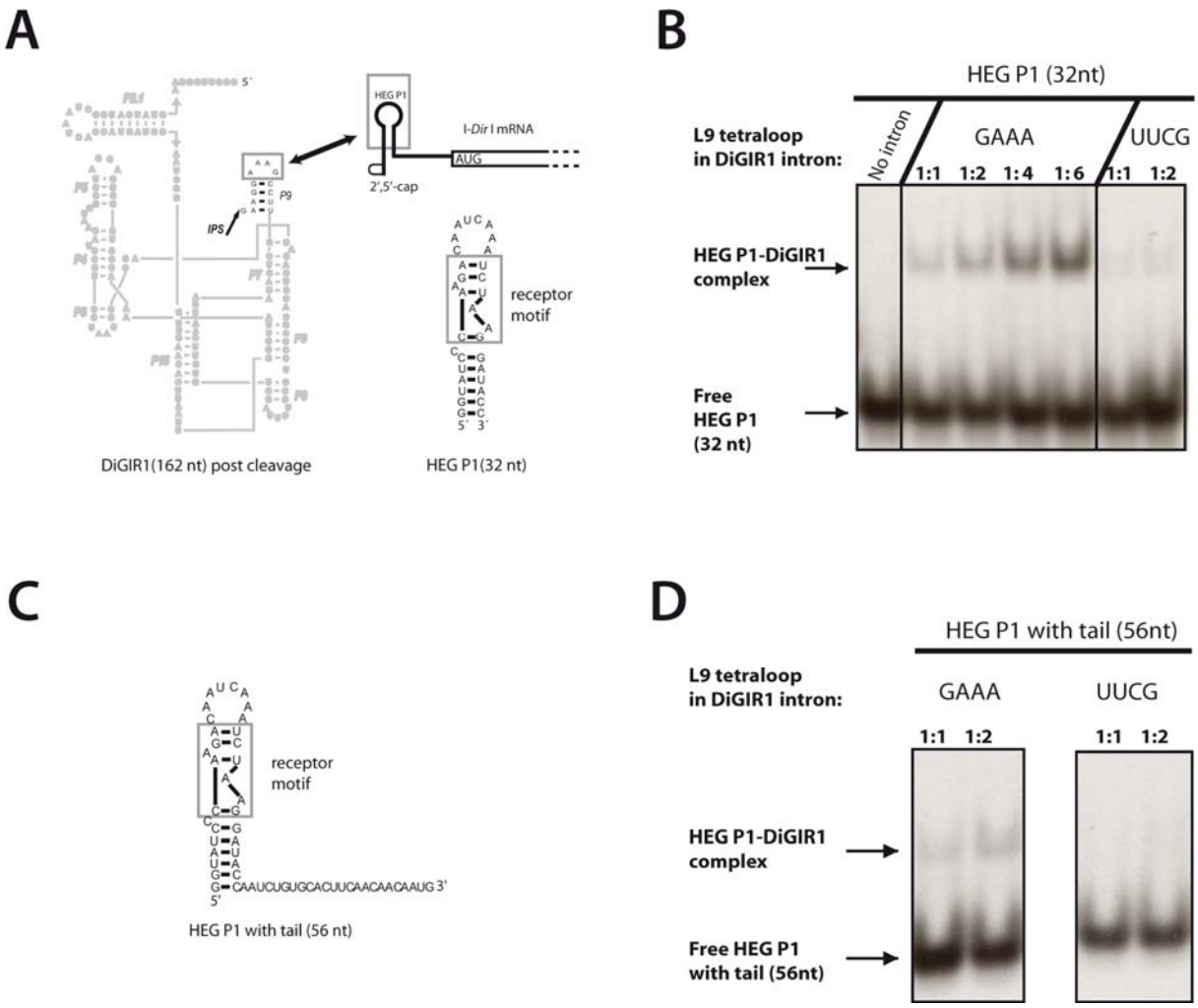

A
HEG P1
I-DirI mRNA
AUG
2',5'-cap
P9
IPS
receptor
motif
DiGIR1(162 nt) post cleavage
HEG P1(32 nt)
B
HEG P1 (32nt)
L9 tetraloop
in DiGIR1 intron:
No intron
GAAA
UUCG
1:1 1:2 1:4 1:6 1:1 1:2
HEG P1-DiGIR1
complex
Free
HEG P1
(32 nt)
C
receptor
motif
CAAUCUGUGCACUUCAACAACAAUG 3'
HEG P1 with tail (56 nt)
D
HEG P1 with tail (56nt)
L9 tetraloop
in DiGIR1 intron:
GAAA
UUCG
1:1 1:2 1:1 1:2
HEG P1-DiGIR1
complex
Free HEG P1
with tail (56nt)

Figure 3

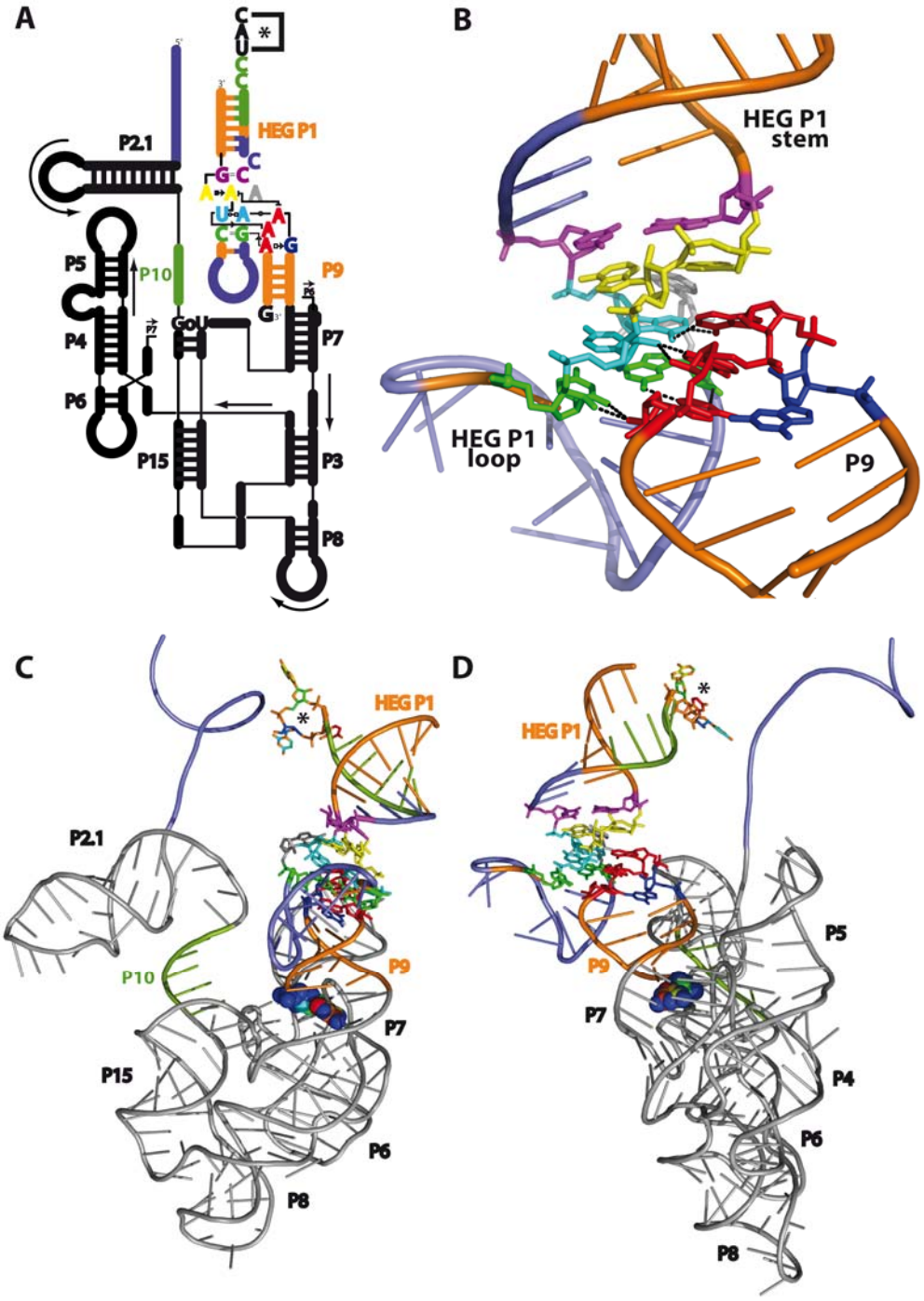

A
B
C
D
HEG P1
P2.1
P5
P10
P9
P4
P7
P6
P15
P3
P8
HEG P1 stem
HEG P1 loop
P9
HEG P1
P2.1
P10
P9
P7
P15
P6
P8
HEG P1
P9
P7
P5
P4
P6
P8

Figure 4

ARTICLE III:

Identification of an on/off switch in the DiGIR1 ribozyme

B. Beckert, M. Marquardt Hedegaard, B. Masquida, H. Nielsen

Manuscript

Identification of an on/off switch in the DiGIR1 ribozyme

Bertrand Beckert[1,2], Mads Marquardt Hedegaard[3], Benoit Masquida[2*], and Henrik Nielsen[1*]

[1] Department of Cellular and Molecular Medicine, The Panum Institute, University of Copenhagen, Denmark

[2]Architecture et Réactivité de l'ARN, Université Louis Pasteur de Strasbourg, IBMC, CNRS, France

[3] Department of Biology, University of Copenhagen, Denmark

*corresponding author

Address:

Henrik Nielsen

Department of Cellular and Molecular Medicine

The Panum Institute, University of Copenhagen

3 Blegdamsvej, DK-2200N

Telephone: +45 35 32 77 63

Fax: +45 35 32 77 32

E-mail: hamra@imbg.ku.dk

Keywords: *Didymium iridis*, DiGIR1 branching ribozyme, conformational switching

Running title:

Abstract

The DiGIR1 branching ribozyme functions in the release and capping of the mRNA encoding a homing endonuclease in the myxomycete *Didymium iridis*. Due to its localization within a group I intron in the small subunit ribosomal RNA, the ribozyme has to fold initially into an inactive conformation to avoid untimely cleavage of the ribosomal RNA precursor. By a combination of native gel electrophoresis and structure probing of DiGIR1 variants, we show that this is accomplished by a regulatory domain that can adopt two different conformations. The domain is fused to a preformed scaffold consisting of most the core responsible for the catalytic reaction. In the inactive conformation, part of the regulatory domain adopts a hairpin structure (HEG P1) that prevents the formation of a stable active site at the core. In the active conformation, HEG P1 is replaced by helices (P2 and P10) that allow the organization of a three-way junction (P2-P2.1-P10) and thus the docking of helix P2.1 that stabilizes the active site. In this way, the regulatory domain acts as an on/off switch that orchestrates the activities of the branching and splicing ribozymes in the twin-ribozyme intron in several different processing pathways and thus regulate the interplay between the intron and the host.

Introduction

RNA catalysis (1) is of fundamental importance for cell function and a few ribozymes are essential (e.g. the ribosome and RNase P) and work as true enzymes with multiple turn-over. However, the majority of ribozymes has a more sporadic phylogenetic distribution, are involved in internal reorganization of RNA molecules and operates as single turn-over catalysts. For these ribozymes, the timing of folding and thus catalysis is of particular importance because it affects dramatically the function of the molecule within which the ribozyme resides. Splicing ribozymes within group I and group II introns catalyze their own removal from precursor transcripts and at the same time promote the ligation of their flanking introns to form mature and biologically functional molecules. In these cases, the ribozyme part would be expected to fold rapidly into its active conformation. Some cleavage ribozymes, such as the hammerhead and hairpin ribozymes found in viroids cleave replication intermediates consisting of concatemers of the viroid genome into monomers and would similarly be expected to fold rapidly into the active conformation. In contrast other cleavage ribozymes have apparently adopted a regulatory function in which the activity of the ribozyme must be regulated. These ribozymes cleave and inactivate pre-mRNAs or mRNAs and thus affect the expression of cellular protein encoding genes. The hammerhead-related Clec2d ribozyme found in rodents, cleaves in the 3' UTR upstream of the polyA tail (2) and the HDV ribozyme-like CPEB3 ribozyme isolated from humans cleaves within an intron and is apparently just one of several such ribozymes in the human genome (3).

An interesting case is presented by the DiGIR1 branching ribozyme found in the myxomycete *Didymium iridis* and in several species of the amoeboflagellate *Naegleria* (4-6). Here, the ribozyme is inserted into a group I intron within the ribosomal RNA precursor (Fig. 1). The ribozyme cleaves the RNA at the internal processing site (IPS) to leave a lariat cap at the 5' end of an intron-encoded mRNA (7). If cleavage occurs prior to splicing of the intron catalysed by the flanking group I ribozyme, the ribosomal precursor will be destroyed which is likely to be detrimental to the cell during normal growth. The cleavage rate of DiGIR1 *in vitro* is only one order of magnitude less than the most optimized cleavage ribozymes (). The *in vivo* rate is unknown, but ribozymes are generally considered to be orders of magnitudes faster *in vivo* than *in vitro* (8). One example of an *in vivo* rate is the *Tetrahymena* intron that self-splices with an apparent $t_{½}$ of approximately 2 sec (9). This number should be compared to the rate of transcription by RNA

polymerase I rate that is in the range 20-50 nt/ sec (10). Given that transcription of DiGIR1 is completed long before the completion of transcription of the splicing ribozyme due to the intervening open reading frame encoding the homing endonuclease (Fig. 1), the cleavage by DiGIR1 should precede splicing. Since this is obviously not the case, we hypothesize that DiGIR1 initially folds into an inactive conformation to allow splicing to occur. Then, DiGIR1 undergoes a conformational change leading to an active ribozyme that caps the resident mRNA.

In this work, we have addressed the hypothesis by studying the folding of DiGIR1 by native gel electrophoresis and Fe-EDTA probing. For practical reasons, we have studied two length variants of DiGIR1 rather than DiGIR1 in the context of the ribosomal RNA precursor. A short variant (DiGIR1-157.22) was previously defined as the minimal version that retains branching activity (7). This variant includes 157 nt upstream and 22 nt downstream of the IPS, respectively. Based on previous cleavage analyses (11), a slightly longer variant (DiGIR1-166.65) was chosen to represent the full-length DiGIR1 cassette that is found inserted into P2 of the group I splicing ribozyme (DiGIR2) (Fig. 1). From a comparison of the two length variants they differ only in the extent of inclusion of the P2-P2.1 domain. Interestingly, this domain can adopt an alternative base pairing scheme in the full-length, but not in the minimal version compared to what was previously characterized in active ribozyme variants (Fig. 3A) (12). This conformation is characterized by the presence of a stem-loop structure, HEG P1, and the consequent disruption of the 3WJ connecting the P2-P2.1 domain and the core. This in turn results in loss of P10 and disassembly of the active site. Thus, the P2-P2.1 domain has the potential to act as a conformational switch that can turn the activity on and off. For this reason, we tentatively refer to the P2-P2.1 domain as the regulatory domain and the remainder of the ribozyme as the core. Interestingly, the switch was recently described to operate in reverse in a mechanism that ensures the release of the HE mRNA following branching (11) and HEG P1 has previously been described as an element in the 5' UTR of the resulting HE mRNA (13), hence the name HEG P1.

The present study reveals that the alternative conformation is a biologically relevant conformation of DiGIR1 rather than a misfolded ribozyme. It co-exists with the catalytically active ribozyme over a wide range of Mg^{2+} concentrations; it is stabilized by a tetraloop/ tetraloop receptor interaction not found in the active conformation, and it plays an important role in the biology of the

twin-ribozyme intron. Furthermore, the Fe-EDTA experiments demonstrate how the P2-P2.1 domain acts as a switch to turn on the ribozyme activity by stabilizing the active site.

Materials and methods

RNA preparation

RNA was prepared by *in vitro* transcription of PCR templates or linearized plasmids using T7 RNA polymerase (Fermentas). Templates for variants of DiGIR1 ribozyme were made by PCR of pDi162G1 (14) using Pfu DNA polymerase and the following combinations of oligos. 157.22: C294 (5' AAT TTA ATA CGA CTC ACT ATA GGG AAG TAT CAT) and OP233 (5' GAT TGT CTT GGG ATA CCG); 166.65: C289 (5′ AAT TTA ATA CGA CTC ACT ATA GGT TGG GTT GGG AAG TAT CAT) and C288 (5' TCA CCA TGG TTG TTG AAG TGC ACA GAT TG); 166.22: C289 and OP233; 157.65: C294 and C288; 166.65HEGP1: C289 and OP879 (5' TCA CCA TGG TTG TTG AAG TGC ACA GAT TTC ATA GGA ATC TTT TGA TTG TCT TGG GAT ACC). The mutant version was previously described as UTR2 (11). The *Azoarcus* intron RNA used in native gel assays was prepared by *in vitro* transcription of linearized pAZ-PREt as described in (15). The transcript is 284nt and is a pre-tRNA consisting of the sequence of the mature tRNA ending in CCA (79 nt) and the intron (205 nt). Full-length twin-ribozyme intron RNA for structure probing was prepared by transcription of linearized DiSSU1 (16) at 5°C over night. This resulted in uniform and unprocessed full-length intron (1749 nt) flanked by 67 nt and 313 nt exons, respectively.

All transcripts were purified on thoroughly buffered S-300 microspin columns (GE Healthcare). Body-labeled RNA was made by including trace amounts of [α-^{32}P]UTP (New England Nuclear) in the transcription reaction. 5' end-labelled RNA was made by treatment of unlabelled *in vitro* transcript with alkaline phosphatase (Fermentas) followed by labelling using T4 PNK (Fermentas) and [γ-^{32}P]ATP (New England Nuclear). 3' end-labelled RNA was made by a fill-in reaction (17) using a short oligo (5'CCG TCA CCA TGG TTG TTG AAG TG ddC 3') with partial complementarity to the 3' end of DiGIR1 166.65 as template. After annealing of the oligo, the RNA was extended by one nucleotide using [α-^{32}P]dCTP (New England Nuclear) and the Klenow fragment of DNA polymerase I (Fermentas). This procedure yields RNA with a well-defined homogenous 3' end label. All radiolabeled RNAs were gel purified on 5% denaturing (urea) polyacrylamide gels, eluted, precipitated, resuspended in DEPC water and stored at -20°C until use.

Native gel assay

To measure the Mg^{2+} dependent folding, body-labelled ^{32}P-RNA was incubated in 1 M KCl, 25 mM sodium acetate (pH 5.5), 0,1 mM EDTA, 10% glycerol, 0,1% xylene cyanol and 0-100 mM $MgCl_2$ for 5 min at 45°C, then immediately loaded on a native 10% polyacrylamide gel (29:1 acrylamide/bis), 34 mM Tris HCl (pH 7.5), 66 mM HEPES (pH 7.5), 0,1 mM EDTA, 3 mM $MgCl_2$) at 4°C (18). Samples were electrophoresed at 15 Watts per gel for 6-7 h at 4°C. Gels were then exposed to Molecular Dynamics PhosphorImager screens and quantified with ImageQuant software. The fraction of native RNA *f*n was determined from the counts in band N relative to the total counts in the lane according to (15). All the RNAs subjected to the native gel experiments were analyzed three times in parallel and the results were highly reproducible.

Hydroxyl radical footprinting

The Fe(II)-EDTA reaction was performed as described in (17). 5' end or 3'end labelled RNA was diluted in 10 µL DEPC water at 20 000 cpm/µL. The RNA was folded by addition of one volume of 2X folding buffer (2 M KCl; 50 mM NaAc (pH 5.5) and 0-100 mM $MgCl_2$) and incubation at 45°C for 5 min. The reaction was started by addition of 1 µL of 75 mM $(NH_4)_2Fe(SO_4)_2$, 1 µL of 150 mM EDTA (pH 8.0), 1 µL of 375 mM DTT and 1 µL of 15 % H_2O_2 (25 µL final volume). Reactions were quenched after 2 min at room temperature by ethanol precipitation at -80°C for 15 min by addition 175 µL H_2O, 1/10 volume of 3 M sodium acetate (pH 5), 2.5 volumes of 96% ethanol and 1 µg of carrier tRNA. Samples were pelleted, dried and dissolved in 6 µL of denaturing loading buffer (95% formamide, 50 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol FF). RNase T1 cleavage ladder was made by incubation of the endlabelled RNA at 50°C for 5 min in buffer T1 (citrate NaOH 20 mM, 1 mM EDTA 7 M urea, 0.05% xylene cyanol FF, 0.05% bromophenol blue). RNase T1 (28 U/µL) was added and incubation continued for 10 min at 50°C after which the reaction was quenched by addition of an equal volume of denaturing loading buffer. Alkaline ladder was carried out by incubation of the end-labelled RNA at 95°C for 3 min in Alkaline buffer (100 mM Na_2CO_3/$NaHCO_3$ (pH 9.2), 0.1 M NaOH, 1 mM EDTA). Before loading on 15% denaturing polyacrylamide gels all the samples were heat-denatured at 95°C for 1 min. Gels were exposed to PhosphorImager screens (Molecular Dynamics) and quantified using ImageQuant and SAFA software as described in (19).

Chemical and enzymatic probing

Structure probing was performed essentially as described in (20). Briefly, 4 µg of *in vitro* transcript in 200 µL (chemical modification) or 40 µL (enzymatic reaction) in probing buffer (270 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 70 mM HEPES-KOH (pH 7.8)) was incubated on ice with the probe. The following specific conditions were applied. DMS: 2µL of 50% DMS for 20 min. DEPC: 7µL for 30 min. Kethoxal: 20 µL of 40 mg/mL for 75 min. CMCT: RNase T1: 0.1 U or 0.2 U. RNase T2: 0.5 U or 1 U. RNase A: 0.05 U or 0.1 U. RNase V1: 1/300 U or 2/300 U. All reactions were terminated by ethanol precipitation and subjected to primer extension reactions as previously described (12;14).

Molecular modeling

The P2-P2.1 domain of the DiGIR1 ribozyme was modeled using the program ASSEMBLE (21). The model of the three-way junction between P2, P2.1 and P10 was based on the RNAJunction database (22) and added onto the 3D model of DiGIR1 core. Interactive modeling followed by refinement steps were performed iteratively until solution data reported in this work could be explained satisfactorily. The color views were generated with program Pymol (23). The accessibilities of the C4' atoms to hydroxyl radicals were computed using the program Naccess.

Results

Native gel assay for Mg^{2+} dependent folding of length variants

Native polyacrylamide gel electrophoresis was used to resolve the folding conformers of the DiGIR1 ribozyme in the minimal (157.22) and full-length (166.65) versions at different Mg^{2+} concentrations. Subsequently, the fraction of native ribozyme as a function of Mg^{2+} concentration was normalized to the extent of folding at saturation and fitted to the Hill equation in order to determine the midpoint of folding transition with respect to the Mg^{2+} concentration (15). The *Azoarcus* group I splicing ribozyme has previously been thoroughly characterized by native gel techniques (15;24;25) and was analyzed in parallel as a reference because of similarities in secondary and tertiary structure (26).

Prior to loading on gels, the samples were pre-incubated in acidic folding buffer with Mg^{2+} for 5 min. At these conditions, the ribozyme is inactive but shows a burst of reactivity upon a pH jump to neutral pH indicating that a fraction of the ribozyme is correctly folded during the 5 min pre-incubation. At low Mg^{2+} concentration the minimal ribozyme migrates as two diffuse bands, one of which shows low mobility (Fig. 3A). These bands most likely represent unfolded and partially folded states consistent with the lack of activity of the ribozyme at low Mg^{2+} concentrations. At higher Mg^{2+} concentration, the ribozyme migrates as a focused band (Fig. 3A) representing a near-active or active conformation. The midpoint of the folding transition was at 1.1 ± 0.37 mM Mg^{2+} and found to be mildly cooperative with respect to Mg^{2+} concentration with a Hill coefficient of $\eta_{157.22} = 2.3$ (Fig. 3C). This implies that the DiGIR1 ribozyme requires 3.5 times more Mg^{2+} than the *Azoarcus* intron with a C_m = 0.35 ± 0.2 mM Mg^{2+} (Fig. 3C). However, the ribozyme requires around 25 mM Mg^{2+} for full activity suggesting that additional folding not revealed by the gel mobility is required.

Next, we analyzed a length variant with nine additional nucleotides at the 5' end (166.22; Fig. 3B). In contrast to the minimal variant that does branching only, this variant also has ligation activity resulting in complete masking of the branching reaction under standard conditions. However, in the native gel assay, the 5' extended variant is similar to the minimal ribozyme ($C_{m166.22}$ = 1.2 ± 0.5 mM Mg^{2+} ; $\eta_{166.22} = 2.9$) suggesting that the 5' extension itself does not affect the pre-cleavage folding of the ribozyme. We then turned to the full-length ribozyme (166.65) that has extensions of 9 nt and 43 nt at the 5' and 3' ends as compared to the minimal version, respectively. The full-length ribozyme is similar in reactivity to the above mentioned 166.22 variant. However, the behaviour of this length variant in the native gel assay is distinctly different. In addition to the diffuse bands at low Mg^{2+} concentrations and the focused band at high Mg^{2+} concentrations, a new and focused band appears at low Mg^{2+} concentration, peaks at Mg^{2+} concentration in the low mM range and is still present at high Mg^{2+} concentration (Fig. 4A; ALT). We speculated that this band corresponds to an inactive conformation that includes HEG P1 (Fig. 2B). To test this band assignment, we analyzed a mutated version of the full-length ribozyme in which the 3' strand contains HEG P1-disrupting substitutions (Fig. 4B) that are not involved in formation of the active core (11). The mutated ribozyme showed similar behaviour in the native gel assay as the 166.22 variant (Fig. 3C), consistent with our assignment of the ALT band as a HEG P1-containing inactive conformer. An additional, diffuse band in the mutant ribozyme was shown in

a side-by-side native gel assay not to co-migrate with the ALT band (Fig. S1). In further support of our assignment of ALT as the inactive HEG P1-containing conformer, the ALT band was dominant irrespective of Mg^{2+} concentrations in the catalytically inactive variant 157.65 (data not shown). Moreover the full-length variant apparently forms dimers at high Mg^{2+} concentrations. The dimerization seems to depend on the formation of HEG P1 since significantly less dimer is formed in the HEG P1 mutant ribozyme.

The overall picture that emerges from the native gel assays is that folding of the DiGIR1 ribozyme *in vitro* is a partitioning between two alternative conformations mediated by formation of a hairpin at the 3′end of the RNA (HEG P1) that destabilizes the active core. The inactive conformation is favoured at low Mg^{2+} concentrations (C_m = 0.65 ± 0.2 mM) and co-exists with the active conformation at physiological and higher Mg^{2+} concentrations. Thus, the formation of the active conformation is in essence shifted towards higher Mg^{2+} concentrations by the ability to form HEG P1 (active conformation $C_{m166.22}$ = 1.2 ± 0.5 mM vs. $C_{m166.65}$ = 4.33 ± 0.32 mM; Fig. 4C).

Fe-EDTA probing of the full-length ribozyme at variable Mg^{2+} concentration

Next, we applied Fe-EDTA probing to the full-length variant of the ribozyme at varying Mg^{2+} concentrations. These included conditions that were shown by native gel assays to be enriched in the inactive and active conformations, respectively. Fe-EDTA probing is based on the generation of hydroxyl radicals that result in ribose oxidation and cleavage of the RNA backbone in solvent-exposed regions whereas the backbone of residues buried in the interior of a folded RNA are protected from cleavage (Fig. 5A) (27).

At low Mg^{2+} concentrations, the core of the ribozyme, including all junctions, is highly accessible indicating an open conformation (Fig. 5B) consistent with the low mobility in the native gel assay. The helical segments are generally accessible. These signals are unaffected by increases in the Mg^{2+} concentration indicating that helices are formed at low Mg^{2+} concentrations. Strikingly, residues at the 3' end show very little reactivity at low Mg^{2+} concentration. These residues are in a single-stranded region in the active conformation and form parts of the loop and stem structure of HEG P1 in the inactive conformation. Thus, the data are more consistent with the inactive conformation and suggests that HEG P1 is involved in formation of specific interactions

with other parts of the ribozyme in order to stabilize the inactive conformation, perhaps by acting as a tetraloop receptor for the L9 GAAA loop (Fig. S2) (Birgisdottir et al., submitted).

Increasing Mg^{2+} concentration results in changes in accessibilities at approximately half of the residues (Figs. 5B and 6) indicating that a major reorganization of the ribozyme takes place between the inactive and the active conformation. Importantly, most of the residues found at junctions and some of the loops become less accessible consistent with the compaction of the ribozyme observed in the native gel assay. For example, the L6 loop that caps P6 has previously been shown to make a tertiary interaction with a receptor located in P3. This is supported by chemical probing, mutation analyses (14) and modelling (26) and is here confirmed by the Fe-EDTA experiment that shows parallel protection of the two elements when the Mg^{2+} concentration increases (Fig. 6).

In summary, Fe-EDTA experiments show that the core of the ribozyme is folded into an open conformation at low Mg^{2+} concentration possibly stabilized by specific interactions involving HEG P1. At increasing Mg^{2+} concentrations, HEG P1 is destabilized leading to the establishment of an alternative base pairing scheme that involves formation of P2 and P10. This in turn allows the organization of the P2-P2.1-P10 3WJ and the subsequent establishment of tertiary interactions that leads to compaction of the ribozyme and formation of the active site.

Modelling of the full-length ribozyme

We have previously modelled the core domain of the DiGIR1 ribozyme based on chemical and enzymatic probing and mutation analysis (14). We also modelled the P2-P2.1-P10 domain but were unable to unambiguously place this domain in relation to the core. With the Fe-EDTA data at hand, we compared the solvent accessibility prediction computed from the 3D coordinates of the core model with the experimental hydroxyl radical results (Fig. 7A). The accessibility prediction was generally in good agreement with the data except at the 5' and 3'ends of P10 and at the two crucial junctions J15/7 and J9/10, which are parts of the active site. These elements were predicted to be accessible in the model while the FE-EDTA experiment showed them to become less accessible when the Mg^{2+} concentration was increased (Fig. 7A). In parallel with this, the 3' part of L2.1 and residues involved in the 3WJ between P2, P2.1 and P10 also became less accessible (Figs. 5B and 6). Structure modelling was applied to find a solution to the 3WJ that

could explain these observations. If P2 stacks coaxially with P10 and P2.1 is parallel to P10 (i.e. family B of 3WJ (28)), P2.1 will fold over the core of the ribozyme and protect the J15/7 junction and two nucleotides of the 3' strands of P15 and P7, respectively (Fig. 7B). This model satisfies the FE-EDTA results (Fig. 7A) and suggests that P2.1 plays an active role in stabilizing the active site of the ribozyme through contacts between the 3' part of the loop and key residues at the active site. The exact nature of these contacts was not revealed by the present experiments.

Inactive conformation in the full length intron

The conformational switching of the DiGIR1 ribozyme mediated by the P2-P2.1-P10 domain is related to reactions that take place in the twin-ribozyme intron context (e.g. transcription and splicing) and may not be faithfully represented in the 166.65 length variant of DiGIR1. Thus, we decided to probe the conformation of DiGIR1 by chemical modification and RNase cleavage of a transcript comprising the full-length twin-ribozyme intron and flanking exons (Fig. 8). The experiment was performed by probing of the transcript as folded during *in vitro* transcription without any denaturation or renaturation steps. This analysis confirmed the general base pairing scheme of most part of the core domain, including the P5-P4-P6 and P7-P3-P8 helical stacks and the P3/P15 pseudoknot. The main difference compared to previous data related to isolated DiGIR1 length variants consists in the chemical modifications of all nucleotides in P9. Of particular importance, P10 was found to be single-stranded by chemical modifications of the 5' strand bases and the absence of V1 cleavage. P2.1 and HEG P1 were supported by V1 cleavages within stems and chemical modifications of loop nucleotides. Thus, in its natural sequence context, DiGIR1 folds primarily into its inactive conformation during transcription *in vitro*.

Discussion

Structure and folding of the active DiGIR1 ribozyme

Previous work has shown the close structural relationship between DiGIR1 and eubacterial group I (subgroup IC3) introns at the second step of splicing (4;5;26). The present work extends this to the folding pathways of the two classes of ribozymes. Furthermore, the Fe-EDTA results allow for incorporation into the model of the essential P2-P2.1 domain that is not found in group I introns. Group I introns are composed of three helical stacks (29;30). The P4-P6 is an early folding domain (31;32) that can act as a scaffolding domain promoting the folding of the P3-P9

catalytic domain (33;34). These domains together form a cleft into which the P1-P2 substrate domain docks (29;30). In addition, a number of optional peripheral domains serve to stabilize the structure (35). Our analysis of the Mg^{2+} dependence of folding by Fe-EDTA probing shows that the very short P4-P6 domain in DiGIR1 is folded at low Mg^{2+} concentration suggesting that it may have a similar role. Similar to group I introns, the three principal domains are largely folded individually at low Mg^{2+} concentration and the tertiary folding of the core established only at increased Mg^{2+} concentration. This involves for example the L6-P3 interaction and the J5/4 junction on which the substrate domain docks. The J6/7 junction becomes buried in the interior whereas the J3/4 becomes more exposed to the solvent (Fig 6). These observations are all consistent with group I intron folding and with our previous work (14;26). In the group I intron from *Tetrahymena*, the folding of the P3/P7 pseudoknot has been shown to be rate limiting due to mispairing of the 3'-strand of P3 and the highly conserved J8/7 junction (36). This mispairing is not found in DiGIR1 which presents a very different topology of the core. Here, the J8/7 sequence is engaged in the formation of the P15 stem with the equivalent of the 5' strands of P1 and P2 as well as J15/7 which participates in the definition of the active site (26). The formation of the P3/P15 pseudoknot prevents the formation of alternative structures within the core of the ribozyme. This pseudoknot as well as the group I-specific P3/P7 pseudoknot are completely folded at low Mg^{2+} concentration.

Another main characteristic of DiGIR1 is the presence of the P2-P2.1 domain that is phylogenetically unrelated to group I intron domains, yet has an important biological role (discussed below). This domain has been shown to be essential for DiGIR1 activity by mutational analysis (4;12) even though it is located away from the active site in the secondary structure. Specifically, the length of the P2.1 stem and the size of the L2.1 loop were found to be important whereas the sequence was of minor importance. In relation to folding of the active conformation, P2 is only formed at intermediate and high Mg^{2+} concentrations. P2 formation allows for the organization of the P2-P2.1-P10 3WJ that imposes architectural constraints on the three-dimensional helix organization (28;37) resulting in the docking of P2.1 onto the core. This view is supported by concomitant Mg^{2+} dependent protection from hydroxyl radical cleavage of parts of the core and parts of the P2-P2.1 domain (Fig. 5). In the core, the active site junctions J15/7 and J9/10 become protected together with parts of P9, P10 and P15. In the P2-P2.1 domain, protection is seen in 3' part of L2.1 and at the 3WJ. Together, these observations allowed us to model the placement of the P2-P2.1 domain on the core (Fig. 7). The domain is organized by a 3WJ of family B (28) that

places the DiGIR1-DiGIR2 connecting stem, P2, away from the core and P2.1 over the core. Presumably, the interactions that take place involve only the backbone because the sequence of P2.1 was found to be relatively unimportant (12). The bases of the 3' part of L2.1 are buried in the structure consistent with previous data from chemical modification experiments (12). Overall, DiGIR1 appears as a ribozyme that folds directly into a near-native structure of the core that requires further stabilization by a regulatory domain in order to be active.

Two biologically relevant folds of DiGIR1

Folding of RNA *in vivo* and *in vitro* differs in several key aspects. Folding *in vivo* is co-transcriptional and strongly dependent on the elongation rate that influences the sequence of the nucleotides available for folding at any given time (38). This is not recapitulated *in vitro* mainly because of use of heterologous RNA polymerases, e.g. the phage polymerases that have much higher polymerization rates than their eukaryotic counterparts. Folding *in vivo* furthermore takes place in an environment that cannot faithfully be reproduced *in vitro* in particular with respect to the proteins that may become associated with the nascent RNA molecules. Based on this, it is not surprising that folding of group I introns *in vivo* is much faster than *in vitro* and occurs essentially without misfolding (although the latter has been challenged (39)). Folding of group I ribozymes *in vitro* is characterized by partitioning between native (or near-native) and misfolded populations of molecules (need a reference i.e. (40)). One key question is whether the misfolded molecules are on a biologically relevant pathway or they simply reflect an *in vitro* artefact. In the *Tetrahymena* ribozyme, the switching between helices P-1 (5' exon only) and helix P1 (41) that contains the 5' splice site is orchestrated along the splicing pathway in a way that appears to secure correct folding of the exons prior to splicing (maybe another reference (36)). Similarly, a misfolded population that involves the *Azoarcus* intron and its flanking tRNA 3' exon has been described. Here, the misfolded species was characterized by pairing of P7-P9 residues with the downstream 3' exon as well as an alternative P3 pseudoknot. It was speculated that this would couple the folding of the tRNA and the intron and prevent 5' and 3' processing of pre-tRNAs that were not competent to splice (25). A different kind of misfolding of the *Tetrahymena* intron involves incorrect organization of helices in the tertiary structure rather than alternative base pairing (36). These misfolded species are entirely intron internal and have no apparent potential for a biological function.

In addition to the active conformation, DiGIR1 can fold into an inactive conformation characterized by correct folding of the three core domains and alternative base pairing of the peripheral domain P2-P2.1 (Fig. 8). Thus, we describe P2-P2.1 as a regulatory domain and conclude that the inactive conformation is biologically relevant because it keeps the branching ribozyme inactive until the twin-ribozyme intron is spliced out from the ribosomal RNA precursor in a reaction catalyzed by DiGIR2. The P2-P2.1 domain is the only part of DiGIR1 that is unrelated to group I introns and appears to have been specifically recruited to perform this function. It is co-transcriptionally favoured in the sense that it is composed of two local hairpins that are expected to form directly upon transcription. The inactive conformation is a distinct species that forms preferentially at intermediate Mg^{2+} concentrations reflecting physiological conditions and is suppressed by the active conformation at higher Mg^{2+} concentrations. This profile is similar to that found for group I introns that can form alternative base pairs and have been suggested to have a function other than splicing and distinctly different from those group I introns that are misfolded at the tertiary structure level. Importantly, the inactive conformation of DiGIR1 appears to be stabilized by a specific tertiary structure that is not found in the active conformation. This involves HEG P1 as a tetraloop receptor and most likely L9 as a tetraloop. Previous work has shown that HEG P1 can indeed function as a GAAA tetraloop-receptor *in vitro* (Birgisdottir, submitted) and it has been speculated that this interaction plays a role in pulling the lariat cap out of the active site following the branching reaction catalyzed by DiGIR1. In accordance with the proposed HEG P1-L9 interaction, mutations that disrupt HEG P1 increase the branching activity of the ribozyme (4). Mutations in P9 are more difficult to interpret because this hairpin also plays a role in the active conformation.

In conclusion, DiGIR1 has integrated two aspects of conformational switching that are represented in other parts of RNA biology. First, the formation of HEG P1 is at the level of the secondary structure and reminiscent of for example the *E. coli* cspA mRNA. Here, the mRNA adopts one secondary structure that promotes translation at low temperatures and a different structure that segregates the translational signals at higher temperatures (42). Second, the activation of DiGIR1 involves a tertiary structural reorganization namely the docking of P2.1 as a consequence of the organization of the 3WJ. Conformational switching involving changes in tertiary contacts are found in riboswitches in general (43).

The role of the P2-P2.1 domain in the twin-ribozyme intron context

The twin-ribozyme intron can be processed in three different ways depending on cellular conditions (13;44;45); Andersen, submitted) (Fig. 9). The three pathways differ in the order of reactions carried out by the splicing (DiGIR2) and the branching (DiGIR1) ribozymes. The splicing pathway is the predominant form of processing during cell growth. In this, DiGIR2 is first active in splicing while the activity of DiGIR1 is suppressed by formation of the inactive HEG P1-containing conformation. Following splicing, DiGIR1 undergoes conformational switching that involves the melting of HEG P1 and formation of P2-P10 and cleaves the intron in the branching reaction that results in lariat capping of the homing endonuclease mRNA. The release of the mRNA involves re-formation of HEG P1 that now becomes a structural element of the 5' UTR of the mRNA (11). There may be a transient association between the two cleavage products mediated by the HEG P1-L9 interaction as part of the release mechanism (Birgisdottir, submitted), but only the mRNA is exported to the cytoplasm (13). Thus, the switch between a P2-P2.1 and a P2.1-HEG P1 containing conformation operates in both directions in the splicing pathway and the tetraloop-tetraloop receptor interaction plays a role in both switching events.

A small fraction of twin-ribozyme intron undergoes 3' SS hydrolysis and subsequent formation of full-length circular intron without exon ligation during cell growth (44). During heat-shock, this processing pathway is significantly up-regulated (Andersen, submitted). In the circularization pathway, DiGIR2 catalyzes both steps whereas DiGIR1 is inactive and presumably folded in the inactive conformation characterized by HEG P1 (Hedegaard, unpublished observations). Only re-opening of the circles, e.g. by hydrolysis at the circularization junction will lead to activation of DiGIR1 in parallel with the activation mechanism operating in the splicing pathway. Finally, during starvation induced encystment, the splicing ribozyme, DiGIR2 is inactive, while DiGIR1 cleaves the ribosomal RNA precursor (45). This reaction is believed to provide the cell with a pre-mRNA encoding the homing endonuclease that can be further processed upon excystment although this has not been experimentally demonstrated.

During *in vitro* processing of the full-length twin-ribozyme intron, products of the splicing and circularization pathways accumulate in parallel and DiGIR1 cleaves the precursor as well as the spliced out intron. Thus, most of the regulatory aspects described above are lost suggesting that the cellular context is critical for regulation, perhaps through the action of specific

factors. Inspection of the *in vivo* processing pathways suggests that events at the 5' splice site are the key to understanding regulation. If a proper 5' splice site is not formed during transcription, the alternative pathways may take over in processing. During heat-shock and starvation, expression of ribosomal RNA is known to be down-regulated (46) (47). One aspect of this may be the inability to form a correct 5' splice site perhaps due to the lack of association of ribosomal proteins with precursor RNA. This may in turn allow the ωG nucleotide to bind the G-binding site in DiGIR2 in place of the guanosine co-factor in splicing and activate the circularization pathway. Similarly, starvation conditions may relieve the repression of DiGIR1 and activate cleavage of the precursor. Together, such relatively simple mechanisms that mostly rely on alternative RNA conformations may constitute an adaptation that regulate host versus intron functions. We suggest that the P2-P2.1 domain is such a regulatory domain that is central to understanding the activity of the branching ribozyme within the twin-ribozyme introns. In this way, our study provides another example that the modular aspect of RNA structure allows for coupling of modules that carry different functions and are of different origin.

Legends to figures

Figure 1: Organization of the DiGIR1 ribozyme. DiGIR1 is located within a group I twin-ribozyme intron in the small subunit (SSU) part of the ribosomal precursor (pre-rRNA) that in addition harbours two group I introns in the large subunit (LSU) (top). Within the twin ribozyme intron, DiGIR1 is inserted together with a downstream homing endonuclease (HE) encoding part inside the splicing ribozyme DiGIR2 (bottom). SS: Splice Site. IPS: Internal Processing Site.

Figure 2: DiGIR1 can be divided into a core and a regulatory domain. The active conformation (**A**) is characterized by a base pairing scheme that incorporates a P10 and a P2 element and has the branch point (BP) nucleotide U232 positioned at the active site for attack at the internal processing site (IPS). (**B**) The inactive conformation is characterized by the presence of the stem-loop structure HEG P1 that excludes the possibility of formation of P10 and P2. As a consequence, the active site is disordered. The stretch of nucleotides involved in alternative base pairing is coloured in red. Open arrows denote the end points of the length variants used in the present study.

Figure 3: Native gel analysis of the minimal version of DiGIR1. (**A**) Gel of minimal version (157.22) and (**B**) of a variant with a slightly longer 5' part (166.22) at different Mg^{2+} concentrations. The labelled bands were interpreted as unfolded (U), partially folded (ni), and near-native (N), respectively. (**C**) The fraction of native RNA (f_N) was determined and fit to the Hill equation. DiGIR1-157.22: $C_{m157.22}$ = 1.1 ± 0.37 mM, DiGIR1-166.22: $C_{m166.22}$ = 1.2 ± 0.5 mM, and the *Azoarcus* ribozyme: C_m = 0.35 ± 0.2 mM. All the RNAs subjected to the native gel experiments were analyzed three times in parallel and the results were highly reproducible.

Figure 4: Native gel analysis of the full-length version of DiGIR1. (**A**) Gel of full-length version (166.65) and (**B**) of a mutant ribozyme (166.65HEG P1). Compared to Fig. 3, two additional bands appeared. These were interpreted as representing dimers, and an alternative, inactive conformation (ALT), respectively. Below the gel picture in B is shown a helix diagram highlighting the substitution in the mutant ribozyme that completely disrupts HEG P1. (**C**) The fraction of native RNA (f_N) was determined and fit to the Hill equation. DiGIR-166.65: $C_{m166.65}$= 4.33 ± 0.32 mM); DiGIR1-166.65HEG-P1:, $C_{m166.65HEGP1}$ = 1.2 ± 0.5 mM. All the RNAs subjected to the native gel experiments were analyzed three times in parallel and the results were highly reproducible.

Figure 5: Fe-EDTA probing of full-length DiGIR1. (**A**) Autoradiogram showing a sample of the data (analysis of 3' end labelled RNA revealing most of the core). (**B**) Summary of the hydroxyl radical reactivity at two selected Mg^{2+} concentrations representing inactive (2 mM Mg^{2+}) and active (25 mM Mg^{2+}) conformations, respectively. The data were plotted on helix diagrams representing the active conformation.

Figure 6: Summary of the hydroxyl radical reactivity showing residues with increased (red) and decreased (blue) reactivity upon increase in the Mg^{2+} concentration. The helix diagram shows the active conformation with an indication of the main tertiary interactions (boxed).

Figure 7: Docking of P2.1 onto the core. (**A**) Comparison of predicted and experimental solvent accessibility between the DiGIR1 model with and without the P2-P2.1 domain and the accessibility as determined by Fe-EDTA probing. Solvent accessibility (in $Å^2$) of C4' atoms were computed from the 3D coordinates of the models by using a rolling sphere of radius 2.8 Å. The straight line at 12 $Å^2$ indicates the average accessibility of a C4' atom within an isolated and regular A-form RNA

helix. Nucleotides protected from hydroxyl radical cleavage are coloured in blue while the accessible residues are in red. (**B**) Core ribozyme 3D model with the same colour code as in (A). The shaded areas represent the parts protected by P2.1.

Figure 8: Structure probing of DiGIR in the full-length twin-ribozyme intron context. (**A**) Summary of structure probing data based on chemical modification or enzymatic cleavage followed by primer extension analysis superimposed on the helix diagram representing the inactive conformation of the ribozyme. BP: Branch Point. IPS: Internal Processing Site. (**B**) Autoradiogram showing structure probing data of the HEG P1 part of the inactive conformation as found in the context of the full-length twin-ribozyme intron.

Figure 9: Summary of the role of HEG P1 in three alternative processing pathways of the twin-ribozyme intron.

Figure S1: Native gel analysis of the full-length version (166.65) and the mutant ribozyme (166.65HEG P1).

Figure S2: Comparison of predicted and experimental solvent accessibility between the DiGIR1 L9 HEG P1 model (Birgisdottir et al., submitted) and the accessibility as determined by Fe-EDTA probing at low Mg^{2+}. (A) Solvent accessibility (in $Å^2$) of C4' atoms were computed from the 3D coordinates of the models by using a rolling sphere of radius 2.8 Å using the DiGIR1 L9 HEG P1 model (Birgisdottir et al., submitted). The straight line at 12 $Å^2$ indicates the average accessibility of a C4' atom within an isolated and regular A-form RNA helix. (B) Plotting accessibility determined by Fe-EDTA probing at low Mg^{2+} onto the HEG P1 secondary structure. (C) DiGIR1 L9 HEG P1 interaction molecular model (Birgisdottir et al., submitted).

Figure S3: Global overview of the DiGIR1 model harbouring the P2P2.1 domain. The DiGIR1 model is in represented in surface while the P2(green) and P2.1 (orange) are represented in stick.

References

1. Lilley DMJ and Eckstein F (2008) *Ribozymes and RNA Catalysis*. The Royal Society of Chemistry.

2. Martick,M., Horan,L.H., Noller,H.F. and Scott,W.G. (2008) A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. *Nature*, **454**, 899-902.

3. Salehi-Ashtiani,K., Luptak,A., Litovchick,A. and Szostak,J.W. (2006) A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science*, **313**, 1788-1792.

4. Johansen,S., Einvik,C. and Nielsen,H. (2002) DiGIR1 and NaGIR1: naturally occurring group I-like ribozymes with unique core organization and evolved biological role. *Biochimie*, **84**, 905-912.

5. Nielsen,H., Beckert,B., Masquida B. and Johansen,S.D. (2008) The GIR1 branching ribozyme. In Lilley,D.M.J. and Eckstein,F. (eds.), *Ribozymes and RNA catalysis*. The Royal Society of Chemistry, London, UK, pp. 229-252.

6. Wikmark,O.G., Haugen,P., Lundblad,E.W. and Haugli,K. (2007) The molecular evolution and structural organization of group I introns at position 1389 in nuclear small subunit rDNA of myxomycetes. *J Eukaryot Microbiol*, **54 %6**, 49-56.

7. Nielsen,H., Westhof,E. and Johansen,S. (2005) An mRNA is capped by a 2', 5' lariat catalyzed by a group I-like ribozyme. *Science*, **309**, 1584-1587.

8. Zhang,F., Ramsay,E.S. and Woodson,S.A. (1995) In vivo facilitation of Tetrahymena group I intron splicing in Escherichia coli pre-ribosomal RNA. *Rna*, **1**, 284-292.

9. Brehm,S.L. and Cech,T.R. (1983) Fate of an intervening sequence ribonucleic acid: excision and cyclization of the Tetrahymena ribosomal ribonucleic acid intervening sequence in vivo. *Biochemistry*, **22**, 2390-2397.

10. Uptain,S.M., Kane,C.M. and Chamberlin,M.J. (1997) Basic mechanisms of transcript elongation and its regulation. *Annu Rev Biochem*, **66**, 117-172.

11. Nielsen,H., Einvik,C., Lentz,T.E., Hedegaard,M.M. and Johansen,S.D. (2009) A conformational switch in the DiGIR1 ribozyme involved in release and folding of the downstream I-DirI mRNA. *Rna*, **15**, 958-967.

12. Einvik,C., Nielsen,H., Nour,R. and Johansen,S. (2000) Flanking sequences with an essential role in hydrolysis of a self-cleaving group I-like ribozyme. *Nucleic Acids Res*, **28**, 2194-2200.

13. Vader,A., Nielsen,H. and Johansen,S. (1999) In vivo expression of the nucleolar group I intron-encoded I-dirI homing endonuclease involves the removal of a spliceosomal intron. *EMBO J*, **18**, 1003-1013.

14. Einvik,C., Nielsen,H., Westhof,E., Michel,F. and Johansen,S. (1998) Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites. *Rna*, **4**, 530-541.

15. Rangan,P., Masquida,B., Westhof,E. and Woodson,S.A. (2003) Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proc Natl Acad Sci U S A*, **100**, 1574-1579.

16. Johansen,S. and Vogt,V.M. (1994) An intron in the nuclear ribosomal DNA of Didymium iridis codes for a group I ribozyme and a novel ribozyme that cooperate in self-splicing. *Cell*, **76**, 725-734.

17. Shcherbakova,I. and Brenowitz,M. (2008) Monitoring structural changes in nucleic acids with single residue spatial and millisecond time resolution by quantitative hydroxyl radical footprinting. *Nat Protoc*, **3**, 288-302.

18. Emerick,V.L. and Woodson,S.A. (1994) Fingerprinting the folding of a group I precursor RNA. *Proc Natl Acad Sci U S A*, **91**, 9675-9679.

19. Laederach,A., Das,R., Vicens,Q., Pearlman,S.M., Brenowitz,M., Herschlag,D. and Altman,R.B. (2008) Semiautomated and rapid quantification of nucleic acid footprinting and structure mapping experiments. *Nat Protoc*, **3**, 1395-1401.

20. Kjems,J., Egebjerg,J. and Christiansen,J. (1998) *Laboratory Techniques in Biochemistry and Molecular Biology: Analysis of RNA-Protein Complexes In Vitro*. Elsevier, Amsterdam, The Netherlands.

21. Jossinet,F., Ludwig,T.E. and Westhof,E. (2010) Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics*.

22. Bindewald,E., Hayes,R., Yingling,Y.G., Kasprzak,W. and Shapiro,B.A. (2008) RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Res*, **36**, D392-D397.

23. DeLano,W.L. The PyMOL Molecular Graphics System. 2002.
Ref Type: Computer Program

24. Chauhan,S. and Woodson,S.A. (2008) Tertiary interactions determine the accuracy of RNA folding. *J Am Chem Soc*, **130**, 1296-1303.

25. Rangan,P., Masquida,B., Westhof,E. and Woodson,S.A. (2004) Architecture and folding mechanism of the Azoarcus Group I Pre-tRNA. *J Mol Biol*, **339**, 41-51.

26. Beckert,B., Nielsen,H., Einvik,C., Johansen,S.D., Westhof,E. and Masquida,B. (2008) Molecular modelling of the GIR1 branching ribozyme gives new insight into evolution of structurally related ribozymes. *EMBO J*, **27**, 667-678.

27. Latham,J.A. and Cech,T.R. (1989) Defining the inside and outside of a catalytic RNA molecule. *Science*, **245**, 276-282.

28. Lescoute,A. and Westhof,E. (2006) Topology of three-way junctions in folded RNAs. *Rna*, **12**, 83-93.

29. Adams,P.L., Stahley,M.R., Kosek,A.B., Wang,J. and Strobel,S.A. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45-50.

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

31. Downs,W.D. and Cech,T.R. (1990) An ultraviolet-inducible adenosine-adenosine cross-link reflects the catalytic structure of the Tetrahymena ribozyme. *Biochemistry*, **29**, 5605-5613.

32. Zarrinkar,P.P. and Williamson,J.R. (1994) Kinetic intermediates in RNA folding. *Science*, **265**, 918-924.

33. Murphy,F.L. and Cech,T.R. (1993) An independently folding domain of RNA tertiary structure within the Tetrahymena ribozyme. *Biochemistry*, **32**, 5291-5300.

34. Zarrinkar,P.P. and Williamson,J.R. (1996) The kinetic folding pathway of the Tetrahymena ribozyme reveals possible similarities between RNA and protein folding. *Nat Struct Biol*, **3**, 432-438.

35. Lehnert,V., Jaeger,L. and Michel,F. (1996) New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the Tetrahymena thermophila ribozyme. *Chem Biol*, **3 %6**, 993-1009.

36. Pan,J. and Woodson,S.A. (1998) Folding intermediates of a self-splicing RNA: mispairing of the catalytic core. *J Mol Biol*, **280**, 597-609.

37. de la Pena,M., Dufour,D. and Gallego,J. (2009) Three-way RNA junctions with remote tertiary contacts: a recurrent and highly versatile fold. *Rna*, **15**, 1949-1964.

38. Koduvayur,S.P. and Woodson,S.A. (2004) Intracellular folding of the Tetrahymena group I intron depends on exon sequence and promoter choice. *Rna*, **10**, 1526-1532.

39. Jackson,S.A., Koduvayur,S. and Woodson,S.A. (2006) Self-splicing of a group I intron reveals partitioning of native and misfolded RNA populations in yeast. *Rna*, **12**, 2149-2159.

40. Woodson,S.A. (2002) Folding mechanisms of group I ribozymes: role of stability and contact order. *Biochem Soc Trans*, **30**, 1166-1169.

41. Woodson,S.A. (1992) Exon sequences distant from the splice junction are required for efficient self-splicing of the Tetrahymena IVS. *Nucleic Acids Res*, **20**, 4027-4032.

42. Giuliodori,A.M., Di,P.F., Marzi,S., Masquida,B., Wagner,R., Romby,P., Gualerzi,C.O. and Pon,C.L. (2010) The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. *Mol Cell*, **37**, 21-33.

43. Serganov,A. (2009) The long and the short of riboswitches. *Curr Opin Struct Biol*, **19**, 251-259.

44. Nielsen,H., Fiskaa,T., Birgisdottir,A.B., Haugen,P., Einvik,C. and Johansen,S. (2003) The ability to form full-length intron RNA circles is a general property of nuclear group I introns. *Rna*, **9**, 1464-1475.

45. Vader,A., Johansen,S. and Nielsen,H. (2002) The group I-like ribozyme DiGIR1 mediates alternative processing of pre-rRNA transcripts in Didymium iridis. *Eur J Biochem*, **269**, 5804-5812.

46. Mackow,E.R. and Chang,F.N. (1983) Correlation between RNA synthesis and ppGpp content in Escherichia coli during temperature shifts. *Mol Gen. Genet*, **192**, 5-9.

47. Cashel,C. and Rudd,K.E. (1987) The stringent response. In Neidhardt,F.C., Ingraham,J.C., Low,K.B., Magasanik,B., Schaechter,M. and Umbarger,H.E. (eds.), *Escherichia coli and Salmonella typhimurium: cellular and molecular biology*. Washington, D.C., pp. 1410-1438.

