Structural studies on Xist RNA in X-inactivation

Malgorzata M. Duszczyk

Structural studies on Xist RNA in X-inactivation

Een wetenschappelijke proeve
op het gebied van de Natuurwetenschappen,
Wiskunde en Informatica

Proefschrift

ter verkrijging van de graad van doctor aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
mede onder auspiciën van het European Molecular Biology Laboratory (EMBL) te
Heidelberg, Duitsland,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op woensdag 6 mei 2009
om 13.30 uur precies

door

Malgorzata Maria Duszczyk

geboren op 26 mei 1978
te Warschau, Polen

Moim rodzicom
Aan mijn ouders

voor Frederik

Preface

This thesis describes the first structural investigation into the molecular mechanism of X-chromosome inactivation. The structure of a novel RNA tetraloop motif within the A-repeats of Xist RNA, a non-coding RNA essential for X-inactivation, is solved by nuclear magnetic resonance (NMR). Additionally, a novel NMR method to unambiguously distinguish between RNA hairpin and duplex conformations is presented. A-repeat duplexes detected by this method *in vitro*, are shown to possibly have a role in X-inactivation *in vivo*. A model is presented in which dimerization of the Xist RNA A-repeats and specific recognition of the novel tetraloop motif contribute to mediate X-inactivation.

Cover Illustration

The cover illustration shows a tortoise-shell, or calico cat flying on an artist's interpretation of the inactive X-chromosome. X-inactivation is initiated by Xist RNA painting the X-chromosome to be silenced. This is visualized in pink, resembling RNA FISH (fluorescence in situ hybridization) images, while DNA is counterstained in blue. The cat's unique coat pattern is a result of X-inactivation. As the gene for either black or red coat color lies on the X-chromosome, random inactivation of one of the two X-chromosomes results in patches of red and black. Males only carry one copy of the X-chromosome, so this occurs exclusively in female cats that carry one each of the two different coat color genes.

Contents

List of abbreviations and symbols

3D	three-dimensional
2D	two-dimensional
Air	Antisense Igf2r RNA
BMRB	Biological Magnetic Resonance Bank
°C	degree Celsius
COSY	Correlation Spectroscopy
CT	Constant Time
δ	chemical shift
DNA	Deoxyribo Nucleic Acid
dsRNA	double stranded RNA
DQF-COSY	Double Quantum Filtered Correlation Spectroscopy
E.COSY	Exclusive Correlation Spectroscopy
EDTA	Ethylene Diamine Tetraacetic Acid
ES cells	Embryonic Stem cells
Ezh2	Enhancer of zeste homologue 2
FID	Free Induction Decay
H3-K27	Histone H3 lysine 27
HMQC	Heteronuclear Multiple Quantum Correlation
HSQC	Heteronuclear Single Quantum Correlation
Igf2r	Insulin-like growth factor 2 receptor
INEPT	Insensitive Nuclei Enhanced by Polarization Transfer
kb	kilobase(pairs) = 10^3 base(s)(pairs)
kHz	kilohertz = 10^3 hertz
MDSA	Molecular Dynamics Simulated Annealing
MHz	Megahertz = 10^6 hertz
miRNA	microRNA
MW	Molecular Weight

MWCO	Molecular Weight Cutoff
ncRNA	non-coding RNA
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect Spectroscopy
NTP	Nucleotide Triphosphate
PAGE	Polyacrylamide Gel Electrophoresis
PRC1/2	Polycomb Repressive Complex 1/2
PDB	Protein Data Bank
Pf1	filamentous phage 1
ppm	parts per million (=10^{-6})
RDC	Residual Dipolar Coupling
RISC	RNA-Induced Silencing Complex
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
RNAi	RNA interference
rox1/2	RNA on the X 1/2
rRNA	ribosomal RNA
siRNA	small interfering RNA
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
TBE	Tris-Borate-EDTA buffer
tRNA	transfer RNA
UV	Ultraviolet
Xa	active X-chromosome
XCI	X-chromosome Inactivation
Xi	inactive X-chromosome
Xic	X-inactivation centre
xiRNA	x-inactivating RNA
Xist	X-inactivation specific transcript

1

General introduction and scope
of the thesis

Chapter 1

General Introduction

RNA mediated regulation of gene expression

Understanding the molecular and genetic mechanisms of regulation of gene expression is fundamental to all areas of biology. Gene expression regulation plays a crucial role in diverse biological phenomena as homeostasis, development, adaptation to the environment and the evolution of multicellular organisms. In these organisms it drives cellular differentiation and morphogenesis: although all cells possess the same genome sequence, different gene expression profiles allow them to develop into different cell types. Disease often involves misregulation of gene expression.

For decades, the central dogma of molecular biology has been that RNA acts as a mere messenger between genetic information and protein expression. Gene regulation was thought to be controlled by protein factors, mostly at the transcriptional level. Although still unchallenged generally, the central dogma has become less comprehensive after an explosion of recent discoveries of non-coding RNAs (ncRNAs) that regulate gene expression.

The idea that sequence-specific non-coding RNA might interact with promoters to regulate genes originates in the earliest days of molecular biology (1, 2). Since then, it took forty years for the discovery of RNA interference (RNAi) (3) to start to revolutionize the understanding of the role of RNA in gene regulation and this process still continues today.

Much of the transcriptional output of eukaryotic genomes consists of RNA that does not encode protein. For example, it has been established that 62% of the mouse genome is transcribed (4) and that of 181.000 unique transcripts, half are ncRNAs. It is speculated that the increased complexity of higher eukaryotes is hidden in this non-coding output. The idea that phenotypic variation between

species (and individuals) results from differences in RNA regulatory networks that control protein expression, not the protein themselves, is an appealing idea, as comparison of genome sizes and protein coding gene numbers does not seem to explain the difference between simple and complex organisms. Humans only have about two or three times as many protein coding genes as the nematode *Caenorhabditis elegans* or the fruit fly *Drosophila melanogaster*, which, in turn, only have about twice as many as the yeast *Saccharomyces cerevisiae*. The homology within protein coding genes between the human and mouse genomes is estimated to be an astonishing 99% (5).

Non-coding RNAs play various roles in prokaryotic and eukaryotic cells. They can be broadly divided into two classes (reviewed in 6): housekeeping RNAs are generally constitutively expressed and required for normal function and viability of the cell. An example are RNAs involved in RNA processing (snRNA, snoRNA), and translation (tRNA, rRNA). In contrast, regulatory non-coding RNAs are expressed at certain stages in development and cell differentiation, or as a response to external stimuli. They can affect the expression of other genes at the level of transcription or translation.

Examples of RNAs involved in translational, or post-transcriptional gene regulation are small regulatory RNAs like short interfering RNAs (siRNAs) or microRNAs (miRNAs) (reviewed in 7). They repress or stimulate translation of mRNAs via antisense RNA:RNA interactions. Long dsRNA (double-stranded RNA) and miRNA precursors are processed to siRNA/miRNA duplexes by the RNase-III-like enzyme Dicer (8). These short dsRNAs are subsequently unwound and assembled into RNA-induced Silencing Complexes (RISCs), which can direct RNA cleavage, mediate translational repression (or activation) and induce chromatin modifications (9-11). Their effect is the regulation of the activity of particular genes with similar sequences to the short RNAs.

RNAs involved in transcriptional regulation are often large. An example is the ~108 kb *Air* ncRNA that is involved in the silencing of the imprinted Igf2r (insulin-like growth factor 2 receptor) gene cluster in the mouse (12). Imprinting is

the preferential expression of genes from one parental allele (reviewed in 13). Other examples are the Drosophila *rox1* (3.7 kb) and *rox2* (1.2 kb) ncRNAs and the mammalian *Xist* (15-17 kb) and *Tsix* (40 kb) ncRNAs. These RNAs are involved in chromatin remodeling associated with dosage compensation (reviewed in 14) and have, different from the earlier described short RNAs, the remarkable ability to associate with chromatin along entire chromosomes in a way that is not sequence specific.

Dosage compensation is a mechanism that evolved to compensate for the difference in X-linked gene expression in species where males and females have different numbers of X-chromosomes. Different solutions to this problem have been achieved in different parts of the animal kingdom. In *Drosophila*, the single male X-chromosome transcribes at a higher rate. In mammals one of the two X-chromosomes of the females is inactivated early in development in a process called X-chromosome inactivation (XCI) (see Fig. 1).

A remarkable similarity between *Drosophila* and mammalian dosage compensation lies in the involvement of the aforementioned large non-coding RNAs: hyperactivity of the X in male *Drosophila* is dependent upon the entire X being painted with the non-coding RNAs *rox1* & *rox2*. Mammalian X-inactivation depends on the non-coding RNA *Xist* that coats the inactive X *in cis*. Both rox and Xist RNAs encompass entire chromosomes and are able to induce epigenetic modifications on these chromosomes to achieve dosage compensation of gene expression (reviewed in 15).

X-chromosome inactivation as a paradigm for epigenetic gene regulation

The term 'epigenetic' defines heritable changes in gene expression which are not coded for in the DNA. Epigenetic mechanisms include histone tail modifications (methylation, acetylation, phosphorylation and ubiquitination) (reviewed in 16),
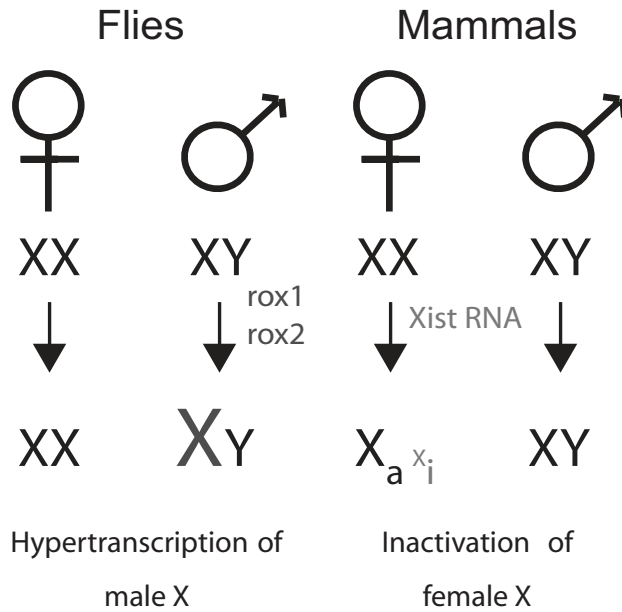
Figure 1 Flies and mammals use different strategies to equalize X-linked gene expression between males and females. While in *Drosophila* the male X is hypertranscribed depending on the rox RNAs, in mammals Xist RNA accumulates *in cis* on one of the female X chromosomes and initiates a silencing process that eventually converts the entire chromosome into a heterochromatic and by large transcriptionally inactive X chromosome (Xi). The other chromosome remains active (Xa) as there *Xist* is repressed.

DNA methylation and RNA-associated silencing. These are mutually related and implicated in initiating changes in chromatin structure resulting in a repressed or activated state that can be sustained through cell division (reviewed in 17).

XCI has been studied extensively as a paradigm for epigenetic gene regulation as it requires a diverse range of epigenetic mechanisms: non-coding RNAs, antisense transcription, histone modifications, and DNA methylation, to selectively silence one of two identical chromosomes within the same nucleus.

Two forms of XCI have been described: in imprinted XCI, limited to extraembryonic tissues in the mouse and early mammals like marsupials, the paternal X-chromosome is preferentially inactivated (reviewed in 18). In the embryo, random XCI occurs. Due to this random inactivation female mammals are mosaic for X-linked traits.

Random XCI is achieved in three stages: first, the X-chromosome-to-autosome ratio is 'counted' to ensure that a single X-chromosome remains active in a cell with a diploid autosomal set. Next, one X-chromosome is 'chosen' to be the future inactive X (Xi) (reviewed in 19). Once the choice is made, silencing is initiated by coating of the future Xi by the ncRNA *Xist*, followed by transcriptional shutdown by exclusion of the transcription machinery from the Xi (20). Subsequently, silencing factors are recruited, and the X-chromatin is condensed. This condensation involves an ordered series of chromatin modifications (see Fig. 2): just after Xist RNA coating, recruitment of the protein complexes PRC2 and PRC1 (the Polycomb Repressive protein Complexes, known to be associated with transcriptional repression of developmental control genes – reviewed in 21) by *Xist* establishes the chromosome-wide histone marks H3K27 trimethylation (22) and H2AK119 ubiquitination (23), respectively. Furthermore, H4K20 methylation and the histone variant macroH2A are enriched on the Xi. Histone H4 is hypoacetylated (reviewed in 24).

Initially, gene silencing is reversible and dependent on *Xist,* but at a later stage in differentiation, X-inactivation becomes independent of *Xist* and irreversible: in the long term, the silenced state of the Xi is maintained through future cell divisions by DNA methylation, a highly persistent epigenetic mark associated with transcriptional silence (25).
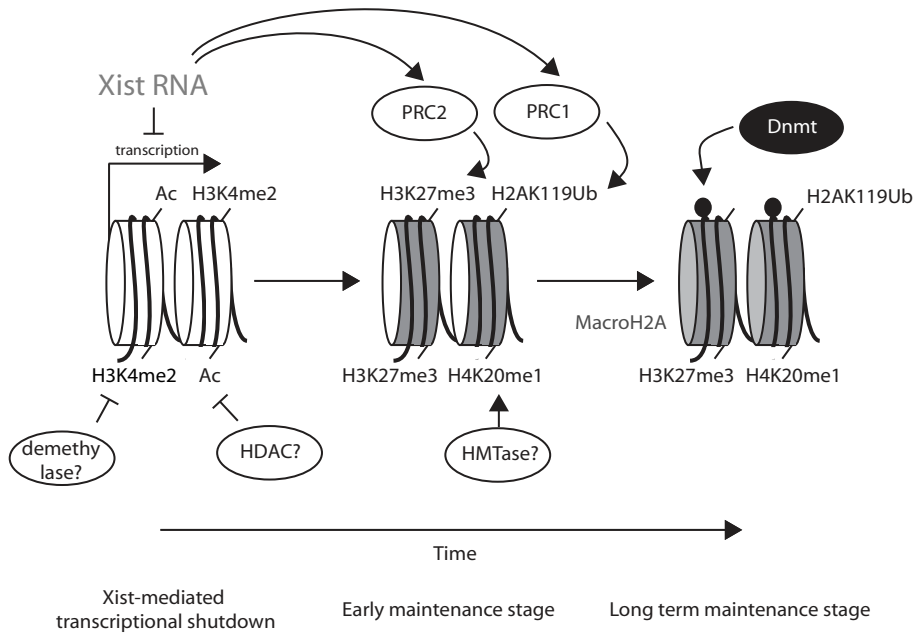
Figure 2 Timeline of epigenetic events on the X-chromosome on the basis of findings in differentiating mouse embryonic stem cells. Expression of Xist RNA immediately shuts down transcription through a not yet elucidated mechanism. Xist RNA coating of the Xi is followed by loss of euchromatic histone marks, through the action of yet unidentified histone deacetylases and demethylases. During this same time window, X-linked gene-silencing initiates. Several histone modifications become enriched on the Xist RNA-coated chromosome at this time. These include H3K27me3, H2AK119ub1 and H4K20me1. The Polycomb complexes PRC1 and PRC2 are recruited at this time in a Xist RNA-dependent manner. The PRC2 complex component Ezh2 is responsible for the appearance of H3K27me3 on the Xi. Mono-ubiquitination of histone H2A is induced by the PRC1 complex. At later stages of differentiation, the PRC2 and PRC1 complexes no longer appear to be present on the Xi, and H3K27 trimethylation levels go down. However, macroH2A becomes associated with the inactive X. The latest mark to appear is DNA methylation of promoters of X-linked genes. This long term maintenance inactivated state is Xist independent and irreversible.

Regulation of X-chromosome inactivation by Xist RNA and its antisense Tsix transcript

To ensure that one X-chromosome stays active, XCI is tightly regulated from a single locus on the X-chromosome called the 'X-inactivation center' (*Xic*). The Xic comprises several genes for ncRNAs including *Xist* and *Tsix* (reviewed in 26). *Xist* RNA is essential for the initiation of X-inactivation (27). It is expressed exclusively from the X-chromosome to be silenced (28) and coats it *in cis* (29), which coincides with transcriptional shutdown by an unknown mechanism.

Xist is negatively regulated by its antisense RNA *Tsix* (30). On the future Xa, persistence of *Tsix* expression prevents upregulation of *Xist*. On the future Xi, *Tsix* expression is lost, which permits upregulation of *Xist* and the initiation of silencing. Studies in mouse ES cells have elucidated two possible mechanisms for how *Tsix* could regulate *Xist*: it has been shown that *Tsix* transcription alters the chromatin structure by modifying histone tails and methylating the DNA at the *Xist* promoter (31-33). In addition, it has been shown that *Xist* and *Tsix* form RNA duplexes *in vivo*, which are processed to small RNAs that have a potential regulatory function (34). The authors propose that these small RNAs (named xiRNAs) could repress *Xist* in an RNAi like manner on the Xa. They show that xiRNA levels are dependent on Dicer and that Dicer, and thus the RNAi pathway, is required to localize *Xist* and target H3K27 methylation to the Xi, indicating a global role for RNA interference in XCI. This apparent link between the RNAi and XCI pathways awaits further investigation.

The Xist RNA A-repeats are essential for silencing

The chromosomal association and gene silencing abilities of Xist RNA are functionally separable (35). Silencing requires a repeat region known as the A-repeats, located at the 5' end of the *Xist* transcript. In the absence of the A-repeats Xist RNA localizes to and spreads over the X-chromosome but does not induce silencing. Association of Xist RNA with chromatin is a prerequisite for silencing and is mediated by other sequences that act synergistically, which are functionally redundant and dispersed throughout the remainder of the transcript (see Chapter 4 Fig. 1).

The A-repeats contain 7.5 copies of a motif that is almost fully conserved between placental mammals and which was predicted to fold into a double hairpin structure, connected by long U-rich linkers (see Chapter 4 Fig. 1). It has long been considered that the A-repeats provide a binding platform for factors that act in gene repression (36), but the exact mechanism of A-repeat mediated silencing remained unknown because no specific interaction factors had been isolated.

Now, a recent report has shown that the Polycomb complex PRC2 is a direct target for the A-repeats (37). The authors show that before the onset of X-inactivation, smaller, separate transcripts of A-repeat RNA initially recruit the PRC2 complex to the X-chromosome, with Ezh2 (Enhancer of zeste homologue 2) serving as the RNA binding subunit. Ezh2 is known to be a H3K27 specific methyltransferase (38) and to bind DNA. Its domain architecture is not defined apart from a single SET (Su(var)3-9, Enhancer of zeste, Trithorax) domain at the C-terminus, a signature motif for histone lysine methyltransferases, and two SANT (SWI3, ADA2, N-CoR and TFIIIB) domains, known to be DNA binding domains.

The authors show that depletion of the A-repeats abolishes full-length *Xist* induction and H3K27 methylation on the Xi. Inversely, depletion of PRC2 compromises *Xist* upregulation. The authors propose that the recruitment of PRC2 by the shorter A-repeat RNA is required for the initiation and spread of XCI. They

further show that *Tsix* RNA inhibits this interaction by competing for PRC2 binding. In their model for A-repeat activity the A-repeats are able to recruit PRC2 after *Tsix* is down-regulated on the future Xi, followed by methylation of the *Xist* promoter which enables full-length *Xist* transactivation. As not only the shorter A-repeat RNA, but also full-length *Xist* binds PRC2, they propose that the spread of Xist RNA along Xi could distribute PRC2 and H3K27 methylation throughout the chromosome.

This story seems to contradict earlier reports that PRC2 is recruited by *Xist* in an A-repeat independent fashion (22). Another report has shown that that H3K27 methylation can be established independent of silencing, by Xist RNA lacking the A-repeats (39). It is also not clear how this model could explain the observation that Xist RNA lacking the A-repeats still is induced and is able to localize to the X-chromosome (35).

Perspectives

X-inactivation has been studied intensively for many years, and progress in understanding the mechanism has been made: several proteins and epigenetic modifications that contribute to silence the Xi have been identified. Only recently proteins that specifically bind the *Xist* RNA have been reported. Although these recent findings are promising, the exact molecular mechanisms that underlie *Xist's* capacity to induce transcriptional silencing, how *Xist* directs the cascade of chromatin changes to the Xi, and the exact role of the A-repeats within *Xist* remain unclear.

At present, no structural information on Xist RNA and the A-repeats is available.

The role of RNA structure in the functional versatility of RNA

The huge versatility in RNA structure and function depends on the chemical properties of RNA that, in contrast to DNA, allow it to form complex tertiary structures based on 2' hydroxyls lining the minor groove of the A-form RNA helix (40). These hydroxyls are available for tertiary interactions. Assembled into higher order structures, RNAs are capable of performing roles that for many years were thought to be exclusive to proteins: they participate in recognition processes and can catalyze numerous chemical reactions in the cell. The ability to perform structural functions, combined with the ability to base pair with genetic information results in a unique versatility of actions in different biological processes.

The recent discoveries of so many new functional RNA molecules are a motivation to analyze new ways in which RNAs can fold, interact with other molecules in recognition processes and be catalytically active. Two methods are available for structural biologists to study RNA structure at the atomic level: X-ray crystallography and Nuclear Magnetic Resonance (NMR).

As RNA molecules are flexible, they are often difficult to crystallize. Moreover, RNA crystal structures are particularly affected by crystal packing forces as RNAs have higher surface areas compared to globular proteins. RNA hairpin structures tend to form biologically irrelevant duplex structures under high concentrations and in high salt conditions (41) that are often required for crystallization. Solution structures, which can be determined by NMR, are therefore particularly important in RNA structural biology as a complement to crystallography. NMR is also particularly suitable to study the dynamic behavior of regions of higher conformational flexibility in nucleic acids (reviewed in 42), and their interactions with other biological macromolecules.

However, the application of NMR to RNA has been limited to molecules less than ~25 kDa because of resonance overlap and signal broadening in NMR spectra of large RNAs. Luckily, biological RNAs can be dissected into structural subunits that stay within this size limit and are still functional in interactions with

proteins, drugs, or other RNAs. These building blocks, or structural motifs, are accessible for NMR structure determination. The next section will describe some of the most important building blocks of RNA structure (for review see 43, 44).

Building blocks of RNA structure

Base pairing

Apart from standard Watson-Crick base pairing in G-C and A-U base pairs, RNA structures have revealed a wide range of so-called non-canonical base pairs (reviewed in 45). When these 'mismatched' pairs are involved in stacking within a RNA helix, they provide recognition sites by presenting functional groups at the base edges and through distortion of the A-form helix (46). G-U wobble pairs and G-A mismatches are the most commonly observed non-canonical base pairs, but in principle all combinations of A, C, G and U are possible if *syn/anti* glycosidic torsion angles and parallel/anti-parallel sugar phosphate backbones are allowed.

Nucleobases have 3 edges where hydrogen-bonding can occur: the Watson-Crick edge, the Hoogsteen edge and the shallow groove edge. The bases can approach each other so that the sugars are on the same side (*cis*) or on opposite sides (*trans,* or *reverse*) of a line median to the hydrogen bonds (see Fig. 3). The local orientation of the backbones may be parallel or antiparallel. (47). Finally, base pairing occurs not only in pairs, but also in triples and quadruples, which can facilitate tertiary interactions.

The A-form helix

Consecutive stacking of canonical Watson-Crick base pairs gives rise to the A-form double helix, the dominating canonical conformation of double stranded RNA. Compared to the dominant B-form in DNA, the A-form helix is shorter and wider, with a deep and narrow major groove and a wide and shallow minor groove that is
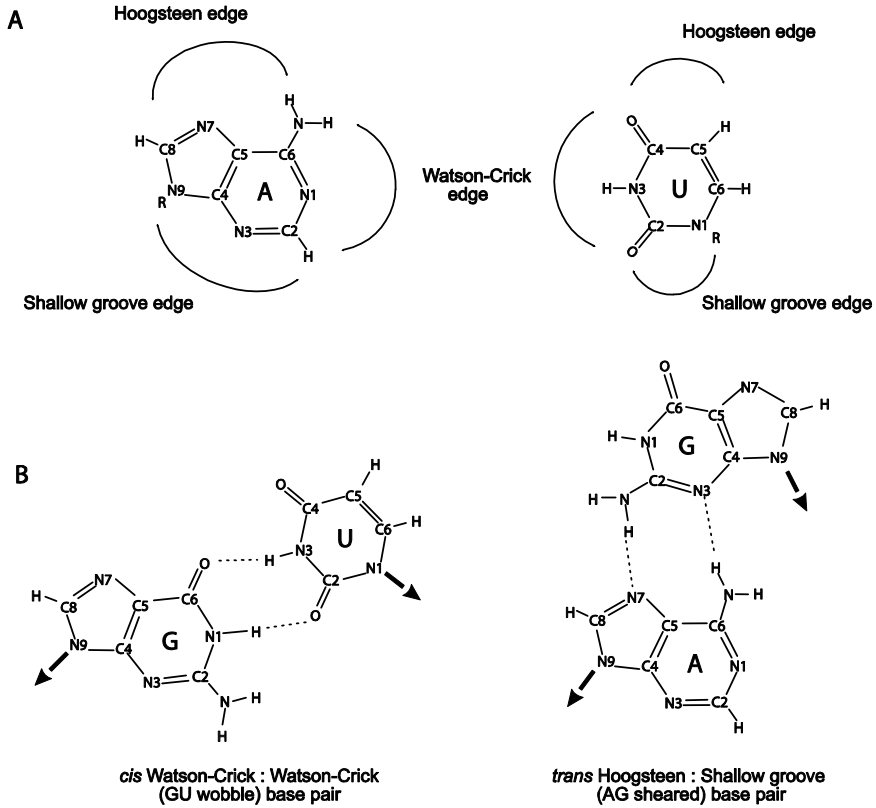
Figure 3 A) the three hydrogen bonding edges of a purine (adenine) and pyrimidine (cytosine) base B) The non-canonical G-U wobble and A-G sheared base pairs as examples of *cis* and *trans / reverse* base pairing. In *cis* pairing the glycosydic bonds (connecting the base with the sugar, represented by arrows) are on the same side of a line median to the hydrogen bonds connecting the bases, in *trans* pairing on opposite sides.

easily accessible for molecular interactions (40). The planes of the nucleobases are not perpendicular to the helix longitudinal axis but are tilted against it at an angle of ~70°. A-helical regions in RNA are separated by 'single-stranded' regions in the form of hairpin-loops, bulges and internal loops. Although shown as single-

stranded in secondary structure representations of RNA, these regions often show non-canonical base pairing and are sometimes involved in buildup of higher order structures.

Hairpin loops

A hairpin consists of a double-stranded helical stem and a single stranded apical (from *apex,* meaning 'top' or 'highest point') loop connecting the two strands of the helix (see Fig. 4). The size of hairpin loops makes them especially suited for solution studies using NMR. Hairpins have been shown to function as nucleation sites for RNA folding and in RNA-RNA as well as RNA-protein recognition. Hairpin loops can be as small as diloops or can be large loops stabilized by mismatched base pairs (48). Phylogenetic analyses have revealed that hairpin loops comprising four residues (tetraloops) are the most abundant (49). Three main families of tetraloops with the sequences GNRA, UNCG and CUUG (N is any nucleotide and R stands for purine) have been extensively characterized biochemically and structurally (50-52). They are exceptionally stable due to base-pairing between the first and fourth nucleotide, stacking of the unpaired nucleotides and extensive hydrogen-bonding networks (see Chapter 4 Fig. 7).

