Potential RNA binding functions for Shadoo through identification and characterisation of an RGG box motif

Susan M. Corley

THE AUSTRALIAN NATIONAL UNIVERSITY

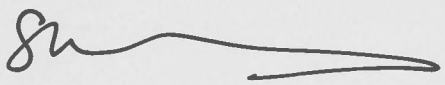
THE JOHN CURTIN SCHOOL
OF MEDICAL RESEARCH

A thesis submitted for the degree of
Doctor of Philosophy
of the Australian National University

November 2010

Statement of Originality

I declare that the work presented in this thesis is my own original work, and contains no material previously written by another person or the result of any work by another person, except as referenced in the text. The material in this thesis has not been submitted in whole or in part for a degree at this university or any other institution of higher education.



Susan Corley

November 2010

PUBLICATIONS ARISING FROM THIS WORK

Corley S.M. and Gready J.E. (2008) Identification of the RGG box motif in Shadoo: RNA-binding and signalling roles? *Bioinform Biol Insights*, 2, 383-400

ACKNOWLEDGEMENTS

Starting at the beginning, I must thank Prof Jill Gready for introducing me to the Prion Protein family and for providing the resources necessary for this project. I appreciate the academic freedom I have had to design my research project and to explore the issues that became of interest to me. I am also grateful to Jill for providing very generous computational resources through her NCI National Facility computer grants.

I want to thank all the people who participated in my supervisory panel, Prof Jill Gready, Dr Max Keniry (Research School of Chemistry), Prof Frances Shannon and Dr Anna Cowan. I have appreciated our panel meeting discussions, and the general support and encouragement I have received. I am particularly grateful to each of my panel members for reading either all, or parts of my thesis and for the helpful comments they have offered.

I thank Prof Nick Dixon at the University of Wollongong who generously hosted my visit to his laboratory where I carried out SPR experiments. I also thank Mr Paul Gugger, Ms Elisabeth Owen (RSC) and Dr Jun Fan for training me to use the CD and Fluorescence instruments.

I am indebted to Dr Deb Crittenden for teaching me computer scripting and for coming to my rescue when I couldn't get my Python scripts to work. I also thank her for our discussions, her comments on my thesis and her tremendous support throughout my PhD.

I have benefitted from the fantastic service provided by the people who run the National (supercomputer) Facility, and my discussions over the years with Andrey Bliznyuk, Vlad Vassiliev and Rika Kobayashi. I am also grateful to Andrey for his comments on Chapter 3.

I feel privileged to have undertaken my PhD at the ANU. I have enjoyed the beauty of the campus, the excellent facilities at the John Curtin School of Medical Research, and the great people I have met during my studies. In particular, I am grateful for the comradeship and support of the other members of the Gready group.

Finally, I thank my partner, family and close friends, for putting up with my 'unavailability' over the past several years, for believing in me and for giving me a huge amount of love and encouragement.

I have the deepest gratitude for the unwavering support of my parents in all my endeavours. This work is dedicated to my mother, Joan, who will no longer have to ask if I can see the light at the end of the tunnel.

ABSTRACT

Shadoo (Sho) is a member of the prion protein (PrP) family, found mainly in the brain. PrP is well-known as the central agent involved in the prion diseases – a form of fatal neurodegenerative disease. Despite decades of intense research, the natural functions of PrP have not been clearly elucidated. The reasons for this are complex, but it is likely that there is a level of redundancy in the functions performed by PrP. In particular, it has been suggested that Sho may have overlapping functions with PrP and that both proteins can compensate for one another. Therefore, investigating the natural function/s of Sho may provide some insight into the roles that the non-pathogenic form of PrP may play *in vivo*. Although the natural function of Sho is currently unknown, recent studies point to it having a role in neural tube development and neuroprotection.

Comparative genomics is a powerful technique for providing insight into functionally important protein domains. Comparing Sho protein sequences over the course of evolution highlighted a strongly conserved sequence at the beginning of the N-terminus. Literature-based data mining led me to hypothesize that it may constitute an RGG box, a known RNA binding motif. This hypothesis provides an interesting link to PrP, which can bind RNA, although an RNA binding function has not yet been identified. However, the RNA binding region of PrP, also located at the beginning of the N-terminal region, does not have the characteristics of an RGG box.

This thesis reports the results of a series of studies designed to test the hypothesis that Sho has an RGG box and establish the plausibility that Sho may play a functional role as an RNA-binding protein.

The RGG box of Sho has strong sequence similarity to the RGG box of the Fragile X Mental Retardation Protein (FMRP), which is known to bind a range of mRNAs and play an important role in neural plasticity. Comparison of the RGG boxes of Sho and FMRP reveals that, like many RNA binding domains, the RGG boxes of Sho and FMRP lie within disordered protein domains. This work examines the nature of these flexible protein domains using molecular dynamics (MD) simulations.

The RGG box of FMRP has an affinity for G-quadruplex RNA, which is also the form of RNA that binds most strongly to PrP. Here, MD simulations and biophysical experiments are used to investigate whether Sho, too, binds G-quadruplex RNA. Binding of 5 different RNA transcripts to the RGG box regions of Sho and FMRP and the N-terminus RNA binding domain of PrP was explored through circular dichroism, fluorescence and surface plasmon resonance experiments.

The results of both MD simulations and biophysical experiments show that a peptide derived from the RGG box region of Sho is capable of binding G-quadruplex RNA with physiologically relevant affinity. The Sho RGG box peptide binds to certain RNA transcripts with similar affinity to peptides comprising the FMRP RGG box and the RNA-binding region of PrP. However, some differences in RNA-binding affinities across all three peptides also indicate their ability to discriminate between different RNA targets.

Overall, the findings reported here suggest that Sho is likely to function, in some capacity, as an RNA-binding protein.

As Sho and PrP are capable of binding G-quadruplex RNA with similar affinities, it is possible that they share an RNA binding function. Given the primary location of Sho and PrP on the outer cell membrane it seems most plausible that they bind extracellular RNA in a signaling context. This finding is particularly interesting as recent research indicates that RNA may be a cofactor in the conversion of PrP to its disease-producing isoform. There is also growing interest in the role of extracellular RNA as a target for cell surface receptors. Future studies to identify likely RNA binding partners for Sho provide a promising avenue for elucidating the natural function of this protein.

Table of Contents

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" (I found it!) but 'That's funny...'
Isaac Asimov

In 2003, Shadoo was discovered, as the third member of the prion protein (PrP) family, through a collaboration involving the Gready and Simonic research groups (Premzl et al., 2003). PrP is well known for its role in a range of fatal neurodegenerative diseases, the 'transmissible spongiform encephalopathies' (TSEs) (Prusiner, 1998). Striking similarity to the topology of PrP led to this new protein being named shadow of PrP (Shadoo). Like PrP, Shadoo (Sho) is found predominantly in the brain. There is currently no evidence directly implicating Sho in the TSEs, although Sho has been found to be diminished in TSE affected tissue. The reason for this and the possibility that Sho may play either a protective or facilitative role in the TSEs is an area of active research (Watts et al., 2007, Daude et al., 2010, Miyazawa and Manuelidis, 2010).

My work in this area started by considering similarities between Sho sequences over the course of evolution. Although a sequence alignment of 5 Sho proteins had been presented in 2003 (Premzl et al., 2003), some additional sequences were available by 2005. I focused my attention on a specific region of the N-terminus of Sho (residues 25-42 in human), which is highly conserved throughout vertebrate evolution. Previously, this basic region of the N-terminus had been described as containing tetrarepeats of consensus (XXRG) (Premzl et al., 2003). I noticed that this region could also be characterized as 'RGG' repeats. Mining the available literature led me to propose that this region may be an 'RGG box'; a known RNA-binding motif, characterized in proteins such as the Fragile X Mental Retardation Protein (FMRP) (Siomi et al., 1993, Darnell et al., 2001, Schaeffer et al., 2001). The idea that Sho may be an RNA-binding protein is novel and particularly interesting given the controversy regarding the possible involvement of nucleic acid (endogenous or exogenous) in the conversion of PrP isoforms, as discussed later in this chapter.

The central hypothesis of this thesis is that the identified region of Sho is an RGG box and that, as a consequence, Sho is competent to bind RNA. This hypothesis is explored using several techniques; bioinformatics, molecular dynamics simulations and laboratory based experimental work.

The varied functions of RNA and the ancillary roles of RNA-binding proteins have become very topical over the past few years. A highlight of this research has been advances made in understanding the role of RNA-binding proteins in the brain, particularly in neural plasticity. The work contained in this thesis contributes to this field by investigating whether Sho may be another neural RNA-binding protein. The results of this research are expected to generate new avenues for investigating Sho, and potentially PrP.

As any process of scientific discovery is iterative, I will start by outlining some of the important milestones in the discovery of the PrP family, and how this story continues to evolve.

1.2 DISCOVERY OF THE PRION PROTEIN AND ITS FAMILY MEMBERS

1.2.1 The Prion Protein

The TSEs are a class of neurodegenerative diseases that affect several species, including human, as noted in Table 1-1. Symptoms associated with the TSEs include loss of motor control, dementia, paralysis, wasting and ultimately death (Collinge, 2001).

Table 1-1. Transmissible spongiform encephalopathies

Species	Disease
Human	Creutzfeldt-Jacob Disease (CJD) which may be sporadic (sCJD), variable (vCJD), iatrogenic (iCJD) or genetic (gCJD) Gerstmann-Strausler-Scheinker disease Fatal Familial Insomnia
Cattle	Bovine spongiform encephalopathy (BSE) commonly known as 'mad cow disease'
Sheep	Scrapie
Deer, mule and elk	Chronic wasting disease (CWD)

Originally, TSEs such as scrapie were thought to be caused by a 'slow virus' (Sigurdsson, 1954). However, when techniques normally effective in destroying

viruses, such as irradiation, proved to have little impact on the infective agent, investigators were motivated to propose some other causative agent such as a protein or polysaccharide (Alper et al., 1967). In response to the questions raised by Alper, JS Griffiths, a mathematician at Bedford College London, published his speculations on the way in which a protein could be the self-replicating agent causing scrapie (Griffith, 1967). This short paper is credited as the first espousal of a 'protein only' theory for the propagation of the TSEs.

However, it was Prusiner and co-workers who ultimately identified the protein that was central to the TSEs. This protein, enriched in diseased neural tissue, was named the 'prion' protein (PrP) due to it being a 'proteinaceous infective particle' (Prusiner, 1982). In 1986, PrP was cloned from human (Kretzschmar et al., 1986, Liao et al., 1986), mouse (Locht et al., 1986) and hamster (Basler et al., 1986). The *PRNP* gene is found on chromosome 20 in human and consists of 3 exons with the entire open reading frame (ORF) found in the third exon. These studies revealed a high level of conservation over species, and also showed that the sequence of PrP derived from diseased tissue did not differ from that derived from healthy tissue. PrP was identified as a 253 residue (including N-terminal and C-terminal signal sequences) glycoprotein, attached to the outer membrane of cells (primarily neurons) (Ford et al., 2002) by a GPI (glycosyl-phosphatidyl-inositol) anchor (Figure 1-1(A)).

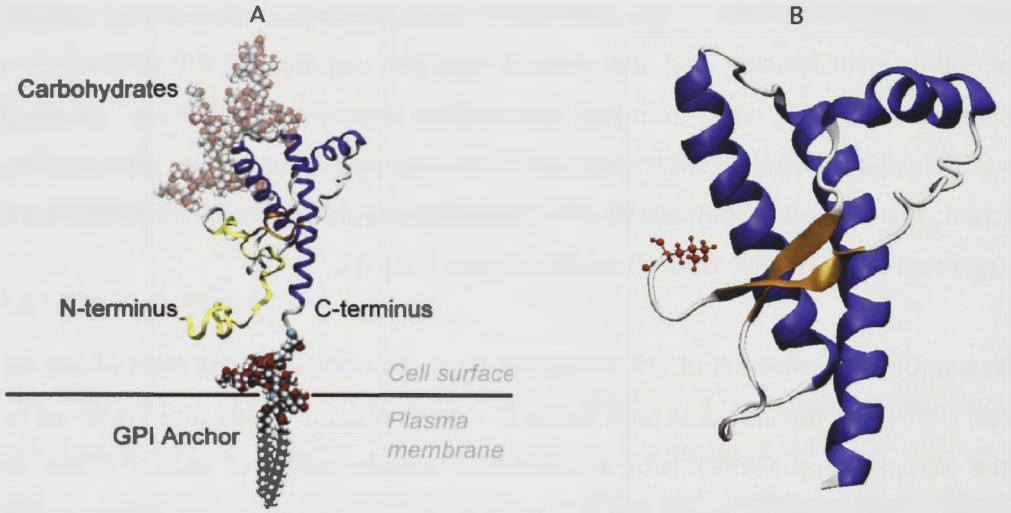
Subsequently, two isoforms of PrP were described; the normal soluble form of cellular protein (PrP^{C}) and the insoluble form found in scrapie infected tissue often referred to as the scrapie form (PrP^{Sc}) with a C-terminal domain (residues 90-231) that is resistant to digestion by proteinase K (Oesch et al., 1990). While the sequences of PrP^{C} and PrP^{Sc} are identical, there are structural differences between the 2 isoforms. The normal isoform (PrP^{C}) has a disordered N-terminus and a C-terminus which folds into 3 helices and 2 short β-strands, as shown in Figure 1-1(B) (Riek et al., 1996). On the other hand the insolubility of the C-terminus of PrP^{Sc} makes structural characterization difficult. However, biophysical studies have shown that PrP^{C} has ~ 43% α-helix and 3% β-sheet, while PrP^{Sc} has ~ 34% α-helix and 43% β-sheet (Aguzzi and Weissmann, 1997). It has been suggested that the increase in β-sheet in PrP^{Sc}, which has a tendency to aggregate into amyloid fibrils, could arise at the expense of the unstructured N-terminus (Aguzzi and Weissmann, 1997) or may involve residues 89-175 (Govaerts et al., 2004).

Prusiner and co-workers suggested that the misfolding of the prion protein from the normal PrP^{C} to the abnormal PrP^{Sc} was the cause of TSEs without any involvement of a viral agent; the 'protein only' theory (Prusiner, 1991). According to this theory, PrP^{Sc} forms a heterodimer with PrP^{C} which leads to PrP^{Sc} dependent conversion of PrP^{C} and subsequent fast formation of a PrP^{Sc} homodimer. This theory was supported by evidence that:

(i) PrP knockout mice are resistant to TSE infection (Bueler et al., 1992), and

(ii) PrP^{Sc} purified from infected brain tissue, incubated with PrP^{C} could facilitate $PrP^{C} \rightarrow PrP^{Sc}$ (albeit at a large molar excess of PrP^{Sc} 50:1 ratio) (Kocisko et al., 1994).

In 1996, Prusiner received the Nobel prize in Medicine and Physiology for his work in advancing the theory that a protein can be an infective agent solely through a change in its conformation.



Source of (A) Bennion et al. (2004)

Figure 1-1. 3D representation of the Prion Protein
(A) Model of the prion protein showing attachment via GPI anchor to the cell membrane. The disordered N-terminus is represented in yellow, the 3 helices of the C-terminus are coloured blue, 2 glycans (pink) are shown binding to the C-terminus (Bennion et al., 2004). (B) NMR structure of the C-terminus of the prion protein (PDB ID 1QLX), the first residue of the sequence (Leu125) is highlighted in red (Zahn et al., 2000).

Later developments by Prusiner and colleagues include the proposition that some other, yet unidentified macromolecule (termed "protein X") is a necessary co-factor in the conversion of the prion protein (Telling et al., 1995, Prusiner, 1998).

However, there has been dissent from the 'protein only' theory. Manuleidis of Yale University has consistently argued that a viral particle must be involved in the TSEs

(Manuelidis et al., 1987, Murdoch et al., 1990, Akowitz et al., 1994, Manuelidis et al., 1995, Manuelidis, 2003, Manuelidis, 2007, Manuelidis et al., 2007, Manuelidis et al., 2009). She suggests a 25 nm viral particle found in diseased tissue interacts with the PrP host protein allowing viral entry or replication (Manuelidis et al., 2007, Manuelidis et al., 2009). Others have proposed that the association of a nucleic acid with PrP^{Sc} could account for the different strains of prion disease (Weissmann, 1991). It has also been suggested that the co-factor proposed by Prusiner and colleagues ('protein X') may in fact be a nucleic acid, which acts as a chaperone facilitating the conversion of PrP^{C} to PrP^{Sc} (Radulescu and Korth, 1996, Cordeiro et al., 2001, Deleault et al., 2003, Grossman et al., 2003, Silva et al., 2008). Recent experimental studies have added significant weight to the latter proposition, as discussed further at section 1.4 of this chapter.

Although many studies have contributed some insight into the potential function of PrP in healthy tissue, its natural role is still not clearly understood (Aguzzi and Calella, 2009). This conundrum is not surprising as PrP is capable of binding numerous ligands, is broadly expressed in many tissues, and studies with PrP knockout mice have been equivocal (Caughey and Baron, 2006). The initial knock-out studies indicated that a strong phenotype is not apparent in the absence of PrP (Bueler et al., 1992). Since then a subtle phenotype has emerged which includes impairment of long term potentiation (Collinge et al., 1994, Criado et al., 2005), interruption of circadian rhythms, (Tobler et al., 1997) and higher sensitivity to various forms of stress. (Walz et al., 1999, Brown et al., 2002). The lack of a strong phenotype may be due to a level of redundancy in key biological processes. That is, other proteins, potentially including Sho, may also perform some of the roles played by PrP (Premzl et al., 2003, Watts et al., 2007).

While the full complement of PrP function is not understood, it is clear that PrP has a role in neural function. PrP is expressed most abundantly in the brain (Bendheim et al., 1992) where it is enriched in structures that have a capacity for functional and structural plasticity such as the dentate gyrus of the hippocampus, which undergoes synaptic rearrangement throughout life. The distribution of PrP in the developing brain suggests that PrP is involved in axonal elongation and synapse formation or remodeling (Sales et al., 2002), while the increased presence of PrP following neuronal injury indicates a role in axon regrowth (Moya et al., 2005). PrP localizes at the

neuronal synapse, both pre and post-synaptically (Sales et al., 1998, Moya et al., 2000) signifying that PrP has a special role at the neuronal synapse.