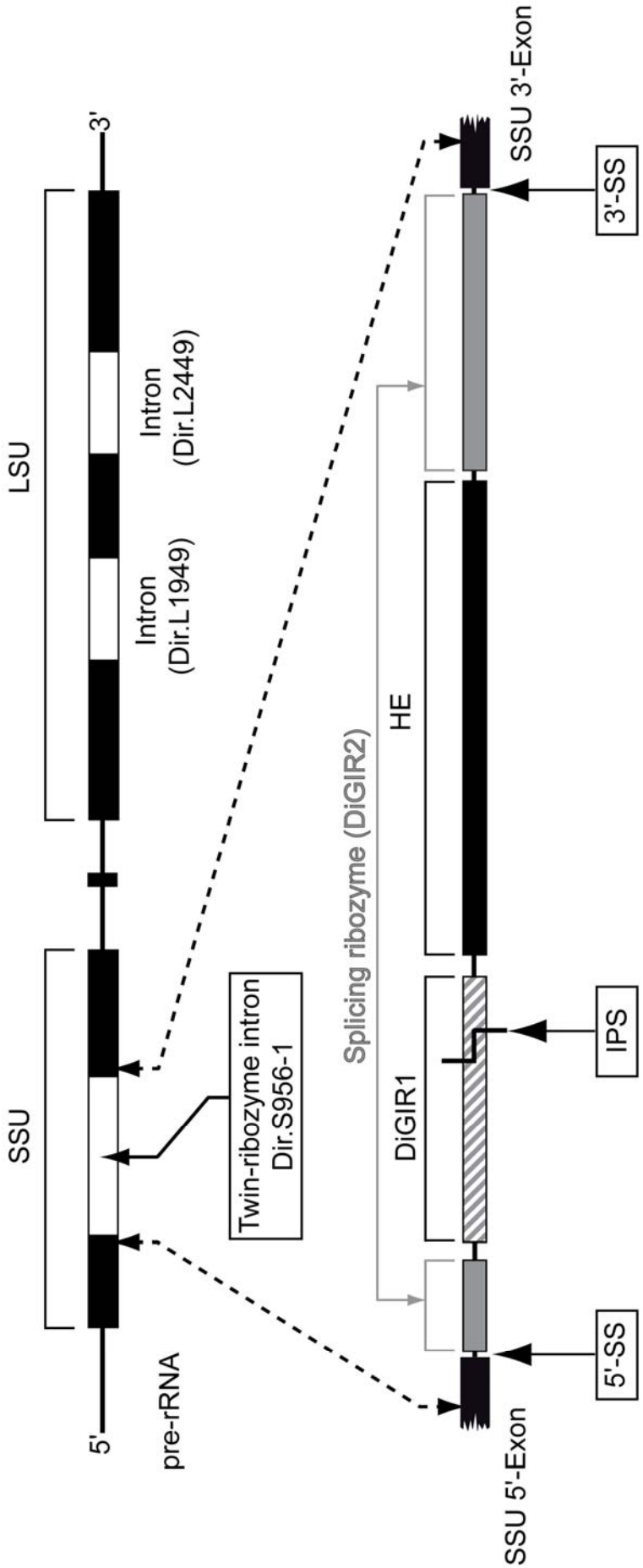


Figure 1 Beckert *et al.*

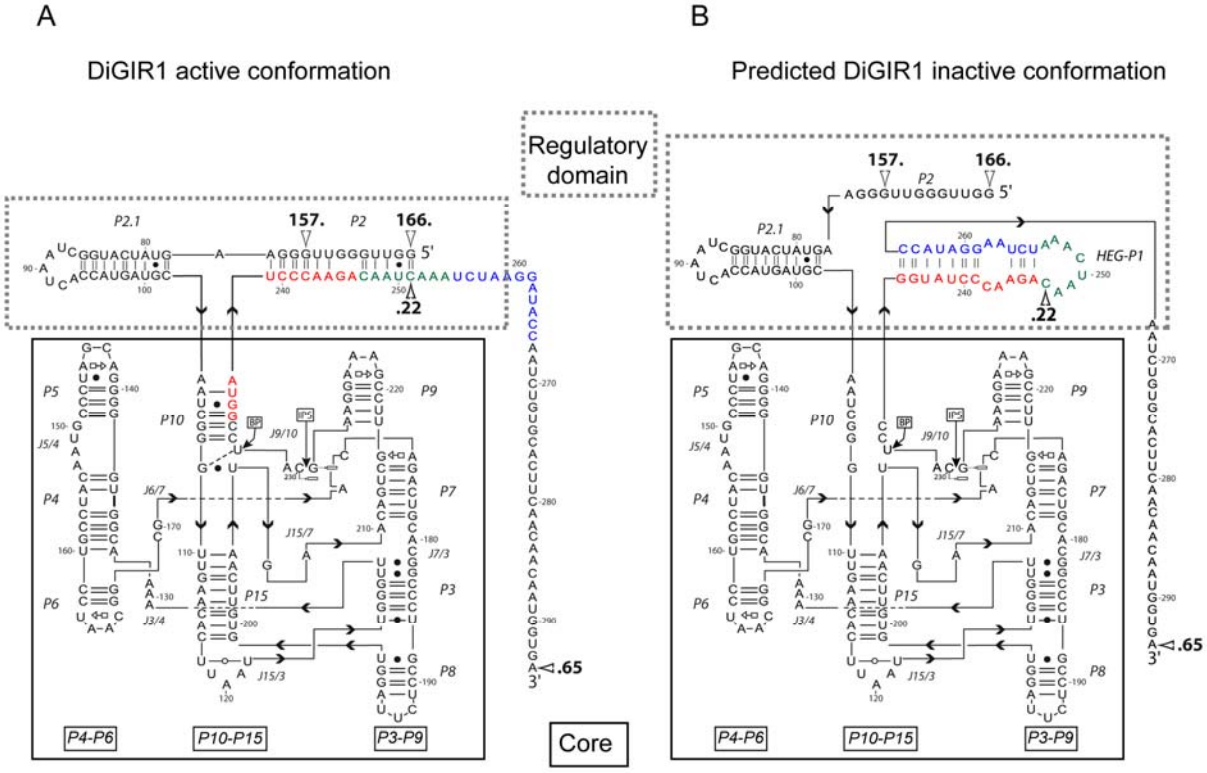


Figure 2 Beckert *et al.*

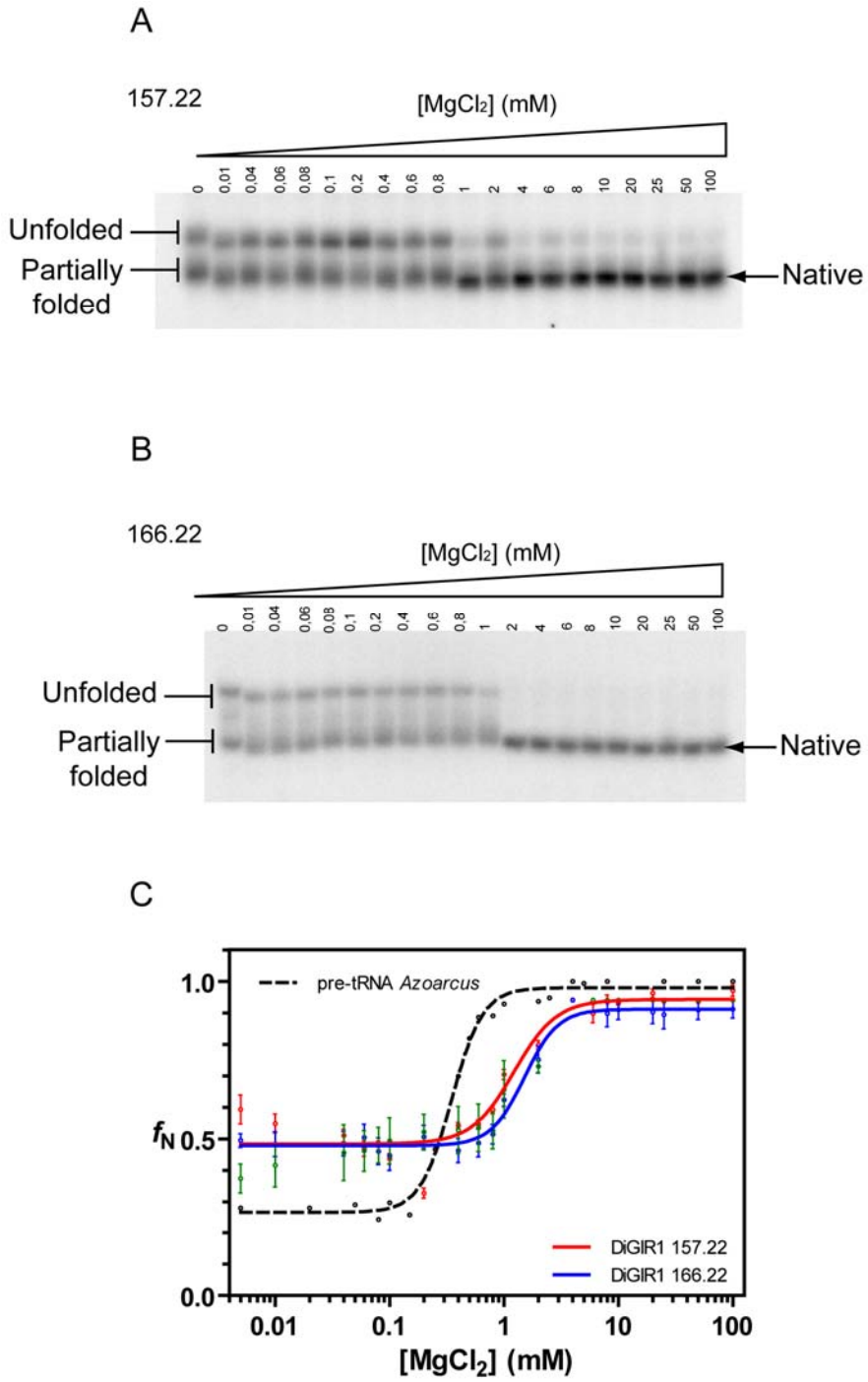


Figure 3 Beckert *et al.*

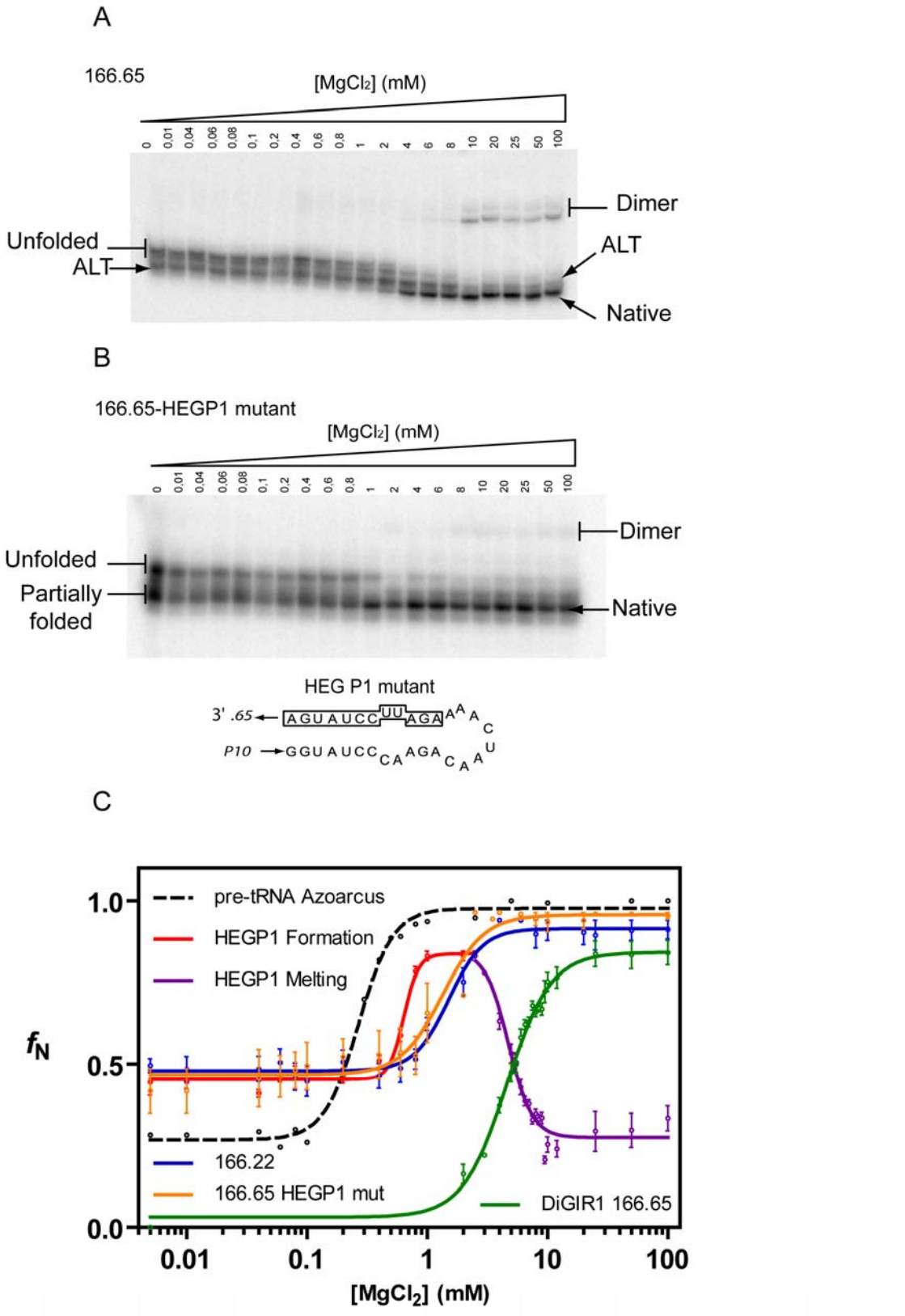


Figure 4 Beckert *et al.*

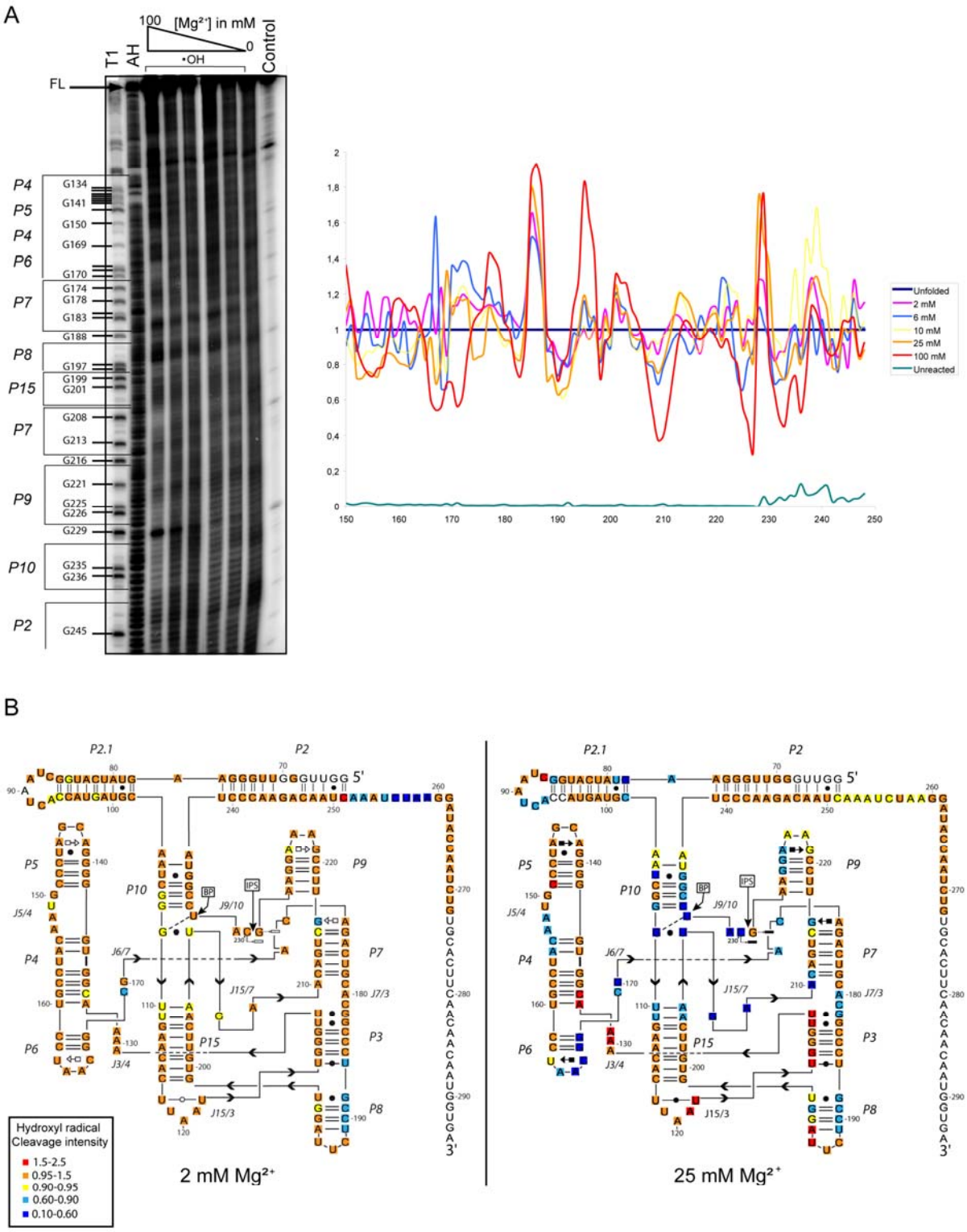


Figure 5 Beckert *et al.*

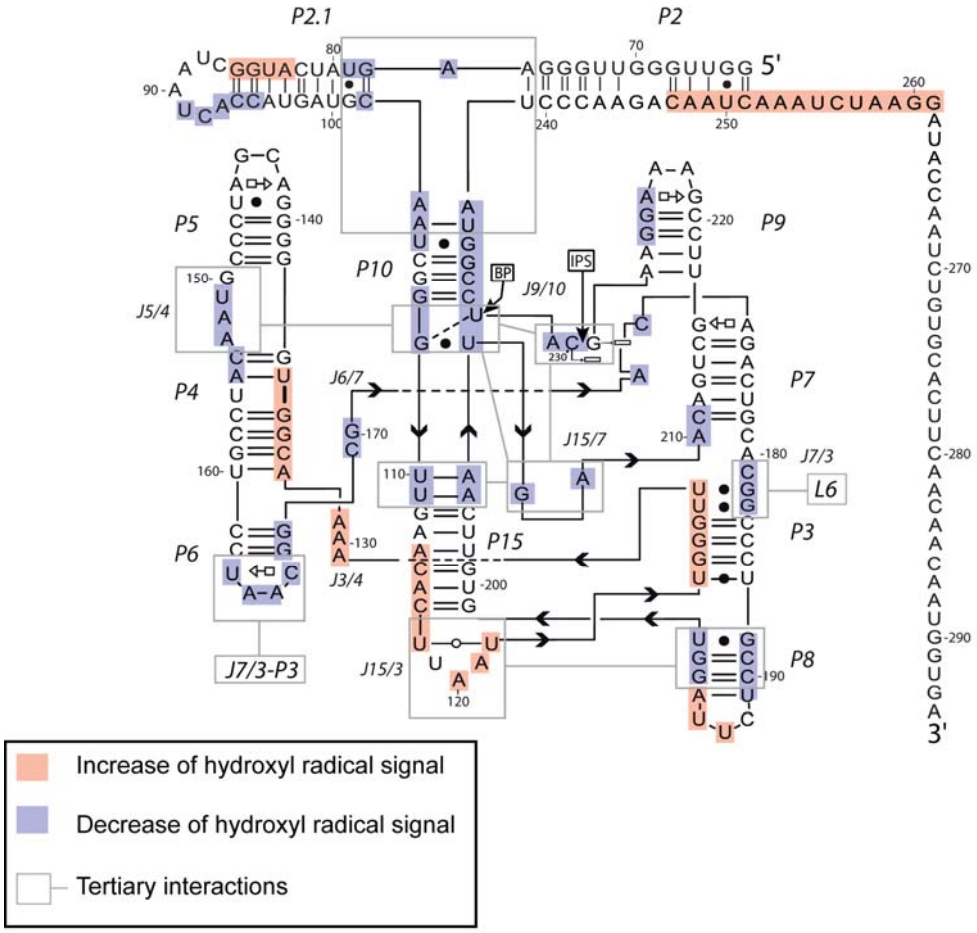


Figure 6 Beckert *et al.*

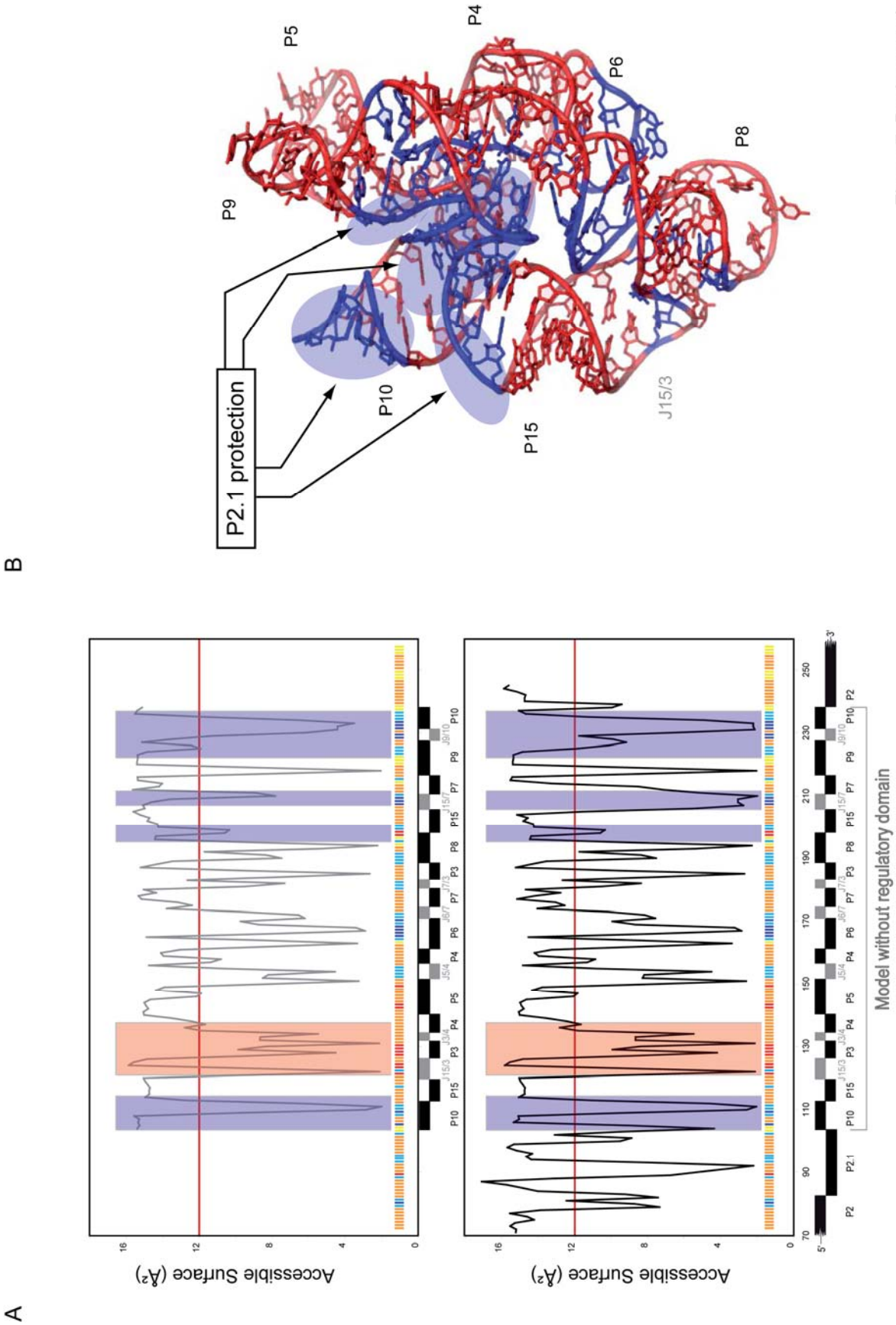


Figure 7 Beckert *et al.*

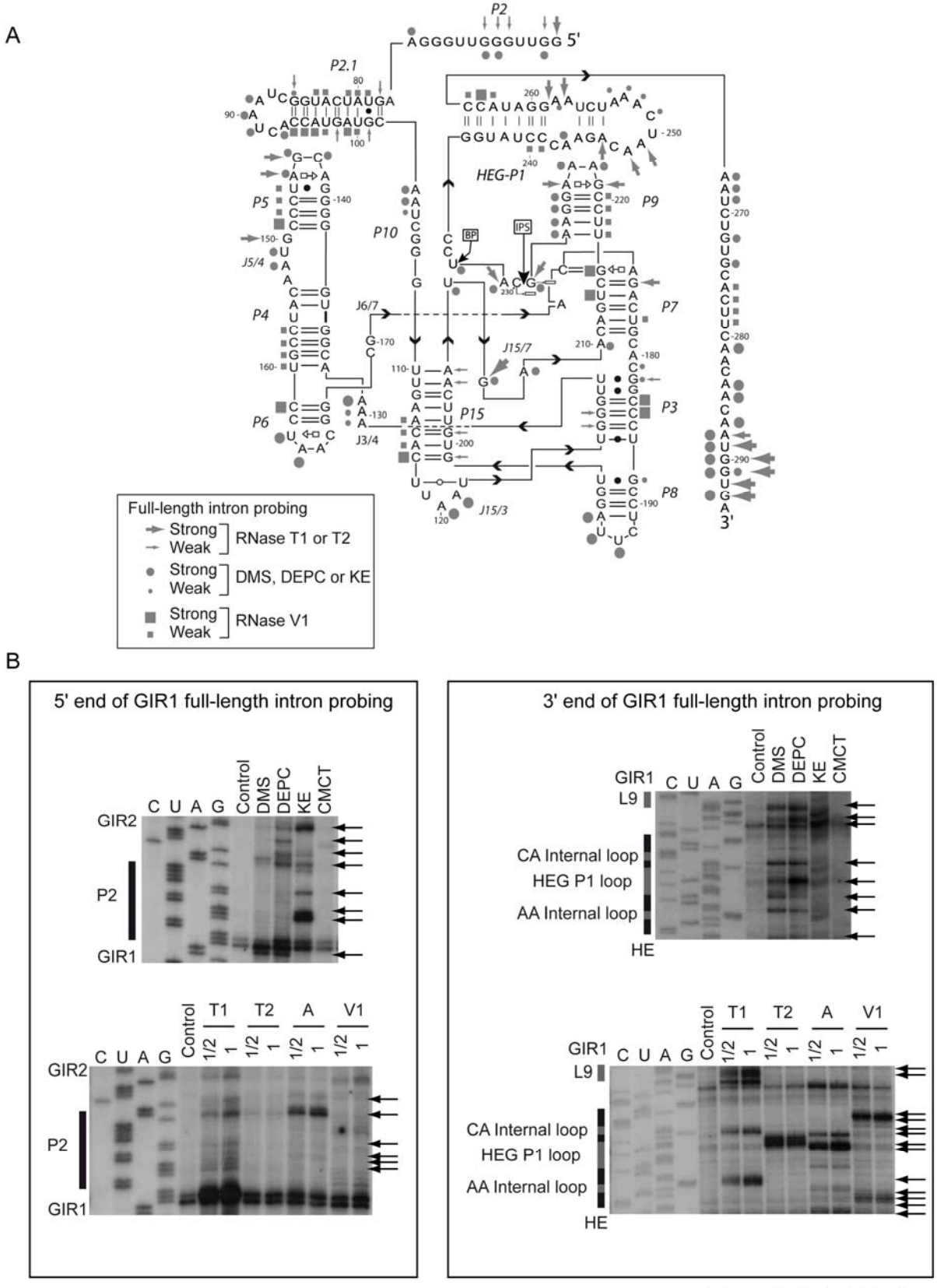


Figure 8 Beckert *et al.*

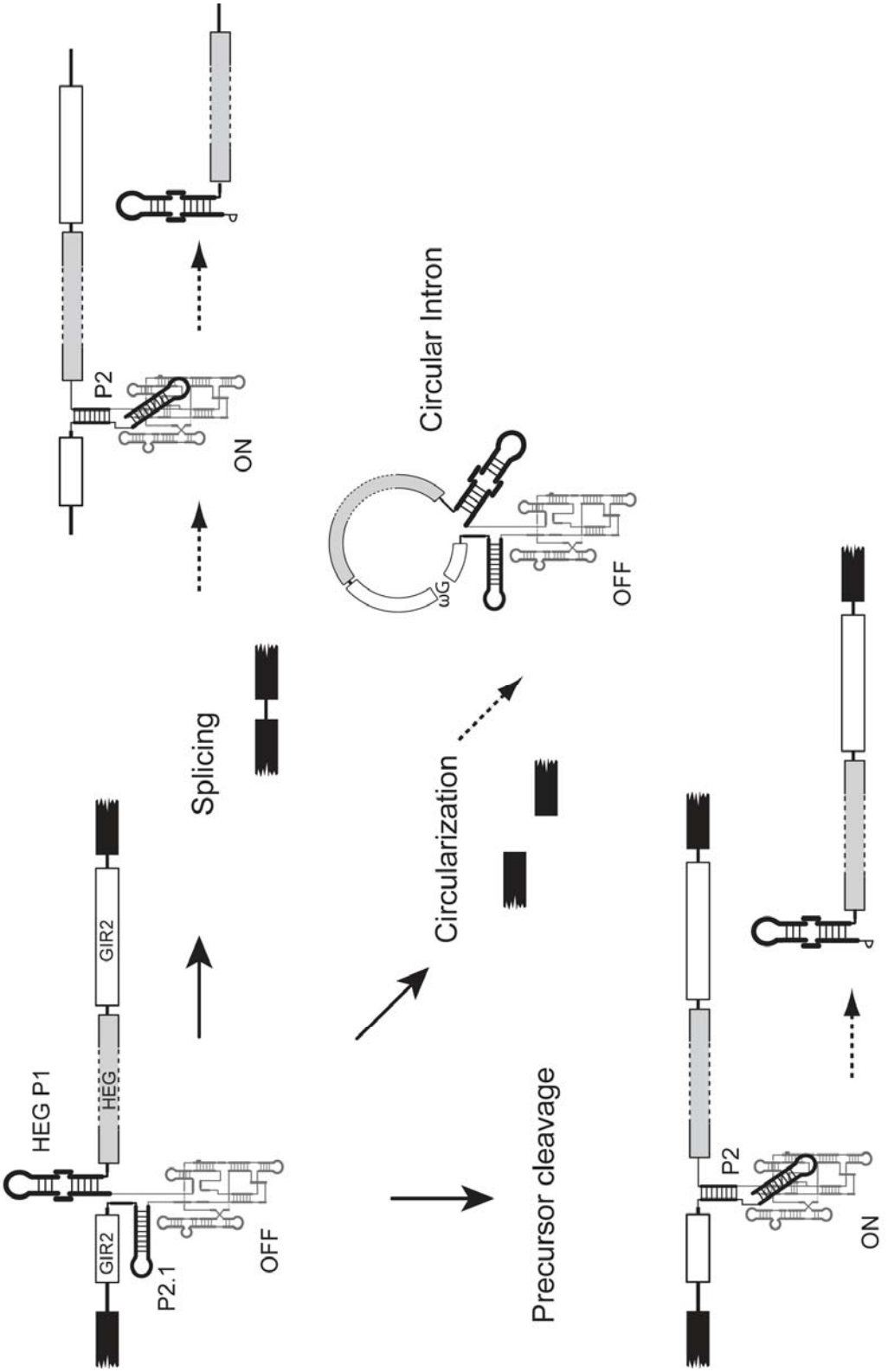


Figure 9 Beckert *et al.*

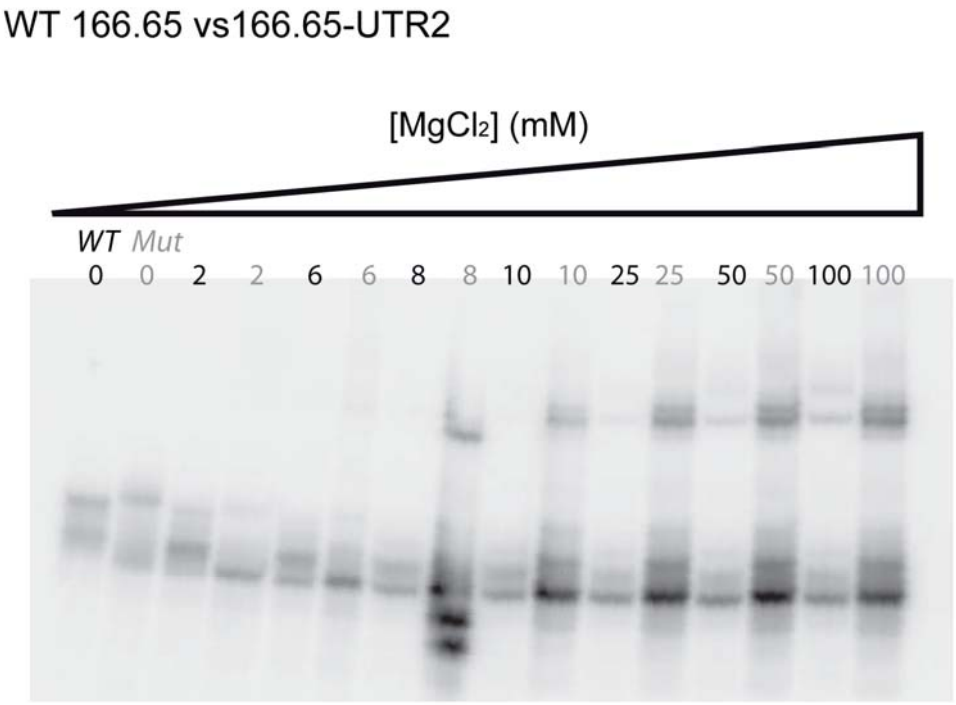


Figure S1 Beckert *et al.*

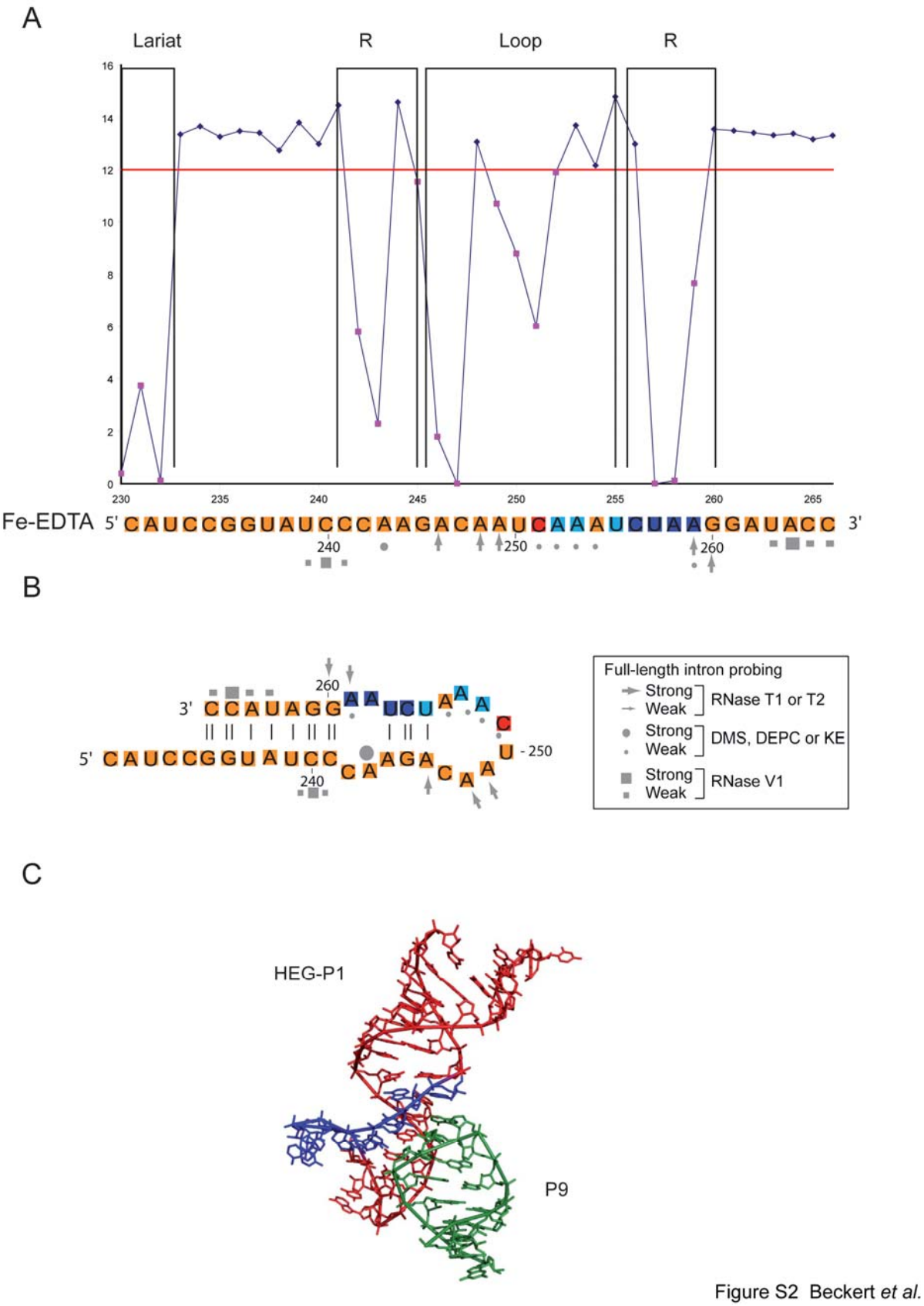


Figure S2 Beckert *et al.*

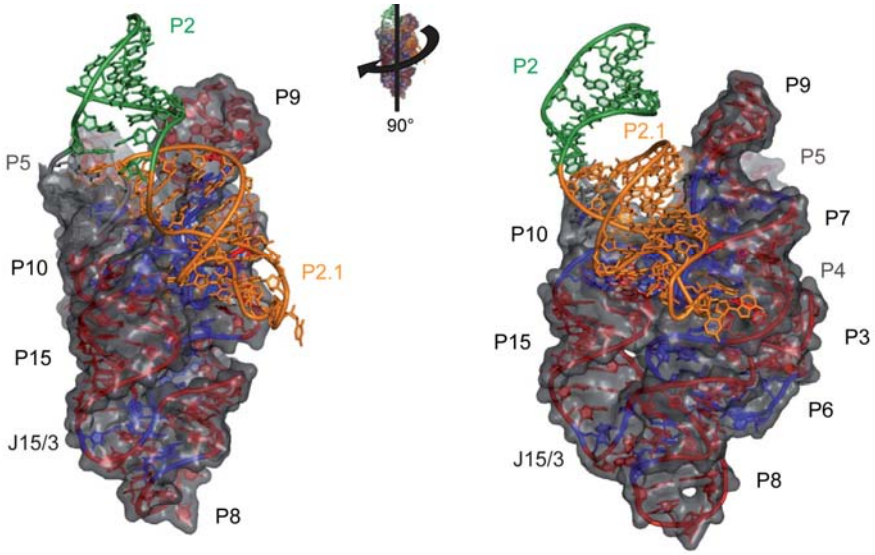


Figure S3 Beckert *et al.*

ARTICLE IV:

Accumulation of stable full-length circular group I introns during heat-shock

K. L. Andersen, B. Beckert, B. Masquida, M. Andreassen, S. D. Johansen, H. Nielsen.

RNA submitted, accepted upon revision

Accumulation of stable full-length circular group I introns during heat-shock

Kasper L. Andersen[1], Bertrand Beckert[1,2], Benoit Masquida[2*], Morten Andreassen[3], Steinar D. Johansen[3], and Henrik Nielsen[1, 3*],

[1] Department of Cellular and Molecular Medicine, The Panum Institute, University of Copenhagen, Denmark

[2]Architecture et Réactivité de l'ARN, Université Louis Pasteur de Strasbourg, IBMC, CNRS, France

[3] Department of Medical Biology – RNA and Transcriptomics Group, University of Tromsø, Norway

*co-corresponding authors

Address:

Henrik Nielsen

Department of Cellular and Molecular Medicine, The Panum Institute,

University of Copenhagen

3 Blegdamsvej, DK-2200N

Telephone: +45 35 32 77 63

Fax: +45 35 32 77 32

E-mail: hamra@sund.ku.dk

Keywords: group I intron, *Didymium iridis*, circular RNA, horizontal gene transfer, molecular modelling, RNA catalysis.

Running title: Stability and expression levels of circular group I introns

Abstract

Group I introns found in the nuclear ribosomal RNA of eukaryotic microorganisms can be processed by splicing or circularization. The latter pathway results in formation of a full-length intron circle without ligation of the exons and has been proposed to be active in intron mobility. We have applied quantitative RT-PCR to estimate the copy number of full-length circular intron RNA from the myxomycete *Didymium iridis*. In exponentially growing amoebae, the circles are predominantly, if not exclusively, nuclear and found in approximately 70 copies/ cell. During heat-shock, the circular form is up-regulated to more than 500 copies/ cell. The intron harbors two ribozymes that both have the potential to linearize the circle. To understand the structural features that maintain circle integrity in this situation, we have performed chemical and enzymatic probing of the splicing ribozyme (DiGIR2) combined with molecular modelling to arrive at models of the inactive circular form and its active linear counterpart. Our results show that the two forms have the same overall structure but differ in key parts, including the catalytic core element P7 and the junctions at which reactions take place. These differences explain the relative stability of the circular species and demonstrate how it is prone to react with a target molecule for circle integration. Together with the observation that the copy number responds to external factors this supports the notion that the circular form is a biologically significant molecule and is consistent with a role in intron mobility.

Introduction

Circular RNA species are found dispersedly in biological systems. Compared to their linear counterparts, they have several features that could convey them new biological functions. First, circularity offers resistance towards exonucleases. Concordantly, an increased resistance of circular RNA to degradation has been observed in several cell types (Chan et al. 1988; Harland and Misher 1988) and the infectious units of highly mobile RNAs such as viroids and satellite RNAs have been found to be circular (Branch and Robertson 1984; Sanger et al. 1976). Second, the joining of the ends creates a unique sequence. This sequence could be used in preferential association of the circular form with protein factors or complementary RNA molecules. Third, end joining by a 2', 5' linkage could create a structural alteration that could serve as a recognition motif. End-joining by a 2', 5' linkage has been observed in group II introns (Murray et al. 2001) and viroids (Cote et al. 2001). In view of these distinct properties, surprisingly few examples of circular RNA molecules are known in biology and little is known about their biological functions.

Self-splicing group I introns (for a recent review, see Nielsen and Johansen 2009) are a particular rich source of circular RNAs. Group I introns catalyze their own splicing by two coupled transesterification reactions that result in ligated exons and a linear intron with a guanosine co-factor added to the 5'end. In addition to this main pathway, three types of circular RNAs have been found as a result of processing of group I introns (Fig. 1). The first type is truncated intron circles that results from circularization of the spliced out intron. Here, the intron terminal residue (ωG) makes an attack at an internal phosphodiester bond at a site near the 5'end and forms circles with concomitant release of a short RNA derived from the 5'end. The second type is full-length circles (FLC) that form as the main product of the circularization pathway that has been extensively characterized in nuclear group I introns that are inserted into the rRNA genes (Nielsen et al. 2003). This is an alternative pathway to splicing initiated by hydrolysis at the 3'SS followed by attack of ωG at the 5'SS. The products are FLC and un-ligated exons. Finally, circles that are one nucleotide longer than full-length were recently described in an *in vitro* study of splicing of the *Anabaena* $tRNA^{Leu}$ intron (Vicens and Cech 2009). These circles are formed by ωG attack at the triphosphate of the GTP that was coupled to the intron 5'end during the first step of splicing. Thus, pyrophosphate is released and the circles incorporate the guanosine co-factor.

The biological significance of group I intron circular RNA is uncertain. Truncated circles from many variants of the *Tetrahymena* rRNA intron have been characterized *in vitro* (Been and Cech 1985; Engberg et al. 1988; Grabowski et al. 1981) and their presence *in vivo* has been documented (Brehm and Cech 1983). Here, the turn-over of the circles was found to be very rapid, similar to that of the linear form of the spliced out intron. Truncated circular forms have also been found *in vitro* and *in vivo* in a number of other group I introns (Nielsen et al. 2003). In contrast to the truncated circles, the FLCs can be considering genomic in the sense that all intronic sequence information is present. They can re-open by hydrolysis or transesterification, and their presence in cellular RNA is well documented (Lundblad et al. 2004; Nielsen et al. 2003). Recent evidence from both *in vivo* (Birgisdottir and Johansen 2005) and *in vitro* (our unpublished data) studies suggests that the FLC can integrate into target RNA by an unknown mechanism. This would be an alternative to the well established mobility of group I introns at the DNA level by a homing mechanism (Stoddard 2005) and at the RNA level by reverse splicing (Birgisdottir and Johansen 2005; Roman and Woodson 1998). The third class of circular group I introns incorporate the guanosine co-factor and thus conserve the bonding energy from the first step of splicing. They may be involved in intron mobility by reverse splicing but their presence *in vivo* has not been documented.

In the present paper, we focus on the structure and expression of the full-length intron circles formed in the circularization pathway. This pathway has been observed in parallel with the splicing pathway *in vitro* and *in vivo* for numerous nuclear group I introns (Inoue et al. 1986; Lundblad et al. 2004; Nielsen et al. 2003). We have studied the Dir.S956-1 intron from the myxomycete *Didymium iridis* because this intron has a much more complex biology than most other introns. Dir.S956-1 is a twin-ribozyme intron (Decatur et al. 1995; Nielsen et al. 2008) composed of a conventional group I splicing ribozyme (DiGIR2) into which is inserted a branching ribozyme (DiGIR1) followed by a homing endonuclease gene (HEG). The splicing ribozyme is entirely responsible for splicing out the intron as well as all steps of the circularization pathway. *In vitro* experiments have demonstrated that processing of Dir.S956-1 is equally partitioned between the two pathways (Decatur et al. 1995; Nielsen et al. 2008). *In vivo* observations show the two pathways to be competing in growing amoebae and flagellates (Decatur et al. 1995; Nielsen et al. 2008). Although the partitioning between the two pathways has so far not been directly addressed in the *in vivo* situation, it is assumed that splicing is by far the dominant pathway.

We have used quantitative RT-PCR (qRT-PCR) to investigate the copy number of FLC in amoebae and flagellates. We estimate that circularization pathway products constitute 1 % of total Dir.S956-1 intron processing products in exponentially growing cells and find that the resulting FLC is predominantly, if not exclusively, nuclear. We show that the copy number of FLC as well as the proportion of primary transcripts undergoing circularization can be influenced by external factors, such as heat-shock. The relative stability of the FLC is surprising in view of its inclusion of two ribozymes that both would be expected to linearize the circle. Structure probing analysis and molecular modelling of the linear and the circular forms of the intron reveals that this is primarily due to relaxation of the active site in the circular form. We furthermore suggest that peripheral structures play a role in regulation of the two reaction pathways in response to external stimuli. Thus, group I ribozymes, like many other RNAs, could function as molecular sensors.

Results

The copy number of FLC is sensitive to external stimuli

We hypothesized that if the FLC is a biologically significant molecule, the copy number should respond to environmental conditions. First, we analyzed the FLC copy number during exponential growth and starvation-induced encystment (Fig. 2A) that is known to affect pre-rRNA processing (Vader et al. 2002). A typical growth curve of *Didymium iridis* in suspension culture is depicted in Fig. 2B. RNA was sampled from exponential growth (mostly amoebae), cells at the transition from exponential growth to stationary phase (mostly flagellates), and cysts. The RNA content of the cells declined during the experiment with 8.3 pg/ cell in the exponential phase, 5.5 pg/ cell in the transition phase and 0.6 pg/ cyst, corresponding to 1:0.66:0.07 ratios. The copy number of FLC also showed a decline during the course of growth (Fig. 2C). However, the decline was more dramatic from 70 copies/ cell in exponential phase to less than one copy per cyst corresponding to ratios 1:0.04:0.003. This demonstrates that less FLC is being produced or that FLC is specifically degraded during starvation-induced encystment. A fractionation study of exponential cells showed that most, if not all, FLC was located in the nucleus (Fig. S1).

Another set of conditions that are relevant to myxomycete biology is low and high temperatures although only few reports on this is found in the literature. *Didymium* diploid

plasmodium forms a macrocyst if exposed to temperatures of 7 to 10 °C for 18 h or 35 °C for 3.5 h (Raub T.J. and Aldrich H.C. 1982). A study of heat-shock protein induction in *Physarum* demonstrated 32 °C as the highest physiological temperature and 37 °C as the highest non-physiological temperature (Wright and Tollon 1982). Based on this we chose temperatures of 5 °C and 10 °C and will refer to this as cold-shock and 34 °C and 40 °C for heat-shock treatment. It should be emphasized that no molecular study was conducted to justify these terms, but it was evident by visual inspection that the *Didymium* cells were affected by the most extreme temperature regimes. Growth for 1 h at 5 °C resulted in 4% of the cells forming cysts, whereas growth at 10 °C had no apparent effect. During growth for 1 h at 34 °C the cells showed a clear tendency to aggregate, and after growth at 40 °C the aggregation was so pronounced that it was impossible to count the cells. The cellular RNA amount was only slightly affected by growth at low or high temperatures. However, the FLC copy number was significantly higher in cells grown at 34 °C or 40 °C compared to the FLC number in cells grown at the standard 25 °C. In Fig. 2D this up-regulation is expressed in relation to amount of input RNA as well as cell count. At 34 °C the up-regulation was 7.3 and 10.4-fold, respectively, and at 40 °C the increase in relation to amount of input RNA is 10.3-fold (at this temperature the cells aggregated and could not be counted). In contrast, cold-shock did not have a significant effect on the steady state level of FLC at either temperature (Fig. 2D) even though growth at 5 °C had an effect in inducing encystment. The ratios of FLC in cold-shocked cells compared to control cells were 1.3 and 1.2 relative to RNA amount and 1.3 and 1.6 relative to the cell count (for 5 °C and 10 °C cold-shock respectively). In conclusion, these experiments show that the copy number of FLC can be significantly up-regulated to more than 500 copies/ cell in response to an environmental factor, demonstrating that FLC formation is unlikely to simply reflect the level of ribosomal RNA synthesis and processing.

FLC is very similar in structure to L-IVS but has an unstable active site

Accumulation of FLC within cells is surprising in view of the reported rapid turn-over of truncated circles in *Tetrahymena* (Brehm and Cech 1983) and the fact that FLC maintain ribozyme activity and is prone to undergo conversion to a linear form by hydrolytic cleavage at the circularization junction. To understand the structural reasons for accumulation of FLC we conducted a structure probing analysis of the splicing ribozyme component (DiGIR2) of the twin-ribozyme intron using a similar analysis of the linear form (L-IVS) as a reference. A DiGIR2 construct in which DiGIR1 and the HEG were removed from the P2 segment of the full-length

intron was used for practical reasons. The resulting P2 stem contains no foreign sequence insertions and has a length comparable to that of other group IE introns (Fig. 3A). The DiGIR2 construct carries out the same reactions and accumulates splicing and circularization products *in vitro* in the same proportions as the full-length intron (Decatur et al. 1995; Johansen and Vogt 1994; Nielsen et al. 2003). Since two different structural variants of the same intron (L-IVS and FLC, respectively) are being compared, it was critical to ensure that the correct folds of the molecules were investigated. Rather than relying on renaturation protocols that would be difficult to verify, structure probing was carried out in the reaction mixture and the relevant molecular species subsequently isolated for analysis.

A secondary structure diagram of the linear DiGIR2 based on the general rules for outlining group I introns and displaying the structure probing data for the L-IVS is depicted in Fig. S3. The structure probing data for the FLC is found in Fig. 3A and a comparison that highlights the differences between FLC and L-IVS is shown in Fig. 3B (examples of primary data are shown in Fig. S2). The probing patterns are mostly similar with a few notable differences. First, several differences were noted in and around the binding pocket for the guanosine co-factor. When ωG is bound into the G-binding pocket located in the narrow groove of P7, it is sandwiched between the last residue of J6/7 and the first residue of P7. Additionally, the other side of the last residue of J6/7 stacks with the 3'residue of J8/7. In the FLC, all these residues are accessible to Watson-Crick probes, meaning that the formation of circles is accompanied by the destructuring of the central column of residues interacting in the narrow groove of P7. The observation of kethoxal modification at G233 and G235 in the FLC demonstrate the overall destabilization of the G-binding pocket as compared to the L-IVS. In further support of this, the V1 cleavages at C182 and G233 in L-IVS were not observed in FLC. Second, the FLC differs from the L-IVS in that the exoG is missing and the 5'- and 3'- nucleotides of the intron are covalently linked. The nucleotides on the 5'- side of the junction are accessible to chemical modification at their Watson-Crick edges and thus appear to be solvent exposed. In contrast, nucleotides 3' to the junction appear inaccessible. The modification pattern of the circle junction region of the FLC is slightly different from that found within the 5' end of the L-IVS indicating structural differences. Finally, increased accessibility of probes in the upper part of the P4-P6 domain in FLC, in particular at J4/5, J5/5a, and J6/7 suggests a less compact packing of the principal domains.

It appears that FLC formation is accompanied by the relaxation of the core of the ribozyme following the second step of the circularization pathway. Once the 5'exon has been released from DiGIR2, the secondary structure of P1 remains solely based on two A-U pairs that are expected to melt readily and trigger the relaxation process. The unfolding of P1 is expected to destabilize the binding of ωG in its pocket and consequently the 4-nt stack interacting in the narrow groove of P7. Thus, P7 becomes more accessible to the solvent, which implies a greater sensitivity to Watson-Crick probes as compared to L-IVS. The unstructured loop encompassing the P1 residues tethered to ωG cannot dock onto J4/5 and provides the driving force to destabilize the catalytic core by conferring higher dynamics to the nucleotides occupying the narrow groove of P7. Thus, we suggest an induced disorder rather than a different position for nucleotides becoming sensitive to Watson-Crick probes specifically in the FLC. As a consequence, these nucleotides have higher dynamics that would confer them higher positional fluctuations around their average position which is in good agreement with FLC-specific but faint reactivity and with the observation that reverse splicing/integration can be stimulated by exon mimicking oligonucleotides.

Pb^{2+} cleavage sites in the catalytic core

The structure probing data indicated that the most significant difference between L-IVS and FLC is found in P7. To further investigate these differences we performed a Pb^{2+} probing experiment. Pb^{2+} probing gives information at two levels. Pb^{2+} can displace Mg^{2+} bound at specific binding sites such as the catalytic Mg^{2+} ions bound to the catalytic core in group I introns and cleave the RNA at a nearby phosphodiester bond (Streicher et al. 1993; Streicher et al. 1996). In addition Pb^{2+} will induce cleavage of the RNA at positions of unconstrained nucleotides. The former type of cleavage site is sensitive to high Mg^{2+} concentrations whereas the latter is not. The results from the Pb^{2+} probing of DiGIR2 L-IVS and FLC core regions are presented in Fig. 4. Analysis of the L-IVS shows a prominent signal at G229 (J8/7) and a less prominent signal at U234 (P7). Both of these signals can be competed out by increasing the Mg^{2+} concentration prior to the addition of Pb^{2+}. In contrast, a signal at U225 (J8/7) and several minor signals outside the core e.g. in L8 are not affected by increases in Mg^{2+} concentration. This result strongly indicates that Pb^{2+} displaces one or more of the Mg^{2+} ions specifically bound in the catalytic core. The Pb^{2+} probing of FLC is similar with respect to the two signals that could be competed with Mg^{2+} ions but differs in the absence of the signal at U225 (J8/7). This result corroborates the previous finding of differences in the core organization of L-IVS and FLC.

3D modelling supports the involvement of the P9 appendages in circle formation

In order to better understand the different organization of the core in FLC, we built a 3D model by homology modelling based on available X-ray crystallographic structures of other group I introns. The details of modelling of GIR2 L-IVS which represents the first whole atom model of a group IE intron are described in the Supplementary Online Materials that also include presentations of the model of the L-IVS and FLC (Fig. S5). Of particular interest was the modelling of the P9 appendages that are one of the characteristics of this subgroup. The P9 domain consists of a four-way junction (4WJ) composed of P9a, P9b, P9.1 and P9.2 that extends the 2-bp P9.0 element stacked onto P7. This allows the ribozyme 3' residue (ωG) to be accommodated in the G-binding pocket. In order to make the tetraloop L9b interact with its receptor in P5, P9a had to be stacked with P9b in the model. Furthermore, P9.2 was stacked with P9.1 to allow the latter to lie along P7/P3 and interact with L2.1 to finally form the P13 pseudoknot. The conformation of the elements encompassing P2/P2.1 and the P9 insertion is supported by the observation that V1 cleavages are only observed in the P13 strand belonging to P9.1 which is indeed exposed to the solvent in the model whereas the opposite strand is buried. The conformation of the 4WJ has two major architectural consequences. First, following the formation of P13, the internal loop of P9.1 is in contact with the narrow groove of P7 and thus wraps the catalytic core. Strikingly, a very similar interaction has already been observed, although originating from a P7 insertion instead of a P9 insertion. In the group I intron of the *td* gene from phage T4 (Waldsich et al. 2002) and in the *Twort* intron crystal structure (Golden et al. 2005), a loop E motif located in the P7.2 extension stabilizes P7 in a similar way. Three nucleotides in the purine-rich internal loop P9.1 of DiGIR2 are fully conserved in a sequence alignment of group I introns of subgroup IE (Fig. 5A) and are thus very likely to directly take part in this interaction. However, the precise arrangement of base-pairs taking place in this loop is difficult to assess due to high variability in the loop size and symmetry. Sequence analysis and structure probing analysis shows that the P9.1 internal loop does not obey the covariation rules and modification patterns observed for the loop E motif (Leontis and Westhof 1998; Waldsich et al. 2002).

The second consequence of the conformation of the 4WJ is that the predicted K-turn motif bends the P9.2 element by about 120°, which orients L9.2 towards the P9.0/P9a region and allows them to interact (Fig. 5B and S5B). The sequence of the K-turn in P9.2 strictly fits to the

phylogenetic rules specific to this motif (Lescoute and Westhof 2006) and the protection pattern corresponds well to the proposed structure with the strongest signals in the two flanking nucleotides A295 and U297 and weaker modification of the central A296. In addition, modification signals from two of the guanosines in the non-canonical region of P9.2 were observed in agreement with the formation of the characteristic *trans* G-A pairs involving their sugar and Hoogsteen edges, respectively (Fig. S2). Interestingly, RNases A and T2 did not cleave the K-turn motif. Thus, it seems that RNases cannot recognize the motif as single stranded. The RNase V1 cleavage pattern shows multiple cleavages on both sides of P9.2 (Fig. 3A) indicating exposure to the solvent. In the model, L9.2 swings in front of J9.0/9a, which becomes buried and thus inaccessible to the solvent as indicated by theoretical accessibility calculations (data not shown). Such a conformation provides a satisfactory rationale for protections from enzymatic and chemical probes of this FLC region (Fig. 3A). Moreover, mutational studies point to L9.2 (Haugen et al. 2004) and also J9.0/9a as major actors in the ability of DiGIR2 to naturally carry out the formation of FLC, suggesting that L9.2 may interact with the region encompassing J9.0/9a, although the molecular details of this interaction have not been addressed.

Pb^{2+} cleavage and in-line probing data show different patterns in the P9.2 region of the L-IVS as compared to the FLC (Table S2). The 3' end of L9.2 is unreactive but the 5' end is strongly cleaved in in-line experiments at position G311 in both RNA species. Moreover, the A307 to A309 nucleotides are significantly less reactive in the FLC than in the L-IVS where this trend propagates up to the K-turn motif. Furthermore, the K-turn motif shows no reactivity at all in the FLC while Pb^{2+} induced cleavages are observed on the 5' side of the K-turn bulge. These observations are in favour of a greater stability of the P9.2 element in the context of the FLC rather than in the L-IVS, which supports the role of the entire P9.2 appendage as a key element in the mechanism of FLC formation.

Discussion

The FLC copy number varies in response to internal and external factors

We have used circle specific qRT-PCR to provide the first quantitative data on the FLC that is the product of the circularization pathway in nuclear group I introns. In exponentially

growing *Didymium* cells we observed an average copy number of 70 FLC/ cell. This is similar to medium abundant mRNA's in eukaryotes. The circularization pathway exists in parallel with the splicing pathway in growing amoebae and flagellates. A rough estimate based on qRT-PCR specific for the products of each of the two pathways suggests that 1 % of the pre-rRNA precursor is allocated to this pathway (Table S1). As the two pathways are equally partitioned during processing *in vitro*, we suggest that the circularization pathway is specifically suppressed *in vivo*. A loss of 1 % of pre-rRNA precursor could be a marginal cost to the cell. On the other hand, the added cost of having 20 group I introns in the precursor as in *Diderma*, most of which produce FLC, could be substantial (Nielsen and Johansen 2009). The mechanism of suppression is not known, but one possibility is an RNA based switching mechanism that is supported by evidence from structure probing and mutagenesis studies (Haugen et al. 2004). Alternatively, individual rDNA copies could be dedicated to one or the other pathway based on the sub-nuclear localization of the extrachromosomal rDNA copies.

It has been speculated that the FLC is used in translation of open reading frames located within the intron. Indeed, translation of open reading frames within circular introns is known from Archaea (Lykke-Andersen et al. 1997). One argument in favor of circular intron-encoded mRNA's is that circularization would offer an alternative stabilization to the m^7G-cap that is not added because of the origin of the mRNA. Our demonstration of the predominantly nuclear localization of the FLC in *Didymium* (Fig. S1) rules out this possibility and is consistent with our previous finding of a linear and processed form of the I-*Dir*I mRNA on polysomes (Vader et al. 1999). The observed nuclear localization is not surprising in view of the fact that FLC contain at least one *bona fide* nuclear retention signal in the form of a spliceosomal intron.

If the FLC were a biologically significant molecule, the abundance would be expected to vary in relation to cell internal (cell cycle or development) or external (e. g. cellular stress) factors. Encystment and excystment are very frequent transformations that *Didymium* cells undergo repeatedly in their natural environment. We find that the FLC copy number decreases during starvation-induced encystment to below a single copy per cell. We have previously shown that starvation-induced encystment results in accumulation of a 7.5 kb RNA species resulting from cleavage at an intron internal processing site (Vader et al. 2002) and speculate that this RNA functions as a pre-mRNA for the intron encoded homing endonuclease during excystment. Our

finding of a dramatic reduction in FLC during encystment is in accordance with the accumulation of the 7.5 kb RNA because this processing pathway is incompatible with FLC formation. Next, we examined the FLC copy number during growth at reduced or elevated temperatures as an example of cellular stress induced by external factors. Low temperatures resulted in cyst formation in a small fraction of the culture. In accordance with this, only minor effects on the FLC copy number were found. In contrast, growth at elevated temperatures resulted in a strong phenotype with pronounced aggregation of cells at both 34 ^{o}C and 40 ^{o}C. In these cultures, the FLC copy number was increased 7-10 fold to more than 500 copies/ cell. This is within the range of infectious, circular viroid molecules in infected plant cells (Schumacher et al. 1983). Thus, the fact that FLC is up-regulated specifically in cells at stress conditions and accumulate to such high numbers is reminiscent of other independent and mobile elements in biology.

Several mechanisms resulting in FLC accumulation can be envisaged. One of the simplest is a shift in allocation of pre-rRNA from the splicing to the circularization pathway resulting from imbalance of pre-rRNA and ribosomal proteins during cellular stress. Studies from plants and *Drosophila* have shown that pre-rRNA synthesis continues at a slightly reduced level after heat-shock, but that the level of ribosomal proteins and rRNA processing is dramatically reduced (Bell et al. 1988; Nover et al. 1986). If such ribosomal factors were mediating the suppression of the circularization pathway in *Didymium*, heat-shock would relieve the suppression and make the reactions of the intron more similar to what is observed *in vitro*.

Structural differences between L-IVS and FLC

The L-IVS and the FLC display similar overall probing patterns implying that the structural alterations following circularization are few. Only two nucleotides were found to be specifically modified by chemicals in the L-IVS compared to 11 positions specific to the FLC. Interestingly, the majority of the differences were clustered in the active site helix P7 and immediate joining segments. In the three crystallized introns from the subgroups IA2, IC1 and IC3 (Adams et al. 2004b; Golden et al. 2005; Guo et al. 2004) G-binding sites in P7 have been found to be almost identical and to involve nucleotides from both J6/7 and J8/7 brought together in the major groove of P7. Hence, the universally conserved G-C pair of P7 forms a base triple with ωG sandwiched in between three other residues in the narrow groove of P7 and participating in base triples. Comparing the probing results between the FLC and the L-IVS (Fig. 3B) it was noted that

nucleotides in P7 (A179, G233 and G235) are modified in the FLC but not in L-IVS. Likewise, neighbouring nucleotides A178 (J6/7) and A231 (J8/7) were modifiable by DMS in the FLC but not in the L-IVS. Notably, A178, A179 and A231 are involved in separate base triple interactions in the sandwich forming the G-binding site (Adams et al. 2004b; Guo et al. 2004). In fact the only base triple found not to be modifiable in FLC is the universally conserved C-G pair in P7 contacting the ωG (Michel et al. 1990). Furthermore, P7 was cleaved by the double-strand specific RNase V1 on both strands in the L-IVS but not in the FLC. Together these results indicate that at least a proportion of the FLC molecules have a relaxed active site with a disassembled G-binding pocket and thus most likely are found in a catalytically inactive conformation following FLC formation. This notion is consistent with the impaired ability of the FLC towards catalysing re-opening at the circularization junction. Interestingly, a relaxed active site was also reported in an X-ray crystallographic study of the *Azoarcus* group I intron following the second step of splicing (Lipchock and Strobel 2008). Here, the metal coordination by the exons was lost even though the exons remained bound to the intron through other contacts. It may be a property of many structured RNA molecules that critical parts of the structure are rendered in a relaxed conformation in the absence of interaction partners or small molecule ligands. Disordered states locally affecting specific RNA regions have been characterized by fluorescence spectroscopy studies of 2-amino purine derivatives of two riboswitches (Lang et al. 2007; Rieder et al. 2007). These studies show a shift from a loose to a tight conformation of loops upon ligand binding.

Two high affinity Pb^{2+} cleavage sites were detected in the catalytic core of DiGIR2 in both the FLC and the linear intron. A high affinity Pb^{2+} cleavage site has previously been detected at the second last nucleotide in the J8/7 of the subgroup IC1 *Tetrahymena*, and the subgroup IA2 *sunY* and the *td* introns (Streicher et al. 1993). This Pb^{2+} cleavage site thus reflects a conserved metal ion binding pocket in the intron core of the DiGIR2 intron that can be expected to be present in all group I intron subgroups. In the crystal structure of the *Azoarcus* intron two Mg^{2+} (M1 and M2) were inferred to be coordinating the catalytic centre (Stahley and Strobel 2005). Five nucleotides in the intron were observed to directly interact with the two ions in the catalytic core and one of these corresponds to the position of the high affinity cleavage site G230 in J8/7 of DiGIR2. This indicates that it is the M1 ion in the catalytic core that is displaced by Pb^{2+} and detected. Additional Pb^{2+} dependent cleavages have been found in P7 and J8/7 of other group I

introns (Streicher et al. 1993) suggesting minor group-specific differences in the structure of the active site.

The Pb^{2+} probing pattern also revealed a clear difference between the linear intron and the FLC. A strong cleavage signal at the 5'end of J8/7 was observed in L-IVS but not in the FLC. This cleavage could not be out-competed by addition of Mg^{2+} ions, and probably reflects a structure of the L-IVS backbone prone to be cleaved by Pb^{2+} ions as is the case in the D-loop of the $tRNA^{phe}$ (Brown et al. 1985). Indeed, the backbone adopts a *gauche⁻/trans* conformation around the phosphate group between residues U225-A226 in the L-IVS very similar to the conformation observed between residues D17 and G18 in the $tRNA^{phe}$. It is worth to note that this specific conformation is due to the recognition of P1 by the 5'end of J8/7, a situation not found in the FLC due to the destructuring of P1 following the circularization. The observed differences in Pb^{2+} probing between L-IVS and FLC thus similarly reflect differences in the engagement of J8/7 with the remaining P1 sequences. This is consistent with the observation of differences in chemical modifications of this part of the molecule.

The L-IVS and FLC differs in the P1-P2 substrate domain in that L-IVS has the guanosine co-factor attached to its 5'end whereas the FLC has the 5′ and 3'ends of the intron covalently linked. The most striking observation is that the 5'end of the intron is readily accessible to chemical modification in both cases indicating that the Watson-Crick face of the internal guide sequence and the immediate upstream sequence is solvent exposed. This is in accordance with the suggestion that the internal guide sequence in both molecular species promotes reverse splicing/ integration by base-pairing with the target. However, several differences in the probing pattern are noted that could reflect differences in interaction with the catalytic site and ability to recognize the target for reverse splicing/ integration. Differences in the V1 cleavage patterns indicate that the P2-P2.1 part of the substrate domain is oriented differently in L-IVS and FLC, probably as a consequence of the differences in the organization of the intron 5'end.

The structure of the P9 domain suggests a role in promotion of circularization

The P9.2 element in DiGIR2 is particularly interesting when considering the circularization pathway. A time course splicing analysis has shown how a DiGIR2-ΔP9.2 mutant had significantly lower 3'SS hydrolysis and thus FLC formation, while the splicing reaction was

unaffected. The major effect was pinpointed to L9.2 and particularly to the 5'-CAAA sequence on the 3'-side (Haugen et al. 2004). P9.2 has previously been modelled as the outermost shell secondary structure element with respect to the core of the intron in both the *Tetrahymena* (IC1) (Lehnert et al. 1996) and group IE introns (Li and Zhang 2005). In this way most, if not all, nucleotides of L9.2 would be accessible to modification. The present study shows that the 3' side of the loop consisting of the sequence that was shown to be important for hydrolysis is protected from modification. Correlated with this, we have verified a previous suggestion that the internal asymmetric loop in P9.2 forms a K-turn motif (Haugen et al. 2004). The present model shows that optimal stacking at the interface of the four helices forming the 4WJ between P9a, P9b, P9.1 and the root of P9.2 orients the K-turn so that the tip of P9.2 swings over P9b opening the possibility for the L9.2 loop to interact with the A-rich strand connecting P9a to the ωG carrier P9.0. The potential interaction may block P9.0 on top of P7 to force the docking of ωG into the G-binding pocket and thus promote circularization. This process may be helped by a second interaction between the internal loop of P9.1 and P7 that leads to clamping the P7 G-binding pocket. In this scenario, the L9b/P5 interaction that leads to the P1 substrate docking along J4/5 and ultimately to splicing is not needed. Thus, the P9 domain seems to orchestrate the interplay between splicing and circularization by favouring the first step of the latter over the first step of the former. In summary, the P9 domain provides an alternative stabilization mechanism for P7 that re-routes the ribozyme activity towards circularization. Furthermore, since the K-turn can be involved in contacts with protein factors or in the formation of tertiary RNA-RNA interactions (Klein et al. 2001), it is possible that P9.2 could be a target for molecules produced in the cell according to specific environmental conditions.

A case for the FLC in intron mobility

The Dir.S956-1 intron analyzed in this study is an example of an intron that is mobile within the species (by homing and possibly reverse splicing) and between species (as revealed by phylogenetic analysis). The FLC has been proposed to be involved in intron mobility by an RNA-based complementary mechanism (Johansen and Vogt 1994). Preliminary reports have indeed shown that reverse integration of FLC molecules at the cognate site in ribosomal RNA is feasible (Birgisdottir and Johansen 2005); HN; unpublished). The FLC has several features that would promote such a role. First, it contains all the sequence information of the intron in contrast to the truncated intron circles that are formed by circularization of the spliced out intron. Second, the ends are protected from degradation by exonucleases due to the covalent joining. Third, the FLC retains

the potential for catalytic activity as evidenced primarily by its ability to reverse integrate into a oligonucleotide mimicking ligated exons. Fourth, the structure analysis shows that the structure of the P4-P6 and P3-P9 domains of the DiGIR2 structure is conserved upon circularization. However, the P1-P2 domain is structurally perturbed with a solvent exposed internal guide sequence and a G-binding site that appears to be disassembled. This structure ties in with the model of an FLC that is inactivated in hydrolysis but can be re-activated for reverse integration by a slight conformational change perhaps induced by binding to its target RNA. Finally, the copy number of FLC is up-regulated during cellular stress similar to what is seen with many other infectious elements. Thus we conclude that the FLC has specific features that are consistent with its proposed role in mobility.

Materials and Methods

Strains and sequences

Didymium iridis strain Lat3-5, derived from Pan2-44 isolate can be obtained from Prof. Steinar D. Johansen, University of Tromsø. The sequence of the twin-ribozyme group I intron can be found in GenBank (Acc. no. AJ938153).

Growth conditions

Amoebae were grown in DS/2 medium (1.0 g D-glucose, 0.5 g yeast extract, 0.1 g $MgSO_4$, 1.0 g KH_2PO_4, 1.5 g K_2HPO_4, dH_2O to 1 liter) at 25 ^oC containing *E. coli* KB as a food source (Johansen et al. 1997). The cells entered starvation-induced encystment by depletion of the food source. Cells were counted in a Coulter Multisizer (Coulter Electronics) or a Neubauer cytometer. Cysts were scored by visual inspection after lysis of amoebae and flagellates in 0.5% Nonidet P-40. For heat-shock or cold-shock, exponentially growing amoeba cells at 25 ^oC were harvested by centrifugation and resuspended in DS/2 medium with *E. coli* KB at 10 ^oC, 25 ^oC, or 34 ^oC, respectively. Aliquots of the cultures incubated at 10 ^oC or 34 ^oC were harvested and re-incubated at 5 ^oC or 40 ^oC, respectively. Temperature shifts were achieved in 10 min and the cells grown for an additional 1 h before RNA isolation.

RNA isolation from cells

2-10 ml of 10^6-10^7 cells/ml was harvested by centrifugation at 450 *g* for 5 min. The pellet was dissolved in 2.5 ml ice-cold RNAzol/10^7 cells. Following vigorous shaking, 0,2 volume of chloroform was added and the sample left on ice for 20 min. After centrifugation at 1800 *g* for 20 min at 5 ^oC, the upper phase was isolated, precipitated by addition of 1 volume of isopropanol, and

centrifugation at 16.500 *g* for 40 min. The RNA pellet was resuspended and extracted twice with Phenol-Chloroform-Isoamyl Alcohol (PCI, pH 6.6) (BDH), once with chloroform, and precipitated with 3 volumes of 96% ethanol. The RNA concentration was determined with RiboGreen Assay (Molecular Probes) and measured with a Flouroscan Ascent FL using the appropriate software (Thermo Electron Corporation). RNA was extracted from each growth condition or cell fraction in 2-5 independent experiments.

Cell fractionation

Cells were harvested, resuspended in 250 µl ice-cold lysis buffer (20 mM Tris-HCl (pH 8.0), 1.5 mM $MgCl_2$, 140 mM KCl, 1.5 mM DTT, 1 mM $CaCl_2$, 0.1 mM EDTA, 0.16 mM cycloheximide, 0.5% Nonidet P-40, 500 U/ml RNasin (Promega; 40 U/µl) and incubated on ice for 5 min. Nuclei were pelleted from the lysate by centrifugation at 10.000 *g* for 10 min at 5 ^{o}C. The supernatant containing cytosolic RNA was extracted twice with PCI and once with chloroform and ethanol precipitated. The pellet containing nuclei was dissolved in RNAzol and treated as for WC RNA isolation described above.

Reverse transcription and quantitative RT-PCR (qRT-PCR)

50 ng of RNA in 20 µl of RT-buffer (Fermentas) containing N_6 random primers (0.2 µg/µl; Pharmacia Biotech) was incubated for 1 min at 80 ^{o}C (heat denaturation) followed by 5 min at 25 ^{o}C (primer annealing). Then, 1 µl of M-MuLV H^- reverse transcriptase (200 U/µl; Fermentas) was added and incubation continued for 1 h at 42 ^{o}C (cDNA synthesis). For each experiment, a reverse transcriptase minus sample and a 10-fold dilution series of the RNA standard was done in parallel. To each of the samples in the standard dilution series, 50 ng of yeast bulk RNA (Ambion) was added to give same total RNA concentration as the experimental sample. The primers for qRT-PCR were designed by use of Primer3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). For FLC specific amplification, two primers that would result in a 127 bp PCR-product spanning the circularization junction were selected (C394: 5'–TGA TCT TGG GAC ACG CTA CA and C395: 5'–TAC TTC CCA ACC CAA CCA AA). This product was sequenced in several experiments showing only the expected circularization site with no evidence of truncated circles.

A template for making the RNA standard used to calibrate the FLC quantifications was constructed from plasmid pDiSSU1 (Johansen and Vogt 1994) containing the Dir.S956-1 intron and flanking exons. 1 µg of linearized plasmid was in vitro transcribed and the transcript processed to yield FLC by incubation in reaction buffer (0,5 M KCl, 25 mM $MgCl_2$, 40 mM Tris-HCl (pH 7.5), 2 mM spermidine, 5 mM DTT) for 45 min. at 45 ^oC. Then, RT-PCR was applied using a 5' primer with a T7 promoter sequence (C396: 5'-AAT TTA ATA CGA CTC ACT ATA GGT GTC TGA AAG TAA GGT CTC AAC) and a 3' primer (C79: 5'-GCC GTT AGG TCG GAT GTT) to give a PCR-product of 283 bp that could be transcribed into an RNA representing a part of the FLC including the FLC junction and the primer sites used in qRT-PCR. The RNA standard was treated with Turbo DNase (Ambion), PCI extracted and precipitated. The concentration was determined using a fluorometric assay based on staining with RiboGreen (Molecular Probes) and a 10-fold dilution series ranging from 25 to 2.5 x 10^{10} molecules/µl was prepared.

qRT-PCR was carried out using the LightCycler Instrument (Roche) and the accompanying software (v.3.5) was used to follow the reaction. The second derivative maximum method was applied to determine threshold cycle and ultimately calculate the FLC copy number. All reagents were from the LightCycler FastStart DNA Master^{plus} SYBR green I kit (Roche). The qRT-PCR parameters were 95 ^oC, 10 min (1 cycle); 95 ^oC, 10s; 60 ^oC, 5s; 72 ^oC, 15s (40 cycles) followed by a melting curve course 95 ^oC, 15s, 65 ^oC, 15s, 65 ^oC to 95 ^oC (continuous measurement). Each RT-PCR reaction on the various RNA extractions was performed in duplicate.

***In vitro* transcription and RNA processing**

In vitro transcription was performed by T7 RNA polymerase (Fermentas) on linearized DiGIR2 plasmid (Decatur et al. 1995) at 5°C over night to avoid processing of the precursor molecules during transcription. The RNA was uniformly labelled using [α-^{32}P]UTP (10 mCi/ml; Amersham Pharmacia Biotech) during transcription. Processing of the RNA (splicing and circularization) was carried out in a buffer that supported structure probing.

Chemical and enzymatic structure probing, RNA purification and primer extension

Chemical and enzymatic probing was performed according to (Christiansen J. et al. 1990). The chemical probes react preferentially with single-stranded nucleotides (abbreviations and the reactions analysed in parenthesis): dimethyl sulphate (DMS; A>C), kethoxal (KE; G), diethyl

pyrocarbonate (DEP; A), 1-cyclohexyl-3-[2-morpholinoethyl]-carbiodiimide (CMCT; U>>G). The enzymatic probes were (supplier and preference of cleavage in parenthesis): RNase T1 (Sigma; single-stranded G), RNase T2 (Sigma; single-stranded A), RNase A (Ambion; single-stranded U and C), RNase V1 (Ambion; double-stranded RNA without sequence preference). Pb^{2+} probing was carried out according to (Brunel and Romby 2000) and in-line probing according to (Winkler et al. 2003). Chemical probing was carried out in the reaction buffer (70 mM Hepes-KOH (pH 7.8), 10 mM $MgCl_2$, 270 mM KCl, 1 mM DTT, 25 µM GTP) and enzymatic probing in the same buffer without DTT. In-line probing was in 50 mM Tris-HCl (pH 8,3), 20 mM $MgCl_2$, 100 mM KCl, 0-0,2 mM GTP. Pb^{2+}-probing was in 25 mM Hepes (pH 7,5), 5 mM $MgAc_2$, 100 mM KAc, 0-0,2 mM GTP. All of the cleavage probes (enzymatic, in-line and Pb^{2+}) were titrated to single-hit conditions in a linearization assay. In this assay, the linearization of a gel-purified circular RNA to a linear species was assessed by gel electrophoresis.

The RNA was probed directly in the reaction mixture. For the non-cleavage probes, the L-IVS and the FLC were subsequently purified on 5% denaturing (urea) polyacrylamide gels, excised and eluted in 1 mM EDTA, 0,25 mM NaAc (pH 6) followed by analysis by primer extension. Probing that involves cleavage of the RNA rules out the above procedure because the fragments originated from L-IVS and FLC cannot be discerned. In this case, most of the FLC could be analysed by primer extension using a circularization junction specific oligonucleotide. Alternatively, the FLC molecules were purified using a Peptide Nucleic Acid (PNA) probe specific for the junction (KLA; unpublished). In short, the biotinylated probe (PNA 1987: (Biotin)-AGC AAT TAC CTT TAT A-Lys-NH_2) was annealed to the FLC in the processed and probed mixture of RNA species and the complex subsequently purified on streptavidin-coated magnetic beads. In the analysis of L-IVS after probing with cleavage probes, signals derived from unprocessed precursor molecules could not be discerned from those originating from L-IVS. The results presented are the consensus from 2-5 independent experiments.