Bulges

Bulges are unpaired nucleotides on one strand within double stranded regions (see Fig. 4). Bulges can be extrahelical or can stack within the double helix, introducing a bend into the helix. Both extrahelical as well as stacked bulges have been shown to function as recognition sites for proteins or other RNAs. Larger bulges are often stabilized by base pairing, stacking or metal ion binding (53, 54).

Internal loops

Internal loops are interruptions within double stranded regions due to nucleotides that cannot form Watson-Crick or G-U base pairs (see Fig. 4).
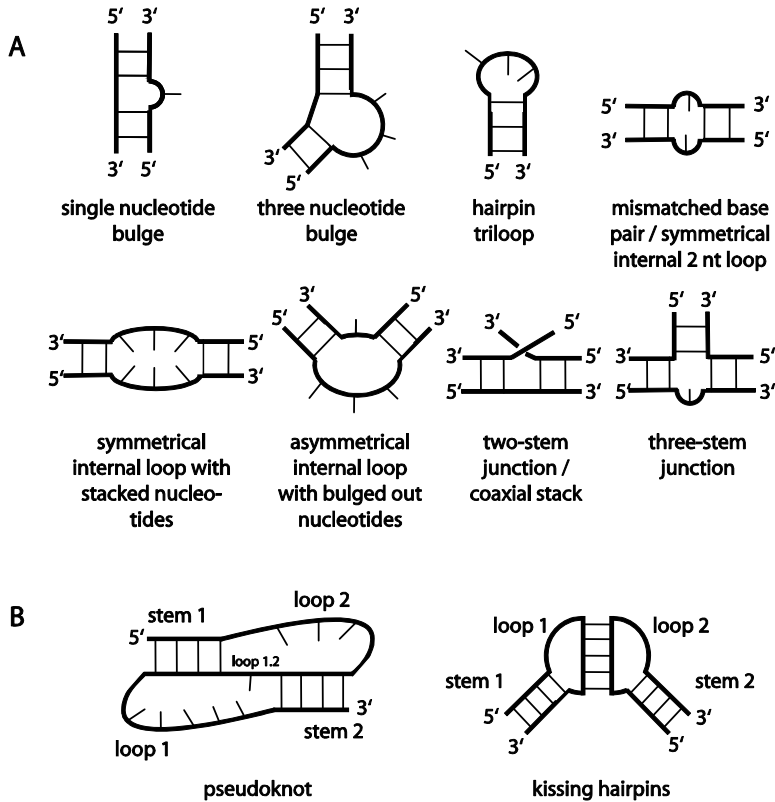
Figure 4: A) schematic representations of some structural building blocks of RNA B) schematic representations of some tertiary interactions of these building blocks. Adapted from Gesteland R.F., Cech T.R. and Atkins J.F. (eds.) (1999) *RNA world*, Cold spring Harbor Laboratory Press, Woodbury NY.

Secondary structures based on Watson-Crick base pairing often give the false impression that internal loops are big floppy 'bubbles' flanked by helical stems. However, structural studies have shown that the loop regions are often highly structured by forming non-canonical mismatched base pairs (55, 56). By these mismatches, distortions in the helical geometry are introduced, allowing the RNA to adopt novel shapes, which are key features in molecular recognition processes. Internal loops alter the flexibility of long helical double-stranded stretches.

Turns

Several structural families of turns are known, classified by the nucleotide preceding the turning phosphate. An example is the U-turn that is found in the GNRA family (52). These turns reverse the RNA strand direction and are thus often found in hairpins. Another family of turns, the S-turn, also referred to as the 'loop E motif' (57) locally reverses the backbone, maintaining an overall linear conformation of the RNA.

Junctions

Junctions are unpaired regions connecting multiple helices (58). They are classified as two-way, three-way, four-way, and higher-order junctions (see Fig. 4) (59).

Tertiary interactions between the structural elements

The secondary structure elements described interact through base pairing, hydrogen bonding and cross-strand base stacking to form complex shaped molecules. Metal ions are often involved in the stabilization of these tertiary structures.

An example of such tertiary interaction is a **coaxial helix** (60). Two separate helical regions stack to form coaxial helices as a pseudo-continuous helix. Coaxial helices are found in several large RNA structures, including tRNA, the Hepatitis Delta Virus (HDV) ribozyme and in many pseudoknots.

Pseudoknots involve base pairing between a hairpin or internal loop and bases outside the loop sequence (61) (see Fig. 4).

Other tertiary interactions are **loop-loop interactions** like the so-called **kissing hairpins**, where the single stranded loops are complementary and base pairing of these results in a quasi-continuous helix (62) (see Fig. 4).

The **tetraloop–tetraloop receptor** is characterized by specific hydrogen-bonding interactions between a tetraloop and a 11-nt internal loop/helical region that forms the receptor (63).

The **ribose zipper** is formed by consecutive hydrogen-bonding between the backbone ribose 2'-hydroxyls from two regions of the RNA chain interacting in an antiparallel manner (64).

In **loop-receptor complexes**, a hairpin loop is docked into a double stranded receptor. An example is the **A-minor motif**, which involves the insertion of minor-groove edges of adenosines into the minor groove of neighboring helices (65).

RNA structure determination by NMR

Structure determination of RNA starts with obtaining a highly concentrated (in the milimolar range) RNA sample in a biologically relevant conformation. RNA is usually produced by in vitro transcription from a DNA template by T7 Polymerase (66).

The actual NMR structure determination process by liquid state NMR consists of three main stages:

1. Chemical shift assignment of the NMR active nuclei (^{1}H, ^{13}C, ^{15}N, ^{31}P) (see Chapter 3).
2. Derivation of structural restraints: inter-proton distances are extracted from NOESY spectra, torsion angles from J-coupling data, long-range angular restraints by measuring residual dipolar couplings (RDCs) and hydrogen bonding restraints for Watson-Crick base pairs are derived from HNN-COSY spectra (see Chapters 3 and 4).
3. Structure calculation using simulated annealing methods in which structures consistent with the structural restraints are generated (see Chapter 4).

The general principles involved in these three steps are identical for the NMR structure determination of proteins and nucleic acids, but due to the chemical properties of nucleic acids, and especially RNA, NMR of RNA has some specific challenges:

- The RNA proton spectrum suffers from extreme spectral overlap, in particular for the chemically equivalent ribose protons. ^{13}C-editing in two- and three-dimensions is usually of vital importance to obtain complete resonance assignments of RNA sugar spin systems.

- Sequential through-bond backbone assignment, as routinely performed in protein NMR, is difficult due to the unfavorable properties of the phosphorous nucleus: ^{31}P chemical shifts are poorly dispersed and phosphorous relaxes rapidly due to its large chemical shift anisotropy. Sequential chemical shift assignments are thus obtained by a homonuclear approach which relies on the observation of regular NOE patterns. As this requires conformational assumptions, special attention is needed to prevent misassignments in non-helical regions, such as hairpin loops, internal loops and bulges.

- An additional problem is that in RNA, the proton density, and therefore the amount of theoretically measurable inter-atomic distances is low. As RNA structure is defined by a large number of torsion angles, of which the ones related to the sugar pucker and five backbone angles can be derived experimentally by measuring hetero- and homonuclear scalar coupling constants (see Chapters 3 and 4), the use of torsion angles as additional structural restraints is of particular importance in the calculation of RNA structures. Furthermore, scalar couplings may help to understand the dynamics of the system as average values of coupling constants indicate conformational exchange processes.

- As RNA A-helical structures are often extended and have little long-distance structural restraints, the use of RDCs is particularly important in defining the global structure of RNA molecules.

The homo- and heteronuclear NMR methods available for chemical shift assignment and the measurement of scalar couplings in nucleic acids have been extensively described in a number of reviews (67-71).

The final structures obtained from the structure calculation process should be validated and checked for quality. This is done by presenting an overview of structural statistics that shows how well the structures fit the experimental restraints. Additionally, back-calculation of parameters that were not used as restraints in the calculations, e.g. chemical shifts (72), should be used as a validation tool.

Finally, the goal of structural biology is to ultimately understand the structure-function relationship of the molecule under study. The three-dimensional structures obtained should be used together with biochemical experiments to understand how their specific molecular architecture is used to perform its function in the cell.

Scope of the thesis

As no molecular information is available on X-chromosome inactivation, we started NMR studies on a single A-repeat with the goal to solve its atomic structure and to obtain molecular insight into X-inactivation. During the initial structural studies we encountered difficulties completing chemical shift assignments of the complete A-repeat. Methodological improvements had to be developed to solve this problem.

Chapter 2 describes a novel NMR based strategy to characterize and distinguish intramolecular and intermolecular base pairing patterns in RNA that was essential to characterize the folding topology of the A-repeat. This topology turned out to be different than expected. We show that only the first predicted hairpin is formed, while the second predicted hairpin is unfolded and mediates dimerization of the A-repeat by duplex formation with a second A-repeat. The strategy described is suitable to unambiguously rule out biologically irrelevant duplex formation in RNA samples at conditions necessary for structural biology. It should also be more generally applicable to identify and quantify populations of hairpin and duplex conformations and to define RNA folding topology from inter- and intra-molecular base-pairing patterns.

Having discovered the peculiar A-repeat architecture, we next concentrated on the solution structure of the first hairpin within the A-repeat. **Chapter 3** reports in detail how chemical shifts for this 14-mer RNA hairpin were assigned and how J-coupling data to be used to restrain dihedral angles in the structure calculations was obtained.

In **Chapter 4** the NMR structure of this hairpin, capped by a fully conserved AUCG tetraloop sequence is presented. The structure reveals a novel well defined AUCG tetraloop fold which is stabilized by base stacking and possible hydrogen bonding. In addition to this structure of 'hairpin 1' within the A-repeat we build on our observation that the 'hairpin 2' sequence forms a duplex with a second

A-repeat *in vitro*. We show that mutants in this sequence that disrupt dimerization of the A-repeats *in vitro* do not initiate silencing *in vivo*.

Considering these *in vivo* data and the high local concentration of A-repeats, as they are connected by U-rich linkers and *Xist* is localized to the X-chromosome, we propose that multimerization of the A-repeats could also be relevant *in vivo*. We propose a model for A-repeat function where this dimerization and specific recognition of the AUCG tetraloop function together in *Xist* regulation and accumulation.

References

[1]   Jacob, F. and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, **3**, 318-356.

[2]   Britten, R.J. and Davidson, E.H. (1969) Gene regulation for higher cells: a theory. *Science*, **165**, 349-357.

[3]   Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. and Mello, C.C. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806.

[4]   Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559-1563.

[5]   Mattick, J.S. and Gagen, M.J. (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol Biol Evol*, **18**, 1611 - 1630.

[6]   Szymanski, M. and Barciszewski, J. (2002) Beyond the proteome: non-coding regulatory RNAs. *Genome Biology*, **3**, reviews5.1 - reviews5.8.

[7]   Matzke, M.A. and Birchler, J.A. (2005) RNAi-mediated pathways in the nucleus. *Nat Rev Genet*, **6**, 24.

[8]   Bernstein, E., Caudy, A.A., Hammond, S.M. and Hannon, G.J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**, 363.

[9]   Hammond, S.M., Bernstein, E., Beach, D. and Hannon, G.J. (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, **404**, 293.

[10]   Zamore, P.D., Tuschl, T., Sharp, P.A. and Bartel, D.P. (2000) RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell*, **101**, 25.

[11]   Pham, J.W., Pellino, J.L., Lee, Y.S., Carthew, R.W. and Sontheimer, E.J. (2004) A Dicer-2-dependent 80S complex cleaves targeted mRNAs during RNAi in Drosophila. *Cell*, **117**, 83.

[12] Sleutels, F., Zwart, R. and Barlow, D.P. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, **415**, 810.

[13] Reik, W. and Walther, J. (2001) Genomic imprinting: parental influence on the genome. *Nat Rev Genet*, **2**, 21 - 32.

[14] Akhtar, A. (2003) Dosage compensation: an intertwined world of RNA and chromatin remodelling. *Curr Opin Genet Dev*, **13**, 161-169.

[15] Kelley, R.L. and Kuroda, M.I. (2000) Noncoding RNA genes in dosage compensation and imprinting. *Cell*, **103**, 9.

[16] Jenuwein, T. and Allis, C.D. (2001) Translating the histone code. *Science*, **293**, 1074-1080.

[17] Turner, B.M. (2002) Cellular memory and the histone code. *Cell*, **111**, 285.

[18] Huynh, K.D. and Lee, J.T. (2001) Imprinted X inactivation in eutherians: a model of gametic execution and zygotic relaxation. *Current Opinion in Cell Biology*, **13**, 690.

[19] Heard, E. and Disteche, C.M. (2006) Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes & Development*, **20**, 1848-1867.

[20] Chaumeil, J., Le Baccon, P., Wutz, A. and Heard, E. (2006) A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes & Development*, **20**, 2223-2237.

[21] Schwartz, Y.B. and Pirrotta, V. (2007) Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet*, **8**, 9.

[22] Silva, J., Mak, W., Zvetkova, I., Appanah, R., Nesterova, T.B., Webster, Z., Peters, A.H.F.M., Jenuwein, T., Otte, A.P. and Brockdorff, N. (2003) Establishment of histone H3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 Polycomb group complexes. *Developmental Cell*, **4**, 481.

[23] de Napoles, M., Mermoud, J.E., Wakao, R., Tang, Y.A., Endoh, M., Appanah, R., Nesterova, T.B., Silva, J., Otte, A.P., Vidal, M. *et al.* (2004) Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. *Developmental Cell*, **7**, 663.

[24] Heard, E. (2005) Delving into the diversity of facultative heterochromatin: the epigenetics of the inactive X chromosome. *Current Opinion in Genetics & Development*, **15**, 482.

[25] Csankovszki, G., Panning, B., Bates, B., Pehrson, J.R. and Jaenisch, R. (1999) Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X-chromosome inactivation. *Nat Genet*, **22**, 323 - 324.

[26] Avner, P. and Heard, E. (2001) X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet*, **2**, 59 - 67.

[27] Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S. and Brockdorff, N. (1996) Requirement for Xist in X chromosome inactivation. *Nature*, **379**, 131-137.

[28] Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R. and Willard, H.F. (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, **349**, 38.

[29] Clemson, C.M., McNeil, J.A., Willard, H. and Lawrence, J.B. (1996) XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J Cell Biol*, **132**, 259 - 275.

[30] Lee, J.T. and Lu, N. (1999) Targeted mutagenesis of Tsix leads to nonrandom X Inactivation. *Cell*, **99**, 47.

[31] Sado, T., Hoki, Y. and Sasaki, H. (2005) Tsix silences Xist through modification of chromatin structure. *Developmental Cell*, **9**, 159.

[32] Navarro, P., Page, D.R., Avner, P. and Rougeulle, C. (2006) Tsix-mediated epigenetic switch of a CTCF-flanked region of the Xist promoter determines the Xist transcription program. *Genes & Development*, **20**, 2787-2792.

[33] Sun, B.K., Deaton, A.M. and Lee, J.T. (2006) A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization. *Molecular Cell*, **21**, 617.

[34] Ogawa, Y., Sun, B.K. and Lee, J.T. (2008) Intersection of the RNA interference and X-inactivation pathways. *Science*, **320**, 1336-1341.

[35] Wutz, A., Rasmussen, T.P. and Jaenisch, R. (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet*, **30**, 167.

[36] Brockdorff, N. (2002) X-chromosome inactivation: closing in on proteins that bind Xist RNA. *Trends Genet*, **18**, 352-358.

[37] Zhao, J., Sun, B.K., Erwin, J.A., Song, J.-J. and Lee, J.T. (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, **322**, 750-756.

[38] Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S. and Zhang, Y. (2002) Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science*, **298**, 1039-1043.

[39] Kohlmaier, A., Savarese, F., Lachner, M., Martens, J., Jenuwein, T. and Wutz, A. (2004) A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS Biology*, **2**, e171.

[40] Saenger, W. (1984) *Principles of Nucleic Acid Structure*. Springer, New York.

[41] Holbrook, S.R., Cheong, C., Tinoco, I. and Kim, S.-H. (1991) Crystal structure of an RNA double helix incorporating a track of non-Watson-Crick base pairs. *Nature*, **353**, 579.

[42] Al-Hashimi, H. (2005) Dynamics-Based Amplification of RNA Function and Its Characterization by Using NMR Spectroscopy. *ChemBioChem*, **6**, 1506-1519.

[43] Leontis, N.B., Lescoute, A. and Westhof, E. (2006) The building blocks and motifs of RNA architecture. *Current Opinion in Structural Biology*, **16**, 279.

[44] Leontis, N.B. and Westhof, E. (2003) Analysis of RNA motifs. *Current Opinion in Structural Biology*, **13**, 300.

[45] Westhof, E. and Fritsch, V. (2000) RNA folding: beyond Watson-Crick pairs. *Structure*, **8**, R55.

[46] Hermann, T. and Westhof, E. (1999) Non-Watson-Crick base pairs in RNA-protein recognition. *Chemistry & Biology*, **6**, R335.

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

[48] Kolk, M., van der Graaf, M., Fransen, C., Wijmenga, S., Pleij, C., Heus, H. and Hilbers, C. (1998) Structure of the 3'-hairpin of the TYMV pseudoknot: preformation in RNA folding. *EMBO J*, **17**, 7498-7504.

[49] Woese, C.R., Winker, S. and Gutell, R.R. (1990) Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 8467-8471.

[50] Allain, F.H.T. and Varani, G. (1995) Structure of The P1 helix from group-I self-splicing introns. *J Mol Biol*, **250**, 333-353.

[51] Jucker, F.M., Heus, H.A., Yip, P.F., Moors, E.H.M. and Pardi, A. (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. *Journal of Molecular Biology*, **264**, 968-980.

[52] Jucker, F.M. and Pardi, A. (1995) Solution structure of the CUUG hairpin loop - a novel RNA tetraloop motif. *Biochemistry*, **34**, 14416-14427.

[53] Hoogstraten, C.G., Legault, P. and Pardi, A. (1998) NMR solution structure of the lead-dependent ribozyme: evidence for dynamics in RNA catalysis. *Journal of Molecular Biology*, **284**, 337.

[54] Aboul-ela, F., Karn, J. and Varani, G. (1995) The structure of the Human Immunodeficiency Virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein. *Journal of Molecular Biology*, **253**, 313.

[55] Wimberly, B., Varani, G. and Tinoco, I. (1993) The conformation of loop E of eukaryotic 5S ribosomal RNA. *Biochemistry*, **32**, 1078-1087.

[56] Leontis, N.B. and Westhof, E. (1998) A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. *Journal of Molecular Biology*, **283**, 571.

[57] Correll, C.C., Wool, I.G. and Munishkin, A. (1999) The two faces of the Escherichia coli 23 S rRNA sarcin/ricin domain: the structure at 1.11 Å resolution. *Journal of Molecular Biology*, **292**, 275.

[58] Lescoute, A.L. and Westhof, E. (2006) Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83-93.

[59] Lilley, D.M.J., Clegg, R.M., Diekmann, S., Seeman, N.C., von Kitzing, E. and Hagerman, P.J. (1995) A nomenclature of junctions and branchpoints in nucleic acids. *Nucl. Acids Res.*, **23**, 3363-3364.

[60] Kim, S., Suddath, F., Quigley, G., McPherson, A., Sussman, J., Wang, A., Seeman, N. and Rich, A. (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science*, **185**, 435-440.

[61] Pleij, C.W.A., Rietveld, K. and Bosch, L. (1985) A new principle of RNA folding based on pseudoknotting. *Nucl. Acids Res.*, **13**, 1717-1731.

[62] Chang, K.-Y. and Tinoco, I. (1997) The structure of an RNA "kissing" hairpin complex of the HIV TAR hairpin loop and its complement. *Journal of Molecular Biology*, **269**, 52.

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

[64] 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.

[65] Nissen, P., Ippolito, J., Ban, N., Moore, P. and Steitz, T. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4899–4903.

[66] Milligan, J.F. and Uhlenbeck, O.C. (1989) Synthesis of small RNAs using T7 RNA-polymerase. *Methods in Enzymology*, **180**, 51-62.

[67] Flinders, J. and Dieckmann, T. (2006) NMR spectroscopy of ribonucleic acids. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **48**, 137.

[68] Fürtig, B., Richter, C., Wöhnert, J. and Schwalbe, H. (2003) NMR Spectroscopy of RNA. *ChemBioChem*, **4**, 936-962.

[69] Latham, M.P., Brown, D.J., McCallum, S.A. and Pardi, A. (2005) NMR methods for studying the structure and dynamics of RNA. *ChemBioChem*, **6**, 1492-1505.

[70] Varani, G., Aboul-ela, F. and Allain, F.H.T. (1996) NMR investigation of RNA structure. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **29**, 51.

[71] Wijmenga, S.S. and van Buuren, B.N.M. (1998) The use of NMR methods for conformational studies of nucleic acids. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **32**, 287.

[72] Cromsigt, J., Hilbers, C.W. and Wijmenga, S.S. (2001) Prediction of proton chemical shifts in RNA - their use in structure refinement and validation. *J Biomol NMR*, **21**, 11-29.

2

A NMR strategy to
unambiguously distinguish
nucleic acid hairpin and duplex
conformations applied
to a Xist RNA A-repeat

A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat

Malgorzata M. Duszczyk, Katia Zanier, Michael Sattler

Abstract

All RNA sequences that fold into hairpins possess the intrinsic potential to form intermolecular duplexes because of their high self-complementarity. The thermodynamically more stable duplex conformation is favored under high salt conditions and at high RNA concentrations, posing a challenging problem for structural studies of small RNA hairpin conformations. We developed and applied a novel approach to unambiguously distinguish RNA hairpin and duplex conformations for the structural analysis of a Xist RNA A-repeat. Using a combination of a quantitative HNN-COSY experiment and an optimized double isotope-filtered NOESY experiment we could define the conformation of the 26-mer A-repeat RNA. In contrast to a previous secondary structure prediction of a double hairpin structure, the NMR data show that only the first predicted hairpin is formed, while the second predicted hairpin mediates dimerization of the A-repeat by duplex formation with a second A-repeat. The strategy employed here will be generally applicable to identify and quantify populations of hairpin and duplex conformations and to define RNA folding topology from inter- and intra-molecular base-pairing patterns.

A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat

Introduction

All RNA sequences that fold into hairpins possess the intrinsic potential to form intermolecular duplexes because of their high self-complementarity. The thermodynamically more stable duplex conformation is favored under high salt and high RNA concentrations. This is a challenging problem for structural studies: crystallization for X-ray crystallography often requires high salt conditions and NMR structural studies require sample concentrations in the milimolar range. Unambiguous identification of sample stoichiometry under experimental conditions is essential as early as possible in RNA structural studies to ensure that the RNA is present in a biologically relevant conformation.

Xist (X-inactivation specific transcript) RNA is a large non-coding RNA essential for the initiation of X-inactivation in mammalian females (1). Early in embryonic development it is expressed from the X-chromosome that will be silenced and coats it *in cis*, which coincides with transcriptional shutdown through an unknown mechanism (2). The conserved so-called 'A-repeats' at the 5' end of *Xist* are essential for its silencing function, while several other regions are redundantly responsible for chromosome association (3) (Fig. 1A). In humans, the A-repeats are constituted of 7.5 copies of a 26 nucleotide motif, connected by long U-rich linkers. An Mfold secondary structure prediction of a single A-repeat suggested a double hairpin structure where the two hairpins possibly stack on top of each other (3) (Fig. 1B). As no structural information on the A-repeats is available, we started NMR studies on a single A-repeat (Fig. 1C) with the goal to solve its atomic structure and to obtain molecular insight into X-inactivation. The construct used in our study shown in Fig. 1C is identical to the 5th human A-repeat, apart from switching the positions of G and C in the third G-C base-pair to facilitate chemical shift assignments. Previous studies have shown that altering the sequence of the stem in hairpin 1 does not influence *Xist* activity as long as base pairing is not disrupted (3).
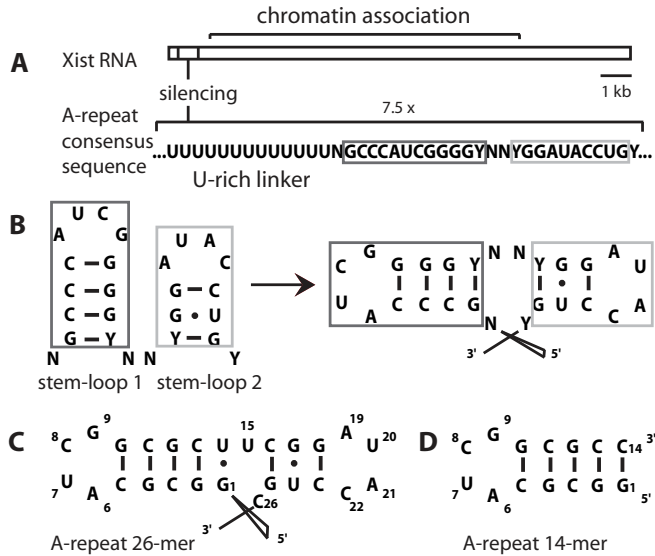
Figure 1 Schematic structure and sequence of Xist RNA and its A-repeats A) Xist RNA is a long (15kb in mouse, 17kb in human) non coding RNA. The A-repeats located at the 5' end are essential for silencing, while other regions are redundantly responsible for chromosome association B) The A-repeats consist of 7.5 copies of a conserved sequence predicted to fold into two hairpins, connected by long U-rich linkers. N = any nucleotide; Y = C/U C) The 26-mer A-repeat construct used in this study, containing both predicted hairpins. This construct is identical to the 5th human Xist RNA A-repeat apart from the reversed G4-C11 base pair as described in the introduction D) The 14-mer A-repeat construct used, containing the first predicted hairpin with a novel tetraloop.

During our structural studies we encountered difficulties completing NMR assignments of the second predicted hairpin. Signals from this hairpin were broad, and sometimes doubled (data not shown), which indicated possible dynamics or sample heterogeneity, although native gel analysis of the 26-mer A-repeat RNA suggested a homogenous monomeric population (Supplementary Material). The strategy described in this paper was essential to characterize and distinguish the intramolecular and intermolecular base pairs in monomeric / dimeric forms of RNA at sample conditions required for structural biology. The approach provided valuable insight into the possible architecture of the A-repeats.

A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat
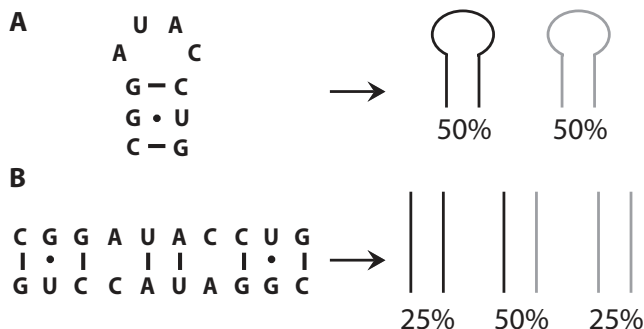

Figure 2 Schematic representation of the relative populations of species present in a 1:1 mixture of isotope labeled (magenta) and unlabeled (black) nucleic acids for A) a hairpin conformation and B) full duplex formation. The sequence of the second predicted hairpin of the Xist RNA A-repeat is shown.

NMR structures are determined from proton-proton distance restraints derived from nuclear Overhauser effects (NOEs), dihedral angle restrains derived from J-coupling constants and Residual Dipolar Couplings (RDCs) (4-8). Since chemical shifts and the pattern of NOEs in NMR spectra of hairpin and duplex species of a given nucleic acid are very similar, standard NMR techniques do not normally suffice to distinguish between the two. Other methods to determine the stoichiometry of nucleic acids include native gel electrophoresis and UV melting (9). However, these experiments are carried out at low concentrations. Hydrodynamic measurements such as ultracentrifugation, light scattering and NMR diffusion (10) are also available, but can be difficult to interpret for elongated molecules like RNA and for low molecular weight systems.