Overall, there is now evidence that PrP is involved in the processes of;

- neurite outgrowth (Santuccione et al., 2005)
- the development of neuronal polarity, including neurite definition and growth axon elongation (Moya et al., 2005)
- the creation of synaptic-like contacts (Kanaani et al., 2005)
- synaptic plasticity (Coitinho et al., 2007)
- neuronal survival (Solforosi et al., 2004)
- adult neurogenesis (Steele et al., 2006)

To date, the neural functions of PrP have been attributed to its role in signaling pathways or through the binding of metal ions. PrP enhances both neurite outgrowth and neuronal survival through its activity in two different signaling pathways; neural outgrowth is associated with tyrosine kinase $p59^{fyn}$, whereas neuronal survival is triggered by other members of the Src-related family of tyrosine kinases (Mouillet-Richard et al., 2000, Chen et al., 2003). PrP interacts directly with the neural cell adhesion molecule (NCAM) and thereby facilitates its recruitment to lipid rafts. NCAM which is bound to the receptor type protein phosphatase α (RPTPα) then activates the fyn tyrosine kinase which leads to neurite outgrowth (Schmitt-Ulms et al., 2001, Santuccione et al., 2005).

Other mechanisms have been proposed by which PrP may influence neural activity; including through the binding of copper at the synapse (Herms et al., 1999, Brown, 2001) and in calcium homeostasis (Krebs et al., 2007).

1.2.2 Doppel

The search for genes related to *PRNP* commenced in the mid 1980s, when very little genomic data was available. Sequencing of the regions flanking *PRNP* failed to reveal additional ORFs. However, when sequencing was extended further downstream of *PRNP*, a new *PRNP*-like gene was found. The new gene, 16 kb downstream of *PRNP*, was thought to be the product of a duplication event involving an ancestoral *PRNP* gene. As such the new gene and its 178 residue protein product were called Doppel "Dpl" (German, "double") (Moore et al., 1999, Watts and Westaway, 2007).

The expression pattern of Dpl is quite different from PrP. Dpl is found mainly in the testis and also in the heart (Moore et al., 1999). The primary function identified for Dpl is in male spermatogenesis and sperm-egg interactions. Male mice deficient in Dpl have a reduced number of spermatozoa, structural abnomalities in the spermatozoa and lack a well defined acrosome resulting in infertility (Qin et al., 2006).

1.2.3 Shadoo

In 2002, a new PrP-like gene was identified in fish (Simonic group, University of Milan). This lead was shared with the Gready group at the ANU where *in silico* work, including EST database searches, genome analyses and searches of tissue expression databases led to the discovery of a new gene present also in mammals and highly conserved from fish to mammals. Analysis of tissue expression data indicated expression mainly in mammalian brain, which was confirmed experimentally with a tissue panel from rat. The new gene, *SPRN* (shadow of *PRNP*), consists of 2 exons with the ORF found in the second exon. In human, *SPRN* is found on chromosome 10 and codes for a 151 residue protein (including signal sequences), with strong topological similarity to PrP, as illustrated in Figure 1-2 (Premzl et al., 2003).

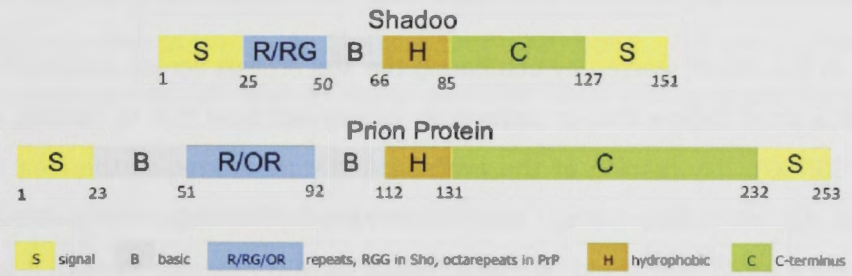


Figure 1-2. Protein topology of human Sho compared with human PrP
The regions of the Sho and PrP proteins are colour coded in accordance with original designation of these areas by Premzl et al. (2003).

Both proteins have N and C-terminal signal sequences indicating GPI anchor attachment, a central hydrophobic area and an N-terminus rich in basic residues. The C-terminus of Sho differs from PrP in that it is much shorter and the length and residue composition suggest it is disordered. Sho's location at the cellular membrane has since been confirmed by Watts et al. (2007).

Unlike Dpl, Sho has a similar expression pattern to PrP. It is expressed most abundantly in the hippocampus and cerebellum (hippocampal neurons and cerebellar

Purkinje cells) but is also found in the cerebral cortex, medulla and thalamus (Watts et al., 2007).

The task of uncovering the function of Sho has only just begun. The main experimental findings to date are summarized in Table 1-2.

Table 1-2. Experimental studies into the function of Sho

2007	Confirmed that Sho is found at the cell surface (as a GPI anchored protein). Mouse *SPRN* expressed most prominently in pyramidal cells and Purkinje cells of hippocampus and cerebellum. Analysis of $Prpn^{0/0}$ mice failed to reveal upregulation of Sho in response to PrP deficiency. Sho found to be depleted in prion infected mice. The hydrophobic region of Sho confers a neuroprotective effect similar to PrP. CD analysis indicates that Sho is disordered.	(Watts et al., 2007)
2009	Demonstrates a positive correlation between *PRNP* and *SPRN* mRNA expression in sheep cerebral and cerebellar cortex, suggesting co-regulation.	(Lampo et al., 2009)
2009	Expression of *PRNP* and *SPRN* transcripts in sheep was not affected by scrapie infection. No positive correlation between expression of *PRNP* and *SPRN* and disease progression.	(Gossner et al., 2009)
2009	Knock down of SPRN expression in *PRNP* knockout mice showed embryonic defects in neurulation and lethality. As absence of *SPRN* alone did not affect normal development, these results suggest overlapping and compensatory functions for PrP and Sho.	(Young et al., 2009)
2010	Showed that with > 10 days incubation at 37°C mouse and ovine Sho gradually forms amyloid fibrils.	(Daude et al., 2010)
2010	The effect of a several different TSE producing agents on Sho levels was tested. Sho reductions were found to be agent specific. Reduction in Sho levels did not correlate with a change in misfolded PrP. The results suggest that reductions of Sho in PrP disease arise from complex cell processing mechanisms i.e. are disease related changes.	(Miyazawa and Manuelidis, 2010)

The work in this thesis started by comparing the Sho protein across species. I focused initially on a short region that is completely conserved from fish to human, residues 25-30 (KGGRGG). My reading of the available literature led me to the idea that this may be an Arg methylation site. I then observed that this sequence is found at the beginning of a series that could be characterized as RGG repeats. Overall, it appeared that this region had the hallmarks of an RGG box, an RNA binding motif. This new observation forms the basis of the hypothesis that Sho is competent to bind RNA and that this capability may be an integral mechanism for Sho's natural function. The first step in investigating this hypothesis is to compare Sho's RGG box with the RGG box domains of other proteins. This is explored fully in Chapter 2.

1.3 THE RGG BOX MOTIF AND THE FRAGILE X MENTAL RETARDATION PROTEIN

One of the proteins identified in Chapter 2 with a similar RGG box motif to Sho's is the Fragile X Mental Retardation Protein (FMRP) (Siomi et al., 1993, Bagni and Greenough,

2005). This protein is well-studied due to its role in 'Fragile X syndrome', the most common form of inherited mental retardation. This disorder occurs when an abnormal expansion of a d(CGG) trinucleotide in the 5' UTR of the *FMR1* gene impacts on the promoter region preventing translation of FMRP (Oberle et al, 1991).

Like PrP and Sho, FMRP is expressed in many tissues but is most abundant in the brain, where it is involved in the transport and translational regulation of neural mRNAs (Brown et al, 2001, Darnell et al, 2001, Schaefer et al 2001), thus playing a role in neural plasticity. Up-regulation of certain mRNAs is seen in FMRP knockout mice, indicating that FMRP functions as a repressor of translation (Laggerbauer et al., 2001, Li et al., 2001, Mazroui et al., 2002). FMRP also interacts with ribosomes (Siomi et al., 1996, Corbin et al., 1997) and has been associated with RNA interference (Ishizuka et al., 2002).

The RGG box of FMRP is found in a disordered protein domain, similar to Sho and has an affinity for G-quadruplex RNA (Darnell et al., 2001, Ramos et al., 2003, Zanotti et al., 2006, Menon and Mihailescu, 2007). This constitutes another interesting similarity to the prion protein family (Weiss et al., 1997, Proske et al., 2002).

Obvious similarities between FMRP and the PrP protein family (Table 1-3) have motivated my comparison of the Sho and FMRP RGG boxes. The structures of these regions are compared (Chapter 4) along with their RNA binding capacities (Chapters 6-9).

Table 1-3. Similarities between FMRP and the PrP Family

FMRP	PrP Family
Found primarily in brain and testes	PrP (Ford et al., 2002) and Sho found primarily in brain (Premzl et al., 2003, Watts et al., 2007) Dpl found primarily in testes (Moore et al., 1999)
Functional role in neural plasticity (Bagni and Greenough, 2005)	PrP has a functional role in neural plasticity (Coitinho et al., 2007)
RGG box has affinity for G-quadruplex RNA (Darnell et al., 2001)	PrP has affinity for G-quadruplex RNA (Weiss et al., 1997)
RGG box has spacing RGG-X9-RGG	Sho RGG box has spacing RGG-X9-RGG
RNA chaperone function (Ivanyi-Nagy et al., 2005)	Proposed RNA chaperone function for PrP (Ivanyi-Nagy et al., 2005)

1.4 NUCLEIC ACID BINDING AND PRP CONVERSION

The identification of a potential RNA binding motif in Sho is of particular interest as research over the past decade has shown that PrP is capable of binding both DNA and RNA. A primary motivation of previous studies has been to identify molecules that may be involved in the conversion of PrP^{C} to other isoforms. To put this in context: it has proven very difficult to convert recombinant PrP, seeded with PrP^{Sc}, into an infective form, capable of producing prion disease. This, of course, is the ultimate test of the protein-only theory. This conversion *in vitro* only seemed possible with a very high excess of PrP^{Sc} to PrP e.g. 50:1 molar ratio (Kocisko et al., 1994). These difficulties fueled the search for other molecules that may participate in, or catalyze the conformational conversion of PrP. As PrP is a positively charged protein the obvious candidate was a polyanionic molecule e.g. sulfated glycans (Caughey et al., 1994) heparin sulfate (Shaked et al., 2001) and nucleic acids. Most attention has focused on the involvement of nucleic acids.

The laboratory test for PrP^{C} conversion has generally been the amplification of protease resistant molecules (PrP^{Res}). Although protease resistance is a feature of the infective molecule (PrP^{Sc}) it does not necessarily follow that all protease resistant isoforms will cause prion disease (Aguzzi and Weissmann, 1997). Accordingly, in this section both terms, PrP^{Res} and PrP^{Sc} will be used. Biophysical techniques have frequently been used to monitor features that accompany PrP^{Sc} formation i.e. protein aggregation and polymerization. There is a substantial body of literature on this topic. Not all can be mentioned here, however the major experimental evidence for nucleic acid binding is summarized in Table 1-4.

In 1999, Nandi and Leclerc were the first to demonstrate that recombinant PrP (recPrP) polymerizes in the presence of DNA to form aggregates with amyloid-like morphologies (Nandi and Leclerc, 1999). Shortly thereafter, it was shown that PrP^{C} binds with specificity to certain synthetic DNA transcripts (oligonucleotides ranged from 24-34 nucleotides) which catalyze the aggregation of PrP^{C} in a concentration-dependent manner (Cordeiro et al., 2001).

At the same time, Gabus and co-workers reported that the disordered N-terminus of recPrP is able to bind RNA as well as DNA, to form condensed nucleoprotein structures similar to those obtained with retroviral nucleocapsid proteins (Gabus et

al., 2001). This work extended earlier suggestions that a retrovirus, shielded by binding proteins, may be involved in prion disease (Akowitz et al., 1994).

Table 1-4. Experimental evidence for PrP nucleic acid binding

1994	Manuelidis and colleagues find endogenous viral RNA in purified CJD infected brain tissue and suggest that RNA-protein complexes are an inherent component of infectivity.	(Akowitz et al., 1994, Manuelidis et al., 1995)
1997	Weiss, Proske and others use SELEX techniques to identify RNA aptamers with high specificity for PrP. They show that the N-terminus of PrP (res 23-52) binds with high affinity to G-quadruplex RNA.	(Weiss et al., 1997)
1999	Shown that low concentration of DNA can induce oligomerization of PrP^{c} into a form similar to PrP^{Sc}. Polymerization is seen to occur in both the PrP and the DNA.	(Nandi and Leclerc, 1999)
2001	Brazilian researchers, Cordeiro, Silva and others, demonstrate that PrP has high affinity for specific DNA. Binding faciltates $PrP^{C} \rightarrow PrP^{Res}$. Where the DNA:PrP ratio is high, PrP aggregation is prevented, but at lower ratios DNA appears to facilitate PrP aggregation.	(Cordeiro et al., 2001)
2001	The French team of Darlix, Gabus and others find that PrP forms condensed nucleoprotein complexes with both DNA and RNA. It is suggested that PrP acts in a similar manner to retroviral nucleocapsid proteins, such as HIV-1 NCp7.	(Gabus et al., 2001, Derrington et al., 2002)
2002	Proske and others identify RNA aptamers which bind to the hydrophobic region of PrP (res 90-129). They find that a G-quadruplex RNA binds with highest affinity and reduces PrP^{Sc} aggregation in prion infected cells.	(Proske et al., 2002)
2003	The Suppatapone group shows that host encoded single strand RNA increases the efficiency of PrP^{C} (from Hamster brain homogenate)$\rightarrow$ PrP^{Res} *vitro* fourfold.	(Deleault et al., 2003)
2003	Identification of RNA aptamers which bind PrP and prevent conversion of PrP^{C} to PrP^{Sc}	(Rhie et al., 2003)
2003	Grossman and colleagues show that small highly structured RNAs < 200 nt (shs RNA), have varying affinity for PrP^{C} in vitro under physiological conditions.	(Zeiler et al., 2003)
2003	Grossman and colleagues extend their study to show PrP converts to PrP^{Res} in complex with shsRNAs and that the RNA is also protected from degradation in this complex.	(Adler et al., 2003)
2005	The Suppatapone group confirm that crude brain extract contains one or more co-factors necessary for the **efficient** conversion of purified $PrP^{C} \rightarrow$ PrP^{Res} RNA molecules were most efficient irrespective of their source.	(Deleault et al., 2005)
2005	The protein misfolding cyclic amplification (PMCA) technique (Saborio et al., 2001) is used to efficiently convert $PrP^{C} \rightarrow PrP^{Res}$ which was found to be infective but at a lower level than brain derived PrP^{Sc}.	(Castilla et al., 2005)
2006	SAXS and NMR structure of PrP bound to DNA aptamer reveals that binding occurs through both the C-terminal globular domain and the disordered N-terminus.	(Lima et al., 2006)
2006	Mercey et al. used a SELEX technique to identify RNA aptamers which bound with high affinity to PrP. SPR experiments showed that high affinity binding occurred at 2 Lys clusters (res 25-34 and 101-110).	(Mercey et al., 2006)
2007	The Suppatapone group show that infectious PrP^{Sc} can be generated *de novo* from PrP^{C} (purified from brain homogenate) in the presence of co-purified lipid molecules and synthetic nucleic acid, using the PMCA technique. The PrP^{Sc} formed was sufficient to produce scrapie in hamster.	(Deleault et al., 2007)
2010	The PMCA technique was used to convert recPrP, in the presence of synthetic lipid and total RNA (from liver) to PrP^{Sc}. The PrP^{Sc} formed was sufficient to produce scrapie rapidly in all inoculated mice.	(Wang et al., 2010)

Grossman and colleagues then tested the affinity of PrP for short RNAs (< 200 nt) with varying secondary structure, including the G-quadruplex RNA identified by Weiss et al. They found that all the RNAs bound PrP with considerable affinity (K_d < 60 nM).

However, competition assays produced a greater differential in binding affinity. The authors concluded that PrP has both a non-specific and specific RNA binding site. They postulated that binding to a non-specific N-terminal site results in a conformational change that allows specific binding at the C-terminus (Zeiler et al., 2003). Some support for this proposition comes from later work, which shows that a DNA aptamer binds PrP at sites on both the N-terminus and C-terminus (Lima et al., 2006). Grossman and colleagues also established that multiple PrP molecules can bind to a single RNA transcript and that this coverage can protect the RNA against degradation by RNase A, similar to viral RNA binding proteins e.g. NCp7. Conversely, binding results in protease resistance of PrP^{C} i.e. conversion of $PrP^{C} \rightarrow PrP^{Res}$ (Adler et al., 2003). Overall, it is suggested that short RNAs, possibly small non-coding RNA may act as a catalyst in the structural transition of PrP^{C} to PrP^{Sc}.

Although the abovementioned experiments demonstrate that RNA and DNA bind at the N-terminus of PrP and facilitate the conversion to PrP^{Res} , with varying efficiency, they do not go the next step of showing that the PrP^{Res} produced can induce TSE disease. However, significant breakthroughs have occurred recently with a number of research groups employing the protein misfolding cyclic amplification (PMCA) technique (Saborio et al., 2001) to convert purified PrP^{C} to highly infective PrP^{Sc} (Castilla et al., 2005, Deleault et al., 2007, Wang et al., 2010). Taken together these studies demonstrate that PrP is central to the propagation of the TSEs but, as suspected for a long time now, other molecules play a necessary role in conversion of PrP to the disease producing isoform. These experiments show that a polyanionic molecule such as RNA is likely to be involved, although a virus is not necessary for the continued propagation of PrP^{Sc} and disease. This process can occur merely in the presence of PrP^{C}, endogenous RNA and lipids.