The primers used for primer extension analysis were: C147: 5'–TTGATCGTTGGCCTCA), C339: 5'–ACCTTAGCGATTCTAA), C340: 5'–CCTTTATACCAGCCT) and C151: 5'–CGGCCTAGCAATTACCTTTATA (junction specific).

Molecular modelling

The DiGIR2 model has been built by homology modeling using the program ASSEMBLE, an extended version of MANIP (Massire et al. 1998) linked to the S2S application (Jossinet and Westhof 2005) (URL: http://paradise-ibmc.u-strasbg.fr). First, eight sequences of representatives of the group IE intron subgroup were aligned using the *Azoarcus* ribozyme as the reference three-dimensional structure (Adams et al. 2004a; Adams et al. 2004b)). Gaps were inserted accordingly in the *Azoarcus* sequence to account for insertions specific to the group IE ribozymes. Following this, ASSEMBLE was used to automatically generate the 3D coordinates of nucleotides from DiGIR2 aligned pair wise to nucleotides from the *Azoarcus* ribozyme. Regions corresponding to insertions (P2, P2.1, P9.1, and P9.2) were built using the interactive functionalities from MANIP implemented in the program Assemble.

Acknowledgements

We would like to thank Dr. Ying C. Lee for assistance and discussions regarding qRT-PCR, Prof. Peter E. Nielsen for supplying PNA, Mr. Franz Frenzel for technical assistance, and Mr. Mads M. Hedegaard for fruitful discussions. The work was supported by The Danish Natural Science Research Council. BB is supported by the Lundbeck Foundation.

References

Adams PL, Stahley MR, Gill ML, Kosek AB, Wang J, Strobel SA. 2004a. Crystal structure of a group I intron splicing intermediate. *RNA* **10:** 1867-1887

Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA. 2004b. Crystal structure of a self-splicing group I intron with both exons. *Nature* **430:** 45-50

Been MD, Cech TR. 1985. Sites of circularization of the Tetrahymena rRNA IVS are determined by sequence and influenced by position and secondary structure. *Nucleic Acids Res* **13:** 8389-8408

Bell J, Neilson L, Pellegrini M. 1988. Effect of heat shock on ribosome synthesis in Drosophila melanogaster. *Mol Cell Biol* **8:** 91-95

Birgisdottir AB, Johansen S. 2005. Site-specific reverse splicing of a HEG-containing group I intron in ribosomal RNA. *Nucleic Acids Res* **33:** 2042-2051

Branch AD, Robertson HD. 1984. A replication cycle for viroids and other small infectious RNA's. *Science* **223:** 450-455

Brehm SL, Cech TR. 1983. Fate of an intervening sequence ribonucleic acid: excision and cyclization of the Tetrahymena ribosomal ribonucleic acid intervening sequence in vivo. *Biochemistry* **22:** 2390-2397

Brown RS, Dewan JC, Klug A. 1985. Crystallographic and biochemical investigation of the lead(II)-catalyzed hydrolysis of yeast phenylalanine tRNA. *Biochemistry* **24:** 4785-4801

Brunel C, Romby P. 2000. Probing RNA structure and RNA-ligand complexes with chemical probes. *Methods Enzymol* **318:** 3-21

Chan WK, Belfort G, Belfort M. 1988. Stability of group I intron RNA in Escherichia coli and its potential application in a novel expression vector. *Gene* **73:** 295-304

Christiansen J., Egebjerg J., Larsen N., Garrett R.A. 1990. Analysis of rRNA structure: Experimental and theoretical considerations. 229-252

Cote F, Levesque D, Perreault JP. 2001. Natural 2',5'-phosphodiester bonds found at the ligation sites of peach latent mosaic viroid. *J Virol* **75:** 19-25

Decatur WA, Einvik C, Johansen S, Vogt VM. 1995. Two group I ribozymes with different functions in a nuclear rDNA intron. *EMBO J* **14:** 4558-4568

Engberg J, Zaug AJ, Nielsen H. 1988. Circularization site choice in the self-splicing reaction of the ribosomal RNA intervening sequence of Tetrahymena silvana. *Molecular Genetics, Life cience Advances* **1:** 50-55

Golden BL, Kim H, Chase E. 2005. Crystal structure of a phage Twort group I ribozyme-product complex. *Nat Struct Mol Biol* **12:** 82-89

Grabowski PJ, Zaug AJ, Cech TR. 1981. The intervening sequence of the ribosomal RNA precursor is converted to a circular RNA in isolated nuclei of Tetrahymena. *Cell* **23:** 467-476

Guo F, Gooding AR, Cech TR. 2004. Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site. *Mol Cell* **16:** 351-362

Harland R, Misher L. 1988. Stability of RNA in developing Xenopus embryos and identification of a destabilizing sequence in TFIIIA messenger RNA. *Development* **102:** 837-852

Haugen P, Andreassen M, Birgisdottir AB, Johansen S. 2004. Hydrolytic cleavage by a group I intron ribozyme is dependent on RNA structures not important for splicing. *Eur J Biochem* **271:** 1015-1024

Inoue T, Sullivan FX, Cech TR. 1986. New reactions of the ribosomal RNA precursor of Tetrahymena and the mechanism of self-splicing. *J Mol Biol* **189:** 143-165

Johansen S, Elde M, Vader A, Haugen P, Haugli K, Haugli F. 1997. In vivo mobility of a group I twintron in nuclear ribosomal DNA of the myxomycete Didymium iridis. *Mol Microbiol* **24:** 737-745

Johansen S, Vogt VM. 1994. An intron in the nuclear ribosomal DNA of Didymium iridis codes for a group I ribozyme and a novel ribozyme that cooperate in self-splicing. *Cell* **76:** 725-734

Jossinet F, Westhof E. 2005. Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* **21:** 3320-3321

Klein DJ, Schmeing TM, Moore PB, Steitz TA. 2001. The kink-turn: a new RNA secondary structure motif. *EMBO J* **20:** 4214-4221

Lang K, Rieder R, Micura R. 2007. Ligand-induced folding of the thiM TPP riboswitch investigated by a structure-based fluorescence spectroscopic approach. *Nucleic Acids Res* **35:** 5370-5378

Lehnert V, Jaeger L, Michel F, Westhof E. 1996. New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the Tetrahymena thermophila ribozyme. *Chem Biol* **3:** 993-1009

Leontis NB, Westhof E. 1998. A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. *J Mol Biol* **283:** 571-583

Lescoute A, Westhof E. 2006. Topology of three-way junctions in folded RNAs. *RNA* **12:** 83-93

Li Z, Zhang Y. 2005. Predicting the secondary structures and tertiary interactions of 211 group I introns in IE subgroup. *Nucleic Acids Res* **33:** 2118-2128

Lipchock SV, Strobel SA. 2008. A relaxed active site after exon ligation by the group I intron. *Proc Natl Acad Sci U S A* **105:** 5699-5704

Lundblad EW, Einvik C, Ronning S, Haugli K, Johansen S. 2004. Twelve Group I introns in the same pre-rRNA transcript of the myxomycete Fuligo septica: RNA processing and evolution. *Mol Biol Evol* **21:** 1283-1293

Lykke-Andersen J, Aagaard C, Semionenkov M, Garrett RA. 1997. Archaeal introns: splicing, intercellular mobility and evolution. *Trends Biochem Sci* **22:** 326-331

Massire C, Jaeger L, Westhof E. 1998. Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis. *J Mol Biol* **279:** 773-793

Michel F, Ellington AD, Couture S, Szostak JW. 1990. Phylogenetic and genetic evidence for base-triples in the catalytic domain of group I introns. *Nature* **347:** 578-580

Murray HL, Mikheeva S, Coljee VW, Turczyk BM, Donahue WF, Bar-Shalom A, Jarrell KA. 2001. Excision of group II introns as circles. *Mol Cell* **8:** 201-211

Nielsen H, Beckert B, Masquida B, Johansen S. 2008. The GIR1 Branching Ribozyme. 229-252

Nielsen H, Fiskaa T, Birgisdottir AB, Haugen P, Einvik C, Johansen S. 2003. The ability to form full-length intron RNA circles is a general property of nuclear group I introns. *RNA* **9:** 1464-1475

Nielsen H, Johansen SD. 2009. Group I introns: Moving in new directions. *RNA Biol* **6:**

Nover L, Munsche D, Neumann D, Ohme K, Scharf KD. 1986. Control of ribosome biosynthesis in plant cell cultures under heat-shock conditions. Ribosomal RNA. *Eur J Biochem* **160:** 297-304

Raub T.J., Aldrich H.C. 1982. Sporangia, spherules, and. 21-75

Rieder R, Lang K, Graber D, Micura R. 2007. Ligand-induced folding of the adenosine deaminase A-riboswitch and implications on riboswitch translational control. *Chembiochem* **8:** 896-902

Roman J, Woodson SA. 1998. Integration of the Tetrahymena group I intron into bacterial rRNA by reverse splicing in vivo. *Proc Natl Acad Sci U S A* **95:** 2134-2139

Sanger HL, Klotz G, Riesner D, Gross HJ, Kleinschmidt AK. 1976. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proc Natl Acad Sci U S A* **73:** 3852-3856

Schumacher J, Sanger HL, Riesner D. 1983. Subcellular localization of viroids in highly purified nuclei from tomato leaf tissue. *EMBO J* **2:** 1549-1555

Stahley MR, Strobel SA. 2005. Structural evidence for a two-metal-ion mechanism of group I intron splicing. *Science* **309:** 1587-1590

Stoddard BL. 2005. Homing endonuclease structure and function. *Q Rev Biophys* **38:** 49-95

Streicher B, von AU, Schroeder R. 1993. Lead cleavage sites in the core structure of group I intron-RNA. *Nucleic Acids Res* **21:** 311-317

Streicher B, Westhof E, Schroeder R. 1996. The environment of two metal ions surrounding the splice site of a group I intron. *EMBO J* **15:** 2556-2564

Vader A, Johansen S, Nielsen H. 2002. The group I-like ribozyme DiGIR1 mediates alternative processing of pre-rRNA transcripts in Didymium iridis. *Eur J Biochem* **269:** 5804-5812

Vader A, Nielsen H, Johansen S. 1999. In vivo expression of the nucleolar group I intron-encoded I-DirI homing endonuclease involves the removal of a spliceosomal intron. *EMBO J* **18:** 1003-1013

Vicens Q, Cech TR. 2009. A natural ribozyme with 3',5' RNA ligase activity. *Nat Chem Biol* **5:** 97-99

Waldsich C, Masquida B, Westhof E, Schroeder R. 2002. Monitoring intermediate folding states of the td group I intron in vivo. *EMBO J* **21:** 5281-5291

Winkler WC, Nahvi A, Sudarsan N, Barrick JE, Breaker RR. 2003. An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nat Struct Biol* **10:** 701-707

Wright M, Tollon Y. 1982. Induction of heat-shock proteins at permissive growth temperatures in the plasmodium of the myxomycete Physarum polycephalum. *Eur J Biochem* **127:** 49-56

Legends to figures

FIGURE 1. Processing of group I introns. The splicing pathway (left part) is initiated by attack at the 5' splice site by an exogenous guanosine cofactor (exoG). By two consecutive transesterification reactions the intron splices out of the precursor RNA resulting in ligated exons (grey boxes) and a free linear intron (L-IVS) with the exogenous guanosine (exoG) coupled to the 5'end. In some introns, attack of the 3' terminal guanosine (ωG) at an internal site in the L-IVS results in formation of a truncated circular intron RNA and release of a small 5'end fragment. Alternatively, the attack takes place at the three phosphates of the exoG leading to formation of a circular RNA that incorporates the guanosine cofactor and release of pyrophosphate. The L-IVS can reverse splice into a cognate site as shown, or alternatively, into a new sequence context (intron transposition). The circularization pathway (right part) is initiated by hydrolysis at the 3' splice site followed by an attack of the ωG at the 5' splice site. This produces a full-length intron circle (FLC) and un-ligated exons. The FLC can integrate into a target RNA.

FIGURE 2. (**A**). The *Didymium iridis* life cycle is divided into microscopic haploid stages (n) and a macroscopic diploid stage (2n). The haploid cell can reversibly transform between amoeba, flagellated cell and cyst depending on environmental factors. Two compatible flagellates or amoeba fusing to form a diploid zygote initiates sexual reproduction. The zygote grows into a multinucleate plasmodium from which fruiting bodies and spores develop. Spores germinate to form amoebae or flagellates thereby completing the cycle. (**B**) Growth course of *Didymium iridis*. The initial growth after parallel inoculation of *D. iridis* and the *Escherichia coli* food source is exponential. Growing in parallel the *Didymium* cells will gradually clear the suspension of *E. coli* and transform from amoebae to flagellates. Then, the growth rate rapidly decreases and the cells transform into dormant and resistant cysts. The cells can be induced to excyst by addition of a bacterial food source. RNA was harvested at the indicated time points: Early (amoeboid cells), Late (flagellated cells) and Cyst. (**C**). Copy number of FLC per 50 ng whole cell (WC) RNA (left y-axis) and per cell (right y-axis) in RNA isolated from the three time points. (**D**). Copy number of FLC per 50 ng WC RNA (left y-axis) and per cell (right y-axis) in exponentially growing cells submitted to cold-shock at 5 ^{o}C or 10 ^{o}C or heat-shock at 34 ^{o}C or 40 ^{o}C, respectively. FLC per cell for heat-shock at 40 ^{o}C cells was

omitted due to impaired cell count. Values in C and D are given as mean ± standard error of the mean.

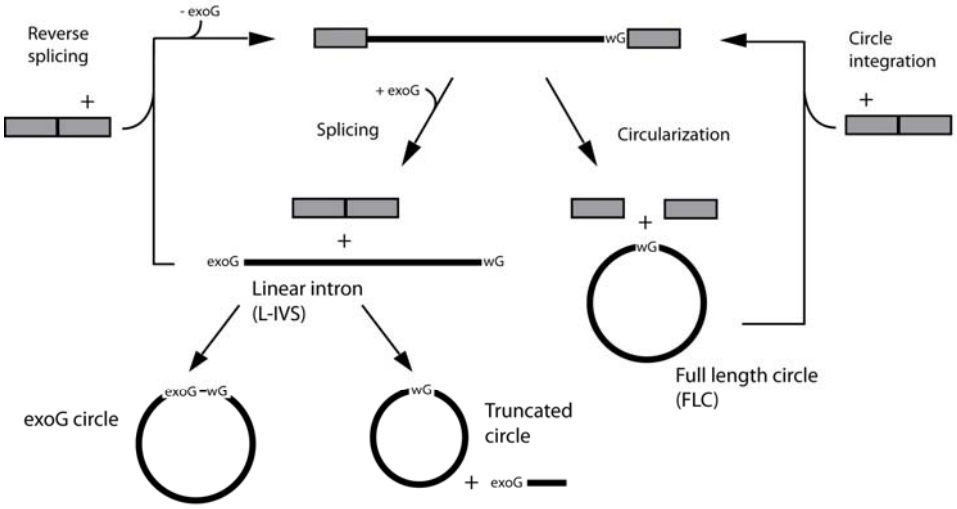
FIGURE 3. Probing results for DiGIR2 FLC and L-IVS. **(A)** The results for the single strand specific modifying probes (DMS (modifying A>C), DEP (A-specific), Kethoxal (G-specific) and CMCT (U>>G) and the single strand specific cleavage probes RNase T1 (G specific), T2 (A preference) and A (U and C specific) as well as the double strand specific RNase V1 are depicted on the proposed secondary structure as specified in the insert box. Prior to this study, it was not known to what extent circularization would result in alterations of the structure. Thus, the diagram of the FLC was made by simply connecting the 5' and 3' end of the intron in a standard representation of the linear form. The extent of probing signal has been scored as 1, 2 or 3 with 3 being the strongest and is represented by the lengths of the arrows. **(B)** Signals in FLC compared to L-IVS. Red indicates increased signal in FLC, green decreased signal. Type of arrow indicates probe specificity as listed in the insert box (ds: double strand, ss: single strand) and letters next to arrows specify the probe (M: DMS, D: DEP, K: Kethoxal, C: CMCT, T1: RNase T1, T2: RNase T2, V1: RNase V1). The paired segments are denoted P1-P13 and the nucleotide position within the DiGIR2 construct is indicated by numbers.

FIGURE 4. Pb^{2+} induced cleavages in the catalytic core. **(A)** Primer extension analysis of DiGIR2 L-IVS probed with increasing concentrations of Pb^{2+} compared to a control (lane 0-5) and cleavage at a constant Pb^{2+} concentration (2 mM) in presence of increasing concentrations of Mg^{2+} (lane 6-9). To the right, grey arrows indicate Mg^{2+} independent Pb^{2+} cleavage sites; black arrows indicate cleavages affected by Mg^{2+} concentration. The length of the arrows as in Fig. 3 represents signal strength (1 to 3). To the left is a DiGIR2 sequencing ladder electrophoresed in parallel with samples and bars indicating structural elements in DiGIR2. **(B)** Primer extension analysis of Pb^{2+} cleavage of DiGIR2 FLC. Figure annotations are identical to (A). **(C)** Results from (A) and (B) depicted on a secondary structure diagram of the J8/7 region. Gray letters indicate nucleotide positions that are not covered in (A) and (B). The open grey arrow indicates the L-IVS specific Pb^{2+} cleavage at U225.

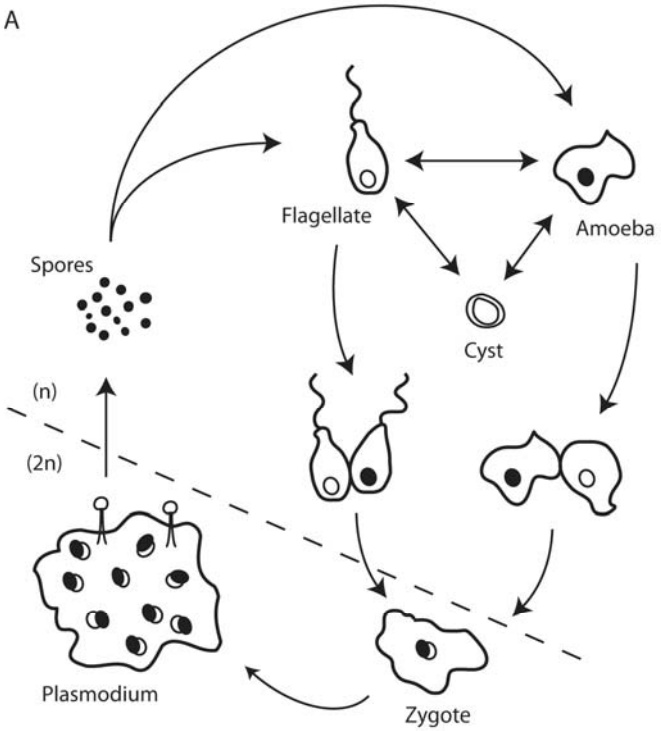
FIGURE 5. Structural analysis of the P9 domain. **(A)** Sequence structure of the internal loop in P9.1 in 16 sequences of group IE introns. Most sequences show a symmetrical internal loop

containing two nucleotides fully conserved and a third one conserved in all but two sequences (Pbi, Lov). These nucleotides (boxed) may be important in the structuring of the loop and/or in its ability to interact with P7. A list of the introns can be found in Fig. S4. **(B)** Ribbon 3D model showing how the internal loop of P9.1 is thought to be involved in stabilizing P7 (red arrow in left panel) and P9.2 is involved in stabilization of P9.0 (red arrow in right panel), respectively. The internal loop (IL9.1) of P9.1 is depicted in cyan, P7 in orange, P9.0, P9a and P9b in lime green, P9.1 and P9.2 in green. ωG is represented as an atom model with coloring according to atom type (CPK).

Andersen et al. 28

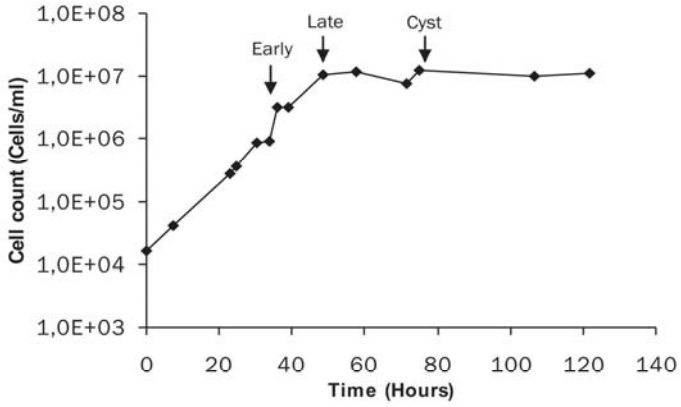


Andersen et al. Figure 1

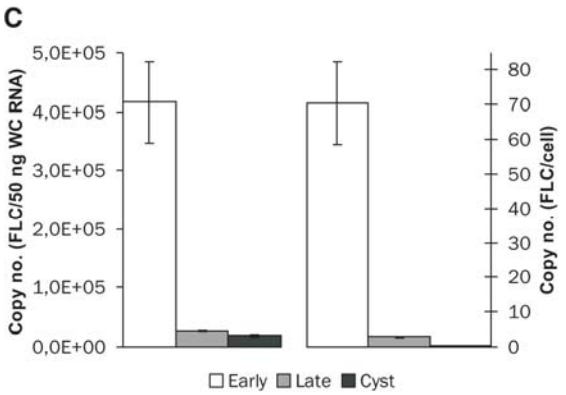


Andersen et al. Figure 2A

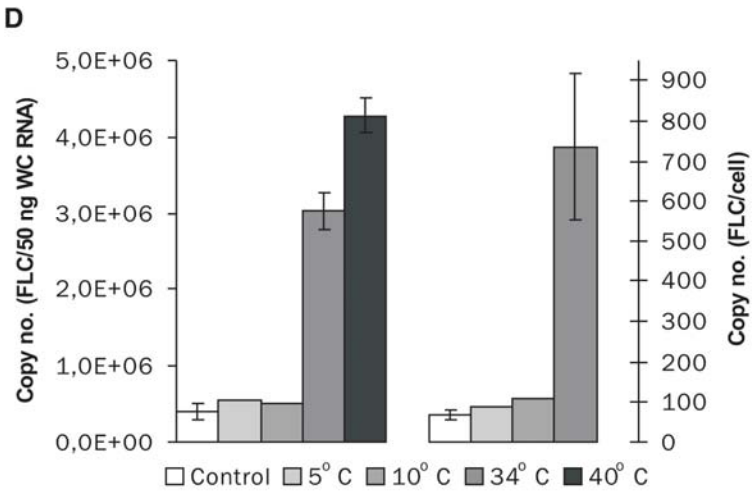
B



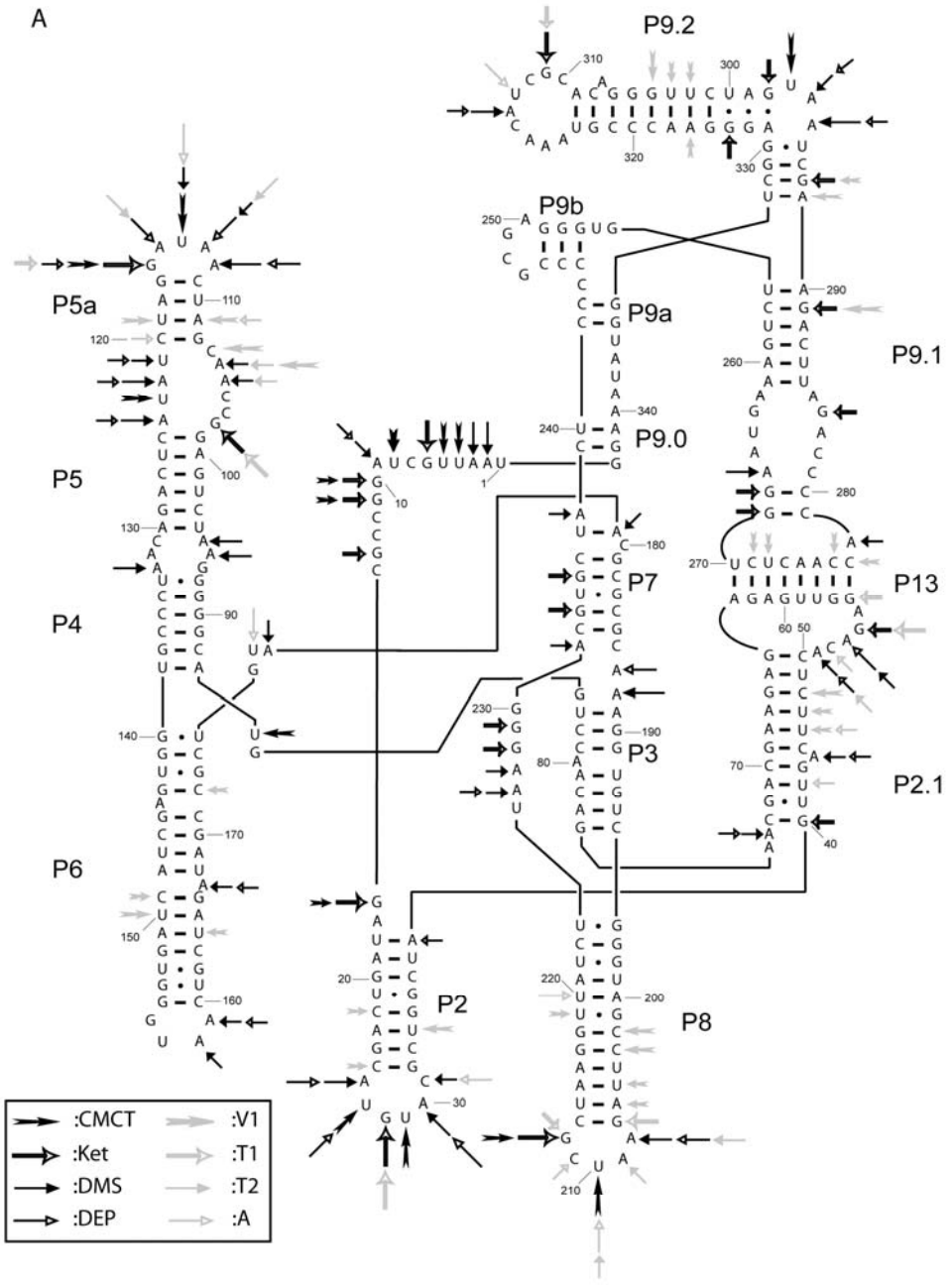
Andersen et al. Figure 2B



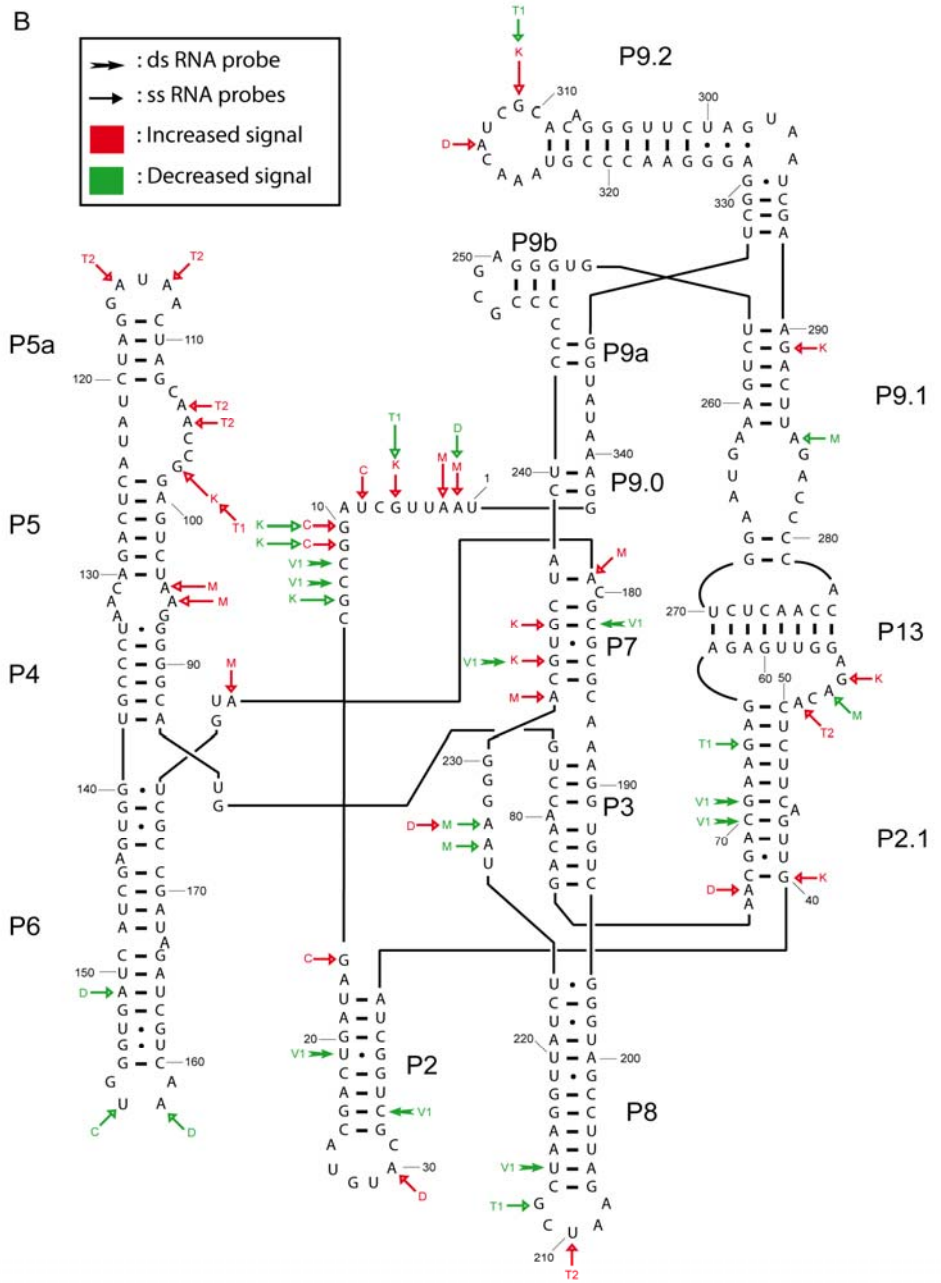
Andersen et al. Figure 2C



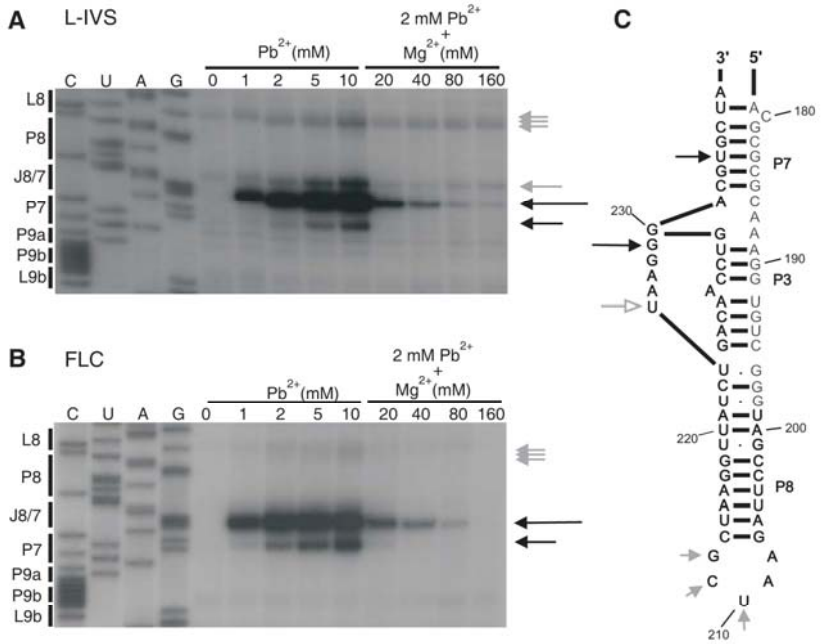
Andersen et al. Figure 2D



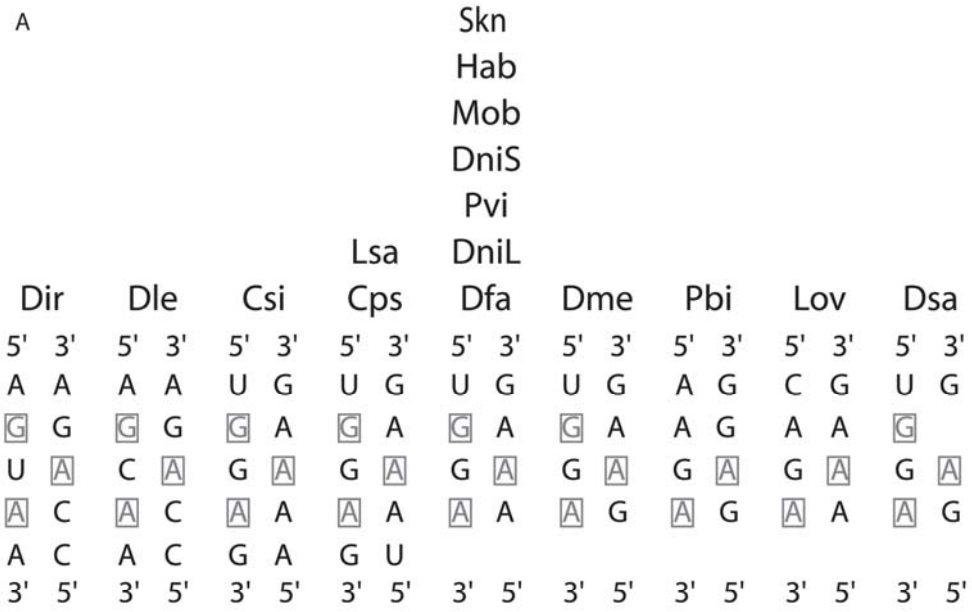
Andersen et al. Fig. 3A



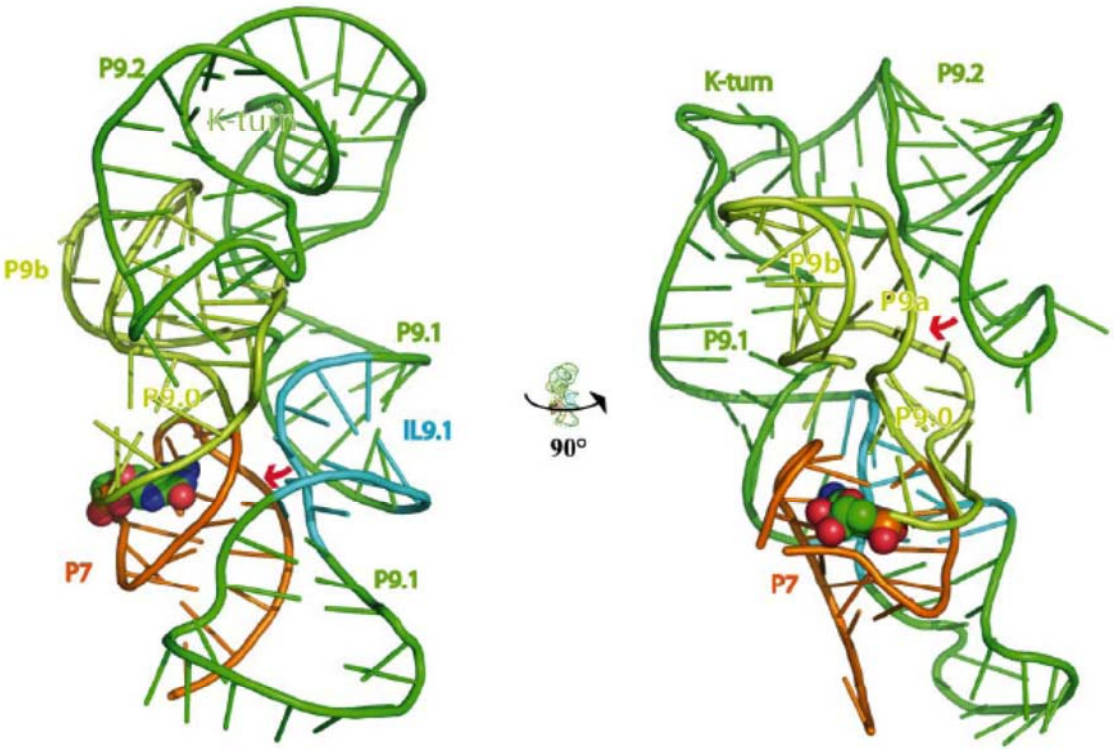
Andersen et al. Figure 3B



Andersen et al. Figure 4



Andersen et al. Figure 5A



Andersen et al. Figure 5 B

Accumulation of stable full-length circular group I introns during heat-shock

Kasper L. Andersen, Bertrand Beckert, Benoit Masquida, Morten Andreassen, Steinar D. Johansen, and Henrik Nielsen

Supplementary on-line information

The full-length circular intron is predominantly nuclear (Fig. S1)

For analysis of the cellular localization of FLC, exponentially growing amoebae were fractionated into nuclei and cytosol and RNA isolated from the two fractions. Then, the RNA was reverse transcribed and qRT-PCR was performed with a FLC specific primer set. In this and similar experiments throughout the study, quantifications were calibrated by a standard curve obtained by analysis of a FLC circle junction sequence produced *in vitro* and quantitated by a fluorometric assay. Control experiments were performed to ensure that the FLC were not produced from linear precursors during the course of the experiment (not shown). FLC is predominantly found in the nucleus and it cannot be ruled out that all FLC copies are in fact nuclear in growing amoebae.

For FLC specific amplification, two primers that would result in a 127 bp PCR-product spanning the circularization junction were selected (C394: 5'–TGA TCT TGG GAC ACG CTA CA and C395: 5'–TAC TTC CCA ACC CAA CCA AA). This product was sequenced in several experiments showing only the expected circularization site with no evidence of truncated circles.

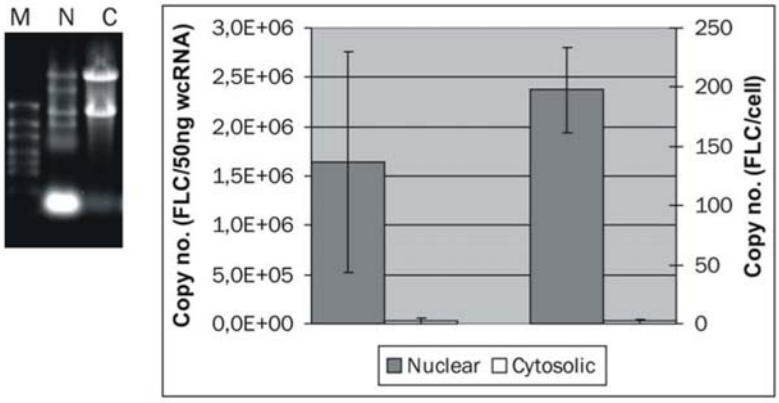


Fig. S1: The full length intron circle (FLC) is predominantly confined to the nucleus. Left panel: 10 µg RNA from cellular fractions used for qRT-PCR detection of FLC was separated on 0.9 % denaturing formaldehyde agarose gel and subsequently stained with ethidium bromide. Marker (M) (Fermenatas), nuclear fraction (N), cytosolic fraction (C). Right panel illustrates the copy number of FLC per 50ng whole cell RNA (wc) and per cell in nuclear and cytosolic fractions. Values are given as mean ± standard error of the mean (n=4).

Products of the circularization pathway constitute approximately 1% of processing products from pre-rRNA processing (Table S1)

Up-regulation of FLC could either be due to an overall increase in the steady-state amounts of intronic RNA species or a change in the partitioning of precursor transcripts between the splicing and the circularization pathways. In order to address the latter possibility, we designed a primer set that would amplify the products of the splicing pathway specifically. The spliced out intron contain a 51-nt spliceosomal intron that is removed as the last step in a series of processing steps leading to the mature mRNA encoding the homing endonuclease (Vader et al. 1999). The splicing event creates an exon-exon junction sequence that is not found in the precursor or in FLC. Thus, one primer was designed to span this junction. When used in qRT-PCR experiments in parallel with the FLC-specific primer set, we were able to determine changes in the partitioning between the two

pathways under the assumption that the relative stabilities of the end-products of the two pathways were not affected. Under all conditions examined, the splicing pathway was by far the dominating pathway. However, the proportion of precursors undergoing circularization was slightly down-regulated during late-phase exponential growth and encystment and up-regulated by a factor of 2.3 and 5.3 at 34 °C and 40 °C, respectively.

	Early	Late	Cyst	Cold-shock 5°C	Cold-shock 10°C	Heat-shock 34°C	Heat-shock 40°C
FLC / HEG mRNA	0.0094	0.0040	0.0077	0.0094	0.0086	0.0214	0.0495
Normalized	1.0	0.4	0.8	1.0	0.9	2.3	5.3

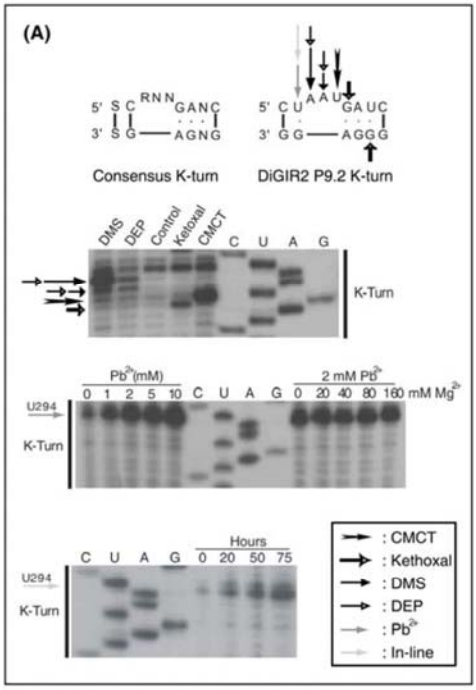
Table S1: Intron processing pathways. The ratios of intron full-length circle (FLC) and homing endonuclease mRNA (HEG mRNA) copy number determined by qRT-PCR for the experimental conditions described in the paper. In the lower row, the ratios are normalized to exponentially growing cells at 25 °C. The primers for FLC specific amplification were similar to those described above. For HEG mRNA specific amplification, the 5' primer was designed to span the exon-exon junction created by splicing of the small spliceosomal intron found in the Dir.S956-1 group I intron (C397: 5'-GCG TTC TAG GCC CGA TG). Amplification using this and a 3' primer (C398: 5' -CCT TGC TTG TCT GGA TCC TC) resulted in a PCR-product of 126 bp.

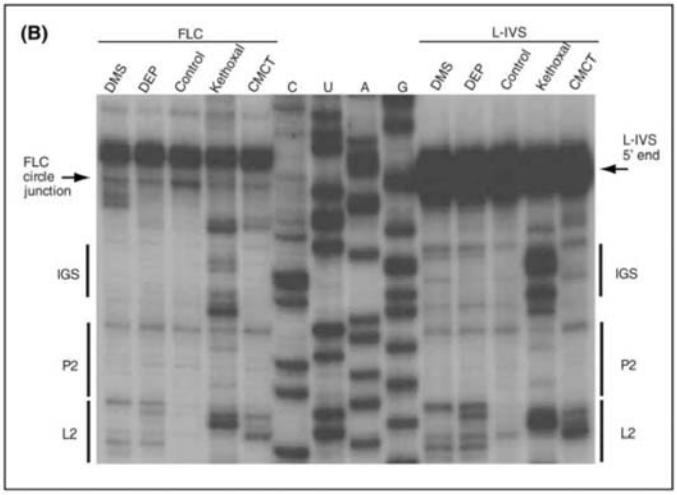
Structure probing analysis of the group IE intron ribozyme from the Dir.S956-1 twin-ribozyme intron in its linear and full-length circular form (Figs. S2 and S3; Table S2).

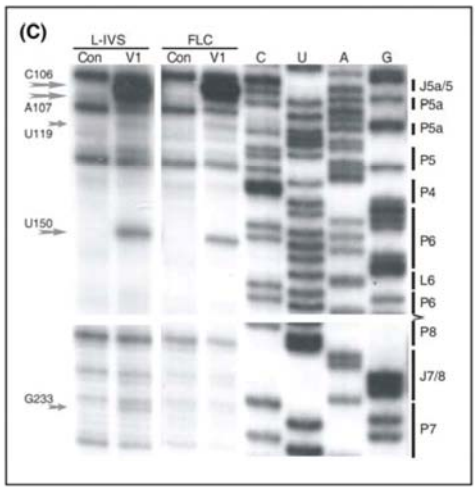
A structure probing analysis was carried out as described in the main paper. In short, RNA was probed after reaction, directly in the reaction mixture and the relevant RNA species subsequently

isolated for analysis. All cleavage probes were titrated to single-hit conditions using a linearization assay based on linearization of DiGIR2 FLC. Fig. S2 presents samples of the autoradiograms from the analysis. In Fig. S3 the results from chemical and RNase probing of the L-IVS are presented on the secondary structure diagram of the linear form of DiGIR2 based on the current rules for representing group I introns. Table S2 lists the results of Pb^{2+} and in-line probing.

Fig. S2: Examples of chemical, RNase, Pb^{2+} and in-line probing of linear (L-IVS) and circular (FLC) DiGIR2. (a): Probing signature of the K-turn in P9.2 of DiGIR2 L-IVS. Top panel: the consensus sequence of the K-turn motif compared to the K-turn sequence of the P9.2 element in GIR2. "R": A\G, "N": A\G\ C\U, "S": G\C. The panels below show results from chemical, Pb^{2+} and in-line probing, respectively. Probe, concentration of Pb^{2+} /Mg^{2+}, hours of in-line treatment and sequence lanes are noted above each panel. The probe specific signals are highlighted by arrows according to the inserted box and are positioned at the corresponding nucleotides in the DiGIR2 K-turn secondary structure diagram above. As in Figure 3 of the main paper arrow length corresponds to the strength of the observed signal. **(b):** Comparison of chemical probing pattern from the 5' region of the L-IVS including the internal guide sequence (IGS) to the corresponding sequence in the FLC. Probes and sequence reaction are indicated above the lanes and the intron paired segment (P2), loop (L2) as well as the internal guide sequence (IGS), the FLC circle junction, and the L-IVS 5'end are indicated on the sides. **(c):** Double strand specific RNase V1 cleavages in L-IVS compared to FLC from paired segments P7, P6 and P5a as well as single stranded joining segment J5a/5. Control (con), probe RNase V1 (V1) and sequence lanes are noted above. Specific RNase V1 cleavage signals in L-IVS and/or FLC as well as structural segments are indicated on the sides.







Andersen et al.; SOM; 5

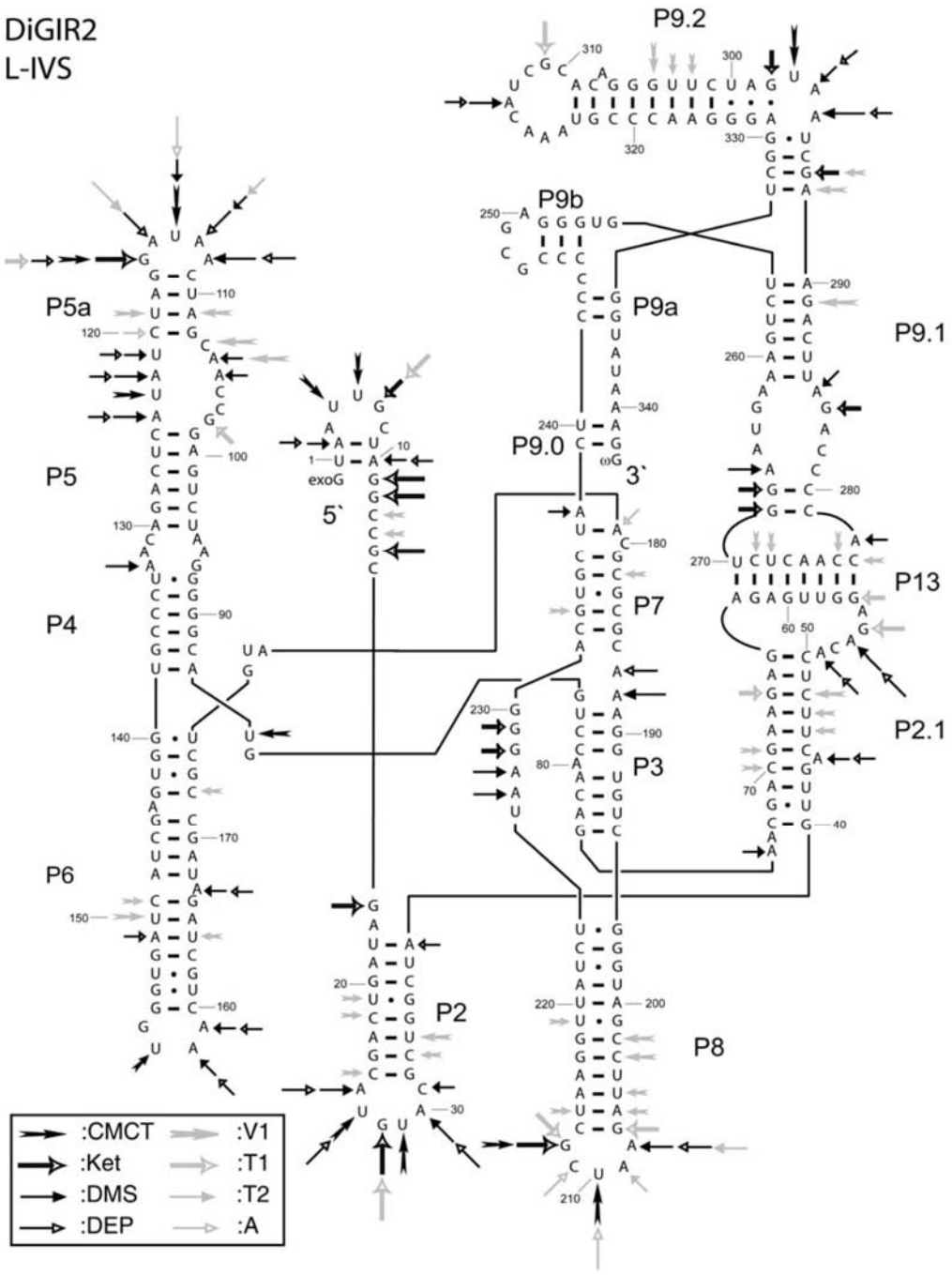


Fig. S3: Chemical and RNase probing of the linear DiGIR2 intron (L-IVS). Signals from modifications and cleavages are indicated with arrows on the proposed secondary structure of DiGIR2 L-IVS. The various probes are specified in the inserted box. The strength of modification/cleavage signals has been scored by as 1, 2 or 3; 3 being the strongest. The lengths of the arrows in the figures correspond to these scores. Thus a long arrow indicates a strong signal.

Andersen et al.; SOM; 6

			1) FLC:		2) L-IVS:	
Position	Element	**Sequence**	Pb(2+)	Inline	Pb(2+)	Inline
200	P8'	**A**	*	*	*	*
201	P8'	**G**	*	*	*	*
202	P8'	**C**	*	*	*	*
203	P8'	**C**	*	*	*	*
204	P8'	**U**	*	*	*	*
205	P8'	**U**	*	*	*	*
206	P8'	**A**	*	*	*	*
207	P8'	**G**	*	*	*	*
208	L8	**A**	*	In-1	*	In-1
209	L8	**A**	*	In-1	*	In-1
210	L8	**U**	Pb-1	In-2	Pb-1	In-2
211	L8	**C**	Pb-1	In-2	Pb-1	In-2
212	L8	**G**	Pb-1	In-1	Pb-1	In-1
213	P8''	**C**	*	*	*	*
214	P8''	**U**	*	*	*	*
215	P8''	**A**	*	*	*	*
216	P8''	**A**	*	*	*	*
217	P8''	**G**	*	*	*	*
218	P8''	**G**	*	*	*	*
219	P8''	**U**	*	*	*	*
220	P8''	**U**	*	*	*	*
221	P8''	**A**	*	*	*	*
222	P8''	**U**	*	*	*	*
223	P8''	**C**	*	*	*	*
224	P8''	**U**	*	In-1	*	In-1
225	J7/8	**U**	*	In-1	Pb-2	In-1
226	J7/8	**A**	*	*	*	*
227	J7/8	**A**	*	*	*	*
228	J7/8	**G**	*	In-1	*	In-2
229	J7/8	**G**	Pb-3 (Mg2+)	In-2	Pb-3 (Mg2+)	In-3
230	J7/8	**G**	*	*	*	*
231	P7''	**A**	*	*	*	*
232	P7''	**C**	*	*	*	*
233	P7''	**G**	*	*	*	*
234	P7''	**U**	Pb-2 (Mg2+)	*	Pb-2 (Mg2+)	*
235	P7''	**G**	*	*	*	*
236	P7''	**C**	*	*	*	*
237	P7''	**U**	*	*	*	*
238		**A**	*	*	*	*
239	P9.0'	**C**	*	*	*	*
240	P9.0'	**U**	*	*	*	*
241	P9a'	**C**	*	*	*	*
242	P9a'	**C**	*	*	*	*
243		**C**	*	*	*	*
244	P9b'	**C**	*	*	*	*
245	P9b'	**C**	*	*	*	*
246	P9b'	**C**	*	*	*	*
247	L9b	**G**	*	*	*	*
248	L9b	**C**	*	*	Pb-1	*
249	L9b	**G**	*	In-1	Pb-1	In-1
250	L9b	**A**	*	*	*	*
251	P9b''	**G**	*	*	*	*
252	P9b''	**G**	*	*	*	*
253	P9b''	**G**	*	*	*	*
254		**U**	*	*	Pb-1	*
255		**G**	*	*	*	*
256	P9.1'	**U**	*	*	*	*
257	P9.1'	**C**	*	*	*	*
258	P9.1'	**U**	*	*	*	*
259	P9.1'	**G**	*	*	*	*
260	P9.1'	**A**	*	*	*	*
261	P9.1'	**A**	*	*	*	*
262	L9.1'	**A**	*	In-1	*	In-1
263	L9.1'	**G**	*	In-1	*	In-1
264	L9.1'	**U**	*	*	*	*

265	L9.1'	A	*	*	*	*
266	L9.1'	A	*	*	*	*
267	L9.1'	G	*	*	*	*
268	L9.1'	G	*	*	*	*
269	P13''	U	*	*	*	*
270	P13''	C	*	*	*	*
271	P13''	U	*	*	*	*
272	P13''	C	*	*	*	*
273	P13''	A	*	*	*	*
274	P13''	A	*	*	*	*
275	P13''	C	*	*	*	*
276	P13''	C	*	*	*	*
277	L9.1''	A	*	*	*	*
278	L9.1''	C	*	*	*	*
279	L9.1''	C	*	*	*	*
280	L9.1''	C	*	*	*	*
281	L9.1''	C	*	In-1	*	In-1
282	L9.1''	A	*	*	*	*
283	L9.1''	G	*	*	*	*
284	L9.1''	A	*	*	*	*
285	P9.1''	U	*	*	*	*
286	P9.1''	U	*	*	*	*
287	P9.1''	C	*	*	*	*
288	P9.1''	A	*	*	*	*
289	P9.1''	G	*	*	*	*
290	P9.1''	A	*	*	*	*
291	P9.2'	A	*	*	*	*
292	P9.2'	G	*	*	*	*
293	P9.2'	C	*	*	*	*
294	P9.2'	U	*	In-1	Pb-2	In-1
295	(P9.2')	A	*	*	Pb-1	*
296	(P9.2')	A	*	*	*	*
297	(P9.2')	U	*	*	*	*
298	P9.2'	G	*	*	*	*
299	P9.2'	A	*	*	*	*
300	P9.2'	U	*	*	*	*
301	P9.2'	C	*	*	*	*
302	P9.2'	U	*	*	*	*
303	P9.2'	U	*	*	*	*
304	P9.2'	G	*	*	*	*
305	P9.2'	G	*	*	*	*
306	P9.2'	G	*	*	*	*
307	(P9.2')	A	*	*	*	In-2
308	P9.2'	C	*	*	Pb-1	In-1
309	P9.2'	A	Pb-1	*	Pb-1	In-1
310	L9.2	C	Pb-1	In-1	Pb-1	In-1
311	L9.2	G	Pb-1	In-2	Pb-1	In-3
312	L9.2	C	Pb-1	*	Pb-1	*
313	L9.2	U	Pb-1	In-1	Pb-2	In-1
314	L9.2	A	Pb-1	In-1	Pb-2	In-1
315	L9.2	C	*	In-1	Pb-1	In-1
316	L9.2	A	*	*	*	*
317	L9.2	A	*	nd	nd	nd
318	L9.2	A	*	nd	nd	nd

Table S2: Results from Pb^{2+} and in-line probing of nucleotide 200-318 of DiGIR2 FLC and L-IVS. Signal strengths are indicated as weak: -1, intermediate: -2, or strong: -3; "*": no signal; nd: not detected due to proximity to primer. Some successive signals could be due to reverse transcriptase "stuttering". This has only been corrected in the most obvious cases.

Description of the overall structure of the L-IVS and the FLC.

P4-P6 domain. This domain has been shown in some introns to fold early and act as a scaffold in folding of subsequent domains (Murphy and Cech 1993). In DiGIR2, P5 is extended by an asymmetrical internal loop and a helix (P5a). The P5 and P6 helices in L-IVS were confirmed by V1 cleavages and their capping loops were accessible to extensive chemical modification. The central part of the molecule appears to be inaccessible to modification and RNase cleavage. The asymmetrical loop connecting P5 and P5a is readily modified by chemical probes and appears not to contain Watson-Crick base-pairs. Interestingly, two of the strongest V1 cleavage sites overall are found in the single-stranded J5/5a at A106 and C107 indicating stacking of these residues. The probing results from FLC were very similar to that of L-IVS, but with some exceptions. An increased accessibility of probes to the 5' strand of the top of the domain in FLC and a decreased accessibility of probes to the 5' strand of P6 and parts of L6 indicate a slightly different positioning of this domain in FLC compared to L-IVS, and is most likely linked to the absence of P1 in FLC. Overall, the structure of the P4-P6 domain was confirmed and did not contain major structural alterations in the FLC.

P3-P9 domain. This domain harbours the binding site for the guanosine co-factor and ωG in P7. In DiGIR2, the P9 part is very complex consisting of a 2-bp helix (P9.0) and three hairpins branching from this, some of which have been specifically assigned a role in the circularization pathway (Haugen et al. 2004). The probing strategy excluded data collection from P9.0'' and P9a'' and the most 3' part of P9.2 in the L-IVS due to overlap or proximity to the primer used in reverse transcription. Likewise, no data is available for 16-22 nucleotides spanning the FLC circle junction due to the circular RNA purification method. From analysis of those parts that are amenable to analysis, P9.0, P9a and P9b appear inaccessible to chemical modification and RNase cleavage, indicating that this part of the structure is buried within the ribozyme. P9.1 and P9.2 are supported

by V1 cleavages and the bulge in P9.2 is accessible to chemical modification. In contrast the 9-nt L9.2 and the 10-nt internal loop in P9.1 are only modifiable in a few positions, indicating that they form structured motifs. An 8-nt stretch of L9.1 is cleaved at several positions by V1 and is not modified by chemical probes, indicating an involvement in a long-range base-pairing (P13; see below). P7 and P8 were confirmed by the presence of V1 cleavages, and the joining segments J8/7, J7/3, J7/9.0, and L8 by accessibility to chemical modification. The probing results for FLC were similar to that of L-IVS in P3, P8, and P9. However, several differences were noted in and around P7. The V1 cleavages at C182 and G233 were not observed in FLC. In accordance with this, the FLC was accessible to DMS modification at A179, and kethoxal modification at G233 and G235. In addition, A231 in J8/7 and A178 in J6/7 were modifiable by DMS in FLC, but not in L-IVS. Taken together, the results indicate a structural alteration of the G-binding site in the FLC compared to the L-IVS.

P1-P2 domain. This is frequently referred to as the substrate domain because it harbours the 5'-splice site. In DiGIR2, the P1-P2 domain consists of an 8 base-pair P1 including 6 base-pairs between the internal guide sequence and the 5'-exon, as well as P2 and P2.1 helices. In the DiGIR2 L-IVS, the 5'-exon is removed and the guanosine co-factor is attached to the 5'-end of the intron. L2 is artificial in the sense that DiGIR1 and HEG have been removed by deletion. However, P2 is roughly similar to P2 known from other group IE introns not carrying an insertion. The remainder of P1 is readily accessible to chemical modification, as expected. P2 and P2.1 are confirmed by numerous V1 cleavages and L2 is modifiable with the chemical probes. The major part of L2.1 appears protected against chemical modification consistent with its involvement in a long-range base-pairing interaction with L9.1. In all, the P1-P2 domain appears as the most accessible domain to the probes applied in this study. Obviously, the FLC differs from the L-IVS in that the exoG is missing and the 5'- and 3'- nucleotides of the intron are covalently linked. The nucleotides on the

5'- side of the junction are accessible to chemical modification and thus appear to be solvent exposed. In contrast, nucleotides 3'- to the junction appear inaccessible. The modification pattern of the circle junction region of the FLC is slightly different from that found within the 5' end of the L-IVS indicating structural differences. The P2-P2.1 part of the domain is mostly similar in the two molecular species.

P13 long-range interaction. This group IE characteristic long-range base-pairing interaction was confirmed by numerous V1 cleavages on the L9.1 strand that consequently is proposed to constitute the surface exposed part of the helix. P13 is found in both the L-IVS and the FLC.

3D modelling of the intron (Figs. S4 and S5).

The model of the core of the L-IVS was built by homology modelling with the *Azoarcus* group I ribozyme crystal structure (Adams et al. 2004a; Adams et al. 2004b). Using this strategy, all nucleotides equivalent to specific positions in the *Azoarcus* ribozyme from the P3-P9 and the P4-P6 domains were built in accordance with secondary and tertiary interactions that form upon domain assembly. In this context, the precise positioning of the P9 hairpin allowed for the identification of the receptor of the L9 GCGA tetraloop. This receptor is located in the P5 region proximal to the J4/5 internal loop and corresponds to the consecutive base-pairs C97=G129 and U98-A128 which interact with the sugar edge of A250 and G249 from the tetraloop, respectively. These interactions are in agreement with the covariation usually observed within this tertiary structure motif (Costa and Michel 1997) and are moreover supported by sequence covariation in a set of representatives of the group IE ribozymes (Fig. S4). Homology modelling was also applied to build the characteristic GoU base-pair of the P1 substrate helix onto the core. The P1 hairpin was further built in the form appearing in L-IVS before the first step of the circularization pathway. The differences between DiGIR2 and the *Azoarcus* ribozyme cores were addressed by taking advantage of the more closely

related *Tetrahymena thermophila* ribozyme crystal structure deprived of P1, P2 and P2.1 (Guo et al. 2004). In DiGIR2, P3 is a 7-bp stem with a bulge between the fourth and fifth base-pairs. J8/7 and J3/4 are 6-nt and 3-nt long, respectively. This situation is specific to DiGIR2 and is closer to introns that contain a P2/P2.1 extension like the *Tetrahymena* IC1 ribozyme. The group IC3 introns, like the *Azoarcus* ribozyme, contain a simple P2 hairpin. In these, P3 is 6-bp long, J8/7 and J3/4 are 6-nt and 4-nt long, respectively. In accordance with the *Tetrahymena* ribozyme crystal structure, a triple interaction was built in the DiGIR2 model by docking the Watson-Crick edge of U224 in the narrow groove of the second base-pair of P3, A77-U194. Concerning J3/4, the 5' residue of J3/4 in the *Azoarcus* ribozyme falls in place with U83 from DiGIR2 involved in the terminal base-pair of P3 with A189. Thus, the additional base-pair of P3 in group IE introns compensates a shorter J3/4 as compared to group IC3 ribozymes.

```
                        P5'          P5"          L9
Dir.S956-1          5' UCUG..(L5)..CAGA.(core).GCGA. 3'
Dsa.S956            5' UCUC........CAGA........GUGA. 3'
Dfa.L1975           5' CCUC........GAGG........GUGA. 3'
Dni.L1975           5' UCUC........GAGA........GCGA. 3'
Pvi.S956            5' CCUG........CAGG........GCGA. 3'
Dle.S956            5' UCUG........CAGA........GUGA. 3'
Dni.S956            5' ACCC........GGGU........GUAA. 3'
Mob.S956            5' AGGG........CCUU........GUAA. 3'
Hab.S956            5' CCCU........AGGG........GUAA. 3'
Csi.S956            5' GC-U........AGGU........GUAA. 3'
Lsa.S956            5' GGCC........GGUU........GUAA. 3'
Lov.S956            5' GCCU........AGGU........GUAA. 3'
Skn.S956            5' GCUA........UGGG........GAAA. 3'
Cps.S956            5' GCCU........AGGU........GUAA. 3'
Dme.S956            5' UCUG........CAGA........GUGA. 3'
```

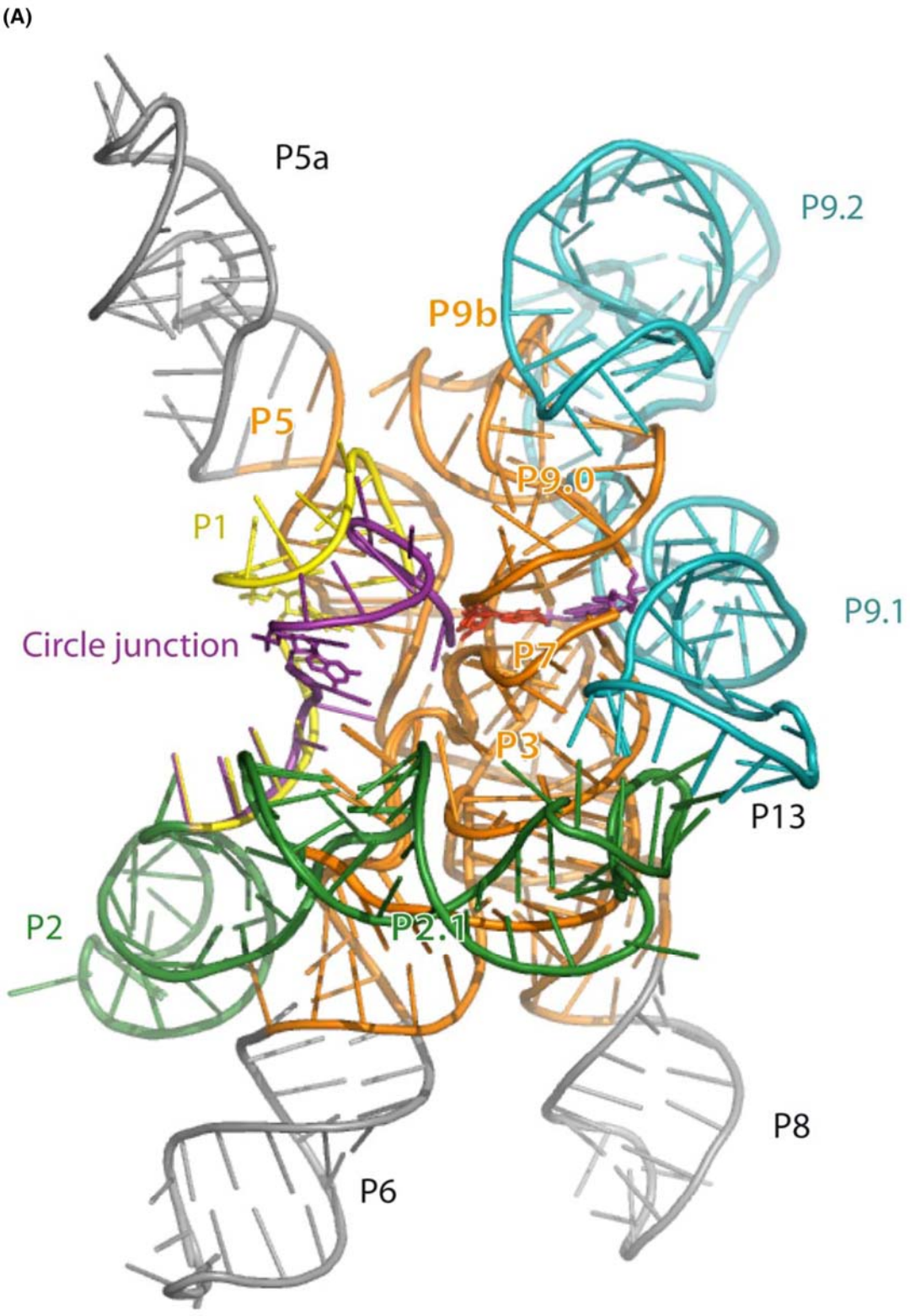
Fig. S4: Sequence covariation between the P5 receptor and L9 sequence. Intron nomenclature is according to (Johansen and Haugen 2001). Abbreviated names stand for *Didymium iridis* (Dir), *Diderma saundersii* (Dsa), *Physarum bivalve* (Pbi), *Diderma fallax* (Dfa), *Diderma niveum* (Dni), *Physarum viresens* (Pvi), *Diachea leucopodia* (Dle), *Cribraria cancellata* (Cca), *Macbrideola oblonga* (Mob), *Hermitrichia abietina* (Hab), *Comatricha sinuatocolumellata* (Csi), *Lamproderma sauteri* (Lsa), *Lamproderma ovoideum* (Lov), *Spizellomyces kniepii* (Skn), *Comatricha pseudoaplina* (Cps), *Diderma meyeriae* (Dme).