Several NMR methods have been developed to unambiguously distinguish between monomers and dimers of nucleic acids under NMR conditions (11-14). All these use an equimolar mixture of isotope labeled and unlabeled RNA/DNA. In such a mixture, different populations will exist for a RNA hairpin conformation with intramolecular base-pairing or for a duplex involving intermolecular base pairs. If the RNA folds into a hairpin, 50% of the molecules will be labeled and 50% will

be unlabeled (Fig. 2A). If the RNA adopts a duplex form, 25% of the molecules will be labeled, 25% will be unlabeled, and 50% of the duplexes will consist of one labeled and one unlabeled strand (Fig. 2B). Nucleic acid duplexes are detected based on differences in NMR parameters such as chemical shifts (11), cross-hydrogen bond $^{h2}J_{NN}$ coupling constants (12, 13), or NOEs (14).

The chemical shift based method (11) requires the introduction of a mutation that shifts a G imino signal to a characteristic frequency in the case of duplex formation. An obvious drawback of this method is that a change in the RNA sequence can possibly alter the monomer-dimer equilibrium.

The method using scalar couplings (12, 13) is based on the HNN-COSY experiment (15), which relies on the transfer of magnetization across the hydrogen bond in nucleic acid base pairs using the two-bond $^{h2}J_{NN}$ coupling. RNA duplexes are detected through a difference in intensities between cross- and diagonal peaks in HNN-COSY spectra, as in a duplex species with one labeled and one unlabeled strand magnetization transfer over the hydrogen bond cannot take place. We have extended this method and demonstrate that the molar fractions of hairpin and duplex species in a mixed population can be determined.

The NOE based method uses isotope editing/filtering techniques to distinguish intermolecular and intramolecular NOEs (14). NOEs arising from pairs of protons where one is bound to ^{15}N and the other one bound to ^{14}N can only arise from a duplex species. This is a complementary approach if an efficient J-based magnetization transfer across the base pair is not possible, for example in G-U pairs. We have employed simultaneous filters for protons attached to nitrogens and to carbon. Thereby, intermolecular NOEs involving imino-imino (in G-U base pairs) and imino-amino (in G-C base pairs) can be discriminated from NOEs involving imino protons and H2s in A-U base pairs.

Here we present the combination of HNN-COSY and NOE based methods to distinguish between monomeric hairpin and duplex conformations of nucleic acids. The approach is demonstrated and was crucial for the determination of the architecture of the Xist RNA A-repeats in our structural studies.

Materials and Methods

Sample Preparation

[13]C,[15]N uniformly labeled and unlabeled r(GGCGC [AUCG] GCGCUUCGG [AUAC] CUGC) A-repeat 26-mer RNA (consisting of both predicted A-repeat hairpins, Fig. 1C) and r(GGCGC [AUCG] GCGCC) A-repeat 14-mer RNA (containing only the first hairpin, Fig. 1D) was prepared by in vitro transcription with T7 RNA polymerase (EMBL Protein Expression and Purification Core Facility, Heidelberg, Germany) using synthetic DNA oligos (MWG Biotech / Eurofins MWG Operon, Ebersberg, Germany) as a template (16) and either ^{13}C,[15]N labeled (Silantes, Munich, Germany) or unlabeled (Sigma-Aldrich, Munich, Germany) NTPs. The positions of G and C in the fourth base pair of the first predicted hairpin were reversed to facilitate NMR assignment. 2'-O-methyl groups were incorporated into the two 5' residues of the DNA template strand to reduce the amount of n+1 transcription products (17). The RNA was purified on preparative denaturing 20% (w/v) polyacrylamide (19:1 acrylamide : bisacrylamide) gels. Gel bands were visualized by UV shadowing using fluorescent thin layer chromatography plates (Whatman) and the band corresponding to the full-length product was excised and electroeluted (Schleicher & Schuell / Whatman, Dassel, Germany). The RNA was precipitated, resuspended and extensively washed in 1K MWCO concentrators (Pall, Dreieich, Germany) with 10 mM NaH_2PO_4/Na_2HPO_4 buffer (pH 6.0) of progressively decreasing NaCl concentration (1M to 0M). Finally the RNA was desalted (PD-10, Amersham / GE Healthcare, Freiburg, Germany) and lyophilized.

NMR samples were prepared in 10 mM NaH_2PO_4/Na_2HPO_4 buffer (pH 6.0), 100 mM NaCl, 0.02mM EDTA, 0.02% Azide in 95% H_2O, 5% D_2O. The RNA concentration of the ^{13}C, ^{15}N uniformly labeled sample for recording the

HNN-COSY reference experiment was 0.8 mM. An equal amount of unlabeled RNA was added to this sample, mixed, lyophilized and resuspended in the same amount of H_2O/D_2O. Both the edited/filtered NOESY and HNN-COSY were run on this 1.6 mM sample. In addition, the HNN-COSY was repeated on the same sample diluted to a total RNA concentration of 0.8 mM. Just before the NMR measurements samples were heated to 95° for 5 minutes followed by snap-cooling on ice with the rationale to trap the kinetically favored intramolecular monomeric hairpin conformation over a possible intermolecular dimer. The stoichiometry of the NMR samples was initially checked with native PAGE (Supplementary Material).

NMR Spectroscopy

HNN-COSY

NMR experiments were recorded at 5°C on a Bruker DRX600 spectrometer equipped with a cryoprobe. HNN-COSY spectra were recorded with the pulse sequence described by Dingley & Grzesiek (15) shown in Supplementary Fig. 1. The delay T for evolution of the $^{h2}J_{NN}$ coupling was set to 20, 30 and 40 ms. 1024 complex points were collected in t_2 with a sweep width of 12.5 kHz, and 140 complex points were recorded in t_1 with a sweep width of 6.25 kHz. 256 scans were recorded for each complex t_1 increment. The experiments were performed with the ^1H carrier positioned at the H_2O resonance and the ^{15}N carrier at 175 ppm. The data were zero filled to $512 \times 2K$ complex data points, followed by apodization using Lorentz-to-Gauss transformation and cosine functions in t_2 and t_1, respectively, before Fourier transformation.

A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat

Isotope-edited/filtered NOESY

The ω_1, ω_2 double isotope-filtered NOESY experiment employed is shown in Fig. 3. Compared to the experiment proposed by Aboul-ela *et al.* (14) for discrimination between intra- and intermolecular NOEs we have added a ^{13}C filter in ω_1. This allows the discrimination of intermolecular imino-imino NOEs in G-U base pairs, and the identification of intermolecular imino-H2 NOEs in A-U base pairs. Moreover, the ω_1 filter is combined with semi-constant time chemical shift evolution (18, 19) in t_1 to exploit the filter delay 2Δ simultaneously for chemical shift labeling and J-coupling evolution (20) for improved sensitivity.

Editing/filtering for both ^{15}N and ^{13}C in ω_1 is accomplished by a 'jumping' 180° pulse (Fig. 3). The pulse is applied before or after a delay Δ and Δ', 1/2J_{HN} and 1/2J_{HC} in which anti-phase magnetization is created for H_N and H_C proton spins, respectively. If the pulse is applied at position (1) in Fig. 3, the signals of ^{15}N/^{13}C bound proton spins are effectively inverted and have an opposite phase compared to other spins. If it is applied at position (2) (Fig. 3), evolution of heteronuclear coupling is completely refocused during the semi-constant time evolution period, and ^{15}N/^{13}C- and ^{14}N/^{12}C-bound protons have the same phase. Transients are stored separately for the two positions of the jumping 180° pulses in order to be able to separate the inter- and intramolecular NOEs. The sum of scans (1) and (2) selects ^{14}N/^{12}C bound protons in ω_1, its difference ^{15}N/^{13}C bound protons.

In ω_2, editing is achieved by applying two consecutive 90° pulses on ^{15}N. The first pulse is applied with phase x and the second with phase ψ alternating between x and $-x$ (21) and the two transients are stored separately. When both pulses have the same phase the signal of ^{15}N bound protons is inverted. Again, the sum of scans with $\psi = x$ and $\psi = -x$ selects ^{14}N bound protons in ω_2, its difference ^{15}N bound protons.

A total of four different FIDs are recorded as described in Table 1, which results in a) inverting signals of ^{15}N/^{13}C bound protons in ω_1 and ^{15}N bound protons in ω_2, b) inverting only ^{15}N bound proton signals in ω_2, c) inverting only

^{15}N/^{13}C bound protons in ω_1 and d) no signal inversion. Linear combinations of these four FIDs yield four subspectra I-IV as shown in Table 2.

The NOESY mixing time was set to 300 ms. The experiment was performed with the ^{1}H carrier positioned at the H$_2$O resonance and the ^{15}N and ^{13}C carriers both at 150 ppm. Pulse lengths for the ^{15}N and ^{13}C 180° pulses were 84 and 30 μs respectively. For the ^{15}N and ^{13}C spins of interest in the G-C and A-U base pairs these pulses achieve better than 98% inversion. Therefore, incomplete inversion which could degrade the editing/filtering performance should not be an issue. 1024 complex points were collected in t_2 with a sweep width of 13.9 kHz, and 256 complex points were recorded in t_1 with a sweep width of 13.9 kHz. 32 scans of 4 separate FIDs were collected for each complex point in t_1. The data was zero filled to 512 × 2K complex data points and apodized using Lorentz-to-Gauss and cosine functions in t_2 and t_1 respectively before Fourier transformation.

Spectra were processed with NMRPipe (22) and analyzed using NMRVIEW (23). Imino protons were assigned with a combination of HNN-COSY (15) and a sequential walk in 2D NOESY spectra (300 ms mixing time) (24) (Duszczyk *et al.*, in preparation – see Chapter 4). All pulse sequences use the WATERGATE sequence (25) and water flipback (26) for water suppression.

Quantitative analysis of HNN-COSY

A description of the magnetization transfer in the HNN-COSY pulse sequence is given in the Supplementary Material. The HNN-COSY spectrum shows cross and diagonal peaks at the chemical shifts of the ^{15}N1 and ^{15}N3 nuclei for each G-C, A-U, and some non-canonical base pairs. If both hydrogen bond acceptor and donor are isotope labeled, intensities of cross and diagonal peaks are proportional to $\sin^2(\pi^{h2}J_{NN}T)$ and $\cos^2(\pi^{h2}J_{NN}T)$ respectively.
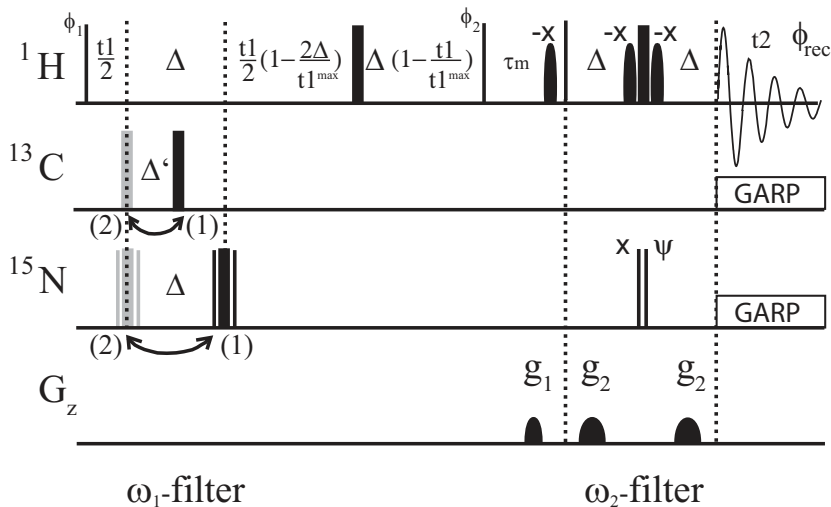
Figure 3 Double (ω_1, ω_2)-filtered NOESY pulse sequence to distinguish intra- and intermolecular NOEs in nucleic acid base pairs. Δ = 5.4 ms, Δ' = 2.5 ms. Narrow and wide bars denote 90° and 180° pulses, respectively, and are applied with phase x unless stated otherwise. ^{13}C, ^{15}N 180° pulses are applied at position (1) or (2) to distinguish ^{13}C/^{15}N and ^{12}C/^{14}N bound protons. The ^{15}N 180° pulses are composite pulses of the form 90°$_y$180°$_x$90°$_y$. ψ = x, -x for distinguishing ^{15}N/^{14}N bound protons. Phase cycle: φ_1 = x, -x + TPPI; φ_2 = x, x, -x, -x; φ_{rec} = x, -x,-x, x. Water suppression is accomplished by WATERGATE (25) combined with water flip back (26). Gradients of 800 µs length were applied with 30% (g_1) and 50% of maximal power (g_2).

Values of $^{h2}J_{NN}$ coupling constants were calculated from $I_{cross}/I_{diagonal}$ extracted from the spectra recorded on the fully labeled RNA samples with:

$$I_{cross}/I_{diagonal} = \tan^2(\pi^{h2}J_{NN}T)$$

$I_{cross}/I_{diagonal}$ was determined as the amplitude ratio of the time domain oscillations using the time domain fitting routine *nlinLS* contained in the NMRPipe package (15, 22). Errors in the intensity ratio and couplings were calculated based on the effect of random noise for the peak height estimated by *nlinLS*. No correction for an underestimation of 10-20% due to the finite excitation bandwidth of the ^{15}N radio frequency pulses (15) was made.

FID	a	b	c	d
Jumping pulse position	(1)	(2)	(1)	(2)
Phase ψ	x	x	-x	-x
Amplitude factor				
(ω_1) ^{13}C, ^{15}N	$-$	$+$	$-$	$+$
^{12}C, ^{14}N	$+$	$+$	$+$	$+$
(ω_2) ^{15}N	$-$	$-$	$+$	$+$
$^{12/13}C$, ^{14}N	$+$	$+$	$+$	$+$
Overall amplitude factor				
^{12}C, ^{14}N (ω_1) - $^{12/13}C$, ^{14}N (ω_2)	$+$	$+$	$+$	$+$
^{13}C, ^{15}N (ω_1) - ^{15}N (ω_2)	$+$	$-$	$-$	$+$
^{13}C, ^{15}N (ω_1) - $^{12/13}C$, ^{14}N (ω_2)	$-$	$+$	$-$	$+$
^{12}C, ^{14}N (ω_1) - ^{15}N (ω_2)	$-$	$-$	$+$	$+$

Table 1 Transfer amplitude factors for the signals in the four different FIDs of the double filtered NOESY experiment shown in Fig. 3.

Discrimination between intra- and intermolecular hydrogen bonds can be accomplished by comparing the relative intensities of cross and diagonal peaks in HNN-COSY spectra of a fully labeled and 50% labeled, 50% unlabeled RNA sample. If the RNA forms monomeric hairpins, the relative intensities of the diagonal and cross peaks ($I_{diagonal}/I_{cross}$) will be equal in both samples, apart from an overall 50% loss of intensity as only the 50% labeled molecules contribute to the signal. In the case of duplex formation $I_{diagonal}/I_{cross}$ will increase as 50% of the RNA population of which one strand is labeled and the other unlabeled will only contribute to the diagonal peak intensity but not to the cross peak intensity.

A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat
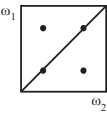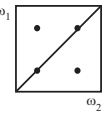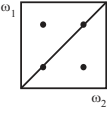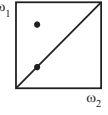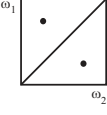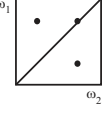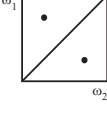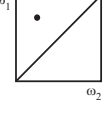
Sub-spectrum	Linear combination of FIDs	NOEs involving protons bound to		NOEs observed	
		ω_1	ω_2	G-U	A-U
I	a + b + c + d	^{12}C, ^{14}N	$^{12/13}$C ^{14}N		
II	a − b − c + d	^{13}C ^{15}N	^{15}N		
III	a − b + c − d	^{13}C ^{15}N	$^{12/13}$C ^{14}N		
IV	a + b − c − d	^{12}C ^{14}N	^{15}N		

Table 2 NOEs observed in the four subspectra that result from linear combinations of the four different FIDs recorded

The ^{1}H and ^{15}N chemical shifts of corresponding positions in hairpin and duplex conformations are usually degenerate. Therefore, if the interconversion between monomer and dimer species is either fast or slow on the time scale of the $^{h2}J_{NN}$ couplings and if there is no significant population of a non-hydrogen bonded form, molar fractions in a mixed population of monomers and dimers can be derived from the intensity ratio of cross and diagonal peaks in a 50% labeled, 50% unlabeled sample. The requirements are: i) corresponding $^{h2}J_{NN}$ coupling constants

are similar in the hairpin and duplex conformations, ii) the size of the $^{h2}J_{NN}$ couplings is known (for example from measurements on a 100% labeled sample) and iii) transverse ^{15}N T2 relaxation times of the monomer and dimer species are known.

Then, the molar fractions of the hairpin ($\chi_{monomer}$) and that of the duplex conformations ($\chi_{duplex} = 1-\chi_{monomer}$) in a 50% labeled, 50% unlabeled sample contribute to the intensities of diagonal and cross peaks as follows:

$$\frac{I_{diagonal}}{I_{cross}} = \frac{\left(\begin{array}{l} \chi_{monomer}\cos^2(\pi^{h2}J_{NN}T)e^{-2T/T2\,(monomer)} \quad + \\ \quad (1-\chi_{monomer})(1+0.5\cos^2(\pi^{h2}J_{NN}T))e^{-2T/T2\,(duplex)} \end{array} \right)}{\left(\begin{array}{l} \chi_{monomer}\sin^2(\pi^{h2}J_{NN}T)e^{-2T/T2\,(monomer)} \quad + \\ \quad (1-\chi_{monomer})0.5\sin^2(\pi^{h2}J_{NN}T)e^{-2T/T2\,(duplex)} \end{array} \right)}$$

Note, that the transverse ^{15}N relaxation has only a very small effect on $I_{diagonal}/I_{cross}$. Moreover, in the case of a fully dimeric or fully monomeric conformation $I_{diagonal}/I_{cross}$ does not depend on the transverse relaxation time at all.

Results and Discussion

Xist RNA A-repeat 26-mer assignment

Assignment of RNA usually starts with confirmation of the secondary structure by identifying Watson-Crick base pairs in A-form helical regions. In long mixing time NOESY spectra in H_2O these cause a characteristic pattern of inter-base NOEs within base pairs and between consecutive stacked base pairs. With the latter a so-called 'sequential walk' through the stem is possible. Sequence specific assignment of the imino protons is facilitated with the HNN-COSY experiment where the imino proton shifts are correlated with base-specific chemical shifts of the nitrogens in the base pairs.

After recording these two experiments on our Xist 26-mer sample, imino assignment could be almost completed. However, we noticed an unusual downfield shifted signal, which did not show any imino-imino NOEs in the NOESY. In the HNN-COSY experiment this imino was correlated to a uridine N3 nitrogen and to a nitrogen at ~222 ppm, indicative of either adenine N1, N3 or purine N7 nitrogens (27). A correlation to adenine N1 would correspond to a canonical A-U base pair, but such a base pair was not expected in the predicted 26-mer double hairpin. We considered two possibilities: either that this imino was involved in a non-canonical base contact in the second predicted loop, or that it was involved in an intermolecular base pair in a duplex species, although native gel analysis of the 26-mer A-repeat RNA suggested a homogenous monomeric population (Supplementary Fig. 2).

Quantification of HNN-COSY

To rule out that our Xist samples are dimerizing under NMR conditions, we used the approach based on the quantitative HNN-COSY experiment (Fig. 4). The HNN-COSY spectrum recorded on a 100% ^{13}C, ^{15}N-labeled sample of the 26-mer is shown in Fig. 4C. Figure 4D shows slices through the maxima of cross- and diagonal peaks for two base pairs (G2-C13 in the first predicted stem and the base pair involving the U20 imino) from HNN-COSY spectra recorded on a fully labeled (blue) and 50% labeled, 50% unlabeled (green) 26-mer RNA.
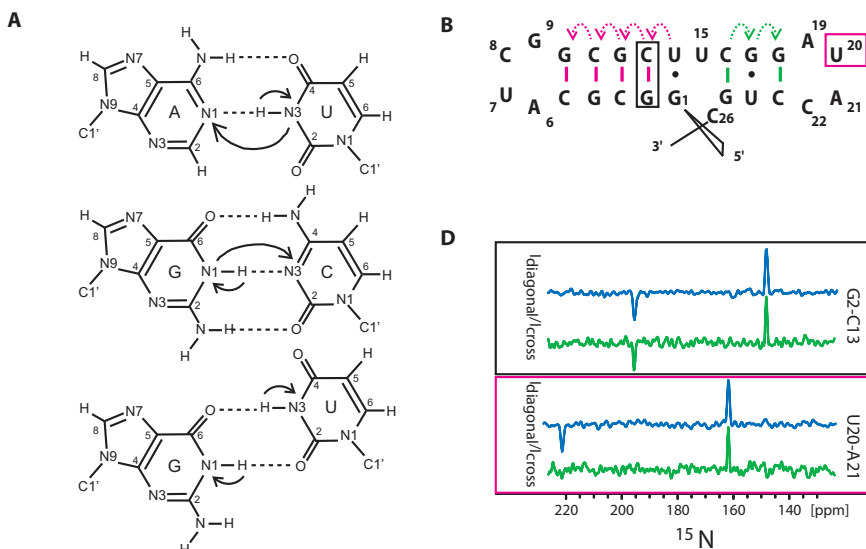
Figure 4 Secondary structure assignment of the Xist RNA A-repeat A) G-C, A-U and G-U base pairs with magnetization transfer in the HNN-COSY. B) The predicted structure of the 26-mer A-repeat construct with sequential imino-imino NOE and HNN transfers as dotted and solid lines respectively, in magenta for the first predicted hairpin, in green for the second one D) Diagonal and cross peak intensities in a fully labeled (blue) and 50% labeled, 50% unlabeled (green) Xist A-repeat 26-mer sample are compared for the intramolecular base pair G2-C13 (boxed in black) and the intermolecular base pair U20-A21 (boxed in magenta). The spectra have been normalized with respect to the diagonal peak height, as experimental times of for the spectra recorded on the fully labeled and 50% labeled samples were different.
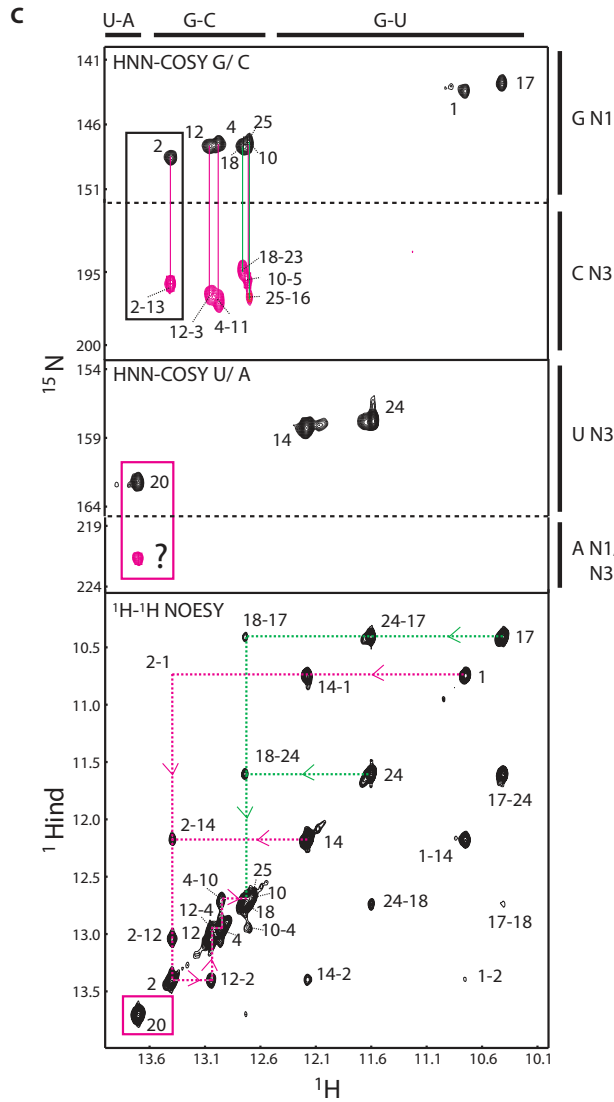
Figure 4 C) Top/middle: 600 MHz ^{1}H, ^{15}N HNN-COSY spectrum recorded on a 100% ^{13}C^{15}N labeled sample in 95:5 H$_2$O:D$_2$O at 278 K. Top: Intra-base imino-N1 correlations leading to diagonal peaks for the guanosines in the stem are shown in black and are connected by solid lines to the inter-base imino-N3 cross peak correlations of the cytosines they are base paired to. Middle: the correlation for the A-U base pair is shown. Bottom: A 2D ^{1}H NOESY spectrum, recorded on an unlabeled sample in H$_2$O under the same conditions, connects the iminos by a sequential walk through the stem as indicated by a dotted line. Connectivities for the first predicted stem are shown in magenta, for the second predicted stem in green as shown in B). Signals arising from the U20 imino proton are boxed in magenta.

Corresponding spectra were recorded on a shorter 14-mer Xist RNA sample containing only the first predicted hairpin (see Fig. 1D) to compare measurements in the first hairpin within the 26-mer sample to the ones in a small and stable hairpin that provided superior spectral quality (data not shown). Only non-overlapped signals were used to measure intensities.

All base pairs in the 14-mer could be analyzed except G1-C14, which is not visible presumably due to end-fraying. Measured intensities for diagonal and cross peaks for the base pairs in the 14-mer and the 26-mer are given in Table 3. It is clear that for the 14-mer Xist RNA $I_{diagonal}/I_{cross}$ ratios are comparable within error for the 100% labeled sample and the 50% labeled, 50% unlabeled sample. This confirms a monomeric state of the 14-mer hairpin.

The non-overlapped signals in the 26-mer show different results for base pairs from the first and second predicted hairpin. For G2-C13, G12-C3 and G4-C11 no significant difference between $I_{diagonal}/I_{cross}$ is seen between the two samples. In contrast, for the base pair involving the U20 imino proton, $I_{diagonal}/I_{cross}$ increases four-fold in the 50% labeled, 50% unlabeled sample. This unambiguously shows that U20 is involved in an intermolecular base pair. The most probable partner in this intermolecular base pair is A21 in a second 26-mer molecule. Thus, A20-U21 is indeed a canonical Watson-Crick base pair, however, from an unexpected duplex RNA species. Although $I_{diagonal}/I_{cross}$ for the G18-C23, G10-C5 and G25-C16 base pairs could not be analyzed quantitatively due to spectral overlap, the signals involving G18-C23 and G25-C16 in the second predicted hairpin show a significant decrease in cross peak intensity in the 50% labeled sample (data not shown), consistent with intermolecular base pairs in the second predicted hairpin.

Calculation of $^{h2}J_{NN}$ coupling constants from a 100% isotope labeled RNA yielded values between 4.9 and 5.6 Hz (see Table 2), which are at the lower end of the range of reported values of 6-7 Hz for Watson-Crick base pairs (15, 28, 29). Repetition of the experiment on a different spectrometer and with different transfer times for the N-N transfer resulted in the same range of couplings.