However, such a simple formula implies that prion disease should be widespread in the community. Fortunately, it is not, and so the question remains as to what instigates the morbid amplification of PrP^{Sc} in the first place. In this context it is difficult to rule out the possibility that the original trigger for conversion is a molecule not found in the vast majority of the population e.g. an exogenous RNA carried by viral proteins.

Efforts have been made to identify specific RNA sequences with high affinity for PrP (Weiss et al., 1997, Proske et al., 2002, Mercey et al., 2006). This research points to

quadruplex forming RNA as having particular affinity for PrP. Most recently, Mercey et al. identified an RNA aptamer termed RM312 with high affinity for PrP (UCAAGACGUCGGGGAUUGGCAAACCCCGUUUCCUUGAG) (Mercey et al., 2006). This aptamer contains the same 21 nt core previously found to bind PrP (aptamer DP7). Interestingly, DP7 is suspected to form a G-quadruplex (Proske et al., 2002).

PrP does not have an RGG box, but RNA binding does occur at the disordered N-terminus. One binding site identified in the N-terminus encompasses residues 23-52 and includes a Lys cluster (Lys23, Lys24, Lys27) (Weiss et al., 1997, Gabus et al., 2001, Mercey et al., 2006).

In this thesis work, a peptide taken from this region of PrP (res 23-52) has been used in RNA-binding studies. Binding of 5 RNA transcripts with the PrP peptide, Sho RGG box peptide and FMRP RGG box peptide are compared. These experiments are discussed in Chapter 9.

1.5 G-QUADRUPLEX NUCLEIC ACID

Both PrP (Weiss et al., 1997) and FMRP (Darnell et al., 2001, Schaeffer et al., 2001, Ramos et al., 2003, Darnell et al., 2004, Menon and Mihailescu, 2007) have affinity for RNA which folds into a G-quadruplex. The G-quadruplex is formed when nucleic acids (DNA and RNA) which are rich in guanine bases self-associate to form structures based on the guanine tetrad also referred to as a 'G-quartet'.

The G-quartet is a planar arrangement of 4 guanine bases bound by Hoogsteen hydrogen bonds. A G-quadruplex is formed when 2 or more tetrads stack on top of each other via π stacking (Simonsson, 2001). The tetrad stack is stabilized by cations which insert between or within the planes (Williamson et al., 1989). This structure may be formed from one (intramolecular), two or four molecules of nucleic acid. G-quadruplexes are therefore a heterogeneous species with diversity arising from the different number of molecules involved, the direction of alternating strands, the arrangement and length of loops and the possibility of *anti* or *syn* glycosidic torsional angles (Burge et al., 2006). See Figures 5-1, 5-2 and Chapter 5 where the G-quadruplex structure is discussed in more detail.

Scientific interest in the G-quadruplex has spiked with the discovery that telomeres at the end of chromosomes and certain areas within the promoter regions of some genes fold into the G-quadruplex. Telomeres are a G-rich stretch of DNA that forms a cap at the end of chromosomes thus protecting the chromosome from exonucleolytic degradation and terminal fusions. Human telomeric DNA consists of tandem repeats of the sequence d(TTAGGG)n terminating in a single strand 3' overhang of 100 - 200 bases (Moyzis et al., 1988). Each cell division results in a shortening of the telomere until a critical level is reached and the cell undergoes apoptosis (Harley et al., 1990). If this normal aging process is disturbed a cell can avoid apoptosis and become immortal, as occurs in cancer. The enzyme telomerase (activated in 80 - 85% of cancer cells) can prevent telomere shortening and is thus associated with cancer (Kim et al., 1994, Hanahan and Weinberg, 2000). The finding that G-quadruplex structure in telomeric DNA can inhibit the action of telomerase provides an important direction for cancer research (Hurley et al., 2000, Neidle and Read, 2000, Bryan and Baumann, 2010).

The promoter regions of certain genes can also form G-quadruplex structure, as exemplified by the promoter region of the MYC gene. The MYC protein is a transcription factor that controls the expression of a large number of genes (Sears, 2004). As overexpression of MYC is linked to many types of cancer (Secombe et al., 2004) the structure of the gene promoter has received a good deal of attention. As a result it is now known that transcription is repressed through stabilization of the intramolecular quadruplex formed by the promoter (Siddiqui-Jain et al., 2002). Similar examples are provided by the promoter regions of PDGF-A (Qin et al., 2007) and KRAS (Cogoi and Xodo, 2006). Accumulating evidence that the G-quadruplex does form *in vivo* together with a better understanding of the effect that ligands, including other proteins, have on the quadruplex structure has seen growth in the search for therapeutic agents based on quadruplex biology (Oganesian and Bryan, 2007).

RNA quadruplexes also play an important role in cellular processes such as the regulation of protein translation from mRNA (Oliver et al., 2000). The identification of the quadruplex motif in the 5' UTR of some 3000 genes has led to the suggestion that this motif may be a common mechanism for translational modulation as demonstrated in the case of the NRAS proto-oncongene (Kumari et al., 2007). Another twist on this theme is the discovery of a G-quadruplex in the mRNA that codes for the RGG box of FMRP and the finding that it is a target for the translated

protein. This suggests a mechanism by which the protein regulates its own translation or the degradation of its coding mRNA (Schaeffer et al., 2001). The G-quadruplex motif has also been found in the mRNA of neural proteins associated with synaptic function and dendritic growth that are high affinity targets for FMRP. FMRP holds these mRNA sequences dormant pending neural signaling that allows translation of the required proteins (Darnell et al., 2001).

1.6 FUNCTIONAL IMPLICATIONS - COMMON FUNCTIONS FOR PRP AND SHO?

While it is clear that PrP is capable of binding both RNA and DNA, there is not yet any direct evidence that PrP has a functional role that involves RNA binding. This has not prevented speculation that such a role may exist and could include the binding of mRNA (Radulescu and Korth, 1996), nucleic acid metabolism (Gabus et al., 2001), and transport of nucleic acid across the cell milieu (Lima et al., 2006).

The similarities between PrP and Sho, namely, similar expression in the brain, and similar protein topology led to the original suggestion of functional overlap between PrP and Sho (Premzl et al., 2003). Since then, experimental studies have shown similar functions in neuroprotection, at least from the toxic effects of Dpl and ΔPrP in *PRNP*^{null} mice (Watts et al., 2007), and in neurulation (Young et al., 2009). Positive correlation between *SPRN* and *PRNP* expression in the cerebellum and cerebrum in sheep, prompted Lampo et al. to suggest the possibility of co-regulation of these proteins (Lampo et al., 2009) which may also be consistent with similar functions or functional pathways.

This thesis is concerned with the idea that Sho may be an RNA-binding protein with a normal function in the brain involving RNA binding. Additional investigation of this idea, alongside further investigation of PrP's RNA-binding potential, could provide important evidence of similarity in PrP and Sho function.

1.7 THESIS OVERVIEW

The primary aim of this study is to establish whether the RGG box of Sho is capable of binding RNA. The starting point is to compare the Sho RGG box with the RGG box regions of other proteins. The other RGG box proteins were identified from searches

of the protein database and draws on the existing literature in relation to their function. This work is presented in Chapter 2.

The most well-studied of the RGG box proteins is FMRP, which has common features with the prion protein family. In particular, its RGG box sequence is similar to the RGG box of Sho. Throughout this work, the RGG box of Sho is compared with that of FMRP.

Initially, molecular dynamics (MD) simulations are used to see how similar these regions are on a molecular level, and how molecular structure and dynamics may influence RNA binding. Generally, MD simulations are used to better understand molecules that have a well-defined structure. Here, MD is used for a very different purpose, that is, to describe molecular domains which are disordered; i.e. do not have a well-defined structure. This analysis calls for a different approach to the conventional molecular dynamics methods that are used to characterize the conformational dynamics of a single well-ordered structure or small set of stable conformers. Investigation of a disordered peptide requires methods that allow a broad sampling of conformational space. This study employs the new method of hybrid replica exchange molecular dynamics (hybrid-REMD) and compares it to implicit solvation REMD (Okur et al., 2006). These methods are discussed in Chapter 3. REMD studies of the RGG box regions of both peptides are presented in Chapter 4.

As the G-quadruplex has emerged as an RNA form likely to bind both PrP and FMRP, this thesis examines whether G-quadruplex RNA is also a target for the RGG region of Sho. MD simulations of the G-quadruplex structure of both DNA and RNA are discussed in Chapter 5. Subsequently the binding of the Sho and FMRP RGG boxes to quadruplex RNA is compared (Chapters 6-8).

The MD studies are complemented by biophysical experiments. Binding assays explore the interaction between peptides derived from Sho, FMRP and PrP and five short RNA transcripts, 3 of which are expected to form a G-quadruplex. Surface Plasmon Resonance and fluorescence studies are used to estimate binding affinity for the different short RNAs. Circular dichroism is used to probe the effect of peptide binding on the G-quadruplex structure (Chapter 9).

Although Sho is the focus of this thesis, these investigations also provide general insight into:

1. The nature of the RGG box binding motif
2. The mode of RNA binding seen in the FMRP RGG box
3. The effect of peptide binding on the G-quadruplex structure

Finally, the conclusions drawn from these studies including functional implications for the PrP family are presented in Chapter 10.

1.8 REFERENCES

Adler, V., Zeiler, B., Kryukov, V., Kascsak, R., Rubenstein, R. & Grossman, A. (2003) Small, highly structured RNAs participate in the conversion of human recombinant PrP(Sen) to PrP(Res) in vitro. *J Mol Biol,* 332, 47-57.

Aguzzi, A. & Calella, A.M. (2009) Prions: protein aggregation and infectious diseases. *Physiol Rev,* 89, 1105-52.

Aguzzi, A. & Weissmann, C. (1997) Prion research: the next frontiers. *Nature,* 389, 795-8.

Akowitz, A., Sklaviadis, T. & Manuelidis, L. (1994) Endogenous viral complexes with long RNA cosediment with the agent of Creutzfeldt-Jakob disease. *Nucleic Acids Res,* 22, 1101-7.

Alper, T., Cramp, W.A., Haig, D.A. & Clarke, M.C. (1967) Does the agent of scrapie replicate without nucleic acid? *Nature,* 214, 764-6.

Bagni, C. & Greenough, W.T. (2005) From mRNP trafficking to spine dysmorphogenesis: the roots of fragile X syndrome. *Nat Rev Neurosci,* 6, 376-87.

Basler, K., Oesch, B., Scott, M., Westaway, D., Walchli, M., Groth, D.F., Mckinley, M.P., Prusiner, S.B. & Weissmann, C. (1986) Scrapie and cellular PrP isoforms are encoded by the same chromosomal gene. *Cell,* 46, 417-28.

Bendheim, P.E., Brown, H.R., Rudelli, R.D., Scala, L.J., Goller, N.L., Wen, G.Y., Kascsak, R.J., Cashman, N.R. & Bolton, D.C. (1992) Nearly ubiquitous tissue distribution of the scrapie agent precursor protein. *Neurology,* 42, 149-56.

Bennion, B.J., Demarco, M.L. & Daggett, V. (2004) Preventing misfolding of the prion protein by trimethylamine N-oxide. *Biochemistry,* 43, 12955-63.

Brown, D.R. (2001) Copper and prion disease. *Brain Res Bull,* 55, 165-73.

Brown, D.R., Nicholas, R.S. & Canevari, L. (2002) Lack of prion protein expression results in a neuronal phenotype sensitive to stress. *J Neurosci Res,* 67, 211-24.

Bryan, T.M. & Baumann, P. (2010) G-quadruplexes: from guanine gels to chemotherapeutics. *Methods Mol Biol,* 608, 1-16.

Bueler, H., Fischer, M., Lang, Y., Bluethmann, H., Lipp, H.P., Dearmond, S.J., Prusiner, S.B., Aguet, M. & Weissmann, C. (1992) Normal development and behaviour of mice lacking the neuronal cell-surface PrP protein. *Nature,* 356, 577-82.

Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. & Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res,* 34, 5402-15.

Castilla, J., Saa, P., Hetz, C. & Soto, C. (2005) In vitro generation of infectious scrapie prions. *Cell,* 121, 195-206.

Caughey, B. & Baron, G.S. (2006) Prions and their partners in crime. *Nature,* 443, 803-10.

Caughey, B., Brown, K., Raymond, G.J., Katzenstein, G.E. & Thresher, W. (1994) Binding of the protease-sensitive form of PrP (prion protein) to sulfated glycosaminoglycan and congo red [corrected]. *J Virol,* 68, 2135-41.

Chen, S., Mange, A., Dong, L., Lehmann, S. & Schachner, M. (2003) Prion protein as trans-interacting partner for neurons is involved in neurite outgrowth and neuronal survival. *Mol Cell Neurosci,* 22, 227-33.

Cogoi, S. & Xodo, L.E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res,* 34, 2536-49.

Coitinho, A.S., Lopes, M.H., Hajj, G.N., Rossato, J.I., Freitas, A.R., Castro, C.C., Cammarota, M., Brentani, R.R., Izquierdo, I. & Martins, V.R. (2007) Short-term memory formation and long-term memory consolidation are enhanced by cellular prion association to stress-inducible protein 1. *Neurobiol Dis.*

Collinge, J. (2001) Prion diseases of humans and animals: their causes and molecular basis. *Annu Rev Neurosci,* 24, 519-50.

Collinge, J., Whittington, M.A., Sidle, K.C., Smith, C.J., Palmer, M.S., Clarke, A.R. & Jefferys, J.G. (1994) Prion protein is necessary for normal synaptic function. *Nature,* 370, 295-7.

Corbin, F., Bouillon, M., Fortin, A., Morin, S., Rousseau, F. & Khandjian, E.W. (1997) The fragile X mental retardation protein is associated with poly(A)+ mRNA in actively translating polyribosomes. *Hum Mol Genet,* 6, 1465-72.

Cordeiro, Y., Machado, F., Juliano, L., Juliano, M.A., Brentani, R.R., Foguel, D. & Silva, J.L. (2001) DNA converts cellular prion protein into the beta-sheet conformation and inhibits prion peptide aggregation. *J Biol Chem,* 276, 49400-9.

Criado, J.R., Sanchez-Alavez, M., Conti, B., Giacchino, J.L., Wills, D.N., Henriksen, S.J., Race, R., Manson, J.C., Chesebro, B. & Oldstone, M.B. (2005) Mice devoid of prion protein have cognitive deficits that are rescued by reconstitution of PrP in neurons. *Neurobiol Dis,* 19, 255-65.

Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T. & Darnell, R.B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell,* 107, 489-99.

Darnell, J.C., Warren, S.T. & Darnell, R.B. (2004) The fragile X mental retardation protein, FMRP, recognizes G-quartets. *Ment Retard Dev Disabil Res Rev,* 10, 49-52.

Daude, N., Ng, V., Watts, J.C., Genovesi, S., Glaves, J.P., Wohlgemuth, S., Schmitt-Ulms, G., Young, H., Mclaurin, J., Fraser, P.E. & Westaway, D. (2010) Wild-type Shadoo proteins convert to amyloid-like forms under native conditions. *J Neurochem,* 113, 92-104.

Deleault, N.R., Geoghegan, J.C., Nishina, K., Kascsak, R., Williamson, R.A. & Supattapone, S. (2005) Protease-resistant prion protein amplification reconstituted with partially purified substrates and synthetic polyanions. *J Biol Chem,* 280, 26873-9.

Deleault, N.R., Harris, B.T., Rees, J.R. & Supattapone, S. (2007) Formation of native prions from minimal components in vitro. *Proc Natl Acad Sci U S A,* 104, 9741-6.

Deleault, N.R., Lucassen, R.W. & Supattapone, S. (2003) RNA molecules stimulate prion protein conversion. *Nature,* 425, 717-20.

Derrington, E., Gabus, C., Leblanc, P., Chnaidermann, J., Grave, L., Dormont, D., Swietnicki, W., Morillas, M., Marck, D., Nandi, P. & Darlix, J.L. (2002) PrPC has nucleic acid chaperoning properties similar to the nucleocapsid protein of HIV-1. *C R Biol,* 325, 17-23.

Ford, M.J., Burton, L.J., Li, H., Graham, C.H., Frobert, Y., Grassi, J., Hall, S.M. & Morris, R.J. (2002) A marked disparity between the expression of prion protein and its message by neurones of the CNS. *Neuroscience,* 111, 533-51.

Gabus, C., Auxilien, S., Pechoux, C., Dormont, D., Swietnicki, W., Morillas, M., Surewicz, W., Nandi, P. & Darlix, J.L. (2001) The prion protein has DNA strand transfer properties similar to retroviral nucleocapsid protein. *J Mol Biol,* 307, 1011-21.

Gossner, A.G., Bennet, N., Hunter, N. & Hopkins, J. (2009) Differential expression of Prnp and Sprn in scrapie infected sheep also reveals Prnp genotype specific differences. *Biochem Biophys Res Commun,* 378, 862-6.

Govaerts, C., Wille, H., Prusiner, S.B. & Cohen, F.E. (2004) Evidence for assembly of prions with left-handed beta-helices into trimers. *Proc Natl Acad Sci U S A,* 101, 8342-7.

Griffith, J.S. (1967) Self-replication and scrapie. *Nature,* 215, 1043-4.

Grossman, A., Zeiler, B. & Sapirstein, V. (2003) Prion protein interactions with nucleic acid: possible models for prion disease and prion function. *Neurochem Res,* 28, 955-63.

Hanahan, D. & Weinberg, R.A. (2000) The hallmarks of cancer. *Cell,* 100, 57-70.

Harley, C.B., Futcher, A.B. & Greider, C.W. (1990) Telomeres shorten during ageing of human fibroblasts. *Nature,* 345, 458-60.

Herms, J., Tings, T., Gall, S., Madlung, A., Giese, A., Siebert, H., Schurmann, P., Windl, O., Brose, N. & Kretzschmar, H. (1999) Evidence of presynaptic location and function of the prion protein. *J Neurosci,* 19, 8866-75.