After building the L-IVS core, the regions non-homologous to the *Azoarcus* ribozyme, namely P2-P2.1 and the P9.1-P9.2 extensions characteristic of group IE introns (Li and Zhang 2005) were modelled step by step in order to form the P13 pseudoknot (unpublished data). The model of the full intron of group IC1 from *Tetrahymena* (Lehnert et al. 1996) which displays a P13 element also formed by the interaction between the loops of P2.1 and P9.1 was used as a starting point to elucidate the conformation of these appendages in DiGIR2 (Figs. S5A and B). Accordingly, the P2 and P2.1 elements were stacked head to head to form a rod roughly orthogonal to the P1 helix. The P2.1 helix was oriented towards the P3-P9 side whereas P2 was directed towards the P4-P6 domain. On the opposite side of the ribozyme, the P9 domain consists in a four-way junction (4WJ) composed by P9a, P9b, P9.1 and P9.2 that extends the 2-bp P9.0 element stacked onto P7 that allows the ribozyme 3'-residue ωG to be accommodated in the G-binding pocket. In order to allow L9b to interact with P5, P9a had to be stacked with P9b. Furthermore, P9.2 was stacked with P9.1 to allow the latter to lie along P7/P3 and interact with L2.1 to finally form the P13 pseudoknot. The conformation of the elements encompassing P2/P2.1 and the P9 insertion is moreover supported by the observation that V1 cleavages are only observed on the P13 strand belonging to P9.1 which is indeed exposed to the solvent in the model whereas the opposite strand is buried.

The P5a appendage in DiGIR2 is much shorter than in the *Tetrahymena* ribozyme (Guo et al. 2004) and about the same size as in the *Twort* ribozyme (Golden et al. 2005). In the crystal structure of the former, the P4-P6 domain is entirely visible in the density map and adopts the same overall fold as the independent domain (Cate et al. 1996) with the tetraloop L5b interlocked with its receptor located at the junction between P6a and P6b. In the latter, the P5a appendage is not visible in the crystal structure beyond the P5 receptor of the L9 loop. Chemical and enzymatic modifications in DiGIR2 show that the P5a extension folds as a hairpin with residues from the L5a loop and from the internal loop connecting to P5 extensively accessible to Watson-

Crick probes. Thus, P5a appears to point into the solvent in an undefined direction rather than folding back to make specific contacts with other regions of the P4-P6 domain. This conclusion is supported by sequences of ribozymes phylogenetically related to DiGIR2 showing that P5 is often extended by Watson-Crick pairs or by motifs unable to kink the helix, such as asymmetrical bulges or C-loops (Lescoute et al. 2005).

The present study represents the first whole atom modelling of a group IE intron ribozyme. Previously, secondary structures and cylinder models of three subgroups of group IE introns have been proposed (Li and Zhang 2005). In addition, the group IE intron from *Candida* has been modelled based on Fe(II)-EDTA and T1 cleavage patterns (Xiao et al. 2005). Our experimental data and modelling generally conform to the previously proposed models (Li and Zhang 2005; Suh et al. 1999). In particular, the characteristic peripheral elements P2.1, P9.1 and P9.2 and the long-range interaction P13 were all incorporated into the model. In addition, our model contributes important new structural features each specific to the splicing or full-length circularization pathway following the observation of two distinct chemical and enzymatic probing patterns.



Andersen et al.; SOM; 15

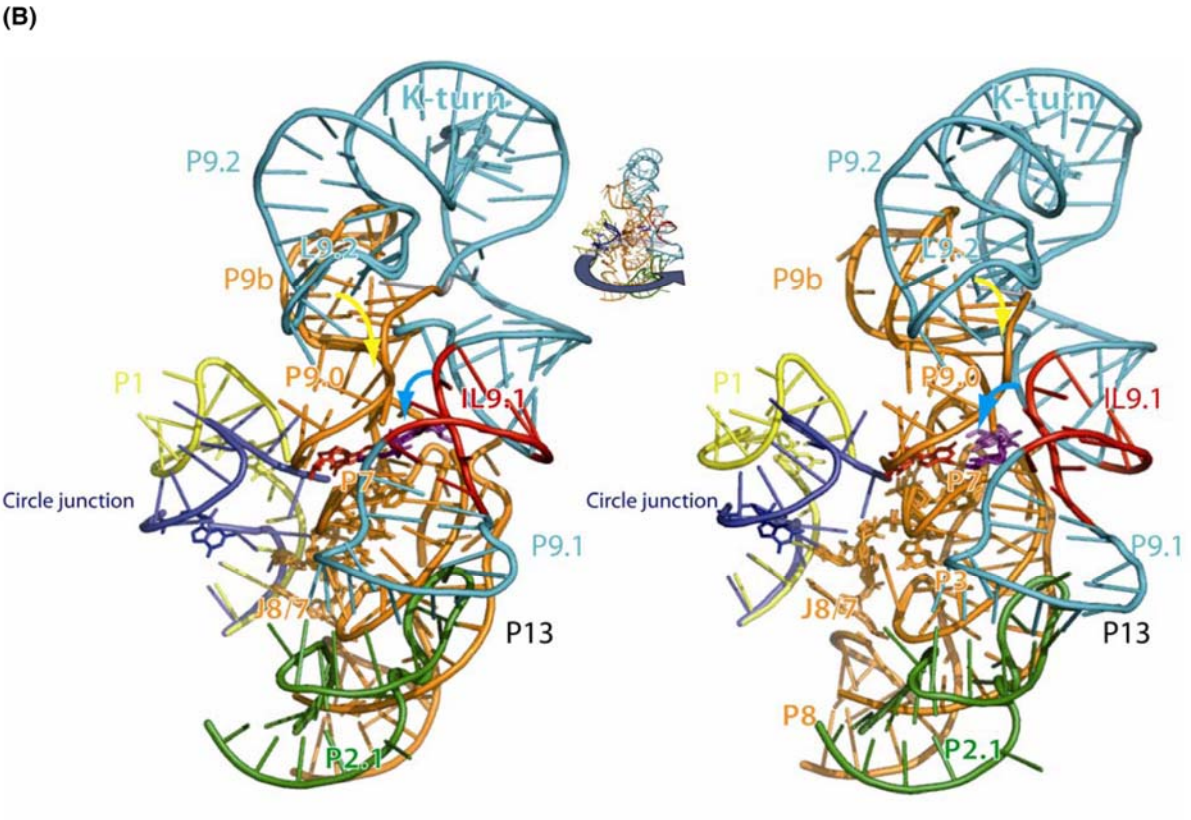


FIG. S5: The FLC and L-IVS DiGIR2 ribozyme 3D model overlay. **(A):** Ribbon nucleotides of the two models were provided by structural homology to the *Azoarcus* group I ribozyme. The structural homologous part to the *Azoarcus* group I ribozyme are depicted in orange (P3, P4, P5, P7, P9b). Nucleotides from the P2/P2.1 and P9.1/P9.2 appendages are depicted in green and cyan, respectively, while the nucleotide extensions of P5a, P6, and P8 specific to DiGIR2 are light gray. The remains of P1 present in the L-IVS are depicted in yellow, while the circle junction of the FLC is materialized in purple. **(B):** Zoom into the catalytic core of the DiGIR2 ribozyme. This overlay shows how the loss of the 5' exon promotes opening of P1 due to the lack of stabilization. Further, the formation of the FLC circle junction leads to a rearrangement of the nucleotides in the region. The loop L9.2 interacts with the J9.0/9a junction promotes and stabilizes the circle junction (yellow arrow). The internal loop in P9.1 (IL9.1 represented in red) stabilizes P7 in both L-IVS and FLC molecules (blue arrow).

References

Adams PL, Stahley MR, Gill ML, Kosek AB, Wang J, Strobel SA. 2004a. Crystal structure of a group I intron splicing intermediate. *RNA* **10:** 1867-1887

Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA. 2004b. Crystal structure of a self-splicing group I intron with both exons. *Nature* **430:** 45-50

Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* **273:** 1678-1685

Costa M, Michel F. 1997. Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution. *EMBO J* **16:** 3289-3302

Golden BL, Kim H, Chase E. 2005. Crystal structure of a phage Twort group I ribozyme-product complex. *Nat Struct Mol Biol* **12:** 82-89

Guo F, Gooding AR, Cech TR. 2004. Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site. *Mol Cell* **16:** 351-362

Haugen P, Andreassen M, Birgisdottir AB, Johansen S. 2004. Hydrolytic cleavage by a group I intron ribozyme is dependent on RNA structures not important for splicing. *Eur J Biochem* **271:** 1015-1024

Johansen S, Haugen P. 2001. A new nomenclature of group I introns in ribosomal DNA. *RNA* **7:** 935-936

Lehnert V, Jaeger L, Michel F, Westhof E. 1996. New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the Tetrahymena thermophila ribozyme. *Chem Biol* **3:** 993-1009

Lescoute A, Leontis NB, Massire C, Westhof E. 2005. Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res* **33:** 2395-2409

Li Z, Zhang Y. 2005. Predicting the secondary structures and tertiary interactions of 211 group I introns in IE subgroup. *Nucleic Acids Res* **33:** 2118-2128

Murphy FL, Cech TR. 1993. An independently folding domain of RNA tertiary structure within the Tetrahymena ribozyme. *Biochemistry* **32:** 5291-5300

Suh SO, Jones KG, Blackwell M. 1999. A Group I intron in the nuclear small subunit rRNA gene of Cryptendoxyla hypophloia, an ascomycetous fungus: evidence for a new major class of Group I introns. *J Mol Evol* **48:** 493-500

Vader A, Nielsen H, Johansen S. 1999. In vivo expression of the nucleolar group I intron-encoded I-dirI homing endonuclease involves the removal of a spliceosomal intron. *EMBO J* **18:** 1003-1013

Xiao M, Li T, Yuan X, Shang Y, Wang F, Chen S, Zhang Y. 2005. A peripheral element assembles the compact core structure essential for group I intron self-splicing. *Nucleic Acids Res* **33:** 4602-4611

CHAPTER V: SUPPLEMENTARY RESULTS

Studies done onto the DiGIR1 ribozyme have permitted to characterize its branching reaction, its secondary structure, its folding and regulation mechanism. Nevertheless, DiGIR1 is not only found in the myxomycete *Dydimium iridis*. Several GIR1s have been isolated from several Strains of *Naegleria* (hence named NaGIR1). As previously presented in chapter II, DiGIR1 and NaGIR1s present some prominent conservation of secondary structural features. Interestingly, NaGIR1s also show some differences with DiGIR1 and also between themselves. Thereby, in this chapter, a recent survey done on the NaGIR1s is presented. This work provides a new picture of NaGIR1 structural elements required for the branching reaction and the regulation mechanism involved in NaGIR1s. Finally, the screening for new natural GIR1 variants, retaining the branching activity, but also selected for their kinetic and folding properties, can be useful candidates for crystallization trials.

1. The GIR1 ribozymes from *Naegleria*:

1.1. Comparative analysis of *Naegleria* specific domain insertions/deletions:

Previous sequence analysis of rDNA from 70 rDNA isolates of *Naegleria* led to the observation that the Nae.S516 twin-ribozyme intron was conserved in 21 of the 70 strains (Wikmark et al., 2006). This study concluded to a vertical inheritance of the Nae.S516 intron within the *Naegleria* phylum (Wikmark et al., 2006). More recently, a new sequence alignment between DiGIR1 and NaGIR1s has been provided (Y. Tang and S. D. Johansen unpublished results). In the light of recent findings that have characterized the structural features required for DiGIR1 branching reaction, this new sequence comparison may give new clues to how the NaGIR1s can perform and regulate their activity.

The sequence comparison between NaGIR1 and DiGIR1 underlines some conservation of secondary structural features within the core of the ribozyme. It also highlights the insertion and/or deletion of domains (Figure 31). The insertions have been pointed out to be mainly localized in few positions: the L9 loop, the P6 stem and J5/4 junction. Interestingly, two of these insertion positions (the L9 loop and the J5/4) have been found to be critical for DiGIR1 activity. As previously demonstrated in **paper II**, the DiGIR1

L9 tetraloop has been shown to be involved in the release mechanism of the homing endonuclease mRNA after the branching reaction. Concerning the J5/4 insertion, molecular modelling of DiGIR1 together with mutagenesis data (**paper I and review 1**) have revealed the importance of this junction in the recognition of the G•U substrate domain. Interestingly in the NaGIR1 ribozymes, this junction is one of the most variable parts (Figure 31). It harbours either an internal loop or insertions of helical segments (see Chapter II).

The NaGIR1 peripheral domain P2 also does not show any similarity with the DiGIR1 peripheral domain P2/P2.1 (Figure 31). The DiGIR1 P2/P2.1 domain has been shown to regulate catalysis by adopting two mutually exclusive alternative conformations (HEG P1 stem-loop structure or P2 domain). Thus, the peripheral domain acts as an on/off switch that orchestrates the branching reaction of the ribozyme (**paper III**). In the case of NaGIR1, the peripheral domain is composed of the P2 domain poorly supported by comparative sequence analysis and a J2/10 internal loop which seems to be apparently unstructured (Jabri et al., 1997; Jabri and Cech, 1998).

Thus, the differences observed between the NaGIR1 and the DiGIR1 raise several questions: Do the various NaGIR1s still retain their branching catalytic activity? What is their current rate of cleavage in comparison with DiGIR1 ribozyme? Is there an impact of the flanking sequence length on the catalytic activity of the NaGIR1 ribozymes? Do these ribozymes adopt 3D structures similar to DiGIR1 despite their rather large extensions in J5/4 and P6? To answer these questions, an original overview of the *Naegleria* GIR1 phylogenetic distribution based on the sequence alignment is provided (Figure 32). From this, NaGIR1s representative of each clade are selected and assayed for the branching activity. Then, a best ribozyme candidate is selected for further characterization. This candidate is further used to carry out folding studies and mutational analysis in order to better understand the role of the flanking sequences on catalysis regulation. Finally, general conclusions will be drawn from these experimental results.

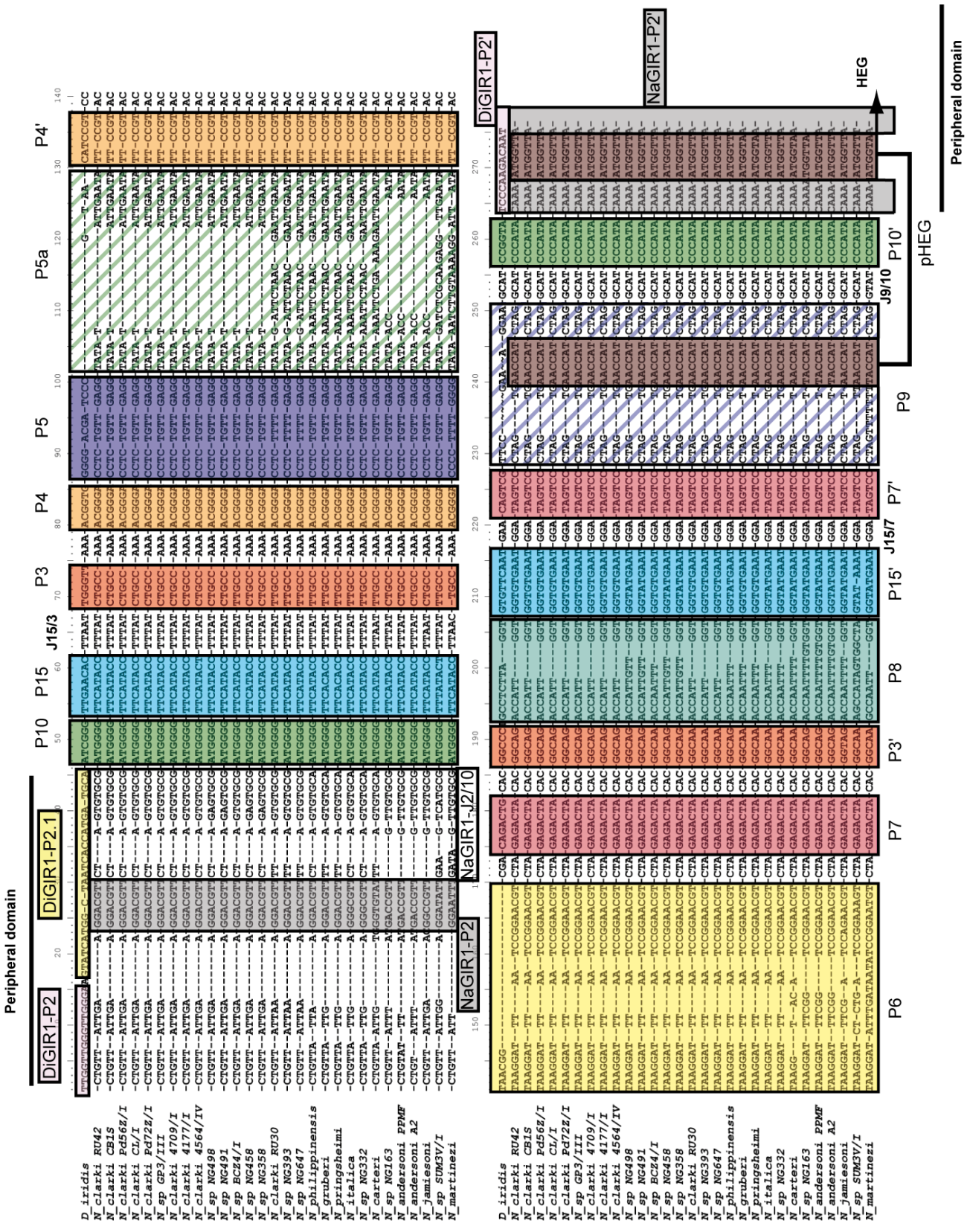


Figure 31
Sequence alignment of NaGIR1s and DiGIR1

1.2. Screening for branching activity in NaGR1s and selection of the NprGIR1:

Sequence analysis of GIR1 core region from 29 natural isolates of *Naegleria*, was performed to gain insight into the genetic relationship among the various strains. From this sequence analysis, a phylogenetic tree, based on the Neighbour Joining method, has been built (see Figure 32, Y. Tang S. D. Johansen, H. Nielsen; unpublished results). The NaGIR1 ribozymes appear to form 6 clusters. Cluster 1 is the most populated with 10 of the 29 strains. Close inspection of the other clusters reveals that they are mainly defined by the sequence insertion in J5/4 (Figure 31).

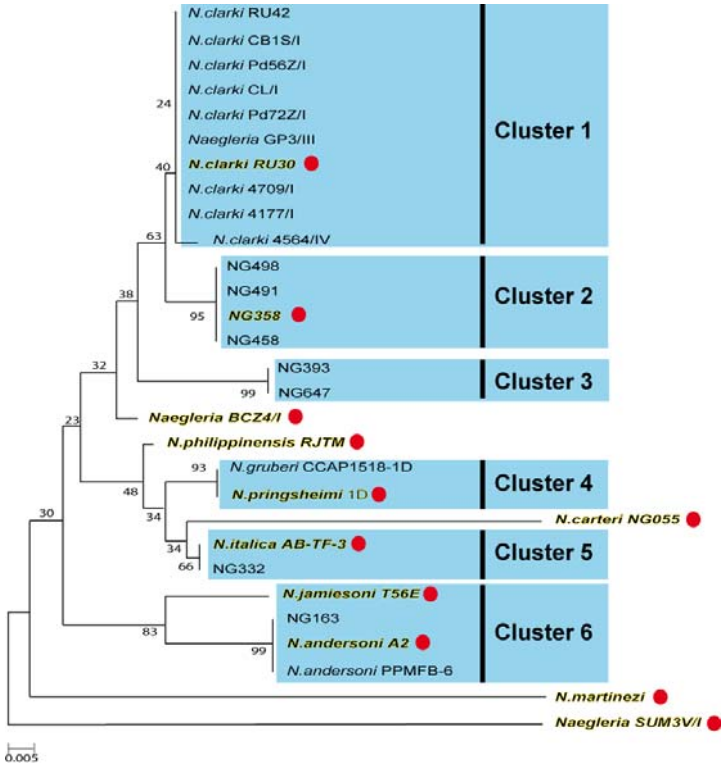


Figure 32
Phylogenetic tree of the natural occurring NaGIR1 ribozymes.

The phylogenetic tree has been built based on NaGIR1s sequence alignment and the NJ method. Interestingly, the NaGIR1s form 6 main clusters (cluster 1-6 in blue). 11 NaGIR1s (red dot) were selected based on the phylogenetic relationship and distribution. (Y. Tang S. D. Johansen, H. Nielsen; unpublished results)

According to their phylogenetic distribution (Figure 32), eleven NaGIR1s have been selected to determine their cleavage reaction catalytic rates and their ability to perform the branching reaction. However, it has been previously shown that the branching activity of DiGIR1 was dependent on the length of the flanking sequences (i.e. 157.22 is the DiGIR1 smallest length variant that retains the branching activity (Nielsen et al., 2005)). Based on this

observation, the length of the flanking sequences in the NaGIR1s may also have an impact on branching. Thereby, preliminary studies done on the *Naegleria andersoni A2* (NanGIR1) have revealed that the minimal length variant required for the branching reaction was 178.28 (i.e. 28 nt after the IPS; Nielsen unpublished result). In this way and in order to compare the 11 different NaGIR1s between each other, all the NaGIR1 clones were constructed according to the NanGIR1 studies all harbouring +28 nucleotides after the IPS in their 3' flanking sequence.

Time-course cleavage reaction experiments of the 11 smallest length variant based on the NanGIR1 studies were then carried out. From these experiments, the various NaGIR1s have been classified according to their cleavage rates (Figure 33 A). Thus, it leads to the observation that the *Naegleria pringsheimi* (NprGIR1) GIR1 variant cleaves with the highest rate (Figure 33 A). However kinetic analysis does not allow to determine whether the ribozymes cleave by hydrolysis or by branching. Further primer extensions were thus performed (Figure 33 B).

Primer extension analysis allows to map and to quantify the 5'end of RNA. The lariat is detected by a premature elongation stop corresponding to the branch point (BP). This method discriminates cleavage by branching from cleavage by hydrolysis that results in a longer primer extension product due to reverse transcription up to the IPS (+3 nt) (Figure 33 B). The NaGIR1 3' cleavage products have been analysed by this method. Data show that most NaGIR1s exhibit branching activity but some of them also cleave by hydrolysis (Figure 33 B). Interestingly, close comparison of the kinetic cleavage rates reveals that NprGIR1 performs branching activity *in vitro* at a higher rate than DiGIR1 (DiGIR1 157.22 K_{obs}: 0.0211, End point: 0.34 (Nielsen et al., 2009); NprGIR1 Kobs: 0.0693, End point: 0.16). Finally, the NprGIR1 was selected as a model system for NaGIR1s for further analysis to better understand the role of the flanking sequences and of the structural elements required for the branching reaction.

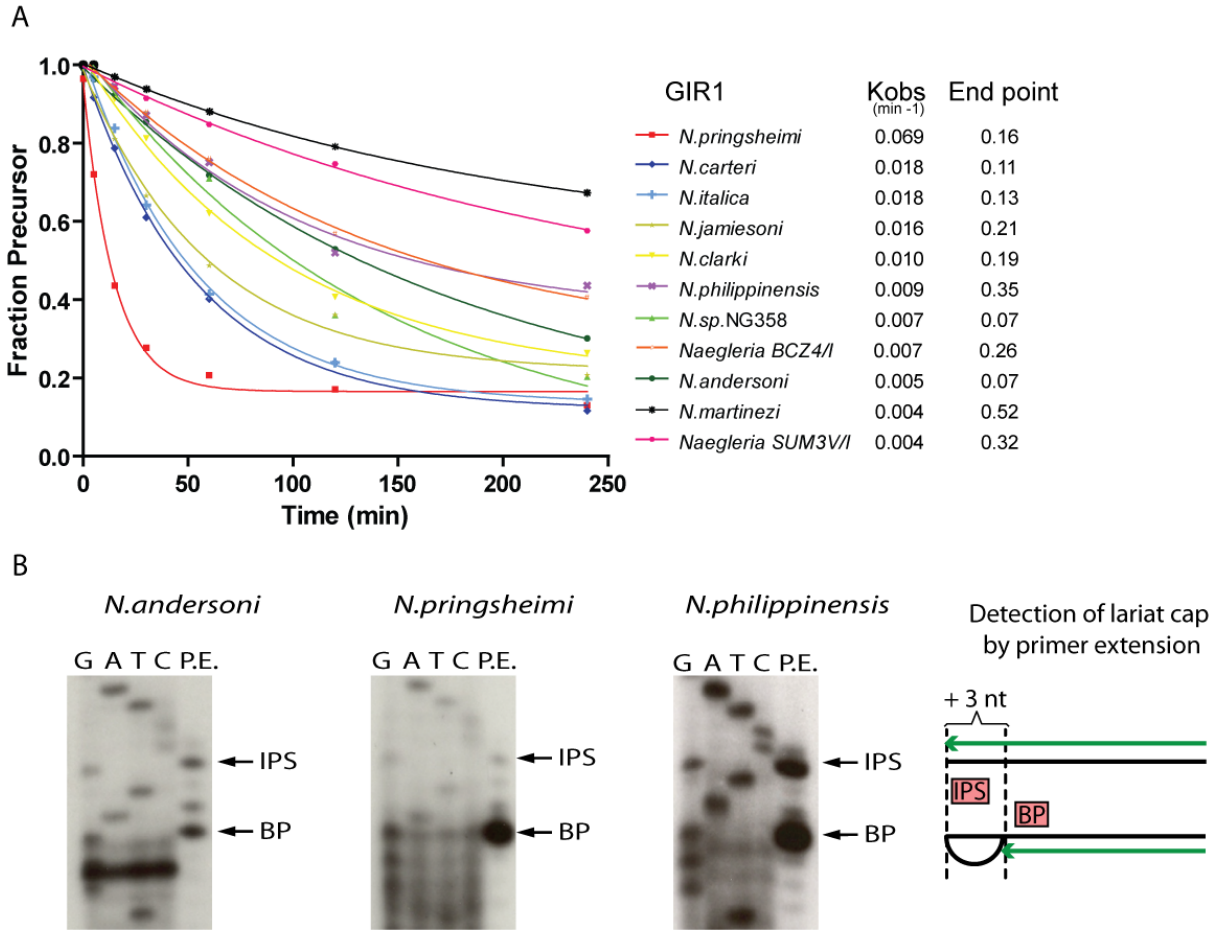


Figure 33
Kinetic cleavage of the 11 NaGIR1s and primer extension analysis.

(A) NaGIR1s were transcribed and subjected to a time-course cleavage experiment directly followed by denaturing gel analyzed at different time points (0-250 min) (Y. Tang, S. D. Johansen, H. Nielsen kinetic cleavage analysis, unpublished results). The fraction of uncleaved precursor was then deduced from the gel analysis and plotted. The cleavage rates of the various NaGIR1s were then deduced. (B) Primer extension analysis on the 3'product of 3 different NaGIR1 variants. The branch point (BP) is detected by a primer extension pausing signal whereas the hydrolytic cleavage at the IPS site results in a primer extension stop signal due to reverse transcription up to the IPS (+3 nt) (see B schematic drawing).

1.3. Study of NprGIR1:

1.3.1. Prediction of two mutually exclusive alternative secondary structures:

Based on sequence alignment and *in silico* predictions, alternative base pairing schemes have been proposed to take place within the ribozyme flanking sequences. These alternative secondary structures involve either the formation of the P2 stem or the formation

of a pseudoknot with P9 (named P9/pHEG) between the terminal part of the 3'end flanking sequence and the L9 loop (Figure 34 A). As seen in the NaGIR1s sequence alignment (Figure 31, Figure 34) the L9 loop varies from 7 to 13 nt. However the last 5 L9 terminal nucleotides (5'-ACCAU-3') are fully conserved (Figure 34 B). Finally, the second half of the pseudoknot located at the 5'end of the HE mRNA (i.e. 5'-AUGGU-3') is also fully conserved Figure 34 B).

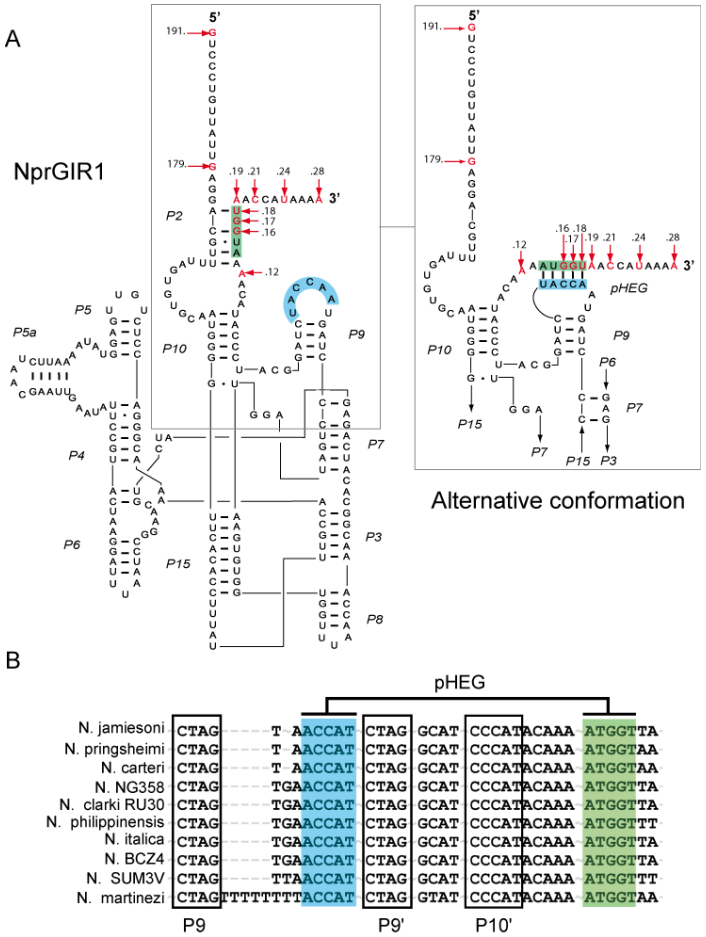


Figure 34
NprGIR1 alternative structure prediction.

(A) Secondary diagram of the NprGIR1 with the various deletion studies plotted and the alternative base pairing prediction. (B) Sequence alignment of the 11 selected NaGIR1 showing the 3'end.

Previously, we have demonstrated that the DiGIR1 flanking sequences were involved in the formation of two mutually exclusive secondary structures depending on the 3'flanking sequence length. The first one implies the formation of the HEG P1 stem that was shown to destabilize the catalytic core. The second one implies first the melting of HEG P1 and then the formation of the P10 and P2 domains which activate the DiGIR1 ribozyme (**Paper III**). Moreover, HEG P1 was also found to promote post-cleavage release of the lariat capped

mRNA (**Paper II**) (Nielsen et al., 2009). Then one could wonder what are the active and inactive conformations of NprGIR1. A solution can be suggested from comparing these two systems. The formation of P2 may promote the formation of the catalytic core whereas the formation of pHEG may promote the inactive folding of the ribozyme or vice versa. Finally, the P9/pHEG pseudoknot could also be involved in post-cleavage release of the product in a still unknown mechanism, albeit different from DiGIR1 (**Paper II**). In order to understand the role of the 3'flanking sequence a deletion study has been performed directly followed by a folding study of the various NprGIR1 length variants. Finally mutational analysis of the P2 stem and of the L9 loop was carried out in order to structurally dissect the ribozyme active and inactive conformations.

1.3.2. Gradual 3'flanking sequence deletion induces a shift from branching to hydrolysis

In order to understand the role of the 3'flanking sequence from the NprGIR1 ribozyme, a systematic deletion study has been performed. The 5'end has been arbitrarily fixed to 191 nt whereas the 3'end has been deleted step by step from 28 nt to 12 nt after the IPS (Figure 34). The primer extension method used to discriminate between cleavage by branching or by hydrolysis cannot be applied in the case of products smaller than 24 nt. This is mainly due to the impossibility to use a primer shorter than 16-18 mer oligonucleotides to bind specifically to the 3'product RNA for the reverse transcription reaction. Thus, a new strategy to detect lariat cap RNA generated by the NprGIR1 ribozyme based on the RNase R assay was developed (Y. Tang, S. D. Johansen, H. Nielsen, RNase R assays, unpublished results) (Figure 35). The RNase R exoribonuclease digests essentially all linear RNAs but does not digest lariat or circular RNA (Vincent and Deutscher, 2006; Suzuki et al., 2006). Thus, most RNAs from a given sample can be completely digested (e.g. all products from the hydrolytic cleavage reaction) as well as lariats 3' tails, only leaving the 3 nt ± 1 nt lariat cap in the case of GIR1 (Figure 35).

In order to compare the branching rates of NprGIR1 and of DiGIR1, the RNase R method was benchmarked using two different DiGIR1 length variants that were selected according to their ability to perform either the branching reaction (i.e. DiGIR1 157.22) or the hydrolytic reaction (i.e. DiGIR1 166.22) (Nielsen et al., 2005)) (Figure 35). The NprGIR1 deletion studies in combination with kinetic cleavage experiments (data not shown) and the

RNase R assays, have permitted to isolate the 191.17 length variant as the minimal length variant that fully retains branching activity (more than 76%) (Figure 35). Moreover, depending of the 3'end length, a clear shift from branching to hydrolysis activity has been observed. This shift appears when the 3'end is reduced to less than 17 nt (191.16: 32% of branching while 191.17: 76% of branching) (Figure 35). This demonstrates that the 3'end flanking sequence length is important for the branching reaction by NprGIR1. Interestingly, these finding are consistent with the previous observation on NanGIR1 (Henrik Nielsen; unpublished data) (Jabri et al., 1997) and the DiGIR1 ribozyme (Einvik et al., 2000; Nielsen et al., 2005; Nielsen et al., 2009). However these deletion studies do not tell whether P2 or pHEG is involved in the ribozyme active conformation. Thus, in NprGIR1 and in NaGIR1s in general, the role of flanking sequence remains unclear at this stage and needs to be experimentally adressed.

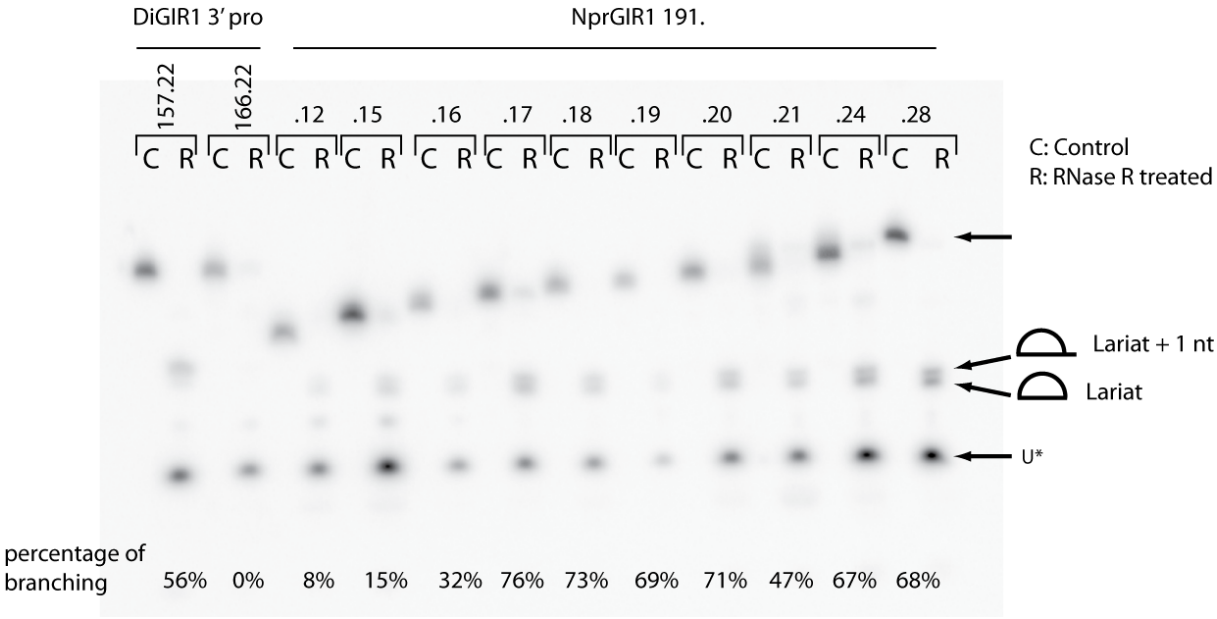


Figure 35
Rnase R experiments

The different length variants were transcribed with α-UTP and subjected to cleavage reaction. The 3' product was then gel purified and then digested by using RNase R enzyme. The digestion product was then analysed by denaturing polyacrylamide gel. The efficiencies of cleavage by branching versus by hydrolyse were then compared (Y. Tang, S. D. Johansen, H. Nielsen, unpublished results)

1.3.3. Impact on folding of 3'flanking sequence deletion

The folding of the NprGIR1 different length variants used in deletion studies was monitored by native polyacrylamide gel electrophoresis. It has been previously shown that this method can be efficiently used to resolve RNA folding conformers of various group I introns including the *Tetrahymena* ribozyme (Emerick and Woodson, 1994; Pan and Woodson, 1998), the *Azoarcus* ribozyme (Rangan et al., 2003; Rangan et al., 2004; Chauhan et al., 2009) and DiGIR1 (157.22 and 166.65) depending on the Mg^{2+} concentration (**Paper III**). Subsequently, the fraction of native ribozymes as a function of Mg^{2+} concentration can be determined and fitted to the Hill equation in order to determine the midpoint of folding transition (Rangan et al., 2003). Eight different NprGIR1 length variants have been analysed by native gel electrophoresis following the protocol described in **Paper III**.

The longest length variant conformers (191.28) were first monitored on native gel. At low Mg^{2+} concentration, the ribozymes form a diffuse band as in the case of the *Azoarcus* ribozyme (Figure 36). This band most likely represents the unfolded state (U) of the ribozyme consistent with the lack of any catalytic activity at such a low Mg^{2+} concentration. At higher Mg^{2+} concentration, the ribozymes migrate as a focused band (N) representing a near-active or active conformation (Figure 36). This correlates with the fact that this ribozyme shows a burst of activity after a pH jump. The midpoint of folding transition, determined at $Cm_{191.28}$=6.3 mM, was found to be cooperative with respect to Mg^{2+} concentration together with a Hill coefficient of $\eta_{191.28}$=1.5 (Figure 37). Interestingly, with a $Cm_{157.22}$= 1.1 mM for the DiGIR1 minimal version (157.22) or $Cm_{166.65}$=4.3mM for the DiGIR1 longer variant (166.65) (**Paper III**), the DiGIR1 ribozyme seems to require less Mg^{2+} ions to reach its near-active conformation. However, the NpGIR1 seems to be less prone to form stable alternative structures that can be separated on native gel. This NprGIR1 folding behaviour is to be compared with the DiGIR1 folding behaviour in which the HEG P1 structure present in the 166.65 length variant was previously shown and isolated by native gel assays (Figure 36) (**Paper III**). Finally, both ribozymes require around 25 mM Mg^{2+} for full activity suggesting additional folding that is not revealed by native gels.

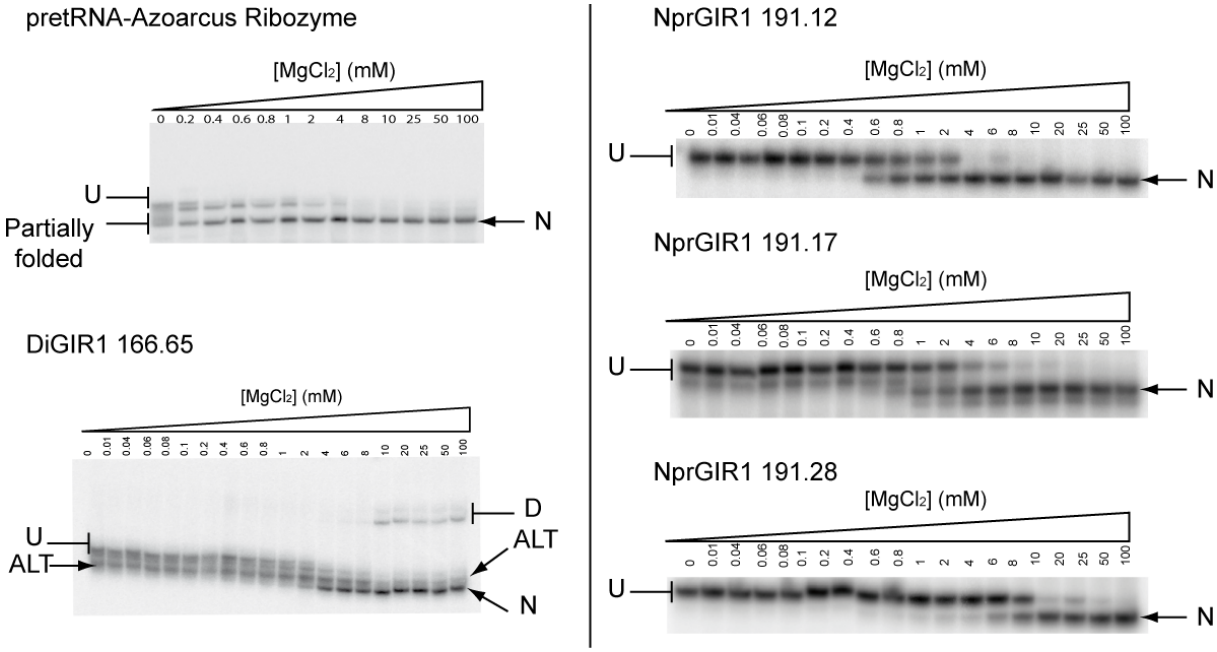


Figure 36
Monitoring of the Azoarcus, DiGIR1 and NprGIR1 folding by native gel.

Next, we decided to monitor the folding of the various length variants harbouring nucleotide deletions at their 3'end (Figure 36 NprGIR1 191.17 and 191.12). Native gel analyses show that the smallest length variant seems to require less Mg^{2+} to reach its native conformation (N) in comparison with the longest length variant (Figure 36 NprGIR1 191.28). After fitting to the Hill equation, this general trend has been confirmed (Figure 37). The midpoints of folding transition of these two length variants were respectively found at $Cm_{191.12}$=0.64 mM and $Cm_{191.17}$=0.67 mM in comparison with the longest length variant that has a $Cm_{191.28}$=6.3 mM. Thus, these results highlight that either the formation of P2 or pHEG in long length variants (i.e. 191.18 to 191.28) requires increasing amount of Mg^{2+} in order to be stabilized. Interestingly, the NprGIR1s that do not harbour full length 3'flanking sequences, require less Mg^{2+} to fold in there native conformation although they do not retain branching activity. In this way, the 3'flanking strand seems to induce branching through folding of the ribozyme catalytic core.

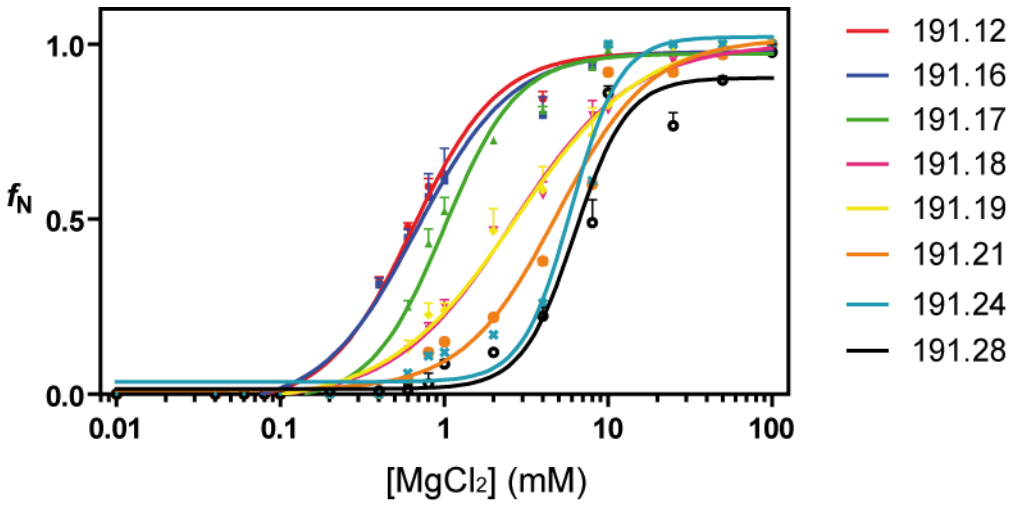


Figure 37
Summary of native gel analysis of the various length variant of the NprGIR1.

The fraction of native RNA (f_N) was determined and fitted to the Hill equation. As a result the midpoint of folding transition for each length variant was determined: $Cm_{191.12}$ =0.64 mM, $Cm_{191.16}$ =0.67 mM, $Cm_{191.17}$ =1 mM, $Cm_{191.18}$ =2.60 mM, $Cm_{191.19}$ =2.72 mM $Cm_{191.21}$ =4.86 mM, $Cm_{191.24}$ =5.93 mM, $Cm_{191.28}$ =6.25 mM. The length variants appear to be clustered in 3 different groups according to their folding properties. The first group is mainly composed of length variants with a 3'end shorter than 17 nt. Remarkably, all these variants were previously characterized to cleave only by hydrolysis (i.e. 191.16 and 191.12). The second group, a transition group, is composed of length variants with 3' ends comprised between 18 to 21 nt (191.18, 191.19 and 191.21). Finally, the third group is composed of length variants with a 3' end longer than 21 nt (191.24, 191.28). These two last groups were shown to cleave mainly by branching (Figure 35).

In summary, *in vitro* folding of NprGIR1 is a portioning between unfolded and near-active states. Native gel assays show that long length variants of NprGIR1 do not promote formation of stable alternative structures that can be separated on native gel. In this respect, the behaviour of NprGIR1 is different from DiGIR1 in which an alternative structure can be isolated (i.e. DiGIR1 166.65 Figure 36 ALT band) (**Paper III**). Finally, the midpoint folding transition, determined for each length variants, reveals that increasing amounts of Mg^{2+} are required when the length of 3'flanking sequences is increased. This observation also correlates with the possibility to form either P2 or either pHEG in longer length variants. The formation of pHEG or P2 seems to require higher Mg^{2+} concentrations to be stabilized. Thus, the 3'flanking sequences seem to directly promote correct folding of the ribozyme catalytic core as proved by the observation that long length variants carry out branching efficiently. However these folding assays cannot prove definitely whether P2 or pHEG is formed in the active conformation.

1.3.4. pHEG: the active conformation revealed by mutagenesis

In order to refine the secondary structure diagram and furthermore to understand the role of P2 and pHEG, we have decided to design a series of mutants that should either stabilize or destabilize pHEG or P2. All mutants were subjected to time-course experiments and the 3' products were analyzed by primer extension.

1.3.4.1. Deletion of 13 nt in the 5' flanking sequence has no impact on branching:

We have started by pruning the 5' flanking sequence from 191 to 179 (deletion of 13 nt) (Figure 38). The 179.28 NprGIR1 new length variant retains branching activity. Its cleavage rate is close to the cleavage rate of the 191.28 NprGIR1 length variant (Figure 38 Kobs and end point). Thus, this first finding highlights that the 13 deleted nucleotides seem to be not involved in the formation of any local secondary structure or either in the stabilization of the catalytic core. This experiment allowed us to fix the length of the 5' end to 179 nt before the IPS while the 3' end was kept to 28 nt after the IPS. This new length variant, NprGIR1 179.28, is now considered the new reference for the NprGIR1 and named accordingly WT NprGIR1 in the following section. All mutants are compared to this minimal WT NprGIR1.

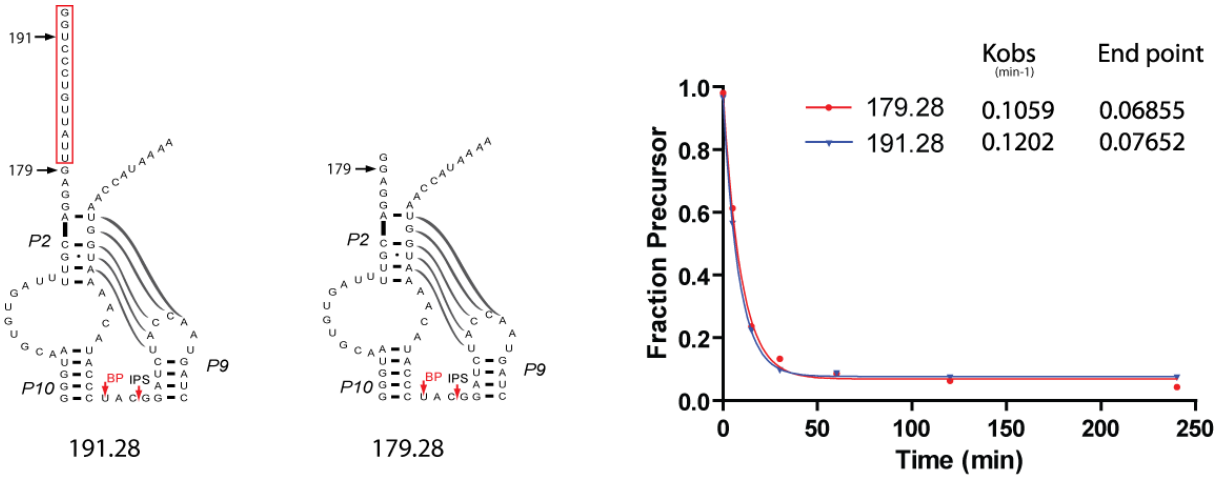


Figure 38
The 13 nt deletion in the 5'end of the NprGIR1 has no impact on catalysis

The two different NprGIR1 length variants (191.28 and 179.28) were transcribed and subjected to a time-course cleavage experiment (at different time points 0-250 min) directly followed by denaturing gel analyzed. The fraction of uncleaved precursor was then deduced

from the gel analysis and plotted. The cleavage rate of two length variants was then deduced. Primer extension reveals that the NpGIR1 179.28 still retains its branching activity (data not shown).

1.3.4.2. Stabilization of P2 shifts the activity toward hydrolytic reaction:

In a first step, we wanted to study the impact on catalysis of stabilising P2 either by introducing point mutations in both the 5' and 3' strands or either by creating an elongated stable P2 stem. To achieve this goal, several mutants were constructed (MutEP2, MutP2'P2") (Figure 39). The results show that the different mutants that harbour a more stable P2 cleave with a slower rate than the WT NprGIR1 (Figure 39, kobs and End point). Moreover, primer extensions of the 3' cleavage products of the two mutants show that the ribozymes mainly cleave by hydrolysis (data not shown). Altogether these results highlight the fact that a stable P2 does not promote branching activity in the NprGIR1 but shifts the cleavage activity toward the hydrolytic reaction.

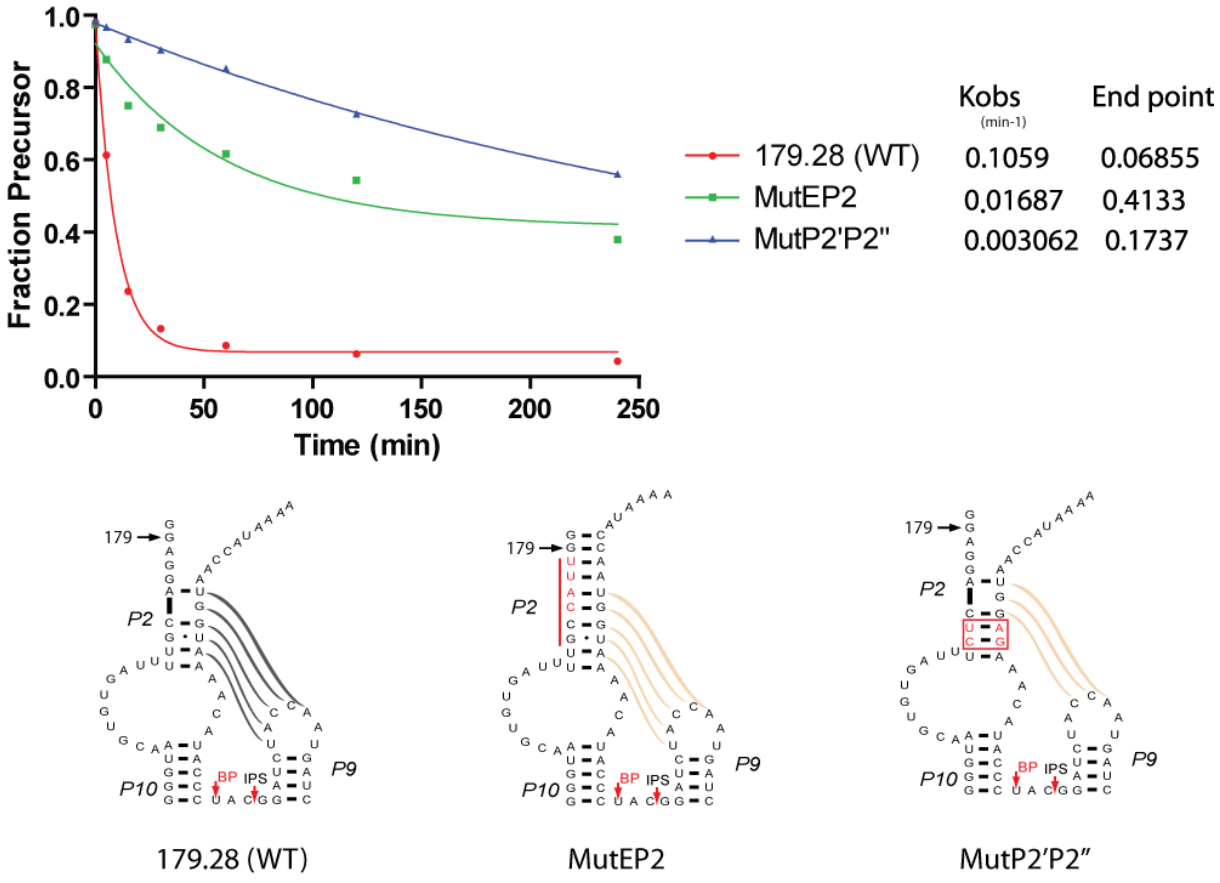


Figure 39
Stabilization of the P2 domain impacts the cleavage rate of the NprGIR1 ribozyme

1.3.4.3. Disruption of pHEG reduce the activity of the NprGIR1:

In a second step, we have constructed mutants to destabilize the pHEG pseudoknot by introducing double mutations in the L9 loop and the P2 3'strand (MutL9, MutP2") or in L9 and the P2 5'strand (MutP2'L9) (Figure 40). Kinetic cleavage analysis in combination with primer extensions reveal that mutated ribozymes cleave at a very slow rate (Figure 40 kobs and End point) and mainly by hydrolysis. Altogether these results underline the fact that the pHEG pseudoknot is essential for branching activity in NprGIR1.

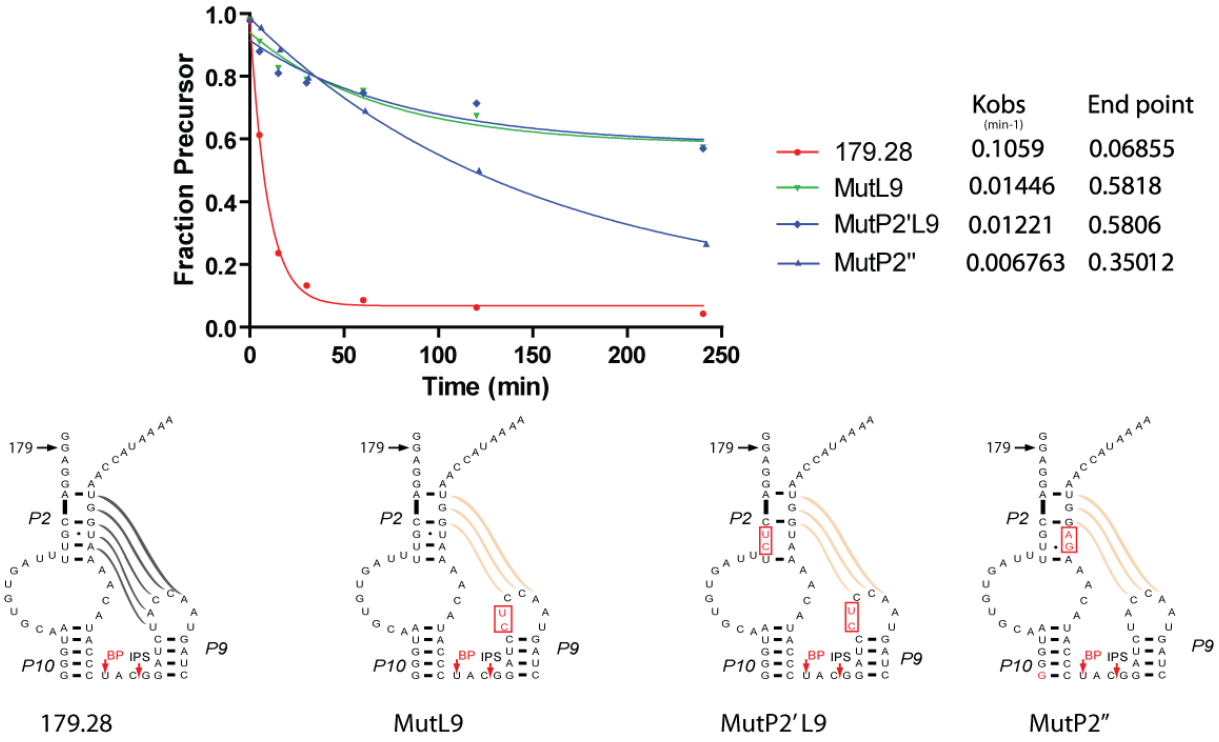


Figure 40
Mutations of the L9 loop or the pHEG formation impair the ribozyme branching activity.

1.3.4.4. Compensatory mutations, come back to a functional ribozyme:

Finally, mutants designed to either stabilize pHEG or restore a WT-like phenotype, have been created (Figure 41 MutP2' and MutP2"L9). Interestingly, the mutant that stabilizes the pseudoknot by destabilizing the P2 domain (single mutant MutP2') has been shown to cleave at the same rate as the WT NprGIR1 (Figure 41). Moreover it has been shown to retain the branching reaction highlighting that pHEG is required for the branching to occur. Next we

decided to monitor the impact of a double mutation in P2 3' strand and L9 loop. Despite the fact that the double mutant (MutP2"L9) has been shown to retain the branching activity, it cleaves with a lower rate than the WT (Figure 41). This finding is consistent the fact that the pHEG is required but it also highlights that the mutations of the L9 loop and the P2 3' strand have an impact on the cleavage rate. Thus, it is possible that the mutations induce either an alternative structure or a less stable pHEG pseudoknot. Finally, a triple mutant has been designed to restore a WT-like ribozyme (MutP2'P2"L9). Interestingly, this triple mutant restores the catalytic activity of the previously tested double mutant (MutP2'P2" kobs=0.003 versus MutP2'P2"L9 kobs=0.010 Figure 39 and Figure 41). This observation emphasizes that pHEG is required for the branching to occur. Despite the fact that this triple mutant has been shown to retain the branching activity, it cleaves with a lower than the WT or the MutP2"L9 double mutant (Figure 41). Thereby, this triple mutant does not restore the WT cleavage rate. Thus, it is possible that the folding of P2 domain has been favoured and stabilized while pHEG has been destabilized.

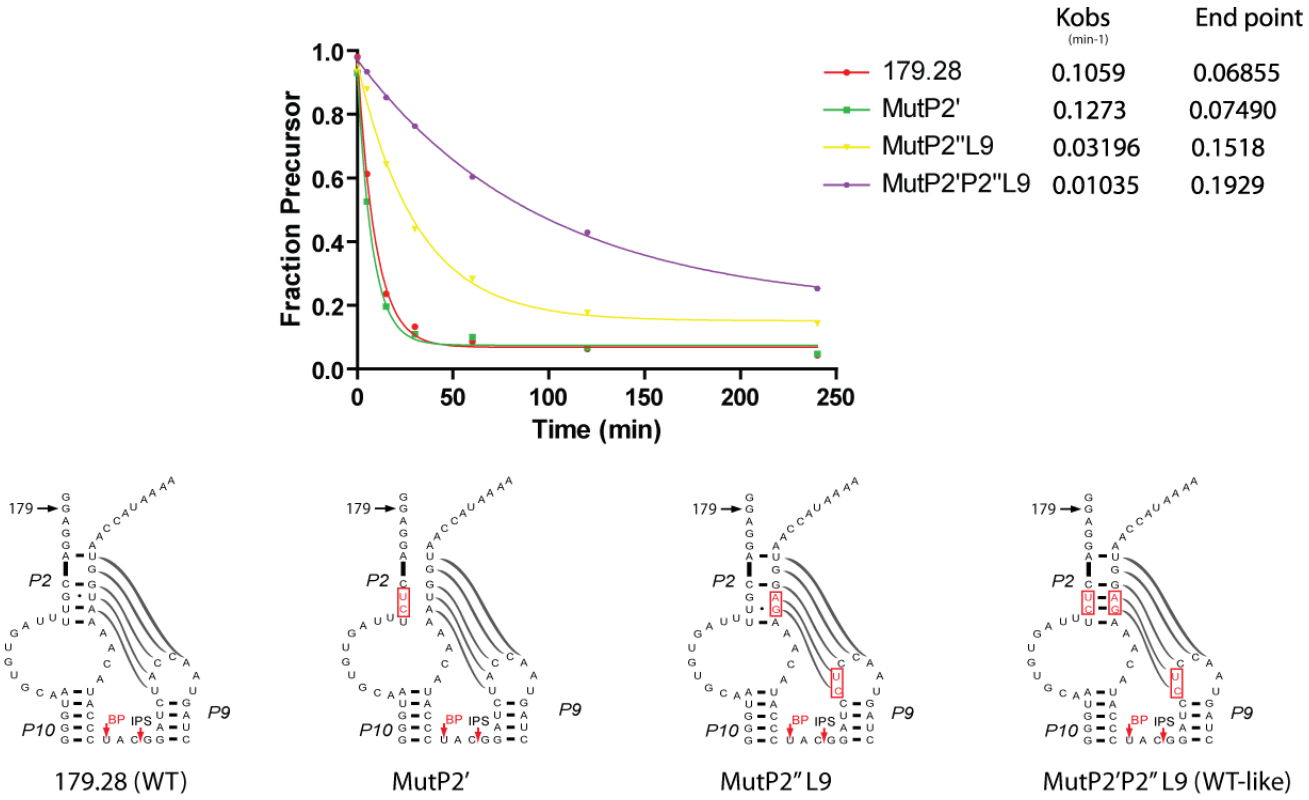


Figure 41
Stabilization of the pHEG pseudoknot and creation of a WT-like NprGIR1

The overall picture that emerges from all these mutational data underline that pHEG formation is required to reach the active conformation of the ribozyme. Moreover, the

stabilisation of P2 promotes the inactivation of the ribozyme by preventing the formation of the pHEG pseudoknot. Thus, the formation of pHEG seems to have an impact on the stabilization of the catalytic core of the ribozyme. Interestingly, from these mutational data it appears straight forward that pHEG has a major implication in the active conformation of NprGIR1 and thus in NaGIR1s due to its sequence conservation. In the opposite, the role of the J2/10 junction still remains unclear and is currently under investigation.

1.3.5. Tertiary interactions revealed by the Fe-EDTA structure probing:

We have previously used the Fe-EDTA probing method on the DiGIR1 ribozyme in order to better understand the spatial organisation of the catalytic core and also the role and the location of the P2P2.1 regulatory domain with respect to the ribozyme core (**Paper III**). In order to understand the role of the peripheral domains from NprGIR1 (i.e. the J5/4 with its P5a extension, P6 extension, the L9 loop, the J2/10) and to compare these two ribozymes together, we performed Fe-EDTA probing on the NprGIR1 191.28 (Figure 42) according to the protocol described in the **Paper III**.

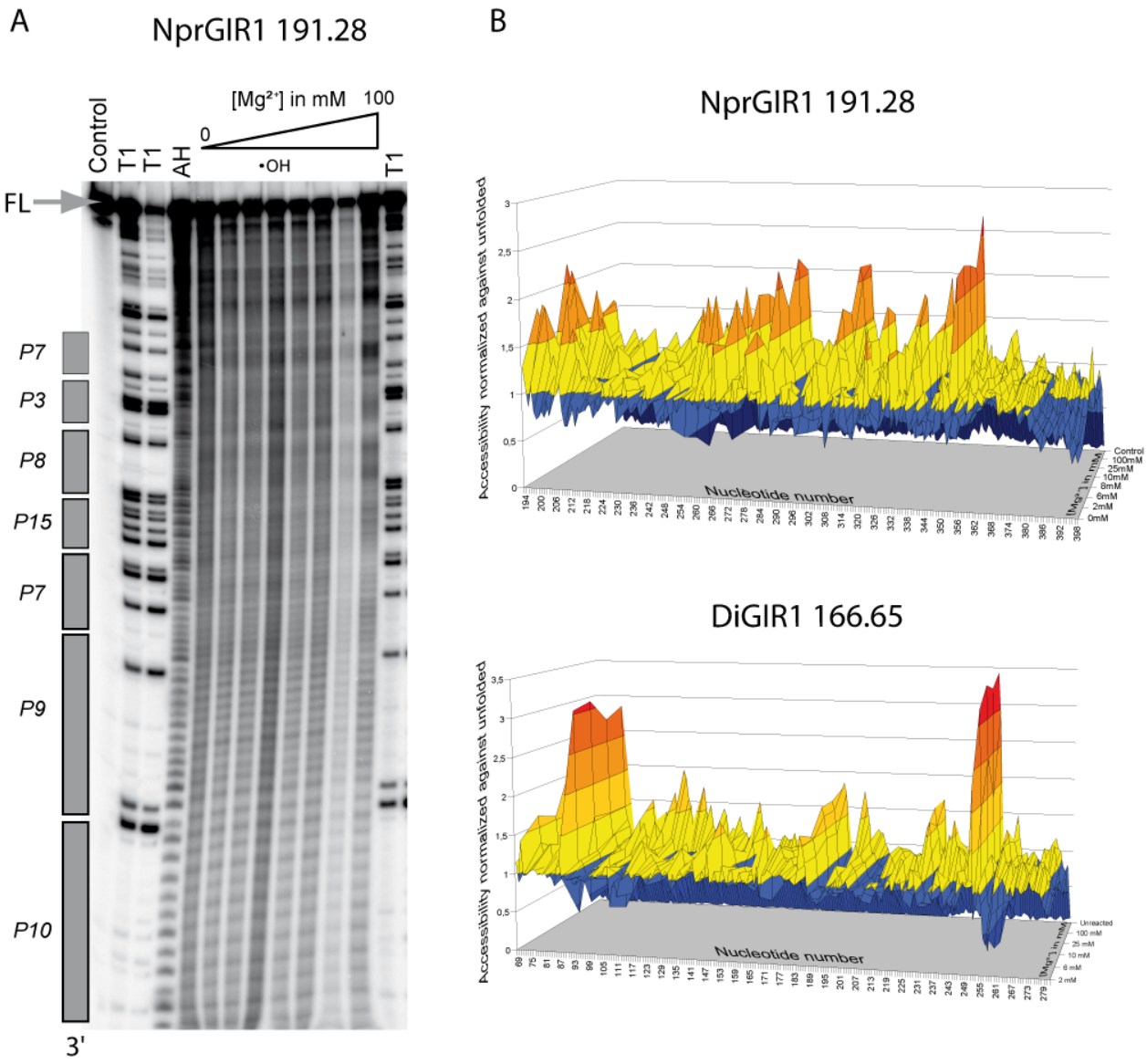


Figure 42
Fe-EDTA structure probing of the NprGIR1 ribozyme

(A) 15% denaturing polyacrylamide gel after NprGIR1 191.28 Fe-EDTA structure probing using 3'end ^{32}P labelled length variant. The gels obtained were then exposed to phosphorImager image plates (Molecular Dynamics) and quantified using SAFA software (Laederach et al., 2008) (B) Graphical view of reactivities in solution of the two ribozymes (NprGIR1 191.28 and DiGIR1 166.65) upon Fe-EDTA treatment. Data were normalized against the band corresponding to the unfolded ribozyme.

At low Mg^{2+} concentration, the core of the ribozyme, including all junctions, is highly accessible to the Fe-EDTA probe (Figure 43 A and B). This indicates an open conformation of the ribozyme core. This observation is consistent with results from native gel assays and with results obtained with the DiGIR1 Fe-EDTA probing at the same Mg^{2+} concentration. Along the same line of evidence, most NprGIR1 parts which are different from DiGIR1 are also highly accessible.

With increasing amounts of Mg^{2+} ions, the ribozyme becomes more compact. This observation corroborated by the emergence of a focused band in the native gel assays, is expressed in the Fe-EDTA probing experiments by a signal decrease for half of the residues (Figure 43 A and B). The residues found in junctions (J9/10, J15/7, J5a/4), internal loop (P6) and loop L9 become less accessible. Most residues from peripheral loops pointing away from the core (L5, L5a and L6) become more accessible which is also consistent with the behaviour of DiGIR1.

Based on results from Fe-EDTA probing experiments, in combination with the previous structure probing done on NanGIR1 (Jabri et al., 1997) and the potential structural homology between the DiGIR1 and NprGIR1, some tertiary interactions taking place within the ribozyme core can be hypothesized. As a first example, decrease of Fe-EDTA signal shows that the P6 internal loop becomes protected which could mean it is involved in a tertiary interaction. According to the DiGIR1 3D model, this internal loop can be proposed to interact with a receptor located in P3 that simultaneously presents a decrease of its reactivity to the probe (Figure 43). In this way, the potential tertiary interaction between the P3 and P6 internal loops mimics the situation observed in DiGIR1 ribozyme with the L6 loop and its receptor located in P3 (**Paper III**). Along the same line of evidence, the three-way junction (3WJ) between P5, P5a and P4 also becomes protected with the increasing amount of Mg^{2+} ions (Figure 43). At this stage the 3WJ classification (Lescoute and Westhof, 2006) left several conformer candidates that need to be further experimentally tested to identify the correct one. However, by combining Fe-EDTA data with the structural homology of the two ribozymes (i.e. P4 and P5 are stacked and J5/4 participates in the recognition of the P10 substrate G•U base pair in the DiGIR1 ribozyme (**Paper I**)) a solution can be proposed. The decrease of reactivity to Fe-EDTA observed in the NprGIR1 experiments in the region around the J5a/4 junction and the 3'strand of P4 (Figure 43 C), indicates that P5a could adopt a parallel orientation or fold over P4. Thus, J5a/4 might participate in the recognition of the G•U base pair of P10. Family C is the only one that could satisfy these biochemical and structural constraints (Lescoute and Westhof, 2006) (Figure 43 C). Molecular modelling of NprGIR1 will be carried out in order to explore this assumption.