A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat

Panel A: A-repeat 14-mer Base pair	$I_{diagonal}/I_{cross}$ 100% labeled 1.2 mM	$^{h2}J_{NN}$ coupling	$I_{diagonal}/I_{cross}$ 50% labeled 50% unlabeled 1.2 mM each
C5-G10	1.6 ± 0.1	5.3 ± 0.1	1.7 ± 0.1
G4-C11	1.4 ± 0.1	5.6 ± 0.1	1.3 ± 0.1
C3-G12	1.4 ± 0.1	5.6 ± 0.1	1.4 ± 0.1
G2-C13	1.9 ± 0.1	5.0 ± 0.1	1.8 ± 0.1

Panel B: A-repeat 26-mer Base pair	$I_{diagonal}/I_{cross}$ 100% labeled 0.8 mM	$^{h2}J_{NN}$ coupling	$I_{diagonal}/I_{cross}$ 50% labeled 50% unlabeled 0.8 mM each / 0.4 mM each
C5-G10	ovl.	n.d.	ovl.
G4-C11	1.6 ± 0.1	5.4 ± 0.1	1.3 ± 0.1 / 1.3 ± 0.1
C3-G12	1.4 ± 0.1	5.6 ± 0.1	1.8 ± 0.1 / 1.7 ± 0.1
G2-C13	1.9 ± 0.1	5.0 ± 0.1	1.7 ± 0.1 / 1.4 ± 0.1
C16-G25	ovl.	n.d.	ovl.
G18-C23	ovl.	n.d.	ovl.
U20-A21	2.0 ± 0.1	4.9 ± 0.1	8.5 ± 1 / noise

Table 3 Experimental ratios between diagonal and cross peak intensities extracted from HNN-COSY spectra of fully labeled and 50% labeled, 50% unlabeled A) Xist RNA A-repeat 14-mer and B) Xist RNA A-repeat 26-mer with calculated J-couplings for each base pair. For the 26-mer, calculated $I_{diagonal}/I_{cross}$ values are given for spectra recorded on samples with a total RNA concentration of 1.6 mM and 0.8 mM. Intensities that could not be analyzed because of overlap are labeled with 'ovl.' for 'overlapped' and their calculated $^2J_{NN}$ couplings are labeled with 'n.d.' for 'not determined'. $I_{diagonal}/I_{cross}$ for U20-A21 could not be determined for the sample with 0.4 mM labeled and unlabeled 26-mer each as the cross peak intensity was in the noise level. Errors in the intensity ratios and couplings were calculated based on the effect of random noise for the peak height estimated by the time-domain fitting routine *nlinLS* contained in the NMRPipe package.

Smaller couplings are usually observed in non-linear H-bond geometries or due to fraying at the interfaces with non-regular secondary structure elements (30). Presumably, apart from a potential underestimation of the couplings due to imperfections of the ^{15}N radio frequency pulses, the size of the couplings in our system is reduced by fraying and/or other conformational dynamics in the small hairpins investigated.

Figure 5 shows the correlation of $I_{diagonal}/I_{cross}$ with the transfer time for full monomeric and full duplex conformations, as well as for the case where only 90% of the sample is either monomeric or dimeric. Transverse ^{15}N T2 relaxation times were estimated to be 50 ms for the dimer and 70 ms for the monomer based on reported values for RNAs of similar size (31) and considering that the experiments were recorded at 5 °C. For a transfer time T of 40 ms and an $^{h2}J_{NN}$ coupling of 5 Hz $I_{diagonal}/I_{cross}$ is calculated to be 7.6 for a pure duplex and 1.9 for a pure hairpin conformation. Thus, a four-fold increase in the ratio is expected for a full duplex species upon mixing with unlabeled RNA. If a mixed population exists, the $I_{diagonal}/I_{cross}$ ratios shift closer together. Variations in the ^{15}N T2 relaxation times also lead to an increase/decrease of $I_{diagonal}/I_{cross}$ for monomer/dimer species, but the effects are negligible (not shown). By comparing the experimentally measured $I_{diagonal}/I_{cross}$ values with the simulated curves it is clear that all base pairs in the first hairpin of the 26-mer A-repeat RNA are consistent with a fully monomeric conformation, while the A-U base pair in the second predicted hairpin exists in a fully dimeric form.

With an A-U base pair originating from a dimeric conformation two possibilities exist for the A-repeat topology. Dimerization could involve a kissing hairpin with two equivalent intermolecular A-U base pairs (Fig. 6B), or correspond to a duplex conformation with full base pairing with the second part of another A-repeat (Fig. 6C).

To distinguish between these two conformations information on the state of the G-U base pair in the second predicted stem would be useful. In a kissing hairpin complex this base pair would be intramolecular while in a duplex this base

pair would be intermolecular. Unfortunately, the analysis of J-couplings across hydrogen bonds is not applicable to G-U base pairs due to a lack of sizable J-couplings (Fig. 4A).
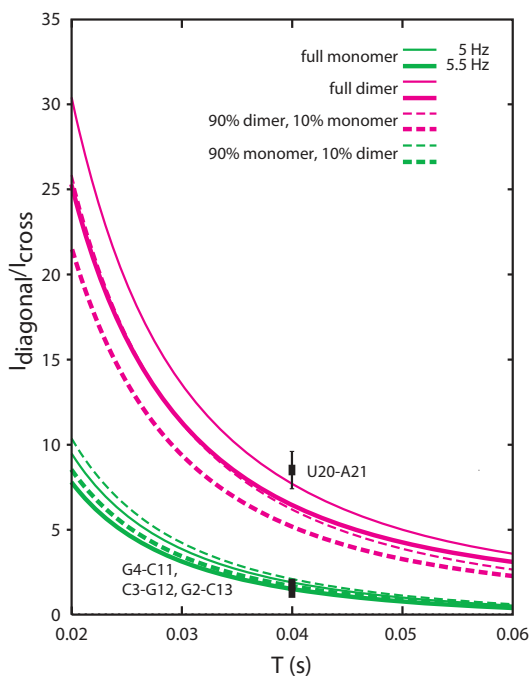
Figure 5 Expected ratio of diagonal and cross peak intensities ($I_{diagonal}/I_{cross}$) as a function of the transfer time T in the HNN-COSY experiment for duplex (magenta) and monomer (green) conformations in a 50% labeled, 50% unlabeled sample, calculated for $^{h2}J_{NN}$ = 5/5.5 Hz (thin/thick lines) and ^{15}N T2 relaxation times of 70 and 50 ms for the monomer and dimer, respectively. Curves for fully monomeric and dimeric species are shown as lines, mixed populations (90% monomer/10% dimer and *vice versa*) are shown as dotted lines. Experimentally measured intensity ratios on the 26-mer RNA are indicated as squares with error bars based on random noise in the spectra. The trans-hydrogen-bond coupling for the U20-A21 base pair is $^{h2}J_{NN}$=4.9 Hz, while the couplings for the G-C base pairs are 5.0-5.5 Hz.

Analysis of the double-isotope-filtered NOESY spectra

As a characteristic and strong NOE pattern is visible between the G and U imino protons in a G-U base pair we analyzed the 50% labeled, 50% unlabeled sample following the approach first proposed by Aboul-ela *et al.* (14) based on NOEs in base pairs, with optimizations and extensions as described in the Materials and Methods section.

Linear combinations of the four FIDs a-d (Table 1) recorded in the double-filtered NOESY experiment yield four NOESY subspectra I-IV (Table 2). The imino-amino regions of these four spectra are shown in Figures 6D and 6E. In Fig. 6D the spectra with NOEs between the unlabeled protons (spectrum I in black) and NOEs between the labeled protons (spectrum II in red) are superimposed. In Fig. 6E the two spectra with NOEs between labeled and unlabeled protons are shown. These NOEs can only originate from a duplex RNA species.

NMR signals of special interest are highlighted: the U20 imino to A21 H2 cross peaks, and the two imino-imino NOEs in the G17-U24 base pair. The latter demonstrate that not only U20-A21, but also the G17-U24 base pair in the second predicted stem of the 26-mer *Xist* construct is intermolecular. For the G1-U14 base pair in the first stem no imino-imino cross peaks are seen in Fig. 6E, which confirms that the first predicted hairpin is intramolecular.

The intensities of the imino to imino NOEs in the G-U base pair are symmetrical in spectra III and IV (considering that spin-diffusion in the long mixing time NOESY can differentially affect the peak intensities of the symmetric cross peaks). In contrast, the imino-H2 NOEs in the A-U base pair are only symmetric in spectrum III since the ^{13}C filter is applied only in ω_1. This asymmetric NOE pattern is only consistent for an intermolecular NOE involving an imino proton and a proton attached to carbon, which independently confirms that these NOEs correspond to an A-U base pair.

The architecture of the Xist RNA A-repeat

The combined data from the HNN-COSY and double-isotope-filtered NOESY
experiments are in agreement with an architecture of the A-repeats as shown in
Figure 6C with a fully dimerized second stem. The fact that the imino-imino NOEs
for the G17-U24 base pair in the spectra shown in Fig. 6D and 6E have comparable
intensities indicates that the 26-mer RNA exists in a fully dimeric form in solution.
If an equilibrium would exist between the kissing hairpin and duplex forms shown
in Fig. 6B and 6C, the G-U imino-imino NOEs in Fig. 6E should have lower
intensities than those in Fig. 6D.

      The question remains if the dimerization seen *in vitro* under NMR
conditions is significant *in vivo*, or if it is merely an artifact of the high
concentrations used in the experimental conditions. In this respect, we note that the
imino signal of U20, which is involved in the intermolecular base pair, is visible in
1D NMR spectra even at concentrations as low as 50 μM. In addition, analytical
ultracentrifugation data of a 0.1 mM sample show that the A-repeat 26-mer exists as
a dimer in solution (Duszczyk *et al.*, in preparation – see Chapter 4). It should also
be considered that *in vivo* there is a high local concentration of the A-repeats as they
are connected by relatively short linkers and are localized to the X-chromosome.
Thus, it is possible that the second region of the A-repeat may function as a
multimerization platform for several A-repeats, either within a single Xist RNA
molecule, or between different ones. Further experiments to address these issues
are underway.

Conclusion

We developed and applied a novel approach to unambiguously define the
stoichiometry of the NMR sample for the structural analysis of a Xist RNA A-
repeat.
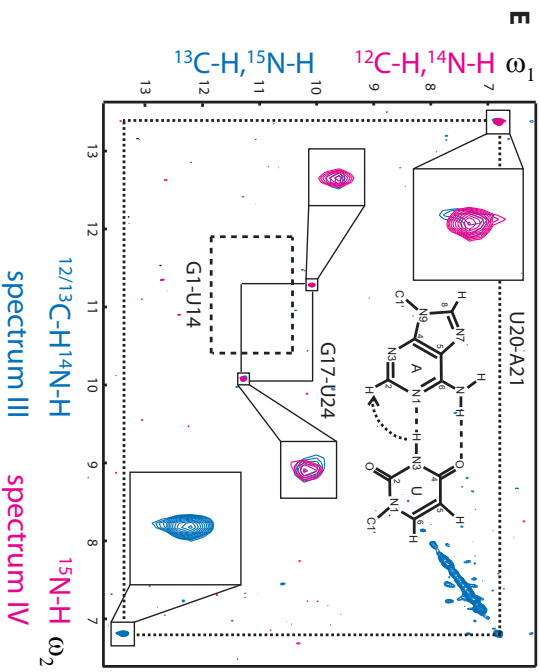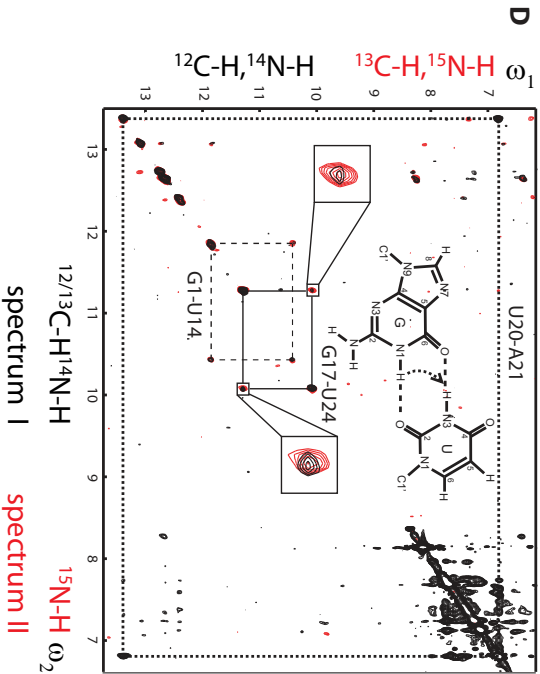
A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat

Using a combination of quantitative HNN-COSY and a novel double isotope-filtered NOESY experiment we could define the conformation of the 26-mer A-repeat RNA. We show that in a single 26-mer A-repeat only the first predicted hairpin is formed, while the second predicted stem-loop forms a RNA duplex and mediates dimerization of the 26-mer A-repeat. Our strategy will be generally applicable to identify and quantify populations of hairpin and duplex conformations in RNAs and to define RNA folding topology from inter- and intra-molecular base-pairing patterns.

Opposite page: Figure 6 A-C) Secondary structure of the Xist RNA A-repeat: A) Stem-loop conformation, B) kissing hairpin and C) full dimerization of the second predicted hairpin. Our results show that the conformation of the 26-mer corresponds to the form shown in C). D) Doubly-filtered and doubly-edited subspectra and E) edited/filtered subspectra derived from the double filtered NOESY. In E) NOEs between pairs of protons where one is attached to an isotope labeled heteroatom and the other one to an unlabeled one are observed. These NOEs can only originate from a duplex RNA species. The corresponding NOE transfers are indicated schematically for G-U and A-U base pairs in panel D and E, respectively. Zoomed-in views of the NOEs involving the intermolecular G17-U24 and U20-A21 base pairs are shown.

Supplementary Material

Materials and Methods

Native polyacrylamide gel mobility assay

Samples were run on a native 12% (w/v) polyacrylamide TBE gel at constant voltage (150V) at 4°C. 1-2 nanomoles of RNA were loaded on each well of a 20 x 20 x 0.1 cm gel. Samples were diluted 1:1 with loading buffer (2xTBE and 0.1% bromophenolblue). To facilitate visualization of small amounts of possible duplexes the gel was silver stained by a 10 min. fix in 10% ethanol followed by incubation for 3 min in 1% nitric acid, washing with water, incubation in 12 mM silver nitrate for 5 min, washing with water and finally development in 280 mM disodium carbonate, 0.017% formaldehyde. The developing reaction was stopped by incubating the gel for 1 min in 5% acetic acid.

Description of magnetization transfer in the HNN-COSY

In a G-C base pair the INEPT transfer from the G imino proton to its attached nitrogen creates $H_{1z}N_{1y}$ anti-phase magnetization at point a) in the HNN-COSY pulse sequence (Supplementary Fig. 1). During the delay T part of this magnetization is transferred to the base-paired ^{15}N nucleus due to the $^{2h}J_{NN}$ coupling. At point b) magnetization proportional to $-2H_{1z}N_{1z}N_{3x}\sin(\pi^{2h}J_{NN}T)$ (where T is the transfer time) has been transferred, and magnetization proportional to $H_{1z}N_{1y}\cos(\pi^{2h}J_{NN}T)$ still remains on the ^{15}N nucleus bonded to the imino proton. During t_1 the transverse terms of this magnetization evolve and are frequency labeled, leading to cross and diagonal peaks at the chemical shifts of the ^{15}N1 and ^{15}N3 nuclei.
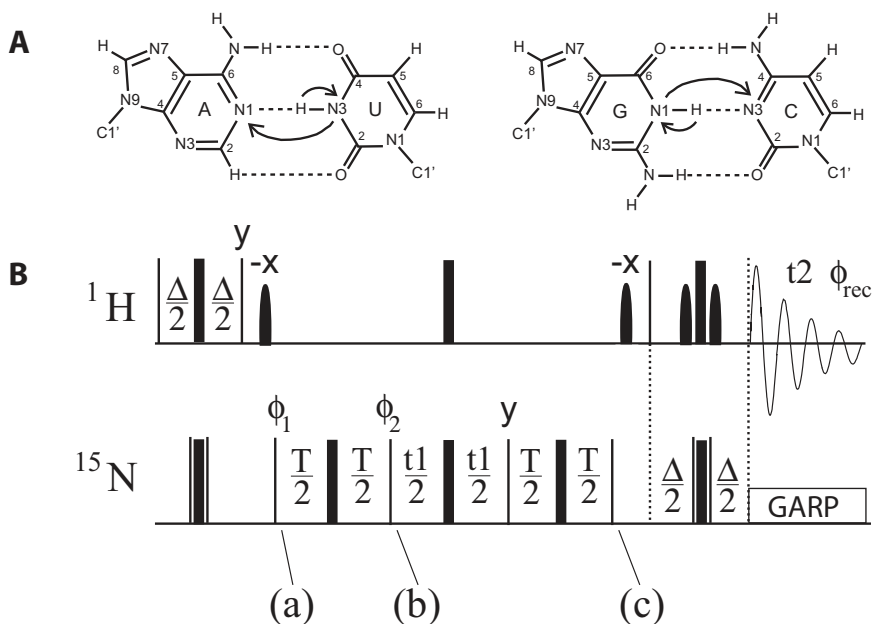
Supplementary Figure 1 The HNN-COSY experiment A) Magnetization transfer in A-U and G-C base pairs B) Pulse sequence: Narrow and wide bars denote 90° and 180° pulses respectively and are applied with phase x if not stated otherwise. 180° pulse on ^{15}N consists of a composite pulse of the form $90°_y 180°_x 90°_y$. Water suppression is achieved by WATERGATE (25) combined with water flip-back (26). Phase cycle: φ_1 = x, -x ; φ_2 = y, -y, φ_{rec} = x, -x. Δ = 2.5ms; T = 20, 30 or 40 ms.

After refocusing during a second T period both magnetization terms give $H_{1z}N_{1y}$ at position c. Intensities of cross and diagonal peaks are proportional to $\sin^2(\pi^{2h}J_{NN}T)$ and $\cos^2(\pi^{2h}J_{NN}T)$ respectively.

Results

Native gel analysis

NMR samples of the Xist 26-mer and the Xist 14-mer, consisting of only the first predicted A-repeat hairpin, were analyzed on native gels.
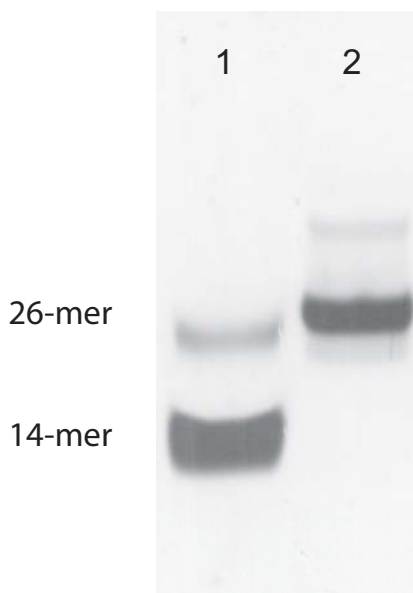
Supplementary Figure 2 Native gel analysis of the A-repeat 26-mer NMR sample suggests a monomeric homogenous conformation.
Lane 1: 14-mer A-repeat first hairpin.
Lane 2: the 26-mer A-repeat NMR sample, both after extensive NMR data collection

Supplementary Fig. 2 shows the native gel analysis of two samples of these constructs after extended NMR data collection. The 14-mer RNA, which we confirmed to be monomeric by our approach, is shown in lane 1. An intense band with higher mobility as well as a small band of lower mobility is shown, which presumably correspond to monomeric and dimeric species. The 26-mer shown in lane 2 has a similar behavior. Initially, we assumed that the intense band represents a monomeric species as it has a similar mobility as the duplex species of the 14-mer in lane 1. However, it should be considered that the conditions (concentration, buffer, matrix/solvent) in the native gel are quite different to those present in the NMR sample. Presumably, the dimerization of the 26-mer is not stable in the gel. Interestingly, 26-mer variants which were designed to stabilize dimerization do show lower mobility as expected for a dimeric species (not shown).

References

[1] Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S. and Brockdorff, N. (1996) Requirement for Xist in X chromosome inactivation. *Nature*, **379**, 131-137.

[2] Sheardown, S.A., Duthie, S.M., Johnston, C.M., Newall, A.E., Formstone, E.J., Arkell, R.M., Nesterova, T.B., Alghisi, G.C., Rastan, S. and Brockdorff, N. (1997) Stabilization of Xist RNA mediates initiation of X chromosome inactivation. *Cell*, **91**, 99-107.

[3] Wutz, A., Rasmussen, T.P. and Jaenisch, R. (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat. Genet.*, **30**, 167-174.

[4] Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*. Wiley, New York.

[5] Varani, G., Aboul-ela, F. and Allain, F.H.-T. (1996) NMR investigation of RNA structure. *Progr. NMR Spectrosc.*, **29**, 51-127.

[6] Wijmenga, S. and van Buuren, B. (1998) The use of NMR methods for conformational studies of nucleic acids. *Prog. Nucl. Magn. Reson. Spectrosc.*, **32**, 287-387.

[7] Fürtig, B., Richter, C., Wöhnert, J. and Schwalbe, H. (2003) NMR spectroscopy of RNA. *Chembiochem*, **4**, 936-962.

[8] Latham, M.P., Brown, D.J., McCallum, S.A. and Pardi, A. (2005) NMR methods for studying the structure and dynamics of RNA. *Chembiochem*, **6**, 1492-1505.

[9] Marky, L.A. and Breslauer, K.J. (1987) Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. *Biopolymers*, **26**, 1601-1620.

[10] Lapham, J., Rife, J.P., Moore, P.B. and Crothers, D.M. (1997) Measurement of diffusion constants for nucleic acids by NMR. *Journal of Biomolecular NMR*, **10**, 255-262.

[11] Cabello-Villegas, J. and Nikonowicz, E.P. (2000) Discriminating duplex and hairpin oligonucleotides using chemical shifts: application to the anticodon stem-loop of Escherichia coli tRNAPhe. *Nucleic Acids Res.*, **28**, e74.

[12] Zanier, K. (2001) Regulation of histone gene expression: solution structure determination by NMR of the 3' histone mRNA hairpin and implications for specific protein-RNA recognition. *Ph.D. thesis*.

[13] Sotoya, H., Matsugami, A., Ikeda, T., Ouhashi, K., Uesugi, S. and Katahira, M. (2004) Method for direct discrimination of intra- and intermolecular hydrogen bonds, and characterization of the G(:A):G(:A):G(:A):G heptad, with scalar couplings across hydrogen bonds. *Nucleic Acids Res.*, **32**, 5113-5118.

[14] Aboul-ela, F., Nikonowicz, E.P. and Pardi, A. (1994) Distinguishing between duplex and hairpin forms of RNA by ^{15}N-^{1}H heteronuclear NMR. *FEBS Lett.*, **347**, 261.

[15] Dingley, A.J. and Grzesiek, S. (1998) Direct observation of hydrogen bonds in nucleic acid base pairs by internucleotide $^{2}J_{NN}$ couplings. *J. Am. Chem. Soc.*, **120**, 8293-8297.

[16] Milligan, J.F. and Uhlenbeck, O.C. (1989) Synthesis of small RNAs using T7 RNA-Polymerase. *Methods Enzymol.*, **180**, 51-62.

[17] Kao, C., Zheng, M. and Ruedisser, S. (1999) A simple and efficient method to reduce nontemplated nucleotide addition at the 3' terminus of RNAs transcribed by T7 RNA polymerase. *RNA*, **5**, 1268-1272.

[18] Grzesiek, S. and Bax, A. (1993) Amino acid type determination in the sequential assignment procedure of uniformly ^{13}C/^{15}N-enriched proteins. *Journal of Biomolecular NMR*, **3**, 185-204.

[19] Logan, T.M., Olejniczak, E.T., Xu, R.X. and Fesik, S.W. (1993) A general method for assigning NMR spectra of denatured proteins using 3D HC(CO)NH-TOCSY triple resonance experiments. *Journal of Biomolecular NMR*, **3**, 225.

[20] Sattler, M., Schleucher, J. and Griesinger, C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.*, **34**, 93-158.

[21] Wörgötter, E., Wagner, G. and Wüthrich, K. (1986) Simplification of two-dimensional H-1-NMR spectra using an X-filter. *J. Am. Chem. Soc.*, **108**, 6162-6167.

[22] Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) NMRPipe - A multidimensional spectral processing system based on Unix pipes. *Journal of Biomolecular NMR*, **6**, 277-293.

[23] Johnson, B.A. and Blevins, R.A. (1994) NMR View - A computer program for the visualization and analysis of NMR data. *Journal of Biomolecular NMR*, **4**, 603-614.

[24] Kumar, A., Ernst, R.R. and Wüthrich, K. (1980) A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochem. Biophys. Res. Commun.*, **95**, 1.

[25] Piotto, M., Saudek, V. and Sklenář, V. (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *Journal of Biomolecular NMR*, **2**, 661.

[26] Grzesiek, S. and Bax, A. (1993) The importance of not saturating H_2O in protein NMR - application to sensitivity enhancement and NOE measurements. *J. Am. Chem. Soc.*, **115**, 12593-12594.

[27] Sklenář, V., Peterson, R.D., Rejante, M.R. and Feigon, J. (1994) Correlation of nucleotide base and sugar protons in a ^{15}N-labeled HIV-1 RNA oligonucleotide by ^{1}H-^{15}N HSQC experiments. *Journal of Biomolecular NMR*, **4**, 117.

[28] Wöhnert, J., Dingley, A.J., Stoldt, M., Görlach, M., Grzesiek, S. and Brown, L.R. (1999) Direct identification of NH...N hydrogen bonds in non-canonical base pairs of RNA by NMR spectroscopy. *Nucleic Acids Res.*, **27**, 3104-3110.

[29] Pervushin, K., Ono, A., Fernandez, C., Szyperski, T., Kainosho, M. and Wüthrich, K. (1998) NMR scalar couplings across Watson-Crick base pair hydrogen bonds in DNA observed by transverse relaxation-optimized spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 14147-14151.

[30] Grzesiek, S., Cordier, F., Jaravine, V. and Barfield, M. (2004) Insights into biomolecular hydrogen bonds from hydrogen bond scalar couplings. *Prog. Nucl. Magn. Reson. Spectrosc.*, **45**, 275.

[31] Zhang, Q., Stelzer, A.C., Fisher, C.K. and Al-Hashimi, H.M. (2007), Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature*, **450**, 1263.

3

NMR assignment of a novel AUCG tetraloop hairpin from a human Xist RNA A-repeat essential for X-inactivation

NMR assignment of a novel AUCG tetraloop hairpin from a human Xist RNA A-repeat essential for X-inactivation

Malgorzata M. Duszczyk and Michael Sattler

Abstract

Initiation of X-inactivation depends on the non-coding RNA *Xist*. We have solved a hairpin structure with a novel AUCG tetraloop fold in an A-repeat of *Xist* that is essential for silencing, and report its full assignments here.

Biological Context

Xist (X-inactivation specific transcript) RNA is a large non-coding RNA essential for the initiation of X-inactivation in mammal females. It is expressed from the X-chromosome that will soon be silenced and coats it *in cis*, which coincides with transcriptional shutdown through an unknown mechanism (1). The conserved so-called 'A-repeats' at the 5' end of *Xist* are essential for its silencing function, while several other regions are redundantly responsible for chromosome association (2). It is believed that these repeats recruit protein factors that induce silencing. In humans the A-repeats are constituted of 7.5 copies of approximately 26 nucleotides length, connected by long U-rich linkers. Although the A-repeat had been predicted to fold into a double hairpin, we have shown that under NMR conditions two A-repeats form an interrepeat dimer through their second predicted hairpins (3) (see Chapter 2). We have determined the three-dimensional structure of the first predicted 14-

mer hairpin and found a novel AUCG tetraloop conformation (Duszczyk *et al.*, in preparation – see Chapter 4). Here we report essentially complete ^{1}H, ^{13}C, ^{15}N and ^{31}P NMR assignments for this first hairpin of a Xist RNA A-repeat.