Hurley, L.H., Wheelhouse, R.T., Sun, D., Kerwin, S.M., Salazar, M., Fedoroff, O.Y., Han, F.X., Han, H., Izbicka, E. & Von Hoff, D.D. (2000) G-quadruplexes as targets for drug design. *Pharmacol Ther,* 85, 141-58.

Ishizuka, A., Siomi, M.C. & Siomi, H. (2002) A Drosophila fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes Dev,* 16, 2497-508.

Ivanyi-Nagy, R., Davidovic, L., Khandjian, E.W. & Darlix, J.L. (2005) Disordered RNA chaperone proteins: from functions to disease. *Cell Mol Life Sci,* 62, 1409-17.

Kanaani, J., Prusiner, S.B., Diacovo, J., Baekkeskov, S. & Legname, G. (2005) Recombinant prion protein induces rapid polarization and development of synapses in embryonic rat hippocampal neurons in vitro. *J Neurochem,* 95, 1373-86.

Kim, N.W., Piatyszek, M.A., Prowse, K.R., Harley, C.B., West, M.D., Ho, P.L., Coviello, G.M., Wright, W.E., Weinrich, S.L. & Shay, J.W. (1994) Specific association of human telomerase activity with immortal cells and cancer. *Science,* 266, 2011-5.

Kocisko, D.A., Come, J.H., Priola, S.A., Chesebro, B., Raymond, G.J., Lansbury, P.T. & Caughey, B. (1994) Cell-free formation of protease-resistant prion protein. *Nature,* 370, 471-4.

Krebs, B., Wiebelitz, A., Balitzki-Korte, B., Vassallo, N., Paluch, S., Mitteregger, G., Onodera, T., Kretzschmar, H.A. & Herms, J. (2007) Cellular prion protein modulates the intracellular calcium response to hydrogen peroxide. *J Neurochem,* 100, 358-67.

Kretzschmar, H.A., Stowring, L.E., Westaway, D., Stubblebine, W.H., Prusiner, S.B. & Dearmond, S.J. (1986) Molecular cloning of a human prion protein cDNA. *DNA,* 5, 315-24.

Kumari, S., Bugaut, A., Huppert, J.L. & Balasubramanian, S. (2007) An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. *Nat Chem Biol,* 3, 218-21.

Laggerbauer, B., Ostareck, D., Keidel, E.M., Ostareck-Lederer, A. & Fischer, U. (2001) Evidence that fragile X mental retardation protein is a negative regulator of translation. *Hum Mol Genet,* 10, 329-38.

Lampo, E., Van Poucke, M., Vandesompele, J., Erkens, T., Van Zeveren, A. & Peelman, L.J. (2009) Positive correlation between relative mRNA expression of PRNP and SPRN in cerebral and cerebellar cortex of sheep. *Mol Cell Probes,* 23, 60-4.

Li, Z., Zhang, Y., Ku, L., Wilkinson, K.D., Warren, S.T. & Feng, Y. (2001) The fragile X mental retardation protein inhibits translation via interacting with mRNA. *Nucleic Acids Res,* 29, 2276-83.

Liao, Y.C., Lebo, R.V., Clawson, G.A. & Smuckler, E.A. (1986) Human prion protein cDNA: molecular cloning, chromosomal mapping, and biological implications. *Science,* 233, 364-7.

Lima, L.M., Cordeiro, Y., Tinoco, L.W., Marques, A.F., Oliveira, C.L., Sampath, S., Kodali, R., Choi, G., Foguel, D., Torriani, I., Caughey, B. & Silva, J.L. (2006) Structural insights into the interaction between prion protein and nucleic acid. *Biochemistry,* 45, 9180-7.

Locht, C., Chesebro, B., Race, R. & Keith, J.M. (1986) Molecular cloning and complete sequence of prion protein cDNA from mouse brain infected with the scrapie agent. *Proc Natl Acad Sci U S A,* 83, 6372-6.

Manuelidis, L. (2003) Transmissible encephalopathies: speculations and realities. *Viral Immunol,* 16, 123-39.

Manuelidis, L. (2007) A 25 nm virion is the likely cause of transmissible spongiform encephalopathies. *J Cell Biochem,* 100, 897-915.

Manuelidis, L., Liu, Y. & Mullins, B. (2009) Strain-specific viral properties of variant Creutzfeldt-Jakob disease (vCJD) are encoded by the agent and not by host prion protein. *J Cell Biochem,* 106, 220-31.

Manuelidis, L., Sklaviadis, T., Akowitz, A. & Fritch, W. (1995) Viral particles are required for infection in neurodegenerative Creutzfeldt-Jakob disease. *Proc Natl Acad Sci U S A,* 92, 5124-8.

Manuelidis, L., Sklaviadis, T. & Manuelidis, E.E. (1987) Evidence suggesting that PrP is not the infectious agent in Creutzfeldt-Jakob disease. *Embo J,* 6, 341-7.

Manuelidis, L., Yu, Z.X., Barquero, N. & Mullins, B. (2007) Cells infected with scrapie and Creutzfeldt-Jakob disease agents produce intracellular 25-nm virus-like particles. *Proc Natl Acad Sci U S A,* 104, 1965-70.

Mazroui, R., Huot, M.E., Tremblay, S., Filion, C., Labelle, Y. & Khandjian, E.W. (2002) Trapping of messenger RNA by Fragile X Mental Retardation protein into cytoplasmic granules induces translation repression. *Hum Mol Genet,* 11, 3007-17.

Menon, L. & Mihailescu, M.R. (2007) Interactions of the G quartet forming semaphorin 3F RNA with the RGG box domain of the fragile X protein family. *Nucleic Acids Res,* 35, 5379-92.

Mercey, R., Lantier, I., Maurel, M.C., Grosclaude, J., Lantier, F. & Marc, D. (2006) Fast, reversible interaction of prion protein with RNA aptamers containing specific sequence patterns. *Arch Virol,* 151, 2197-214.

Miyazawa, K. & Manuelidis, L. (2010) Agent-specific Shadoo responses in transmissible encephalopathies. *J Neuroimmune Pharmacol,* 5, 155-63.

Moore, R.C., Lee, I.Y., Silverman, G.L., Harrison, P.M., Strome, R., Heinrich, C., Karunaratne, A., Pasternak, S.H., Chishti, M.A., Liang, Y., Mastrangelo, P., Wang, K., Smit, A.F., Katamine, S., Carlson, G.A., Cohen, F.E., Prusiner, S.B., Melton, D.W., Tremblay, P., Hood, L.E. & Westaway, D. (1999) Ataxia in prion protein (PrP)-deficient mice is associated with upregulation of the novel PrP-like protein doppel. *J Mol Biol,* 292, 797-817.

Mouillet-Richard, S., Ermonval, M., Chebassier, C., Laplanche, J.L., Lehmann, S., Launay, J.M. & Kellermann, O. (2000) Signal transduction through prion protein. *Science,* 289, 1925-8.

Moya, K.L., Hassig, R., Breen, K.C., Volland, H. & Di Giamberardino, L. (2005) Axonal transport of the cellular prion protein is increased during axon regeneration. *J Neurochem,* 92, 1044-53.

Moya, K.L., Sales, N., Hassig, R., Creminon, C., Grassi, J. & Di Giamberardino, L. (2000) Immunolocalization of the cellular prion protein in normal brain. *Microsc Res Tech,* 50, 58-65.

Moyzis, R.K., Buckingham, J.M., Cram, L.S., Dani, M., Deaven, L.L., Jones, M.D., Meyne, J., Ratliff, R.L. & Wu, J.R. (1988) A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proc Natl Acad Sci U S A,* 85, 6622-6.

Murdoch, G.H., Sklaviadis, T., Manuelidis, E.E. & Manuelidis, L. (1990) Potential retroviral RNAs in Creutzfeldt-Jakob disease. *J Virol,* 64, 1477-86.

Nandi, P.K. & Leclerc, E. (1999) Polymerization of murine recombinant prion protein in nucleic acid solution. *Arch Virol,* 144, 1751-63.

Neidle, S. & Read, M.A. (2000) G-quadruplexes as therapeutic targets. *Biopolymers,* 56, 195-208.

Oesch, B., Teplow, D.B., Stahl, N., Serban, D., Hood, L.E. & Prusiner, S.B. (1990) Identification of cellular proteins binding to the scrapie prion protein. *Biochemistry,* 29, 5848-55.

Oganesian, L. & Bryan, T.M. (2007) Physiological relevance of telomeric G-quadruplex formation: a potential drug target. *Bioessays,* 29, 155-65.

Okur, A., Wickstrom, L., Layten, M., Geney, R., Song, K., Hornak, V. & Simmerling, C. (2006) Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model. *Journal of Chemical Theory and Computation,* 2, 420-433.

Oliver, A.W., Bogdarina, I., Schroeder, E., Taylor, I.A. & Kneale, G.G. (2000) Preferential binding of fd gene 5 protein to tetraplex nucleic acid structures. *J Mol Biol,* 301, 575-84.

Premzl, M., Sangiorgio, L., Strumbo, B., Marshall Graves, J.A., Simonic, T. & Gready, J.E. (2003) Shadoo, a new protein highly conserved from fish to mammals and with similarity to prion protein. *Gene,* 314, 89-102.

Proske, D., Gilch, S., Wopfner, F., Schatzl, H.M., Winnacker, E.L. & Famulok, M. (2002) Prion-protein-specific aptamer reduces PrPSc formation. *Chembiochem,* 3, 717-25.

Prusiner, S.B. (1982) Novel proteinaceous infectious particles cause scrapie. *Science,* 216, 136-44.

Prusiner, S.B. (1991) Molecular biology of prion diseases. *Science,* 252, 1515-22.

Prusiner, S.B. (1998) Prions. *Proc Natl Acad Sci U S A,* 95, 13363-83.

Qin, K., O'donnell, M. & Zhao, R.Y. (2006) Doppel: more rival than double to prion. *Neuroscience,* 141, 1-8.

Qin, Y., Rezler, E.M., Gokhale, V., Sun, D. & Hurley, L.H. (2007) Characterization of the G-quadruplexes in the duplex nuclease hypersensitive element of the PDGF-A promoter and modulation of PDGF-A promoter activity by TMPyP4. *Nucleic Acids Res,* 35, 7698-713.

Radulescu, R.T. & Korth, C. (1996) Prion function and dysfunction: a structure-based scenario. *Med Hypotheses,* 46, 225-8.

Ramos, A., Hollingworth, D. & Pastore, A. (2003) G-quartet-dependent recognition between the FMRP RGG box and RNA. *Rna,* 9, 1198-207.

Rhie, A., Kirby, L., Sayer, N., Wellesley, R., Disterer, P., Sylvester, I., Gill, A., Hope, J., James, W. & Tahiri-Alaoui, A. (2003) Characterization of 2'-fluoro-RNA aptamers that bind preferentially to disease-associated conformations of prion protein and inhibit conversion. *J Biol Chem,* 278, 39697-705.

Riek, R., Hornemann, S., Wider, G., Billeter, M., Glockshuber, R. & Wuthrich, K. (1996) NMR structure of the mouse prion protein domain PrP(121-321). *Nature,* 382, 180-2.

Saborio, G.P., Permanne, B. & Soto, C. (2001) Sensitive detection of pathological prion protein by cyclic amplification of protein misfolding. *Nature,* 411, 810-3.

Sales, N., Hassig, R., Rodolfo, K., Di Giamberardino, L., Traiffort, E., Ruat, M., Fretier, P. & Moya, K.L. (2002) Developmental expression of the cellular prion protein in elongating axons. *Eur J Neurosci,* 15, 1163-77.

Sales, N., Rodolfo, K., Hassig, R., Faucheux, B., Di Giamberardino, L. & Moya, K.L. (1998) Cellular prion protein localization in rodent and primate brain. *Eur J Neurosci,* 10, 2464-71.

Santuccione, A., Sytnyk, V., Leshchyns'ka, I. & Schachner, M. (2005) Prion protein recruits its neuronal receptor NCAM to lipid rafts to activate p59fyn and to enhance neurite outgrowth. *J Cell Biol,* 169, 341-54.

Schaeffer, C., Bardoni, B., Mandel, J.L., Ehresmann, B., Ehresmann, C. & Moine, H. (2001) The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. *Embo J,* 20, 4803-13.

Schmitt-Ulms, G., Legname, G., Baldwin, M.A., Ball, H.L., Bradon, N., Bosque, P.J., Crossin, K.L., Edelman, G.M., Dearmond, S.J., Cohen, F.E. & Prusiner, S.B. (2001) Binding of neural cell adhesion molecules (N-CAMs) to the cellular prion protein. *J Mol Biol,* 314, 1209-25.

Sears, R.C. (2004) The life cycle of C-myc: from synthesis to degradation. *Cell Cycle,* 3, 1133-7.

Secombe, J., Pierce, S.B. & Eisenman, R.N. (2004) Myc: a weapon of mass destruction. *Cell,* 117, 153-6.

Shaked, G.M., Meiner, Z., Avraham, I., Taraboulos, A. & Gabizon, R. (2001) Reconstitution of prion infectivity from solubilized protease-resistant PrP and nonprotein components of prion rods. *J Biol Chem,* 276, 14324-8.

Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. & Hurley, L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A,* 99, 11593-8.

Sigurdsson, B. (1954) Observations on three slow infections of sheep, Maedi, Paratuberculosis, Rida, a chronic encephalitis of sheep with general remarks on infections, which develop slowly, and some of their special characteristics. *British Veterinary Journal,* 110.

Silva, J.L., Lima, L.M., Foguel, D. & Cordeiro, Y. (2008) Intriguing nucleic-acid-binding features of mammalian prion protein. *Trends Biochem Sci,* 33, 132-40.

Simonsson, T. (2001) G-quadruplex DNA structures--variations on a theme. *Biol Chem,* 382, 621-8.

Siomi, H., Siomi, M.C., Nussbaum, R.L. & Dreyfuss, G. (1993) The protein product of the fragile X gene, FMR1, has characteristics of an RNA-binding protein. *Cell,* 74, 291-8.

Siomi, M.C., Zhang, Y., Siomi, H. & Dreyfuss, G. (1996) Specific sequences in the fragile X syndrome protein FMR1 and the FXR proteins mediate their binding to 60S ribosomal subunits and the interactions among them. *Mol Cell Biol,* 16, 3825-32.

Solforosi, L., Criado, J.R., Mcgavern, D.B., Wirz, S., Sanchez-Alavez, M., Sugama, S., Degiorgio, L.A., Volpe, B.T., Wiseman, E., Abalos, G., Masliah, E., Gilden, D., Oldstone, M.B., Conti, B. & Williamson, R.A. (2004) Cross-linking cellular prion protein triggers neuronal apoptosis in vivo. *Science,* 303, 1514-6.

Steele, A.D., Emsley, J.G., Ozdinler, P.H., Lindquist, S. & Macklis, J.D. (2006) Prion protein (PrPc) positively regulates neural precursor proliferation during developmental and adult mammalian neurogenesis. *Proc Natl Acad Sci U S A,* 103, 3416-21.

Telling, G.C., Scott, M., Mastrianni, J., Gabizon, R., Torchia, M., Cohen, F.E., Dearmond, S.J. & Prusiner, S.B. (1995) Prion propagation in mice expressing human and chimeric PrP transgenes implicates the interaction of cellular PrP with another protein. *Cell,* 83, 79-90.

Tobler, I., Deboer, T. & Fischer, M. (1997) Sleep and sleep regulation in normal and prion protein-deficient mice. *J Neurosci,* 17, 1869-79.

Walz, R., Amaral, O.B., Rockenbach, I.C., Roesler, R., Izquierdo, I., Cavalheiro, E.A., Martins, V.R. & Brentani, R.R. (1999) Increased sensitivity to seizures in mice lacking cellular prion protein. *Epilepsia,* 40, 1679-82.

Wang, F., Wang, X., Yuan, C.G. & Ma, J. (2010) Generating a prion with bacterially expressed recombinant prion protein. *Science,* 327, 1132-5.

Watts, J.C., Drisaldi, B., Ng, V., Yang, J., Strome, B., Horne, P., Sy, M.S., Yoong, L., Young, R., Mastrangelo, P., Bergeron, C., Fraser, P.E., Carlson, G.A., Mount, H.T., Schmitt-Ulms, G. & Westaway, D. (2007) The CNS glycoprotein Shadoo has PrP(C)-like protective properties and displays reduced levels in prion infections. *Embo J,* 26, 4038-50.

Watts, J.C. & Westaway, D. (2007) The prion protein family: diversity, rivalry, and dysfunction. *Biochim Biophys Acta,* 1772, 654-72.

Weiss, S., Proske, D., Neumann, M., Groschup, M.H., Kretzschmar, H.A., Famulok, M. & Winnacker, E.L. (1997) RNA aptamers specifically interact with the prion protein PrP. *J Virol,* 71, 8790-7.

Weissmann, C. (1991) A 'unified theory' of prion propagation. *Nature,* 352, 679-83.

Williamson, J.R., Raghuraman, M.K. & Cech, T.R. (1989) Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell,* 59, 871-80.

Young, R., Passet, B., Vilotte, M., Cribiu, E.P., Beringue, V., Le Provost, F., Laude, H. & Vilotte, J.L. (2009) The prion or the related Shadoo protein is required for early mouse embryogenesis. *FEBS Lett,* 583, 3296-300.

Zahn, R., Liu, A., Luhrs, T., Riek, R., Von Schroetter, C., Lopez Garcia, F., Billeter, M., Calzolai, L., Wider, G. & Wuthrich, K. (2000) NMR solution structure of the human prion protein. *Proc Natl Acad Sci U S A,* 97, 145-50.

Zanotti, K.J., Lackey, P.E., Evans, G.L. & Mihailescu, M.R. (2006) Thermodynamics of the fragile X mental retardation protein RGG box interactions with G quartet forming RNA. *Biochemistry,* 45, 8319-30.