Two regions of NprGIR1 belonging to P2, J2/10 and L9 remain structurally unclear after interpretation of Fe-EDTA data. Interestingly, these zones were previously proposed to be involved in the formation of alternative structures either P2 or pHEG.

In summary, the Fe-EDTA experiments performed on NprGIR1 in combination with native gel assays and previous data related to DiGIR1, show that the core of the ribozyme becomes more compact with increasing amount of Mg^{2+} ions. Based on the structural homology between DiGIR1 and NprGIR1, several tertiary interactions can be proposed for regions J9/10, J15/7 and J15/3. For insertions specific to NprGIR1, the P6 internal loop and J5a/4, structure probing together with the structural constraints required to form the lariat fold required for the branch point will guide us to propose a model of NprGIR1. Moreover, molecular modelling may help us to resolve the locations and orientations of the P2 domain, the J2/10 junction and the L9 loop which are not clarified by Fe-EDTA data.

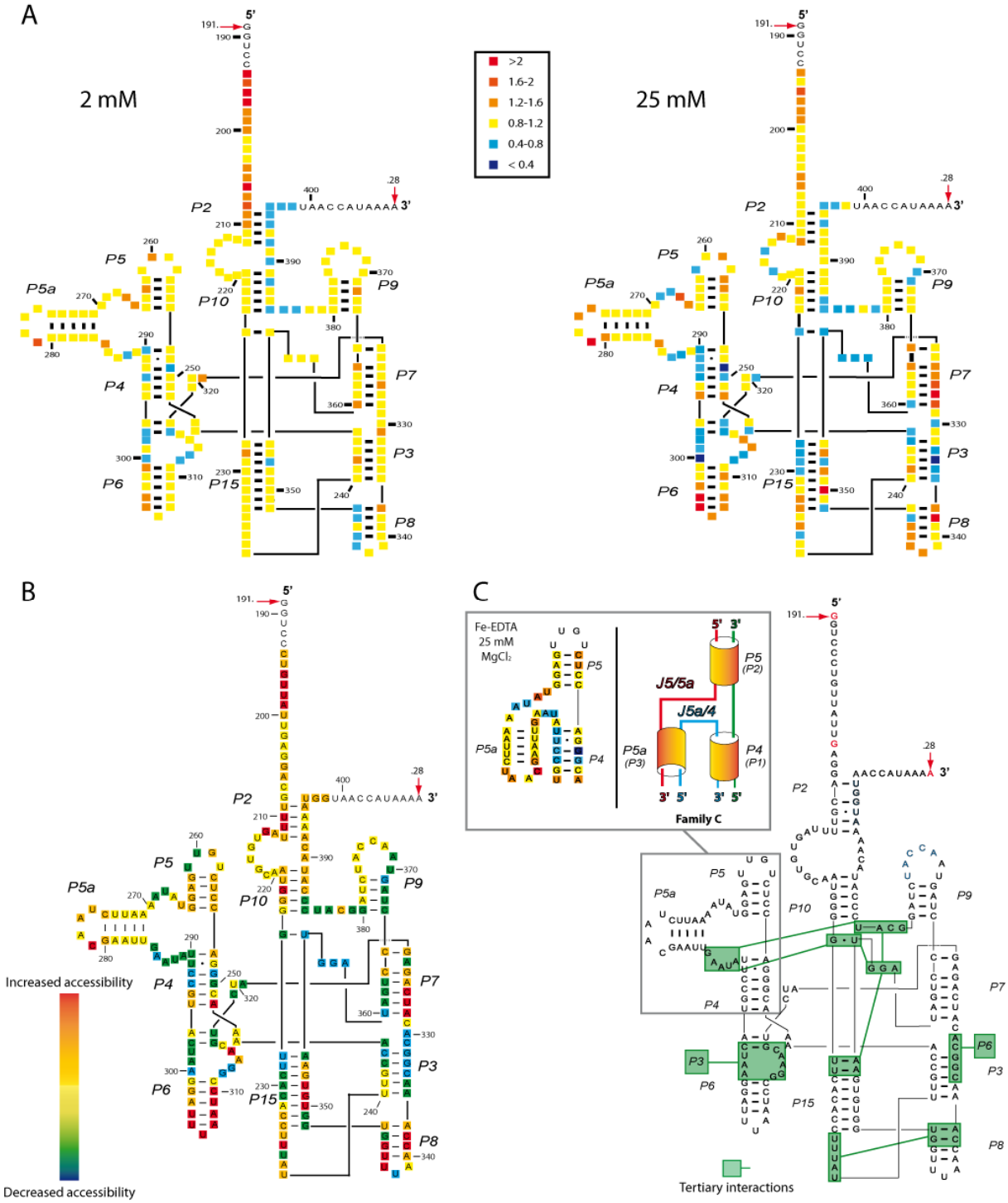


Figure 43
Fe-EDTA structure probing summary.

(A) The hydroxyl radical reactivity at two selected Mg^{2+} concentrations (2 mM and 25 mM). Data were plotted on the secondary structure diagram representing the inactive conformation, harbouring the P2 domain. (B) Global overview of the hydroxyl radical reactivity showing the residues with increased (from yellow to red) and decreased (from yellow to blue) reactivity upon increase of the Mg^{2+} concentration. (C) Predicted tertiary interactions in the core of the ribozyme.

1.4. NaGIR1 studies: conclusion and perspectives:

Screening the different *Naegleria* GIR1 ribozymes for branching activity has led to identify the NaGIR1 variant with the best catalytic rates: the NprGIR1 ribozyme. Interestingly, NprGIR1 appears to have better branching capabilities than DiGIR1 (DiGIR1 157.22 Kobs: 0.0211, End point: 0.34 (Nielsen et al., 2009); NprGIR1 Kobs: 0.0693, End point: 0.16). Deletion studies performed on NprGIR1 by a combination of deletion studies and mutational analysis helps to better understand the role of the flanking sequences and the involvement of the pHEG pseudoknot in the catalytically active conformation of *Naegleria* GIR1s.

The folding studies have also revealed that the longest length variant of NprGIR1 (191.28) requires a higher Mg^{2+} concentration than the shortest one (191.12). It is clear from the secondary structure diagrams of these two length variants that this effect is due to the formation of the pseudoknot and maybe also to the likely melting of the P2 stem. Thus, it can be speculated that there is a folding order in NprGIR1 that could be applied to the NaGIR1s in general since their secondary structures are highly similar. First, the P2 domain has to fold in order to regulate the *in vivo* activity of the ribozyme by a different mechanism from the one found in DiGIR1 (**Paper III**) (Figure 44). In the same time it also gives enough flexibility to the catalytic core to reach its active conformation. Second, P2 melting will create the pseudoknot pHEG that stabilizes and locks the catalytic core of the ribozyme in its active configuration (Figure 44). Thus, the formation of pHEG may participate in the folding of the ribozyme active catalytic core by orientating correctly the branch nucleotide in order to allow catalysis. Furthermore, pHEG might also be involved in the release of the 3'product from the catalytic core by a different mechanism from the one observed in DiGIR1. Even if the release mechanism is different in these two ribozymes, it leads to the same important goal: the expression of the Homing endonuclease after the branching reaction (Figure 44).

The current role of the J2/10 junction still remains unclear and is currently under investigation. However, based on mutational, and Fe-EDTA probing data a molecular model can be built to better understand the role of the various insertions found in the catalytic core of the NaGIR1s ribozyme. Finally, monitoring the catalytic properties of the NaGIR1s enriches the pool of ribozymes that can be crystallised.

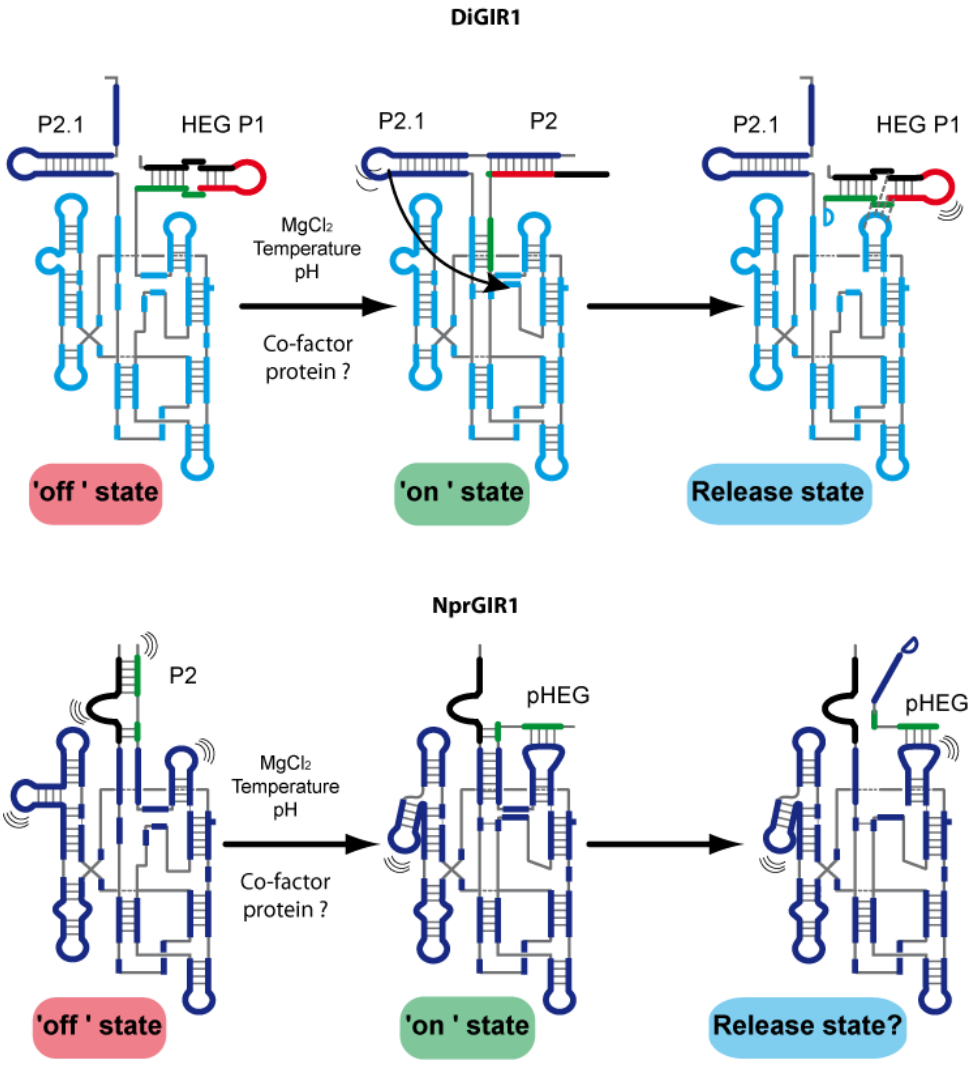


Figure 44
Comparison between "on/off" state in DiGIR1 and NaGIR1

2. The structure of the DiGIR1 ribozyme, crystallization assays:

DiGIR1 performs branching right after folding under its active conformation. Crystallization of DiGIR1 thus requires catalysis inactivation by incorporation of chemical modifications without preventing the formation of the native structure of the catalytic core. In order to easily incorporate those, a strategy based on truncated ribozymes has been designed inspired by the crystallization strategy of the *Azoarcus* ribozyme (Adams et al., 2004a; Adams et al., 2004b). These strategies will be briefly described in the following sections as well as the results of the first crystallization trials.

2.1. Example of the *Azoarcus* crystallization strategy:

The crystal structure of the ribozyme naturally found within the pre-tRNA^{Ile} anticodon loop of the purple bacterium *Azoarcus sp* BH72 was obtained by Adams *et al.* (Adams et al., 2004a; Adams et al., 2004b) in the splicing intermediate prior to the exon ligation reaction. In this crystal structure, the 5'exon has been cleaved but still remains base-paired to P1 while the 3'exon is covalently connected to the intron by forming the P10 helix (Adams et al., 2004b) (Figure 45). In order to obtain crystals in this trapped state, the ribozyme was extensively engineered. The 5'exon was trimmed down to 3 nucleotides complementary to the IGS (Figure 45). The ribozyme was also truncated in the L9 loop which was then reconstructed using an oligoribonucleotide in *trans,* complementary to P9 and to the first 6 nucleotides of P10 forming the 3'exon. Most importantly, in order to lock the ribozyme in this trapped state, introduction of 2'-deoxy substitutions at several points was necessary. Finally a U1A binding site was engineered in P6a and the ribozyme was co-crystallized with the RNA binding protein U1A (Rupert and Ferre-D'Amare, 2001). Thus, a crystal structure of the active ribozyme with both 5'and 3' exons was solved at 3.1 Å of resolution (pdb accession code: 1U6B/1ZZN (Adams et al., 2004a; Adams et al., 2004b)) (Figure 45). A similar strategy was applied for crystallising DiGIR1.

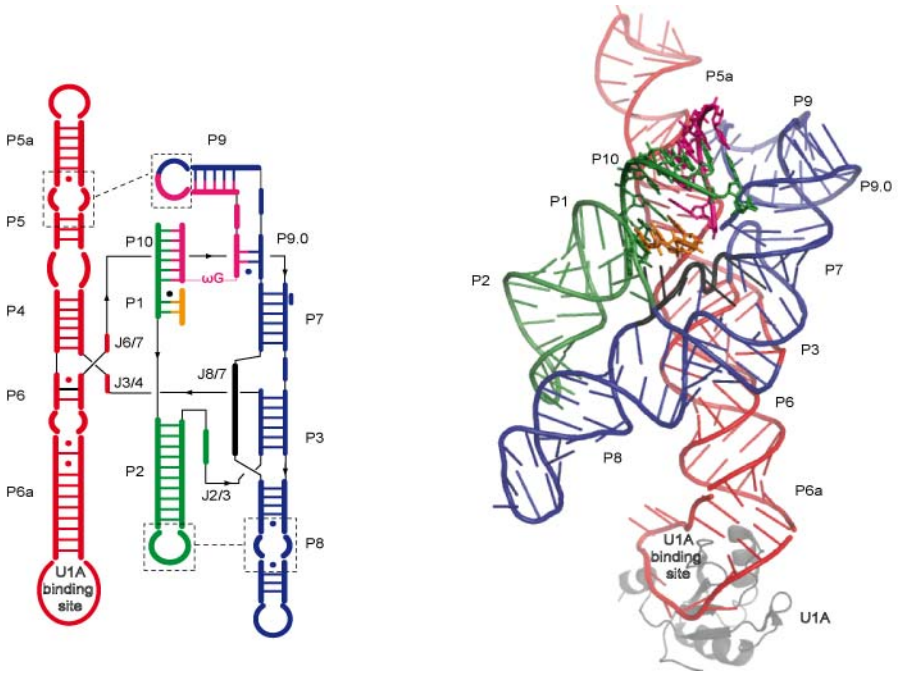


Figure 45
Schematic secondary structure and tertiary structure *Azoarcus* group I ribozyme

2.2. Construction of truncated DiGIR1 and insertion of the U1A site:

Due to its ability to perform branching upon folding, inactivation was carried out by truncating DiGIR1 after the second nucleotide of L9. Three kinds of ribozymes were then designed according to the secondary structure elements contained in their 5' flanking sequence. The wild-type (WT) contains the full P2/P2.1 5' sequence stretch. The ΔP2 constructs start at nucleotide 18 and the ΔP2/P2.1 starts in P10 (Figure 49). The DiGIR1 missing part is restored by an oligoribonucleotide complementary to the 5' strands of P9, P10 and P2 (in the WT constructs). All ribozymes remain partially unfolded until the substrate corresponding to different RNA oligomers, is added in *trans* to reform P9, P10 and P2 (Figure 46). In order to obtain DiGIR1 truncated, precisely in the middle of L9, the Hepatitis Delta Virus ribozyme (HDV) has been added immediately downstream A222 to cleave during the *in vitro* transcription process (Figure 46).

The crystallization of a homogenous RNA population is by essence a difficult task. Interestingly, several successful examples of co-crystallizations ribozymes with the RNA binding protein U1A have been recently reported in the literature (Adams et al., 2004b; Ferre-D'Amare and Doudna, 2000; Ferre-D'Amare and Rupert, 2002; Ferre-D'Amare, 2010; Rupert and Ferre-D'Amare, 2001; Rupert et al., 2003). Based upon these results, a U1A binding site was engineered at the end of P8 of the three kinds of constructs described above. Note that P8 was chosen because it seems to be quite neutral both in terms of folding and catalysis in DiGIR1 (Figure 46). Moreover, additional structural diversity was generated by varying the length of P8 fused to the U1A binding site.

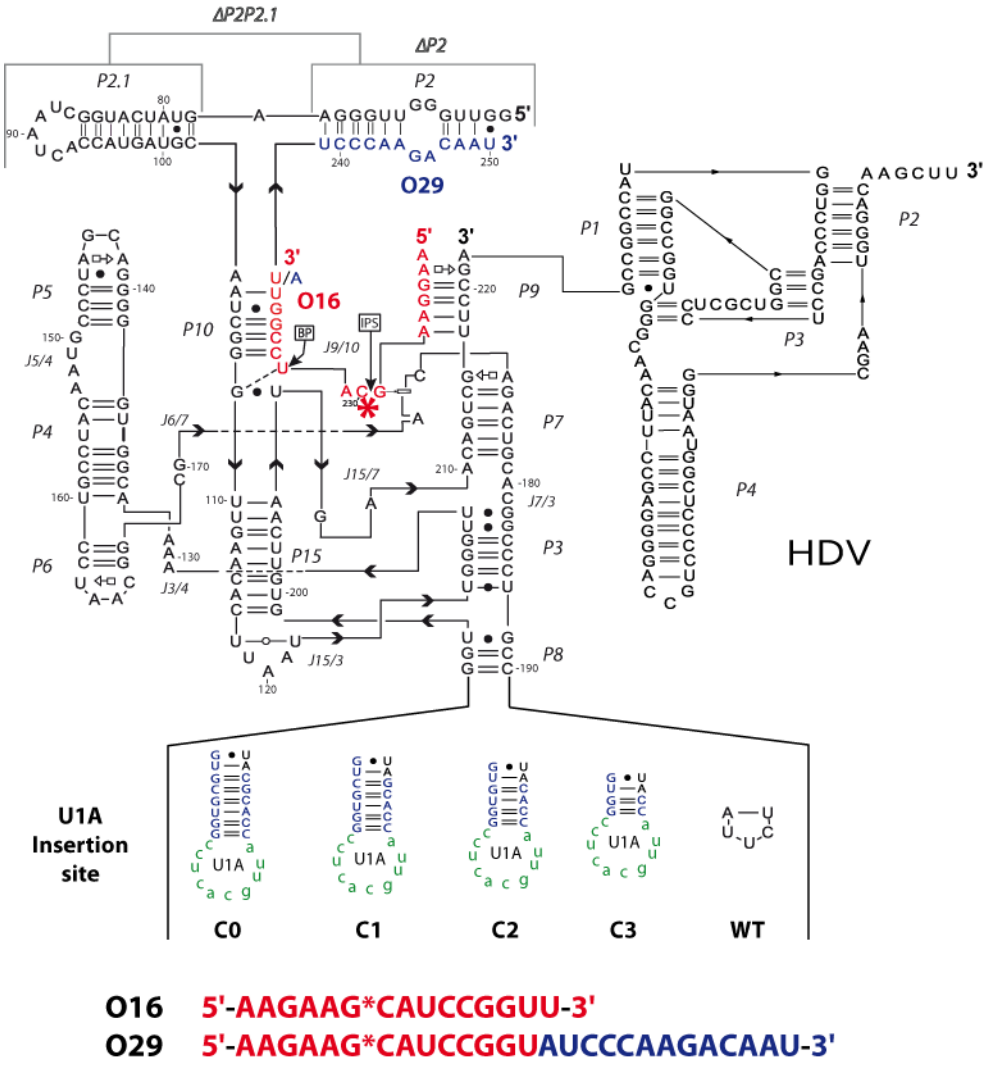


Figure 46
Engineering of truncated DiGIR1 ribozyme

Presentation of the P9-truncated DiGIR1s fused to the HDV ribozyme and containing the U1A insertion site in P8.

2.3. Test of the HDV activity in the various construction:

Before starting the large-scale transcription in order to produce truncated DiGIR1 ribozyme for the crystallization assays, small-scale transcriptions were set up to test the activity of the HDV ribozyme fused to the various DiGIR1 truncated ribozyme construction (Figure 47). As expected, the quantity of truncated-DiGIR1 released after cleavage by HDV was found to be about 85% after quantification of the bands (precursor, truncated-DiGIR1, HDV) during the time-course transcription assays. Then, the truncated ribozyme can be purified by either polyacrylamide denaturing gel electrophoresis or gel exclusion

chromatography on a Superdex® 200 10/300 GL (small scale) or HiLoad 26/60 Superdex® (large scale) (Amersham Pharmacia Biotech column).

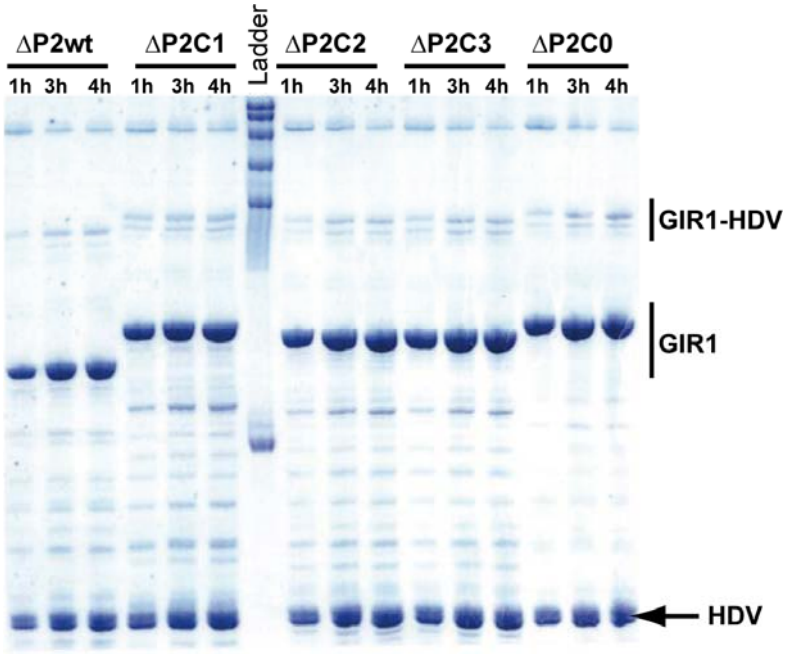


Figure 47
Time-course T7 polymerase *in vitro* transcriptional assays.

Time-course of T7 polymerase *in vitro* transcription assays of the various constructions on run 8% denaturing polyacrylamide gel. Major bands respectively correspond from the top to the bottom to the precursor (GIR1-HDV), the truncated DiGIR1 (GIR1) and finally the cleaved HDV ribozyme from the top to the bottom.

2.4. Preparation of ternary complex containing the truncated-DiGIR1, the substrate and U1A protein:

2.4.1. Formation of binary complex DiGIR1-Oligo:

After purification, truncated-DiGIR1s need to adapt to the two distinct RNA substrates (Figure 46). The first oligomer, composed of 16 nucleotides (hence named O16) was designed to form P9 and P10. The second oligomer, the 29-mer (hence named O29), was designed to form P9, P10 and P2. Oligonucleotides were synthesized by Dharmacon® with or without chemical modification in order to prevent catalysis or to check for ribozyme activity, respectively. Chemical modifications were incorporated in residues directly involved in catalysis. rU232 was replaced by dU232 to prevent the nucleophilic attack leading to the formation of the lariat (i.e. O29d11: O29 harbors a deoxyribose substitution at position 11). A

phosphorothioate substitution (*) was incorporated at the scissile bond of G229 (i.e. O16*8, O29*8). After purification of the different oligonucleotides, band shift assays were performed in order to check the formation of the binary complex.

Band shift assays were done by mixing increasing concentrations (0, 100 nM, 1 µM, 25 µM) of unlabelled truncated-DiGIR1 with a constant concentration (1 nM) of a 5' end labelled oligoribonucleotide. In order to follow the formation of the binary complex, samples were then subjected to a non-denaturing polyacrylamide gel electrophoresis (protocol from **Paper III**) (Figure 48). Band-shift assays show that the lengths of the oligonucleotides (O16 or O29) have a different impact on the formation of the binary complex (Figure 48). Interestingly, O16 did not bind the truncated-DiGIR1 under the conditions used whereas O29 did. Thus, O29 came out as the best candidate to form binary complexes with the WT GIR1 or the ΔP2 series (Figure 48). Dissociation constants (Kd) for the complexes were determined in the range of 1 to 3µM.

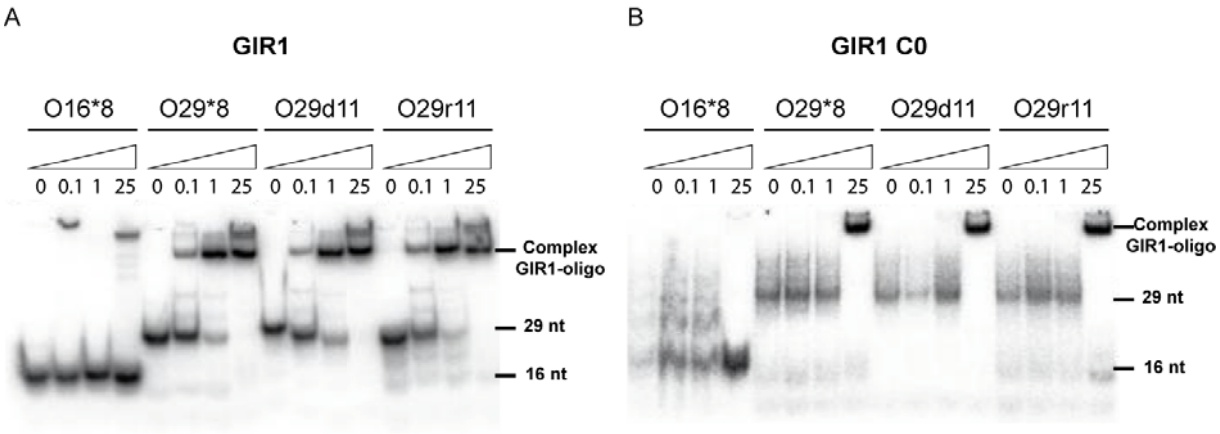


Figure 48
Band shift assays and formation of binary complex DiGIR1-Oligo.

The band shift assays (Mira Tawk, Masters student 2009-2010, unpublished results) revealed the possibility to form a binary complex composed of the truncated-DiGIR1 ribozyme and the substrate added in *trans*. The concentration of truncated-DiGIR1 ribozyme was increased from 0 to 25 µM while the concentration of 5' end labeled oligoribonucleotide was fixed to 1 nM. (A) Truncated DiGIR1 ribozyme. (B) Truncated DiGIR1 ribozyme with the U1A site inserted in P8.

Interestingly, the WT GIR1 ribozyme was able to perform cleavage of the unmodified substrate O29 (data not shown). Nevertheless, since reverse transcription tests were not performed, it was not assessed whether the ribozyme cleaves predominantly by branching or

hydrolysis. As expected, the phosphorothioate-modified oligoribonucleotide was shown to be marginally cleaved by the ribozyme (data not shown).

2.4.2. Formation of the ternary complex between DiGIR1, the RNA substrate and U1A protein:

In order to form the ternary complex (DiGIR1-Oligo-U1A), the double mutant (Y31H, Q36R) of the U1A RNA binding domain (RDB) was first purified (strain and protocol from Pr Kyoshi Nagai MRC, Cambridge, UK and Dr. A. Ferré d'Amaré, Fred Hutchinson Cancer Research Center, Seattle, WA, USA). The ability of the U1A RDB to bind to the binary complex was assessed by band-shift assays (Figure 49). Two concentrations of a particular truncated-DiGIR1 construct harbouring the U1A binding site were chosen (10 µM and 20 µM). The concentration of 5'end labelled oligonucleotide was also fixed to 1 nM. Finally, the concentration of the added protein was increased from 0 to 40µM. Upon addition of the U1A protein a mobility shift resulting from the formation of the ternary complex was observed on native gels (Figure 49).

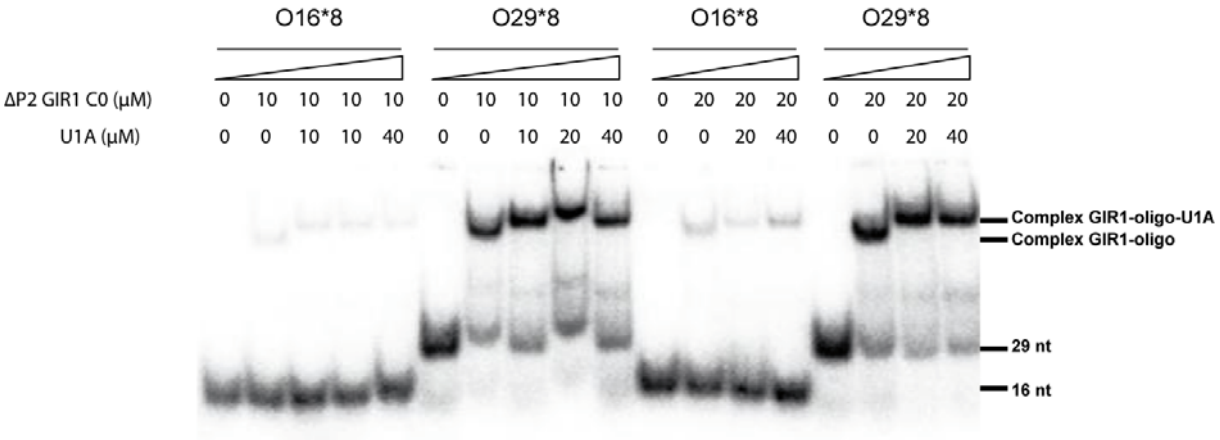


Figure 49
Band-shift assays revealing the formation of ternary complex.

Band shift assays reveal formation of a binary complex composed of the truncated-DiGIR1 ribozyme, the RNA substrate and the U1A protein that binds to its site located in P8.

2.5. Screening for crystallization condition:

Since ΔP2 constructs were not catalytically active and ΔP2P2.1 could not bind the RNA substrate, only samples containing different concentrations of the WT GIR1 series in

complex with the RNA substrate (DiGIR1-Oligo) or the ternary complex (DiGIR1-Oligo-U1A) were tested for crystallization. Crystallisation trays were set up using commercially available high-throughput (HT) screening suites (MPD suite and Nucleix Suite from Qiagen or Index and PEG-Rx from Hampton Research). Crystallization trays were set up using a mosquito robot (TTP Labtech), stored at constant temperature (22°C), and observed weekly using a microscope AX10 Zeiss.

For most constructs, only precipitates were obtained. The only construct that apparently yields microcrystalline precipitates is formed by the WTGIR1C2, the O29*8 substrate and the U1A RBD (See Figure 51). Microcrystals were observed under few conditions of the Index HT and PEG HT screens from Hampton Research after one week of equilibration. In order to improve the homogeneity of the samples, they were subjected to HPLC purification prior to set up the crystallisation trays. The column used was a superdex 200 PC 3.2/30 run on an Akta system (Pharmacia). This column allows for loading concentrated small volume samples that are slightly diluted during the purification process to the right concentration (between 50 to 100 μM) and that can be used for crystallisation right after elution. Strikingly, this additional purification step significantly increased the number of conditions under which precipitates apparently containing microcrystals could be obtained (See **Table 3**).

Crystal quality will be improved by varying physico-chemical factors (temperature, pH) as well as modifying the concentrations of crystallants composing the conditions yielding the best results. More engineering of the RNA will also be carried out aiming at integrating distinct motifs capable of promoting intermolecular interactions capable of improving crystal packing (Ferré-d'Amaré 1998 JMB).

Table 3: Increase of the number of positive crystallisation hits following purification of the ternary complex by HPLC. Conditions identified initially are indicated in italic bold face. HPLC purification of the complex strikingly increases the number and the diversity of conditions under which the complex forms comparable microcrystalline precipitates.

Index HT	WTGIR1C2/O29*8/U1A
C6	0.1 M Sodium chloride, 0.1 M BIS-TRIS pH 6.5, 1.5 M Ammonium sulfate
D6	0.1 M BIS-TRIS pH 5.5, 25% w/v Polyethylene glycol 3,350
D7	0.1 M BIS-TRIS pH 6.5, 25% w/v Polyethylene glycol 3,350
D10	0.1 M BIS-TRIS pH 6.5, 20% w/v Polyethylene glycol monomethyl ether 5,000
E1	0.2 M Calcium chloride dihydrate, 0.1 M BIS-TRIS pH 6.5, 45% v/v (+/-)-2-Methyl-2,4-pentanediol
E5	0.2 M Ammonium acetate, 0.1 M Tris pH 8.5, 45% v/v (+/-)-2-Methyl-2,4-pentanediol
E6	0.05 M Calcium chloride dihydrate, 0.1 M BIS-TRIS pH 6.5, 30% v/v Polyethylene glycol monomethyl ether 550
E10	0.1 M BIS-TRIS pH 6.5, 45% v/v Polypropylene glycol P 400
F1	0.2 M L-Proline, 0.1 M HEPES pH 7.5, 10% w/v Polyethylene glycol 3,350
F3	5% v/v TacsimateTM pH 7.0, 0.1 M HEPES pH 7.0, 10% w/v Polyethylene glycol monomethyl ether 5,000
H3	0.2 M Sodium malonate pH 7.0, 20% w/v Polyethylene glycol 3,350
H7	0.15 M DL-Malic acid pH 7.0, 20% w/v Polyethylene glycol 3,350
H9	0.05 M Zinc acetate dihydrate, 20% w/v Polyethylene glycol 3,350

PEG HT	WTGIR1C2/O29*8/U1A
B9	0.1 M BICINE pH 8.5, 15% w/v Polyethylene glycol 1,500
C7	0.1 M Sodium acetate trihydrate pH 4.0, 10% w/v Polyethylene glycol 4,000
C12	0.1 M BICINE pH 8.5, 8% w/v Polyethylene glycol monomethyl ether 5,000
D8	0.1 M BIS-TRIS pH 6.5, 16% w/v Polyethylene glycol 10,000
D9	0.1 M BICINE pH 8.5, 20% w/v Polyethylene glycol 10,000
D12	0.1 M BIS-TRIS propane pH 9.0, 8% w/v Polyethylene glycol 20,000
E6	10% v/v 2-Propanol, 0.1 M Sodium citrate tribasic dihydrate pH 5.0, 26% v/v Polyethylene glycol 400
G1	0.10% w/v n-Octyl-β-D-glucoside, 0.1 M Sodium citrate tribasic dihydrate pH 5.5, 22% w/v Polyethylene glycol 3,350
H12	3% w/v Dextran sulfate sodium salt, 0.1 M BICINE pH 8.5, 15% w/v Polyethylene glycol 20,000

Table 3

CONCLUSION AND PERSPECTIVES

The GIR1 branching reaction forms a short lariat with a 3-nt loop at the 5'end of the mRNA encoding the homing endonuclease (HE) (Nielsen et al., 2005). In this way, GIR1 is involved in the maturation of the 5'end of the mRNA. The resulting lariat cap helps protecting the mRNA from degradation by 5'-3' endonucleases and thus seems to confer a selective advantage to the HE. In terms of catalysis, the GIR1 branching reaction is typical of the first step of splicing performed by group II ribozymes. However, from a structural point of view, GIR1 is clearly related to group I ribozymes. Moreover, RNA molecular modelling together with biochemical and mutational data reveal that the catalytic core of this ribozyme constitutes a specific fold. Thus, GIR1s with their 2',5'-phosphodiester bond-forming activity together with a structural context related to group I intron can be listed as an independent class of naturally occurring ribozymes: the group-I-like ribozyme (Jabri et al., 1997; Nielsen et al., 2008).

GIR1 alternatively adopts active and inactive conformations:

The main function of GIR1 is to process the 5'end of the HE mRNA after GIR2 splicing. Alternatively, GIR1 seems to be involved in the regulation of the SSU ribosomal RNA synthesis by promoting its branching reaction before the splicing of the GIR2 ribozyme, mainly during the encystment pathway (Vader et al., 2002). Thus, it appears that there is a switch mechanism responsible for the coupling between the activities of the two ribozymes GIR1 and GIR2 according to environmental conditions.

Studies performed on the DiGIR1 ribozyme reveal the critical role of the P2P2.1 peripheral domain. The folding of P2P2.1 promotes the active conformation of GIR1 by stabilizing the catalytic core. In the NaGIR1 case, deletion studies together with structure probing and mutagenesis have permitted to emphasize the role of the flanking sequences. Thus, the active conformation relies on the ability to form the pHEG pseudoknot between the

L9 loop and the 3'strand of the P2 domain. Interestingly, forming the pHEG pseudoknot might also release the cleaved product from the ribozyme core. Thereby, the *Naegleria* and *Didymium* systems seem to have adopted two different mechanisms to stabilise their core to achieve the same goal: the maturation of the 5'end of the mRNA. Along the same line of evidence, they also seem to have adopted two different mechanisms involved in the release of the mRNA.

Based on biochemical and mutational data, the GIR1 peripheral domain (P2P2.1 in DiGIR1 and P2 in NaGIR1s) undergoes a conformational switching resulting in distinct inactive and active states of the ribozyme. In DiGIR1, the folding of stem loop structure HEG P1 regulates the activity of the ribozyme by destabilizing the catalytic core. Alternatively the melting of HEG P1 promotes the folding of the P2 P2.1 domain inducing the active conformation of the ribozyme. In the NaGIR1 case, the inactive conformation seems to involve the presence of the P2 stem while the active form requires the formation of the pseudoknot pHEG. In both cases, the catalytic site is proximal to the 3'end of GIR1 and the alternative conformation prevents the formation of the catalytic site. Thereby, the active conformation of GIR1 relies on conformational changes that involve the melting of key structural elements located in RNA regions resulting in the interplay between GIR1 and GIR2.

The GIR1 peripheral domain acts as an on/off switch that controls the activity of the branching ribozyme. However, *in vivo* it is possible that *trans* acting factor(s) (i.e. the homing endonuclease protein as an example) could be involved in stabilising either the active or the inactive conformation. Thus, one of our goals will be to affinity purify putative *trans* factor(s) from *D. iridis* and Naegleria cell extracts using the full intron as a bait in order to identify the individual components of the system. Along the same line of idea, the *in vivo* structure of GIR1 and the twin-ribozyme intron can be also elucidated by using first a combination of classical RNA structure probing reagent (DMS (Brunel and Romby, 2000), BzCN (Mortimer and Weeks, 2009), and SHAPE (Selective 2' Hydroxyl Acylation analysed by Primer Extension) (Wilkinson et al., 2006; Merino et al., 2005)) (Figure 50). Following that, primer extension reactions can be analysed by using automated sequencing strategies to better understand the structure of the inactive conformation of GIR1 in the precursor rRNA and also

the structure of the twin-ribozyme intron. Thus, by comparing *in vitro* and *in vivo* data, structural variations could be identified and alternatively it could also give some clue for understanding the folding of the HE mRNA.

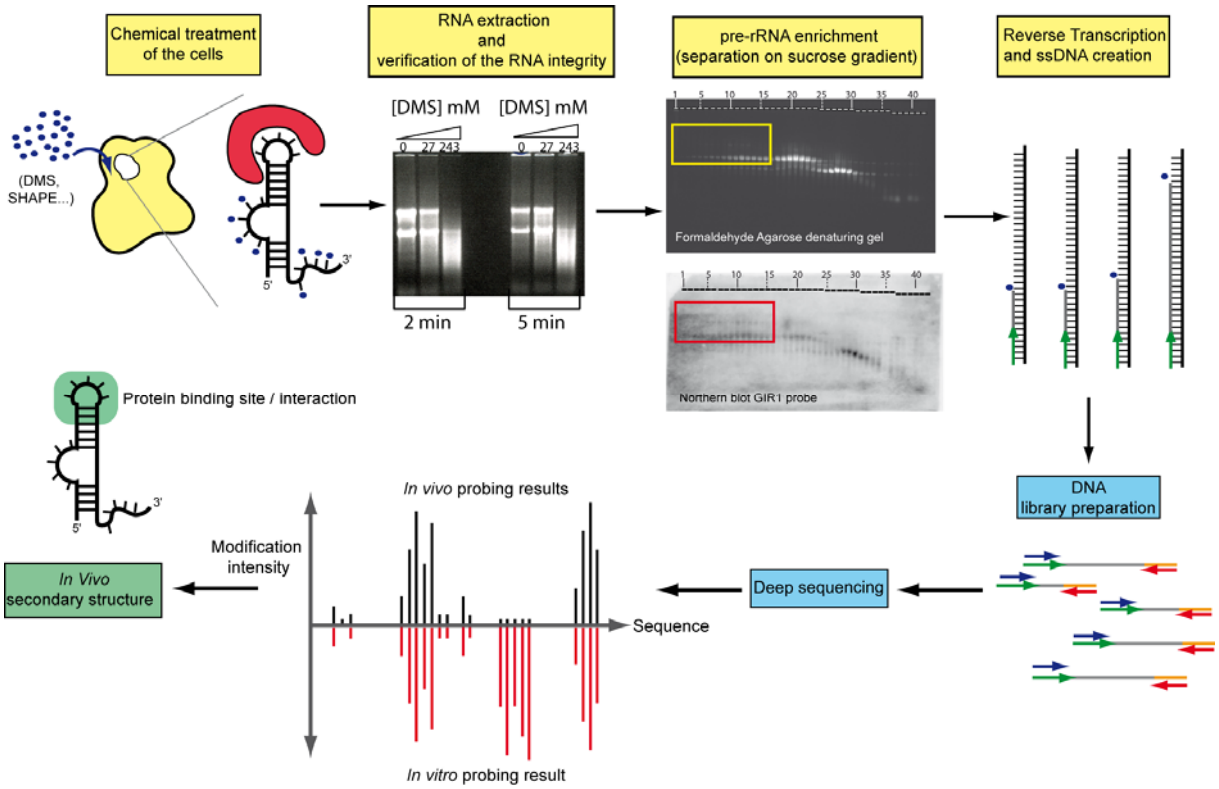


Figure 50
In vivo probing and deep sequencing strategy

Group-I-like ribozyme structure and evolution:

Interestingly, our DiGIR1 3D molecular model suggests an evolutionary link with group I introns. The evolutionary mechanism of GIR1 suggests it could derive from a group I ribozyme that incidentally inserted itself within a group I ribozyme that was already present in the SSU gene. Strikingly, the DiGIR1 model shows how two distinct RNA junctions functionally replaced the critical junction of the splicing ribozyme in a process driven by the appearance of a double pseudoknot. Thereby, this shuffling between functional motifs promoted by sequence drift, strand exchange and alternative base pairing have arisen in the GIR1 ribozyme due to the absence of selection pressure for the splicing activity. Finally and as a general observation, the process involving the appearance or disappearance of

pseudoknots can also be instrumental in the evolution of other large structured RNAs like RNase P.

Our future goals will be first to determine the structure of GIR1. We want to solve the crystal structure of the GIR1 ribozyme with its 5' and 3' appendages to be able to compare the organization of its catalytic site to group I ribozymes. GIR1 crystal structure would constitute the first crystal structure of a branching ribozyme. The crystal structure should also teach us how the RNA elements tethering GIR1 to the twin-intron are conformed. Alternatively, this will help us to better understand the molecular basis of the GIR1/GIR2 cross-talk. As previously highlighted by the DiGIR1 model, an evolutionary link exists between the group I like ribozyme and the group I splicing ribozyme. Thus, we want to emphazise this evolutionary link by transforming the GIR1 ribozyme into an *Azoarcus* like ribozyme by *in vitro* selection. By doing this, we want to understand how an RNA molecule can evolve to gain a new function. Interestingly, the *in vitro* selection strategy has already permitted to explore the catalytic potential of RNA (Bartel and Szostak, 1993; Szostak et al., 2001). It is possible to explicitly reproduce various chemical activities by succeeding in the (re)creation of ribozymes (Bartel and Unrau, 1999) capable of aminoacylating tRNA (Lee et al., 2000), accelerating peptide bond formation (Zhang and Cech, 1997), catalysing steps in nucleotide synthesis (Unrau and Bartel, 1998), acting as ligase (Lawrence and Bartel, 2005) and even acting as a processive RNA replicase (Ekland and Bartel, 1996; Johnston et al., 2001).

Figure legend:

(A) *In vitro* selection strategy developed for transforming GIR1 into a splicing like group I ribozyme. In a preliminary step, the peripheral domain is removed from GIR1 and a P1 domain is grafted. A 6 nt transposition of the P15 3' strand is then performed in order to restore the topology of group I ribozymes. Then by *in vitro* selection or SELEX (systematic evolution of ligands by exponential amplification) rare functional ribozymes can be isolated from pools of over 10^{15} different sequences (Wilson and Szostak, 1999). Thus, functional ribozymes, promoting the first step of group I intron self splicing pathway in presence of co-factor (biotin-GTP), can be selected. In this way, ribozymes retaining the splicing activity can be affinity purified based on the biotin-GTP covalently bound to the 5' end. Then the pool is enriched in a new selection round until selected ribozymes have reached a significant level of activity. However, this strategy relies on ribozyme ability to perform the splicing activity in presence of the biotin-GTP. Thus, before starting the SELEX experiments, the possibility of

the biotin-GTP to be used as a splicing substrate needs to be checked. (B) Proof of concept. The *Azoarcus* pre-tRNA group I splicing ribozyme was used to monitor the feasibility of using biotin-GTP compound to promote the self spicing pathway. (B1) The azido GTP analog (8-Azidoguanosine 5'-triphosphate [g]-biotinyl-3,6,9-trioxaundecanediamine AB14 from Affinity Photoprobe LLC®) used for the fishing experiments. (B2) The ribozyme self splicing activity was monitored by using GTP or the biotin-GTP as cofactor. The cleavage condition of the pre-tRNA Azoarcus were done according to (Tanner and Cech, 1996). (B3) Preliminary fishing experiment results using streptavidin magnetic beads. After splicing, the biotin-GTP covalently bound to the ribozyme can be used to affinity purify the ribozyme using streptavidin magnetic beads (protocol according to Roch® streptavidin bead specification). This method allows us to affinity purify the ribozyme after its splicing reaction.

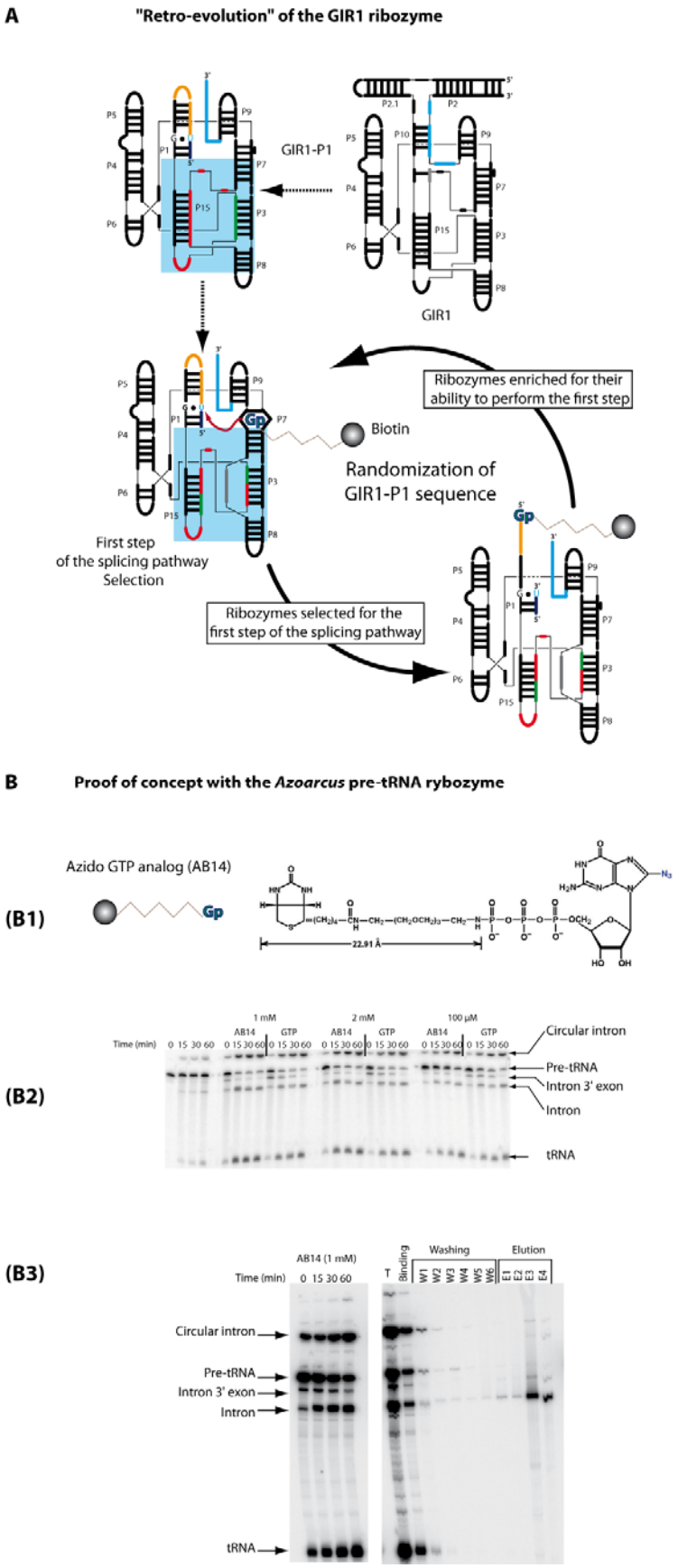


Figure 51
Illustration of the *in vitro* selection strategy and the proof of concept.

Understanding the implication of the 5'end lariat cap in translation initiation:

The GIR1 ribozyme maturates the 5'end of an mRNA by catalyzing a branching reaction that leads to the formation of a lariat cap (Nielsen et al., 2005). The resulting lariat cap protects then the mRNA from the 5' end degradation by exonucleases. Thereby, this lariat cap constitutes a substitute for the conventional cap m7G that is formed co-transcriptionally by enzymes in eukaryotic system. The m7G cap has been shown to be essential for the translation initiation in eukaryotic cells. Despite the fact that the lariat cap seems to constitute a substitute for the m7G, its role in the translation initiation remains unknown. Thereby, mRNA lariat capped product released by GIR1 can be used be for *in vitro* cap dependant translational studies (Johannes et al., 1999). As an example, the ribozyme can be fused to the sequence encoding a functional RNA. Following the transcription GIR1 folds in its catalytic active conformation and catalyzes the formation of the lariat cap at the 5' end of the downstream mRNA. In this way, various *in vitro* translation systems (yeast, insect, human and prokaryote) can be used in order to monitor the effect of the lariat cap onto the recruitment of translational initiation factors (Figure A).

Understanding of the implication of the 5'end lariat cap in RNA decay:

mRNAs are generally transient molecules and their amounts in the cell is a function of both rates of synthesis and degradation. Thus, regulated degradation is an important component of gene expression in prokaryotes and eukaryotes. The mRNA of prokaryotes and eukaryotes display several different characteristics. The eukaryotic mRNAs are generally capped by in the m7G in their 5'end while their 3' end harbours a poly(A) tail (Figure A). In contrast, prokaryotic mRNAs carry a triphosphate at their 5' end and they usually end with a stem-loop structure despite the fact that some of them can be polyadenylated (Figure A). Thus, mRNAs degradation mechanisms differ from prokaryotes to eukaryotes. Nevertheless, in both cases mRNA degradation mainly starts at the 5' end but can also start by the 3' end or even by internal cleavage.

In prokaryotic systems the mRNA undergo a rapid turnover. As previously mentioned the RNA transcripts can be degraded in a 3'-5' direction. This process is catalysed by a conserved multi enzyme complex known as the RNA degradosome (Carpousis, 2002; Carpousis, 2007). However focusing on mRNA decay via the 5'-3' direction, it has been shown in *E. coli* that the degradation process is initiated by triphosphorylated 5' end removal from primary transcript by the RppH pyrophosphohyfrolase (Celesnik et al., 2007; Deana et al., 2008) (Figure A). Thereby, RNAs carrying monophosphate in their 5' end are recognized first by RNase E and then by a number of RNases (RNase E and other RNases like RNase III, RNase G, RNase P, RNase Z…) (Deana et al., 2008). GIR1, with its ability to maturate the 5'end of an mRNA by the formation of a lariat cap, can be used to produce *in vivo* lariat capped mRNA. Thereby, the lariat cap can be an alternative to the 5' triphosphate residue from prokaryotes. The lariat cap can be used to escape the 5'-3' RNA degradation due to the fact that it is a foreign element to the cellular system. Thus it could help stabilising the mRNA. Along the same line of idea, the lariat cap can be used in combination with hairpin structures located at both 5' and 3' ends known to stabilize the mRNA in prokaryote system (Emory et al., 1992) (Figure B).

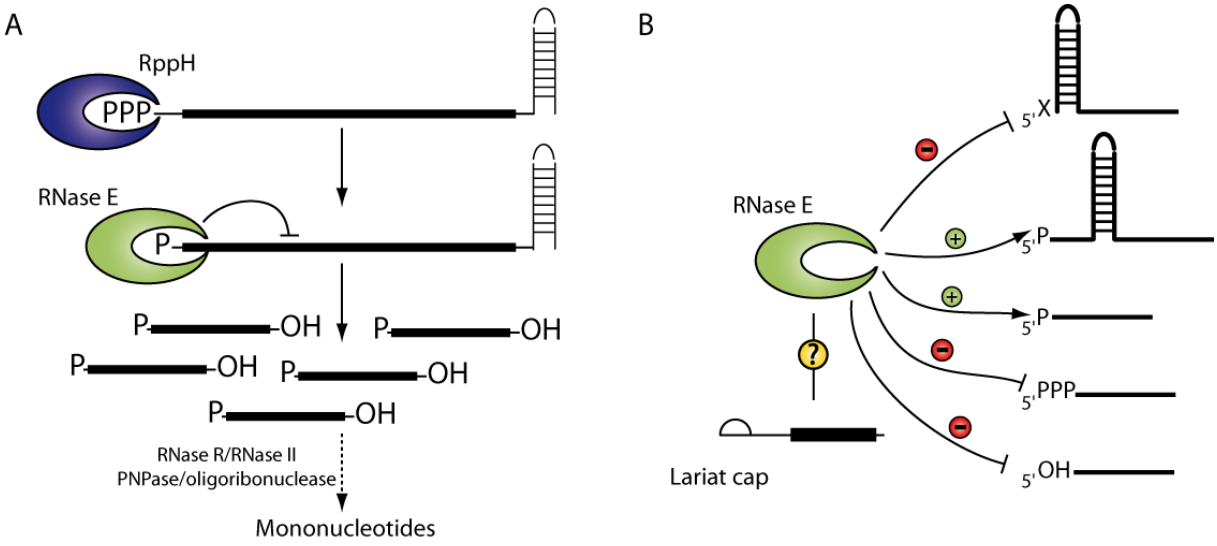


Figure 52
Degradation mechanism in prokaryotes.

(A) General scheme of the 5' end dependant mRNA degradation pathway in *E. coli*. (B) The influence of the nature of the 5' end on RNase E activity and thus on RNA turnover.

As previously mentioned, eukaryotic mRNAs are generally capped by a m7G at their 5' end and polyadenylated at their 3' end (Figure 53 A). Due to those characteristics, the mRNA degradation mechanism is different from the prokaryotic mRNA degradation process.

The major pathways of mRNA turnover in eukaryote organisms are initiated with shortening of the poly(A) tail by a poly(A)-specific exoribonuclease "Pan" (exoribonuclease member of the RNaseD family reviewed in (Tucker and Parker, 2000; Parker and Song, 2004)) (Figure 53 A). Following de-adenylation, RNA transcripts can be degraded in a 3'-5' direction (Muhlrad et al., 1995). This process, catalysed by a conserved complex of multiple 3'-5' exonucleases termed the exosome (Mitchell et al., 1997), does not involved de-capping of mRNA 5' end (reviewed in (van and Parker, 1999)). In contrast, the 5'-3' mRNA degradation involves RNA de-capping. Thereby, mRNA de-adenylation promotes mRNA de-capping mediated by two de-capping proteins Dcp1p and Dcp2p (Dunckley and Parker, 2001) which bind RNA as a prerequisite for cap recognition and are remarkably specific to the m7G cap (Piccirillo et al., 2003) (Figure 53 A). De-adenylation and de-capping steps are directly followed by the 5'-3' exonucleolytic decay by recruitment of the 5'-3' exonucleases Xrn1 (reviewed in (Parker and Song, 2004)) (Figure A). Interestingly, this 5'-3' RNA degradation pathway involving RNA transcripts de-capping, has been shown in yeast to be faster than the 3'-5'degradation process (reviewed in (Tucker and Parker, 2000)). Moreover, the currently available evidence suggests that the major mechanism of mRNA decay in *Saccharomyces cerevisiae* is done by de-capping directly followed by 5'-3' degradation (reviewed in (Tucker and Parker, 2000; Parker and Song, 2004)). Thus, the 5' end mRNA lariat capped product released by GIR1 which is a foreign cap to the cellular system, may interfere with the 5'-3' degradation process (Figure 53 B). In this way, the lariat cap may not be recognized by the RNA degrading enzymes and thus the lariat capped mRNA could escape the normal RNA turn-over and be stabilized. Based on RNA turn-over studies, it also appears that the translation initiation and the turn-over of mRNA are closely linked. Interestingly, it has been recently shown that both RNA de-capping and 5'-3' degradation process can occur when the transcripts are associated with actively translating ribosome (Hu et al., 2009). Thus, by by-passing the normal turn-over of capped RNAs, the various RNA-turnover cellular back-up systems could be revealed and thus it could also lead to gain insight into factors involved in these RNA degradation processes. It could also lead to identify new factors (Figure 53 B). Finally, GIR1 and its branching activity can be used to increase the stability of functional RNAs like miRNAs, siRNAs to extend the temporary knock down of the targetted gene or group of genes.

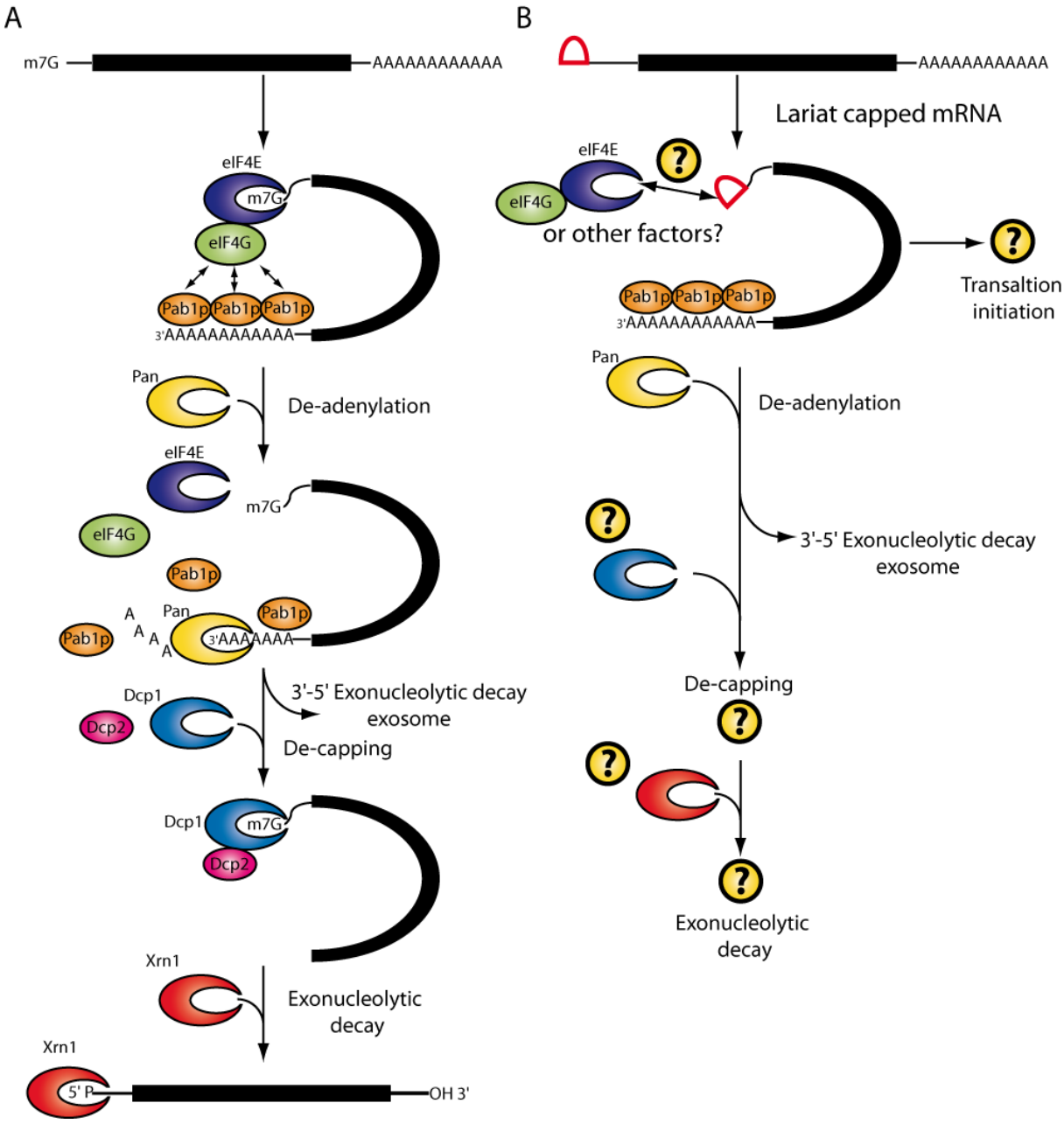


Figure 53
Illustration of the eukaryotic mRNA degradation pathway.

(A) mRNA turnover in eukaryotic cells adapted from (Tourriere et al., 2002). (B) Lariat capped mRNA turnover.

MATERIAL AND METHODES

1. Plasmid and cell strains:

1.1. Plasmids:

Three different plasmids were used in order to prepare various construction of the GIR1 ribozyme. For the PCR amplification, domain grafting the pGEM-DiGIR1 (original pG1-163) was used (Einvik et al., 1998b) (Figure 54 A). The GIR1-truncated fused to HVD ribozyme in 3' end was cloned Puc19 (Figure 54 B). Previously the insert was selected by using the plasmid vector pCR®II-TOPO (invitrogen®).

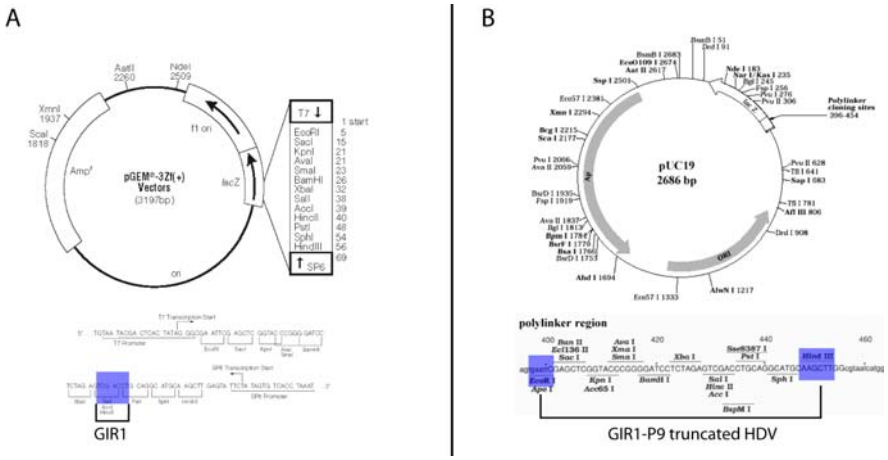


Figure 54

Restriction map of the two plasmids, pGEM and pUC19. The bleu boxes represent the location of the insert in the poly-linker region. In (A) original pGEM-DiGIR1 plasmid from (Einvik et al., 1998b), where GIR1 fragment is inserted in the HindII site. In (B) pUC19-DiGIR1-P9 truncated with HDV fusion, inserted between EcoRI and HindIII.

1.2. Cells strains:

One shot® chemically competent DH5α TOP 10 *E. coli* (Invitrogen®) were used for transformation. Cells were grown in classical LB-media containing 50 µg/mL ampicillin or on LB plates.

2. General methods for the study of DNA:

2.1. General technique of DNA manipulation:

2.1.1. Quantification of the DNA concentration:

DNA concentrations were routinely quantified spectrophotometrically on a Nanodrop ND-1000 (Thermo Scientific) by measuring the optic density (OD) at 260nm. OD260 = 1.0 equals 50µg/ml ssDNA. The absorbance values at 280nm were also measured. The

OD260/OD280 ratio provides an estimate of the purity of the RNA solution. Pure DNA has a ratio of 1.8-2.1 in 10mM Tris-HCl, pH 7.5.

2.1.2. Phenol chloroform extraction (PCI-extraction):

PCI-extraction was carried out using commercially available Phenol: Chloroform: IsoAmylalcohol (PCI) by adding 1x volume PCI, vortexing vigorously followed by centrifugation for 1min. at 5000xg. The upper DNA containing water phase was transferred to a new Eppendorf tube.

2.1.3. Ethanol precipitation:

DNA was precipitated by addition of 1/10 volume of NaAc 3M pH 6 and 2,5x volumes of 96% and leave on dry ice for minimum 15min or at -20°C O/N. Next the DNA was precipitated by centrifugation for 30min at 16.000xg. The supernatant was removed and the pellet was washed in 150µl of 70% EtOH centrifuge for 10min. Remaining ethanol was evaporated by drying the pellet under vacuum in a SpeedVac. The pellet was then uptake in depc water.

2.1.4. Basic protocols: digestion and enzymatic modification of DNA:

Procedures:

Plasmid digestion, linearization:

5,0 µg	DNA
2,0 µL	Buffer 10X
1,0 µL	Enzyme (5 U/µL)
XX µL	H_2O
20 µL	Total

The mix was then incubated at 37°C for 1 h to O/N (over night) and heat-inactivated at 65°C for 5min. The plasmidic DNA was then recovered by PCI extraction and ethanol precipitation. The pellet was then dissolved in DEPC H_2O.

Plasmid dephosphorylation:

5,0 µg	Digested plasmidic DNA
2,0 µL	BAP-Buffer 10X
1,0 µL	BAP enzyme (5U/µL)
XX µL	H_20
20 µL	Total

The reaction mix was put at 60°C for 1hrs. After 1 hrs, 2,6 µL of Proteinase K were added to stop the reaction. The mix was incubated at 37°C during 30 min. DNA was recovered by a PCI extraction and ethanol precipitation. The pellet was then dissolved in DEPC H_2O.

Insert plasmid ligation:

X µg	dephosophorylated plasmidic DNA
X µg	PCR products, insert
2,0 µL	10X buffer T4 DNA ligase
1,0 µL	T4 DNA Ligase
X µL	H2O
20 µL	Total

The mix was then incubated at RT for 1 h or at 16°C O/N

5'end labelling of primer with [γ-^{32}P] ATP:

2,0 µL	PNK-A buffer 10X
1,0 µL	PNK enzyme (5 U/µL)
1,0 µL	Primer (25 µM)
5,0 µL	[γ -^{32}P] ATP (3000 Ci/mmol)
11 µL	H_2O
20 µL	Total

The mix was then incubated at 37°C for 30min and heat-inactivated at 65°C for 5min. The 5'end labelled primer can then be used directly.

Set of primers phosphorylation:

2,5 µL	PNK-A buffer 10X
2,0 µL	PNK enzyme (5 U/µL)
1,0 µL	Primer A (25 µM)
1,0 µL	Primer B (25 µM)
5,5 µL	ATP (100 µM)
13 µL	H_2O
25 µL	Total

The mix was then incubated at 37°C for 30min and heat-inactivated at 65°C for 5min.

2.1.5. Gel-electrophoresis of DNA:

Agarose gel electrophoresis was used to separate and identify DNA fragments of different sizes. DNA of interest was mixed with approximately ¼ volume of loading buffer () and run on an 1% agarose gel at 100V in 1X TBE. The DNA fragments were visualized under UV illumination after an ethidium bromide bath.

2.2. Amplification, cloning, extraction and DNA sequencing:

2.2.1. Polymerase chain reaction (PCR):

Polymerase chain Reaction was performed to amplify specific DNA sequences. This technique is based on the knowledge of the flanking sequence and the use of two oligonucleotides surrounding of the interest region in combination with a temperature resistant DNA-polymerase enzyme. In general, the PCR reaction mix contains: DNA

template, dNTPs, DNA-polymerase, 2 set of oligonucleotide complementary to the flanking region of interest. The amplification of the sequence is done by three major steps repeated 30-40 times: (1) denaturation of the double stranded DNA, (2) annealing of the oligonucleotides to their complementary sequence, (3) extension, the oligonucleotides serve as primer for the chain synthesis provided by the DNA polymerase.