Methods and Experiments

Sample Preparation

^{13}C,^{15}N uniformly labeled and unlabeled r(GGCGCAUCGGCGCC) Xist RNA 14-mer was prepared by in vitro transcription with T7 RNA polymerase as previously described (3).

NMR samples were prepared in 10 mM NaH$_2$PO$_4$/Na$_2$HPO$_4$ buffer (pH 6.0), 100 mM NaCl, 0.02mM EDTA, 0.02% Azide in 95% H$_2$O, 5% D$_2$O. Final sample volumes were 300 µl (Shigemi) and RNA concentrations between 0.8 and 1.2 mM. Samples were heated to 95° for 5 minutes before snap-cooling on ice to trap the kinetically favored intramolecular monomeric hairpin conformation over a possible intermolecular dimer.

The unlabeled sample used to measure RDCs was prepared by adding filamentous Pf1 phage solution (Asla Biotech, Riga, Latvia) to a total phage concentration of 18 mg/ml.

Data collection and assignments

NMR spectra were acquired at 5 °C (experiments in H$_2$O for assignment of exchangeable protons and their bonded nitrogens) or 25 °C (experiments in D$_2$O for all other assignments) on Bruker DRX600, DRX800 or DRX900 spectrometers equipped with cryogenic probes. Spectra were processed with NMRPipe (4) and analyzed using NMRVIEW (5).

The homogenous monomeric hairpin conformation of the 14-mer was confirmed by measuring the relative intensity of cross- and diagonal peaks in a HNN-COSY experiment recorded on a 50% ^{13}C,^{15}N - labeled, 50% unlabeled sample (3) (see Chapter 2) in H_2O.

Standard methods were used for resonance assignments (6). Figure 1 shows 2D (^{1}H, ^{13}C) CT-HSQC spectra with assignments for the sugar protons and their bonded carbons. Additional 2D (^{1}H, ^{13}C) CT-HSQC spectra with base assignments are shown in Supplementary Fig. 1. Assignment of non-exchangeable protons was started by identifying all protons and carbons belonging to the same sugar ring using 3D (^{1}H, ^{13}C, ^{1}H) HCCH-COSY and TOCSY. Intra-nucleotide correlations between the individual sugar rings and bases were obtained from HCN experiments. Sequential NOE connectivities observed in 2D (^{1}H, ^{1}H) NOESY and 3D (^{1}H, ^{13}C, ^{1}H) NOESY-HMQC were used to assign each nucleotide sequence-specifically using a H1' to H6/H8 sequential walk (Duszczyk *et al.*, in preparation – see Chapter 4).

Notably, the ^{13}C chemical shifts of U7, C8 and G9 are shifted upfield (C1') or downfield (all other carbons) compared to the other residues. This reflects their non A-helical conformations within the loop. Other unusual shifts are the downfield shifted H2' and H3' protons of G9 and the downfield shifted C5' shifts of G9 and G10. H5' and H5" of these two residues have similar shifts. This reflects the backbone conformation around the unusually flipped-out G9 loop nucleotide (Duszczyk *et al.*, in preparation – see Chapter 4).

A TROSY relayed HCCH-COSY (7) correlated H2 and H8 resonances of the adenine and provided some assignments of the non-protonated carbons in the adenine and guanosines. Sequential assignments were confirmed and ^{31}P shifts were assigned with the HCP experiment (Supplementary Fig. 2).

Assignments of the exchangeable imino protons were provided by a sequential walk (Supplementary Fig. 4) in 2D NOESY spectra (50 – 300 ms mixing time) employing the WATERGATE sequence and water flip back for water suppression.

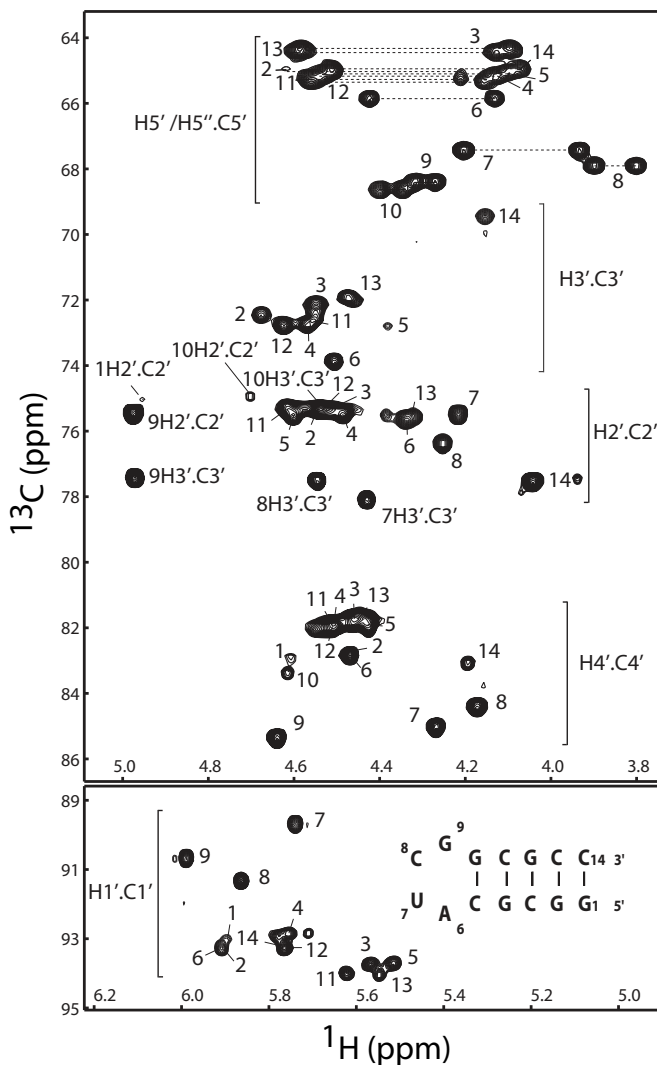**Figure 1** 900 MHz 2D ^{1}H–^{13}C CT-HSQC spectra of the sugar region of the Xist RNA A-repeat first hairpin, with assignments as shown in the secondary structure insert

These experiments also linked the guanosine iminos to the NH_2 and H5 of the cytidines they are base-paired to. The base pairing scheme in the A-form helical stem was further confirmed by the HNN-COSY experiment (Supplementary Fig. 3)

that also provided ^{15}N chemical shifts for the nitrogens bonded to the iminos. Assignments of nitrogens in NH_2 groups were provided by 2D (^{15}N, ^{1}H) HSQC.

The base pairing scheme was independently confirmed in D_2O in a new experiment directly detecting hydrogen bonds in G-C base pairs that correlates guanosine N1 and cytidine N3 chemical shifts to the cytidine H5 (manuscript in preparation) and a long-range HNN-COSY (8). This last experiment also completed ^{15}N assignments for the nucleobases.

Homonuclear $^{3}J_{HH}$ couplings for the sugar spin systems were measured in 3D HCCH-E.COSY (9, 10) and forward directed HCC-TOCSY-CCH-E.COSY (11) experiments and $^{2}J_{C2'P}$ couplings were measured by a 2D spin echo difference CT-HSQC experiment (12). ^{13}C–^{1}H Residual dipolar couplings were measured in natural abundance 2D (^{13}C–^{1}H) TROSY spectra (13).

Extent of assignments and data deposition

The quality of the NMR data obtained is illustrated by the 2D CT-HSQC spectrum shown in Fig. 1. Assignments were obtained for all 14 nucleotides, including 98% of the non-exchangeable protons, 48% of all possible exchangeable protons (keeping in mind that exchangeable protons in non-base paired regions are largely invisible because of rapid exchange with the solvent), 99% of the ribose ^{13}C, 100% of the proton-attached nucleobase ^{13}C, 100% of glycosidic N1/N9, 83% of cytidine amino N4, 80% of guanosine imino N1 / cytidine N3 nitrogens involved in base pairs, and 93% of ^{31}P resonances. Overall 78% of all ^{13}C shifts and 65% of ^{15}N shifts were assigned, keeping in mind that many quaternary carbon shifts were not accessible in the experiments used for assignment, and nitrogens attached to rapidly exchanging protons are not visible.

^{1}H chemical shifts were referenced to H_2O, with heteronuclear ^{13}C, ^{15}N, and ^{31}P chemical shifts referenced indirectly according to the X/^{1}H ratio (14). The ^{1}H, ^{13}C, ^{15}N and ^{31}P chemical shifts, J-couplings and RDCs (see Tables 1-3) will be

deposited in the BMRB (*http://www.bmrb.wisc.edu/*). The coordinates of the Xist RNA A-repeat novel AUCG tetraloop will be deposited at the Protein Data Bank.

Acknowledgements

Supplementary Figures
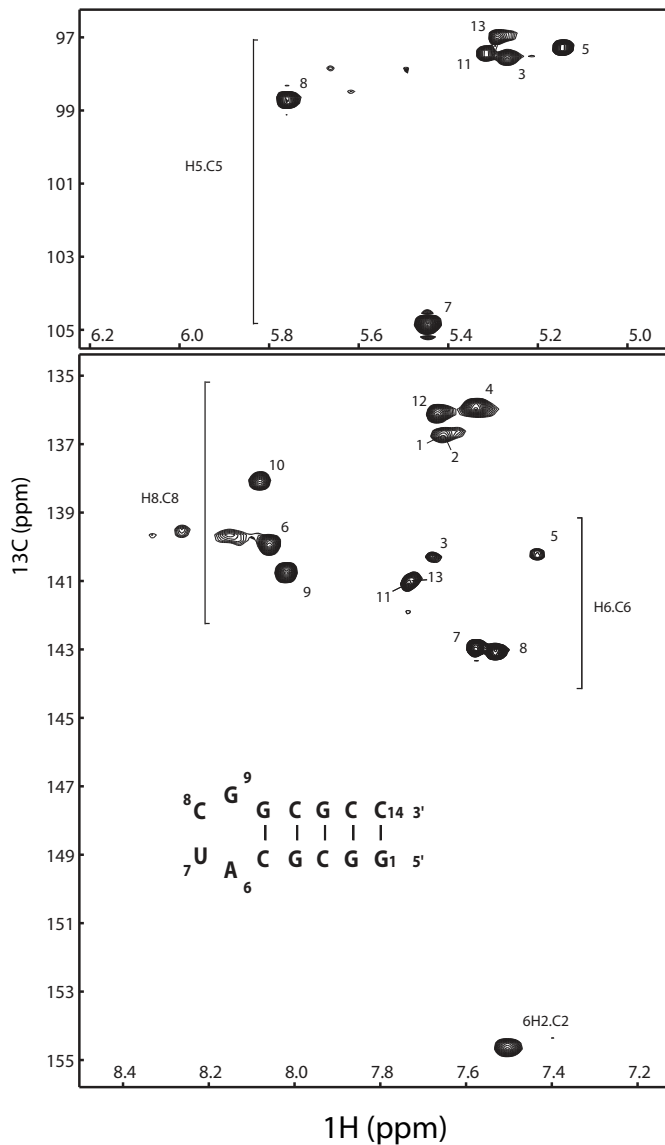
Supplementary Figure 1 900 MHz 2D ^{1}H–^{13}C CT-HSQC base spectra of the Xist RNA A-repeat first hairpin, with assignments as shown in the secondary structure insert
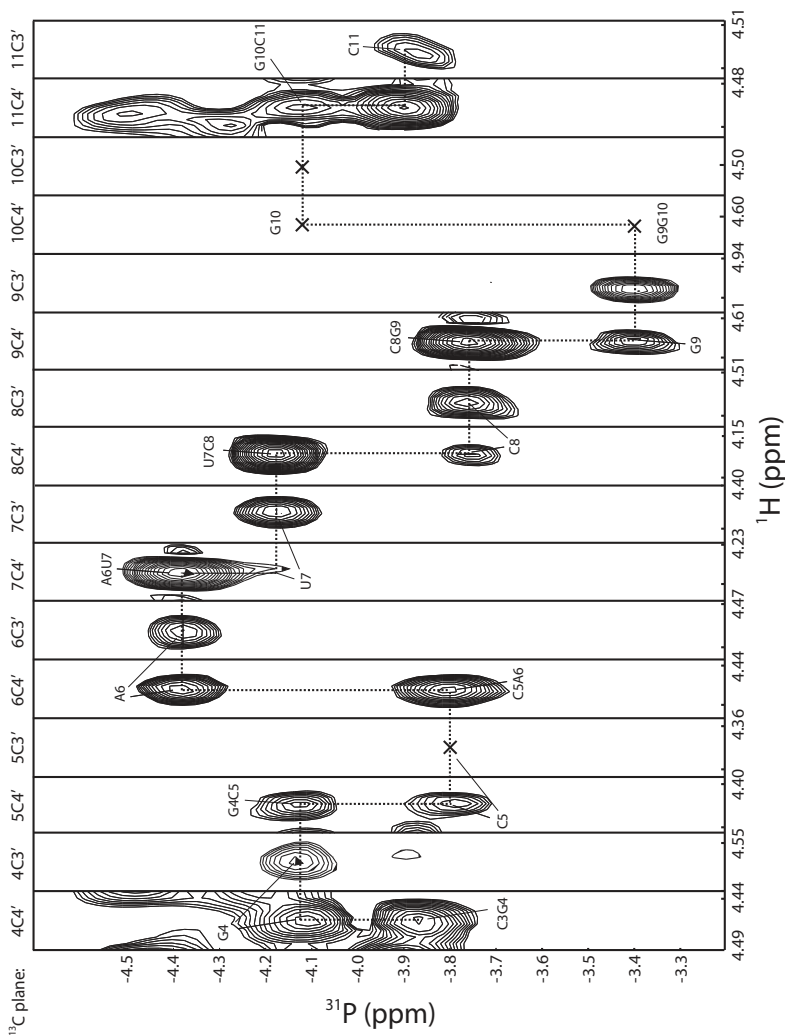
Supplementary Figure 2 C3' and C4' planes of a 900 MHz ^{1}H–^{31}P–^{13}C 3D HCP spectrum in 100% D$_2$O. Assignments as shown in Fig. 1. The dotted line shows intra-residual H3'/H4'-P and sequential H4'-P correlations of residues 4-11. Crosses indicate missing cross peaks – most notable are the missing cross peaks for G9-G10 and G10 due to line broadening due to conformational dynamics of the backbone between 9 and 10 (Duszczyk *et al.*, in preparation).
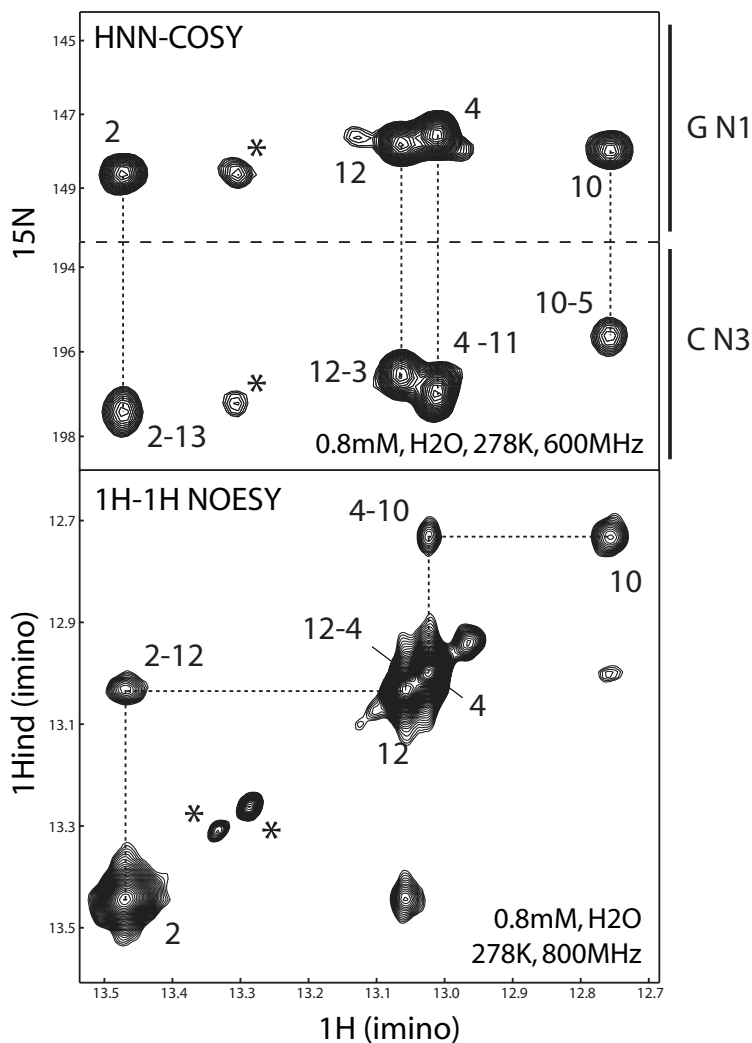
Supplementary Figure 3 Confirmation of the secondary structure of the Xist RNA A-repeat first hairpin: on top is a 600 MHz ^{1}H–^{15}N HNN-COSY spectrum recorded on a ^{13}C^{15}N labeled sample in 95:5 H$_2$O:D$_2$O at 278K. Assignments as shown in Fig. 1. Intra-base imino-N1 correlations for the G's in the stem are connected by dotted lines to the N3's of the C's they are base paired to. The 2D ^{1}H NOESY spectrum at the bottom, recorded on an unlabeled sample at 800 MHz and 278K, connects the G imino atoms by a sequential walk through the stem shown by the dotted line. In both spectra signals for the first G-C base pair are not visible due to end-fraying. Peaks originating from an n+1 RNA contamination are marked with (*).

Table 1A

Residue	P / H1'	C1' / H2'	C2' / H3'	C3' / H4'	C4' / 1H5'	C5' / 2H5'	Base assignments
G1	— / 5.92	93.1 / 4.96	75.0 / 4.83	73.5 / 4.61	83.0 / —	— / —	C8 136.8; H8 7.65; N9 169.3
G2	−4.2 / 5.90	93.1 / 4.54	75.6 / 4.68	72.5 / 4.48	82.6 / 4.61	64.7 / 4.20	C8 136.7; H8 7.65; N1 147.1; N7 232.0; N9 169.5
C3	−3.9 / 5.56	93.6 / 4.53	75.6 / 4.55	72.0 / 4.46	81.8 / 4.59	64.2 / 4.11	C5 97.4; H5 5.27; C6 140.4; H6 7.69; N1 150.2; N3 196.7
G4	−4.1 / 5.76	92.7 / 4.49	75.4 / 4.56	72.7 / 4.48	82.3 / 4.54	64.9 / 4.14	C8 136.0; H8 7.58; C5 118.3; N1 146.7; C4 151.3; N7 235.1; C6 160.7; N9 169.2
C5	−3.8 / 5.51	93.7 / 4.60	75.6 / 4.38	72.8 / 4.43	82.0 / 4.51	64.9 / 4.10	C5 97.3; H5 5.15; C6 140.3; H6 7.44; N1 150.2; N3 195.8
A6	−4.4 / 5.91	93.2 / 4.34	75.6 / 4.49	73.7 / 4.48	83.0 / 4.43	66.0 / 4.14	C8 140.0; H8 8.06; C5 120.7; N1 149.0; N7 231.2; C6 157.0; N9 170.3; C2 154.6; H2 7.51
U7	−4.2 / 5.74	89.5 / 4.22	75.4 / 4.43	78.2 / 4.27	85.0 / 4.21	67.3 / 3.94	C5 104.7; H5 5.44; C6 143.0; H6 7.58; N1 142.8; N3 156.9
C8	−3.8 / 5.86	91.2 / 4.26	76.4 / 4.55	77.6 / 4.18	84.4 / 3.91	67.8 / 3.81	C5 98.7; H5 5.76; C6 143.1; H6 7.54; N1 151.2; N3 199.5
G9	−3.4 / 5.99	90.6 / 4.97	75.4 / 4.97	77.5 / 4.64	85.3 / 4.32	68.3 / 4.27	C8 140.8; H8 8.02; C5 119.1; C4 153.4; N7 235.8; C6 160.7; N9 167.7
G10	−4.1 / 5.61	92.5 / 4.71	75.1 / 4.52	75.0 / 4.62	83.6 / 4.40	68.6 / 4.35	C8 138.1; H8 8.08; N1 147.2; C4 151.8; N7 233.2; N9 168.7
C11	−4.0 / 5.62	94.0 / 4.61	75.4 / 4.54	72.5 / 4.51	81.9 / 4.57	65.3 / 4.21	C5 97.4; H5 5.31; C6 141.0; H6 7.74; N1 150.7; N3 197.1
G12	−4.5 / 5.76	92.6 / 4.53	75.3 / 4.63	72.9 / 4.51	82.0 / 4.57	65.2 / 4.15	C8 136.2; H8 7.67; C5 118.2; N1 147.1; C4 151.4; N7 234.8; C6 160.7; N9 169.2
C13	−4.2 / 5.55	94.0 / 4.32	75.6 / 4.46	72.0 / 4.44	81.8 / 4.59	64.2 / 4.11	C5 97.1; H5 5.29; C6 140.9; H6 7.73; N1 151.2; N3 197.0
C14	−4.2 / 5.76	93.0 / 4.05	77.6 / 4.16	69.4 / 4.19	83.1 / 4.52	64.9 / 4.09	C5 98.5; H5 5.62; C6 141.2; H6 7.69; N1 152.4

Table 1A Chemical shifts (D₂O) for the non-exchangeable protons in the Xist RNA A-repeat 14-mer AUCG tetraloop hairpin

Tables 1A (previous page) / 1B Chemical shifts measured at 25º in D_2O (1A) and at 5º in H_2O (1B) in ppm. Shifts not assigned in 1A are missing due to overlap or because the non-proton-bonded heteroatoms were invisible in the experiments recorded. Shifts missing in 1B are unassigned because of line broadening due to exchange with the solvent. Estimated errors are 0.02 ppm for proton, 0.1 ppm for nitrogen, and 0.2 ppm for carbon shifts.

Table 1B

G1	N1		H1		N2		1H2		2H2	
G2	N1	148.6	H1	13.47	N2		1H2		2H2	
C3	N3	196.5	N4	99.2	1H4	8.60	2H4	6.71		
G4	N1	147.6	H1	13.00	N2		1H2		2H2	
C5	N3	195.6	N4	98.2	1H4	7.93	2H4	6.83		
A6	N6	80.5	1H6		2H6	6.72				
U7	N3		H3							
C8	N3		N4	95.0	1H4	7.23	2H4	6.78		
G9	N1		H1		N2		1H2		2H2	
G10	N1	148.0	H1	12.74	N2	73.5	1H2		2H2	6.54
C11	N3	197.0	N4	98.0	1H4	8.60	2H4	6.69		
G12	N1	147.8	H1	13.05	N2		1H2		2H2	
C13	N3	197.4	N4	100.0	1H4	8.64	2H4		6.98	
C14	N3		N4		1H4		2H4			

Table 1B Chemical shifts (H_2O) for the exchangeable protons in the Xist RNA A-repeat 14-mer AUCG tetraloop hairpin

Table 2

	H1'H2'	H3'H4'	H4'H5'	H4'H5''	C2'P
G1					
G2	0.8	9.7	4.4	<1.0	
C3	0.5				
G4	0.1	10.7			
C5	1.9	10.1			<1.0
A6	0.8	10.4	<1.0	<1.0	
U7	7.2	1.7	<1.0	<1.0	6.3
C8	5.0	2.5	<1.0	2.1	4.6
G9	4.9	2.8	3.3	<1.0	4.2
G10	2.3	7.1	<1.0	3.2	<1.0
C11	0.9	11.8			
G12	0.1	10.6			
C13	0.5	10.9			
C14	1.1	9.9			<1.0

Table 2 Experimental $^3J_{HH}$ and $^3J_{CP}$ couplings measured in D_2O in Hz. Missing couplings could not be measured because of spectral overlap or line broadening. Estimated errors are in the range of 1 Hz

Table 3

	H1'C1'	H2'C2'	H3'C3'	H4'C4'	H5C5	H6C6	H8C8
G1							
G2			0.2				
C3	5.7						
G4							9.8
C5	2.4		13.6		7.3	2.7	
A6			3.8	7.8			4.6
U7	2.6	3.9	-9.4	0.3	4.9	4.6	
C8	-0.7		-2.0	-0.4	-0.7	4.5	
G9	1.1	-0.7	3.3	1.8			1.5
G10	-6.5	4.2					4.7
C11	-13.1		3.2				
G12							2.8
C13	1.2		5.8		5.0		
C14		2.5	6.2	11.8	12.5		

Table 3 Experimental ^{13}C-^1H RDCs for the Xist RNA A-repeat 14-mer AUCG tetraloop hairpin measured in D_2O in Hz for non-overlapped signals not affected by line broadening. RDCs for C8, G9 and C14 were not used in the refinement of the AUCG tetraloop because of suspected mobility (Duszczyk *et al.*, in preparation – see Chapter 4). Estimated errors are in the range of 1 Hz.

References

[1]   Sheardown, S.A., Duthie, S.M., Johnston, C.M., Newall, A.E.T., Formstone, E.J., Arkell, R.M., Nesterova, T.B., Alghisi, G.-C., Rastan, S. and Brockdorff, N. (1997) Stabilization of Xist RNA mediates initiation of X chromosome inactivation. *Cell*, 91, 99.

[2]   Wutz, A., Rasmussen, T.P. and Jaenisch, R. (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat. Genet.*, 30, 167-174.

[3]   Duszczyk, M.M., Zanier, K. and Sattler, M. (2008) A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat. *Nucl. Acids Res.*, 36, 7068-7077.

[4]   Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) NMRpipe - A multidimensional spectral processing system based on Unix pipes. *Journal of Biomolecular NMR*, 6, 277-293.

[5]   Johnson, B.A. and Blevins, R.A. (1994) NMRView - A computer program for the visualization and analysis of NMR data. *Journal of Biomolecular NMR*, 4, 603-614.

[6]   Wijmenga, S.S. and van Buuren, B.N.M. (1998) The use of NMR methods for conformational studies of nucleic acids. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 32, 287.

[7]   Simon, B., Zanier, K. and Sattler, M. (2001) A TROSY relayed HCCH-COSY experiment for correlating adenine H2/H8 resonances in uniformly 13C-labeled RNA molecules. *Journal of Biomolecular NMR*, 20, 173.

[8]   Hennig, M. and Williamson, J.R. (2000) Detection of N-HN hydrogen bonding in RNA via scalar couplings in the absence of observable imino proton resonances. *Nucleic Acids Res.*, 28, 1585-1593.

[9]   Marino, J.P., Schwalbe, H., Glaser, S.J. and Griesinger, C. (1996) Determination of gamma and stereospecific assignment of H5' protons by measurement of (2)J and (3)J coupling constants in uniformly C-13 labeled RNA. *Journal of the American Chemical Society*, 118, 4388-4395.

[10] Zimmer, D.P., Marino, J.P. and Griesinger, C. (1996) Determination of homo- and heteronuclear coupling constants in uniformly C-13,N-15-labeled DNA oligonucleotides. *Magnetic Resonance in Chemistry*, 34, S177-S186.