Zeiler, B., Adler, V., Kryukov, V. & Grossman, A. (2003) Concentration and removal of prion proteins from biological solutions. *Biotechnol Appl Biochem,* 37, 173-82.

CHAPTER 2

UNCOVERING THE RGG BOX IN SHADOO

2.1 INTRODUCTION

2.1.1 Background

Over the past two decades the amount of available genomic sequence data has increased hugely, providing the potential to compare genomic sequences as they have evolved over millions of years (Lander et al., 2001, Aparicio et al., 2002, Waterston et al., 2002, Zdobnov et al., 2002, Kellis et al., 2003, Thomas et al., 2003, Consortium, 2004, Consortium, 2005). This has led to a better understanding of the features that both distinguish and are common to the species in our environment. Comparative genomics has been embraced as a powerful tool (O'Brien et al., 1999, Hedges and Kumar, 2002) for investigating the importance of particular protein domains or motifs (Bork et al., 1998, Ponting et al., 2000, Copley et al., 2003). A central premise of this approach is that functional DNA/protein sequences should be conserved over time and shared among closely related species whereas nonfunctional or neutral sequences are free to change (Fay and Wu, 2003).

In its most simplistic application, an alignment of protein sequences from species spanning a significant evolutionary period will reveal conserved domains or motifs that may have functional importance. An alignment of 5 Sho sequences (human, mouse, rat, zebra fish and puffer fish) has previously been used to demonstrate conservation in particular regions of the Sho protein (Premzl et al., 2003).

2.1.2 Aims of this chapter

The initial aim of this study was to perform a comparative analysis of the Sho protein using all available Sho sequences (from fish to human) to identify areas of potential functional significance. This led to my observation that the N-terminus of Sho has a well conserved region of RGG repeats that conforms to the RGG box motif, a well-known RNA binding sequence. The second aim follows from the first, and has been to mine existing data repositories in order to compare the Sho RGG region to the RGG box domains of other proteins. The goal of this endeavour has been to gain preliminary insights into the potential significance of the RGG domain in Sho.

METHODS

The amino acid sequences of 14 Sho proteins ranging from fish to human were used in this study. Twelve of these sequences are available from GenBank [*Homo sapiens* Np_001012526, *Pan troglodytes* XP_001146049, *Mus musculus* NP_898970, *Rattus norvegicus* NP_001027015, *Canis lupus familiaris* CAJ43798, *Bos taurus* CAJ43799, *Monodelphis domestica* CAF43800, *Gallus gallus* CAJ43796, *Xenopus tropicalis* CAJ43801, *Danio rerio* CAD35503, *Takifugu rubripes* CAG34291, *Tetraodon nigroviridis* CAG30521]. The sequences for *Ornithorhynchus anatinus* (platypus), *M. domestica* (American opossum), *G. gallus* (chicken), and *X. tropicalis* were initially extracted from the genomic databases (N. Chakka, unpublished work of this research group). The sequences for *X. tropicalis* and *X. laevis* were also verified experimentally (T. Vassilieva and N. Chakka, unpublished work of this research group). The sequences were aligned using ClustalW (Chenna et al., 2003). Subsequent manual adjustments in the N-terminal region were made to the alignment.

The Swiss-Prot protein database was searched using the program Prosite (Hofmann et al., 1999) http://au.expasy.org/ for known motifs within the Sho sequences. The Swiss-Prot data base was also searched for all proteins that have an RGG box motif, which was defined as a sequence of at least 3 RGG repeats with no more than 6 residues between the repeats. This search produced 10 archaeal, 229 bacterial, 14 viral and 1632 eukaryotic sequences, within which there are 607 fungal, 300 plant and 70 human sequences. Examination of the human sequences showed that some well-known RGG-box proteins had not been picked up by this search. The search was then broadened to include proteins with 2 RGG repeats separated by 9, 8, 7, 6 or 5 residues. The results were visually inspected and those proteins with at least one 'RG' between the RGG repeats were retained. All uncharacterized proteins or redundant sequences were excluded and the remaining human proteins are collected in Table 2-1. Only the sequence beginning and ending with an RGG repeat is presented, however, the functional RGG box may extend beyond this. The RGG sequences were subsequently aligned using ClustalW.

2.2 RESULTS AND DISCUSSION

2.3.1 Sho Sequence Alignment

A sequence alignment of the N-terminal segment from residue 25 to 42 (the mature protein starts at residue 24) of Shos from different species, spanning an evolutionary period of 450 million years (Aparicio et al., 2002), is presented in Figure 2-1. This Figure highlights a stringently conserved 'GGRGG' sequence of residues at the N-terminus. This motif is a recognized arginine methylation site (Lee and Bedford, 2002), commonly found in Arg rich proteins. Looking more closely at the Sho sequence alignment, it is apparent that the GGRGG sequence is followed by a sequence of RGG repeats that are also well conserved throughout evolution. These sequences adhere to the RGG box format, suggesting that Sho may belong to the RGG box class of RNA-binding proteins.

In Shos from human and the other Eutherian mammals there are three RGG repeats, with the first and third separated by 9 residues (**RGG**ARGSA**RGG**V**RGG**). The RGG box of human Sho consists of 15 residues: 4 positively charged Arg residues, 7 Gly residues - which are expected to confer a large degree of flexibility, and 4 small intervening residues. This pattern diverges slightly in other species; the first RGG repeat is conserved but the second and third RGG repeats are sometimes truncated to RG. Overall, although there is some variability in the number of Gly residues in the Sho RGG box, all species from fish to human exhibit conservation of Lys25 and the following 3 Arg residues which are each separated by 3 intervening residues. The increased prevalence of Gly-Gly dipeptides in the higher Eutherian mammals could suggest evolutionary pressure for increased flexibility in this domain.

Species	Start	End	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
Human	25	42	K	G	G	R	G	G	A	R	G	S	A	R	G	G	V	R	G	G
Chimp	25	42	K	G	G	R	G	G	A	R	G	S	A	R	G	G	V	R	G	G
Mouse	25	42	K	G	G	R	G	G	A	R	G	S	A	R	G		V	R	G	G
Rat	25	42	K	G	G	R	G	G	A	R	G	S	A	R	G		V	R	G	G
Dog	25	42	K	G	G	R	G	G	A	R	G	S	A	R	G	G	L	R	G	G
Bovine	25	42	K	G	G	R	G	G	A	R	G	S	A	R	G	G	R	G	A	A
Possum (Mdl)	25	42	K	G	G	R	G	G	A	R	G	A	A	R	G	R	S	R	S	S
Platypus	25	42	K	G	G	R	G	G	A	R	G	A	A	R	G	A	A	R	G	A
Chicken	25	42	K	G	G	R	G	G	S	R	G	A	A	R	G	M	A	R	G	A
Frog (Xl)	25	42	K	G	G	R	G	G	A	R	G	G	A	R	G	S	S	R	G	T
Frog (Xt)	25	42	K	G	G	R	G	G	A	R	G	G	A	R	G	A	S	R	G	A
Fish (Danio)	25	42	K	G	G	R	G	G	A	R	G	S	A	R	G	T	A	R	G	G
Fish (Fugu)	25	42	K	G	G	R	G	G	S	R	G	S	S	R	G	S	P	S	R	S
Fish (Tetraodon)	26	43	K	G	G	R	G	G	S	R	G	S	S	R	G	S	P	S	R	S

Figure 2-1. Sho Sequence Alignment
Alignment of the RGG-box sequence at the N-terminal end of Shos from fish to mammals; Md, *Monodelphis domestica*; Xl, *Xenopus laevis*; Xt, *Xenopus tropicalis*; Zebra, *Danio rerio*; Fugu, *Fugu rubripes*, Tetraodon, *Tetraodon nigroviridis*.

The sequence displayed in Figure 2-1 appears to fit the description of the RGG box motif which has been defined as a sequence of closely spaced Arg-Gly-Gly (RGG) repeats interspersed with other, often aromatic, amino acids (Kiledjian and Dreyfuss, 1992) . The RGG box proteins are one of the classes of RNA-binding proteins (RBPs) involved in various aspects of RNA processing, including splicing, stabilizing, transport and translation of mRNAs (Burd and Dreyfuss, 1994). In addition to being an RNA-binding motif, the RGG box of some proteins is known to mediate interactions with other proteins (Christensen and Fuxa, 1988, Cobianchi et al., 1988, Bouvet et al., 1998, Said et al., 2002, Lukasiewicz et al., 2007).

The human Sho RGG sequence is now compared to other human RGG sequences derived from 44 human proteins listed in Table 2-1.

2.3.2 Comparative Analysis of RGG-box proteins – Structure and Composition

Proteins with an RGG-box motif, as defined for the purpose of this study (Methods), are presented in Table 2-1. Most (#2-#34) are known to have an RNA-binding function.

Table 2-1. Survey of RGG box proteins

	Database name, name (ID)	RGG domain (residue numbers)	Other RNA binding motifs[a]	Function/Comments
1	**SHO_HUMAN**, Shadoo (Q5BIV9)	RGGARGSARGGVRGG (28-42)		PrP family member. Likely attached to cell membrane by GPI anchor (Premzl et al.)
2	**ROA0_HUMAN**, hnRNP A0 (Q13151)	RGGNFSGRGGFGGSRGG (192-202)	2 RRM[b]	Found in splicesome C; expected involvement in splicing, pre-mRNA processing. Similar to hnRNP A/B but less abundant. Component of ribonucleosomes (Myer and Steitz, 1995).
3	**ROA1_HUMAN**, hnRNP A1 (P09651)	RGGNFSGRGGFGGSRGG (218-234)	2 RRM	Found in splicesome C; expected involvement in splicing, pre-mRNA processing. Transport of poly(A) mRNA from nucleus to cytoplasm (Biamonti et al., 1989, Siomi and Dreyfuss, 1995, Jurica et al., 2002).
4	**ROA2_HUMAN**, hnRNP A2 (P22626)	RGGNFGFGDSRGGGGNF GPGPGSNFRGG (203-230)	2 RRM	Found in splicesome C; expected involvement in splicing, pre-mRNA processing. Trafficking of RNAs containing the cis-acting A2 response element (A2RE) (Jurica et al., 2002).

5	**ROA3_HUMAN**, hnRNP A3 (P51991)	RGGGSGNFMGRGGNFGG GGGNFGRGG (216-241)	2 RRM	Found in splicesome C; expected involvement in splicing (Jurica et al., 2002), pre-mRNA processing. Trafficking of RNAs containing the cis-acting A2 response element (A2RE) (Jurica et al., 2002).
6	**HNRPD_HUMAN**, hnRNP D0 (Q14103)	RGGFAGRARGRGG (272-284)	2 RRM	Binds to mRNA with AU-rich elements (AREs) in 3'-UTR. Transcription regulator; binds to ds/ssDNA sequences. Possibly Involved in translationally coupled mRNA turnover (Tay et al., 1992, Kajita et al., 1995).
7	**HNRPG_HUMAN**, hnRNP G (P38159)	RGGSGGTRGPPSRGG (113-126)	1 RRM	Found in splicesome C; expected involvement in splicing, pre-mRNA processing (Soulard et al., 1993, Jurica et al., 2002).
8		RGGGRGGSRSDRGG (373-386)		
9	**HNRPK**, hnRNP K (P61978)	RGGFDRMPPGRGGRPMP PSRRDYDDMSPRRGPPPP PPGRGGRGGSRARNLPLP PPPPPRGG (258-318)	3 KH[c]	Found in splicesome C, as above. Major poly(C) RNA binding hnRNP. Also binds poly(C) ssDNA. (Jurica et al., 2002; Matunis et al., 1992)
10	**HNRPQ_HUMAN**, hnRNP Q (O60506)	RGGPGSARGVRGARGGA QQQRGRGVRGARGGRGG (526-559)	3 RRM	3 isoforms. Pre-mRNA processing. Associated with splicing intermediates and mature mRNA. Interacts preferentially with poly(A) and poly (U) RNA sequences. Region 518-549 (RGRAGYSQRGGPGSARGVRGARG GAQQQRGRG) sufficient to bind RNA (Mourelatos et al., 2001).
11	**HNRPR_HUMAN**, hnRNP R (O43390)	RGGRGGPAQQQRGRGSR GSRGNRGG (543-567)	3 RRM	Found in splicesome C; expected involvement in splicing, pre-mRNA processing (Hassfeld et al., 1998, Jurica et al., 2002).
12	**HNRPU_HUMAN**, hnRNP U (Q00839)	RGGGHRGRGGFNMRGGN FRGGAPGNRGG (701-728)		RGG box motif first identified in this protein. Expected involvement in splicing, pre-mRNA processing. Stabilizes specific mRNAs. Aka SAF-A as it has high affinity for DNA with scaffold attachment regions (SAR). Found in splicesome C; expected involvement in splicing, pre-mRNA processing (Kiledjian and Dreyfuss, 1992, Fackelmayer and Richter, 1994, Jurica et al., 2002, Helbig and Fackelmayer, 2003, Yugami et al., 2007).
13	**HNRL1**, hnRNP U like protein 1 (Q9BUJ2)	RGGGGFRGRGGGGGFQR YENRGPPGGNRGGFQNR GGGSGGGGNYRGG (612-658)		Pre-mRNA processing and transport. Binds poly(G) and poly(C) RNA. Represses transcription driven by viral and cellular promoters. Associated with RBRD7, activates transcription (Gabler et al., 1998).
14	**PURG_HUMAN**, Purine-rich element-binding protein gamma (Q9UJV8)	RGGGGGRGRGG (7-17)		Found in purine-rich element binding protein family (PUR). Binds ssDNA and RNA. Highly expressed in many tumor lines (Liu and Johnson, 2002).

15	**DDX4_HUMAN**, DEAD box protein 4 (Q9NQI0)	RGGRGSFRGCRGG (147-159)		Found in DEAD box helicase family. Helicase activity, RNA unwinding, needed in splicing, ribosome biogenesis and RNA degradation (Luking et al., 1998, Castrillon et al., 2000).
16	**THOC4_HUMAN**, Tho complex subunit 4 (Q86V81)	RGGGAQAAARVNRGG (38-52)	1 RRM	In THO/TREX complex, promotes transcriptional activation, recruited to RNA polymerase during elongation. Associated with spliced mRNA; roles in mRNA export and decay. May mediate interactions of proteins and/or RNA (Virbasius et al., 1999, Strasser et al., 2002).
17	**NOLA1_HUMAN**, Nucleolar protein family A member 1 (Q9NY12)	RGGGRGGFNRGGGGGGFNRGGSSNHFRGGGGGGGGNFRGGGRGGFGRGGGRGG (4-57)		Aka GAR1. Required for ribosome biogenesis and telomere maintenance. Processing or intranuclear trafficking of TERC, the RNA component of the telomerase reverse transcriptase (TERT). RGG box accessory to RNA binding. Interaction with SMN1 requires at least one of the RGG-box regions (Bagni and Lapeyre, 1998, Whitehead et al., 2002).
18		RGGGRGGRGGGRGGGGRGGGRGGGFRGGRGGGGGFRGGRGG (169-210)		
19	**SFPQ_HUMAN**, Splicing factor proline- and glutamine-rich (P23246)	RGGGGGGFHRRGGGGGRGG (9-27)	2 RRM	Pre-mRNA splicing factor. Binds to intronic polypyrimidine tracts. Possible role in nuclear retention of defective RNAs. Regulates basal and cAMP-dependent transcription (Patton et al., 1993).
20	**FBRL_HUMAN**, Fibrillarin (P22087)	RGGGFGGRGGFGDRGGRGGRGGFGGGRGRGGGFRGRGRGG (8-47)		Involved in pre-rRNA processing (Aris and Blobel, 1991, Jansen et al., 1991).
21	**HABP4_HUMAN**, Hyaluronan binding protein (IHAPB4, Ki-1/57) (Q5JVS0)	RGGPRGGMRGRGRGG (185-199)		This sequence also constitutes a hyaluronan binding motif, (R/K–X(7)-R/K) where X is not acidic (Yang et al., 1994). Domain in IHAPB4 binds strongly and specifically to hyaluronan and weakly to RNA. Involved in mRNA transport, chromatin remodeling, regulation of transcription. Interacts with chromodomain DNA helicase binding protein 3(CHD3) (Kobarg et al., 1997, Lemos et al., 2003, Passos et al., 2006a)
22	**PAIRB_HUMAN**, Plasminogen activator inhibitor 1 RNA binding protein (Q8NC51)	RGGRGGRGGRGRGG (367-380)		Aka CG1-55. Regulation of mRNA stability/decay. Interacts with CHD3, similar to HABP4 (Lemos et al., 2003).
23	**FUS_HUMAN**, RNA binding protein FUS (P35637)	RGGGRGGRGGMGGSDRGG (244-261)	1 RRM	Component of nuclear riboprotein complexes. Binds ds/ssDNA.