Procedures:

Reaction mix PCR preparation:

50 pg 1 µg	DNA template
1,0 µL	Reverse primer (25 µM)
1,0 µL	Pfu-polymerase (2,5 U/µL)
XX µL	H_2O
25 µL	Total

Amplification step description:

Step	Temperature	Time	Cycle
Polymerase Activation	95°C	2 min	x 1
Denaturation	95°C	30 sec	x 30
Annealing	primer Specific	30 sec	x 30
Extension	68°C	Size dependant	x 30
Final Extension	68°C	7 min	x 1

PCR products were then analysed on 1% agarose 1X TBE gel and then purified using the GenElute™PCR Clean-Up kit. For the cloning purpose, the PCR products were purified on 1% agarose by using QIAquick gel extraction from Qiagen®.

2.2.2. PCR *in vitro* mutagenesis:

In vitro PCR site mutagenesis was the method used to introduce point mutation into the DNA at a desired position or large sequence. This method requires two primers containing the desired mutation which should be previously phosphorylated in 5' end. The primer are extends by using DNA-polymerase with a high 3'-5' exonucleases proofreading activity in order to keep the error rate as low as possible. The PCR was performed in the same way as described in the previous section but with an extended time for the extension step in order to amplified the entire plasmid. After the PCR was completed the PCR products were treated for 1h with *Dpn I* in order to digest the parental DNA template and enrich the mutated synthesized plasmid DNA. The PCR products were then ligated used to transform *E. coli* cells.

2.2.3. Cloning of PCR products into a plasmid vector:

The cloning of PCR products into a plasmid vector involved the following steps: (1) Construction of recombinant DNA molecule by ligation of the desired DNA fragment into the

selected plasmid (2) Transformation of the recombinant DNA molecules into cells (3) Selection and isolation of the transformants.

The purified PCR products were cloned in the Topo TA PCR 2.1 from Invitrogen® (see Invitrogen® Topo TA cloning protocol). The Topo TA reaction mix was then used to transform DH5α *E. coli* competent bacteria. The discrimination between true positive and false positive was provided by a simple digestion of the plasmid by the restriction enzyme *Nco*I, which give two fragments (first at 2644 nt, second at 1564 nt). In order to extract the insert form Topo TA vectors, a double digestion by *EcoR*I and *Hind*III was used. Then the insert was purified on Agarose gel 1% by using QIAquick gel extraction from Qiagen®. The purified insert was then cloned in linearized dephosphorilated Puc19 plasmid. The mix (insert-plasmid) was then used to transform DH5α *E. coli* competent bacteria The discrimination between true positive and false positive was provided by a simple digestion of the plasmid by the restriction enzyme *EcoR*I and *Nco*I, which give two fragments (first at 192 nt Gir1 P9 truncated, second at 2718 nt).

2.2.4. Plasmid extraction:

The procedure used to extract plasmid DNA from bacterial cell cultures is based on the alkaline lysis procedure. Extracted plasmids were used for most purpose e. g. sequencing, PCR templates and transcription templates.

Procedures:

Material:

M-solution I:	50 mM glucose, 25 mM Tris-HCl ph 8.0, 10 mM EDTA
M-solution II:	0,2 N NaOH, 1% SDS
M-solution III:	3 M KAc, 11,5 mL Acetic Acid, add water to 100 mL

Plasmid extraction:

1) 2-10 mL cell-suspension was transferred into a new falcon tube.
2) Cells were centrifuged for 5min at 6000 rpm, and the supernatant was discarded.
3) The cells pellet was resuspended in 100 µL ice-cold M-solution I by vortexing.
4) The cells were then lysed under alkaline conditions by adding 200 µL of fresh M-solution II. The solution was mixed by gently inverting the tube a few times.
5) Neutralization was then provided by adding 150 µL of M-solution III.
6) The mixture was spun down at maximum speed for 5 min, the supernatant was then transferred into a new Eppendorf tube and treated with1 µL of RNase A (2 mg/mL)
7) The plasmid was then recovered by PCI extraction followed by ethanol precipitation.

2.2.5. Sequencing:

Sequencing was done using the Thermo Sequenase Cycle Sequencing Kit from USB. Reactions were carried out according to the supplied protocol.

Procedures:

Sequencing mix:

2 µL	Buffer 5X
1 µL	Primer 5' end labelled
1 µL	Enzyme (Thermo Sequenase DNA polymerase)
X µL	DNA template
X µL	H_2O
17 µL	Total

4 µL of the sequencing mix was added to each PCR tube that contains 4 µL of the one of the dd-nucleotide. The PCR was run with cycling parameters of 50 cycles (95°C, 30 sec; 48°C, 30 sec; 72°C, 1 min). The reaction was stopped by adding 8,0 µL of UBB loading buffer.

3. Methods for the study of RNA:

3.1. General technique of RNA manipulation:

Generally water used for all experiments was DEPC treated. DEPC inactivates RNases by modifying histidine and tyrosine residues. DEPC is added to a final concentration of 0.1% (v/v). The solution is vigorously shaken and left shaking O/N.DEPC is degraded by autoclaving.

3.1.1. Quantification of the RNA concentration:

RNA concentrations were routinely quantified spectrophotometrically on a Nanodrop ND-1000 (Thermo Scientific) by measuring the optic density (OD) at 260nm. OD260 = 1.0 equals 40µg/ml ssRNA. The absorbance values at 280nm were also measured. The OD260/OD280 ratio provides an estimate of the purity of the RNA solution. Pure RNA has a ratio of 1.8-2.1 in 10mM Tris-HCl, pH 7.5.

3.1.2. Phenol chloroform extraction (PCI-extraction):

PCI-extraction was carried out using commercially available Phenol: Chloroform: IsoAmylalcohol (PCI) by adding 1x volume PCI, vortexing vigorously followed by centrifugation for 1min. at 5000xg. The upper RNA containing water phase was transferred to a new Eppendorf tube. Generally, the PCI was buffered to pH 8.0, but for extraction of total cell RNA pH was kept at 6.6 to exclude DNA.

3.1.3. Ethanol precipitation:

RNA was precipitated by addition of 1/10 volume of NaAc 3M pH 6 and 2,5x volumes of 96% and leave on dry ice for minimum 15min or at -20°C O/N. Next the RNA was precipitated by centrifugation for 30min at 16.000xg. The supernatant was removed and

the pellet was washed in 150µl of 70% EtOH centrifuge for 10min. Remaining ethanol was evaporated by drying the pellet under vacuum in a SpeedVac. If the RNA concentration of the mixture desired to precipitate was low, glycogen was added as carrier.

3.1.4. Gel-electrophoresis of RNA:

3.1.4.1. Denaturing formaldehyde agarose gels:

Didymium Iridis total RNA was separated on a denaturing formaldehyde agarose gel. 1,2% formaldehyde agarose gel was made by boiling 1,2 g of agarose in 90 mL of water. The solution was cooled to 60°C. Then, 4 mL of 25X MOPS and 5,36 mL of 37% formaldehyde were added, and after at least half an hour at room temperature the gel polymerized. 2,7 vols. Of Formaldehyde-agarose loading buffer per µL of RNA were added. The mix was heated for 10 min at 70°C and then loaded on the gel. The gel was run at 2 V/ cm for approx. 3 hrs.

3.1.4.2. Polyacrylamide gels:

3.1.4.2.1. Denaturing polyacrylamide gels (UPAG-gel):

Denaturing polyacrylamide gel electrophoresis is a method commonly used to separate DNA/RNA fragments of different size. Transcription, sequencing reactions, primer extension reactions and product of cleavage reaction were separated on denaturing polyacrylamide gels. TEMED and APS 10% were added to the gel mix in the amount of 10 µL and 30 µL per 10 mL mix, respectively.

Analyzing gels (40x32cm): 6%, 8%, 10% or 15% acrylamide (29:1), 50% urea, 1xTBE

Preparative gels I (28x15cm): 4% or 5% acrylamide (29:1), 50% urea, 1xTBE

The polymerization of the gel was done at room temperature for 45 min to 1h at room temperature. The gel was then run at 2,1 kV (65 Watt) in 1X TBE.

3.1.4.2.2. Native polyacrylamide gels (PAG-gel):

Native polyacrylamide gel electrophoresis is a method commonly used to study the folding dynamics and the conformation homogeneity of a RNA population. For those experiments, the protocol was adapted from various examples found in the literature and some

tests were provided with the *Azoarcus* pre-tRNA ribozyme in order to reproduce the results from various publication (Lilley, 2008a; Pan et al., 1997).

Procedures:

Running buffer:

132 mL	66 mM HEPES pH 7,5 (1 M)
68 mL	34 mM Tris HCl pH 7,5 (1 M)
6 mL	3 mM $MgCl_2$ (1 M)
1794 mL	H_2O
2000 mL	Total

Native gel preparation:

17,5 mL	10 % Acrylamide (40%)
2,4 mL	34 mM Tris HCl pH 7,5 (1 M)
4,6 mL	66 mM HEPES pH 7,5 (1M)
210 µL	3 mM $MgCl_2$ (1M)
45,3 mL	H_2O
70 mL	Total

TEMED and APS 10% were added to the gel mix in the amount of 10 µL and 30 µL per 10 mL mix, respectively. After polymerization the gel was placed in the operating system and cooled down with the running buffer at 5°C (15 to 30 minutes) before the loading of the samples. After the loading of the samples the gel was run for 6-8 hrs at 4°C at 15 Watts.

3.2. Preparation of RNA by *In vitro* transcription:

The basic strategy in order to prepare a template for *in vitro* transcription is to place the sequence of interest downstream of the T7 promoter. The promoter covers the sequence ranging from -17 to +6 with +1 being the first nucleotide of the transcribed region (Figure 55). Thus, there is not complete freedom in the choice of the sequence at the very 5'-end of the *in vitro* transcript. Most T7 promoters, like class III promoters (Milligan *et al.*, 1987), have G's at +1, +2, and +3, and the first two G's are critical for transcriptional yield. The alternative class II promoters initiate with an A and have a similar preference for G's at +2 (Huang and Yarus, 1997).The termination of the *in vitro* transcription occurs usually by "run off", that is when the polymerase falls off at the very end of the template. With the PCR and oligo templates this is defined by the ends of the template products. With cloned templates this is achieved by linearizing the plasmid by restriction enzyme digestion downstream of the sequence of interest. However during run-off transcription T7 RNAP has a tendency to incorporate one or several non-templated nucleotides at the 3'-end, thus leaving the pool of transcripts with heterogeneous 3'-ends.

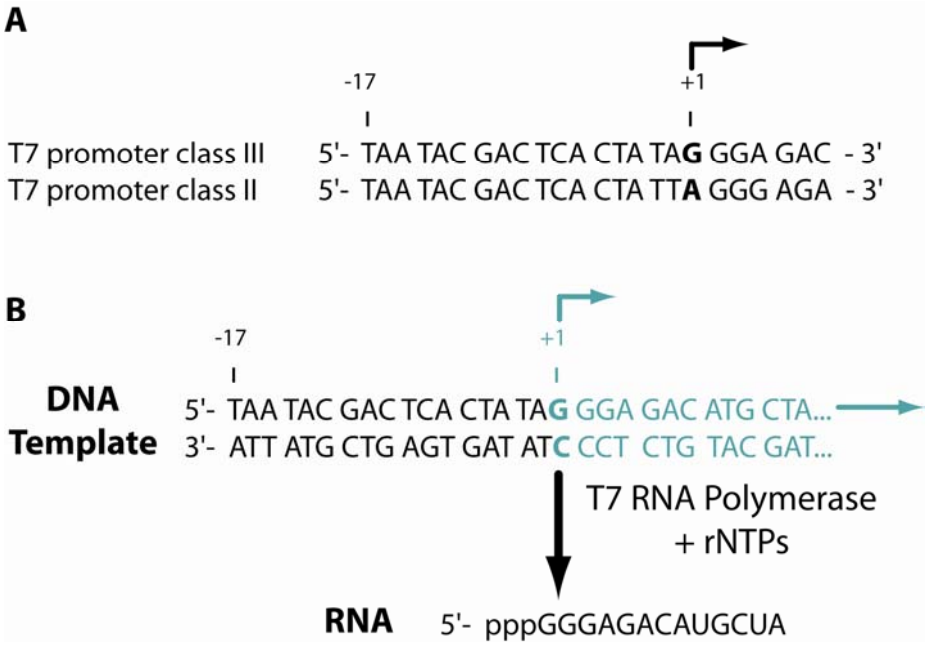


Figure 55

(A) Consensus sequence of (class III and class II) T7 RNA polymerase promoter with indication of the +1 nucleotide (bold; corresponds to the first nucleotide in the transcript). (B) When the DNA template is incubated in the presence of T7 RNA polymerase and rNTPs, a transcript is made as indicated with a triphosphate at the 5'-end.

3.2.1. Template preparation:

Two different strategies were used for the template DNA. The first strategy implies the use of DNA templates that were generated by PCR. In this case the T7 promoter was added to the PCR products by including the promoter sequence at the 5' end of the forward primer. When small amounts were needed, PCR-products were probably the most convenient due to the flexibility in design of the template and the ease its production. The second strategy consists in the cloning of the DNA template including a T7 promoter immediately 5' of the sequence to be transcribed in a pUC18 or pUC19 plasmid that doesn't contain a built-in T7 promoter in opposition with other plasmids (e.g. the pBluescript (Stratagene) and pGEM (Promega)). The second strategy was used when large amounts of RNA were required, using simple and economical techniques based on bacterial culture and plasmid extraction to create the template.

3.2.2. Transcription by using the T7-polymerase:

In vitro transcription was performed by using commercial T7 RNA polymerase, and the reaction was carried out according to the supplied protocol. However, the commercial T7 RNA polymerase could be easily replaced by an in-house T7 RNA polymerase made by expression and purification of an His-tagged T7 RNA polymerase (plasmid pT7-911Q) (Ichetovkin *et al.*, 1997) in order to prepare large scale amount of RNA.

Procedures:

Transcription mix:

5,0 µL	Transcription Buffer 5X
5,0 µL	rNTPs (2,5 mM each)
0,8 µL	DTT (1M)
3,0 µL	PCR Template
1,0 µL	T7-polymerase (20 U/µL)
10,2 µL	DEPC H_2O
25 µL	Total

1 to 2 µL of [α-^{32}P] UTP can be added to get body labelled RNA.

The mix was then incubated at 37°C for 1-4 hrs. After the transcription reaction PCR product were digested by using 1 µL of D*nase*I for 15min at 37°C. Then RNA was recovered by PCI extraction and ethanol precipitation. Alternatively the RNAs can be purified on 5% UPAG gel.

3.3. End labelling of the RNA:

3.3.1. 5' end labelling:

In vivo or *in vitro* synthesised RNA generally contains a 5' phosphate(s). Prior to labelling these phosphates must be removed. However, the dephosphorylation of *in vitro* transcribed RNA can be avoided if a diribonucleotide, such as ApG is included in the transcription mixture. This diribonucleotide will be incorporated at the 5'end of the transcript.

Procedures:

Dephosphorylation of the in vitro transcribed RNAs:

5,0 µg	RNA
2,5 µL	Antarctica phosphatase buffer 10X
2,0 µL	Antarctica phosphatase enzyme (5 U/µL)
25 µL	Total

Before the dephosphorylation reaction the RNAs are heat denatured at 95°C for 30s to 1min and cool down on ice. The mix reaction was then incubated at 37°C for 30min and heat-inactivated at 65°C for 5min. The RNAs were then recovered by PCI extraction and ethanol precipitation. The RNAs were then dissolved in 10µL DEPC H_2O.

5' end labelling of the RNA by [γ-^{32}P] ATP:

10 µL	Deposphorylated RNA
2,5 µL	PNK-A buffer 10X
1,0 µL	PNK enzyme (10 U/µL)
5,0 µL	[γ-^{32}P] ATP (3000 Ci/mmol)
11,5µL	H_2O
25 µL	Total

Before the 5'end labelling reaction the RNAs are heat denatured at 95°C for 30s to 1min and cool down on ice. The mix reaction was then incubated at 37°C for 30min to 1h and heat-inactivated at 65°C for 5min directly followed by purification of the 5'end labelled transcript on UPAG 5%.

3.3.2. 3' end labelling:

3.3.2.1. By using the T4 RNA ligase:

The 3' end labelling of RNA was done by ligating [α-^{32}P] pCp to the 3' end using the T4 RNA ligase. The advantage of this ligase method is that the radioactive pCp is relatively easy to produce from standard radioactive nucleotides (Kjems *et al.*, 1998). However during run-off transcription T7 RNAP has a tendency to incorporate one or several non-templated nucleotides at the 3' end, thus leaving the pool of transcripts with heterogeneous 3' ends. By using this ligase we obtain a pool of transcripts with heterogeneous 3' end labelling.

Procedures:

Preparation of [α-^{32}P] pCp from Cp and [γ-^{32}P] ATP:

2,0 µL	PNK-A buffer 10X
1,0 µL	Cp 3 mM
1,0 µL	PNK enzyme (5 U/µL)
10 µL	[γ -^{32}P] ATP (3000 Ci/mmol)
1,0 µL	ATP 25 µM (can be omitted if higher activity is needed)
20 µL	Total

Incubation at 37°C for 30 min and heat-inactivate at 70°C for 5 min. The [α-^{32}P] pCp is directly use in the ligase reaction.

3' end labelling of the RNA by ligating [α-^{32}P] pCp:

2,0 µg	RNA
2,0 µL	T4 RNA Ligase Buffer 10X
5,0 µL	pCp Mix previously prepared
2,0 µL	ATP 1 mM
1,0 µL to 2,5µL	DMSO 100%
1,5 µL	T4 RNA Ligase (5 U/µL)
20 µL	Total

The RNAs are previously heat denatured at 95°C between 30s to 1min before the reaction. Then the incubation takes place at 4°C overnight or 2h at RT directly followed by purification of the 3'end labelled transcript on UPAG 5%.

3.3.2.2. By using the Klenow Fragment:

In order to obtain RNA with a specific 3' end labelling that avoids non-precise termination of transcription by the T7 polymerase, we have use short DNA oligonucleotide

complementary to the 3' end of the targeted RNA in combination with the use of Klenow fragment of DNA polymerase I and [α-^{32}P] dCTP (Huang and Szostak, 1996). However, the Klenow fragment of DNA polymerase I may non-specifically extend the 3' end of the DNA oligonucleotide, thus making the labelling inefficient. To increase the efficacy of the 3' end RNA labelling, the 3' end of the DNA oligonucleotide has to be modified (i.e. dideoxy modification) and the 5' end should contain 3' GCC 5' overhang (Figure 56) (Shcherbakova and Brenowitz, 2008).

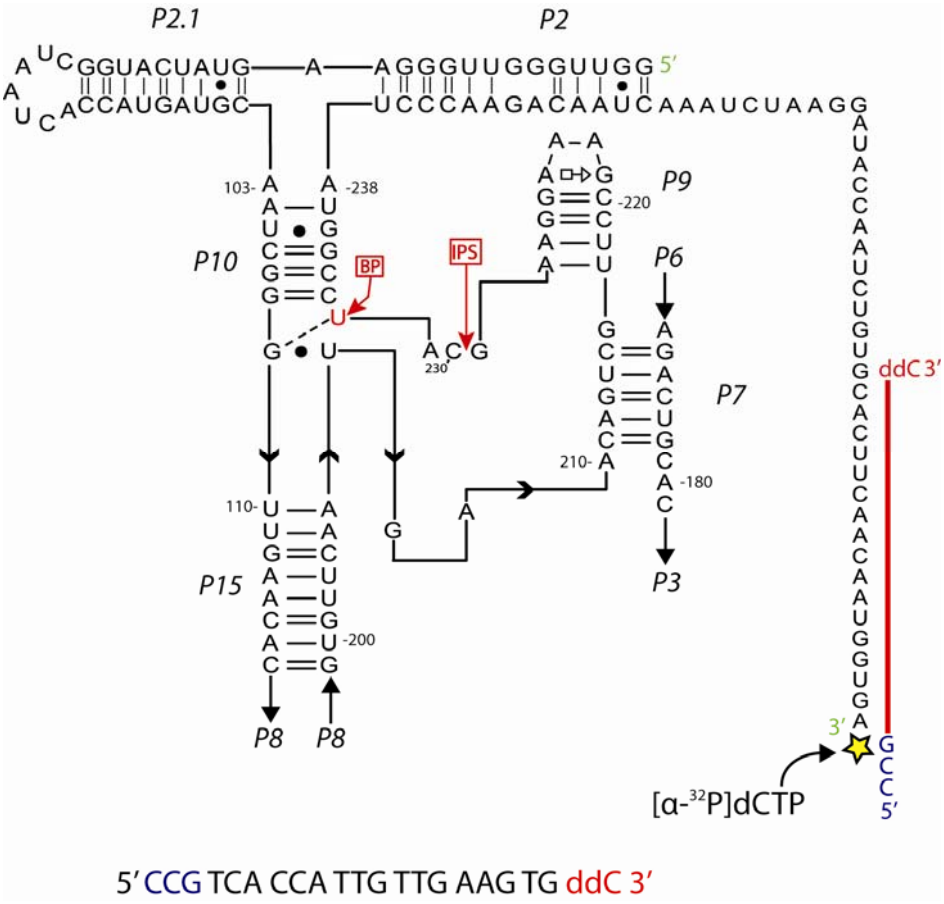


Figure 56

Schematic representation of 3'-end labelling strategy. The 3' end domain of the DiGIR1 ribozyme is presented as an example. Its 3' end is labelled by addition of a [α-^{32}P] dCTP by the Klenow fragment of DNA of the polymerase, following the annealing of DNA oligonucleotide (red part) with 3'-GCC-5' overhang (blue letters) and the dideoxy modified in 3'-end of the primer.

Procedures:

Mix 1:

3,0 µg	RNA
1,0 µL	Annealing buffer 10X (140 mM Tris HCl, pH 7, 400 mM NaCl, 2 mM EDTA)
10 µL	Total

The Mix 1 should be heat denatured at 95°C for 1min and slowly cool down to 37°C (1°C/sec) and subsequent incubation on ice for 10 min. Then to the Mix 1 add Mix 2:

Mix2:

3,0 µL	Klenow Fragment 10X buffer
8,0 µL	[α-^{32}P] dCTP (3000 Ci/mmol)
1,0 µL	Klenow Enzyme (10 U/µL)
30 µL	Total

Incubation at 37°C for 2h and heat-inactivate at 95°C for 2 min by adding UBB directly followed by purification of the 3'end labelled transcript on UPAG 5%.

3.4. Working with GIR1 ribozymes:

DiGIR1 catalyses three different reactions. The natural reaction is the branching reaction (1 in Figure 57) in which a transesterification at the IPS results in the cleavage of the RNA with a 3'OH and a downstream lariat cap made by joining of the first and the third nucleotide by a 2',5' phosphodiester bond. In vitro, DiGIR1 catalyses the reverse reaction (2 in Figure 57), referred to as the ligation reaction. It is very efficient to the extent that the forward reaction is completely masked in reactions with full-length intron and length variants that include more than 166 nucleotides upstream of the IPS. The branching reaction is isolated from the reverse reaction by cleavage in the presence of 2 M urea that inhibits the ligation reaction. Finally, DiGIR1 catalyses hydrolytic cleavage at the IPS (3 in Figure 57) at a relatively low rate. This is the cleavage reaction observed with the full-length intron and several length variants. The hydrolytic cleavage is irreversible and is considered an *in vitro* artefact resulting from a failure to present the branch nucleotide (BP) correctly for catalysis.

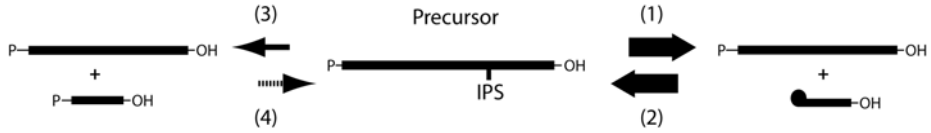


Figure 57

Reactions known to be catalysed by GIR1 branching ribozyme. The main activity (1) is the branching activity. However, the branching reaction is highly reversible and can even be masked by ligation reaction (2). A hydrolytic cleavage reaction (3) is less pronounced and only observed *in vitro*. (4) Hypothetical reaction that has not been observed (dashed arrows)

Procedures:

Cleavage kinetic of the GIR1 ribozyme:

Material:

Refolding buffer 2X:	2 M KCl, 50 mM $MgCl_2$, 20 mM Acetat (pH 5.47)
Start buffer 25 mM:	47,5 mM Hepes (pH 7.5), 1 M KCl, 25 mM $MgCl_2$

Mix reaction:

X µL	RNA
5,0 µL	Refolding buffer 2X
X µL	H2O
10 µL	Total

The RNAs are previously heat denatured at 95°C between 30s to 1min before then the RNA are refolded at 45°C for 5 min. The cleavage reaction starts by adding 4 volumes of Start buffer. Aliquot of the reaction mixture were then uptake and mix with UBB loading buffer to stop the reaction. The samples were heat denatured before loading on 8% UPAG gel.

3.5. Probing in solution of the RNA structures:

Structure probing in solution is based on the reactivity of the RNA molecules toward chemical or enzymes that have specific target on it. All the probes are used under statistical conditions, where less than one cleavage or modification occurs per molecule. The identification of the cleavages or modifications can be done by two different techniques depending on the length of the RNA molecule and the nature of the nucleotide positions probed. The first technique, which requires end labeled RNA, only detects scissions in the RNA and is limited to molecules containing less than 300 nucleotides. The second approach is most likely an indirect method, by using primer extension in order to detect stops of the reverse transcription at modified or cleaved nucleotides.

3.5.1. Fe-EDTA probing:

Footprinting describes assays which investigate ligand binding or conformational changes by monitoring the accessibility of a nucleic acid backbone to an exogenous probe. Quantitation of the accessibility is achieved by chemical and/or enzymatic probes which modify or cleave the nucleic acids. Amongst the available chemical footprinting probes (Brunel and Romby, 2000) •OH radicals offer a unique combination of properties as they are highly reactive and cleave the RNA backbone without nucleotide or base pairing specificity (Latham and Cech, 1989). Due to the similar size of water molecules and hydroxyl radical (•OH), •OH footprinting reports the solvent accessible surface of a RNA molecule, and provides quantitative information about the structural changes associated with macromolecular folding, interactions and ligand binding. Structure probing in solution of the GIR1 ribozyme was done by using the hydroxyl radical footprinting in order to get a better understanding of the ribozyme folding.

Procedures:

Fe-EDTA reaction:

Mix reaction:

X µL	RNA 5' or 3' end labelled (~ 50 000 cpm/µL)
1,0 µL	Fe ($(NH_4)_2Fe(SO_4)_2$ 75 mg/ml)
1,0 µL	EDTA pH 8.0 (150 mM)
1,0 µL	DTT (375 mM)
1,0 µL	H_2O_2 (15 %)
X µL	H_2O
25 µL	Total

Reactions were quenched after 2 min at room temperature by ethanol precipitation at -80°C for 15 min by adding 175 µL H_2O, 1/10 volume of 3 M sodium acetate (pH 5), 3 volume of ethanol 96% and 1µg of carrier tRNA. Samples were pelleted, dried and dissolved in 6 µL of denaturing blue loading buffer with 5M urea. Before loading on 15 % denaturing polyacrylamide gels the samples were heated at 95°C.

RNase T1 ladder preparation:

Material:

Buffer ΔT1: Citrate NaOH 20 mM, EDTA 1mM, Urea 7 M, 0,05 % Xylene Cyanol, 0,05 % Bromophenol blue

Mix reaction:

X µL	RNA 5' or 3' end labelled (~100 000 cpm/µL)
5 µL	Buffer ΔT1
1 µL	Cold RNA 2 mg/mL

The mix was preincubated at 50°C for 5 min, then 1µL of RNase T1 (28 U/µl) was added. The mix was then incubated for 10 min at 50°C. The reaction was stopped by adding 4 µL UBB and heated denatured before loading on 15 % denaturing polyacrylamide gels.

Alkaline ladder preparation:

Material:

Buffer L: carbonate de NaOH 0,1 M (Na2CO3/NaHCO3 100 mM each) pH 9,2; 1 mM EDTA

Mix reaction:

X µL	RNA 5' or 3' end labelled (~100 000 cpm/µL)
5 µL	Buffer L
1 µL	Cold RNA 2 mg/mL

The mix was incubated for 3 min at 95°C. The reaction was stopped by adding 4 µL UBB and loaded on 15 % denaturing polyacrylamide gels.

3.5.2. Chemical probing:

Structure probing was performed essentially as described in (Kjems 1998). Briefly, 4 µg of *in vitro* transcript in 200 µL (chemical modification) or 40 µL (enzymatic reaction) in

probing buffer on ice (270 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 70 mM HEPES-KOH (pH 7.8)) was incubated with the probe. The following specific conditions were applied. DMS: 2µL of 50% DMS for 20 min. DEPC: 7µL for 30 min. Kethoxal: 20 µL of 40 mg/mL for 75 min. CMCT: RNase T1: 0.1 U or 0.2 U. RNase T2: 0.5 U or 1 U. RNase A: 0.05 U or 0.1 U. RNase V1: 1/300 U or 2/300 U. All reactions were terminated by ethanol precipitation and subjected to primer extension reactions as previously described (Einvik 1998, Einvik 2002).

3.6. Primer extension and RNA direct sequencing:

Primer extension is one of the most common methods used to measure the amount, the size of the RNA. This method is also used to map and quantify the 5' end of RNAs. Generally, an end labeled oligonucleotide, between 10 to 18 nucleotides, is hybridized to RNA and then extended using reverse transcriptase to produce single-stranded cDNA. The reverse transcriptase stops at points either where the RNA was modified by chemical probe or strong secondary structure, or either where a break is introduced e.g. self-cleavage reaction, enzymatic reaction or degradation. The stops due to the pause of the reverse transcriptase are mapped in a denaturing gel electrophoresis by comparing the cDNA length with a sequence ladder (e. g. direct RNA sequencing or basic DNA sequencing). This general method was used first to detect the modification induce by a chemical probe and second to detect and quantify the branching reaction of the GIR1 ribozyme.

Procedure:

Basic Primer extension:

Annealing mix :

X µL	RNA template
X µL	H_2O
1,0 µL	primer 5' end labeled
1,2 µL	RT-Buffer 5X (Fermentas M-MuLV)
6,0 µL	Total

The annealing was heated at 81°C for 1min and slowly cooled down to 42°C and incubated for 10 min at 42°C. Then to the annealing mix was added:

RT mix:

0,8 µL	RT-Buffer 5X (Fermentas M-MuLV)
1,0 µL	dNTPs (2 mM)
0,1 µL	RT-Enzyme (Fermentas M-MuLV H^-, 200U/µL)
2,1 µL	H_2O
4,0 µL	Total

The final mix was incubated at 42°C for at least 1 hrs. The reverse transcription reaction was then stooped by adding 10 µL UBB loading buffer and denatured at 95°C prior to run on gel with appropriate sequencing ladder.

RNA direct sequencing:

Materials:

dNTPs	0,5 mM
ddGTP	0,25 mM
ddATP	0,5 mM
ddTTP	1,0 mM
ddCTP	0,5 mM

Annealing mix :

X µL	RNA template
X µL	H_2O
2,0 µL	primer 5' end labeled
2,4 µL	RT-Buffer 5X (Fermentas M-MuLV)
12,0 µL	Total

In 4 Eppendorf tubes 0,5 µL of dd-NTP was mixed with 3 µL of the annealing mix. The mix was heated at 81°C for 1min and slowly cooled down to 42°C and incubated for 10 min at 42°C. Then to the mix 1 was added 1,5 µL of the RT mix

RT mix:

1,6 µL	RT-Buffer 5X (Fermentas M-MuLV)
2,0 µL	dNTPs (0,5 mM)
0,3 µL	RT-Enzyme (Fermentas M-MuLV H^-, 200U/µL)
2,1 µL	H_2O
6,0 µL	Total

The final mix was incubated at 42°C for at least 1 hrs. The reaction was then stooped by adding 5 µL UBB loading buffer and denatured at 95°C prior to run on gel.

4. Culture of the slime mould *Didymium iridis*:

4.1. *E .coli*-KB culture:

The *E. coli*-KB (strain given by Professor S. Johansen) served as source food for the slime mould Didymium iridis. This *E. coli*-KB was grown like any other strain of *E. coli*.

Procedures:

Buffer:

LB-media: 25g yeast extract, 50g Tryptone, 25g NaCl, dH_20 ad 5 liters, autoclave for ½ an hour at 120°C. Or 20g LB Broth (Difco) dH_20 ad 1liter, autoclave for ½ an hour at 120°C.

TM-buffer: 1.0ml 1M Tris-HCl (pH 8.0), 2.5ml 1M $MgSO_4$, dH_2O ad 100ml, autoclave for ½ an hour at 120°C.

Growing and harvesting *E. coli*-KB:

1) 400ml LB-media was inoculated with 20µl of 10X *E. coli*-KB from freeze stock or previous 5°C-stock.
2) The cells were incubated with shaking at 37°C O/N.
3) The *E. coli*-KB was harvested 2x 180ml by centrifugation at 1450*g*.
4) Cells pellet was resuspended in 2x 18ml TM-buffer and stored at 5°C.

4.2. Growing of the slime mould *Didymium iridis*:

The slime mould *Didymium iridis* Lat3-5 strain derived from the Pan2-44 isolate as previously described in (Johansen et al., 1997a) was cultivated. The cells were grown at 25°C by calmly shaking in liquid media (DS/2) with 1/5 of the volume being 10X *E. coli*-KB functioning as source food for *Didymium*. The cells concentration in the growing media was determined by using Burker-Turk cell-counter plate (manual method) or by using Casy® cell-counter from Innovatis® (automatic cell-counter). According to time of growing, the cells concentration, the position in the grow curve (Figure 58) and the cells differentiation stage, the cells were harvested as previously described in (Vader and Nielsen, 1999).

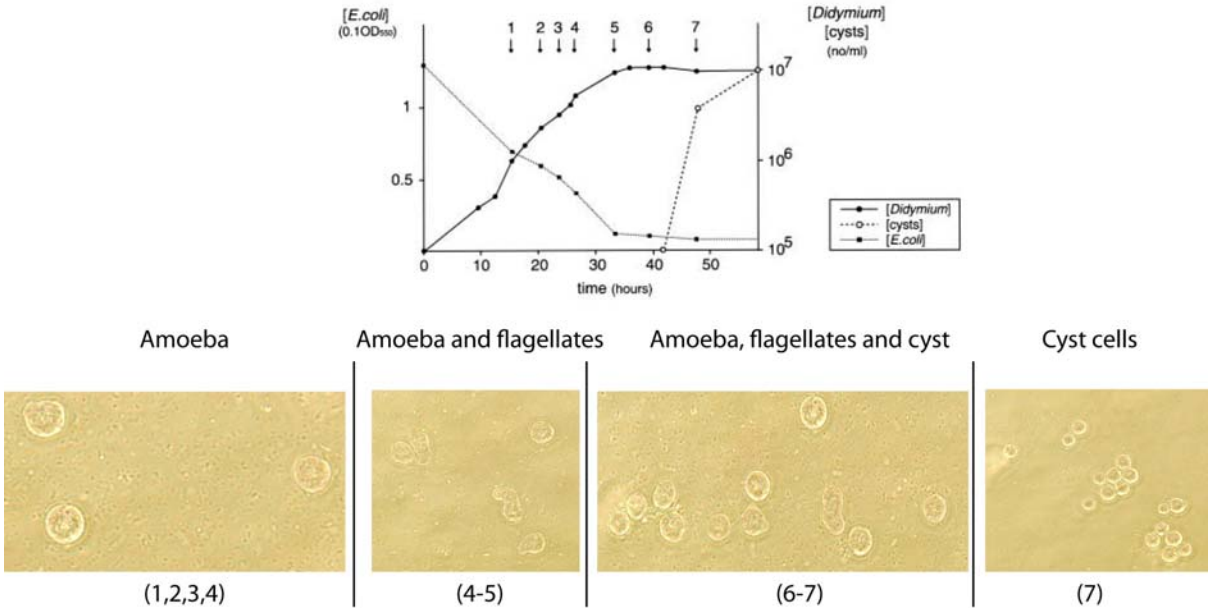


Figure 58

Time course of culture growth, showing starvation and subsequent encystment of vegetative *Dydimium iridis* Lat 3-5 cells with the cells differentiation according the growing time. The time points when total number of Didymium cells, number of encysted cells or amount of E. coli food was measured are indicated.

Procedures:

Buffer:

DS/2-media: 1.0g D-glucose, 0.5g yeast extract, 0.1g $MgSO_4$, 1.0g KH_2PO_4, 1.5g K_2HPO_4, dH_2O ad 1 liter, autoclave for ½ an hour at 120°C.

Growing *Didymium* cells:

1) 24 mL of 25°C pre-warmed DS/2 was poured in 200 mL PYREX glass previously washed and baked at 200°C.
2) 5 mL of 10X E.coli-KB was added to the DS/2 media.
3) 1 mL of 10^6-10^7 cell/ml was added to the mix DS/2 media-*E.coli*-KB.
4) The cells were grown at 25°C by calmly shaking.

Harvesting cells procedure:

1) 2-10 mL cell-suspension (~10^6-10^7 cells/ml) was transferred into a new falcon tube.
2) Cells were harvested cells by spinning at 450*g* for 5min, at the same temperature than the growing condition in order to avoid cold/Heat shock stress.
3) The supernatant that contains the majority of bacteria was poured in a new falcon tube for determination of harvest efficiency.
4) The Didymium cells were then washed with 20 mL of DS/2-media and spin down at 450g for 5min.
5) The supernatant was then poured in a new falcon tube for determination of harvest/washing efficiency

Cells pellet was then resuspended in DS/2 media with no yeast extract. The washing of cells (Step 4-5) can be repeated if the cells need to be absolutely free from the bacteria.

4.3. *In vivo* probing:

***In vivo* DMS modification** (Additional steps after step 4 in protocol for "Harvest of Didymium cells") (Modified from Zaug & Cech 1995)

4a) Resuspend cell pellet in 4ml DS2 (Excluded yeast extract). At least 3 tubes necessary.
4b) Add to one tube 10-90 µl DMS (27-240mM final concentration). Incubate at RT for 2 min. gently rocking the cell suspension.
4c) Quench the DMS reaction by adding 200 µl β-Mercaptoethanol (0,7M final concentration) and add equally 200 µl β-Mercaptoethanol to one control tube, whilst last control tube is blank.
4d) Harvest cells as described from step 3 in protocol for "Harvest of Didymium cells".

4.4. Total RNA extraction:

GuSCN-stock: 152g GuSNC, 5.3ml 1.5M NaCl, 8ml 20%sarcosyl (filtrated not autoclaved), 6.4ml 0.5M EDTA (pH 8.0), dH_2O ad 200ml. Heat to 65°C to dissolve. While warm filter through 45µm nitrocellulose filter. Increase volume to 197ml.

RNazol: 80ml phenol (1g 8-hydroxyquinoline in 500g phenol), 49.2ml GuSCN-stock, 0.8ml β-mercaptoethanol, 8ml 2M NaAc (pH 4.0), 30ml dH_2O.

1) Resuspend cell pellet in RNAzol (2,5ml per 10^7 cells). Shake vigorously.
2) Add 0,2x vol. of chloroform. Shake vigorously.
3) Rest on ice for 20 min., occasionally turning the tube up-side down.
4) Spin 1.811*g*, 20 min.

5) Transfer H_20 phase to glass Corex tube. Add 1x vol. isopropanol. Incubate O/N at 5°C.
6) Centrifuge at 16.500*g*, 40 min.
7) Discard supernatant; add 1ml 70% EtOH, 0,25M NH_4Ac. Centrifuge at 16.500g, 10 min.
8) Discard supernatant and dry pellet at room temperature for around 30 min.
9) Resuspend pellet in around. 600µl H_20.
10) Determine concentration and purity by 260nm absorbance and 260/280 ratio respectively. For precise determination of concentration: confirm absorbance with Ribogreen Assay™.

5. Oligonucleotide table:

oligos	sequence	size	Made for	DiGIR
C287	AAT TTA ATA CGA CTC ACT ATA GGG TTG GGA AGT ATC AT	38	DiGIR1; T7prom + 5'end (162)	162.
C288	TCA CCA TGG TTG TTG AAG TGC ACA GAT TG	29	DiGIR1; 3'-oligo same as OP12	.65
C289	TTA ATA CGA CTC ACT ATA GGT TGG GTT GGG AAG TAT CAT	42	DiGIR1; T7prom + 5'end (166)	166.
C291	GAT TGT CTT GGG AT	14	DiGIR1; for PX of IPS1/2	PX
C292	AAT TTA ATA CGA CTC ACT ATA GTT TTG GTT GGG TTG GGA AGT ATC AT	47	DiGIR1; T7prom + 5'end (171)	171.
C293	AAT TTA ATA CGA CTC ACT ATA GTT GGG AAG TAT CAT	36	DiGIR1; T7prom + 5'end (160)	160.
C294	AAT TTA ATA CGA CTC ACT ATA GGG AAG TAT CAT	33	DiGIR1; T7prom + 5'end (157)	157.
C295	AAT TTA ATA CGA CTC ACT ATA GGT TTT GGT TGG GTT GGG AAG TAT CAT	48	DiGIR1; T7prom + 5'end (G171)	171.
C296	AAT TTA ATA CGA CTC ACT ATA GGG TTG GGA AGT ATC ATA GCT AAT CAC TAT GAT GCA ATC GGG TTG AA	68	DiGIR1; T7prom + 5'end (162)/ G85A;C95T-mut P2.1	162. (P2.1 Mut)
C301	TCA CCA TGG TTG TTG A	16	GIR1; short version of OP12 (C288) for px	PX
C302	TTC CTT TCA CCA TTG T	16	GIR1; short version of C298 for px	PX
C303	GAT TGT CTT GGG ATA CCG	18	GIR1; re-named OP233	
C321	GCA TCC GGT ATC TCA AGA CAA TCA AAT CTA AGG	33	GIR1; C241T (P2) by QC, RNA-like	
C322	CCT TAG ATT TGA TTG TCT TGA GAT ACC GGA TGC	33	GIR1; C241T (P2) by QC, α-	

			sense	
C323	GCA TCC GGT ATC CTA AGA CAA TCA AAT CTA AGG	33	GIR1; C242T (P2) by QC, RNA-like	
C324	CCT TAG ATT TGA TTG TCT TAG GAT ACC GGA TGC	33	GIR1; C242T (P2) by QC, α-sense	
C352	TAA TAC GAC TCA CTA TAG GAA CAC TTA ATT GGG TTA	36	GIR1; T7-start P15	P15.
C354	AAT TTA ATA CGA CTC ACT ATA GGT TGT CTT GGG AAG TAT CAT	42	GIR1;T7mut GG69TC	
C363	CGT TCC GAA AGG AAG TAT CCG GTA TCC CAA G	31	GIR1;C230U QC (sense)	
C364	CTT GGG ATA CCG GAT ACT TCC TTT CGG AAC G	31	GIR1;C230U QC (a-sense)	
C365	CGT TCC GAA AGG AAG CGT CCG GTA TCC CAA G	31	GIR1;A231G QC (sense)	
C366	CTT GGG ATA CCG GAC GCT TCC TTT CGG AAC G	31	GIR1;A231G QC (a-sense)	
C513	GCA CGG CCC TGC CTC TTA GGT AAT GAA CAG TCG TTC CGA AAG G	43	DiGIR1D15'' (sense)	
C514	CCT TTC GGA ACG ACT GTT CAT TAC CTA AGA GGC AGG GCC GTG C	43	DiGIR1D15'' (antisense)	
C515	GCA ATC GGG TTG AAC ACT TAA GTG TTC TTG GGT TAA AAC GGT GGG GGA	48	DiGIR1insP15'' (sense)	
C516	TCC CCC ACC GTT TTA ACC CAA GAA CAC TTA AGT GTT CAA CCC GAT TGC	48	DiGIR1insP15'' (antisense)	
C557	CCT AAG CGC CCG GAC GGG CGT ATG GCC GTA ACA TCC GTC CTA A	43	Insertion of Azo P5a into GIR1 5'oligo	
C558	CCC ACC GTT TTA ACC CAA	18	Insertion of Azo P5a into GIR1 3'oligo	
C559	CGG GAG GCG AAA GCC CCG GGA AGC ATC CGG TAT CC	35	Insertion of Azo P9 into GIR1 5'oligo	
C560	GGA ACG ACT GTT CAT TGA AC	20	Insertion of Azo P9 into GIR1 3'oligo	
C594	CCG TAA CAT CCG TCG ACA GAC T	22	GIR1 for deletion of P6	
C595	CCC ACC GTT TTA ACC CAA TTA	21	GIR1 for deletion of P6	
C596	GAT GAA GGT CGA CAG ACT GCA CGG CCC T	28	GIR1 insertion of Azo P6 ok	
C597	GGC GCA GGC GCC GAA GCT TGG CAG GGA TGT TAC	33	GIR1 insertion of Azo P6 mut P4	
C598	TAA TAC GAC TAC CTA TAG ATC CGG TAT TCG AAT CGG GTT GAA CAC CTT AA	50	GIRAZO P1-Wrong T7	P1.
C599	GGC GCA GGC GCC GAA GCT TGG CAC GGA TGT TAC	33	Azo P6 (1) corrected	
C600	TAA TAC GAC TAC CTA TAG GTC CGG TAT TCG AAT CGG GCT GAA CAC CTT AA	50	GIRAZO P1-mut1-Wrong T7	P1.
C601	TAA TAC GAC TAC CTA TAG GTC CGG TAT TCG AAT CGG GCC GAA CAC CTT AA	50	GIRAZO P1-mut2- Wrong T7	P1.
C602	TAC CGG ATG CTT CCT TTC GGA ACG ACT GTT	30	GIR1 IPS.6	.6WT
C603	GGA CCG AAA TCC TTA GTA CGG ATG TTA C	28	Nae-P6ins (1)	
C604	GGA ACG TCG ACA GAC TGC ACG GCC GT	26	Nae-P6ins (2)	

C605	GTG TTC AAT GAA TCG TTC C	19	GIR1 P8 deletion (1)	
C606	AGG GCC GTG CAG TCT GTC T	19	GIR1 P8 deletion (2)	
C607	CAC CAT TGC GTT TGC CTT AGG TGA GGG CCG TGC AGT CTG TCG	42	Azo P8 ins (1)	
C608	AAT GAA CAG TCG TTC C	16	GIR1 for 6-nt transposition	
C609	AAT TTA ATA CGA CTC ACT ATA GAT CCG GTA TTC GAA TCG GGT TGA ACA CCT TAA		P1 T7 promoter OK	P1.
C610	AAT TTA ATA CGA CTC ACT ATA GGT CCG GTA TTC GAA TCG GGC TGA ACA CCT TAA		P1-mut1 promoter OK	P1.
C611	AAT TTA ATA CGA CTC ACT ATA GGT CCG GTA TTC GAA TCG GGC CGA ACA CCT TAA		P1-mut2 promoter OK	.6WT
C614	CCA TAG CGT TTG CCT TAG GCA GGG CCG TGC AGT CTG TCG		Azo P8 ins (1) for Peripheral element add	
C615	ATT TAA TAC GAC TCA CTA TTA GGG AAG TAT CAT		DiGIR1; ApG T7prom + 5'end (157)	157.
C616	ATT TAA TAC GAC TCA CTA TTA GTT GGG AAG TAT CAT		DiGIR1; ApG T7prom + 5'end (160)	160.
C617	ATT TAA TAC GAC TCA CTA TTA GGG TTG GGA AGT ATC AT		DiGIR1; ApG T7prom + 5'end (162)	162.
C618	ATT TAA TAC GAC TCA CTA TTA GGT TGG GTT GGG AAG TAT CAT		DiGIR1; ApG T7prom + 5'end (166)	166.
C619	CCG TCA CCA TGG TTG TTG AAG TGdC.	24	Oligo for 3' end labeling .65 GIR1 with Klenow enzyme !!	
C620	AAT TTA ATA CGA CTC ACT ATA GGT CCC TGT TAT TGA GGA C	40	Npr 5'-oligo for T7-191.28	191.
C621	TTT TAT GGT TAC CAT TTT GTA	21	Npr 3'-oligo for T7-191.28	.28
OP878	TCA CCA TGG TTG TTG AAG TGC ACA GAT TGG TAT CCG GAG ATT TGA TTG TCT TGG GAT	57	GIR1;mut of hairpin I ("UTR1")	mut UTR1
OP879	TCA CCA TGG TTG TTG AAG TGC ACA GAT TTC ATA GGA ATC TTT TGA TTG TCT TGG GAT ACC	60	GIR1;mut of hairpin I ("UTR2")	mut UTR2
OP619	GGA TGC TTC CTT TCG GAA	18	GIR1;3'deletion G1-.5	.5
OP620	ACC GGA TGC TTC CTT TCG	18	GIR1;3'deletion G1-.8	.8
OP78	GAT ACC GGA TGC TTC CTT	18	GIR1;3'deletion G1-.11	.11
OP315	CTT GGG ATA CCG GAT GCT TCC TTT	24	GIR1;3'deletion G1-.16	.16
OP233	GAT TGT CTT GGG ATA CCG	18	GIR1;3'deletion G1-.22	.22
OP353	TTA GAT TTG ATT GTC TTG	18	GIR1;3'deletion G1-.30	.30
OP314	GGT ATC CTT AGA TTT GAT TGT CTT	24	GIR1;3'deletion	.37

			G1-.37	
OP235	GCA CAG ATT GGT ATC CTT	18	GIR1;3'deletion G1-.46	.46
OP12	TCA CCA TGG TTG TTG AAG TGC ACA GAT TG	29	GIR1;3'deletion G1-.65	.65
OP233	GAT TGT CTT GGG ATA CCG	18	GIR1;3'deletion G1-.22	.22-mut
OP233	GAT TGT CTT GGG ATA CCG	18	GIR1;3'deletion G1-.22	.22-mut

	Use for PCR for transcription templates preparation
	Use for 3' end labeling of RNA by using a DNA oligo modified in 3' end with the Klenow fragment of the DNA polymerase 1

REVIEW II:

Exploring RNA structure by integrative molecular modelling

B. Masquida, B. Beckert, F. Jossinet

New Biotechnology, Volume 23, Number 3, July 2010

RESEARCH PAPER

New Biotechnology • Volume 27, Number 3 • July 2010

ELSEVIER

Exploring RNA structure by integrative molecular modelling

Benoît Masquida[1], Bertrand Beckert[2] and Fabrice Jossinet[1]

[1] Architecture et Réactivité de l'ARN, Université de Strasbourg, IBMC, CNRS, 15 rue René Descartes, 67084 Strasbourg, France
[2] Department of Cellular and Molecular Medicine, The Panum Institute, University of Copenhagen, Copenhagen, Denmark

RNA molecular modelling is adequate to rapidly tackle the structure of RNA molecules. With new structured RNAs constituting a central class of cellular regulators discovered every year, the need for swift and reliable modelling methods is more crucial than ever. The pragmatic method based on interactive all-atom molecular modelling relies on the observation that specific structural motifs are recurrently found in RNA sequences. Once identified by a combination of comparative sequence analysis and biochemical data, the motifs composing the secondary structure of a given RNA can be extruded in three dimensions (3D) and used as building blocks assembled manually during a bioinformatic interactive process. Comparing the models to the corresponding crystal structures has validated the method as being powerful to predict the RNA topology and architecture while being less accurate regarding the prediction of base–base interactions. These aspects as well as the necessary steps towards automation will be discussed.

Contents

Introduction

The extensive use of high-throughput methods such as tiling arrays has modified our view of the transcriptome [1]. The observations that the transcribed portion of some mammalian genomes is more important by about one order of magnitude than the translated portion [2] and that the length of 3′ UTRs is correlated with organism complexity [3] illustrate the broad biological perspectives of the RNA field. RNA is everywhere in cells having key roles at every step from replication to translation and control of genetic expression through mechanisms that had not even been imagined just 10 years ago [4–6]. Moreover, since RNA

Corresponding author: Masquida, B. (b.masquida@ibmc.u-strasbg.fr)

New Biotechnology • Volume 27, Number 3 • July 2010 REVIEW

mediates its function through its structure and since sequences accumulate much faster than crystal and NMR structures are solved, RNA modelling methods undergo a new impetus for development.

RNA polymers weave specific networks of non-bonded interactions to fold and finally form their functional 3D structure in a hierarchical process. The folding schematically proceeds through the entropically driven stacking of base rings that leads to the exclusion of water molecules from the RNA chain. This process is coupled to the transcription and can be viewed as a molecular collapse which is promoted by ions that counteract the electrostatic repulsive effects due to condensing the negative charges brought by the individual phosphate group of each nucleotide by specific or diffuse interactions with RNA sites [7,8]. Condensing of the RNA can be promoted by proteins as is the case for ribosomal RNAs [9] or by RNA chaperones [10]. Within this complex network of interactions, most of the specific contacts present in the condensed folded RNA are composed by hydrogen bonds. They result in base-pairs that stack or top of each others to form an assembly of Watson–Crick (WC) helices interspersed by single-stranded regions represented under the form of a tree-like secondary structure (2D) diagram (Figure 1). The apparently single-stranded regions also form hydrogen bonds that lead to tertiary contacts. Because these interactions take place out of the helical context, they usually form non-WC contacts and contribute ~25% of the total hydrogen bonds [11]. These are crucial for the native 3D structure and subsequently for the RNA function and dynamics. Single-stranded regions (terminal loops, internal loops, junctions between helices) very often form specific motifs associated with specific sequence signatures that are recurrently found in unrelated RNAs. Under an adaptive selection scenario, the RNA sequence undergoes neutral sequence drift [12] so as to preserve the structure and therefore RNA function.

In proteins, sheets and helices are stabilized by hydrogen bonds mainly mediated by the chemical groups from the backbone

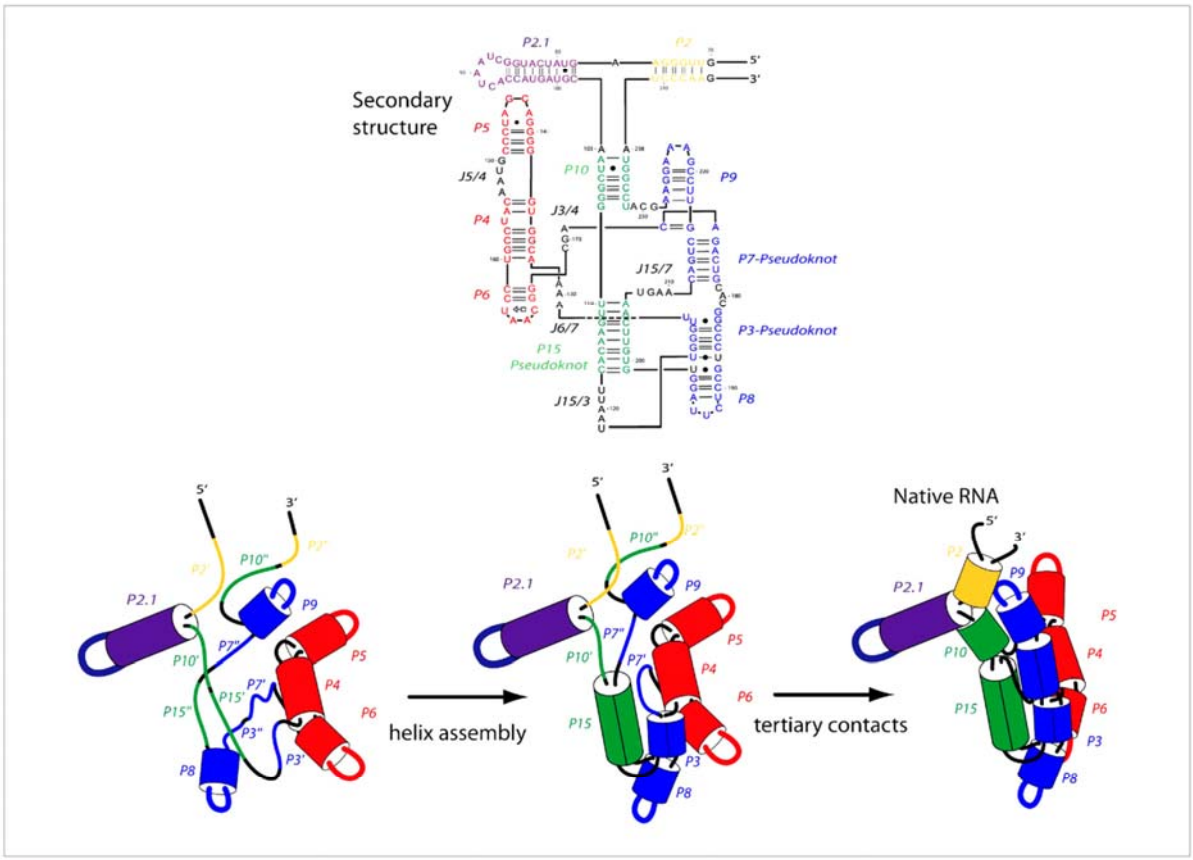


FIGURE 1

A view of the RNA folding process of the GIR1 ribozyme. The folding is initiated by the recognition between complementary strands within the RNA chain in the course of the transcription. This can lead to the formation of transient helices that are resolved upon completion of the transcription. Ions and formation of RNA motifs promote spatial proximity between segments that can then interact (formation of P3 and P15). Finally, the formation of tertiary interactions induces a higher compaction of the RNA that is then ready to accomplish its function. The secondary structure corresponding to the native state of the RNA is indicated in the upper panel.

Confident deduction of the secondary structure of a protein from a sequence alignment without prior knowledge of the structure of at least one of the members is thus a difficult exercise [13]. On the contrary, nucleotide identity in RNA is related to interaction specificity. This is a necessary and sufficient condition to deduce the secondary structure from the sequence. The secondary structure formed by RNA homologs can thus be deduced by comparative sequence analysis (CSA) using the rules governing the formation of WC base-pairs [14]. Covariation analysis facilitates the identification of long-range tertiary interactions such as pseudoknots that consist in WC base-pairs between residues from a loop and a *cis* remote 5′ or 3′ single-stranded segment [15–17]. Moreover, since RNA sequences are evolutionary not only selected for their structural information, modelling is also appropriate to address questions related to folding, stability, and catalytic properties (ribozymes). Building complexes involving RNA and proteins and/or small organic ligands can also be attempted. Modelling is particularly indicated to study conformational intermediates that are by essence difficult to isolate and crystallize [18].

RNA modelling methods are numerous. They are either based on conformational space searching [19–22] or on interactive modelling [23]. This review focuses on the interactive and semi-automatic all-atom molecular modelling method originally developed to model the catalytic core of group I introns [24], which is known as the MANIP package [25]. The general ideas and concepts on which the method is based [23,26] will be discussed as well as the recent technical improvements undertaken to make the method more integrative towards the enormous amount of biochemical and structural data available today.

From secondary structures to 3D models

Identifying the secondary structure using bioinformatics and biochemical methods

Identifying the secondary structure of an RNA family constitutes a vast research field by itself. The methods implemented are based on thermodynamic or comparative approaches or combine both. Pioneering thermodynamic approaches using dynamic programming like Mfold [27] and RNAfold [28] predict secondary structures for a unique RNA sequence without pseudoknot. The main drawback is that folding of any single RNA sequence may not produce the biologically active RNA structure. Moreover, the strong occurrence of pseudoknots in RNA structures underlines the intrinsic weakness of these programs hence fostering the development of algorithms capable of their prediction like Kinefold [29] or others [30,31]. For RNAs longer than few tens of nucleotides (with a higher probability of pseudoknot occurrence), more reliable results can be obtained by comparative approaches exploiting sequence and structure conservation. They take into account the strong correlation existing between columns of interacting nucleotides from a sequence alignment explicitly represented as base-pairs in the RNA secondary structure. In a first attempt, the construction of such 'structural alignments' can be done automatically using different approaches (for a recent review, see [32]). Some tools align secondary structures precomputed independently for each RNA sequence [33–35]. Others try to fold and align RNA sequences simultaneously using the Sankoff algorithm approach [36–39]. Starting from curated 'seed' alignments, it is also possible to deduce probabilistic profiles (a.k.a. covariance models (CMs)) allowing to search and align new RNA candidates from sequence databases. The Rfam website provides such alignments for any of the well-known RNA families [40]. Unfortunately, due to computational limitations, the structural alignment produced by these tools can rarely be used *per se*. The identification of the consensus secondary structure needs generally a second step performed by hand using human expertise. Several tools have been released recently to leverage this iterative task (4SALE [41], ConStruct [42], S2S [43], SARSE [44]). If curated secondary structures or solved tertiary structures are available for a given RNA family, tools like S2S allow to infer a new secondary structure for an orthologous sequence of the same family. Curated secondary structures are available from databases like the 'Comparative RNA Website' [45] or RNA STRAND [46]. Solved tertiary structures are available from the Nucleic Acid Database [47]. Their usage needs a previous annotation step done by algorithms like RNAVIEW [48] or MCAnnotate [49].

Secondary structures can be refined experimentally. Specific enzymatic and chemical probes are capable of discriminating between rigid/flexible and/or single-stranded/double-stranded regions. RNA flexibility can be simply assessed by in-line probing [50]. Phosphodiester bonds in the flexible regions (loops, junctions) present increased sensitivity to hydrolysis because the nucleophilic O2′ group is statistically more prone to be placed in-line with the phosphorus atom and the O5′ leaving group than in conformationally restricted helical regions. Backbone flexibility can also be investigated by selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) [51] to determine nucleotides forming helices. Secondary structure-specific enzymes can also be employed. RNase V1 is specific of double-stranded regions and other RNases are specific of unpaired nucleotides yet sequence specific, for example RNase T1 cleaves 3′ of unpaired G residues [52]. Base-specific chemical probes are also very useful. They react mostly with chemical groups that are accessible to the solvent leading to conclude about the interactions in which they participate. Chemical probes also help detecting tertiary interactions like pseudoknots.

RNA architectures are built from recurrent structural motifs

An RNA motif can be defined as a recurrent RNA element characterized by a set of specific RNA–RNA interactions. The most common RNA motif is the regular A-form WC helix. A variety of distinct non-WC motifs in the single-stranded regions of the secondary structure are associated with sequence-specific signatures. In the absence of NMR or crystal structures for these motifs, the base-pair geometries governing their formation are difficult to elucidate solely from sequence alignments. Indeed, the three edges of the nucleotides (WC, Hoogsteen (H), sugar (S)) can interact resulting in six possible pairs with anti-parallel orientation of the riboses and six possible pairs with parallel ribose orientation [53], a situation significantly more complex than for the WC pairs. These twelve base-pairs described in the so-called Leontis–Westhof (LW) nomenclature (Figure 2) circumscribe the structural diversity of RNA base-pairs. The anti-parallel pairs are naturally adaptable to helix ends whereas parallel ones imply additional unpaired

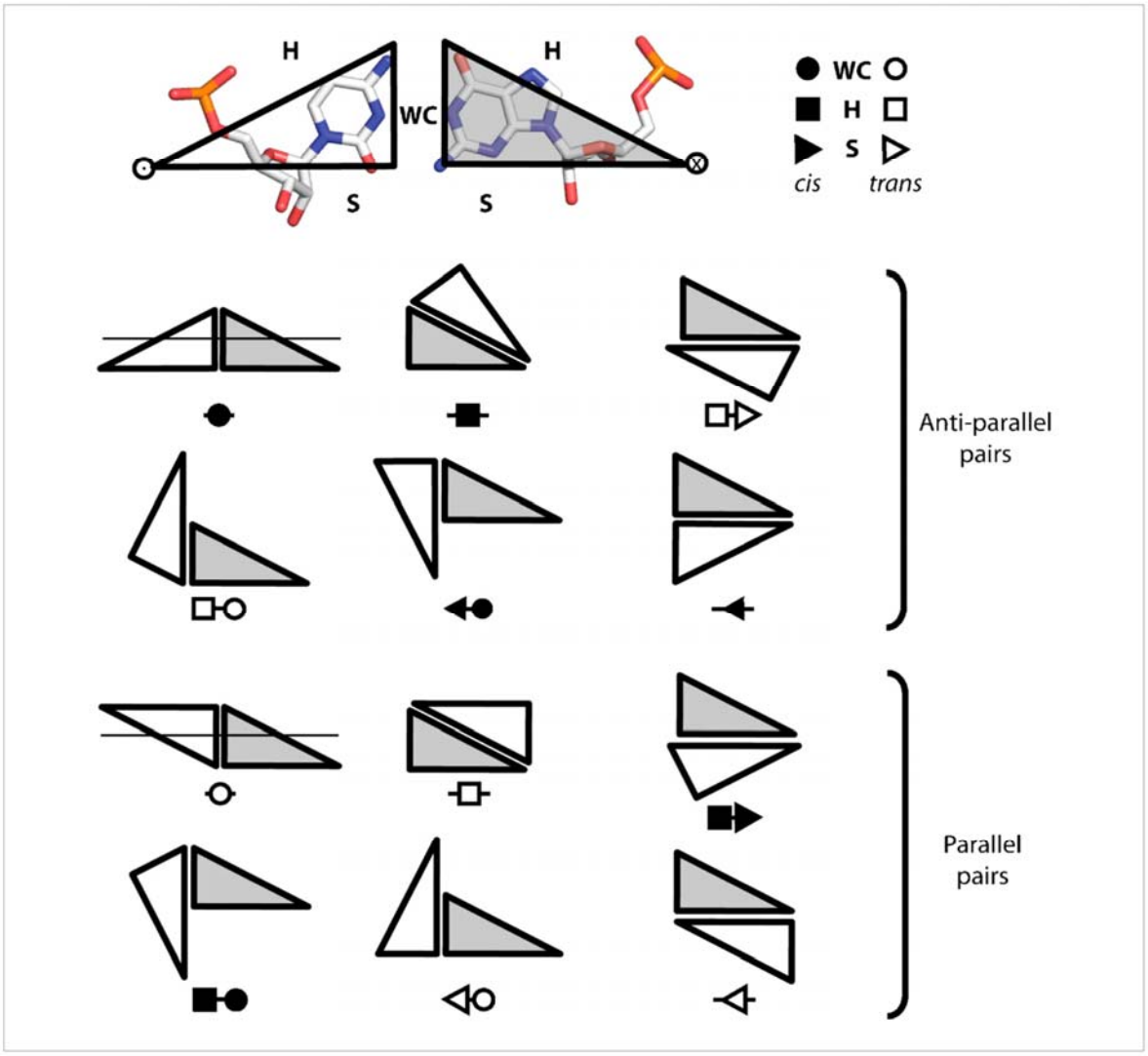


FIGURE 2

Representation of the limited base-pair diversity schematized according to the LW nomenclature [51]. The 12 interaction types can be clustered into anti-parallel and parallel pairs. Pairs are *cis*/*trans* when the position of the ribose rings of the two nucleotides involved are located on the same/opposite sides of a line parallel to the hydrogen bonds, respectively. Anti-parallel pairs are likely to participate in pseudo-secondary structure motifs tethering anti-parallel helices whereas parallel pairs are by essence more prone to participate in tertiary structure motifs with a notable exception for the loop E and reverse-kink-specific *trans* HH. Parallel pairs can be reverted to anti-parallel and vice versa if the sugar is reoriented in *syn* with respect to the base instead of *anti* as depicted on the top panel.

nucleotides in order to compensate for the local reversal of the backbone polarity or a *anti* to *syn* shift for the torsion between the base and the ribose. Consequently, parallel pairs are more prone to build tertiary interactions than to be part of structure motifs embedded within helices. Sequence of RNA base-pairs is also limited by base-pair isostericity [54,55]. Isostericity of non-canonical pairs suggests covariation rules that are helpful to grab the sequence diversity authorized for each base-pair taking place in a given recurrent motif [56]. In a 5′ *ij*…*kl* 3′ strand, the *i*–*l* base-pair can be tethered to the *j*–*k* base-pair only if the distance between the phosphate group of *j* and the O3′ group of *k* is similar to the distance between the O3′ group of *i* and the phosphate group of *l*. This condition is always true in WC base-pairs but often wrong for non-canonical pairs. This property means that by constraining the ribose-phosphate backbone, non-canonical pairing interactions control the base-pairs that can be associated to build RNA motifs. Consequently when necessary, bulge residues achieve some flexibility in non-canonical motifs in order to release constraints on

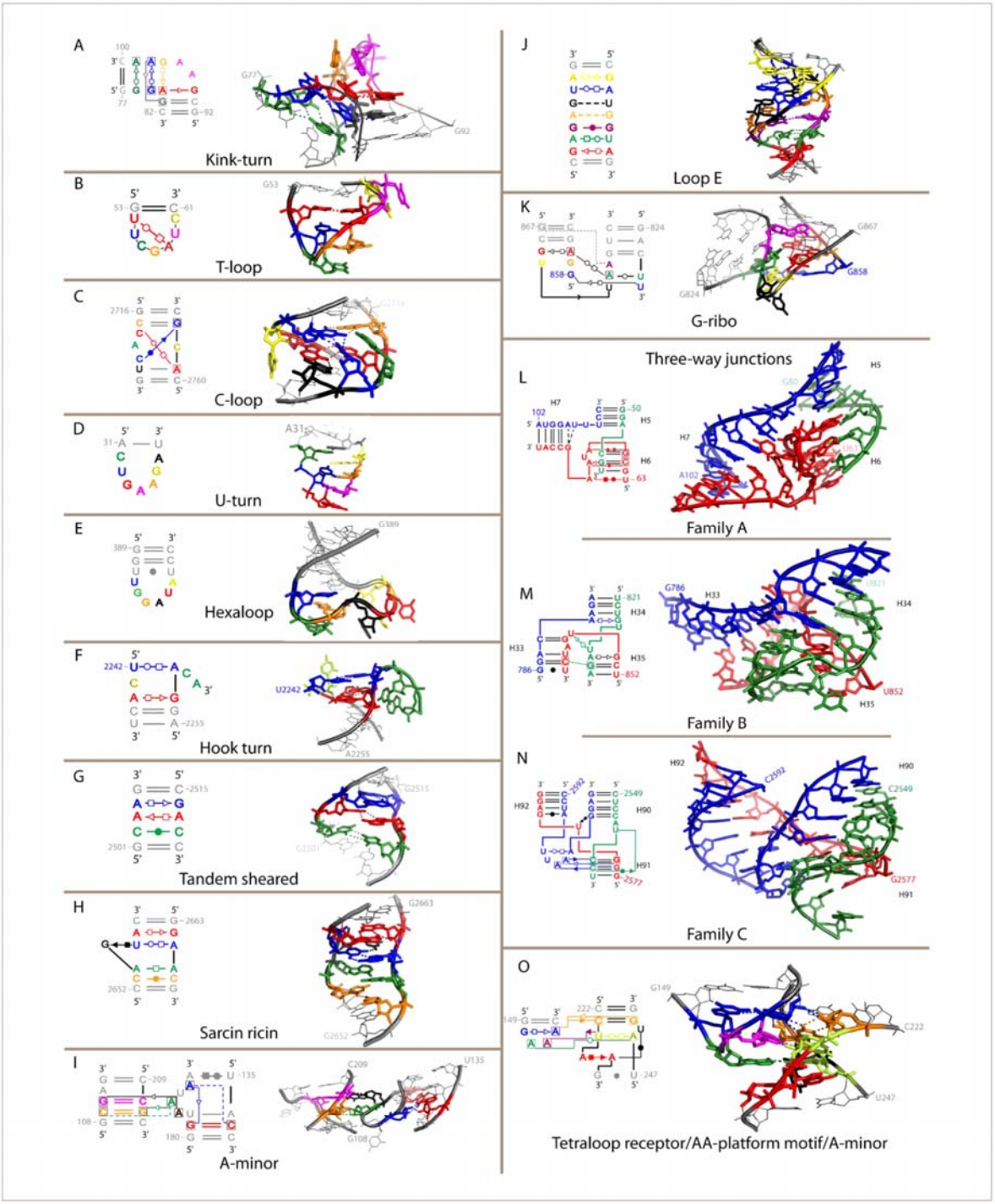


FIGURE 3

View of the secondary and 3D structure from various motifs recurrently found within RNA structures. PDB codes are indicated. (A) Kink-turn (1JJ2 [64]), (B) T-loop (1EHZ [143]), (C) C-loop (1S72 [144]), (D) U-turn (1EHZ), (E) hexaloop (1S72), (F) hook-turn ([145] 1JJ2), (G) tandem sheared (1S72), (H) eukaryotic loop E or sarcin ricin (480D [146]), (I) A-minor (1GID [66]), (J) bacterial loop E (354D [147]), (K) G-ribo (2AVY [135]), (L) three-way junction (3WJ), (1S72), family A (1S72), (M) 3WJ, family B (1S72), (N) 3WJ, family C (1S72), (O) tetraloop receptor (1GID).

the backbone. Another way of circumventing backbone constraints is to change the puckering of the ribose ring, or to change the conformation of the ribose with respect to the base from *anti* to *syn* (Figure 2). Non-canonical motifs are thus associated with a significant occurrence of C2′ endo conformation as opposed to the usual C3′ endo typical of the A-form helix. Moreover, the participation of the O2′ group or of the phosphate group in H-bonds involving the S edge or the H edge of the nucleotides, respectively offers stabilizing interactions unseen in WC pairs. NMR or crystal structures of RNA thus provide the necessary link to decipher the selection pressure that a sequence undergoes to form a given structural motif.