[11] Schwalbe, H., Marino, J.P., Glaser, S.J. and Griesinger, C. (1995) Measurement of H,H-coupling constants associated with nu-1, nu-2, and nu-3 in uniformly C-13-labeled RNA by HCC-TOCSY-CCH-E.COSY. *Journal Of The American Chemical Society*, 117, 7251-7252.

[12] Legault, P., Jucker, F.M. and Pardi, A. (1995) Improved measurement of ^{13}C, [31]P J-coupling constants in isotopically labeled RNA. *FEBS Letters*, 362, 156.

[13] Meissner, A. and Sorensen, O.W. (1999) The role of coherence transfer efficiency in design of TROSY-type multidimensional NMR experiments. *Journal of Magnetic Resonance*, 139, 439.

[14] Wishart, D.S. and Sykes, B.D. (1995) ^{1}H, ^{13}C and ^{15}N chemical shift referencing in biomolecular NMR. *Journal of Biomolecular NMR*, 6, 135-140

4

The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization

The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization

Malgorzata M. Duszczyk, Anton Wutz, Michael Sattler

Abstract

X-chromosome inactivation (XCI) in female mammals depends on the non-coding RNA *Xist* (X inactivation specific transcript). The mechanism through which *Xist* initiates this unique chromosome-wide silencing event is largely unknown. Protein factors that induce silencing are thought to be recruited by the conserved A-repeats in the 5' end of *Xist* which are essential for its silencing function. We report the solution structure of a 14-mer hairpin within a single A-repeat. The structure reveals a novel well defined AUCG tetraloop fold. The novel fold is stabilized by 5' stacking of the A and U on top of the helical stem and possibly hydrogen bonding between A and the phosphate backbone between C and G. The C is folded back into the minor groove and G is solvent exposed. Additional to this AUCG tetraloop hairpin, a single A-repeat contains a sequence that was earlier predicted to fold into a second hairpin but that we have shown previously to be unfolded and to form a duplex with a second A-repeat *in vitro*. Here we show that mutants in this sequence that disrupt dimerization of the A-repeats *in vitro* do not initiate silencing *in vivo*. Considering these *in vivo* data and the high local concentration of A-repeats as they are connected by U-rich linkers and *Xist* is localized to the X-chromosome, we propose that multimerization of the A-repeats could also be relevant *in vivo*. We

propose a model for A-repeat function where this dimerization and specific recognition of the AUCG tetraloop function together in *Xist* regulation and accumulation.

Introduction

Dosage compensation is a mechanism that evolved to compensate for the difference in X-linked gene expression in species with different numbers of X chromosomes between the sexes (1). The mammalian solution to dosage compensation is X-chromosome inactivation (XCI), the inactivation of one of the two X chromosomes in females. Initiation of XCI takes place early in development and depends on the large non-coding RNA *Xist* (X inactivation specific transcript), which is unique to placental mammals (2). *Xist* is expressed exclusively from the X-chromosome that will be silenced and coats it *in cis*, which coincides with transcriptional shutdown through an unknown mechanism (3).

Differential treatment of two X chromosomes in a single cell is facilitated by regulatory mechanisms of so-called 'counting' of the number of X chromosomes and 'choice' to inactivate all but one of them (4). These mechanisms are controlled from a single locus on the X chromosome, the X inactivation centre (Xic), which contains several regulatory elements including the *Xist* gene and an overlapping gene for *Tsix*, another non-coding RNA that is transcribed in antisense orientation and down-regulates *Xist* expression (5). Recently *Xist-Tsix* sense-antisense duplexes have been detected *in vivo*. During XCI these are processed to small RNAs in a Dicer-dependent manner. It has been proposed that this might only take place on the active X (Xa) which could be key to the differential treatment of the X chromosomes: on the Xa *Xist* expression is locally repressed in an RNAi-like manner and on the inactive X (Xi) *Xist* is allowed to accumulate and induces silencing (6).
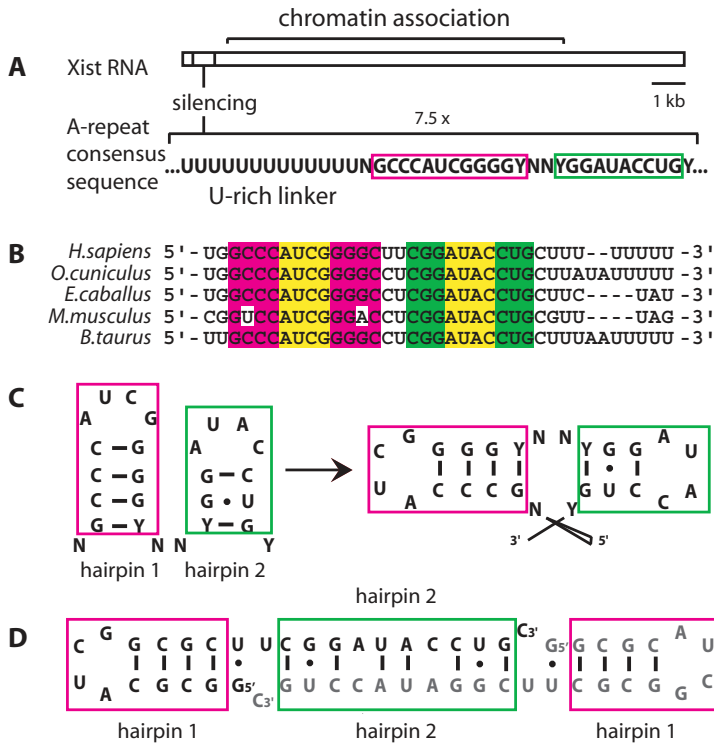
Fig. 1 Schematic structure and sequence of Xist RNA and its A-repeats. A) Xist RNA is a long (15kb in mouse, 17kb in human) non-coding RNA. The A-repeats, located at the 5' end are essential for silencing, while other regions are redundantly responsible for chromosome association. B) Sequence alignment of the 5th A-repeat shows that it is highly conserved between placental mammals. C) The A-repeats consist of 7.5 copies of a conserved sequence predicted to fold into two hairpins, connected by long U-rich linkers. N = any nucleotide; Y = C/U. D) *In vitro*, 'hairpin 2' is unfolded and mediates dimerization with a second A-repeat.

Xist accumulation on the Xi and its ability to trigger silencing are functionally separable (7). The so-called 'A-repeats' at the 5' end of *Xist*, conserved among all placental mammals, are essential for initiation of silencing, while several other regions are redundantly responsible for chromosome association (Fig. 1A). In humans, the A-repeats are constituted of 7.5 copies of a highly conserved 26-nucleotide motif, connected by long U-rich linkers (Fig. 1B). Xist RNA lacking the

The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization

A-repeats accumulates on the Xi but is not able to induce silencing. It has been proposed that the A-repeats are a recruitment site for factors that act in gene repression. Part of the reason why the mechanism by which *Xist* induces silencing remains unknown is because these interacting factors had not yet been isolated. However, recently it was reported that the Polycomb complex PRC2 is a direct target for the A-repeats as a small separate transcript, before accumulation of full length *Xist* occurs, with Ezh2 (Enhancer of Zeste homologue 2) serving as the RNA binding subunit (8). It was also shown that *Tsix* RNA inhibits this interaction by competing for PRC2 binding. It was recognized previously that Polycomb complexes are associated with the Xi and are responsible for the establishment of the histone H3-K27 methylation marks required for long-term X inactivation (9). These data show that it is the A-repeats specifically that target the Polycomb complexes to the Xi. It is suggested that upon *Xist* accumulation binding of PRC2 to the A-repeats within the full length Xist RNA spreads H3-K27 methylation along the Xi.

At present no molecular insight into XCI and no structural information on the A-repeats is available. A single A-repeat has been predicted to fold into a double hairpin structure where the two hairpins possibly stack on top of each other (Fig. 1C). Recently we reported that *in vitro* only 'hairpin 1' is formed, while 'hairpin 2' does not fold as predicted but forms an RNA duplex as shown in Fig. 1D, which mediates dimerization of the 26-mer A-repeat (10) (see chapter 2).

Here we present the NMR structure of a 14-mer hairpin with a fully conserved novel AUCG tetraloop fold, corresponding to 'hairpin 1' of the A-repeats. We correlate our structure with *in vivo* data from a mouse ES cell system with an inducible *Xist* transgene into which mutations can be introduced. Mutants that disrupt the AUCG tetraloop fold impair *Xist* function. In addition, 'hairpin 2' mutants that destabilize dimerization of the A-repeats *in vitro* and are unable to induce silencing *in vivo*. Our results suggest that the dimerization seen *in vitro* may be relevant for *Xist* function *in vivo*. We propose a model for possible Xist RNA A-repeat architecture and function.

Results

Construct design for structure elucidation

The 14-mer sequence investigated in the work presented here (see insert in Fig. 2) is identical to the one in the 5th human A-repeat (see Fig. 1B), apart from switching the positions of G and C in the third G-C base-pair to facilitate chemical shift assignments and with G-C in stead of G-U as a closing base pair because of superior spectral quality (data not shown). Previous studies have shown that changing the sequence of the stem in hairpin 1 while keeping the base pairing intact does not influence Xist activity (7), moreover, a 26-mer construct with this exact hairpin 1 sequence was able to induce silencing in an *in vivo* assay (see 'NMR-XCR' in Fig. 6D). The homogenous formation of a monomeric hairpin was directly confirmed by a method based on the HNN-COSY experiment (10, 11) (see Chapter 2). Comparison of 2D (^{1}H, ^{1}H) TOCSY and NOESY spectra confirmed that the 14-mer AUCG tetraloop adopts the same conformation as within the context of the full 26-mer single A-repeat (see Suppl. Fig. 1).

Chemical shift assignment

Part of the assignment procedure, which is described in full elsewhere (Duszczyk *et al.*, to be submitted to Biomolecular NMR assignments – see Chapter 3) and the quality of the NMR data are illustrated by the aromatic – anomeric sequential walk using H6/H8 to H1' connectivities shown in Fig. 2. To aid this sequential walk the strong H6-H5 correlations in a (^{1}H, ^{1}H) TOCSY were used to distinguish between the purine and pyrimidine spin systems and carbon shifts from a (^{1}H, ^{13}C) CT-

HSQC were used to distinguish the H5 pyrimidine residues of cytosine from uracil. Sequential connectivities could be traced from G1 to C5 (shown in green) and G9 to C14 (shown in magenta), confirming the predicted secondary structure. The intensities of intra-nucleotide H6/H8 – H1' correlations reflect the *syn* and *anti* conformation of the nucleotides. The strong intensity of its H8-H1' cross peak suggests that G9 adopts a *syn* conformation, the medium intensity of C8 suggests a possible intermediate conformation, while all other nucleotides are in the *anti* conformation.

Structure determination

We determined the structure of the Xist RNA A-repeat 14-mer AUCG tetraloop hairpin using 414 NOE-derived distance restraints, 96 torsion angle restraints and 25 restraints derived from residual dipolar couplings (RDCs). NOEs used to restrain the AUCG tetraloop and neighboring residues are shown in Fig. 3. NOEs that were visible only in NOESY spectra with mixing times longer than 80 ms are shown in dotted lines, while NOEs visible at 80 ms, the mixing time of the spectra from which distance restraints were extracted, are shown in solid lines. Experimental torsion angle restraints were derived from analysis of scalar coupling data which is summarized in Table 1A for the AUCG tetraloop and neighboring residues. This analysis revealed that the γ angle is *gauche⁺* for all residues, the ε angle *trans* for C5 and G10, *gauchē* for U7 and mixed *trans / gauchē* for C8 and G9. The ε torsion angle of A6 could not be extracted due to spectral overlap. Sugar puckers were determined to be in the C3'-endo/N conformation for C5 and A6, in the C2'-endo/S conformation for U7 and in a mixed conformation for C8 to G10. Based on averaged coupling data the sugar pucker and torsion angle ε for C8, and all dihedral angles for G9 and G10 were left unrestrained in the structure calculations.
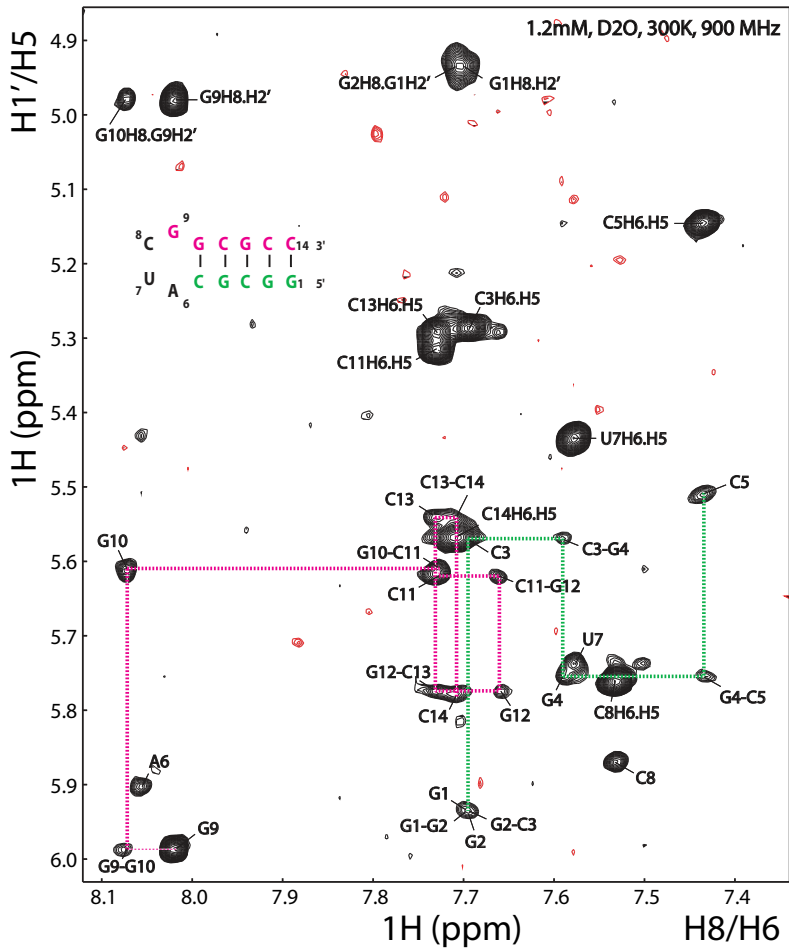
Figure 2 (^{1}H,^{1}H) 2D NOESY spectrum in 100% D$_2$O of the Xist RNA A-repeat 'hairpin 1' recorded at 900 MHz, showing the aromatic (H6/H8) to anomeric (H1' and some H2') region. The lines illustrate the 'sequential walk' up one side of the RNA stem in green and down the other in magenta, with H6/H8-H1' intra- and interresidual cross peak assignments as shown in the inserted secondary structure representation. Intra-base H6-H5 cross peak assignments are shown as well, as are some assignments of cross peaks to the H2' atoms of G1 and G9.
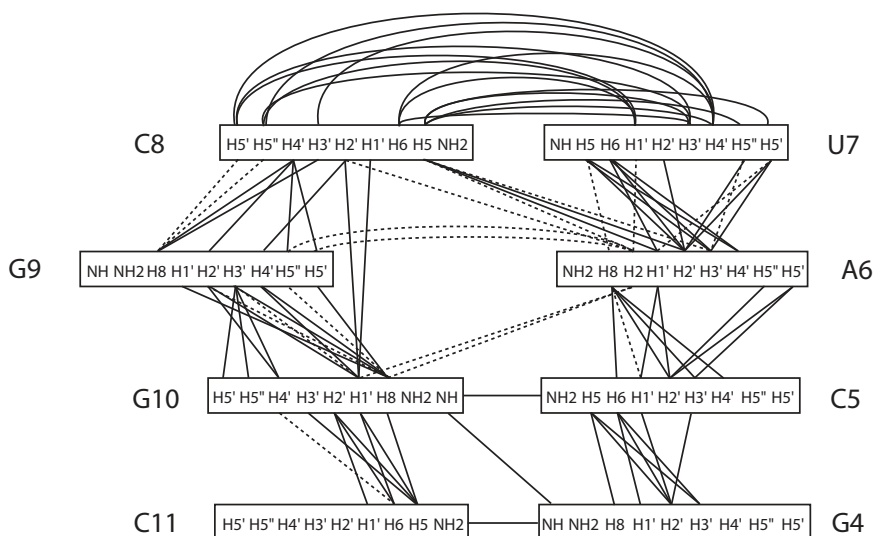
Figure 3 NOE connectivities between protons within the AUCG tetraloop and neighboring residues. NOEs shown in solid lines are visible in NOESY spectra with mixing times of 80 ms, NOEs shown in dotted lines only at longer mixing times.

The dihedral angles in the final ensemble of structures are given in Table 1B. Because of possible conformational dynamics due to fast N/S interconversion, RDCs for C8 and G9 were not included in the structure refinement.

The structure of the 14-mer Xist RNA AUCG tetraloop

The 14-mer adopts a well defined stem–loop structure with a standard A-form helix capped by a structured AUCG tetraloop, stabilized by base stacking interactions (Fig. 4). The structure is well defined by the NMR data as demonstrated by the final ensemble of 10 structures (Fig. 4D) that were selected based on restraint violation

analysis and back-calculation of chemical shifts from an AMBER-refined ensemble of 20 lowest energy structures calculated with ARIA/CNS (12,13).

A summary of structural statistics for the final ensemble, with, and without inclusion of RDCs in the final refinement is given in Table 2. The final ensemble without RDC refinement converges to a RMSD value of 0.54 Å (superimposing all heavy atoms), which improves only slightly to 0.52 Å by including RDCs. The small improvement could be explained by the limited number of RDCs available for refinement or because the tight dihedral restraints used for the stem residues already impose good convergence. The long range RDCs, although consistent with the structures, do not improve the RMSD much in this case. The potential impact of including the tight A-form restraints for residues 1-5 and 10-14 on the AUCG tetraloop structure was analyzed by comparing the structures with these tight ($\pm$ 10°) and also with looser ($\pm$ 30°) dihedral restraints. It was found that the tetraloop structures are essentially identical with either tight or loose A-form dihedral restraints on the Watson–Crick base pairs (results not shown), showing that the structure of the tetraloop is not influenced by the tight A-form restraints on the adjacent helical nucleotides.

The AUCG tetraloop adopts a fold in which A6 and U7 continue 5' base stacking over the A-form helix. At the U7-C8 step the phosphodiester backbone is reversed allowing C8 to fold back into the minor groove, with the H5/H6 side pointing inwards. Following C8 the backbone makes a twist towards the major groove before G9 is fully bulged out in a *syn* conformation. This last backbone twist orients one of the G9 phosphate oxygens towards A6, bringing its H2 within hydrogen-bonding distance. The chemical shift of the A6 H2 is not unusual, but deshielding effects originating from such an H-bond could be compensated for by the ring currents of the bases stacked above and below the A6. Exceptionally striking about the tetraloop structure is the twist in the backbone between C8 and G9 and the conformation of G9, which is rarely found flipped out in RNA hairpin loops.

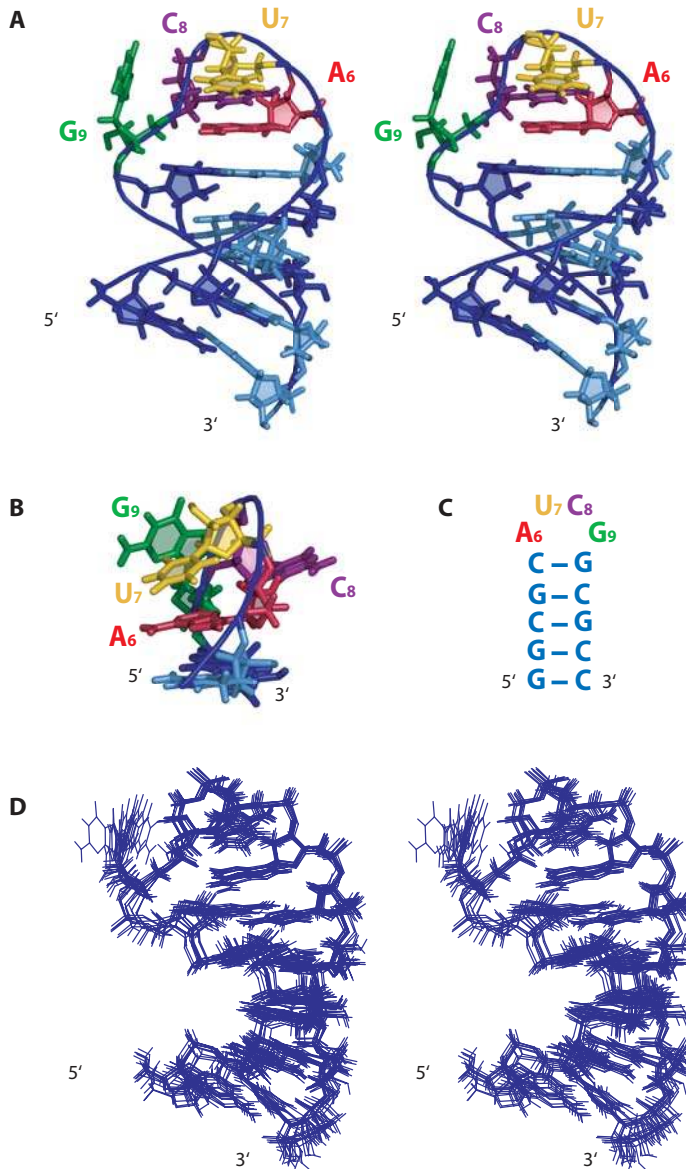**Figure 4** Solution structure of the Xist RNA A-repeat AUCG tetraloop. A) Stereo view of the lowest energy structure and B) view of the AUCG tetraloop from the side. The loop nucleotides are colored red (A), yellow (U), purple (C) and green (G) as shown in the secondary structure in C). D) Stereo view of the NMR ensemble of 10 final AMBER-refined structures, superimposed over all heavy atoms.

	C5	A6	U7	C8	G9	G10
A) Experimental J-couplings						
$^3J_{H1'H2'}$	1.9	0.8	7.2	5.0	4.9	2.3
$^3J_{H3'H4'}$	10.1	10.4	1.7	2.5	2.8	7.1
Pucker	**N**	**N**	**S**	mixed	mixed	mixed
$^3J_{H4'H5'}$	nd	<1.0	<1.0	<1.0	3.3	<1.0
$^3J_{H4'H5''}$	nd	<1.0	<1.0	2.1	<1.0	3.2
γ	**g⁺**	**g⁺**	**g⁺**	**g⁺**	g⁺	g⁺
$^3J_{C2'P}$	<1.0	nd	6.3	4.6	4.2	<1.0
ε	**t**	na	**g⁻**	t/g⁻	t/g⁻	t
B) Torsion angles in the final NMR ensemble						
α	-69.3 ± 1.2	-73.2 ± 1.9	-73.0 ± 1.8	89.3 ± 3.0 trans	49.9 ± 1.7 trans	-122.8 ± 64.7 trans
β	178.2 ± 1.8	172.7 ± 3.0	178.8 ± 2.6	177.3 ± 2.3	177.8 ± 3.3	139.8 ± 57.3
γ	60.4 ± 1.3 gauche⁺	68.2 ± 3.7 gauche⁺	67.5 ± 2.1 gauche⁺	65.0 ± 2.0 gauche⁺	**-177.3 ± 5.2 trans**	142.3 ± 61.4 g⁺/t
δ	78.7 ± 1.6 N	84.3 ± 3.7 N	143.4 ± 1.7 S	**85.6 ± 2.7 N**	**166.4 ± 2.2 S**	**95.3 ± 10.4 N**
ν1	-24.4 ± 4.6 N	-18.7 ± 4.4 N	38.7 ± 1.4 S	-39.2 ± 1.2 mixed	10.0 ± 6.2 mixed	-33.7 ± 1.9 mixed
ν2	36.6 ± 3.4 N	31.8 ± 1.0 N	-36.3 ± 1.5 S	41.5 ± 0.8 mixed	-30.3 ± 3.6 mixed	**33.7 ± 7.0 N**
ε	-161.6 ± 1.6 trans	-159.8 ± 2.3 trans	-65.5 ± 1.0 gauche⁻	**-176.5 ± 2.5 trans**	-78.8 ± 10.4 t/g⁻	-153.0 ± 7.1 trans
ζ	-62.8 ± 2.1	-61.5 ± 2.1	102.5 ± 1.7 trans	-128.4 ± 4.3 trans	-5.5 ± 51.3 trans	-67.7 ± 2.1
χ	-162.4 ± 3.5 anti	-163.2 ± 1.9 anti	-124.9 ± 3.3 anti	-68.5 ± 1.2 syn	-38.3 ± 10.3 syn	-175.8 ± 5.4 anti

Table 1 Experimental J-couplings and torsion angles for the AUCG tetraloop A) Sugar pucker conformations (C3'-endo/N and C2'-endo/S) and the backbone dihedral angles γ and ε as determined by the J-couplings are shown. The angles that were actually restrained during the structure calculations (as described in the Materials and Methods section) are highlighted in bold.
B) Torsion angles and the rotamers they define in the final ensemble of NMR structures. Highlighted in bold are torsion angles that converged to a single conformation while showing averaged J-coupling data. An explanation is given in the results section.

The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization

NMR restraints	No RDCs	RDCs
Total restraints	414 (30/residue)	445 (32/residue)
NOEs	293	
Intra-residual	177	
Inter-residual	116	
Torsion angles	96	
H-bonds for paired residues	25	
RDCs	0	25
Structural statistics		
NOE violations		
Number (>0.2Å)	0.4	0.6
Maximum violations (Å)	0.33	0.35
Torsion violations		
Number (>5°)	0.2	0.1
Maximum violations (°)	20.7	5.9
RCDs violations		
Number (>2 Hz)	12.5	1.6
Q-factor	0.61 ± 0.34	0.23 ± 0.01
Maximum violations (Hz)	20	3.8
R.M.S.D. from the mean coordinates (Å)		
All heavy atoms	0.54	0.52
5-10 (loop)	0.44	0.45
1-5 & 10-14 (stem)	0.44	0.42
R.M.S.D. from ideal geometry		
Bonds lengths (Å)	0.01 ± 0.00	0.01 ± 0.00
Bonds angles (°)	2.03 ± 0.03	2.07 ± 0.03
Average Amber Energy (kcal/mol)	-2422 ± 3	-2410 ± 11

Table 2 NMR experimental restraints and structural statistics. Structural statistics are averages calculated for the bundle of 10 selected AMBER refined lowest energy structures

Fig. 3 shows that this tetraloop conformation is defined by a large number of NOEs. NOEs between the base of A6 and the sugar of C5 and the base of U7 and the sugar of A6 indicate continuation of stacking on the 5' side of the loop. Numerous sugar/sugar contacts between U7 and C8 deviate from A-helical stacking. The location of C8 in the minor groove is supported by NOEs between the H5 of C8 and the A6 sugar protons. The flipped out orientation of G9 is supported by lack of NOEs from its base to G10 and by direct long range sugar/sugar NOEs between C8 and G10. The unusual conformation of the G9 sugar is also supported by its unusually downfield shifted H2' and H3' protons and the unusual shifts of H5'/H5" of both G9 and G10, that are almost degenerate for both residues.

The experimental backbone dihedral angles and sugar puckers (Table 1B) only have A-helical geometries for A6, which extends A-form stacking on the 5' side of the tetraloop, while dihedral angles of U7 to G9 deviate from A-form values, allowing for reversal of the phosphodiester backbone. The phosphates of C8, G9 and G10 are flanked by unusual α and ζ torsion angles, which for G9 is also supported by a downfield ^{31}P chemical shift and missing sequential connectivities between the 9 and 10 H4' and the 10 phosphorous in a HCP spectrum (Duszczyk *et al.*, to be submitted to Biomolecular NMR assignments – see Chapter 3).