23		RGGGRGGRGGMGGSDRGG (244-261)	1 RRM	al., 1993).
24		RGGGNGRGGRGRGGPMGRGG (377-396)		
25		RGGRGGYDRGGYRGRGGDRGGFRGGRGGGDRGG (473-505)		
26	**EWS_HUMAN**, Ewing sarcoma (EWS) protein (Q01844)	RGGFDRGGMSRGGRGGGRGGMGSAGERGG (304-332)	1 RRM	Found on cell surface as well as in the nucleus and cytoplasm. Is a transcriptional activator but this activity can be repressed by the RGG box. May be involved in pre-mRNA splicing and transport. May act as a receptor or binding protein for ligands on the cell surface, such as nucleic acids, and thus might mediate extracellular and nuclear events. Interacts with PTK2B/FAK2 then relocates from cytoplasm to ribosomes ((Plougastel et al., 1993, Ohno et al., 1994, Belyanskaya et al., 2001, Belyanskaya et al., 2003).
27		RGGPGGMRGGRGGLMDRGGPGGMFRGGRGGDRGGFRGGRGMDRGGFGGGRRGG (565-617)		
28	**RB56_HUMAN**, TATA-binding protein-associated factor 2N (Q92804)	RGGYRGRGGFQGRGG (337-351)	1 RRM	Binds RNA and ssDNA. Role in transcriptional regulation. Found in RNA polymerase II transcriptional multiprotein complex. Similar to EWS and FUS/TLS (Morohoshi et al., 1996).
29		RGGGYGGDRGGGYGGDRGGGYGGDRGGYGGDRGGGYGGDRGGYGGDRGGYGGDRGGYGGDRSRGGYGGDRGG (459-537)		
30	**CIRPB_HUMAN**, Cold-inducible RNA-binding protein (Q14011)	RGGSAGGRGFFRGGRGRGRGFSRGG (94-118)	1 RRM	Role in cold-induced suppression of cell proliferation. Activates the ERK pathway (Nishiyama et al., 1997).
31	**PP1RA_HUMAN**, Serine/threonine-protein phosphatase 1 regulatory subunit (Q96QC0)	RGGPGPGPGPYHRGRGGRGGNEPPPPPPPFRGARGGRSGGGPPNGRGG (693-740)		Aka p99. Binds mRNA, ssDNA, poly(A) and poly(G). Inhibits phosphatase activities when phosporylated (Kreivi et al., 1997, Totaro et al., 1998).
32	**FMR1_HUMAN**, Fragile X Mental Retardation Protein (FRMP) (Q06787)	RGGGGRGQGGRGRGG (534-548)	2 KH	Binds many mRNA transcripts. Transports mRNA from nucleus to cytoplasm. Involved in neural plasticity through translational repression (Darnell et al., 2001, Zalfa et al., 2003, Bagni and Greenough, 2005, Ule and Darnell, 2006).
33	**NUCL_HUMAN**, Nucleolin (P19338)	RGGGRGGFGGRGGGRGGRGGFGGRGRGGFGGRGGFRGGRGG (656-696)	4 RRM	Found on cell surface as well as in the nucleus and cytoplasm. RGG box necessary for efficient RNA binding and may unstack RNA bases, but the RRMs are required for specific RNA recognition. Duplex DNA, ssDNA and

				RNA are all effective ligands for nucleolin. Associated with intranucleolar chromatin and preribosomal particles. Binds to histone H1 to induce chromatin decondensation. When attached to the cell surface, nucleolin binds the proteins cytokine MK and HB-19 through its RGG box and acts as cell surface receptor . (Ghisolfi et al., 1992b, Said et al., 2002, Hirano et al., 2005)
34	**G3BP1_HUMAN**, Ras GTPase-activating protein-binding protein 1 (Q13283)	RGGLGGGMRGPPRGG (435-449)	1 RRM	Role in ras-signaling pathway affecting cell proliferation and survival as well as being involved in RNA metabolism. Cleaves MYC mRNA and has Helicase activity - unwinds DNA/DNA, RNA/DNA and RNA/RNA. Suggested that the G3BPs are members of a novel sub-class of RNA-binding proteins which act at the level of RNA metabolism in response to cell signaling allowing the cell to rapidly control protein activity at a stage after transcription. Also involved in formation of stress granules. (Kennedy et al., 2001, Tourriere et al., 2001, Tourriere et al., 2003, Irvine et al., 2004)
35	**RGMC_HUMAN**, Hemojuvelin (precursor) (Q6ZVN8)	RGGGSSGALRGGGGGGRGG (54-72)		Member of repulsive guidance molecule (RGM) family; RGMa and RGMb involved in neural development. GPI anchored. Interacts with neogenin which regulates shedding of GPI anchor. Binding cytokines BMP2 and BMP4 affects BMP signaling pathway and expression of hepcidin. Function of RGG domain unknown. (Matsunaga and Chedotal, 2004, Zhang et al., 2005, Zhang et al., 2007)
36	**ZNH14_HUMAN**, Zinc finger HIT domain-containing protein 4 (Q9C086)	RGGRGGARGERRGG (238-251)		Aka PAPA-1. Induces growth and cell cycle arrest at G1 phase. Interacts with splicing factors altering pre-mRNA splicing. Complexes with other nucleolar proteins. (Kuroda et al., 2004, Maita et al., 2004).
37	**K1C9_HUMAN**, Keratin type 1 cytoskeletal 9 (P35527)	RGGSGGSYGRGSRGG (478-492)		Cytoskeletal and microfibrillar keratin. Function in mature or developing palmar and plantar skin (Langbein et al., 1993).
38	**MRE11_HUMAN**, Double strand break repair protein MRE 11A (P49959)	RGGRGQNSASRGG (577-589)		Found in MRN complex, role in dsDNA repair, recombination, maintenance of telomere integrity and meiosis (Petrini et al., 1995).
39	**WBP7_HUMAN**, WW domain-binding protein 7 (Q9UMN6)	RGGQSSRGGRGGRGRGRGG (281-299)		WW domain-binding (Trithorax homolog 2). Possible transcriptional regulator (FitzGerald and Diaz, 1999).

40	**BRWD3_HUMAN**, Bromodomain and WD repeat-containing protein 3 (Q6RI45)	RGGGGTRGRGRGRGG (1699-1713)		Found in WD repeat protein family involved in cell-cycle progression, signal transduction, apoptosis, gene regulation. Possible transcription factor with 2 bromodomains and 9 WD repeats. May be involved in Jak/Stat pathway (Vodermaier, 2001).
41	**CA077_HUMAN**, Uncharacterised protein C1orf77 (Q9Y3Y2)	RGGVRGRGGPGRGG (153-166)		NA
42	**FA98A_HUMAN**, Protein FAM98A (Q8NCA5)	RGGHEQGGGRGGRGGYD HGGRGG (352-374)		NA
43	**FA98A_HUMAN** (Q8NCA5)	RGGGRGGRGGRGGRGG (458-473)		
44	**LS14A_HUMAN**, LSM14 protein homolog A (Q8ND56)	RGGYRGRGGLGFRGGRG RGGGRGG (406-429)		Lsm (like Sm RNA-binding protein) protein family are involved in RNA processing tasks (Albrecht and Lengauer, 2004). AKA FAM61A and RAP55, may repress translation (Tanaka et al., 2006).

[a] RNA-binding motifs in addition to the RGG box. [b] RRM (RNA Recognition Motif) a 80 - 90 amino acid sequence containing a RNP-1 (octapeptide) and RNP-2 (6 amino acid) consensus sequences.
[c] K homology region as in hnRNP K

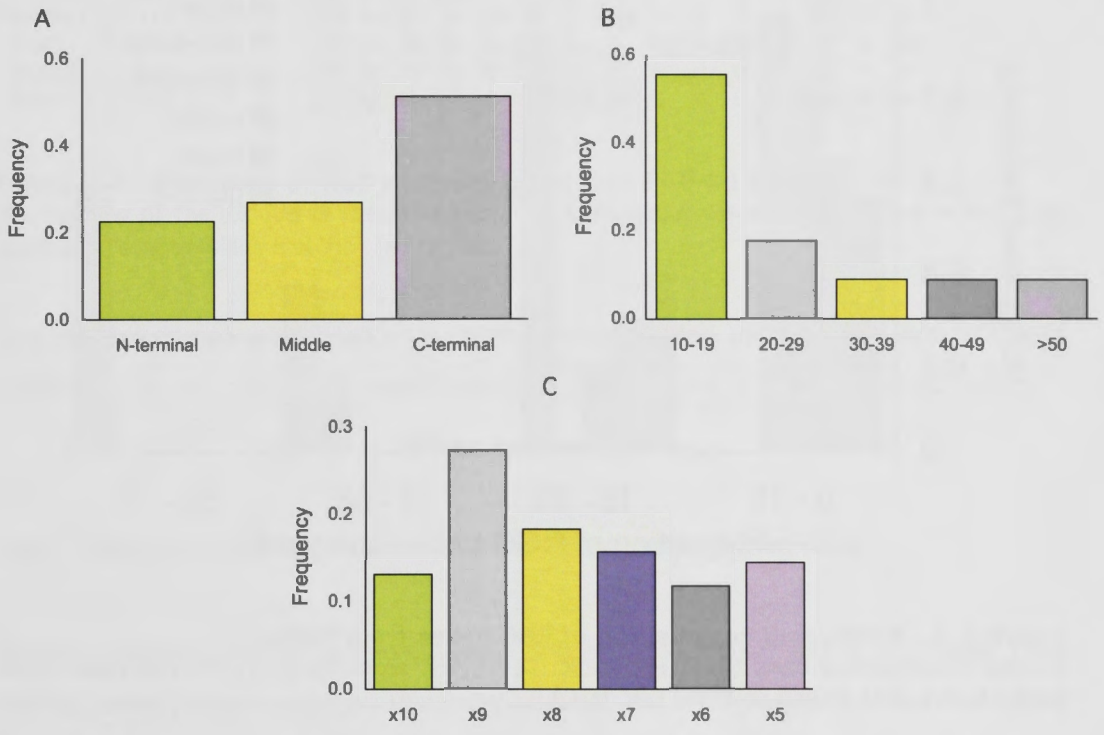


Figure 2-2. Structural and compositional features of RGG sequences in Table 2-1
Frequency histograms of (A) Position of the RGG box region in the proteins. N-terminus (first 35% of protein sequence), C-terminus (last 35%), Middle (rest). (B): Length of the RGG box. (C) Spacing between RGG repeats where x may be any residue including Arg and Gly.

Analysis of all the proteins listed in Table 2-1 reveals that the RGG box is generally found at the end of the protein sequence (Figure 2-2(A)), particularly at the C-terminus, and is mostly 10-19 residues in length (Figure 2-2(B)). There is a slight preference for RGG repeats to be separated by 9 intervening residues (RGG-X9-RGG), as in Sho, but overall the spacing is variable (Figure 2-2(C)).

The amino acid composition of the sequences was analysed by calculating the proportion of basic (Arg, Lys and His), acidic (Glu and Asp), aromatic (Phe, Trp and Tyr), polar (Ser, Thr, Asn, Gln and Cys), Gly and the other non-polar amino acids (Ala, Val, Leu, Ile, Met and Pro) which make up each sequence and then producing a frequency distribution for the entire set of proteins (Figure 2-3). As expected, the sequences are Gly rich, with peaks in frequency at 50-60% Gly composition while basic residues peak at 20-30% composition. Although a significant number of sequences do not contain an aromatic amino acid between the RGG repeats, it is possible that there are aromatic residues in close sequence or spatial proximity to this domain. Very few sequences contain acidic residues.

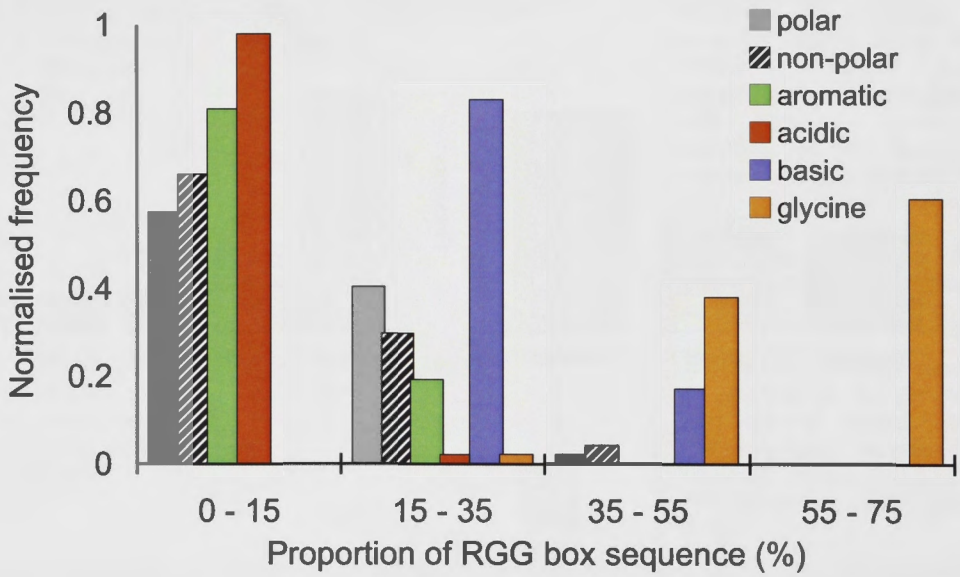


Figure 2-3. Amino acid composition of RGG boxes from Table 2-1
Residue classification: basic (R, K, H); acidic (D, E); Gly: aromatic (F, Y, W); non-polar amino acids (A, V, L, I, M, P) and polar (S, T, N, Q, C).

The Sho RGG sequence conforms to these general structural and compositional parameters. It is found at the end of the protein (N-terminus), is 15 residues long and is comprised of 47% Gly, 27% basic, 20% non-polar and 7% polar residues, and has no acidic or aromatic residues.

The first and last RGG repeat in Sho is separated by 9 intervening residues, similar to the Fragile X Mental Retardation Protein (FMRP) (#32, Table 2-1). The RGG sequence of Sho was aligned against other sequences with RGG-X9-RGG spacing in order to identify those most similar to Sho (Figure 2-4). As can be seen from Figure 2-4, several sequences have at least 50% sequence identity to those in the Sho RGG box.

Protein	Start	End	#	Sequence
Shadoo	28	42		**R G G** A **R G** S A **R G G** V **R G G**
HABP4	185	199	10	**R G G** P **R G G** M **R G R G R G G**
FUS	491	505	10	**R G G** F **R G G R G G** D **R G G**
K1C9	478	492	9	**R G G** S **G G** S Y **G R G** S **R G G**
G3BP1	435	449	9	**R G G** L **G G G** M **R G** P P **R G G**
LS14A	406	420	9	**R G G** Y **R G R G G** L **G** F **R G G**
NOLA1	173	187	9	**R G G R G G G R G G G G R G G**
NOLA1	185	199	9	**R G G G R G G G** F **R G G R G G**
EWS	309	323	9	**R G G** M S **R G G R G G G R G G**
NOLA1	30	44	8	**R G G G G G G G G G** N F **R G G**
FBRL	24	38	8	**R G G R G G** F **G G G R G R G G**
BRWD3	1699	1713	8	**R G G G G** T **R G R G R G R G G**
RBP56	337	351	8	**R G G** Y **R G R G G** F Q **G R G G**
THOC4	38	52	7	**R G G G** A Q A A A **R** V N **R G G**
FBRL	15	29	7	**R G G** F **G** D **R G G R G G R G G**
FA98A	352	366	7	**R G G** H E Q **G G G R G G R G G**
PP1RA	726	740	7	**R G G R** S **G G G** P P N **G R G G**
HNRPG	113	127	7	**R G G** S **G G** T **R G** P P S **R G G**
FMR1	534	548	7	**R G G G G R G** Q **G G R G R G G**
HNRPU	714	728	7	**R G G** N F **R G G** A P **G** N **R G G**
CA077	128	142	6	**R G G** M S L **R G G** N L L **R G G**

Figure 2-4. Alignment of RGG box sequences with RGG-X9-RGG spacing
The number of the residue at the start and end of the sequence is given, as well as the total number of exact residue matches (#) to Sho.

Overall, this comparative analysis supports the prediction that the RGG box of Sho is competent to bind RNA in a manner similar to the RGG box of other proteins.

2.3.3 Sho - predicted arginine methylation and phosphorylation sites

The Arg methylation site in Sho (GGRGG) is completely conserved in all species from fish to human, suggesting functional importance. Arg methylation is a common post-translational modification in RGG-box domains (Liu and Dreyfuss, 1995) which affects protein-protein interactions (Boisvert et al., 2005) and RNA binding (Dolzhanskaya et al., 2006). Arg methylation occurs in a range of proteins and as such affects diverse cellular processes, including cellular location of proteins. For example, the

nuclear/cytoplasmic distribution of hnRNP Q (#10, Table 2-1) (Passos et al., 2006b) and IHABP4 (#21) (Passos et al., 2006a) and the relocation of EWS protein (#26) to the cell surface (Belyanskaya et al., 2003) is changed by methylation. Methylation of certain identified RNA-binding proteins also affects transcription, processing and transport of mRNAs (Yu et al., 2004). Signaling pathways, in particular, the pathway which results in the cAMP response element-binding protein (CREB) dependent gene expression and the cytokine signaling pathway (JAK-STAT) (Boisvert et al., 2005) are also regulated in this manner.

Phosphorylation is another common post-translational modification found in RNA binding proteins. Three potential protein kinase C (PKC) phosphorylation sites (SAR (34-36 huSho), SLR (63-65 huSho) and SYR (119-121 huSho)) are identified for Sho as depicted in Figure 2-5.

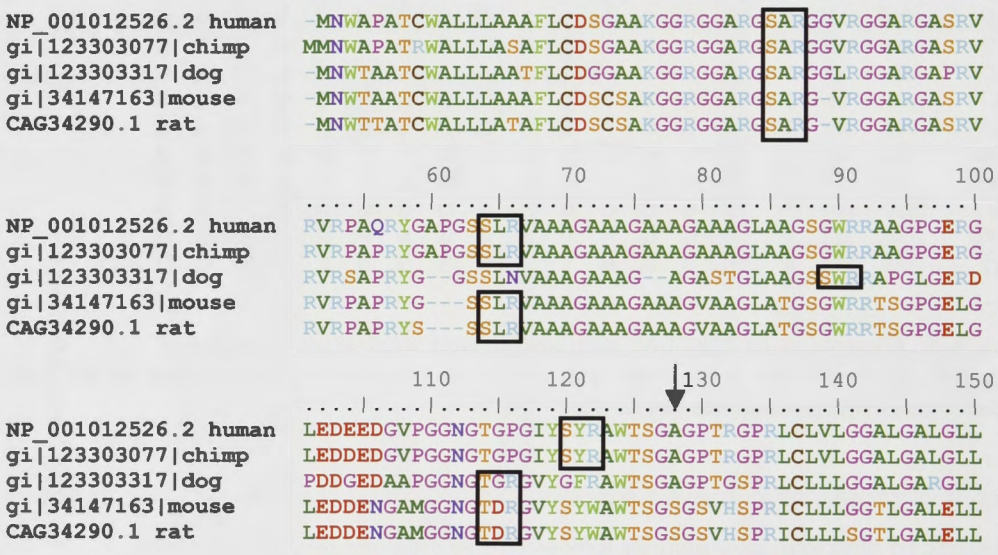


Figure 2-5. Sho sequence alignments
Alignment of Sho sequences for 5 Eutherian species (human, chimpanzee, dog, mouse and rat) with the 3 PKC phosphorylation sites indicated by boxes.