Every complex RNA crystal structure unravels new motifs (see Figure 3 for detailed description of motifs that have been mainly described in the literature), giving scientists opportunities to improve sequence alignments, refine RNA secondary structures and finally build more accurate 3D models. Motif diversity is expected to grow with more and more solved RNA structures, although motif diversity should be limited because the structural diversity resulting from their assemblage in complex 3D architectures is expected to offer combinatorial solutions to any RNA function.

The first category encompasses motifs folding locally. One of these motifs, the loop E, was originally found in the eukaryotic 5S ribosomal RNAs [57]. Sequence analysis later on showed its widespread occurrence in rRNAs [58] which was later confirmed by the crystal structures of the 50S and 30S subunits [59,60] and reported in models of other RNAs [61–63]. Since the loop E is sandwiched between two regular helices, it constitutes a pseudo-secondary structure motif. Other examples of such kind encover the kink-turn [64], and the C-loop which have been phylogenetically analyzed [65]. The loop E and the C-loop preserve the coaxiality of the flanking helices whereas the kink-turn mediates a 120° kink between them. A sharper kink motif obeying distinct sequence requirements has also been observed in the crystal structure of the P4-P6 domain of the Tetrahymena group I ribozyme [66]. Regarding multiple way junctions, specific motifs direct the formation of three distinct clusters of three-way junctions characterized by the respective orientation of the three helices they are composed of as well as the length and sequence of the linkers [67]. The G-ribo motif [68,69] is one of these motifs, albeit also found in kinks mediated by 2-way junctions and pseudoknots. A recent survey of the organization of four-way junctions characterized by the coaxial stacking patterns identifies nine distinct structural families [70]. The UA handle [71] and the S-turn [72,73] intervene as minimal building blocks or can be submotifs of larger motifs. The UA-handle is for example a submotif of the T loop. The T loop is characteristic of the tRNA [74–76] but has been identified as a key element from the RNase P ribozyme [77,78].

Local motifs enforce the position of helices with respect to each others upon folding, hence programming long-range interactions leading to the formation of tertiary structure motifs. They also constitute platforms for interacting with proteins [64,79,80]. An archetype of these long-range interacting composite motifs consists in the interaction between a GNRA tetraloop and its 11-nt receptor [66], although the loop E and the Kink-turn are also known to weave tertiary interactions. It is worth to note that GNRA, UNCG and CUYG tetraloops constitute a class of very stable apical loops that help forming the secondary structure by indicating preferred kinking points of the RNA chain [81]. A fourth class of tetraloop (GANC) has been recently reported in the group II intron crystal structure [82] where it seems to have specifically evolved to build a critical tertiary contact [83]. GNRA tetraloops are versatile since the sugar edge of at least two consecutive A residues can interact with the shallow groove of helices containing G=C pairs [84] to form a motif called A-minor [85]. A-minor interactions really glue the RNA chain by forming hydrogen bonds between the sugar edge of at least two consecutive A residues in the shallow groove of helices. The A-minor motif thus constitutes a submotif of the GNRA tetraloop or of the cross-strand AAA stack [86] where A residues mediate identical interactions with WC pairs, albeit in different general contexts. Yet, the A-minor motif can also originate from any RNA motif presenting a patch of consecutive adenosines as in the A-rich bulge of the P4–P6 domain of the Tetrahymena ribozyme [87]. Strikingly, the A-minor motif is responsible for the decoding of the codon–anticodon helix by the ribosomal A-site [88].

Recently, a systematic method has been developed to identify motifs by an algorithmic clustering approach based on the LW nomenclature [89]. Interestingly this approach did not only identify already known motifs but also suggested putative new motifs. Another approach based on a symbolic 3D query allows for retrieving occurrences of known motifs [90]. Although this approach is obviously not intended to find new motif candidates, it is well suitable to find composite motifs built from three or more strands. These two approaches are thus complementary. Several initiatives exist to find, store and diffuse the recurrent patterns observed in solved RNA structures. At present, the FR3D project [90] is the only one to provide data compatible with the definition of RNA motifs exposed above. Others like SCOR [91] or RNA-As-Graphs [92,93] are more focused on recurrent topological features. The DARTS database provides a clustering of RNA structures on the basis of the identification of highly identical fragments [94]. The NCIR database lists all the non-canonical interactions found in RNA structures [95].

Importantly, the recurrency of RNA motifs suggests they constitute individual building blocks that can be used individually to assemble large RNA architectures. This assembly is programmed by local pseudo-secondary structure motifs that guide long-range interactions. A proof of concept of this idea comes from the characterization of artificial RNAs designed by molecular modelling to incorporate a subset of these motifs to control their structure and oligomerization [96].

The role of the O2′ in base-pairing

Involvement of at least one S edge accounts for around two thirds of the non-canonical base-pairs stressing the preponderant role of the O2′ group of the ribose [55]. 27% are owing to S–H interactions typically found in G–A pairs formed within pseudo-secondary structure motifs like GA tandems or isolated G–A or A–A pairs [97,98] or AA-platforms [99]. 18% form WC–H pairs and are also typical of pseudo-secondary motifs. Amazingly, 46% involve the S edge of one residue with the S or WC edge of the second. These are typical of tertiary contacts involving the shallow groove of a helix

with a facing motif such as the A-minor motif. In the 50S ribosomal subunit crystal structure [59], the A-minor motif accounts for nearly one third of all adenosine doublet both involved in tertiary contacts and 90% conserved [87]. When considering non-WC tertiary interactions, the geometry of the A-form RNA helix mostly favors the approach of nucleotides on the shallow/minor groove of helices. The resulting H-bond interaction network systematically involves participation of the O2′ groups of the facing residues as well as the N3(R)/O2(Y). Since these chemical groups are always present in nucleotides, detection of shallow groove tertiary interactions solely by comparative sequence analysis is made very challenging.

RNA modelling workflow

Interactive model assembly proceeds through the following steps. After careful analysis of the secondary structure of the RNA, individual secondary structure elements are extruded in 3D using the MANIP package [25]. Any sequence can be superimposed onto a ribose-phosphate backbone corresponding to a given motif. Helices and non-canonical motifs are then assembled interactively on the computer monitor by rotating and translating domains or individual nucleotides or by using the torsion angles. Once a working model has been successfully assembled, several cycles of geometrical least-square refinement are performed in order to avoid nucleotides with deviating geometries that could favor wrong models [100]. At this step H-bonds are explicitly defined as constraints. When regions of the model resist geometric refinement, interactive modelling is resumed and the whole process iterated until the models show no more stereochemical outliers. During the whole process, the working models are confronted to experimental data that are useful to both guide the modelling and ultimately validate the final model.

Several remarks arise from this workflow. (i) Superimposition of crystallographic occurrences of each motif shows they can be considered as rigid building blocks. Consequently, this strategy simplifies the modelling problem to the assembly of these building blocks without considering modelling their dynamical properties but rather emphasizes how they cluster and orient one to each others. (ii) Interactive modelling can suggest the overall topology (the spatial organization of single-stranded junctions) and architecture of the RNA (the way helices stack in multi-branched junctions) if it is fairly long and constrained by tertiary interactions like pseudoknots. In this case, exploring the conformational space of the RNA can be summarized as finding the rare conformations that allow satisfactory tethering of all the secondary structure elements. Hence, there is a direct relationship between the degree of constraints of the secondary structure and the uniqueness of the final model. (iii) Interactions made by single-stranded nucleotides of the secondary structure are difficult to assign. However, satisfaction of the spatial constraints (essentially pseudoknots) often leads to suggesting tertiary contacts between single-stranded regions that can be assembled in non-canonical motifs that were not apparent from the secondary structure. (iv) The power of the method relies on the fact that all-atom modelling integrates the knowledge of nucleic acid stereochemistry. Stereochemical rules guarantee models with structural significance owing to the implementation of intrinsic constraints that superimpose to the secondary structure independently of any other data. (v) The integration of biochemical data is very helpful to locally make the model more accurate by suggesting interactions between nucleotides and/or suggesting pairing status, especially when considering modelling of motifs with unknown structure.

Towards automation of the process

Since several years, the development of an integrated bioinformatics framework dedicated to the study and the construction of RNA architectures has been pursued in the laboratory. This initiative can be seen as equivalent to the Bio* toolkits used to construct bioinformatics tools and databases focused mainly on sequence and genomic data [101]. The features of the framework can be divided into two main categories: (i) the computational representation of the biological objects needed to study and construct RNA structures (helices, single-strands, pseudoknots, base-pairs, motifs) and (ii) a communication layer to consume remote RNA algorithms hosted as Web Services [102,103]. Two graphical tools have been developed on this framework (Figure 4). S2S improves the construction of RNA structural alignments [43] (http://bioinformatics.org/s2s), and Assemble is an RNA 3D modeller (http://bioinformatics.org/assemble).

Assemble provides two windows dedicated to the construction of the two main components of an RNA architecture: a first one contains a 2D panel displaying the secondary structure and a second one renders the 3D fold (Figure 4). These windows are interconnected, meaning that a selection done in a first one will highlight its counterpart in the second one. Within the 2D panel, the user can alter the secondary structure to fit his presumptions and/or experimental data. The helices and single-stranded regions define the building blocks that can be exported in the 3D scene. In a first step, they are exported with a default helical fold. Then, the 3D scene provides two options: (i) the ability to modify the folding of the building blocks and (ii) the ability to reorganize them in the 3D space to produce the overall shape of the 3D model. The 3D model can be altered interactively by modifying the torsion angles for any single residue or semi-automatically by applying to a set of residues a local fold observed in a solved 3D structure. Assemble provides functionalities to extract and store such folds.

Since several structural inconsistencies can appear within the 3D model, a geometrical refinement of the model can be launched from Assemble. The geometric constraints used for this refinement are automatically deduced from the set of base-pairs defined in the 2D panel. In contrast to the previous tool MANIP, the modelling process with Assemble is done exclusively through its graphical interface. Some tasks delegate the work to external algorithms that are called transparently over the network. Their results are integrated and displayed in the current working session.

Comparison between models and crystal structures

The ultimate way of validating a model is to compare it to a homologous crystal structure. In this section, approaches to compare RNA models to corresponding crystal structures are presented. The simplest method consists in determining the normalized root

New Biotechnology • Volume 27, Number 3 • July 2010 **REVIEW**

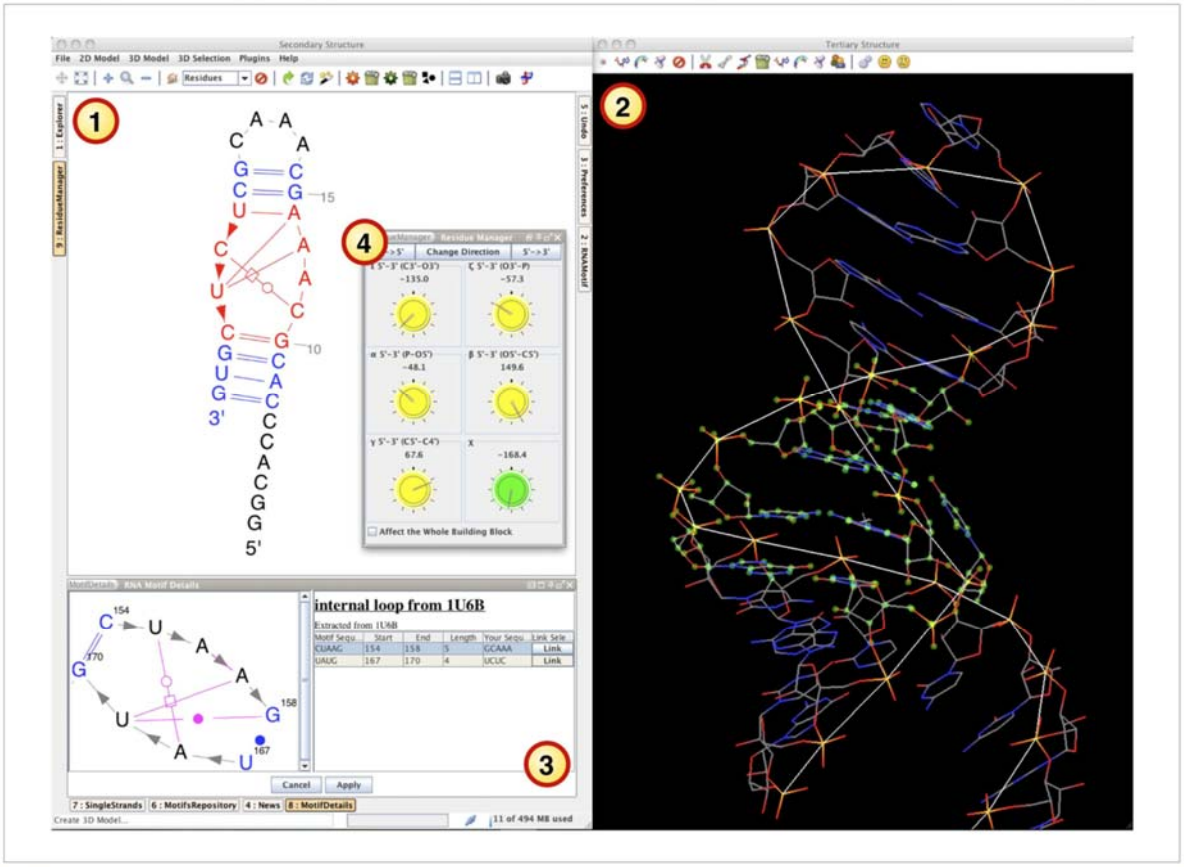


FIGURE 4

A screenshot of the graphical interface of Assemble (http://bioinformatics.org/assemble). The 2D panel (1) is dedicated to display and edit the secondary structure which constitutes the scaffold of the 3D model. The structural domains (helices or single-strands) are exported to the 3D panel (2) with a default helical fold. Several options are available to the user to reorganize them in the 3D space. Their folding can be fitted to the RNA sequence peculiarities by applying an RNA motif previously observed in a crystal structure and saved (3). The folding of this motif will be reproduced in the 3D scene. The base–base interactions stabilizing it will be automatically added to the secondary structure using the Leontis–Westhof classification. At any time during the construction process, the model can be improved by geometric refinement. The structural constraints of this refinement will be automatically deduced from the base-base interactions described in the 2D panel. Finally, the user can also manually alter the torsion angles of any residue selected in the 3D scene (4).

mean square deviation (*nrmsd*) [104] between the model and the crystal structure using Lsqman [105]. Normalization prevents the *rmsd* value to increase with the size of the molecule hence allowing the comparison of values obtained for RNAs of different size. As a global value, *nrmsd* tells information on the average distance between atom pairs but gives no information about base-pair prediction accuracy. New indicators have been described recently to circumvent this problem [106]. However, when *nrmsd* values are less than 10 Å (the distance roughly corresponding to the half of the A-form helix width), they still constitute a valid indicator to verify if the various structural elements of the overall model occupy spatial locations in agreement with the crystal structure. Several RNA models have been built using the MANIP package and published before the release of the corresponding crystal structures (Table 1).

Commenting on the 3D model of the *Azoarcus* pre-tRNA^{Ile} containing a group I ribozyme [107] well illustrates the necessity to integrate biochemical data. Since this model is representative of the models generated by the workflow described above, it will be the only one discussed here. Similar observations are applicable to other models from the laboratory. When dealing with group I intron modelling, the first task is to pinpoint the G-cofactor binding site. In the case of the *Azoarcus* ribozyme, it was modelled accordingly to the original 3D model of the group I intron [24] as lying in the deep groove (H edge) of the second G=C pair of the P7 stem (Figure 5A and B). This was demonstrated by phylogenetic and genetic evidence [108]. On the opposite, the absence of data related to the position of the P7 bulge (A129) prevented its correct modelling. Nonetheless, the *nrmsd* of 3.85 Å between the model and the crystal structure of the group I

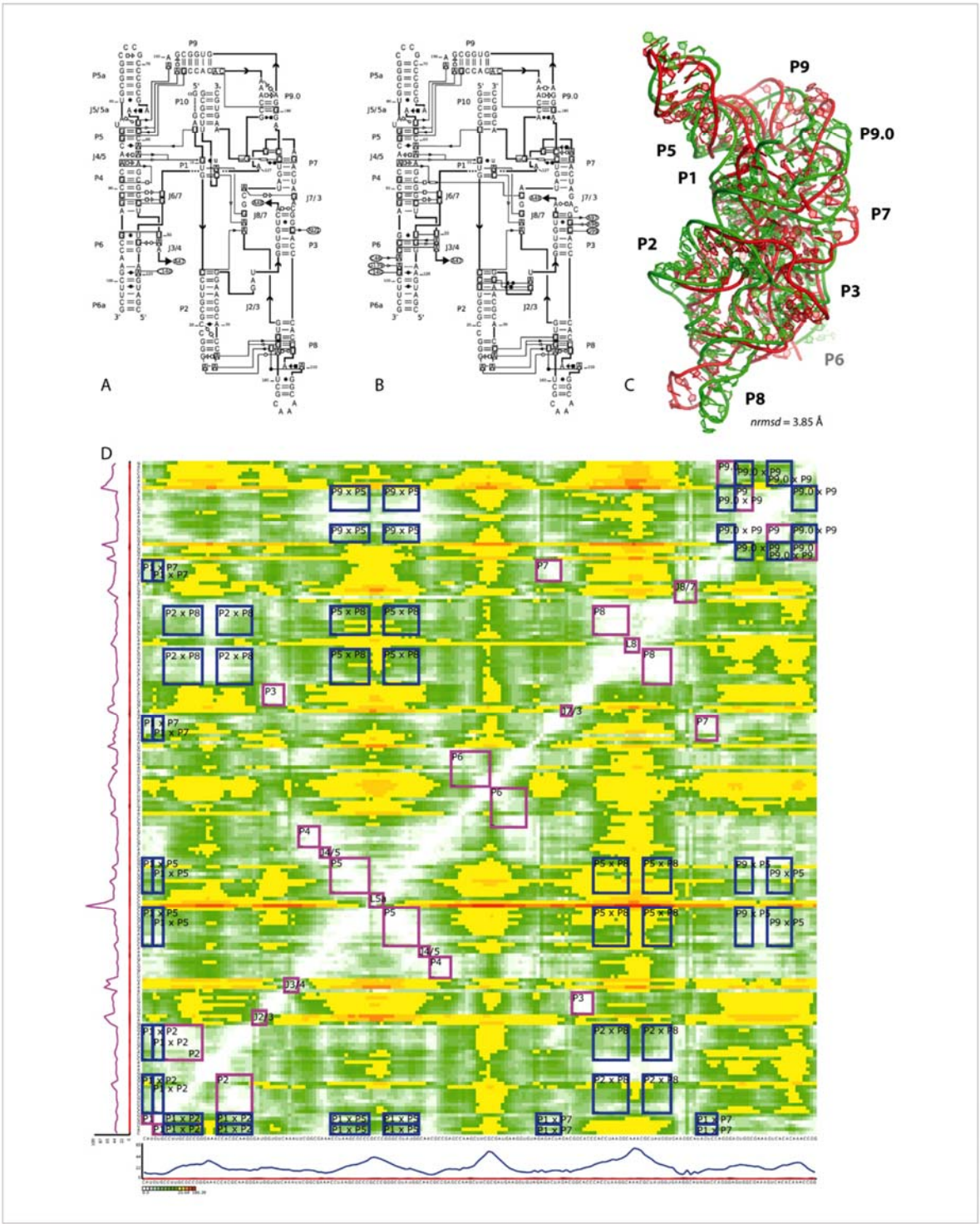


FIGURE 5

Comparison between the model and the crystal structure of the group I ribozyme embedded in the anticodon loop of the pre-ARNt^{Ile}. (A) secondary structure of the RNA structure model. (B) Secondary structure of the RNA crystal structure. The LW nomenclature specific to each secondary structure diagram is indicated by

TABLE 1

Summary of 3D models for which a corresponding crystal structure exists

RNA	Model reference	X-ray structure reference
Group I intron	[24,115–117]	[109,113,114,118]
U1 snRNA	[119]	[120]
5S rRNA	[121]	[59]
Hammerhead ribozyme	[122]	[123,124]
Hepatitis delta virus ribozyme	[125]	[126]
Hairpin ribozyme	[61]	[127]
Diels-alderase ribozyme	[128]	[129]
Class I ligase	[130]	[131]
Bacterial 16S domains	[132,133]	[134,135]
ThreRS operator	[136]	[80]
RNase P	[137,138]	[78,139–141]
Group II intron	[142]	[82]

ribozyme [109] shows that (i) the architectural elements were correctly predicted and (ii) that their relative positions with respect to one another are in agreement with the overall ribozyme architecture (Figure 5C).

The deformation profile (DP, Figure 5D) [106] allows pointing out regions of the model in good agreement or significantly deviating from the reference crystal structure. In the DP, each row represents the average distance between pairs of nucleotides once one pair of nucleotides has been superimposed. Consequently, each column represents the average distance between one pair of nucleotides for each superimposition. As denoted by purple squares annotated *Pi* or *Ji/j* on Figure 5D, helices, junctions and loops are in good agreement when considered individually. They occupy white regions denoting short distances between nucleotides composing them. Yet, some short junctions integrate green pixels showing that although the distances resulting from individual nucleotide superimpositions are minimal, adjoining nucleotides are starting to deviate from the crystal structure. This is also observed for the P7 stem in which the bulge was poorly modelled. A wrong conformation for the first residue of loop L5 (G71) has deep consequences for the average distances between individual pairs of residues as represented by a line of orange and red pixels. Deviations in P6 and P9.0 are due to poor modelling of internal loops as shown by comparison of non-WC base-pair symbols in Figure 5A and B. The crystal structure of the intron revealed a slippage of the 5′ strand of P9.0 that resulted in one additional *trans* H–H pairing between A183 and A201. On the contrary, the knowledge of the structure of tetraloop receptors [66] allowed building very good models for helices P5 and P8.

Interpretation of the DP becomes very useful when checking the consequences on the superimposition of a second element following the superimposition of a first one. Such relationships are represented by blue squares annotated *Pi* × *Pj* in Figure 5D. Despite the secondary structure discrepancies in the model of P9.0, squares addressing the distances of P9 resulting from the superimposition of P9.0 (and vice versa) show that the relative position of P9 is not more affected than the model of P9.0 itself. Although corresponding to an alternative choice, the kink used to model the internal loop between P9.0 and P9 (the 1730 kink of the H. Marismortui 23S RNA) was compatible with the reverse-kink-turn observed in the crystal structure [110]. This kink correctly orients the tetraloop of P9 towards its tetraloop receptor located in P5. The correct modelling of the tetraloops interacting with their receptor results in very good correlation for the positioning of P5 versus P9 and P8 versus P2. Following the same general trend P1 versus P2, P5, or P7 also show good relative locations notably because they are quite close in space. When checking relative locations of distal elements, larger discrepancies start to appear. Optimal superimposition of the P5 or P8 strands result in poor or acceptable superimposition of P8 or P5 strands, respectively, as denoted by the systematic appearance of orange pixels. The conclusion is that P8 is better oriented with respect to the overall model than is P5.

These observations lead to the conclusion that every secondary structure element of the *Azoarcus* ribozyme model co-localizes with its corresponding element in the crystal structure. This situation is reached because the fold of the model's backbone follows the same path than in the crystal structure. Elucidating the topology thus led to deduce the correct architecture of this structurally complex ribozyme. However, the model displays regions that were accurately modelled alternating with regions containing errors. In this case, errors could relate to mistakes in the base-pairing scheme that were either directly inherited from mistakes in the secondary structure or indirectly owing to erroneous non-canonical base-pairing choices. Minimization of these errors strongly depends on the extent of biochemical data that can be integrated under the form of additional constraints to drive modelling of flexible regions.

Homology modelling with reference crystal structures

Integrative modelling can be brought one step beyond to build RNA models by structure homology. The recent progress towards the modelling of the GIR1 branching ribozyme relies on the structural similarity of this ribozyme to regular group I ribozymes [111]. GIR1 is a natural ribozyme that does not perform a splicing reaction but rather catalyses a unique transesterification reaction leading to the formation of a lariat with 3 nucleotides in the loop [112]. Comparative sequence analysis allowed exhibiting features common or distinctive of both ribozymes. Then, common parts were built taking advantage of group I ribozyme crystal structures [109,113,114] and dis-

symbols described in Figure 2. (C) Superimposition between the model (red) and the crystal structure (green). The overall normalized *rms* deviation (*nrmsd*) calculated between backbone atoms including phosphate groups and riboses is 3.85 Å. P10 and the tip of P6 have been omitted from the comparison because secondary structures were different. (D) The deformation profile between the model and the crystal structure gives a qualitative view of the model. The curve on the left indicates the average distance between all nucleotide pairs for each pairwise superimposition (see text). The curve on the bottom indicates the average distance for one pair of nucleotides for each superimposition.

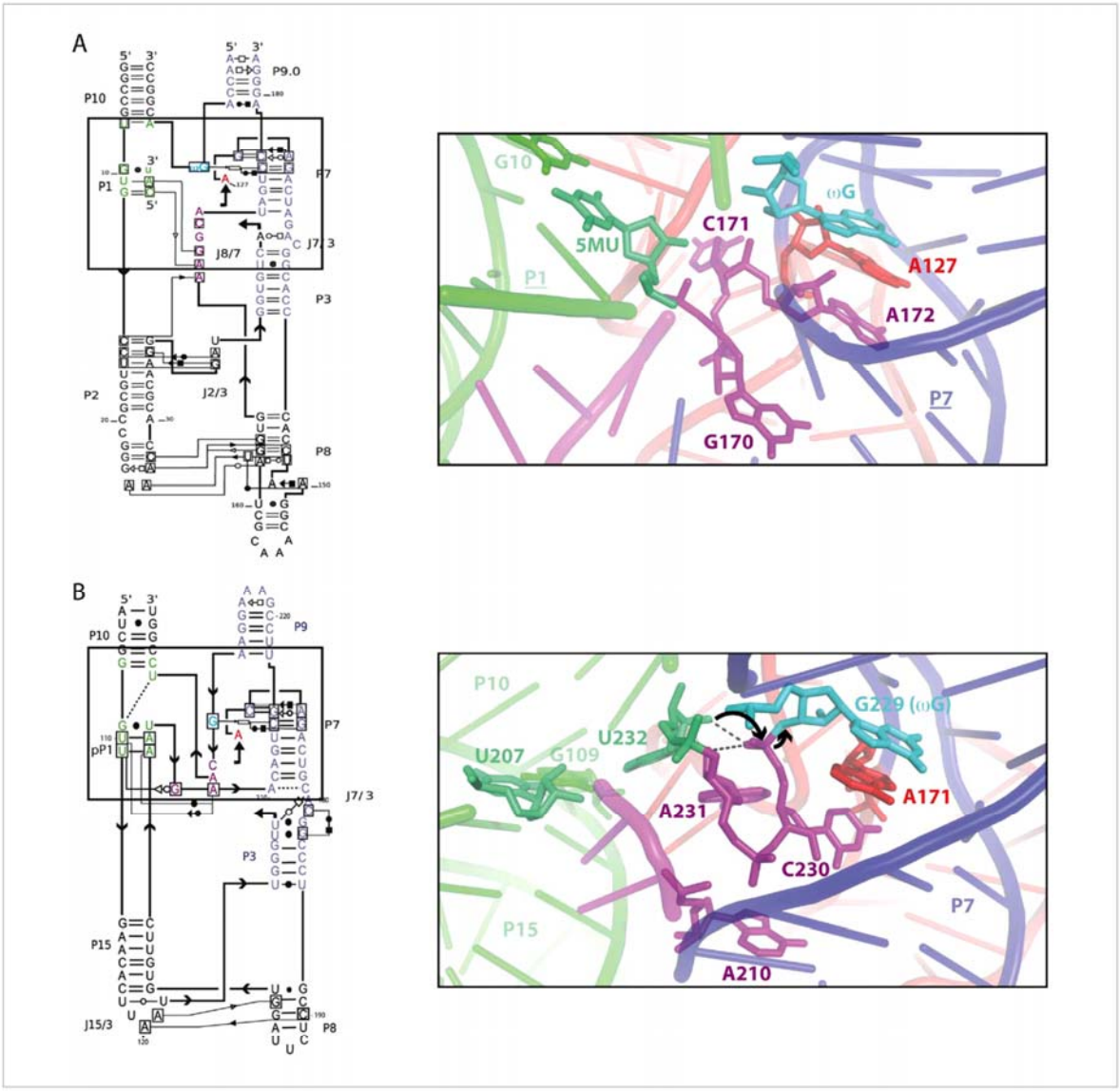


FIGURE 6

Structural homology between the *Azoarcus* group I ribozyme (A) and the GIR1 ribozyme (B). Each panel represents the minimal part of the secondary structure necessary to emphasize the topological differences between the ribozymes. Rectangles indicate the focused region in the corresponding crystal structure or 3D model. In the GIR1 model, P2 is replaced by P15 which forms a pseudoknot with P3 additionally to the P3/P7 pseudoknot characterizing all group I introns. P15 extends itself so as to form a pseudo-P1 (pP1) stem closed by a GoU pair (G109oU207). Arrows indicate positions where the secondary structure diagrams have been truncated. Homology modelling suggests how this topology promotes the remodelling of the critical junction between P8 and P7 enforcing its 3′ end to be replaced by nucleotides from the junction intended to form the lariat following the nucleophilic attack by the 2′ hydroxyl group of U232 (see curved arrows on the 3D representation).

tinctive parts were interactively added. The modelling of the common parts with an unprecedented level of details for the crucial junctions of GIR1 (J3/4, J6/7, J7/9) constituted a set of additional constraints that guided the modelling of the distinctive parts. This process enabled spotting the location of the residues responsible for the formation of the lariat which were strikingly taking the place of key residues from the catalytic junction J8/7 from group I ribozymes within the catalytic site (Figure 6). Beyond the structure, this more accurate model shed light on evolutionary aspects and suggested a pathway for the appearance of a second pseudoknot crucial for the branching reaction within the core of the ribozyme.

Conclusions

Today, crystallography or NMR methods cannot catch up with the growing need for structural data that derive from the rapid harvest of RNA sequences of biological interest and electronic microscopy is still limited by the size of the objects under scrutiny and the resolution that can be reached. Hence, RNA models are necessary because they provide a 3D view of a given RNA that is closer to the reality than the purely intellectual planar secondary structures. This review describes an interactive RNA modelling method that represents a good trade-off between speed and accuracy in order to obtain RNA models unravelling global architectures rationalizing experimental data of all kind.

RNA models can only be finally validated by confronting to the upcoming crystal structures. Hence, metrics such as *rmsd* [104] or deformation profiles [106] can be used for this purpose only retrospectively. To be used efficiently as a heuristic tool, the model should be validated on a reasonable time-scale by checking how it rationalizes experimental data that it directly inspires. In a given RNA model, specific architectural features leading to local deformations are inserted by non-canonical motifs representing around one third of an RNA chain. These regions are thus crucial to orient the WC helices and guide them so as to form tertiary interactions that ultimately lead to the formation of the native and active 3D structure. From an ontological point of view, RNA motifs can be classified according to the effect they have on the relative positions of the adjoining motifs. Do they preserve helical coaxiality or do they mediate bends or kinks with associated specific angles between flanking helices? Do they promote tertiary interactions like pseudoknots? Whatever the base-pair scheme proposed for a non-WC motif, a model will be considered as correct only if the proposed non-WC motifs predict stacking and orientation of adjacent secondary structure elements that are compatible with the biological knowledge of the system. Since sequences encode structural determinants but also the information allowing specific folding, dynamics and interaction properties, it seems very likely that distinct motifs should share identical structural properties regarding how they locally carve the RNA to allow the formation of the final 3D architecture.

The modelling approach described in this review is meant to be pragmatic. It considers that the RNA structure problem can be pinpointed as elucidating one or a couple of biologically relevant conformations built from an assembly of individual rigid building blocks that simultaneously integrate data obtained from virtually unlimited approaches (phylogenetics, probing, footprint, crosslink studies, catalytic activity measurements). The predictive power of a model critically depends on the completeness of these data. Because the integration of data from such various sources is difficult to implement in computer programs, this part of the work is left to an expert that has to identify the folds that are the best candidates to assemble a model fulfilling the experimental dataset. By internally constraining the working models independently from experimental data, stereochemical rules guide the expert in his choices. This perspective has to be replaced in a context where the RNA structure repository of the NDB [47] not only constitutes an unprecedented source of inspiration for modelling non-canonical regions but also provides limits beyond which inspiration becomes speculation. RNA structural databases bring new generation models to an unprecedented level of accuracy. Nonetheless, these large amounts of data create the need for computer tools that can couple secondary structure analysis to 3D interactive modelling platforms able to deal with mining non-canonical motifs so as to fast suggest the expert with reasonable solutions to their specific modelling problems. Decision-aided softwares such as Assemble are triggering the start of a new era for RNA modelling. Furthermore, the observation that the models generated by this method are generally in agreement with upcoming crystal structures shows that the final result is only slightly biased by the expert. This supports the fact that the modelling workflow described here can be used by a large part of the research community.

Acknowledgements

The authors thank Pascale Romby, Pascal Auffinger for critical reading of the manuscript and Eric Westhof for helpful discussions and sharing ideas. Pictures 5 and 6 have been designed using Pymol (http://www.pymol.org/). BB is supported by the Danish Lundbeck Foundation.

References

1 Kapranov, P. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488
2 Frith, M.C. *et al.* (2005) The amazing complexity of the human transcriptome. *Eur. J. Hum. Genet.* 13, 894–897
3 Taft, R.J. *et al.* (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29, 288–299
4 Tucker, B.J. and Breaker, R.R. (2005) Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.* 15, 342
5 Mattick, J.S. and Makunin, I.V. (2006) Non-coding RNA. *Hum. Mol. Genet.* 15 Spec No. 1, R17–R29
6 Mattick, J.S. (2009) The genetic signatures of noncoding RNAs. *PLoS Genet.* 5, e100045910.1371/journal.pgen.1000459
7 Woodson, S.A. (2005) Metal ions and RNA folding: a highly charged topic with a dynamic future. *Curr. Opin. Chem. Biol.* 9, 104–109
8 Draper, D.E. (2008) RNA folding: thermodynamic and molecular descriptions of the roles of ions. *Biophys. J.* 95, 5489–5495
9 Sykes, M.T. and Williamson, J.R. (2009) A complex assembly landscape for the 30S ribosomal subunit. *Annu. Rev. Biophys.* 38, 197–215
10 Rajkowitsch, L. *et al.* (2007) RNA chaperones, RNA annealers and RNA helicases. *RNA Biol.* 4, 118–130
11 Masquida, B. and Westhof, E. (2006) A modular and hierarchical approach for all-atom RNA modeling. In *The RNA world* (third edition) (Gesteland, R.F., Cech, T.R., Atkins, J.F., eds), pp. 659–681, CSHL Press
12 Kimura, M. (1989) The neutral theory of molecular evolution and the world view of the neutralists. *Genome* 31, 24–31
13 Cozzetto, D. *et al.* (2008) The evaluation of protein structure prediction results. *Mol. Biotechnol.* 39, 1–8
14 Gutell, R.R. *et al.* (2002) The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* 12, 301–310
15 Pleij, C.W.A. *et al.* (1985) A new principle of RNA folding based on pseudoknotting. *Nucl. Acids Res.* 13, 1717–1731
16 Westhof, E. and Jaeger, L. (1992) RNA pseudoknots: Structural and functional aspects. *Curr. Opin. Struct. Biol.* 2, 327–333
17 Gautheret, D. *et al.* (1995) Identification of base-triples in RNA using comparative sequence analysis. *J. Mol. Biol.* 248, 27–43
18 Baird, N.J. *et al.* (2005) Structure of a folding intermediate reveals the interplay between core and peripheral elements in RNA folding. *J. Mol. Biol.* 352, 712–722
19 Ding, F. *et al.* (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA* 14, 1164–1173

20 Das, R. and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14664–14669
21 Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452, 51–55
22 Jonikas, M.A. *et al.* (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* 15, 189–199
23 Masquida, B. and Westhof, E. (2005) Modeling the architecture of structured RNAs within a modular and hierarchical framework. In *Handbook of RNA biochemistry* (Hartmann, R.K., Bindereif, A., Schön, A., Westhof, E., eds), pp. 536–545, Wiley VCH Verlag Gmbh & Co.
24 Michel, F. and Westhof, E. (1990) Modelling of the three-dimensional architecture of group-I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* 216, 585–610
25 Massire, C. and Westhof, E. (1998) MANIP: an interactive tool for modelling RNA. *J. Mol. Graph. Model.* 16 197–205, 255–257
26 Westhof, E. *et al.* (1996) RNA tectonics: towards RNA design. *Fold. Des.* 1, R78–R88
27 Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415
28 Hofacker, I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chem. Monthly* 125, 167–188
29 Xayaphoummine, A. *et al.* (2003) Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15310–15315
30 Rivas, E. and Eddy, S.R. (2000) The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics* 16, 334–340
31 Reeder, J. and Giegerich, R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics* 5, 104
32 Jossinet, F. *et al.* (2007) RNA structure: bioinformatic analysis. *Curr. Opin. Microbiol.* 10, 279–285
33 Tabei, Y. *et al.* (2006) SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. *Bioinformatics* 22, 1723–1729
34 Dalli, D. *et al.* (2006) STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics* 22, 1593–1599
35 Siebert, S. and Backofen, R. (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* 21, 3352–3359
36 Sankoff, D. (1985) Simultaneous solution of the RNA folding. Alignment and protosequence problems. *SIAM J. Appl. Math.* 45, 810–825
37 Gorodkin, J. *et al.* (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.* 25, 3724–3732
38 Havgaard, J.H. *et al.* (2005) The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res.* 33 (Web Server issue), W650–W653
39 Mathews, D.H. (2005) Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics* 21, 2246–2253
40 Gardner, P.P. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.* 37 (Suppl. 1), D136–140
41 Seibel, P.N. *et al.* (2006) 4SALE—a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics* 7, 498
42 Wilm, A. *et al.* (2008) ConStruct: improved construction of RNA consensus structures. *BMC Bioinformatics* 9, 219
43 Jossinet, F. and Westhof, E. (2005) Sequence to structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* 21, 3320–3321
44 Andersen, E.S. *et al.* (2007) Semiautomated improvement of RNA alignments. *RNA* 13, 1850–1859
45 Cannone, J.J. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* 3, 2
46 Andronescu, M. *et al.* (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics* 9, 340
47 Berman, H.M. *et al.* (2003) The nucleic acid database. *Methods Biochem. Anal.* 44, 199–216
48 Yang, H. *et al.* (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.* 31, 3450–3460
49 Lemieux, S. and Major, F. (2002) RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res.* 30, 4250–4263
50 Soukup, G.A. and Breaker, R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA* 5, 1308–1325
51 Watts, J.M. *et al.* (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 460, 711–716
52 Brunel, C. and Romby, P. (2000) Probing RNA structure and RNA-ligand complexes with chemical probes. *Methods Enzymol.* 318, 3–21
53 Leontis, N.B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA* 7, 499–512
54 Leontis, N.B. *et al.* (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* 30, 3497–3531
55 Stombaugh, J. *et al.* (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.* 37, 2294–2312
56 Leontis, N.B. *et al.* (2002) Motif prediction in ribosomal RNAs Lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie* 84, 961–973
57 Branch, A.D. *et al.* (1985) Ultraviolet light-induced crosslinking reveals a unique region of local tertiary structure in potato spindle tuber viroid and HeLa 5S RNA. *Proc. Natl. Acad. Sci. U.S.A.* 82, 6590–6594
58 Leontis, N.B. and Westhof, E. (1998) A common motif organizes the structure of multi-helix loops in 16S and 23S ribosomal RNAs. *J. Mol. Biol.* 283, 571–583
59 Ban, N. *et al.* (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289, 905–920
60 Wimberly, B.T. *et al.* (2000) Structure of the 30S ribosomal subunit. *Nature* 407, 327–339
61 Earnshaw, D.J. *et al.* (1997) Inter-domain cross-linking and molecular modelling of the hairpin ribozyme. *J. Mol. Biol.* 274, 197–212
62 Collier, A.J. *et al.* (2002) A conserved RNA structure within the HCV IRES eIF3-binding site. *Nat. Struct. Biol.* 9, 375–380
63 Waldsich, C. *et al.* (2002) Monitoring intermediate folding states of the td group I intron in vivo. *EMBO J.* 21, 5281–5291
64 Klein, D.J. *et al.* (2001) The kink-turn: a new RNA secondary structure motif. *EMBO J.* 20, 4214–4221
65 Lescoute, A. *et al.* (2005) Recurrent structural RNA motifs, Isostericity matrices and sequence alignments. *Nucleic Acids Res.* 33, 2395–2409
66 Cate, J.H. *et al.* (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* 273, 1678–1684
67 Lescoute, A. and Westhof, E. (2006) Topology of three-way junctions in folded RNAs. *RNA* 12, 83–93
68 Steinberg, S.V. and Boutorine, Y.I. (2007) G-ribo motif favors the formation of pseudoknots in ribosomal RNA. *RNA* 13, 1036–1042
69 Steinberg, S.V. and Boutorine, Y.I. (2007) G-ribo: a new structural motif in ribosomal RNA. *RNA* 13, 549–554
70 Laing, C. *et al.* (2009) Tertiary motifs revealed in analyses of higher-order RNA junctions. *J. Mol. Biol.* 393, 67–82
71 Jaeger, L. *et al.* (2009) The UA_handle: a versatile submotif in stable RNA architectures. *Nucleic Acids Res.* 37, 215–230
72 Wimberly, B.T. *et al.* (1999) A detailed view of a ribosomal active site: the structure of the L11-RNA complex [see comments]. *Cell* 97, 491–502
73 Conn, G.L. *et al.* (1999) The crystal structure of the RNA/DNA hybrid r(GAAGAGAAGC)center dot d(GCTTCTCTTC) shows significant differences to that found in solution. *Nucleic Acid Res.* 27, 555–561
74 Suddath, F.L. *et al.* (1974) Three-dimensional structure of yeast phenylalanine transfer RNA at 3.0angstroms resolution. *Nature* 248, 20–24
75 Robertus, J.D. *et al.* (1974) Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* 250, 546–551
76 Moras, D. *et al.* (1980) Crystal structure of yeast $tRNA^{Asp}$. *Nature* 288, 669–674
77 Krasilnikov, A.S. and Mondragon, A. (2003) On the occurrence of the T-loop RNA folding motif in large RNA molecules. *RNA* 9, 640–643
78 Krasilnikov, A.S. *et al.* (2004) Basis for structural diversity in homologous RNAs. *Science* 306, 104–107
79 Lu, M. and Steitz, T.A. (2000) Structure of Escherichia coli ribosomal protein L25 complexed with a 5S rRNA fragment at 1.8-Å resolution. *Proc. Natl. Acad. Sci. U.S.A.* 97, 2023–2028
80 Torres-Larios, A. *et al.* (2002) Structural basis of translational control by Escherichia coli threonyl tRNA synthetase. *Nat. Struct. Biol.* 9, 343–347
81 Woese, C.R. *et al.* (1990) Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci. U.S.A.* 87, 8467–8471
82 Toor, N. *et al.* (2008) Crystal structure of a self-spliced group II intron. *Science* 320, 77–82
83 Keating, K.S. *et al.* (2008) The GANC tetraloop: a novel motif in the group IIC intron structure. *J. Mol. Biol.* 383, 475–481
84 Pley, H.W. *et al.* (1994) Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. *Nature* 372, 111–113

85 Nissen, P. *et al.* (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4899–4903
86 Lee, J.C. *et al.* (2006) The UAA/GAN internal loop motif: a new RNA structural element that forms a cross-strand AAA stack and long-range tertiary interactions. *J. Mol. Biol.* 360, 978–988
87 Doherty, E.A. *et al.* (2001) A universal mode of helix packing in RNA. *Nat. Struct. Biol.* 8, 339–343
88 Kondo, J. *et al.* (2006) Two conformational states in the crystal structure of the Homo sapiens cytoplasmic ribosomal decoding A site. *Nucleic Acids Res.* 34, 676–685
89 Djelloul, M. and Denise, A. (2008) Automated motif extraction and classification in RNA tertiary structures. *RNA* 14, 2489–2497
90 Sarver, M. *et al.* (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.* 56, 215–252
91 Tamura, M. *et al.* (2004) SCOR: structural Classification of RNA, version 2.0. *Nucleic Acids Res.* 32 (Database issue), D182–D184
92 Fera, D. *et al.* (2004) RAG: RNA-As-Graphs web resource. *BMC Bioinformatics* 5, 88
93 Gan, H.H. *et al.* (2004) RAG: RNA-As-Graphs database—concepts, analysis, and features. *Bioinformatics* 20, 1285–1291
94 Abraham, M. *et al.* (2008) Analysis and classification of RNA tertiary structures. *RNA* 14, 2274–2289
95 Nagaswamy, U. *et al.* (2002) NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res.* 30, 395–397
96 Chworos, A. *et al.* (2004) Building programmable jigsaw puzzles with RNA. *Science* 306, 2068–2072
97 Gautheret, D. *et al.* (1994) A major family of motifs involving G.A mismatches in ribosomal RNA. *J. Mol. Biol.* 242, 1–8
98 Elgavish, T. *et al.* (2001) AA.AG@helix.ends: A:A and A:G base-pairs at the ends of 16 S and 23 S rRNA helices. *J. Mol. Biol.* 310, 735–753
99 Cate, J.H. *et al.* (1996) RNA tertiary structure mediation by adenosine platforms. *Science* 273, 1696–1699
100 Westhof, E. *et al.* (1985) Crystallographic refinement of yeast aspartic acid transfer RNA. *J. Mol. Biol.* 184, 119–145
101 Mangalam, H. (2002) The Bio* toolkits—a brief overview. *Brief Bioinform.* 3, 296–302
102 Curcin, V. *et al.* (2005) Web services in the life sciences. *Drug Discov. Today* 10, 865–871
103 Stein, L. (2002) Creating a bioinformatics nation. *Nature* 417, 119–120
104 Carugo, O. and Pongor, S. (2001) A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.* 10, 1470–1473
105 Kleywegt, G.J. (1996) Use of non-crystallographic symmetry in protein structure refinement. *Acta Cryst.* D52, 842–857
106 Parisien, M. *et al.* (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA* 15, 1875–1885
107 Rangan, P. *et al.* (2004) Architecture and folding mechanism of the Azoarcus Group I Pre-tRNA. *J. Mol. Biol.* 339, 41–51
108 Michel, F. *et al.* (1989) The guanosine binding site of the tetrahymena ribozyme. *Nature* 342, 391–395
109 Adams, P.L. *et al.* (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature* 430, 45–50
110 Strobel, S.A. *et al.* (2004) RNA kink turns to the left and to the right. *RNA* 10, 1852–1854
111 Beckert, B. *et al.* (2008) Molecular modelling of the GIR1 branching ribozyme gives new insight into evolution of structurally related ribozymes. *EMBO J.* 27, 667–678
112 Nielsen, H. *et al.* (2005) An mRNA is capped by a 2′, 5′ lariat catalyzed by a group I-like ribozyme. *Science* 309, 1584–1587
113 Guo, F. *et al.* (2004) Structure of the tetrahymena ribozyme; base triple sandwich and metal ion at the active site. *Mol. Cell* 16, 351–362
114 Golden, B.L. *et al.* (2005) Crystal structure of a phage Twort group I ribozyme-product complex. *Nat. Struct. Mol. Biol.* 12, 82–89
115 Lehnert, V. *et al.* (1996) New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the Tetrahymena thermophila ribozyme. *Chem. Biol.* 3, 993–1009
116 Caprara, M.G. *et al.* (1996) A Tyrosyl-tRNA synthetase protein induces tertiary folding of the group I intron catalytic core. *J. Mol. Biol.* 257, 512–531
117 Rangan, P. *et al.* (2003) Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proc. Natl. Acad. Sci. U.S.A.* 100, 1574–1579
118 Paukstelis, P.J. *et al.* (2008) Structure of a tyrosyl-tRNA synthetase splicing factor bound to a group I intron RNA. *Nature* 451, 94–97
119 Krol, A. *et al.* (1990) Solution structure of human U1 snRNA. Derivation of a possible three-dimensional model. *Nucleic Acids Res.* 18, 3803–3811
120 Pomeranz Krummel, D.A. *et al.* (2009) Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature* 458, 475–480
121 Brunel, C. *et al.* (1991) Three-dimensional model of E. coli ribosomal 5S RNA as deduced from structure probing in solution and computer modeling. *J. Mol. Biol.* 221, 293–308
122 Tuschl, T. *et al.* (1994) A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. *Science* 266, 785–789
123 Pley, H.W. *et al.* (1994) Three-dimensional structure of a hammerhead ribozyme. *Nature* 372 (November), 68–74
124 Scott, W.G. *et al.* (1995) The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell* 81, 991–1002
125 Tanner, N.K. *et al.* (1994) A three-dimensional model of hepatitis delta virus ribozyme based on biochemical and mutational analysis. *Curr. Biol.* 4, 488–498
126 Ferré-D'Amaré, A.R. *et al.* (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature* 395, 567–574
127 Rupert, P.B. and Ferre-D'Amare, A.R. (2001) Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature* 410, 780–786
128 Keiper, S. *et al.* (2004) Architecture of a Diels-Alderase ribozyme with a preformed catalytic pocket. *Chem. Biol.* 11, 1217–1227
129 Serganov, A. *et al.* (2005) Structural basis for Diels-Alder ribozyme-catalyzed carbon-carbon bond formation. *Nat. Struct. Mol. Biol.* 12, 218–224
130 Bergman, N.H. *et al.* (2004) The three-dimensional architecture of the class I ligase ribozyme. *RNA* 10, 176–184
131 Bagby, S.C. *et al.* (2009) A class I ligase ribozyme with reduced Mg^{2+} dependence: Selection, sequence analysis, and identification of functional tertiary interactions. *RNA* 15, 2129–2146
132 Masquida, B. *et al.* (1997) Context dependent RNA-RNA recognition in a three-dimensional model of the 16S rRNA core. *Bioorgan. Med. Chem.* 5, 1021–1035
133 Serganov, A.A. *et al.* (1996) The 16S rRNA binding site of Thermus thermophilus ribosomal protein S15: comparison with Escherichia coli S15, minimum site and structure. *RNA* 2, 1124–1138
134 Nikulin, A. *et al.* (2000) Crystal structure of the S15-rRNA complex. *Nat. Struct. Biol.* 7, 273–277
135 Schuwirth, B.S. *et al.* (2005) Structures of the bacterial ribosome at 3.5 Å resolution. *Science* 310, 827–834
136 Caillet, J. *et al.* (2003) The modular structure of Escherichia coli threonyl-tRNA synthetase as both an enzyme and a regulator of gene expression. *Mol. Microbiol.* 47, 961–974
137 Massire, C. *et al.* (1998) Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis. *J. Mol. Biol.* 279, 773–793
138 Tsai, H.Y. *et al.* (2003) Molecular modeling of the three-dimensional structure of the bacterial RNase P holoenzyme. *J. Mol. Biol.* 325, 661–675
139 Krasilnikov, A.S. *et al.* (2003) Crystal structure of the specificity domain of ribonuclease P. *Nature* 421, 760–764
140 Torres-Larios, A. *et al.* (2005) Crystal structure of the RNA component of bacterial ribonuclease P. *Nature* 437, 584–587
141 Kazantsev, A.V. *et al.* (2005) Crystal structure of a bacterial ribonuclease P RNA. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13392–13397
142 Costa, M. *et al.* (2000) A three-dimensional perspective on exon binding by a group II self-splicing intron. *EMBO J.* 19, 5007–5018
143 Shi, H. and Moore, P.B. (2000) The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: a classic structure revisited. *RNA* 6, 1091–1105
144 Klein, D.J. *et al.* (2004) The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.* 340, 141–177
145 Szep, S. *et al.* (2003) The crystal structure of a 26-nucleotide RNA containing a hook-turn. *RNA* 9, 44–51
146 Correll, C.C. *et al.* (1999) The two faces of the Escherichia coli 23 S rRNA sarcin/ricin domain: the structure at 1.11 A resolution. *J. Mol. Biol.* 292, 275–287
147 Correll, C.C. *et al.* (1997) Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell* 28, 705–712

REVIEW III:

Synthesis of RNA by *in vitro* transcription

B. Beckert, B. Masquida

The Humana Press Inc.
Methods in Molecular Biology.
In press

Synthesis of RNA by *in vitro* transcription

Bertrand Beckert[1,2] and Benoit Masquida[1]

[1] Université Louis Pasteur de Strasbourg, Institut de Biologie Moléculaire et Cellulaire, CNRS Strasbourg, France

[2] Department of Cellular and Molecular Medicine, University of Copenhagen, Denmark

Corresponding author: Benoit Masquida, Architecture et Réactivité de l'ARN, Université Louis Pasteur de Strasbourg, Institut de Biologie Moléculaire et Cellulaire, CNRS, 15 rue René Descartes, 67084 Strasbourg, France

Summary

In vitro transcription is a simple procedure that allows for template-directed synthesis of RNA molecules of any sequence and of lengths from short oligos to several kilobases in µg to mg quantities. It is based on the engineering of a template that include a bacteriophage promoter sequence (e.g. from the T7 coliphage) upstream of the sequence of interest followed by transcription using the corresponding RNA polymerase. *In vitro* transcripts are used in analytical techniques (e.g. hybridization analysis), structural studies (for NMR and X-ray crystallography), in biochemical and genetic studies (e.g. as antisense reagents), and as functional molecules (ribozymes and aptamers).

Keywords: T7 RNA polymerase, *in vitro* transcription, template purification

1. Introduction

RNA is conveniently synthesized by *in vitro* transcription using the components of bacteriophage systems. The RNA polymerase (RNAP) is a single subunit of about 100 kDa that is highly specific for its 23-bp promoter sequence. With these two simple components, it is possible to make transcripts ranging in size from less than 30 nt to well over 10^4 nt in scales from μg to mg amounts. The most frequently used systems are the T3, T7, and SP6 systems. Here, *in vitro* transcription is examplified by the T7 system derived from the T7 phage of *E. coli* established many years ago (1). *In vitro* transcripts can be used as hybridization probes, in RNase protection or interference experiments, as antisense reagents, for analysis of RNA-binding proteins, to elucidate RNA structure by structure probing, NMR or X-ray crystallography, or as functional molecules (e.g. aptamers and ribozymes). The emphasis in this chapter is the synthesis of transcripts in small scale for probes and simple biochemical applications. For a more comprehensive discussion of *in vitro* transcription, see Gruegelsiepe et al. (2).

The basic strategy is to place the sequence of interest downstream of the T7 promoter. The promoter covers the sequence ranging from -17 to +6 with +1 being the first nucleotide of the transcribed region (*see* **Fig. 1**). Thus, there is not complete freedom in the choice of the sequence at the very 5'-end of the *in vitro* transcript. Most T7 promoters, like class III promoters (3), have G's at +1, +2, and +3, and the first two G's are critical for transcriptional yield. The alternative class II promoters initiate with an A and have a similar preference for G's at +2 (4). The template for transcription can be 1) a plasmid that typically has the promoter for *in vitro* transcription immediately upstream of a polylinker for cloning of the sequence to be transcribed, 2) a PCR product that has the T7 promoter as part of the 5'-oligonucleotide used in the PCR reaction, and 3) two annealed oligonucleotides that carries the T7 promoter sequence and the template to be transcribed (in this case, only the T7 promoter part of the oligo's needs to be double-stranded) (*see* **Fig. 2**). Most plasmid cloning vectors have one or more promoters for *in vitro* transcription upstream of multiple cloning sites (MCS) (e.g. the pBluescript (Stratagene) and pGEM (Promega) series). An alternative strategy consists in cloning a DNA fragment including a T7 promoter immediately 5' of the sequence to be transcribed in order to avoid the presence of nucleotides derived from the MCS in the transcript. In this case plasmids like pUC18 and pUC19 are preferred due to the absence of a built-in T7 promoter. Cloned templates are used for long transcripts (> 100

nt) and annealed oligo's for very short transcripts. When large amounts of RNA are needed, it is better to use a cloned template in order to generate enough template using simple and economical techniques based on bacterial culture and plasmid extraction. When small amounts are needed, PCR-products are probably the most convenient due to the flexibility in design of the template and the ease its production.

Transcription termination in the natural setting occurs at specific terminator sites called Rho-independent terminators (5). In this mechanism, the 3' end of the mRNA forms a hairpin structure about 7-20 base pairs in length directly followed by an U-rich stretch (6). The hairpin formation promotes pausing of the RNA polymerase and leads to disruption the transcription complex. However, for *in vitro* transcripts, termination is usually by "run off", that is when the polymerase falls off at the very end of the template. With the PCR and oligo templates this is defined by the ends of the template products. With cloned templates this is achieved by linearizing the plasmid by restriction enzyme digestion downstream of the sequence of interest.

The average rate of transcription *in vitro* is 200-260 nt/ sec and the error frequency about 6 x 10^{-6} (7). In addition, the use of artificial templates for T7 transcription can result in sequence heterogeneities at the 5' and 3' ends of transcripts. For some applications, like in NMR or X-ray crystallography, homogeneity of the ends is crucial. Some sequences located at the 5' end of DNA templates render the T7 RNAP inaccurate during the initiation of transcription. For example when the template sequence starts with a stretch of 5 to 6 G residues, untemplated G residues can be integrated in the transcripts (8). If the 5' end of the sequence does not start with guanine residues but with 5'C_{+1}AC/G as in the human mitochondrial lysyl and prolyl-tRNAs, transcription will occur but leads to incorporation of one additional nucleotide (preferentially a purine) or to skipping of the +1 and +2 residues (9). It is likely that other sequences could present similar transcription defects. One solution to problems like these is to fuse a cleavage ribozyme 5' to the RNA of interest (10) (11). In this case, the natural +1-+6 residues of the natural T7 promoter can be used regardless of the starting sequence of the RNA of interest guaranteeing efficient transcription and efficient control of the 5' sequence content. The 3' end of the transcript can similarly be heterogeneous. During run-off transcription T7 RNAP has a tendency to incorporate one or several non-templated nucleotides at the 3'-end, thus leaving the pool of transcripts with heterogeneous 3'-ends. This problem is addressed by incorporating a sequence that encodes a *cis*-acting cleavage ribozyme like

the Hepatitis delta virus (HDV ribozyme) at the 3'-end of the template (*see* **Fig. 3**) (11). By using an optimized HDV ribozyme, homogenous RNA 3'ends can be easily generated (12). During transcription, the HDV ribozyme folds into an active conformation and cleaves the transcript (*see* **Fig. 3**). The self-cleavage reaction, even efficient at low Mg^{2+} concentration, releases well-defined 3'-ends. However, the competition between the folding of the RNA of interest and the folding of the HDV ribozyme could lead to reduced cleavage efficiency. This problem normally be can be tackled optimization of temperature, pH and salt conditions (13).

Another concern can be the concentration of rNTPs in the course of the transcription reaction. This problem arises when one of the nucleotides is used at limiting concentrations e. g. during synthesis of radioactive body-labelled transcripts. During the initiation process, the RNA polymerase initially produces short, abortive oligoribonucleotides of 9-12 nt in length. At some point, the polymerase switches to processive transcription leading to full-length products. If the first 9-12 nucleotides are rich in a nucleotide that is used at limiting concentrations (e.g. several U's when attempting to make a transcript labelled at high specific activity with [α-^{32}P]UTP), the switch to processive transcription is made more difficult and the ratio between full length and abortive transcripts decreases. As a consequence of this phenomenon, [α-^{32}P]GTP is frequently avoided as a label because G's are inherently rich at the 5'-end of the transcripts.

In vitro transcription protocols are easily modified to allow for synthesis of modified transcripts. T7 RNAP can initiate transcription with guanosine or GMP to obtain 5'-OH or 5'-monophosphate ends. The latter gets more easily dephosphorylated as compared to a triphosphate 5'end for subsequent 5'end labelling using [γ-^{32}P]ATP and T4 polynucleotide kinase. Dinucleotides (e.g. ApG) or various cap analogues, e.g. 7-methyl-guanosine (to obtain mRNA transcripts with native-like 5'-ends) can also be used for transcription initiation. The cap nucleotide protects the transcript against degradation by 5' exonucleases present in extracts and supports translation of the transcript. T7 RNA polymerase use variety of modified nucleoside 5' triphosphates for internal modification by incorporation. Biotinylated or digoxigenylated nucleotides can be incorporated to make non-radioactive probes for hybridization. Photoreactive nucleotides can be incorporated for synthesis of modified RNAs for various biochemical analyses. The nucleotide analogue interference mapping method (NAIM, See (Cochrane and Strobel (14) for review) also relies on the property of the T7 RNA polymerase to incorporate modified nucleotides in transcripts. In this method, 5'-O-(1-

thio)-nucleoside triphosphate analogues that are commercially available (GlenResearch, VA, USA) are incorporated at a 5% rate by transcription. After purification of the RNA using an activity assay specific to the studied RNA, iodine cleavage is performed so as to identify the residues that are important for activity. The wild-type T7 RNA polymerase or the mutant Y639F (15) (Epicentre, WI, USA) which also allows efficient incorporation of nucleotides with a modified 2' position, such as 2'-deoxy or 2'-fluoro can be used in this case (See Gruegelsiepe (2) for a more detailed discussion of the applications of modified transcripts).

All the protocols below describe the various procedures for *in vitro* transcription from plasmid- and PCR-derived templates (Figure 2). All these protocols provide simple methods to produce RNA by using a commercial T7 RNA polymerase. However, the commercial T7 RNA polymerase could be easily replaced by an in-house T7 RNA polymerase made by expression and purification of an His-tagged T7 RNA polymerase (plasmid pT7-911Q) (16)). Then follow protocols for making unlabelled and ^{32}P-labelled transcripts. The protocols are for small-scale transcriptions, but they can be scaled up without problems. Similarly, the specific activity of the radioactive transcripts can be altered by adjusting the ratio between UTP and [α-^{32}P]UTP. Depending on the use of the transcript, a simple phenol:chloroform extraction directly followed by a ethanol precipitation of the transcript may be sufficient. Transcripts that are used as hybridization probes are purified by gel-filtration to get rid of the unincorporated nucleotides for reasons of radiation hazards and to allow for a simple evaluation of the probe. A protocol for gel filtration and a simple calculation of the specific activity of the probe is included. In other cases, gel purification of the transcripts is required and a simple protocol for this ends the chapter (*see* **Fig. 3**).

2. Materials

2.1.1 Plasmid DNA templates for *in vitro* transcription

1. Plasmid including the sequence to be transcribed downstream of a T7 promoter and upstream of a unique restriction enzyme site to be used for linearization (*see* **Note 1**).
2. Restriction enzyme and corresponding buffer.
3. Proteinase K.
4. Phenol:chloroform:isoamylalcohol (25:24:1).
5. 96% ethanol.

6. 70% ethanol.
7. TE 8.0 (10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA).

2.1.2 PCR templates for *in vitro* transcription

1. Template DNA (genomic DNA, cDNA or a cloned fragment inserted into a vector).
2. Oligonucleotides designed to amplify the sequence of interest (*see* **Note 2**).
3. Thermostable DNA polymerase with proof-reading activity such as *Pfu*I.
4. 10X polymerase buffer (usually provided by the supplier of the polymerase; *see* **Note 3**).
5. 10X dNTP-mix (2 mM of each dNTP).
6, PCR clean-up kit (e.g. GenElute^{TM} PCR Clean-Up Kit Sigma).

2.2.1 *In vitro* transcription of unlabelled transcripts

1. Template DNA (*see* **Subheading 2.1.2**) at 1 µg/ µL of a 3 kb linearized plasmid or 0.2 µg/ µL of a 600 bp PCR-product. This will result in a final concentration of T7 promoter in the transcription of ~20 nM.
2. 10X polymerase buffer: 100 mM NaCl, 80 mM $MgCl_2$, 20 mM spermidine, 800 mM Tris-HCl, pH 8.0.
3. 100 mM DTT.
4. 10X rNTP mix: 10 mM of each rNTP.
5. T7 RNA polymerase (20 U/ µL).

2.2.2 *In vitro* transcription of ^{32}P-labelled transcripts (*see* **Note 4**)

1. Template DNA at 1 µg/ µL of a 3 kb linearized plasmid or 0.2 µg/ µL of a 600 bp PCR-product. This will result in a final concentration of T7 promoter in the transcription of ~20 nM.
2. 10X polymerase buffer: 100 mM NaCl, 80 mM $MgCl_2$, 20 mM spermidine, 800 mM Tris-HCl, pH 8.0.
3. 100 mM DTT.
4. 10X rNTP mix "low UTP" for radio-labelled transcripts: 1 mM UTP, 10 mM of each of ATP, CTP, and GTP (*see* **Note 5**).
5. T7 RNA polymerase (20 U/ µL).
6. [α-^{32}P]UTP (3000 Ci/ mmol; 10 mCi/ mL) (this corresponds to ~3 µM in UTP).

2.3.1 Purification of transcripts by gel filtration

1. Sephacryl S-200 columns (GE Healthcare).

2.3.2 Gel purification of transcripts

1. Denaturing polyacrylamide gel.
2. TBE 10X electrophoresis buffer.
3. Ethidium bromide staining solution.:
4. Elution buffer: 0.25 M sodium acetate, pH 6.0, 1 mM EDTA.
5. Phenol saturated with elution buffer.
6. Glycogen
7. 96% ethanol.
8. TE 7.6 (10 mM Tris-HCl pH 7.6, 0.1 mM EDTA).

3. Methods

3.1.1 Plasmid templates for *in vitro* transcription

1. Digest the (RNase-free) plasmid DNA (e.g. 100 µg) with an appropriate restriction enzyme that cleaves downstream of the T7 promoter and the segment to be transcribed.
2. Add proteinase K to a final concentration of 50 µg/ mL and incubate for 30 min at 37°C in order to remove the restriction enzyme from the template DNA.
3. Extract twice with one volume of phenol-chloroform (*see* **Note 6**).
4. Precipitate the template with 2.5 vols. of 96% ethanol.
5. Resuspend the DNA to 1 µg/ µL in TE 8.0.
6. Run an aliquot (e.g. 0.5 µg) of the DNA on an agarose gel to check the linearization of the plasmid (*see* **Note 7**).