Dihedral angles in the final ensemble are in good agreement with the experimental J-couplings for C5 to U7. Experimental couplings for C8 to G10 indicate averaged sugar puckers and ε / γ angles. However, this is not reflected in the final structural ensemble. Presumably, the good convergence of the final ensemble does not reflect conformational dynamics / flexibility of the loop between C8 and G10. Apart from the averaged J-coupling data, other evidence supporting such flexibility are the missing connectivities in the HCP mentioned before and a double resonance for the G10 H3' sugar proton (not shown).

In summary, the Xist RNA AUCG tetraloop structure is primarily defined by base stacking interactions, possibly further stabilized by hydrogen bonding at the 5' side,

and shows a less defined structure at the 3' side, possibly involving dynamic residues C8 and G9 of which especially G9 is highly solvent exposed and easily accessible for intermolecular interactions.

Structural validation based on chemical shifts

Proton chemical shifts are highly sensitive for structural variation due to large aromatic ring current effects and can be used as such as an independent structural validation tool for structures calculated based on NOEs, J-couplings and RDCs (14). As a detailed comparison of the predicted versus the experimental shifts can detect local discrepancies between the derived and the actual structure, we back-calculated chemical shifts of our 20 initial structures refined by AMBER with the program NUCHEMICS (15) and selected the best 10 structures based on violation analysis and best fit between calculated and experimental chemical shifts.

Experimental shifts and the back calculated shifts from the final structural ensemble for the AUCG tetraloop and its neighboring residues are shown in Fig. 5. Typical A-form chemical shifts and good agreement with experimental data are seen for C5, A6 and G10. Other residues show some deviations from the experimental data: for U7, calculated H1' and H4' shifts are shifted significantly upfield in comparison to the experimental shifts. It should be noted that both these atoms are directed towards the C8 sugar that, as discussed above, possibly shows N/S conformational averaging that might not be reflected in the final structural ensemble. For C8 most notable are the H5 and H6 shifts that are shifted downfield of the experimental shifts. An explanation for this is possible mobility of C8 that is not reflected in the final structural ensemble.

In agreement with Cromsigt *et al.*, (14) the downfield H2' and H3' shifts can be correlated with a *syn* base conformation for G9. Its pronounced flipped-out conformation is supported by excellent agreement between calculated and experimental non-A-form chemical shifts for all sugar protons except H5".

Regarding this proton, it should be noted that, as also stated by Cromsigt *et al.*, chemical shifts strongly depend on distance and relative small structural adjustments can change the values of the shifts dramatically. This is clearly visible for the G9 H5" proton, where the most extreme chemical shifts calculated based on different members of our final structural ensemble differ almost a whole ppm unit. The structure from which the best H5" shift was back calculated is the one where the G9 base is oriented the furthest away from the G9 H5" proton.
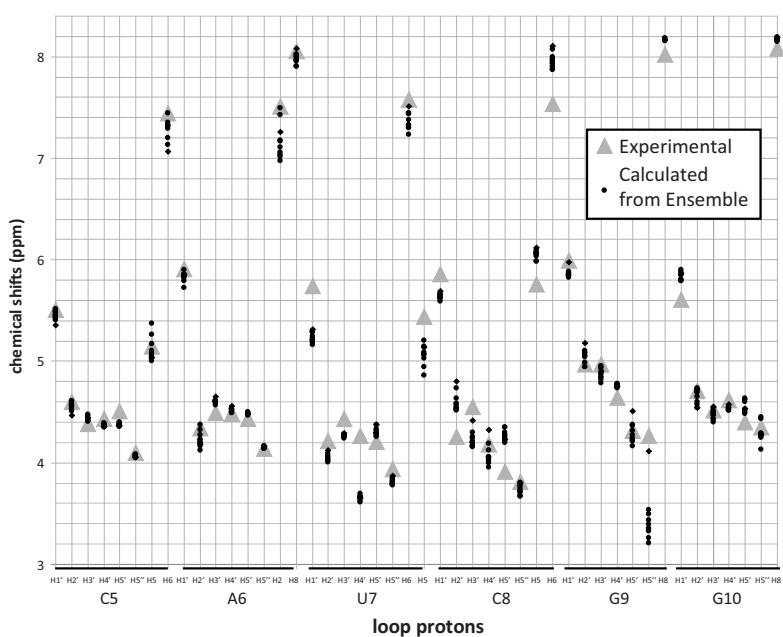

Figure 5 Experimental proton chemical shifts (triangles) and chemical shifts calculated with NUCHEMICS (15) from the final AMBER-refined structural ensemble (dots) are plotted against the sequence of the AUCG tetraloop and neighboring residues.

Cromsigt *et al.*, state that cases of bad agreement between calculated and experimental shifts for certain structural elements can be attributed to either their flexibility or lower definition due to insufficient NMR restraints. Both of these explanations are reasonable for the orientation of the G9 base as due to its flipped-out orientation it is restrained by few NOEs and as stated earlier is probably flexible.

Design of mutants that disrupt Xist 26-mer dimerization

Initially, it was our goal to solve the structure of a complete 26-mer A-repeat. However, we found that a single A-repeat does not fold as predicted into a double hairpin conformation, but its 'hairpin2' is unfolded and mediates dimerization with a second A-repeat *in vitro* (10) (see Fig. 1D and Chapter 2). It is an interesting question if this dimerization could play a role in *Xist* function *in vivo*, as it has been shown that at least 5 repeats are needed for activity and *in vivo* there is a high local concentration of repeats as they are connected by relatively short linkers and are targeted to the X chromosome (7).

We designed 'hairpin 2' mutants to disrupt the dimerization seen *in vitro*. Part of the dimerization platform in 'hairpin 2' is formed by 2 intermolecular AU base pairs as shown in Fig. 1D. The 'U20C' mutant has a single 'point' – mutation of the U that is involved in these intermolecular base pairs to a C, with the rationale to disrupt these base pairs. In the second mutant, 'hairpin 2' is capped by a GUAA tetraloop, a member of the very stable GNRA tetraloop family, which should fold 'hairpin 2' into a stable hairpin (see Fig. 6A). We used sedimentation velocity analytical ultracentrifugation (AUC) to confirm the *in vitro* oligomeric state of the A-repeat 26-mer NMR construct we used in our initial studies and the two mutants (Fig. 6B) at 0.1 mM concentration in NMR buffer. As expected, both mutants were monomeric and we reconfirmed that our Xist 26-mer NMR construct ('wt-NMR') as described in Duszczyk *et al.* (10) is a dimer at 0.1 mM concentration.
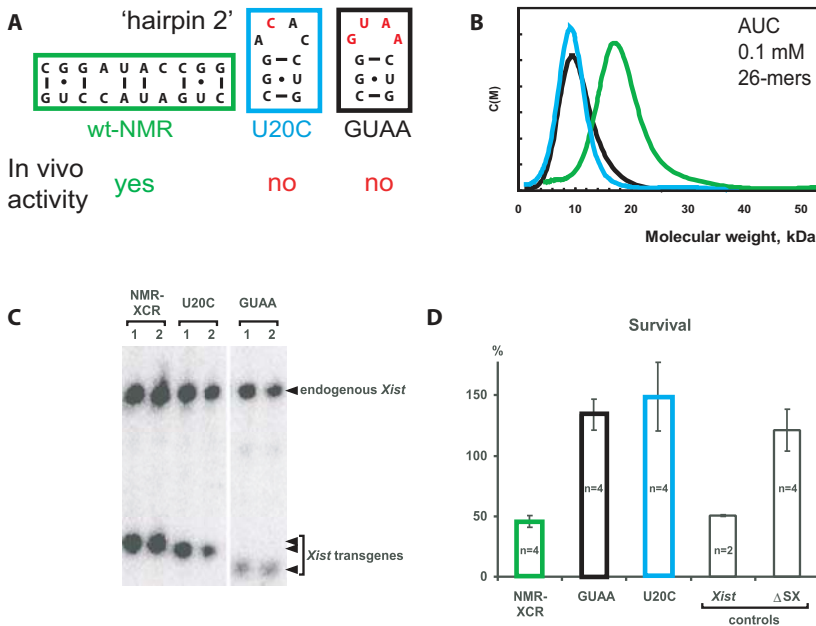
Figure 6 A) We have shown preciously that the 'wt-NMR' A-repeat 26-mer construct dimerizes *in vitro* through 'hairpin 2' (10). We designed 'hairpin' mutants that disrupt this dimerization, U20C and GUAA and confirmed their oligomeric state with analytical ultracentrifugation (B). We tested these mutants *in vivo*: C) Southern blot analysis shows that mutant *Xist* RNA transgenes were successfully introduced into a mouse ES cell system D) The percentage of cells (*n* is the number of measurements in triplicates) surviving in differentiating cultures in the presence of doxycyclin is given as a measure of silencing. Wild type *Xist* and a mutant where the A-repeats are deleted (ΔSX) are used as controls. Cell survival of the U20C and GUAA mutants in cell culture is comparable with the mutant where the A-repeats are deleted.

A-repeat mutants that influence dimerization are inactive in vivo

Next we tested if the A-repeat dimerization mutants are active *in vivo* in the same inducible *Xist* expression system in mouse ES cells that was first used to show that chromosomal association and spreading of Xist RNA can be functionally separated from silencing (7). In this male mouse ES cell system, single-copy mutant *Xist* RNA

transgenes can be introduced into a locus on the X chromosome under control of an inducible promoter. When *Xist* expression is induced by addition of doxycyclin, the activity of the *Xist* transgene can be monitored by measuring cell survival in cell culture, as successful *Xist*-mediated repression of the single male X causes cell death.

Figure 6C shows Southern Blot analysis of the endogenous *Xist* DNA and the *Xist* transgenic DNA, confirming that the transgenes were successfully introduced into the ES cell system. In each case there is a single copy insertion of the *Xist* transgene (lanes 1 and 2 show two independent clones). NMR-XCR contains 12 copies of the A-repeat connected by short linkers of 8 uridines, as does U20C, causing the bands to run at approximately equal height. GUAA does appear as a shorter and weaker band as it has only 7 copies of the A-repeat, joined by the same linkers, and might be lost by diffusion through the membrane. Concerning expression of the transgenes all three constructs show comparably sized *Xist* clusters by RNA FISH (results not shown), which means they are expressed at similar levels.

Figure 6D shows cell survival of differentiating ES cells after 5 days in cell culture. The two controls shown are wild type *Xist* (with 7.5 copies of the A-repeat) and ΔXS in which the A-repeats are deleted, a construct that is known to be unable to induce silencing. The stronger appearance of NMR-XCR silencing compared to wild type *Xist* is probably due to the larger A-repeat copy number, so although NMR-XCR has clear silencing activity, it may have a little less per A-repeat unit. The U20C A-repeat construct clearly shows no activity compared to NMR-XCR. It should be noted that both mutants have an identical 'hairpin 1' sequence as NMR-XCR, but even if a somehow lower per repeat activity of NMR-XCR is assumed, some silencing activity should have been detected for 7 copies of the GUAA construct, as normally 5 wild type A-repeats would be required for efficient silencing (7). We conclude that also GUAA is severely impaired if not entirely inactive. These data are consistent with the idea that dimerization could be required for silencing *in vivo*.

Discussion

Structural comparison to other tetraloops reveals that the AUCG tetraloop is a novel tetraloop motif

We compared the AUCG tetraloop to other known tetraloops to look for fold similarities. Tetraloops are some of the most abundant RNA structural elements and early sequence comparisons on ribosomal RNAs revealed three families as hyperabundant (16) : the UNCG, CUUG and GNRA-type tetraloops. All of them are structurally characterized and Figure 7 shows a schematic comparison between our AUCG tetraloop (Fig. 6A) and 3 members of the other families: UUCG (17) (Fig. 6B), CUUG (18) (Fig. 7C) and GAGA (19) (Fig. 6D). The AUCG tetraloop is distinct from these three abundant tetraloop families in many features: comparison of the base pairing pattern reveals that all these three loops are in fact diloops (a base pair is formed between the first and fourth base within the tetraloop), while the AUCG loop is truly a tetraloop. Other hydrogen bonding patterns are not comparable between the loops. If an A6 H2 to G9 phosphate hydrogen bond really exists within the AUCG tetraloop, this would be another distinct feature. Furthermore, only in the AUCG tetraloop two bases are in the 5' stack, while none is in the 3' stack. In the other tetraloops two or more bases are found in the 3' stack. The position of the turning phosphate, the phosphate where the backbone changes direction, between the second and third residue of the tetraloop is also unique for the AUCG tetraloop: in the other tetraloops it is found between the first and second residue. A fully solvent exposed flipped out purine (G9) is exclusive to the AUCG tetraloop. Therefore we conclude that it represents a novel RNA tetraloop motif.

Correlation of the AUCG tetraloop structure with functional data

In previous studies of *Xist* function several A-repeat mutants have been tested in the earlier described *in vivo* system (7). Several sequence-function relationships emerged from these studies. It was shown that the length and sequence of the spacer by which the A-repeats are joined do not influence *Xist* activity and that at least 5 A-repeats are required. Further it was shown that the sequence of the hairpin 1 stem is not crucial as long as base pairing is not disrupted. This is fully consistent with our 3D structure that shows a canonical Watson-Crick base paired stem as expected. One of the mutants described where a scrambled hairpin 1 loop with the sequence UAGC still is able to induce silencing is puzzling in relation to the results presented here, where the fully conserved AUCG hairpin loop folds into a well defined tetraloop structure. A possibility is that in *Xist* function not the specific sequence of the AUCG tetraloop is recognized, but its global fold and that the UAGC and AUCG tetraloops fold into similar overall structures, which should be investigated. The sequence must be somehow important though, as in the same study, repeats with a fully antisense sequence do not induce silencing and mutants with the hairpin 1 tetraloop extended to a hexaloop, or a hairpin 1 capped by a GNRA loop were also non-functional (A. Wutz, unpublished data). The question then still remains why the AUCG sequence is fully conserved if not completely specific.

With the AUCG tetraloop structure known, more rational hairpin 1 mutants can be designed to address this question, for example by targeting the flipped out G9 that in our structure is solvent exposed and freely available for intermolecular interactions. To further investigate specificity of the AUCG tetraloop, interaction studies with the recently proposed A-repeat binding partners should be helpful in investigating specific recognition.
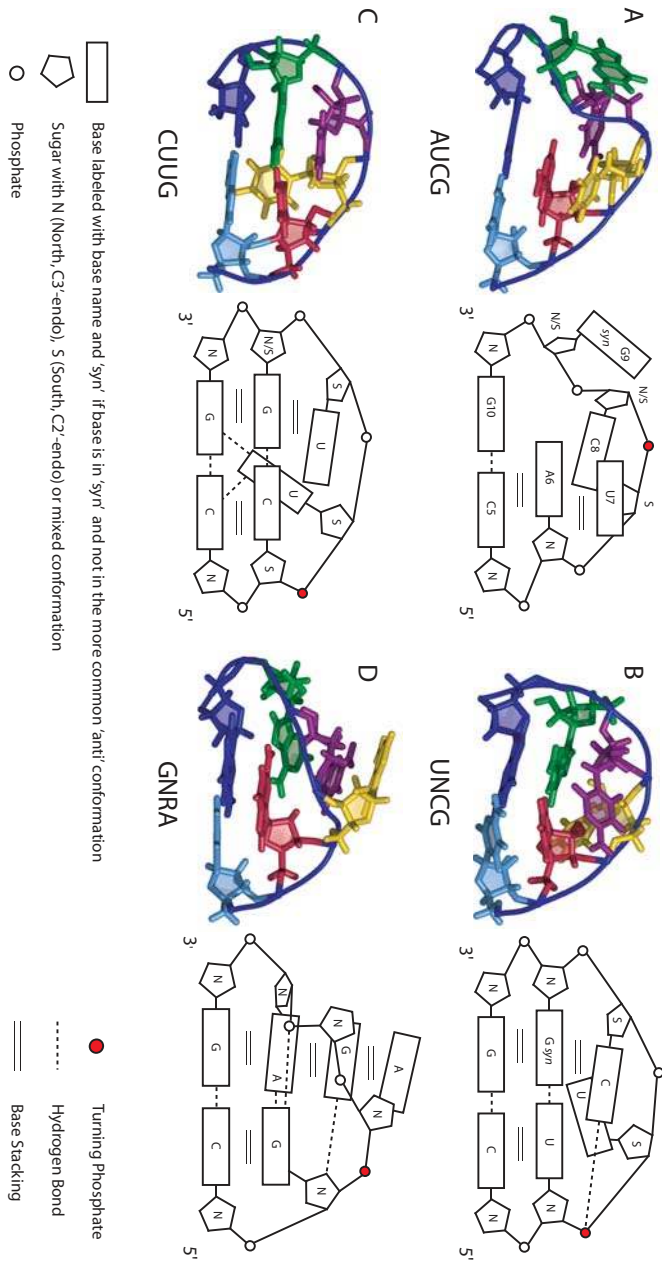
Figure 7 Comparison of the structures of the A-repeat AUCG tetraloop (A) with members of the UNCG (B), CUUG (C) and GNRA (D) tetraloop families (N is any nucleotide, R is a purine) shows that the AUCG fold is a novel tertraloop motif.
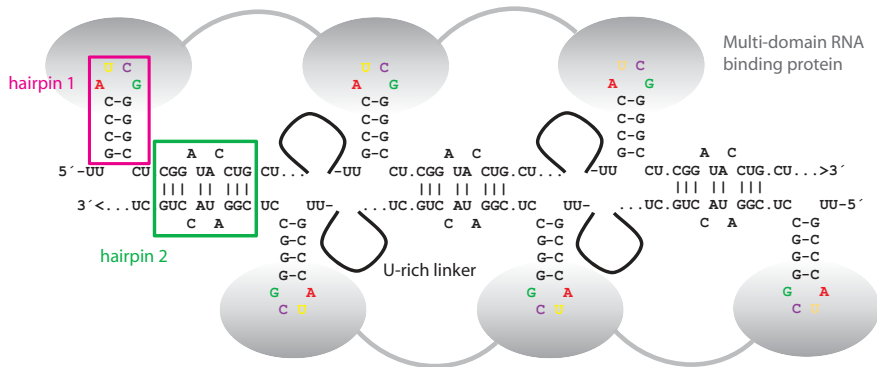
Figure 8 A model for the Xist RNA A-repeats: dimerization of the 'hairpin 2' sequence brings together the AUCG tetraloops of different A-repeats, either within the same, or between different *Xist* molecules. The AUCG tetraloops are recognized by a multidomain RNA-binding protein partner.

In view of the competition between the A-repeats and *Tsix* in PRC2 binding proposed by Zhao *et al.*, (8) it would also be interesting to look into the structure of the antisense GCUA hairpin that should be found in *Tsix* and test if repeats involving this hairpin are the competitors in this binding event.

Functional implications for 'hairpin 2' dimerization: a model of the Xist RNA A-repeats

The earlier studies on mutant A-repeats did not elucidate any sequence-function relationships within 'hairpin 2', apart from that its complete removal abolished *Xist* function. The results presented here are consistent with the idea that dimerization through this sequence could be required for silencing *in vivo*. A way to strengthen this hypothesis would be to find 'hairpin 2' dimerizers with an unrelated sequence that are active in our *in vivo* assay, but this might prove difficult as it is hard to imagine that the specific sequence of 'hairpin 2' is not important at all as it is also

fully conserved. On one hand it is possible that the dimerizing AUAC sequence has a different specific functional structure within some 'hairpin 2' conformation that we do not see *in vitro* in our NMR studies, either because of the relatively high concentrations, or maybe because an interaction partner involved in this fold is missing. Here again, interaction studies with the recently proposed binding partners could show if structural rearrangement of the A-repeat occurs upon protein binding. On the other hand, if this is true, this would not rule out that at some other time dimerization through the second part of the A-repeat is important. Interaction between different A-repeats is a reasonable hypothesis as a minimum of 5 is required for *Xist* function and there is a high local concentration of A-repeats *in vivo* as they are connected by relatively short linkers and are localized to the Xi.

In view of this it could be considered that multimerization of several A-repeats, either within a single Xist RNA molecule, or between different ones, brings the AUCG tetraloops close together as shown in Fig. 8. This alone, or possibly in combination with a multi-domain RNA binding protein partner specifically recognizing the AUCG tetraloops, could prevent the formation of duplexes with *Tsix* as described by Ogawa *et al.* (6) and could thus prevent short RNA formation and the RNAi like *Xist* repression on the Xi, allowing Xist RNA to accumulate and to initiate silencing. Of course, to validate this model the hypothetical protein partner should be first identified.

Another possibility is that by multimerizing in this way the A-repeats themselves serve as a nucleation center by which silencing is spread over the Xi, possibly mediated by the PRC2 complex that could directly interact with the AUCG tetraloops. Further experimental work, both *in vivo* as well as structural is needed to validate these models. As the dimerization model has been derived based on structural work on a single A-repeat, the next step for the structural work will be to study the 3D topology of multiple A-repeats, to see if they behave in a similar way, alone, as well in complex with the proposed binding partners.

Conclusion

We have solved the NMR structure of a novel AUCG tetraloop motif within the Xist RNA A-repeat that is essential for initiation of X-Inactivation. Our studies show that the second part of the A-repeat is involved in multimerization *in vitro* that could also be required for X-inactivation *in vivo*. Further experiments are needed to address the structure-function relationship of the AUCG tetraloop in X-inactivation and validation of the multimerization hypothesis. Both will require and should benefit from interaction studies with recently proposed binding partners.

Materials and Methods

Sample Preparation

[13]C,[15]N uniformly labeled and unlabeled r(GGCGC[AUCG]GCGCC) A-repeat 14-mer RNA, containing the A-repeat AUCG tetraloop 'hairpin 1', was prepared as described previously (10) (see Chapter 2). For AUC measurements of the oligomeric state of the full A-repeat, unlabeled r(GGCGC [AUCG] GCGCUUCGG [AUAC] CUGC) 'wt-NMR', r(GGCGC [AUCG] GCGCUUCGG [ACAC] CUGC) 'U20C', and r(GGCGC [AUCG] GCGCUUCGG [GUAA] CUGC) 'GUAA' 26-mer RNA was prepared in the same way.

NMR samples were prepared in 10 mM NaH_2PO_4/Na_2HPO_4, pH 6.0, 100 mM NaCl, 0.02 mM EDTA, 0.02% Azide in 95% H_2O, 5% D_2O or 100% D_2O. RNA concentrations ranged between 0.8 and 1.2 mM. AUC samples were prepared in NMR buffer at 0.1 mM concentration. Prior to all measurements samples were

heated to 95° for 5 minutes followed by snap-cooling on ice with the rationale to trap the kinetically favored intramolecular monomeric hairpin conformation over a possible intermolecular dimer. The homogenous formation of a monomeric 14-mer hairpin was unambiguously confirmed by measuring the relative intensity of cross- and diagonal peaks in a HNN-COSY spectrum (11) recorded on a 50% ^{13}C,^{15}N - labeled, 50% unlabeled sample in H$_2$O (10). Partial alignment of RNA for residual dipolar coupling measurements was achieved by adding 18 mg ml^{-1} filamentous bacteriophage Pf1 (ASLA Biotech, Riga, Latvia) to an unlabeled sample.

NMR Spectroscopy and generation of structural restraints

NMR experiments were recorded at 5°C (assignment and NOESY spectra involving exchangeable protons) or 25°C (assignment, NOESY spectra involving non-exchangeable protons, measurements of J-couplings and RDCs) on Bruker DRX600, 800 or 900 spectrometers equipped with cryoprobes. Spectra were processed with NMRPipe (20) and analyzed using NMRVIEW (21). Chemical shifts in the Xist RNA 14-mer AUCG hairpin were assigned as described previously (Duszczyk *et al.,* to be submitted to Biomolecular NMR assignments – see Chapter 3).

Inter-proton distance restraints

NOEs were manually assigned. Distance restraints from non-exchangeable protons were derived from a 3D NOESY-HMQC experiment with 80 ms mixing time measured in D$_2$O. Distance restraints between exchangeable protons were extracted from 2D NOESY (mixing time 150 ms) in 90% H2O, 10% D2O. From this spectrum only imino-imino and imino-amino distance restraints were extracted and subsequently used in the structure calculations. Integration of NOE volumes, calibration of distances (using a relaxation matrix spin diffusion correction) and the setting of the distance restraint lower and upper bounds was performed by

ARIA1.2 (12). The automated NOE assignment features of ARIA were not used as due to the limited number of available NOE based distance restraints in RNA the introduction of an assignment error could have a high impact on the overall fold of the molecule. Also, wrong assignments within highly overlapped spectral regions can lead to miscalibration of distances. In manual NOE assignment symmetrical peaks were carefully inspected and only fully assigned NOEs were used for the distance calibration.

Torsion angle restraints

Experimental endocyclic torsion angle restraints for the sugar puckers were derived from homonuclear $^3J_{H1'H2'}$ and $^3J_{H2'H3'}$ couplings measured in 3D HCCH-E.COSY (22, 23) and 3D forward-directed HCC-TOCSY-CCH-E-COSY (24) experiments. Sugar puckers were defined as C2'-endo if $^3J_{H1'H2'} > 7$ (Hz) and $^3J_{H2'H3'} < 2$ (Hz) or C3'-endo if the opposite was true. In the case of C3'-endo (U7) the torsion angle δ was restrained to $140° \pm 10°$, $\phi_{1',2'}$ to $160° \pm 10°$ and $\phi_{2',3'}$ to $-25° \pm 10°$. For intermediate couplings (C8, G9 and G10) these angles were left unrestrained. In the case of C2'-endo (all other residues) the torsion angle δ was restrained to $80° \pm 10°$, $\phi_{1',2'}$ to $99° \pm 10°$, and $\phi_{2',3'}$ to $38° \pm 10°$, respectively.

The backbone angle γ and stereospecific assignments for H5' and H5" were obtained based on $^3J_{H4'H5'/H5''}$ and $^3J_{C4'H5'/H5''}$ coupling constants (22). Due to overlap these could only be measured for the loop residues A6 to G10. As all measured $^3J_{H4'H5'}$ and $^3J_{H4'H5''}$ couplings were < 5 Hz, γ was restricted to the *gauche*$^+$ conformation ($60° \pm 30°$), but left unrestrained for the suspected mobile residues G9 and G10.

The backbone angle ε was restrained based on $^3J_{C2'P}$ coupling constants extracted from the relative difference in signal volumes measured in a 2D spin echo difference CT-HSQC experiment (25) and qualitative analysis of C4'P cross-peak intensities in a HCP experiment (26). Due to overlap, experimental restraints for ε could only be obtained for C5, U7-G10 and C14. For C5 and C14 ε was defined as *trans* ($210° \pm 30°$) as $^3J_{C2'P} < 5$ (Hz) and the C4'P cross peak was visible in the HCP

spectrum. The angle ε for U7 was restrained to *gauche⁻* (260° ± 40°) as $^3J_{C2'P} > 5$ (Hz) and the C4'P cross peak was almost invisible in the HCP spectrum. As the $^3J_{C2'P}$ couplings were intermediate for C8 and G9 and the C4'P cross peak for C10 was invisible the angle ε was left unrestrained for these residues.

The glycosidic χ angles for C5-C8 and G10 were loosely restrained to the *anti* (190° ± 100°) and to the *syn* conformation (20° ± 100°) for G9 consistent with the NOE data.