Methylation and phosphorylation mechanisms co-regulate a number of RGG box proteins, possibly including Sho. A good example of co-regulation of an RNA-binding protein through methylation and phosphorylation, is the mechanism by which phosphorylation affects the RGG box activity of the yeast protein Npl3p. This mechanism has recently been elucidated at a structural level (Lukasiewicz et al., 2007). Npl3p is an RNA-binding protein with 2 RRMs and an RGG box. It is thought to play a role in splicing and also transports mRNAs from the nucleus to the cytoplasm.

To enter the nucleus Npl3p must be phosphorylated in the cytoplasm by the kinase Sky1p. The RGG box of Npl3p binds to the docking site of the kinase allowing phosphorylation of Ser411 at the C-terminus of Npl3p. Once phosphorylated, Npl3p is able to bind to the nuclear import receptor Mtr10p and is translocated to the nucleus. By contrast, arginine methylation inhibits phosphorylation of Ser411 and consequently nuclear import. In the nucleus, Npl3p is dephosphorylated and methylated, which allows it to bind to the nuclear export factor Mex76p. Re-phosphorylation in the cytoplasm allows Npl3p to release its mRNA cargo and to commence a new cycle (Lukasiewicz et al., 2007).

One of the phosphorylation sites predicted for Sho, SAR34-36, is within the RGG box and is found in all the Eutherian mammal sequences analysed (Figure 2-1). The other 2 predicted sites are not as well conserved. Conservation of the SAR34-36 site and its proximity to an Arg methylation site suggest that this site may function as a phosphorylation site *in vivo*. Phosphorylation of Ser34 would have a direct affect on the structure of the RGG box and most likely affect its function. It is plausible that methylation of Arg28 would affect phosphorylation of Ser34 and vice versa, allowing co-regulation, as for the example provided above. The close proximity of phosphorylation sites to an RGG box which is likely to undergo arginine methylation has also been reported in other proteins, including FMRP (Mazroui et al., 2003), hnRNP A1 (Cobianchi et al., 1993) and IHABP4 (Passos et al., 2006a).

Although the phosphorylation site motifs are patterns with a high probability of occurrence it is interesting to note that the presence of at least one phosphorylation site has been experimentally confirmed in 70% of the RGG box proteins surveyed (Table 2-2). This is a high proportion even taking into account the over-representation of nuclear proteins in the phosphoproteome (Olsen et al., 2006) and suggests that phosphorylation is particularly prevalent in RGG-box proteins. The finding of potential methylation and phosphorylation sites in Sho is another point of similarity with other RGG box proteins. The existence of phosphorylation sites within Sho raises the possibility that Sho may be involved in a signaling pathway that is regulated by phosphorylation.

Table 2-2. Number of phosphorylation sites in RGG box proteins surveyed [a]

Protein	Id	PKC[b]	CK2[c]	TYR[d]	Expt[e]
SHO	Q5BIV9	3	0	0	
ROA0	Q13151	4	3	0	2
ROA1	P09651	10	10	0	9
ROA2	P22626	9	4	0	6
ROA3	P51991	9	7	0	6
HNRPD	Q14103	9	6	1	7
HNRPG	P38159	18	16	1	7
HNRPK	P61978	7	12	1	6
HNRPQ	O60506	7	3	2	2
HNRPR	O43390	5	5	2	
HNRPU	Q00839	10	5	0	5
HNRL1	Q9BUJ2	6	9	0	3
PURG	Q9UJV8	6	2	0	1
DDX4	Q9NQI0	16	14	0	
THOC4	Q86V81	4	5	0	1
NOLA1	Q9NY12	3	1	0	
SFPQ	P23246	7	4	2	1
FBRL_	P22087	5	3	0	
HABP4	Q5JVS0	5	8	2	2
PAIRB	Q8NC51	6	10	0	12
FUS	P35637	7	6	1	
EWS	Q01844	4	5	0	
RB56	Q92804	7	12	2	1
CIRPB	Q14011	4	2	1	
PP1RA	Q96QC0	10	12	2	4
FMR1	Q06787	9	12	1	1[f]
NUCL	P19338	8	23	0	14
G3BP1	Q13283	2	6	1	5
RGMC	Q6ZVN8	11	2	0	
ZNH14	Q9C086	3	0	0	
K1C9	P35527	7	14	3	
MRE11	P49959	16	17	0	4
WBP7	Q9UMN6	40	36	3	4
BRWD3	Q6RI45	34	42	4	4
CA077	Q9Y3Y2	4	2	0	1
FA98A	Q8NCA5	5	9	1	
LS14A	Q8ND56	4	7	1	11

[a] Searches were conducted using the ScanProsite program available on the ExPASy Proteomics Server of the Swiss Institute of Bioinformatics website http://au.expasy.org/. [b] Number of protein kinase C phosphorylation sites (PS00005).[c] Number of casein kinase II phosphorylation sites (PS00006).[d] Number of tyrosine kinase phosphorylation sites (PS00007).[e] Experimentally confirmed sites as annotated in the SwissProt database. [f] (Mazroui et al., 2003).

2.4 FUNCTIONAL IMPLICATIONS

2.4.1 Functional diversity of the RGG box proteins

The most prominent category of proteins identified in Table 2-1 comprises the hnRNPs which shuttle between the nucleus and cytoplasm and are involved in pre-mRNA processing, splicing and RNA transport (Krecic and Swanson, 1999). Otherwise, the proteins identified in Table 2-1 are quite functionally diverse as demonstrated by 6 examples, Nucleolin ((#33 Table 2-1), the Ewing Sarcoma (EWS) protein (#26), FMRP(#32), IHABP4(#21), G3PB1(#34) and RGMc (#35). Four of these proteins,

FMRP, IHABP4, EWS and G3PB1 have an RGG box with RGG-X9-RGG spacing similar to Sho while RGMc is the only other protein with a GPI anchor.

Nucleolin (#33) is a multifunctional protein found in the nucleolus, cytoplasm and the surface of certain cells (Srivastava and Pollard, 1999). In the nucleus it is involved in the synthesis and maturation of ribosomes required for cell proliferation. It can bind to both RNA and DNA and also to other proteins, which facilitates its function as a shuttling protein moving between the nucleus and cytoplasm, carrying ribosomal proteins to the nucleus or assisting in the export of ribosomal subunits to the cytoplasm. Nucleolin has recently been identified as a cell-surface receptor on macrophage cells capable of recognizing early apoptotic cells, in addition to binding ligands such as lipoproteins, laminin, midkine, L-selectin, lactoferrin, viruses and bacterium (Hirano et al., 2005). Nucleolin has 4 RRM binding domains as well as the RGG box. The RGG box is known to affect RNA binding affinity but not specificity (Ghisolfi et al., 1992b) and is particularly effective in disturbing RNA secondary structure (Ghisolfi et al., 1992a). The RGG box has also been found to interact with a subset of ribosomal proteins which may explain its proposed role in shuttling ribosomal proteins to the nucleus (Bouvet et al., 1998). When attached to the cell surface, the RGG box has been found to bind the proteins cytokine MK and HB-19 (Said et al., 2002).

The EWS protein (#26), a member of the TET family (Bertolotti et al., 1996), is also a multifunctional protein found in the nucleus, cytoplasm and on the cell surface. Its functions within the cell include transcriptional activation and repression and possibly pre-mRNA splicing and transport. As it is found on the cell surface, it is also suggested that it may act as a receptor or binding protein for extracellular ligands such as nucleic acids, and thus might mediate extracellular and nuclear events (Belyanskaya et al., 2001). The RGG box regions of EWS are both necessary and sufficient to repress several transcriptional activation domains (Deepa and Lee, 2005).

FMRP (#32) is expressed most abundantly in the brain where it is highly concentrated in neurons and involved in synaptic plasticity. It has 3 RNA binding domains; 2 KH domains as well as the RGG box (Siomi et al., 1993). A primary function is to bind mRNA transcripts and shuttle these between the nucleus and cytoplasm and then to distal parts of the neuron in the form of cytoplasmic granules (Mazroui et al., 2002).

In addition, FMRP represses translation of certain proteins involved in the out-growth of dendritic spines, until these are required at the active synapse (Bagni and Greenough, 2005). This translational repression may occur in conjunction with microRNAs (Corbin et al., 2009). Unlike Nucleoin where the RGG box contributes to affinity but not specificity, here the RGG box binds specifically to several mRNA transcripts (Darnell et al., 2001, Schaeffer et al., 2001). While the most notable function of FMRP is in translation repression, a recent study suggests that in some cases FMRP may activate translation (Bechara et al., 2009).

IHABP4 (#21) (also known as Ki-1/57) is found in the cytoplasm and the nucleus with its location being affected by arginine methylation and phosphorylation (Passos et al., 2006a). The RGG domain of IHABP4 identified in this study, also constitutes a hyaluronan binding motif, (R/K–X(7)-R/K) where X is not acidic (Yang et al., 1994). Hyaluronan is a ubiquitous glycosaminoglycan (GAG) present in the extracellular and pericellular matrices known to play a critical role in tissue development and remodeling. GAGs are long negatively charged polysaccharide chains consisting of repeating disaccharide units. Two of the most common GAGs are; heparan sulfate, and hyaluronan. There is evidence that GAGs are present in the cytoplasm and nucleus of various cell types and are likely to regulate cell behaviour including structural changes in the cytoskeleton or nuclear matrix during cell division or motility (Huang et al., 2000). It is interesting to note that while the RGG domain of IHABP4 binds weakly to RNA it also binds strongly and specifically to hyaluronan (Huang et al., 2000). Overall, it is suggested that IHAPB4 is involved in mRNA splicing and transport (Bressan et al., 2009), and possibly chromatin remodeling, and regulation of transcription (Huang et al., 2000, Nery et al., 2006).

G3BP1 (#34) also has a diverse range of functions (Irvine et al., 2004) although its most prominent role is in the ras-signaling pathway which affects cell proliferation and survival. It is also involved in RNA metabolism; it has helicase activity by which it unwinds DNA/DNA, RNA/DNA and RNA/RNA duplexes, and is involved in mRNA degradation and translational control. Combining its roles in signaling and mRNA biogenesis, it has been suggested the G3BPs are members of a novel sub-class of RNA-binding proteins which act at the level of RNA metabolism in response to cell signaling (Kennedy et al., 2001). Thus providing a mechanism by which the cell can rapidly control protein activity at a stage after transcription.

RGMc (#35), is a member of the Repulsive Guidance Molecule (RGM) family and is the only other GPI anchored protein found in this survey. RGMc (also known as hemojuvelin) is found primarily in skeletal muscle, heart and liver. It interacts with the dependence receptor, neogenin, which regulates the shedding of its GPI anchor and the release of a soluble protein. The soluble form of RGMc then competes with the GPI-anchored form for interaction with members of the BMP family of signaling molecules. Stimulation of this signaling pathway ultimately leads to expression of the protein hepcidin which plays an essential role in iron homeostasis (Zhang et al., 2005, Babitt et al., 2006, Zhang et al., 2007, Zhang et al., 2009). The other members of the RGM family, RGMa and RGMb, lack the RGG domain seen in RGMc. To date there is no indication of the relevance of the RGG region in this protein.

While the above examples demonstrate functional diversity, common themes also emerge. Most striking is the broad range of actions and locations that are suggested for these proteins. Secondly, many appear to be involved in a signaling pathway as well as in RNA binding. In fact, there is growing evidence that some RNA-binding proteins have additional roles as cell surface receptors (Bajenova et al., 2001, Belyanskaya et al., 2001, Said et al., 2002, Bajenova et al., 2003, Hirano et al., 2005, Chang et al., 2007). Thirdly, location and function are commonly modulated through phosphorylation and arginine methylation.

2.4.2 Functional implications for Sho

Sho differs from most of the other RGG box proteins surveyed in 2 main ways; first, as a GPI anchored protein it is located primarily at the outer cell membrane, secondly, it has no other RNA binding motif apart from the RGG box.

The cellular location of Sho will determine its opportunities to bind RNA and/or other ligands. Sho is a GPI anchored protein, as confirmed in mouse (Watts et al., 2007) and for a Sho-like protein (Sho2) (Premzl et al., 2004, Strumbo, 2006) in zebrafish (Miesbauer et al., 2006). This indicates a primary location in the extracellular space where it is attached to the cell membrane by a GPI anchor. However, as already noted in the case of RGMc, GPI-anchored proteins may undergo anchor cleavage ('shedding') (Parkin et al., 2004, Zhang et al., 2005), resulting in formation of soluble proteins which can relocate to other cellular destinations. It is quite possible that Sho will share

a similar fate and may be found in different cellular locations, as already established for PrP (Campana et al., 2005) Alternatively, Sho may undergo proteolytic cleavage before attachment to the cell membrane (Watts et al., 2007) freeing the RGG box domain to participate in intracellular activity. It is plausible that Sho has multiple cellular locations, similar to other proteins identified in this study, such as nucleolin, the EWS protein and RGMc.

The first RGG box protein to be identified by Dreyfuss and co-workers was hnRNP U (#12) which also has a single RNA binding motif, so Sho is not unique in this respect. The RGG box is typically associated with binding to single-stranded nucleic acids (Zhang and Grosse, 1997), whereas additional RNA-binding motifs may allow binding of a broader range of targets as is the case for nucleolin (Ghisolfi et al., 1992a). As Sho has no other RNA binding motif, its RGG box may be expected to bind to single stranded nucleic acid. However, like FMRP it could have a range of single stranded RNA targets, possibly with a preferred 3D structure, such as the G-quadruplex favored by FMRP (Schaeffer et al., 2001, Ramos et al., 2003, Zanotti et al., 2006, Menon and Mihailescu, 2007).

The Sho RGG box may also interact electrostatically with other proteins or anionic molecules such as GAGs in similar fashion to IHAPB4, which has high sequence similarity to Sho (Figure 2-4). Although it is not surprising to find the hyaluronan binding motif (R/K–X(7)-R/K) in an Arg-rich sequence (in fact it is present in most of the proteins included in Table 2-1), it may have particular relevance in the case of Sho, given its primary location as a GPI-anchored cell surface protein. Two examples of GAGs binding to GPI anchored proteins are: heparan sulfate binding to PrP (Horonchik et al., 2005) and hyaluronan binding to the GPI anchored neural protein brevican which plays a role in the structural plasticity of neural tissue (Seidenbecher et al., 1995, Brakebusch et al., 2002, Rauch, 2004).

If Sho were to shed its GPI anchor and re-enter the cell or if the RGG domain is cleaved prior to expression at the cell surface, the RGG box could potentially interact with local mRNA or indeed, as a small protein of no more than 123 residues, Sho would be capable of diffusing in and out of the nucleus (Cyert, 2001) and potentially interacting with a broad range of RNA. In this form it could have a function similar to other RNA-binding proteins involved in neural plasticity (Ule and Darnell, 2006). In general these proteins bind mRNA and affect its transport and local translation at the

neuronal dendrites. Another interesting possibility is that it may bind the mRNA of PrP thus serving to regulate its translation.

Notwithstanding the proposition that Sho may be found in different cellular locations, it is most likely that Sho's central activity as a cell surface molecule is to act as a mediator between the intracellular and extracellular environments. When attached to the cell surface Sho could act as a receptor for nucleic acids or other ligands as suggested for nucleolin and EWS. It is plausible that Sho has a role in signaling which is mediated through RNA binding. When attached to the cell surface by a GPI anchor the free N-terminus would have a good deal of flexibility to bind a range of extracellular RNA. Several RNA-binding proteins with dual roles as cell surface receptors have now been identified (Benimetskaya et al., 1997, Said et al., 2002, Bajenova et al., 2003, Belyanskaya et al., 2003, Hirano et al., 2005, Chang et al., 2007). There is also emerging evidence that extracellular RNA may comprise a new class of signaling molecule mediating cell-to-cell communication (Dinger et al., 2008). Sho could interact with extracellular RNA and assist in its transport across the cell membrane, or RNA binding could initiate a signaling pathway with Sho as the protein mediator in this process.

Extracellular RNA and DNA are also associated with pathological conditions e.g. stroke, head injury and tumor burden (Fischer et al., 2007). It has previously been shown that PrP has a neuroprotective role in stroke, most likely mediated through a signaling mechanism (Spudich et al., 2005). However, it is plausible that this neuroprotective function involves the recognition of extracellular nucleic acids released as a result of trauma. Both PrP and Sho could play a neuroprotective role in this manner.

2.5 CONCLUSION

In this chapter it has been demonstrated that Sho has a conserved RGG box motif with similar composition to other known RGG-box proteins. This work is the first to draw attention to the RGG domain at the N-terminus of Sho and to formulate the hypothesis that Sho may have a function involving RNA binding.

The RGG box of Sho will be explored in this thesis. The RGG boxes of Sho and FMRP are compared using molecular dynamics studies in Chapter 4 and then the binding of these regions to G-quadruplex RNA is explored in Chapters 6-9.

2.6 REFERENCES

Albrecht, M. & Lengauer, T. (2004) Novel Sm-like proteins with long C-terminal tails and associated methyltransferases. *FEBS Lett,* 569, 18-26.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J., Doggett, N., Zharkikh, A., Tavtigian, S.V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D. & Brenner, S. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science,* 297, 1301-10.

Aris, J.P. & Blobel, G. (1991) cDNA cloning and sequencing of human fibrillarin, a conserved nucleolar protein recognized by autoimmune antisera. *Proc Natl Acad Sci U S A,* 88, 931-5.

Babitt, J.L., Huang, F.W., Wrighting, D.M., Xia, Y., Sidis, Y., Samad, T.A., Campagna, J.A., Chung, R.T., Schneyer, A.L., Woolf, C.J., Andrews, N.C. & Lin, H.Y. (2006) Bone morphogenetic protein signaling by hemojuvelin regulates hepcidin expression. *Nat Genet,* 38, 531-9.