3.1.2 PCR templates for *in vitro* transcription

1. Design the oligos for PCR-amplification.
2. Make a standard PCR reaction.
3. Purify the PCR product using a commercial PCR clean-up kit (GenElute™ PCR Clean-Up Kit Sigma) according to the manufacturer's instructions.

3.2.1 *In vitro* transcription of cold (i.e. unlabelled) transcripts

1. Set up the transcription reaction by at adding at room temperature the components in a siliconized or Teflon-coated tube in the following order (*see* **Note 8**):

- 5 µL of 5X transcription buffer
- 4 µL of 10X rNTP mix
- 2.5 µL of 100 mM DTT
- 11.5 µL DEPC-treated dH_2O
- 1 µL of template DNA (linearized plasmid or PCR-product)
- 1 µL 10U of the appropriate (in this case T7) RNA polymerase (*see* **Note 9**)

2. Incubate for 30 to 60 min at 37°C.

3.2.2 *In vitro* transcription of ^{32}P-labelled transcripts (*see* **Note 4** for ^{32}P-handling)

1. Set up the transcription reaction by adding at room temperature the components in a siliconized or Teflon-coated tube in the following order:

- 5 µL of 5X transcription buffer
- 4 µL of 10X rNTP mix "low UTP"
- 2.5 µL of 100 mM DTT
- 6.5 µL DEPC-treated dH_2O
- 1 µL of template DNA (linearized plasmid or PCR-product)
- 5 µL of 3000 Ci/ mmol, 10 mCi/ ml [α-^{32}P]UTP
- 1 µL 10U of the appropriate (in this case T7) RNA polymerase

2. Incubate for 30 to 60 min at 37°C.

3.3.1 Purification of transcripts by gel filtration

1. Prepare the column according to the manufacturer's recommendation (usually a brief, low-speed spin to remove storage buffer).
2. Add the transcription reaction on top the column and spin briefly (typically 2 min at low (735 x *g*) speed.
3. Collect the eluate containing the transcript. Most of the unincorporated nucleotides are retained in the column. If the transcript is radioactive, an aliquot can be removed and used for estimation of the specific activity without further purification (*see* **Note 10**).

9

3.3.2 Gel purification of transcripts (*see* **Note 11**)

1. Run the transcription mixture on a denaturing polyacrylamide gel. The type of gel depends on the size of the transcript to be purified, but in most cases, a 5% polyacrylamide gel will be appropriate.
2. Visualize the RNA by ethidium bromide staining or UV_{254}-shadowing over Xerox paper. Radioactive transcripts are detected by autoradiography using fluorescent markers to help in alignment of the gel and autoradiogram.
3. Excise the full-length transcript using a scalpel. Avoid carrying over excessive amounts of polyacrylamide.
4. Place the gel slice in a tube containing 400 µL of elution buffer and an equal volume of phenol. (*see* **Notes 12 and 13**).
5. Shake the tubes at room temperature for several hours or over night in the cold room (4°C). The time required will depend on the size of the RNA and the composition of the gel.
6. Spin and transfer all of the liquid to a new tube.
7. Spin and transfer the aqueous phase to a new tube. Add 4 µL of glycogen and 1200 µL of ethanol to precipitate the RNA.
8. Resuspend in dH_2O or TE buffer.

4. Notes

1. A restriction enzyme that leaves 5'-protruding ends is preferred in the linearization of the plasmid because T3 and T7 polymerases can initiate transcription from the ends of DNA fragments. This type of initiation is most prevalent with 3'-protruding termini followed by blunt ends and 5'-protruding termini. Non-specific initiation is suppressed in transcription buffers with increased (100 mM) NaCl concentration. However, this will also result in a decrease of the total transcription efficiency by approximately 50%.
2. The 5'-oligo should incorporate the class I T7 promoter sequence: 5´- TAATACGACTCACTAT AGG(G) or the class II promoter sequence for ApG transcription starter: 5'- TAATACGACTCACT ATTAG (*see* **Fig. 1**) both of them directly followed by specific target sequence. For this and the 3'-oligo, we typically use 15-20-mer sequences with a T_m around 50°C as calculated adding 2 °C for each A or T in the sequence and 4 °C for each C or G. This simple approach for designing oligos rarely fails. However, it is also possible to use software made to optimize primer design, such as Primer3 found at http://frodo.wi.mit.edu.

3. The free [Mg^{2+}] must be adjusted according to the nucleotide concentration. Since each nucleotide chelates one Mg^{2+} ion, the total [Mg^{2+}] should exceed the total nucleotide concentration by approximately 5 mM.

4. **CAUTION !** ^{32}P is a high energy β-emitter. Avoid exposure to the radiation and radioactive contamination. Wear disposable gloves when handling radioactive solutions. Check your gloves and pipettes frequently for radioactive contamination. Use protective laboratory equipment (protective eyeglasses, Plexiglas shields) to minimize exposure to radiation. Dispose of radioactive waste in accordance with the rules and regulations established at your institution.

5. Any of the four rNTPs can be used as label. The main concern is to avoid using a nucleotide that is prevalent in the first 10-12 nucleotides of the transcript and this criteria will in many cases argue against GTP because G's are required at +1 and +2 and preferred at +3 positions.

6. To increase the recovery in extractions of small volumes it is sometimes advisable to increase the volume of the sample prior to extraction. For DNA samples this can be done by addition of DEPC-water.

7. Incomplete digestion can be due to suboptimal conditions or the possibility that some of the DNA was not exposed to the enzyme. As a result, subsequent transcription will lead to transcripts of the full plasmid including vector sequences. To avoid this, siliconized or Teflon-treated tubes should be used in the restriction enzyme digestion and the sample should be given a brief spin after the addition of the enzyme to collect all of the components in the bottom of the tube. One other possibility is to transfer the sample to a new tube before the next step. In this way, droplets on the side of the tube that were not exposed to the enzyme are not carried over to subsequent steps.

8. The order of assembling the reaction is to avoid spermidine precipitation of the template DNA, especially at low temperatures.

9. Alkaline pyrophosphatase can be added to the transcription reaction at 2 ng/µL. The phosphatase we use is purified from *E. coli* and commercially available at Sigma-Aldrich. This hydrolase cleaves the insoluble pyrophosphate into phosphate. Hence, the RNA pellet obtained by ethanol precipitation of the transcription reaction is free of pyrophosphate which greatly facilitates further solubilisation in an appropriate buffer. Furthermore, the hydrolysis of pyrophosphate drives the chemical equilibrium towards the formation of pyrophosphate which means enhancing the polymerization of the RNA by the T7 RNAP and improving the transcription yield.

10. Note on calculation of yield. RNA labelled to a high specific activity is unstable and should be used within a couple of weeks if full-length RNA is required.
11. As an alternative to elution by diffusion, the RNA can be electro-eluted from the gel slice placed in a dialysis bag in an electrophoresis chamber (1 h at 10 V/ cm in TBE) or using dedicated commercial equipment.
12. In some protocols the gel slice is crushed or freeze-thawed. In our experience this will give rise to difficulties with small pieces of polyacrylamide in downstream steps. We prefer to avoid this and have not experienced less recovery of transcript from this.
13. Break the hinge of the tube by pressing it against the table and wrap in parafilm. This will prevent leakage from the tube during shaking.

References

1. Milligan, J. F. & Uhlenbeck, O. C. (1989) Synthesis of small RNAs using T7 RNA polymerase. *Methods Enzymol, 180*, 51--62
2. Gruegelsiepe, H., Schön, A., Kirsebom L. A. and Hartmann, R. K. (2005) Handbook of RNA Biochemistry (Hartmann, R. K., Bindereif, A., Schön A., Westhof E., ed.), WILEY-VCH Verlag GmbH & Co. KGaA, Germany, pp. 3--21.
3. Milligan, J.F., Groebe, D.R., Witherell, G.W. and Uhlenbeck, O.C. (1987) Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic Acids Res*, **15**, 8783--8798.
4. Huang, F. and Yarus, M. (1997) 5'-RNA self-capping from guanosine diphosphate. *Biochemistry*, **36**, 6557--6563.
5. Jeng, S.T., Gardner, J.F. and Gumport, R.I. (1990) Transcription termination by bacteriophage T7 RNA polymerase at rho-independent terminators. *J Biol Chem*, **265**, 3823--3830.
6. Dunn, J.J. and Studier, F.W. (1983) Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J Mol Biol*, **166**, 477--535.
7. Brakmann, S. and Grzeszik, S. (2001) An error-prone T7 RNA polymerase mutant generated by directed evolution. *Chembiochem*, **2**, 212--219.
8. Pleiss, J.A., Derrick, M.L. and Uhlenbeck, O.C. (1998) T7 RNA polymerase produces 5' end heterogeneity during in vitro transcription from certain templates. *RNA*, **4**, 1313--1317.

9. Helm, M., Brule, H., Giege, R. and Florentz, C. (1999) More mistakes by T7 RNA polymerase at the 5' ends of in vitro-transcribed RNAs. *RNA*, **5**, 618--621.
10. Fechter, P., Rudinger, J., Giege, R. and Theobald-Dietrich, A. (1998) Ribozyme processed tRNA transcripts with unfriendly internal promoter for T7 RNA polymerase: production and activity. *FEBS Lett*, **436**, 99--103.
11. Price, S.R., Ito, N., Oubridge, C., Avis, J.M. and Nagai, K. (1995) Crystallization of RNA-protein complexes I. Methods for the large-scale preparation of RNA suitable for crystallographic studies. *J. Mol .Biol.*, **249**, 398--408.
12. Schurer, H., Lang, K., Schuster, J. and Morl, M. (2002) A universal method to produce in vitro transcripts with homogeneous 3' ends. *Nucleic Acids Res*, **30**, e56.
13. Bevilacqua, P.C., Brown, T.S., Nakano, S. and Yajima, R. (2004) Catalytic roles for proton transfer and protonation in ribozymes. *Biopolymers*, **73**, 90--109
14. Cochrane, J.C. and Strobel, S.A. (2004) Probing RNA structure and function by nucleotide analog interference mapping. *Curr Protoc Nucleic Acid Chem*, **Chapter 6**, Unit 6 9.
15. Sousa, R. and Padilla, R. (1995) A mutant T7 RNA polymerase as a DNA polymerase. *Embo J*, **14**, 4609--4621.
16. Ichetovkin, I.E., Abramochkin, G. and Shrader, T.E. (1997) Substrate recognition by the leucyl/phenylalanyl-tRNA-protein transferase. Conservation within the enzyme family and localization to the trypsin-resistant domain. *J Biol Chem*, **272**, 33009--33014.

Figure legends

Figure 1. (A) Consensus sequence of (class III and class II) T7 RNA polymerase promoter with indication of the +1 nucleotide (bold; corresponds to the first nucleotide in the transcript). (B) When the DNA template is incubated in the presence of T7 RNA polymerase and rNTPs, a transcript is made as indicated with a triphosphate at the 5'-end.

Figure 2. Three different types of DNA templates for *in vitro* transcription. In the upper panel, a circular plasmid with the insert of interest cloned between a T7 promoter and a unique restriction enzyme site is linearized and transcribed from the promoter to yield multiple RNA transcripts terminated by "run-off" on the template. In the middle panel, a DNA template (genomic DNA, cDNA or a cloned fragment) acts as a template in PCR with a 5'-primer containing a T7 promoter

(with no complementarity to the template) fused to a specific sequence complementary to the sequence of interest and a similarly specific 3'-primer. The resulting PCR-product is transcribed into RNA. In the lower panel, a short oligo corresponding to the T7 promoter sequence is annealed to an oligo that has the complementary sequence fused to a template sequence of interest. The partially double-stranded oligos can be transcribed into short RNAs.

Figure 3. The 3' cassette for the creation of homogeneous RNA 3' end. The DNA molecule to be transcribe (linearized plasmid, PCR product) includes extra cassette downstream of the sequence encoding the RNA of interest. This cassette (in grey) is transcribed into self-splicing ribozyme structure (the HDV ribozyme). The cleavage activity of the HDV ribozyme leads to the release of a 2',3'-cyclic phosphate goup at the 3' end terminus of the synthesized RNA.

Figure 4. Gelelectrophoretic separation of a transcription reaction. In addition to the full-length transcript, several prematurely terminated transcripts are seen. The full-length transcript can be excised from the gel and eluted into a buffer from which it can be recovered. Premature termination is typical when the concentration of one nucleotide is lowered to favour synthesis of radioactive transcripts of high specific activity. The presence of sequences in the template that resemble terminators or other sequences that are difficult to transcribe will similarly result in short transcripts.

14

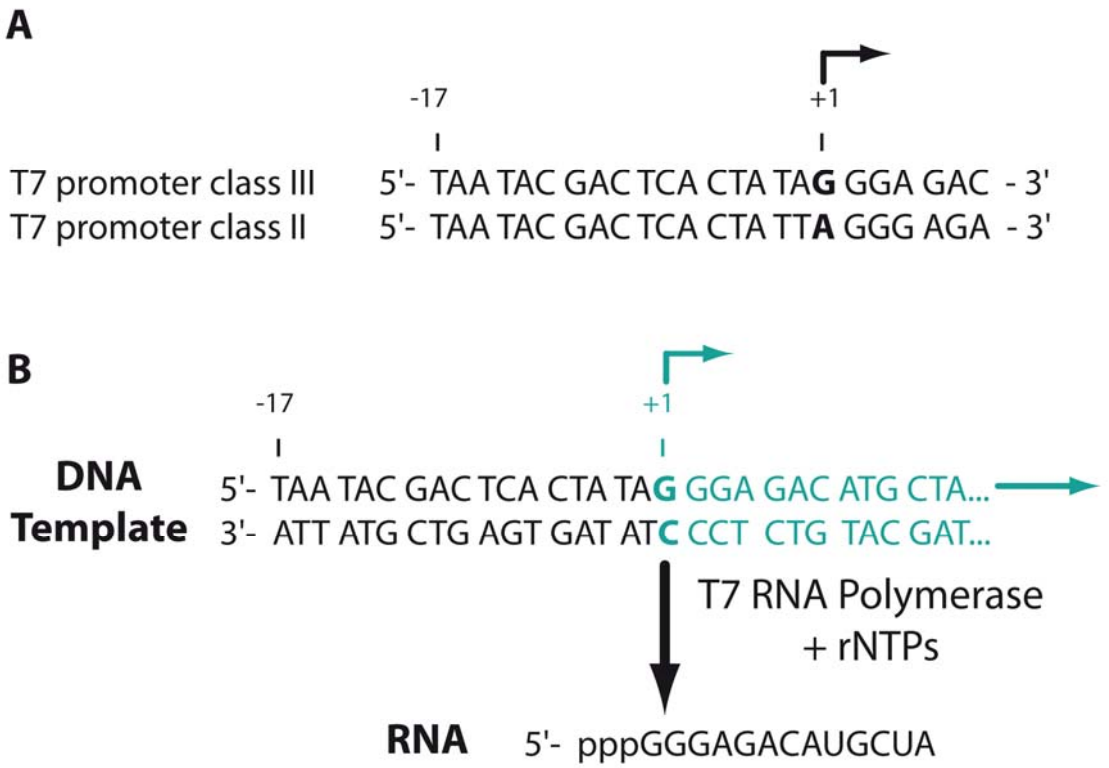


Figure 1

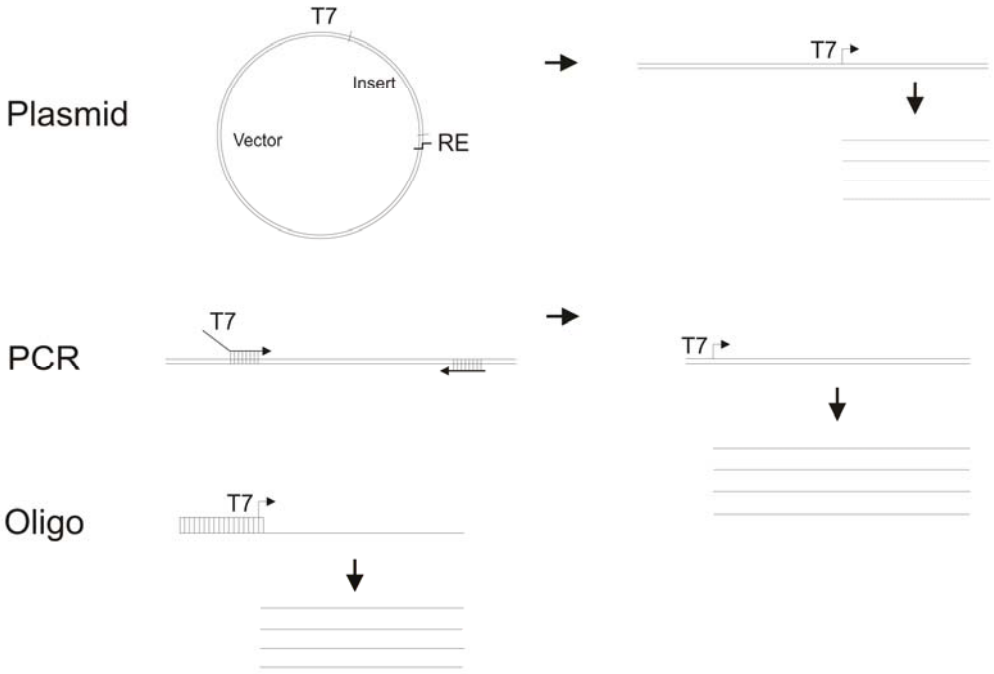


Figure 2

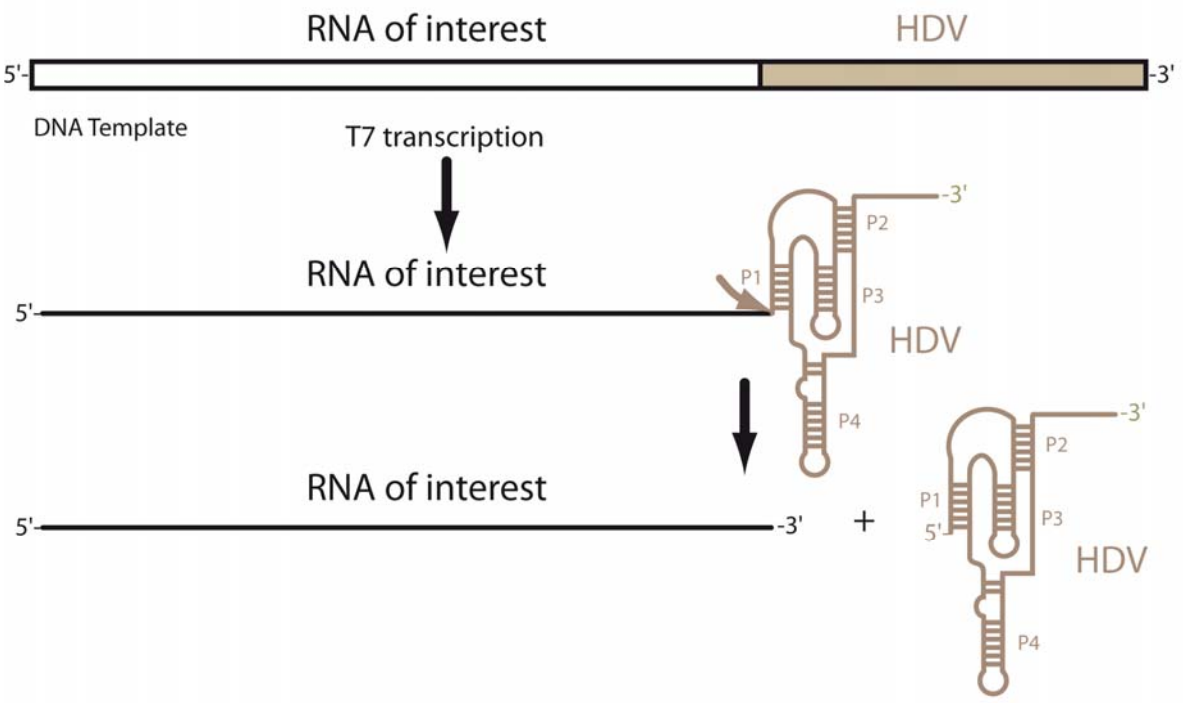


Figure 3

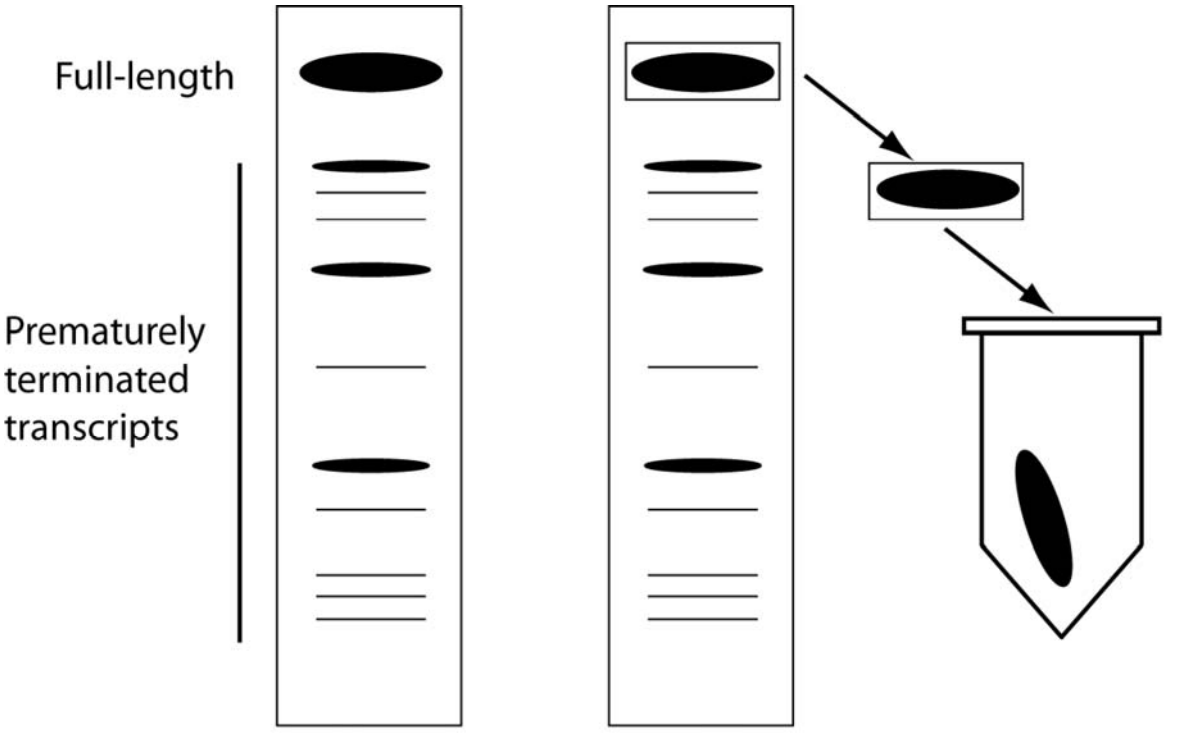


Figure 4

REFERENCES

1. Adams,P.L., Stahley,M.R., Kosek,A.B., Wang,J., and Strobel,S.A. (2004b). Crystal structure of a self-splicing group I intron with both exons. Nature *430*, 45-50.

2. Adams,P.L., Stahley,M.R., Kosek,A.B., Wang,J., and Strobel,S.A. (2004a). Crystal structure of a self-splicing group I intron with both exons. Nature *430*, 45-50.

3. Amaral,P.P., Dinger,M.E., Mercer,T.R., and Mattick,J.S. (2008). The eukaryotic genome as an RNA machine. Science *319*, 1787-1789.

4. Bartel,D.P. and Szostak,J.W. (1993). Isolation of new ribozymes from a large pool of random sequences [see comment]. Science *261*, 1411-1418.

5. Bartel,D.P. and Unrau,P.J. (1999). Constructing an RNA world. Trends Cell Biol *9*, M9-M13.

6. Bassi,G.S., de Oliveira,D.M., White,M.F., and Weeks,K.M. (2002). Recruitment of intron-encoded and co-opted proteins in splicing of the bI3 group I intron RNA. Proc Natl Acad Sci U S A *99*, 128-133.

7. Bassi,G.S. and Weeks,K.M. (2003). Kinetic and thermodynamic framework for assembly of the six-component bI3 group I intron ribonucleoprotein catalyst. Biochemistry *42*, 9980-9988.

8. Beagley,C.T., Okada,N.A., and Wolstenholme,D.R. (1996). Two mitochondrial group I introns in a metazoan, the sea anemone Metridium senile: one intron contains genes for subunits 1 and 3 of NADH dehydrogenase. Proc Natl Acad Sci U S A *93*, 5619-5623.

9. Beckert,B., Nielsen,H., Einvik,C., Johansen,S.D., Westhof,E., and Masquida,B. (2008). Molecular modelling of the GIR1 branching ribozyme gives new insight into evolution of structurally related ribozymes. EMBO J *27*, 667-678.

10. Belfort,M. (2003). Two for the price of one: a bifunctional intron-encoded DNA endonuclease-RNA maturase. Genes Dev *17*, 2860-2863.

11. Belfort,M. and Perlman,P.S. (1995a). Mechanisms of intron mobility. J Biol Chem *270*, 30237-30240.

12. Belfort,M. and Perlman,P.S. (1995b). Mechanisms of intron mobility. J Biol Chem *270*, 30237-30240.

13. Belfort,M. and Roberts,R.J. (1997). Homing endonucleases: keeping the house in order. Nucleic Acids Res *25*, 3379-3388.

14. Bevilacqua,P.C. (2008). Proton Transfer in Ribozyme Catalysis. In Ribozymes and RNA Catalysis, D.M.J.Lilley and Eckstein F., eds. (London, UK: The Royal Society of Chemistry), pp. 11-36.

15. Bhaskaran,H. and Russell,R. (2007). Kinetic redistribution of native and misfolded RNAs by a DEAD-box chaperone. Nature *449*, 1014-1018.

16. Birgisdottir,A.B. (2005). Site-specific reverse splicing of a HEG-containing group I intron in ribosomal RNA. Nucleic Acids Res *33 %6*, 2042-2051.

17. Birgisdottir,A.B. and Johansen,S. (2005). Site-specific reverse splicing of a HEG-containing group I intron in ribosomal RNA. Nucleic Acids Res *33*, 2042-2051.

18. Brehm,S.L. and Cech,T.R. (1983). Fate of an intervening sequence ribonucleic acid: excision and cyclization of the Tetrahymena ribosomal ribonucleic acid intervening sequence in vivo. Biochemistry *22*, 2390-2397.

19. Brion,P. and Westhof,E. (1997). Hierarchy and dynamics of RNA folding. Annu Rev Biophys Biomol Struct *26*, 113-137.

20. Brunel,C. and Romby,P. (2000). Probing RNA structure and RNA-ligand complexes with chemical probes. Methods Enzymol *318*, 3-21.

21. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M., Pande,N., Shang,Z., Yu,N., and Gutell,R.R. (2002). The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics *3*, 2.

22. Cao,Y. and Woodson,S.A. (1998). Destabilizing effect of an rRNA stem-loop on an attenuator hairpin in the 5' exon of the Tetrahymena pre-rRNA. Rna *4*, 901-914.

23. Caprara,M.G., Lehnert,V., Lambowitz,A.M., and Westhof,E. (1996a). A tyrosyl-tRNA synthetase recognizes a conserved tRNA-like structural motif in the group I intron catalytic core. Cell *87*, 1135-1145.

24. Caprara,M.G., Mohr,G., and Lambowitz,A.M. (1996b). A tyrosyl-tRNA synthetase protein induces tertiary folding of the group I intron catalytic core. J Mol Biol *257*, 512-531.

25. Caprara,M.G., Myers,C.A., and Lambowitz,A.M. (2001). Interaction of the Neurospora crassa mitochondrial tyrosyl-tRNA synthetase (CYT-18 protein) with the group I intron P4-P6 domain. Thermodynamic analysis and the role of metal ions. J Mol Biol *308*, 165-190.

26. Carpousis,A.J. (2007). The RNA degradosome of Escherichia coli: an mRNA-degrading machine assembled on RNase E. Annu Rev Microbiol *61*, 71-87.

27. Carpousis,A.J. (2002). The Escherichia coli RNA degradosome: structure, function and relationship in other ribonucleolytic multienzyme complexes. Biochem Soc Trans *30*, 150-155.

28. Cech,T.R. (1990). Self-splicing of group I introns. Annu Rev Biochem *59*, 543-568.

29. Cech,T.R., Damberger,S.H., and Gutell,R.R. (1994). Representation of the secondary and tertiary structure of group I introns. Nat Struct Biol *1*, 273-280.

30. Celander,D.W. and Cech,T.R. (1991). Visualizing the higher order folding of a catalytic RNA molecule. Science *251*, 401-407.

31. Celesnik,H., Deana,A., and Belasco,J.G. (2007). Initiation of RNA decay in Escherichia coli by 5' pyrophosphate removal. Mol Cell *27*, 79-90.

32. Chan,W.K., Belfort,G., and Belfort,M. (1988). Stability of group I intron RNA in Escherichia coli and its potential application in a novel expression vector. Gene *73*, 295-304.

33. Chauhan,S., Behrouzi,R., Rangan,P., and Woodson,S.A. (2009). Structural rearrangements linked to global folding pathways of the Azoarcus group I ribozyme. J Mol Biol *386*, 1167-1178.

34. Chauhan,S., Caliskan,G., Briber,R.M., Perez-Salas,U., Rangan,P., Thirumalai,D., and Woodson,S.A. (2005). RNA tertiary interactions mediate native collapse of a bacterial group I ribozyme. J Mol Biol *353*, 1199-1209.

35. Chauhan,S. and Woodson,S.A. (2008). Tertiary interactions determine the accuracy of RNA folding. J Am Chem Soc *130*, 1296-1303.

36. Chen,X., Gutell,R.R., and Lambowitz,A.M. (2000). Function of tyrosyl-tRNA synthetase in splicing group I introns: an induced-fit model for binding to the P4-P6 domain based on analysis of mutations at the junction of the P4-P6 stacked helices. J Mol Biol *301*, 265-283.

37. Chevalier,B.S. and Stoddard,B.L. (2001). Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. Nucleic Acids Res *29*, 3757-3774.

38. Clementi,N., Chirkova,A., Puffer,B., Micura,R., and Polacek,N. (2010). Atomic mutagenesis reveals A2660 of 23S ribosomal RNA as key to EF-G GTPase activation. Nat Chem Biol *6*, 344-351.

39. Clementi,N. and Polacek,N. (2010). Ribosome-associated GTPases: The role of RNA for GTPase activation. RNA Biol *7*.

40. Costa,M. and Michel,F. (1997). Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution. EMBO J *16*, 3289-3302.

41. Costa,M. and Michel,F. (1995). Frequent use of the same tertiary motif by self-folding RNAs. EMBO J *14*, 1276-1285.

42. Crick,F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.

43. Cruz,J.A. and Westhof,E. (2009). The dynamic landscapes of RNA architecture. Cell *136*, 604-609.

44. Damberger,S.H. and Gutell,R.R. (1994). A comparative database of group I intron structures. Nucleic Acids Res *22*, 3508-3510.

45. Das,R., Kwok,L.W., Millett,I.S., Bai,Y., Mills,T.T., Jacob,J., Maskel,G.S., Seifert,S., Mochrie,S.G., Thiyagarajan,P., Doniach,S., Pollack,L., and Herschlag,D. (2003). The fastest global events in RNA folding: electrostatic relaxation and tertiary collapse of the Tetrahymena ribozyme. J Mol Biol *332*, 311-319.

46. Deana,A., Celesnik,H., and Belasco,J.G. (2008). The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. Nature *451*, 355-358.

47. Decatur,W.A., Einvik,C., Johansen,S., and Vogt,V.M. (1995). Two group I ribozymes with different functions in a nuclear rDNA intron. EMBO J *14*, 4558-4568.

48. Doudna,J.A. and Cech,T.R. (2002). The chemical repertoire of natural ribozymes. Nature *418*, 222-228.

49. Doudna,J.A. and Lorsch,J.R. (2005). Ribozyme catalysis: not different, just worse. Nat Struct Mol Biol *12*, 395-402.

50. Downs,W.D. and Cech,T.R. (1990a). An ultraviolet-inducible adenosine-adenosine cross-link reflects the catalytic structure of the Tetrahymena ribozyme. Biochemistry *29*, 5605-5613.

51. Downs,W.D. and Cech,T.R. (1990b). An ultraviolet-inducible adenosine-adenosine cross-link reflects the catalytic structure of the Tetrahymena ribozyme. Biochemistry *29*, 5605-5613.

52. Draper,D.E. (2004). A guide to ions and RNA structure. Rna *10*, 335-343.

53. Draper,D.E., Grilley,D., and Soto,A.M. (2005). Ions and RNA folding. Annu Rev Biophys Biomol Struct *34*, 221-243.

54. Dror,O., Nussinov,R., and Wolfson,H. (2005). ARTS: alignment of RNA tertiary structures. Bioinformatics *21 Suppl 2*, ii47-ii53.

55. Duncan,C.D. and Weeks,K.M. (2008). SHAPE analysis of long-range interactions reveals extensive and thermodynamically preferred misfolding in a fragile group I intron RNA. Biochemistry *47*, 8504-8513.

56. Duncan,C.D. and Weeks,K.M. (2010). The Mrs1 splicing factor binds the bI3 group I intron at each of two tetraloop-receptor motifs. PLoS. One. *5*, e8983.

57. Dunckley,T. and Parker,R. (2001). Yeast mRNA decapping enzyme. Methods Enzymol *342*, 226-233.

58. Einvik,C., Decatur,W.A., Embley,T.M., Vogt,V.M., and Johansen,S. (1997). Naegleria nucleolar introns contain two group I ribozymes with different functions in RNA splicing and processing. Rna *3*, 710-720.

59. Einvik,C., Elde,M., and Johansen,S. (1998a). Group I twintrons: genetic elements in myxomycete and schizopyrenid amoeboflagellate ribosomal DNAs. J Biotechnol *64*, 63-74.

60. Einvik,C., Nielsen,H., Nour,R., and Johansen,S. (2000). Flanking sequences with an essential role in hydrolysis of a self-cleaving group I-like ribozyme. Nucleic Acids Res *28*, 2194-2200.

61. Einvik,C., Nielsen,H., Westhof,E., and Michel,F. (1998b). Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites. Rna *4 %6*, 530-541.

62. Einvik,C., Nielsen,H., Westhof,E., Michel,F., and Johansen,S. (1998c). Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites. Rna *4*, 530-541.

63. Ekland,E.H. and Bartel,D.P. (1996). RNA-catalysed RNA polymerization using nucleoside triphosphates. Nature *382*, 373-376.

64. Emerick,V.L. and Woodson,S.A. (1993). Self-splicing of the Tetrahymena pre-rRNA is decreased by misfolding during transcription. Biochemistry *32*, 14062-14067.

65. Emerick,V.L. and Woodson,S.A. (1994). Fingerprinting the folding of a group I precursor RNA. Proc Natl Acad Sci U S A *91*, 9675-9679.

66. Emory,S.A., Bouvet,P., and Belasco,J.G. (1992). A 5'-terminal stem-loop structure can stabilize mRNA in Escherichia coli. Genes Dev *6*, 135-148.

67. Ferre-D'Amare,A.R. (2010). Use of the spliceosomal protein U1A to facilitate crystallization and structure determination of complex RNAs. Methods.

68. Ferre-D'Amare,A.R. and Doudna,J.A. (2000). Crystallization and structure determination of a hepatitis delta virus ribozyme: use of the RNA-binding protein U1A as a crystallization module. J Mol Biol *295*, 541-556.

69. Ferre-D'Amare,A.R. and Rupert,P.B. (2002). The hairpin ribozyme: from crystal structure to function. Biochem Soc Trans *30*, 1105-1109.

70. Fire,A.Z. (2007). Gene silencing by double-stranded RNA (Nobel Lecture). Angew Chem Int Ed Engl *46*, 6966-6984.

71. Flores,R., Gas,M.E., Molina-Serrano,D., Nohales,M.A., Carbonell,A., Gago,S., and De La Pena,W.D.J.A. (2009). Viroid Replication: Rolling-Circles, Enzymes and Ribozyme. Viruses *1*, 317-334.

72. Flores,R., Hernandez,C., de la,P.M., Vera,A., and Daros,J.A. (2001). Hammerhead ribozyme structure and function in plant RNA replication. Methods Enzymol *341*, 540-552.

73. Forster,A.C. and Symons,R.H. (1987). Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. Cell *50*, 9-16.

74. Galburt,E.A. and Stoddard,B.L. (2002). Catalytic mechanisms of restriction and homing endonucleases. Biochemistry *41*, 13851-13860.

75. Garriga,G. and Lambowitz,A.M. (1986). Protein-dependent splicing of a group I intron in ribonucleoprotein particles and soluble fractions. Cell *46*, 669-680.

76. Goddard,M.R. and Burt,A. (1999a). Recurrent invasion and extinction of a selfish gene. Proc Natl Acad Sci U S A *96*, 13880-13885.

77. Goddard,M.R. and Burt,A. (1999b). Recurrent invasion and extinction of a selfish gene. Proc Natl Acad Sci U S A *96*, 13880-13885.

78. Golden,B.L. (2008). Group I Introns: Biochemical and Crystallographic Characterization of the Active Site Structure. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 178-200.

79. Golden,B.L. and Cech,T.R. (1996). Conformational switches involved in orchestrating the successive steps of group I RNA splicing. Biochemistry *35*, 3754-3763.

80. Golden,B.L., Gooding,A.R., Podell,E.R., and Cech,T.R. (1998). A preorganized active site in the crystal structure of the Tetrahymena ribozyme. Science *282*, 259-264.

81. Golden,B.L., Kim,H., and Chase,E. (2005). Crystal structure of a phage Twort group I ribozyme-product complex. Nat Struct Mol Biol *12*, 82-89.

82. Graber,D., Moroder,H., Steger,J., Trappl,K., Polacek,N., and Micura,R. (2010). Reliable semi-synthesis of hydrolysis-resistant 3'-peptidyl-tRNA conjugates containing genuine tRNA modifications. Nucleic Acids Res.

83. Guerrier-Takada,C. and Altman,S. (1984). Catalytic activity of an RNA molecule prepared by transcription in vitro. Science *223*, 285-286.

84. Guerrier-Takada,C., Gardiner,K., Marsh,T., Pace,N., and Altman,S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell *35*, 849-857.

85. Guhan,N. and Muniyappa,K. (2003). Structural and functional characteristics of homing endonucleases. Crit Rev Biochem Mol Biol *38*, 199-248.

86. Guo,F., Gooding,A.R., and Cech,T.R. (2004). Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site. Mol Cell *16*, 351-362.

87. Guo,Q. and Lambowitz,A.M. (1992). A tyrosyl-tRNA synthetase binds specifically to the group I intron catalytic core. Genes Dev *6*, 1357-1372.

88. Harland,R. and Misher,L. (1988). Stability of RNA in developing Xenopus embryos and identification of a destabilizing sequence in TFIIIA messenger RNA. Development *102*, 837-852.

89. Hasselmayer,O., Braun,V., Nitsche,C., Moos,M., Rupnik,M., and von Eichel-Streiber,C. (2004a). Clostridium difficile IStron CdISt1: discovery of a variant encoding two complete transposase-like proteins. J Bacteriol. *186*, 2508-2510.

90. Hasselmayer,O., Nitsche,C., Braun,V., and von Eichel-Streiber,C. (2004b). The IStron CdISt1 of Clostridium difficile: molecular symbiosis of a group I intron and an insertion element. Anaerobe. *10*, 85-92.

91. Haugen,P., Andreassen,M., Birgisdottir,A.B., and Johansen,S. (2004a). Hydrolytic cleavage by a group I intron ribozyme is dependent on RNA structures not important for splicing. Eur J Biochem *271*, 1015-1024.

92. Haugen,P., De Jonckheere,J.F., and Johansen,S. (2002). Characterization of the self-splicing products of two complex Naegleria LSU rDNA group I introns containing homing endonuclease genes. Eur J Biochem *269*, 1641-1649.

93. Haugen,P., Reeb,V., Lutzoni,F., and Bhattacharya,D. (2004b). The evolution of homing endonuclease genes and group I introns in nuclear rDNA. Mol Biol Evol *21*, 129-140.

94. Haugen,P., Simon,D.M., and Bhattacharya,D. (2005a). The natural history of group I introns. Trends Genet *21*, 111-119.

95. Haugen,P., Wikmark,O.G., Vader,A., Coucheron,D.H., Sjottem,E., and Johansen,S.D. (2005b). The recent transfer of a homing endonuclease gene. Nucleic Acids Res *33*, 2734-2741.

96. Heilman-Miller,S.L. and Woodson,S.A. (2003). Effect of transcription on folding of the Tetrahymena ribozyme. Rna *9*, 722-733.

97. Herschlag,D. (1995). RNA chaperones and the RNA folding problem. J Biol Chem *270*, 20871-20874.

98. Hougland ,J.L., Piccirilli,J.A., Forconi,A., Lee,J., and Herschlag,D. (2006). How the Group I Intron Works: A Case Study of RNA Structure and Function. In The RNA World, 3rd Ed., R.F.Gesteland, T.R.Cech, and J.F.Atkins, eds., pp. 133-205.

99. Hu,W., Sweet,T.J., Chamnongpol,S., Baker,K.E., and Coller,J. (2009). Co-translational mRNA decay in Saccharomyces cerevisiae. Nature *461*, 225-229.

100. Huang,F. and Yarus,M. (1997). 5'-RNA self-capping from guanosine diphosphate. Biochemistry *36*, 6557-6563.

101. Huang,Z. and Szostak,J.W. (1996). A simple method for 3'-labeling of RNA. Nucleic Acids Res *24*, 4360-4361.

102. Ichetovkin,I.E., Abramochkin,G., and Shrader,T.E. (1997). Substrate recognition by the leucyl/phenylalanyl-tRNA-protein transferase. Conservation within the enzyme family and localization to the trypsin-resistant domain. J Biol Chem *272*, 33009-33014.

103. Ikawa,Y., Naito,D., Aono,N., Shiraishi,H., and Inoue,T. (1999). A conserved motif in group IC3 introns is a new class of GNRA receptor. Nucleic Acids Res *27*, 1859-1865.

104. Ikawa,Y., Nohmi,K., Atsumi,S., Shiraishi,H., and Inoue,T. (2001). A comparative study on two GNRA-tetraloop receptors: 11-nt and IC3 motifs. J Biochem *130*, 251-255.

105. Inoue,T. and Cech,T.R. (1985). Secondary structure of the circular form of the Tetrahymena rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase. Proc Natl Acad Sci U S A *82*, 648-652.

106. Jabri,E., Aigner,S., and Cech,T.R. (1997). Kinetic and secondary structure analysis of Naegleria andersoni GIR1, a group I ribozyme whose putative biological function is site-specific hydrolysis. Biochemistry *36*, 16345-16354.

107. Jabri,E. and Cech,T.R. (1998). In vitro selection of the Naegleria GIR1 ribozyme identifies three base changes that dramatically improve activity. Rna *4*, 1481-1492.

108. Jackson,S., Cannone,J., Lee,J., Gutell,R., and Woodson,S. (2002). Distribution of rRNA introns in the three-dimensional structure of the ribosome. J Mol Biol *323*, 35-52.

109. Jackson,S.A., Koduvayur,S., and Woodson,S.A. (2006). Self-splicing of a group I intron reveals partitioning of native and misfolded RNA populations in yeast. Rna *12*, 2149-2159.

110. Jaeger,L., Michel,F., and Westhof,E. (1994). Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. J Mol Biol *236*, 1271-1276.

111. Johannes,G., Carter,M.S., Eisen,M.B., Brown,P.O., and Sarnow,P. (1999). Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. Proc Natl Acad Sci U S A *96*, 13118-13123.

112. Johansen,S., Einvik,C., and Nielsen,H. (2002). DiGIR1 and NaGIR1: naturally occurring group I-like ribozymes with unique core organization and evolved biological role. Biochimie *84*, 905-912.

113. Johansen,S., Elde,M., Vader,A., Haugen,P., and Haugli,K. (1997a). In vivo mobility of a group I twintron in nuclear ribosomal DNA of the myxomycete Didymium iridis. Mol Microbiol *24 %6*, 737-745.

114. Johansen,S., Elde,M., Vader,A., Haugen,P., Haugli,K., and Haugli,F. (1997b). In vivo mobility of a group I twintron in nuclear ribosomal DNA of the myxomycete Didymium iridis. Mol Microbiol *24*, 737-745.

115. Johansen,S., Embley,T.M., and Willassen,N.P. (1993). A family of nuclear homing endonucleases. Nucleic Acids Res *21*, 4405.

116. Johansen,S. and Haugen,P. (2001). A new nomenclature of group I introns in ribosomal DNA. Rna *7*, 935-936.

117. Johansen,S. and Vogt,V.M. (1994). An intron in the nuclear ribosomal DNA of Didymium iridis codes for a group I ribozyme and a novel ribozyme that cooperate in self-splicing. Cell *76*, 725-734.

118. Johnston,W.K., Unrau,P.J., Lawrence,M.S., Glasner,M.E., and Bartel,D.P. (2001). RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. Science *292*, 1319-1325.

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

120. Kjems,J., Egebjerg,J., and Christiansen,J. (1998). Laboratory Techniques in Biochemistry and Molecular Biology: Analysis of RNA-Protein Complexes In Vitro. Elsevier, Amsterdam, The Netherlands,).

121. Koo,S., Novak,T., and Piccirilli,J.A. (2008). Catalysis Mechanism of the HDV Ribozyme. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 92-122.

122. Kuhsel,M.G., Strickland,R., and Palmer,J.D. (1990). An ancient group I intron shared by eubacteria and chloroplasts. Science *250*, 1570-1573.

123. Laederach,A., Das,R., Vicens,Q., Pearlman,S.M., Brenowitz,M., Herschlag,D., and Altman,R.B. (2008). Semiautomated and rapid quantification of nucleic acid footprinting and structure mapping experiments. Nat Protoc *3*, 1395-1401.

124. Lambert,D. and Burke,J.M. (2008). Finding the Hammerhead Ribozyme Active Site. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 37-47.

125. Lambert,D. and Draper,D.E. (2007). Effects of osmolytes on RNA secondary and tertiary structure stabilities and RNA-Mg2+ interactions. J Mol Biol *370*, 993-1005.

126. Lambowitz,A.M. and Belfort,M. (1993). Introns as mobile genetic elements. Annu Rev Biochem *62*, 587-622.

127. Lambowitz,A.M. and Caprara,M.G. (1999). Group I and group II ribozymes as RNPs: Clues to the Past and Guides to the Future. In The RNA World, R.F.Gesteland, T.Cech, and J.F.Atkins, eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 451-485.

128. Latham,J.A. and Cech,T.R. (1989). Defining the inside and outside of a catalytic RNA molecule. Science *245*, 276-282.

129. Lawrence,M.S. and Bartel,D.P. (2005). New ligase-derived RNA polymerase ribozymes. Rna *11*, 1173-1180.

130. Lee,E.R., Baker,J.L., Weinberg,Z., Sudarsan,N., and Breaker,R.R. (2010). An allosteric self-splicing ribozyme triggered by a bacterial second messenger. Science *329*, 845-848.

131. Lee,N., Bessho,Y., Wei,K., Szostak,J.W., and Suga,H. (2000). Ribozyme-catalyzed tRNA aminoacylation. Nat Struct Biol *7*, 28-33.

132. Lehnert,V., Jaeger,L., and Michel,F. (1996). New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the Tetrahymena thermophila ribozyme. Chem Biol *3 %6*, 993-1009.

133. Lescoute,A. and Westhof,E. (2006). Topology of three-way junctions in folded RNAs. Rna *12*, 83-93.

134. Levinthal C. (1969). Proceeding of a Meeting held at Allerton House. In Mossbauer Spectroscopy in Biology Systems, DeBrunner J.T.P. and Munck E., eds. (Champaign-Urbana IL: University of Illinois Press), p. 22.

135. Li,Z. and Zhang,Y. (2005). Predicting the secondary structures and tertiary interactions of 211 group I introns in IE subgroup. Nucleic Acids Res *33*, 2118-2128.

136. Lilley,D.M. (2008a). Analysis of branched nucleic acid structure using comparative gel electrophoresis. Q. Rev Biophys *41*, 1-39.

137. Lilley,D.M.J. (2008b). The Hairpin and Varkud Satellite Ribozymes. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 66-91.

138. Lim,J., Grove,B.C., Roth,A., and Breaker,R.R. (2006). Characteristics of ligand recognition by a glmS self-cleaving ribozyme. Angew Chem Int Ed Engl *45*, 6689-6693.

139. Link,K.H. and Breaker,R.R. (2008). The Structure and Action of glmS Ribozyme. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 134-152.

140. Longo,A., Leonard,C.W., Bassi,G.S., Berndt,D., Krahn,J.M., Hall,T.M., and Weeks,K.M. (2005). Evolution from DNA to RNA recognition by the bI3 LAGLIDADG maturase. Nat Struct Mol Biol *12*, 779-787.

141. Lorsch,J.R. (2002). RNA chaperones exist and DEAD box proteins get a life. Cell *109*, 797-800.

142. Luptak,A., Ferre-D'Amare,A.R., Zhou,K., Zilm,K.W., and Doudna,J.A. (2001). Direct pK(a) measurement of the active-site cytosine in a genomic hepatitis delta virus ribozyme. J Am Chem Soc *123*, 8447-8452.

143. Luptak,A. and Szostak,J.W. (2008). Mammalian Self-Cleaving Ribozymes. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 123-132.

144. Maquez,S.M., Evans,D., Kazantsev,A.V., and Pace,N.R. (2008). A Structural Analysis of Ribonuclease P. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 153-177.

145. Mattick,J.S. (2009). Deconstructing the dogma: a new view of the evolution and genetic programming of complex organisms. Ann. N. Y. Acad Sci *1178*, 29-46.

146. Medina,M., Collins,A.G., Takaoka,T.L., Kuehl,J.V., and Boore,J.L. (2006). Naked corals: skeleton loss in Scleractinia. Proc Natl Acad Sci U S A *103*, 9096-9100.

147. Mello,C.C. (2007). Return to the RNAi world: rethinking gene expression and evolution (Nobel Lecture). Angew Chem Int Ed Engl *46*, 6985-6994.

148. Merino,E.J., Wilkinson,K.A., Coughlan,J.L., and Weeks,K.M. (2005). RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). J Am Chem Soc *127*, 4223-4231.

149. Michel,F. (1990). Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J Mol Biol *216 %6*, 585-610.

150. Michel,F., Ellington,A.D., Couture,S., and Szostak,J.W. (1990). Phylogenetic and genetic evidence for base-triples in the catalytic domain of group I introns. Nature *347*, 578-580.

151. Michel,F., Hanna,M., Green,R., Bartel,D.P., and Szostak,J.W. (1989). The guanosine binding site of the Tetrahymena ribozyme. Nature *342*, 391-395.

152. Michel,F., Jacquier,A., and Dujon,B. (1982). Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. Biochimie *64*, 867-881.

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

154. Milligan,J.F., Groebe,D.R., and Witherell,G.W. (1987). Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. Nucleic Acids Res *15 %6*, 8783-8798.

155. Mitchell,P., Petfalski,E., Shevchenko,A., Mann,M., and Tollervey,D. (1997). The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. Cell *91*, 457-466.

156. Mohr,G., Caprara,M.G., Guo,Q., and Lambowitz,A.M. (1994). A tyrosyl-tRNA synthetase can function similarly to an RNA structure in the Tetrahymena ribozyme. Nature *370*, 147-150.

157. Mohr,S., Stryker,J.M., and Lambowitz,A.M. (2002). A DEAD-box protein functions as an ATP-dependent RNA chaperone in group I intron splicing. Cell *109*, 769-779.

158. Moroder,H., Steger,J., Graber,D., Fauster,K., Trappl,K., Marquez,V., Polacek,N., Wilson,D.N., and Micura,R. (2009). Non-hydrolyzable RNA-peptide conjugates: a powerful advance in the synthesis of mimics for 3'-peptidyl tRNA termini. Angew Chem Int Ed Engl *48*, 4056-4060.

159. Mortimer,S.A. and Weeks,K.M. (2009). Time-resolved RNA SHAPE chemistry: quantitative RNA structure analysis in one-second snapshots and at single-nucleotide resolution. Nat Protoc *4*, 1413-1421.

160. Muhlrad,D., Decker,C.J., and Parker,R. (1995). Turnover mechanisms of the stable yeast PGK1 mRNA. Mol Cell Biol *15*, 2145-2156.

161. Murphy,F.L. and Cech,T.R. (1993). An independently folding domain of RNA tertiary structure within the Tetrahymena ribozyme. Biochemistry *32*, 5291-5300.

162. Nakano,S., Chadalavada,D.M., and Bevilacqua,P.C. (2000). General acid-base catalysis in the mechanism of a hepatitis delta virus ribozyme. Science *287*, 1493-1497.

163. Nielsen,H., Beckert,B., Masquida B., and Johansen,S.D. (2008). The GIR1 branching ribozyme. In Ribozymes and RNA catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 229-252.

164. Nielsen,H., Einvik,C., Lentz,T.E., Hedegaard,M.M., and Johansen,S.D. (2009). A conformational switch in the DiGIR1 ribozyme involved in release and folding of the downstream I-DirI mRNA. Rna *15*, 958-967.

165. Nielsen,H., Fiskaa,T., Birgisdottir,A.B., Haugen,P., Einvik,C., and Johansen,S. (2003). The ability to form full-length intron RNA circles is a general property of nuclear group I introns. Rna *9*, 1464-1475.

166. Nielsen,H. and Johansen,S.D. (2009). Group I introns: Moving in new directions. RNA Biol *6*, 375-383.

167. Nielsen,H., Westhof,E., and Johansen,S. (2005). An mRNA is capped by a 2', 5' lariat catalyzed by a group I-like ribozyme. Science *309*, 1584-1587.

168. Noller,H.F., Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Lancaster,L., Dallas,A., Fredrick,K., Earnest,T.N., and Cate,J.H. (2001). Structure of the ribosome at 5.5 A resolution and its interactions with functional ligands. Cold Spring Harb Symp Quant Biol *66*, 57-66.

169. Pan,J., Thirumalai,D., and Woodson,S.A. (1997). Folding of RNA involves parallel pathways. J Mol Biol *273*, 7-13.

170. Pan,J. and Woodson,S.A. (1998). Folding intermediates of a self-splicing RNA: mispairing of the catalytic core. J Mol Biol *280*, 597-609.

171. Parker,R. and Song,H. (2004). The enzymes and control of eukaryotic mRNA turnover. Nat Struct Mol Biol *11*, 121-127.

172. Paukstelis,P.J., Chen,J.H., Chase,E., Lambowitz,A.M., and Golden,B.L. (2008). Structure of a tyrosyl-tRNA synthetase splicing factor bound to a group I intron RNA. Nature *451*, 94-97.

173. Paukstelis,P.J., Coon,R., Madabusi,L., Nowakowski,J., Monzingo,A., Robertus,J., and Lambowitz,A.M. (2005). A tyrosyl-tRNA synthetase adapted to function in group I intron splicing by acquiring a new RNA binding surface. Mol Cell *17*, 417-428.

174. Piccirilli,J.A., Vyle,J.S., Caruthers,M.H., and Cech,T.R. (1993). Metal ion catalysis in the Tetrahymena ribozyme reaction. Nature *361*, 85-88.

175. Piccirillo,C., Khanna,R., and Kiledjian,M. (2003). Functional characterization of the mammalian mRNA decapping enzyme hDcp2. Rna *9*, 1138-1147.

176. Prathiba,J. and Malathi,R. (2008). Group I introns and GNRA tetraloops: remnants of 'The RNA world'? Mol Biol Rep *35*, 239-249.

177. Pyle,A.M. (2010). The tertiary structure of group II introns: implications for biological function and evolution. Crit Rev Biochem Mol Biol *45*, 215-232.

178. Pyle,A.M. (2008). Group II Introns: Catalysts for Splicing, Genomic Change and Evolution. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 201-228.

179. Rangan,P., Masquida,B., Westhof,E., and Woodson,S.A. (2004). Architecture and folding mechanism of the Azoarcus Group I Pre-tRNA. J Mol Biol *339*, 41-51.

180. Rangan,P., Masquida,B., Westhof,E., and Woodson,S.A. (2003). Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. Proc Natl Acad Sci U S A *100*, 1574-1579.

181. Rho,S.B. and Martinis,S.A. (2000). The bI4 group I intron binds directly to both its protein splicing partners, a tRNA synthetase and maturase, to facilitate RNA splicing activity. Rna *6*, 1882-1894.

182. Rodnina,M.V. (2008). Peptidyl Transferase Mechanism: The Ribosome as a Ribozyme. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 270-294.

183. Roman,J., Rubin,M.N., and Woodson,S.A. (1999). Sequence specificity of in vivo reverse splicing of the Tetrahymena group I intron. Rna *5*, 1-13.

184. Roman,J. and Woodson,S.A. (1995). Reverse splicing of the Tetrahymena IVS: evidence for multiple reaction sites in the 23S rRNA. Rna *1*, 478-490.

185. Roman,J. and Woodson,S.A. (1998). Integration of the Tetrahymena group I intron into bacterial rRNA by reverse splicing in vivo. Proc Natl Acad Sci U S A *95*, 2134-2139.

186. Rupert,P.B. and Ferre-D'Amare,A.R. (2001). Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. Nature *410*, 780-786.

187. Rupert,P.B., Xiao,H., and Ferre-D'Amare,A.R. (2003). U1A RNA-binding domain at 1.8 A resolution. Acta Crystallogr. D. Biol Crystallogr. *59*, 1521-1524.

188. Salehi-Ashtiani,K., Luptak,A., Litovchick,A., and Szostak,J.W. (2006). A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. Science *313*, 1788-1792.

189. Saville,B.J. and Collins,R.A. (1990). A site-specific self-cleavage reaction performed by a novel RNA in Neurospora mitochondria. Cell *61*, 685-696.

190. Schroeder,R., Barta,A., and Semrad,K. (2004). Strategies for RNA folding and assembly. Nat Rev Mol Cell Biol *5*, 908-919.

191. Sclavi,B., Sullivan,M., Chance,M.R., Brenowitz,M., and Woodson,S.A. (1998). RNA folding at millisecond intervals by synchrotron hydroxyl radical footprinting. Science *279*, 1940-1943.

192. Scott,W.G. (2008). Hammerhead Ribozyme Crystal Structure and Catalysis. In Ribozymes and RNA Catalysis, D.M.J.Lilley and F.Eckstein, eds. (London, UK: The Royal Society of Chemistry), pp. 48-65.

193. Shan,S., Kravchuk,A.V., Piccirilli,J.A., and Herschlag,D. (2001). Defining the catalytic metal ion interactions in the Tetrahymena ribozyme reaction. Biochemistry *40*, 5161-5171.

194. Shan,S., Yoshida,A., Sun,S., Piccirilli,J.A., and Herschlag,D. (1999). Three metal ions at the active site of the Tetrahymena group I ribozyme. Proc Natl Acad Sci U S A *96*, 12299-12304.

195. Shan,S.O. and Herschlag,D. (1999). Probing the role of metal ions in RNA catalysis: kinetic and thermodynamic characterization of a metal ion interaction with the 2'-moiety of the guanosine nucleophile in the Tetrahymena group I ribozyme. Biochemistry *38*, 10958-10975.

196. Shcherbakova,I. and Brenowitz,M. (2008). Monitoring structural changes in nucleic acids with single residue spatial and millisecond time resolution by quantitative hydroxyl radical footprinting. Nat Protoc *3*, 288-302.

197. Simon,D., Fewer,D., Friedl,T., and Bhattacharya,D. (2003). Phylogeny and self-splicing ability of the plastid tRNA-Leu group I Intron. J Mol Evol *57*, 710-720.

198. Steitz,T.A. and Steitz,J.A. (1993). A general two-metal-ion mechanism for catalytic RNA. Proc Natl Acad Sci U S A *90*, 6498-6502.

199. Suh,S.O., Jones,K.G., and Blackwell,M. (1999). A Group I intron in the nuclear small subunit rRNA gene of Cryptendoxyla hypophloia, an ascomycetous fungus: evidence for a new major class of Group I introns. J Mol Evol *48*, 493-500.

200. Suzuki,H., Zuo,Y., Wang,J., Zhang,M.Q., Malhotra,A., and Mayeda,A. (2006). Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. Nucleic Acids Res *34*, e63.

201. Szostak,J.W., Bartel,D.P., and Luisi,P.L. (2001). Synthesizing life. Nature *409*, 387-390.

202. Tanner,M. and Cech,T. (1996). Activity and thermostability of the small self-splicing group I intron in the pre-tRNA(lle) of the purple bacterium Azoarcus. Rna *2*, 74-83.

203. Tinoco,I., Jr. and Bustamante,C. (1999). How RNA folds. J Mol Biol *293*, 271-281.

204. Toor,N., Keating,K.S., Fedorova,O., Rajashankar,K., Wang,J., and Pyle,A.M. (2010). Tertiary architecture of the Oceanobacillus iheyensis group II intron. Rna *16*, 57-69.

205. Tourriere,H., Chebli,K., and Tazi,J. (2002). mRNA degradation machines in eukaryotic cells. Biochimie *84*, 821-837.

206. Treiber,D.K. and Williamson,J.R. (2001). Beyond kinetic traps in RNA folding. Curr Opin Struct Biol *11*, 309-314.

207. Tucker,B.J. and Breaker,R.R. (2005). Riboswitches as versatile gene control elements. Curr Opin Struct Biol *15*, 342-348.

208. Tucker,M. and Parker,R. (2000). Mechanisms and control of mRNA decapping in Saccharomyces cerevisiae. Annu Rev Biochem *69*, 571-595.

209. Turk,E.M. and Caprara,M.G. (2010). Splicing of yeast aI5beta group I intron requires SUV3 to recycle MRS1 via mitochondrial degradosome-promoted decay of excised intron ribonucleoprotein (RNP). J Biol Chem *285*, 8585-8594.

210. Unrau,P.J. and Bartel,D.P. (1998). RNA-catalysed nucleotide synthesis. Nature *395*, 260-263.

211. Uptain,S.M., Kane,C.M., and Chamberlin,M.J. (1997). Basic mechanisms of transcript elongation and its regulation. Annu Rev Biochem *66*, 117-172.

212. Vader,A., Johansen,S., and Nielsen,H. (2002). The group I-like ribozyme DiGIR1 mediates alternative processing of pre-rRNA transcripts in Didymium iridis. Eur J Biochem *269*, 5804-5812.

213. Vader,A. and Nielsen,H. (1999). In vivo expression of the nucleolar group I intron-encoded I-dirI homing endonuclease involves the removal of a spliceosomal intron. EMBO J *18 %6*, 1003-1013.

214. Vader,A., Nielsen,H., and Johansen,S. (1999). In vivo expression of the nucleolar group I intron-encoded I-dirI homing endonuclease involves the removal of a spliceosomal intron. EMBO J *18*, 1003-1013.

215. van Oppen,M.J., Catmull,J., McDonald,B.J., Hislop,N.R., Hagerman,P.J., and Miller,D.J. (2002). The mitochondrial genome of Acropora tenuis (Cnidaria; Scleractinia) contains a large group I intron and a candidate control region. J Mol Evol *55*, 1-13.

216. van,H.A. and Parker,R. (1999). The exosome: a proteasome for RNA? Cell *99*, 347-350.

217. Vicens,Q. and Cech,T.R. (2006). Atomic level architecture of group I introns revealed. Trends Biochem Sci *31*, 41-51.

218. Vicens,Q. and Cech,T.R. (2009). A natural ribozyme with 3',5' RNA ligase activity. Nat Chem Biol *5*, 97-99.

219. Vicens,Q., Paukstelis,P.J., Westhof,E., Lambowitz,A.M., and Cech,T.R. (2008). Toward predicting self-splicing and protein-facilitated splicing of group I introns. Rna *14*, 2013-2029.

220. Vincent,H.A. and Deutscher,M.P. (2006). Substrate recognition and catalysis by the exoribonuclease RNase R. J Biol Chem *281*, 29769-29775.

221. Waldsich,C., Grossberger,R., and Schroeder,R. (2002a). RNA chaperone StpA loosens interactions of the tertiary structure in the td group I intron in vivo. Genes Dev *16*, 2300-2312.

222. Waldsich,C., Masquida,B., Westhof,E., and Schroeder,R. (2002b). Monitoring intermediate folding states of the td group I intron in vivo. EMBO J *21*, 5281-5291.

223. Wallweber,G.J., Mohr,S., Rennard,R., Caprara,M.G., and Lambowitz,A.M. (1997). Characterization of Neurospora mitochondrial group I introns reveals different CYT-18 dependent and independent splicing strategies and an alternative 3' splice site for an intron ORF. Rna *3*, 114-131.

224. Wang,J.F. and Cech,T.R. (1992). Tertiary structure around the guanosine-binding site of the Tetrahymena ribozyme. Science *256*, 526-529.

225. Wang,J.F., Downs,W.D., and Cech,T.R. (1993). Movement of the guide sequence during RNA catalysis by a group I ribozyme. Science *260*, 504-508.

226. Weinstein,L.B., Jones,B.C., Cosstick,R., and Cech,T.R. (1997). A second catalytic metal ion in group I ribozyme. Nature *388*, 805-808.

227. Wikmark,O.G., Einvik,C., De Jonckheere,J.F., and Johansen,S.D. (2006). Short-term sequence evolution and vertical inheritance of the Naegleria twin-ribozyme group I intron. BMC Evol Biol *6*, 39.

228. Wilkinson,K.A., Merino,E.J., and Weeks,K.M. (2006). Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat Protoc *1*, 1610-1616.

229. Wilson,D.S. and Szostak,J.W. (1999). In vitro selection of functional nucleic acids. Annu Rev Biochem *68*, 611-647.

230. Winkler,W.C., Nahvi,A., Roth,A., Collins,J.A., and Breaker,R.R. (2004). Control of gene expression by a natural metabolite-responsive ribozyme. Nature *428*, 281-286.

231. Wong,T.N., Sosnick,T.R., and Pan,T. (2007). Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. Proc Natl Acad Sci U S A *104*, 17995-18000.

232. Woodson,S.A. (2000a). Recent insights on RNA folding mechanisms from catalytic RNA. Cell Mol Life Sci *57*, 796-808.

233. Woodson,S.A. (1992). Exon sequences distant from the splice junction are required for efficient self-splicing of the Tetrahymena IVS. Nucleic Acids Res *20*, 4027-4032.

234. Woodson,S.A. (2005). Metal ions and RNA folding: a highly charged topic with a dynamic future. Curr Opin Chem Biol *9*, 104-109.

235. Woodson,S.A. (2000b). Compact but disordered states of RNA. Nat Struct Biol *7*, 349-352.

236. Woodson,S.A. and Cech,T.R. (1989). Reverse self-splicing of the tetrahymena group I intron: implication for the directionality of splicing and for intron transposition. Cell *57*, 335-345.

237. Woodson,S.A. and Cech,T.R. (1991). Alternative secondary structures in the 5' exon affect both forward and reverse self-splicing of the Tetrahymena intervening sequence RNA. Biochemistry *30*, 2042-2050.

238. Wu,H.N., Lin,Y.J., Lin,F.P., Makino,S., Chang,M.F., and Lai,M.M. (1989). Human hepatitis delta virus RNA subfragments contain an autocleavage activity. Proc Natl Acad Sci U S A *86*, 1831-1835.

239. Xu,M.Q., Kathe,S.D., Goodrich-Blair,H., Nierzwicki-Bauer,S.A., and Shub,D.A. (1990). Bacterial origin of a chloroplast intron: conserved self-splicing group I introns in cyanobacteria. Science *250*, 1566-1570.

240. Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H., and Noller,H.F. (2001). Crystal structure of the ribosome at 5.5 A resolution. Science *292*, 883-896.

241. Yusupova,G., Yusupov,M., Spirin,A., Ebel,J.P., Moras,D., Ehresmann,C., and Ehresmann,B. (1991). Formation and crystallization of Thermus thermophilus 70S ribosome/tRNA complexes. FEBS Lett. *290*, 69-72.

242. Zarrinkar,P.P. and Sullenger,B.A. (1998). Probing the interplay between the two steps of group I intron splicing: competition of exogenous guanosine with omega G. Biochemistry *37*, 18056-18063.

243. Zarrinkar,P.P. and Williamson,J.R. (1994). Kinetic intermediates in RNA folding. Science *265*, 918-924.

244. Zarrinkar,P.P. and Williamson,J.R. (1996). The kinetic folding pathway of the Tetrahymena ribozyme reveals possible similarities between RNA and protein folding. Nat Struct Biol *3*, 432-438.

245. Zaug,A.J., Grabowski,P.J., and Cech,T.R. (1983). Autocatalytic cyclization of an excised intervening sequence RNA is a cleavage-ligation reaction. Nature *301*, 578-583.

246. Zhang,A., Rimsky,S., Reaban,M.E., Buc,H., and Belfort,M. (1996). Escherichia coli protein analogs StpA and H-NS: regulatory loops, similar and disparate effects on nucleic acid dynamics. EMBO J *15*, 1340-1349.

247. Zhang,B. and Cech,T.R. (1997). Peptide bond formation by in vitro selected ribozymes. Nature *390*, 96-100.

248. Zhang,F., Ramsay,E.S., and Woodson,S.A. (1995). In vivo facilitation of Tetrahymena group I intron splicing in Escherichia coli pre-ribosomal RNA. Rna *1*, 284-292.