As both NOESY sequential walk, HNN-COSY and dihedral angle data for residues 1-5 and 10-14 are in agreement with a canonical A-helical conformation, backbone torsion angle restraints for 1-4 and 11-14 were set to standard values (300/180/50/210/290° ± 10° for α/β/γ/ε/ζ respectively). Only α/β/γ were restrained non-experimentally in 5 and only ζ in 10. Additionally for 1-4 and 11-14 χ angles were restrained to 200 ± 20°. The tight A-form restraints were necessary to achieve a better convergence for the stem conformation due to an insufficient amount of interresidual NOEs because of high overlap for the resonances within the stem. For residues 5-10 only NMR-derived torsion angle restraints were used, only if the data did not indicate conformational averaging as described above. Measured J-coupling values and restraints used for residues 5-10 are summarized in Table 1A. Other measured J-couplings are reported elsewhere (Duszczyk *et al.,* to be submitted to Biomolecular NMR assignments – see Chapter 3).

Residual Dipolar Couplings

Splittings for the determination of one-bond ^{1}H-^{13}C couplings were measured as the distance between the maxima of only well resolved peaks of doublets in natural abundance (^{13}C–^{1}H) TROSY spectra (27). RDCs were calculated as the difference between splittings in isotropic and partially aligned sample (Pf1) and are reported elsewhere (Duszczyk *et al.,* to be submitted to Biomolecular NMR assignments – see Chapter 3). RDCs from G1, C8, G9 and C14 were excluded from the

refinement due to potential dynamics for these parts of the molecule and 6H2, 3H6, 10H4' due to line broadening in the spectra.

Other restraints

Hydrogen-bond distance restraints were used for the four Watson–Crick pairs detected in the HNN-COSY experiment (11) and G1-C14, assuming that it was not visible in this experiment due to end-fraying.

Structure Calculation and refinement

As a first step, 100 initial structures were calculated with ARIA/CNS 1.2 (12) with a mixed Cartesian and torsion angle dynamics simulated annealing protocol from an extended starting structure using NOE, torsion angle, hydrogen bonding and planarity restraints.

The standard CNS dna-rna-allatom topology and energy parameter files were used with uniform energy constants for all bond, angle, and improper dihedral energy terms of the force field. The simulated annealing protocol in CNS consisted of four stages: a high-temperature torsion angle simulated annealing phase, (50000 steps at 10000 K with a time step of 27 fs), a first torsion angle dynamics cooling stage from 10,000 K to 2,000 K (10000 steps), a Cartesian dynamics cooling phase 1 from 2,000 K to 1,000 K (50000 steps) and a Cartesian dynamics cooling phase 2 from 1,000 K to 50 K (20000 steps) with a time step of 3 fs. Energy constants for distance restraints and H-bonds were 10 kcal mol^{-1}Å^{-2} during the high temperature dynamics phase, were ramped up to 50 kcal mol^{-1} during cooling phase 1 and were kept at this level until the end of cooling phase 2. Energy constants for dihedral restraints were 5, 25 and 200 kcal mol^{-1} rad^{-2} during the high temperature dynamics and the cooling phases 1 and 2 respectively.

Weak planarity restraints (25 kcal mol^{-1}Å^{-2}) were applied for the stem Watson-Crick base pairs during cooling phase 2. Planarity within a base pair was

defined for a plane that involves one atom of the H-bond acceptor and four atoms of the donor base to allow for propeller twist and tilt. Structures were calculated in the default number of nine iterations, although the option to automatically assign NOEs and to iteratively detect and remove inconsistent restraints based on these automatically assigned NOEs through these nine iterations was not used. Hydrogen bond, planarity, torsion angle, and distance restraints were simultaneously applied during all nine iterations. Structures were checked for close proton–proton distances that are inconsistent with experimental NOEs. Based on this and on low energy criteria, 20 best structures were selected for refinement with the SANDER module of AMBER 9 (University of California, San Francisco, http://ambermd.org). The AMBER force field was chosen for refinement because of its treatment of electrostatics that results in superior simulation of base stacking and hydrogen bonding (28, 29). In our experience this better performance is e.g. illustrated by the fact that in AMBER, as opposed to ARIA/CNS no planarity restraints for the base pairs are necessary to correctly shape the A-form helical stem.

Before refinement, ARIA structures were energy minimized. To use RDCs in AMBER, initial estimates of the alignment tensors were obtained from the preliminary ARIA/CNS structures as described (30). To prevent high violations of local geometries while accommodating RDC restraints, additional angle restraints to maintain proper local geometries were employed. 25 RDCs were included in the refinement with a single floating alignment tensor.

The restrained MD refinement was performed with the Cornell *et al.* (1995) force field (31) with the generalized-Born solvation model (32). Identical hydrogen bond, distance, experimental and artificial torsion angle, and NOE-derived distance restraints were employed as in the CNS/ARIA calculations. The SA protocol involved a 20 ps restrained molecular dynamics run, with heating from 0 to 500 K during the first 5 ps, followed by a first cooling step to 100 K (13 ps) and a final cooling step to 0 K (2 ps) as described by Padrta *et al.* (29). Square-well penalty functions with force constants as described by Stefl *et al.* (33) were used for all experimental restraints with the difference that force constants for RDC

restraints were ramped up from 0.1 kcal mol^{-1}Hz^{-2} to 1 kcal mol^{-1}Hz^{-2}. The final selection step on the 20 AMBER-refined structures was composed of back calculation of proton chemical shifts with the program NUCHEMICS (15) for the loop nucleotides and selection of 10 structures that agreed best with chemical shift values while maintaining low restraint violations and force field energies.

RMSDs and angles in the final structures were calculated using MOLMOL (34). Molecular graphics was generated using Pymol (35).

In vivo studies

Cloning of mutant *Xist* constructs, generation of transgenic cell lines and growing of ES cell cultures was performed as described previously (7).

Deposition of coordinates, chemical shifts and restraints

Coordinates for the ensemble of 10 AMBER-refined structures of the 14-mer AUCG tetraloop hairpin will be deposited at the Protein Data Bank. NMR chemical shift assignments and other restraints used to calculate the structural ensemble will be deposited into the BMRB.

Acknowledgements

help with the RNA sample preparation, Vladimir Rybin for recording the AUC experiments and Elena Conti, Asifa Akhtar and Sybren Wijmenga for useful discussions. We acknowledge the Bayerisches NMR Zentrum Garching (München), BNMRZ Frankfurt and Wolfgang Bermel (Bruker, Karlsruhe) for spectrometer time.

Funding

Supplementary Figure

Supplementary Figure 1 Overlay of (^{1}H,^{1}H) 2D NOESY spectra in 100% D_2O of the Xist RNA A-repeat 14-mer AUCG tetraloop (in black) and the 26-mer full A-repeat (in grey) recorded at 900 MHz, showing the aromatic (H6/H8) to anomeric (H1' and some H2') region. H6/H8-H1' intra- and interresidual and intra-base H6-H5 cross peak assignments for the AUCG tetraloop hairpin are shown as in the inserted secondary structure representation. Solid and dotted lines represent the 'sequential walk' in the stem of hairpin 1. Excellent overlay of the peaks involved in the AUCG tetraloop between the two constructs show that the AUCG tetraloop has the same conformation in the 14-mer hairpin as in the full 26-mer A-repeat.

References

[1]   Lucchesi, J.C., Kelly, W.G. and Panning, B. (2005) Chromatin remodeling in dosage compensation. *Annual Review of Genetics*, **39**, 615-651.

[2]   Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S. and Brockdorff, N. (1996) Requirement for Xist in X chromosome inactivation. *Nature*, **379**, 131-137.

[3]   Sheardown, S.A., Duthie, S.M., Johnston, C.M., Newall, A.E., Formstone, E.J., Arkell, R.M., Nesterova, T.B., Alghisi, G.C., Rastan, S. and Brockdorff, N. (1997) Stabilization of Xist RNA mediates initiation of X chromosome inactivation. *Cell*, **91**, 99-107.

[4]   Avner, P. and Heard, E. (2001) X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet*, **2**, 59.

[5]   Lee, J.T. and Lu, N. (1999) Targeted mutagenesis of Tsix leads to nonrandom X Inactivation. *Cell*, **99**, 47.

[6]   Ogawa, Y., Sun, B.K. and Lee, J.T. (2008) Intersection of the RNA Interference and X-Inactivation Pathways. *Science*, **320**, 1336-1341.

[7]   Wutz, A., Rasmussen, T.P. and Jaenisch, R. (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet*, **30**, 167.

[8]   Zhao, J., Sun, B.K., Erwin, J.A., Song, J.-J. and Lee, J.T. (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, **322**, 750-756.

[9]   Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A. and Panning, B. (2002) Xist RNA and the mechanism of X-Chromosome Inactivation. *Annual Review of Genetics*, **36**, 233-278.

[10] Duszczyk, M.M., Zanier, K. and Sattler, M. (2008) A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat. *Nucl. Acids Res.*, **36,** 7068-7077.

[11] Dingley, A.J. and Grzesiek, S. (1998) Direct observation of hydrogen bonds in nucleic acid base pairs by internucleotide $^2J_{NN}$ couplings. *J. Am. Chem. Soc.*, **120**, 8293-8297.

[12] Linge, J.P., Habeck, M., Rieping, W. and Nilges, M. (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics*, **19**, 315-316.

[13] Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S. *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. Sect. D-Biol. Crystallogr.*, **54**, 905-921.

[14] Cromsigt, J., Hilbers, C.W. and Wijmenga, S.S. (2001) Prediction of proton chemical shifts in RNA - their use in structure refinement and validation. *J. Biomol. NMR*, **21**, 11-29.

[15] Wijmenga, S.S., Kruithof, M. and Hilbers, C.W. (1997) Analysis of H-1 chemical shifts in DNA: Assessment of the reliability of H-1 chemical shift calculations for use in structure refinement. *J. Biomol. NMR*, **10**, 337-350.

[16] Woese, C.R., Winker, S. and Gutell, R.R. (1990) Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 8467-8471.

[17] Allain, F.H.T. and Varani, G. (1995) Structure of The P1 helix from group-I self-splicing introns. *J Mol Biol*, **250**, 333-353.

[18] Jucker, F.M. and Pardi, A. (1995) Solution structure of the CUUG hairpin loop - a novel RNA tetraloop motif. *Biochemistry*, **34**, 14416-14427.

[19] Jucker, F.M., Heus, H.A., Yip, P.F., Moors, E.H.M. and Pardi, A. (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.*, **264**, 968-980.

[20] Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) NMRpipe - A multidimensional spectral processing system based on Unix pipes. *Journal of Biomolecular NMR*, **6**, 277-293.

[21] Johnson, B.A. and Blevins, R.A. (1994) NMRView - a computer program for the visualization and analysis of NMR data. *J. Biomol. NMR*, **4**, 603-614.

[22] Marino, J.P., Schwalbe, H., Glaser, S.J. and Griesinger, C. (1996) Determination of gamma and stereospecific assignment of H5' protons by measurement of (2)J and (3)J coupling constants in uniformly C-13 labeled RNA. *Journal of the American Chemical Society*, **118**, 4388-4395.

[23]   Zimmer, D.P., Marino, J.P. and Griesinger, C. (1996) Determination of homo- and heteronuclear coupling constants in uniformly C-13, N-15-labeled DNA oligonucleotides. *Magnetic Resonance in Chemistry*, **34**, S177-S186.

[24]   Schwalbe, H., Marino, J.P., Glaser, S.J. and Griesinger, C. (1995) Measurement of H,H-coupling constants associated with nu-1, nu-2, And nu-3 in uniformly C-13 labeled RNA by HCC-TOCSY-CCH-E.COSY. *Journal Of The American Chemical Society*, **117**, 7251-7252.

[25]   Legault, P., Jucker, F.M. and Pardi, A. (1995) Improved measurement of ^{13}C, [31]P J-coupling constants in isotopically labeled RNA. *FEBS Letters*, **362**, 156.

[26]   Marino, J.P., Schwalbe, H., Anklin, C., Bermel, W., Crothers, D.M. and Griesinger, C. (1994) Three-dimensional triple-resonance ^{1}H, ^{13}C, ^{31}P experiment: sequential through-bond correlation of ribose protons and intervening phosphorus along the RNA oligonucleotide backbone. *J. Am. Chem. Soc.*, **116**, 6472-6473.

[27]   Meissner, A. and Sorensen, O.W. (1999) The role of coherence transfer efficiency in design of TROSY-type multidimensional NMR experiments. *Journal of Magnetic Resonance*, **139**, 439.

[28]   Cheatham, T.E. and Kollman, P.A. (2000) Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.*, **51**, 435-471.

[29]   Padrta, P., Stefl, R., Kralik, L., Zidek, L. and Sklenar, V. (2002) Refinement of d(GCGAAGC) hairpin structure using one- and two-bond residual dipolar couplings. *J. Biomol. NMR*, **24**, 1-14.

[30]   Tsui, V., Zhu, L.M., Huang, T.H., Wright, P.E. and Case, D.A. (2000) Assessment of zinc finger orientations by residual dipolar coupling constants. *J. Biomol. NMR*, **16**, 9-21.

[31]   Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P.A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, **117**, 5179-5197.

[32]   Bashford, D. and Case, D.A. (2000) Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, **51**, 129-152.

[33]   Stefl, R. and Allain, F.H.T. (2005) A novel RNA pentaloop fold involved in targeting ADAR2. *RNA*, **11**, 592-597.

[34]   Koradi R, B.M., Wuethrich K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graphics*, **14**, 51-55.

[35]   DeLano, W.L. (2002). DeLano Scientific, San Carlos, CA, USA.

Summary
Samenvatting

Summary

Xist RNA is a large non-coding RNA that is required for the initiation of X-chromosome inactivation in female mammals. This is a mechanism necessary to compensate for the difference in X-linked gene expression between XY males and XX females. X-inactivation has been studied extensively as it is a fascinating paradigm for epigenetic regulation of gene expression involving non-coding RNA. This thesis presents the first structural investigation into the molecular mechanism by which X-inactivation takes place.

First, RNA-mediated regulation of gene expression in general, and Xist RNA mediated X-inactivation specifically, are introduced. Basic building blocks of RNA structure are described, and RNA-specific methods for structural studies using nuclear magnetic resonance (NMR) are summarized.

Next, the *in vitro* folding topology of the A-repeats, conserved repeats of a 26-nucleotide motif within Xist RNA, essential for its silencing function, is presented. Using NMR it is shown that only half of a single A-repeat folds into a stable hairpin, while the other half is unfolded and mediates dimerization of the A-repeats by duplex formation with a second A-repeat. A novel NMR-based method that was necessary to characterize this peculiar folding topology is described. This novel method is generally applicable in structural studies of other RNAs.

A detailed description of the NMR strategy to assign chemical shifts for the stable hairpin within the A-repeat is given, before finally, the NMR structure of this hairpin is presented. It reveals a novel well defined AUCG tetraloop conformation. Additionally it is shown that the multimerization mediated by the second part of the A-repeat *in vitro*, could possibly be required for X-inactivation *in*

vivo, as mutants that disrupt this dimerization are unable to initiate silencing in an *in vivo* assay.

With these novel molecular insights, a model for Xist RNA function is proposed where dimerization of the A-repeats and specific recognition of the AUCG tetraloops therein, work together to achieve X-inactivation.

Samenvatting

Xist RNA is een groot, niet-coderend RNA dat nodig is voor de initiatie van X-chromosoom inactivatie in vrouwelijke zoogdieren. Dit mechanisme is nodig om het verschil in genexpressie van X-gebonden genen tussen XY mannen en XX vrouwen te compenseren. X-inactivatie is uitvoerig onderzocht omdat het een fascinerend paradigma is voor epigenetische genregulering waarbij niet-coderend RNA betrokken is. Dit proefschrift beschrijft het eerste structuuronderzoek naar het moleculaire mechanisme waarmee X-inactivatie plaatsvindt.

Eerst wordt genregulering door RNA in het algemeen, en X-inactivatie door Xist RNA in het bijzonder, ingeleid. De structurele bouwstenen van RNA worden beschreven, en methoden voor structuuronderzoek met kernspinresonantie (NMR) die specifiek zijn voor RNA worden kort samengevat.

Vervolgens wordt de topologie gepresenteerd met welke de A-repeats, geconserveerde repeats van een 26-nucleotide motief in Xist RNA, die essentieel zijn voor stillegging van het X-chromosoom, *in vitro* gevouwen zijn. Met behulp van NMR wordt getoond dat alleen de helft van een A-repeat gevouwen is in een stabiele hairpin, terwijl de tweede helft ongevouwen is en zorgt voor dimerisatie van de A-repeats door een duplex te vormen met een tweede A-repeat. Een nieuwe methode, gebaseerd op NMR, die nodig was om deze eigenaardige topologie te bepalen, wordt beschreven. Deze methode kan in het algemeen toegepast worden bij structuuronderzoek aan andere RNAs.

Een gedetailleerde beschrijving wordt gegeven van de NMR strategie die is toegepast om de chemical shifts van het stabiele hairpin in de A-repeats toe te kennen, voordat uiteindelijk de NMR structuur van deze hairpin wordt gepresenteerd. Deze structuur toont een nieuwe, goed gedefinieerde AUCG

tetraloop. Verder wordt getoond dat de multimerisatie door het tweede gedeelte van het A-repeat *in vitro*, ook nodig zou kunnen zijn voor X-inactivatie *in vivo*, omdat mutanten die deze dimerisatie verbreken geen X-inactivatie kunnen inleiden in een *in vivo* assay.

Met behulp van deze nieuwe moleculaire inzichten wordt een model gepresenteerd voor de functie van Xist RNA, waarbij dimerisatie van de A-repeats en specifieke herkenning van de AUCG tetraloops samenwerken om X-inactivatie te bewerkstelligen.

Curriculum Vitae

De auteur van dit proefschrift werd geboren op 26 mei 1978 in Warschau. Op vijfjarige leeftijd verhuisde ze met haar ouders naar Delft, waar zij het gymnasium aan het Stanislascollege doorliep. Daar viel ze, mede dankzij fantastische leraren, als een blok voor de vakken geschiedenis en scheikunde. Ook stortte zij zich op het schooltoneel, waar ze haar talent en passie voor theater ontdekte. Van 1996 tot 2003 studeerde ze, onderbroken door een jaar aan de Toneelacademie Maastricht, Scheikunde en Wijsbegeerte aan de Universiteit Leiden. Haar interesse voor de biomoleculaire NMR werd gewekt tijdens doctoraalcolleges gegeven door Marcellus Ubbink en Gregg Siegal, onder wiens begeleiding ze ook haar eerste afstudeeronderzoek deed in de Metalloproteinen groep van Prof. Gerard Canters. Ze werkte daar aan de eiwitexpressie van BRCT domeinen voor structuuronderzoek. Ondersteund door o.a. het LUF Internationaal Studie Fonds en the British Council, verbleef ze voor een tweede onderzoeksstage zeven maanden in de groep van Prof. Iain Campbell aan de universiteit van Oxford, waar ze onder begeleiding van Jörn Werner onderzoek deed naar de orientatie van domeinen binnen het eiwit fibronectine met behulp van Residual Dipolar Couplings en NMR relaxatie. Ze behaalde het doctoraalexamen scheikunde met veel genoegen in december 2002. Haar promotieonderzoek, beschreven in dit proefschrift, voerde zij tussen december 2003 en december 2008 uit onder begeleiding van Michael Sattler in Duitsland, grotendeels aan het European Molecular Biology Laboratory (EMBL) in Heidelberg en vervolgens aan het Helmholtz Zentrum en de Technische Universität in München. Ze presenteerde haar werk op verschillende congressen, waaronder een lezing op de jaarlijkse meeting van de RNA society in 2008 in Berlijn. Tijdens haar promotie was ze twee jaar gekozen studentlid van het bestuur van het EMBL international PhD programme en bovendien actief binnen de Science & Society commissie van het EMBL.

The author of this thesis was born on the 26th of May 1978 in Warsaw. At the age of 5 she moved with her parents to Delft in the Netherlands where she attended gymnasium at the Stanislascollege. Thanks to great teachers, her favorite subjects were history and chemistry. She also discovered her talent and passion for the theatre in the school drama club. From 1996 to 2003 she studied Chemistry and Philosophy at the University of Leiden, with a year break to attend Maastricht Theatre Academy. Her interest for biomolecular NMR was sparked during Master's level lectures taught by Marcellus Ubbink and Gregg Siegal, who also supervised her during her first Master's research project in the Metalloprotein group of Prof. Gerard Canters. There she worked on protein expression of BRCT domains for structural studies. With financial support from, amongst others, the Leiden University International Study Fund and the British Council, she spent seven months at the University of Oxford to do a second Master's research project, working in the lab of Prof. Iain Campbell on the orientation of domains within the protein fibronectin using Residual Dipolar Couplings and NMR relaxation under supervision of Jörn Werner. She got her Master's degree in Chemistry with distinction in December 2002. She did her PhD research, described in this thesis, between December 2003 and December 2008 under the supervision of Michael Sattler, in Germany, for the most part at the European Molecular Biology Laboratory (EMBL) in Heidelberg and later at the Helmholtz Zentrum and Technische Universität München. She presented her work at several meetings, including an oral presentation at the yearly meeting of the RNA society in 2008 in Berlin. During her PhD she was elected student representative in the EMBL international PhD programme graduate committee for two years. She was also active within the EMBL Science & Society committee.

List of Publications

Duszczyk, M.M., Zanier, K. and Sattler, M. (2008) A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a Xist RNA A-repeat. *Nucl. Acids Res.*, **36,** 7068-7077.

Duszczyk, M.M., Wutz, A and Sattler, M. The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization. *Manuscript in preparation.*

Duszczyk, M.M. and Sattler, M. NMR assignment of a novel AUCG tetraloop hairpin from a human Xist RNA A-repeat essential for X-inactivation. *Manuscript in preparation.*

Bermel, W., Simon, B., Duszczyk, M.M. and Sattler, M. A novel H5C5C4N3N1 NMR experiment correlating pyrimidine H5 with N1 across Watson-Crick base pairs using adiabatic refocusing. *Manuscript in preparation*

Acknowledgements

These last few pages are possibly the most important as this thesis could not have been completed without the help and support of many people. They will certainly be the most widely read pages of this publication!

Michael, you undoubtedly had the greatest influence on this piece of work. Thank you for the opportunity to embark on the Xist project, for your enthusiasm for science and for your style of mentoring. In addition to all you taught me about NMR, being offered such a degree of freedom and independence, mixed with your continuing encouragement and interest, was with the benefit of hindsight exactly what I needed to gain the self-confidence without doubt necessary in my further career. Not in the last place, thanks for giving me the opportunity to do my PhD at EMBL which has been the best environment for PhD studies I could wish, not only scientifically, but also, maybe even more important, socially (more on that later).

Thank you Sybren, for agreeing to be my promotor, giving the opportunity to defend my thesis 'at home' in Holland, making things a lot easier for friends & family to join in on the celebrations. I'm certainly looking forward for them to finally find out what on earth it was that I have been doing that made me virtually disappear from their lives for the last 5 years! In addition, thank you (and Elena Conti and Asifa Akhtar) for taking the time to discuss my PhD work during the yearly TAC meetings. I'm also grateful to Asifa, Marco Tessari and Arno Kentgens for evaluating my thesis.

The initiation of the Xist project and a great part of the last chapter of this thesis was the result of a fruitful collaboration: thanks Anton, for such an interesting project and for your wonderful *in vivo* system. I can't believe that finding the A-repeats kissing in the NMR tube, which was a small tragedy at the time, actually turned out to be something potentially very interesting!

Now, let's go on to EMBL: Anne, Matthias, Lars, Milanka & Tizi – thanks for running the EMBL PhD programme. I've learned a lot from the years working

with you in the graduate committee. Thanks Halldor, for all the Science & Society events, which have been essential in keeping my right brain (at least partly) alive, and have been an amazing opportunity to meet and discuss with so many writers and philosophers: I certainly never thought that choosing a PhD in science would give me the opportunity to meet and even have dinner with Dan Dennett!

Thank you, Vladimir, for recording the AUC experiments. Rainer, Frank & Chris (in reversed chronological order): thanks for all the hardware & software support (and Chris for all the smoking breaks!). Tamara & Waltraud: thank you for the best secretarial support. To the Carlomagno group: thanks for adopting me during my six orphan months after the Sattler group left for Munich!

Now we've come to you, past & present members of the Sattler group: I still haven't figured out what it is exactly that Michael did to put together a group of people that get along so well! Even in times of serious Xist-ential problems, you were a reason to look forward to come up the hill (or up to Garching). Katia – you were the guinea pig in the group for tackling RNA with NMR – without your well documented work I would have had to go through so much more trouble. Robi (I hope to get to celebrate my birthday in Napoli again someday), Ana, André, Phinny, Lorenzo, Erika, Koen & Divita – thanks for sharing an office with me! (the reader might think that is a hell of a lot people in one office but they have to consider that I've spent a loooooong time in the Sattler group). Alex, thank you for your wet lab support in the first months and for laughing away all initial RNA catastrophes with me. Gunter, I'm really sorry I never got a chance to learn from your protein experience! Cameron: thanks for your friendship and for reassuring me over and over again that I *will* eventually get my PhD (yay! you were right!). The parties at your place were the best in Heidelberg! Bernd, thanks for all NMR, software & structure calculation support and for your patience in explaining obviously simple (though not always instantly to me…) NMR topics like chemical shift referencing over and over (and over…) again. Christian (x2), Julia, Serge, Anders, Anna, Frank, Fabi, Helge, Irina and the most recent Munich additions: Tobias, Alex, Fatiha, Hamed, Kostas, Gülden, Peijian, Gerd, Yun & Uli - all of you: thanks for your

advice, for listening to me complain, for working on weekends with me, having dinner, drinking beer and generally enjoying life in and outside of the lab!

Was there life outside of the lab? Indeed! To such extent that even the dearest friends made in Heidelberg are far too many to mention, especially as I'm running out of time (and pages and money). Thanks to all of you for relieving the PhD stress over holidays, brunches, dinners and nights drinking beer (or rosé) at the Sonderbar or dancing at the Rollercoaster! You're the reason why life in the village that Heidelberg really is was bearable even after 5 long years. Fred & I really appreciate that so many of you joined us in Holland to celebrate our wedding! Special thanks to the girls for their friendship, fun & the Girls' nights (we really should do another one in Munich…).

To my friends at home: thanks for keeping in touch – your mails and visits were a welcome distraction from day to day business. Here, Krista & Marit (nog een Glühwijntje meiden?) deserve special praise for making it to the Heidelberger Weihnachtsmarkt *every single year* that I was living there (and this not only to complete their mug collection…).

Krista and Esther, thanks for agreeing to support me during my defence as my paranymphs – you are the friends connecting me most to my undergraduate years in Leiden. Forgive me for not renting tuxedos as promised – we will have to dress up like Marlene Dietrich some other time…

Mamo, Tato, without the solid basis you gave me in life this thesis would not exist. Thank you for your unconditional support, trust and for your continuing interest in my career.

Ania, thanks for being the best little sister and friend and for your visits to Heidelberg and Munich!

Fred, you've stuck with me from the very start through the end of this project – who could imagine at the beginning where we would end up? Thank you, lieve Fredje, for everything! Let's enjoy this accomplishment together…

Malgosia   München, March 2009

„Das Leben ein Mittel der Erkenntnis"

- mit diesem Grundsatze im Herzen

kann man nicht nur tapfer,

sondern sogar fröhlich leben

und fröhlich lachen!

from Nietzsche's „Die fröhliche Wissenschaft" (324)