Bagni, C. & Greenough, W.T. (2005) From mRNP trafficking to spine dysmorphogenesis: the roots of fragile X syndrome. *Nat Rev Neurosci,* 6, 376-87.

Bagni, C. & Lapeyre, B. (1998) Gar1p binds to the small nucleolar RNAs snR10 and snR30 in vitro through a nontypical RNA binding element. *J Biol Chem,* 273, 10868-73.

Bajenova, O., Stolper, E., Gapon, S., Sundina, N., Zimmer, R. & Thomas, P. (2003) Surface expression of heterogeneous nuclear RNA binding protein M4 on Kupffer cell relates to its function as a carcinoembryonic antigen receptor. *Exp Cell Res,* 291, 228-41.

Bajenova, O.V., Zimmer, R., Stolper, E., Salisbury-Rowswell, J., Nanji, A. & Thomas, P. (2001) Heterogeneous RNA-binding protein M4 is a receptor for carcinoembryonic antigen in Kupffer cells. *J Biol Chem,* 276, 31067-73.

Bechara, E.G., Didiot, M.C., Melko, M., Davidovic, L., Bensaid, M., Martin, P., Castets, M., Pognonec, P., Khandjian, E.W., Moine, H. & Bardoni, B. (2009) A novel function for fragile X mental retardation protein in translational activation. *PLoS Biol,* 7, e16.

Belyanskaya, L.L., Delattre, O. & Gehring, H. (2003) Expression and subcellular localization of Ewing sarcoma (EWS) protein is affected by the methylation process. *Exp Cell Res,* 288, 374-81.

Belyanskaya, L.L., Gehrig, P.M. & Gehring, H. (2001) Exposure on cell surface and extensive arginine methylation of ewing sarcoma (EWS) protein. *J Biol Chem,* 276, 18681-7.

Benimetskaya, L., Loike, J.D., Khaled, Z., Loike, G., Silverstein, S.C., Cao, L., El Khoury, J., Cai, T.Q. & Stein, C.A. (1997) Mac-1 (CD11b/CD18) is an oligodeoxynucleotide-binding protein. *Nat Med,* 3, 414-20.

Bertolotti, A., Lutz, Y., Heard, D.J., Chambon, P. & Tora, L. (1996) hTAF(II)68, a novel RNA/ssDNA-binding protein with homology to the pro-oncoproteins TLS/FUS and EWS is associated with both TFIID and RNA polymerase II. *Embo J,* 15, 5022-31.

Biamonti, G., Buvoli, M., Bassi, M.T., Morandi, C., Cobianchi, F. & Riva, S. (1989) Isolation of an active gene encoding human hnRNP protein A1. Evidence for alternative splicing. *J Mol Biol,* 207, 491-503.

Boisvert, F.M., Chenard, C.A. & Richard, S. (2005) Protein interfaces in signaling regulated by arginine methylation. *Sci STKE,* 2005, re2.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. (1998) Predicting function: from genes to genomes and back. *Journal of Molecular Biology,* 283, 707-725.

Bouvet, P., Diaz, J.J., Kindbeiter, K., Madjar, J.J. & Amalric, F. (1998) Nucleolin interacts with several ribosomal proteins through its RGG domain. *J Biol Chem,* 273, 19025-9.

Brakebusch, C., Seidenbecher, C.I., Asztely, F., Rauch, U., Matthies, H., Meyer, H., Krug, M., Bockers, T.M., Zhou, X., Kreutz, M.R., Montag, D., Gundelfinger, E.D. & Fassler, R. (2002) Brevican-deficient mice display impaired hippocampal CA1 long-term potentiation but show no obvious deficits in learning and memory. *Mol Cell Biol,* 22, 7417-27.

Bressan, G.C., Quaresma, A.J., Moraes, E.C., Manfiolli, A.O., Passos, D.O., Gomes, M.D. & Kobarg, J. (2009) Functional association of human Ki-1/57 with pre-mRNA splicing events. *Febs J,* 276, 3770-83.

Burd, C.G. & Dreyfuss, G. (1994) Conserved structures and diversity of functions of RNA-binding proteins. *Science,* 265, 615-21.

Campana, V., Sarnataro, D. & Zurzolo, C. (2005) The highways and byways of prion protein trafficking. *Trends Cell Biol,* 15, 102-11.

Castrillon, D.H., Quade, B.J., Wang, T.Y., Quigley, C. & Crum, C.P. (2000) The human VASA gene is specifically expressed in the germ cell lineage. *Proc Natl Acad Sci U S A,* 97, 9585-90.

Chang, L.Y., Mohd Ali, A.R., Hassan, S.S. & Abubakar, S. (2007) Human neuronal cell protein responses to Nipah virus infection. *Virol J,* 4, 54.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. & Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res,* 31, 3497-500.

Christensen, M.E. & Fuxa, K.P. (1988) The nucleolar protein, B-36, contains a glycine and dimethylarginine-rich sequence conserved in several other nuclear RNA-binding proteins. *Biochem Biophys Res Commun,* 155, 1278-83.

Cobianchi, F., Calvio, C., Stoppini, M., Buvoli, M. & Riva, S. (1993) Phosphorylation of human hnRNP protein A1 abrogates in vitro strand annealing activity. *Nucleic Acids Res,* 21, 949-55.

Cobianchi, F., Karpel, R.L., Williams, K.R., Notario, V. & Wilson, S.H. (1988) Mammalian heterogeneous nuclear ribonucleoprotein complex protein A1. Large-scale overproduction in Escherichia coli and cooperative binding to single-stranded nucleic acids. *J Biol Chem,* 263, 1063-71.

Consortium, C.S.A.A. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature,* 437, 69-87.

Consortium, I.C.G.S. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature,* 432, 695-716.

Copley, R.R., Goodstadt, L. & Ponting, C. (2003) Eukaryotic domain evolution inferred from genome comparisons. *Curr Opin Genet Dev,* 13, 623-8.

Corbin, R., Olsson-Carter, K. & Slack, F. (2009) The role of microRNAs in synaptic development and function. *BMB Rep,* 42, 131-5.

Cyert, M.S. (2001) Regulation of nuclear localization during signaling. *J Biol Chem,* 276, 20805-8.

Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T. & Darnell, R.B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell,* 107, 489-99.

Deepa, A. & Lee, K. (2005) RGG-boxes of the EWS oncoprotein repress a range of transcriptional activation domains. *Nucleic Acids Res,* 33, 1323-1331.

Dinger, M.E., Mercer, T.R. & Mattick, J.S. (2008) RNAs as extracellular signaling molecules. *J Mol Endocrinol,* 40, 151-9.

Dolzhanskaya, N., Merz, G., Aletta, J.M. & Denman, R.B. (2006) Methylation regulates the intracellular protein-protein and protein-RNA interactions of FMRP. *J Cell Sci,* 119, 1933-46.

Fackelmayer, F.O. & Richter, A. (1994) hnRNP-U/SAF-A is encoded by two differentially polyadenylated mRNAs in human cells. *Biochim Biophys Acta,* 1217, 232-4.

Fay, J.C. & Wu, C.I. (2003) Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet,* 4, 213-35.

Fischer, S., Gerriets, T., Wessels, C., Walberer, M., Kostin, S., Stolz, E., Zheleva, K., Hocke, A., Hippenstiel, S. & Preissner, K.T. (2007) Extracellular RNA mediates endothelial-cell permeability via vascular endothelial growth factor. *Blood,* 110, 2457-65.

Fitzgerald, K.T. & Diaz, M.O. (1999) MLL2: A new mammalian member of the trx/MLL family of genes. *Genomics,* 59, 187-92.

Ghisolfi, L., Joseph, G., Amalric, F. & Erard, M. (1992a) The glycine-rich domain of nucleolin has an unusual supersecondary structure responsible for its RNA-helix-destabilizing properties. *J Biol Chem,* 267, 2955-9.

Ghisolfi, L., Kharrat, A., Joseph, G., Amalric, F. & Erard, M. (1992b) Concerted activities of the RNA recognition and the glycine-rich C-terminal domains of nucleolin

are required for efficient complex formation with pre-ribosomal RNA. *Eur J Biochem,* 209, 541-8.

Hassfeld, W., Chan, E.K., Mathison, D.A., Portman, D., Dreyfuss, G., Steiner, G. & Tan, E.M. (1998) Molecular definition of heterogeneous nuclear ribonucleoprotein R (hnRNP R) using autoimmune antibody: immunological relationship with hnRNP P. *Nucleic Acids Res,* 26, 439-45.

Hedges, S.B. & Kumar, S. (2002) Genomics. Vertebrate genomes compared. *Science,* 297, 1283-5.

Helbig, R. & Fackelmayer, F.O. (2003) Scaffold attachment factor A (SAF-A) is concentrated in inactive X chromosome territories through its RGG domain. *Chromosoma,* 112, 173-82.

Hirano, K., Miki, Y., Hirai, Y., Sato, R., Itoh, T., Hayashi, A., Yamanaka, M., Eda, S. & Beppu, M. (2005) A multifunctional shuttling protein nucleolin is a macrophage receptor for apoptotic cells. *J Biol Chem,* 280, 39284-93.

Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res,* 27, 215-9.

Horonchik, L., Tzaban, S., Ben-Zaken, O., Yedidia, Y., Rouvinski, A., Papy-Garcia, D., Barritault, D., Vlodavsky, I. & Taraboulos, A. (2005) Heparan sulfate is a cellular receptor for purified infectious prions. *J Biol Chem,* 280, 17062-7.

Huang, L., Grammatikakis, N., Yoneda, M., Banerjee, S.D. & Toole, B.P. (2000) Molecular characterization of a novel intracellular hyaluronan-binding protein. *J Biol Chem,* 275, 29829-39.

Irvine, K., Stirling, R., Hume, D. & Kennedy, D. (2004) Rasputin, more promiscuous than ever: a review of G3BP. *Int J Dev Biol,* 48, 1065-77.

Jansen, R.P., Hurt, E.C., Kern, H., Lehtonen, H., Carmo-Fonseca, M., Lapeyre, B. & Tollervey, D. (1991) Evolutionary conservation of the human nucleolar protein fibrillarin and its functional expression in yeast. *J Cell Biol,* 113, 715-29.

Jurica, M.S., Licklider, L.J., Gygi, S.R., Grigorieff, N. & Moore, M.J. (2002) Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. *Rna,* 8, 426-39.

Kajita, Y., Nakayama, J., Aizawa, M. & Ishikawa, F. (1995) The UUAG-specific RNA binding protein, heterogeneous nuclear ribonucleoprotein D0. Common modular structure and binding properties of the 2xRBD-Gly family. *J Biol Chem,* 270, 22167-75.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature,* 423, 241-54.

Kennedy, D., French, J., Guitard, E., Ru, K., Tocque, B. & Mattick, J. (2001) Characterization of G3BPs: tissue specific expression, chromosomal localisation and rasGAP(120) binding studies. *J Cell Biochem,* 84, 173-87.

Kiledjian, M. & Dreyfuss, G. (1992) Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. *Embo J,* 11, 2655-64.

Kobarg, J., Schnittger, S., Fonatsch, C., Lemke, H., Bowen, M.A., Buck, F. & Hansen, H.P. (1997) Characterization, mapping and partial cDNA sequence of the 57-kD intracellular Ki-1 antigen. *Exp Clin Immunogenet,* 14, 273-80.

Krecic, A.M. & Swanson, M.S. (1999) hnRNP complexes: composition, structure, and function. *Curr Opin Cell Biol,* 11, 363-71.

Kreivi, J.P., Trinkle-Mulcahy, L., Lyon, C.E., Morrice, N.A., Cohen, P. & Lamond, A.I. (1997) Purification and characterisation of p99, a nuclear modulator of protein phosphatase 1 activity. *FEBS Lett,* 420, 57-62.

Kuroda, T.S., Maita, H., Tabata, T., Taira, T., Kitaura, H., Ariga, H. & Iguchi-Ariga, S.M. (2004) A novel nucleolar protein, PAPA-1, induces growth arrest as a result of cell cycle arrest at the G1 phase. *Gene,* 340, 83-98.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., Mcewan, P., Mckernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., Mcmurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., Mcpherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* 409, 860-921.

Langbein, L., Heid, H.W., Moll, I. & Franke, W.W. (1993) Molecular characterization of the body site-specific human epidermal cytokeratin 9: cDNA cloning, amino acid sequence, and tissue specificity of gene expression. *Differentiation,* 55, 57-71.

Lee, J. & Bedford, M.T. (2002) PABP1 identified as an arginine methyltransferase substrate using high-density protein arrays. *EMBO Rep,* 3, 268-73.

Lemos, T.A., Passos, D.O., Nery, F.C. & Kobarg, J. (2003) Characterization of a new family of proteins that interact with the C-terminal region of the chromatin-remodeling factor CHD-3. *FEBS Lett,* 533, 14-20.

Liu, H. & Johnson, E.M. (2002) Distinct proteins encoded by alternative transcripts of the PURG gene, located contrapodal to WRN on chromosome 8, determined by differential termination/polyadenylation. *Nucleic Acids Res,* 30, 2417-26.

Liu, Q. & Dreyfuss, G. (1995) In vivo and in vitro arginine methylation of RNA-binding proteins. *Mol Cell Biol,* 15, 2800-8.

Lukasiewicz, R., Nolen, B., Adams, J.A. & Ghosh, G. (2007) The RGG domain of Npl3p recruits Sky1p through docking interactions. *J Mol Biol,* 367, 249-61.

Luking, A., Stahl, U. & Schmidt, U. (1998) The protein family of RNA helicases. *Crit Rev Biochem Mol Biol,* 33, 259-96.

Maita, H., Kitaura, H., Keen, T.J., Inglehearn, C.F., Ariga, H. & Iguchi-Ariga, S.M. (2004) PAP-1, the mutated gene underlying the RP9 form of dominant retinitis pigmentosa, is a splicing factor. *Exp Cell Res,* 300, 283-96.

Matsunaga, E. & Chedotal, A. (2004) Repulsive guidance molecule/neogenin: a novel ligand-receptor system playing multiple roles in neural development. *Dev Growth Differ,* 46, 481-6.

Mazroui, R., Huot, M.E., Tremblay, S., Boilard, N., Labelle, Y. & Khandjian, E.W. (2003) Fragile X Mental Retardation protein determinants required for its association with polyribosomal mRNPs. *Hum Mol Genet,* 12, 3087-96.

Mazroui, R., Huot, M.E., Tremblay, S., Filion, C., Labelle, Y. & Khandjian, E.W. (2002) Trapping of messenger RNA by Fragile X Mental Retardation protein into cytoplasmic granules induces translation repression. *Hum Mol Genet,* 11, 3007-17.

Menon, L. & Mihailescu, M.R. (2007) Interactions of the G quartet forming semaphorin 3F RNA with the RGG box domain of the fragile X protein family. *Nucleic Acids Res,* 35, 5379-92.

Miesbauer, M., Bamme, T., Riemer, C., Oidtmann, B., Winklhofer, K.F., Baier, M. & Tatzelt, J. (2006) Prion protein-related proteins from zebrafish are complex glycosylated and contain a glycosylphosphatidylinositol anchor. *Biochem Biophys Res Commun,* 341, 218-24.

Morohoshi, F., Arai, K., Takahashi, E.I., Tanigami, A. & Ohki, M. (1996) Cloning and mapping of a human RBP56 gene encoding a putative RNA binding protein similar to FUS/TLS and EWS proteins. *Genomics,* 38, 51-7.

Nery, F.C., Rui, E., Kuniyoshi, T.M. & Kobarg, J. (2006) Evidence for the interaction of the regulatory protein Ki-1/57 with p53 and its interacting proteins. *Biochem Biophys Res Commun,* 341, 847-55.

Nishiyama, H., Higashitsuji, H., Yokoi, H., Itoh, K., Danno, S., Matsuda, T. & Fujita, J. (1997) Cloning and characterization of human CIRP (cold-inducible RNA-binding protein) cDNA and chromosomal assignment of the gene. *Gene,* 204, 115-20.

O'brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E. & Marshall Graves, J.A. (1999) The promise of comparative genomics in mammals. *Science,* 286, 458-62, 479-81.

Ohno, T., Ouchida, M., Lee, L., Gatalica, Z., Rao, V.N. & Reddy, E.S. (1994) The EWS gene, involved in Ewing family of tumors, malignant melanoma of soft parts and desmoplastic small round cell tumors, codes for an RNA binding protein with novel regulatory domains. *Oncogene,* 9, 3087-97.

Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P. & Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell,* 127, 635-48.

Parkin, E.T., Watt, N.T., Turner, A.J. & Hooper, N.M. (2004) Dual mechanisms for shedding of the cellular prion protein. *J Biol Chem,* 279, 11170-8.

Passos, D.O., Bressan, G.C., Nery, F.C. & Kobarg, J. (2006a) Ki-1/57 interacts with PRMT1 and is a substrate for arginine methylation. *Febs J,* 273, 3946-61.

Passos, D.O., Quaresma, A.J. & Kobarg, J. (2006b) The methylation of the C-terminal region of hnRNPQ (NSAP1) is important for its nuclear localization. *Biochem Biophys Res Commun,* 346, 517-25.

Patton, J.G., Porro, E.B., Galceran, J., Tempst, P. & Nadal-Ginard, B. (1993) Cloning and characterization of PSF, a novel pre-mRNA splicing factor. *Genes Dev,* 7, 393-406.

Petrini, J.H., Walsh, M.E., Dimare, C., Chen, X.N., Korenberg, J.R. & Weaver, D.T. (1995) Isolation and characterization of the human MRE11 homologue. *Genomics,* 29, 80-6.

Plougastel, B., Zucman, J., Peter, M., Thomas, G. & Delattre, O. (1993) Genomic structure of the EWS gene and its relationship to EWSR1, a site of tumor-associated chromosome translocation. *Genomics,* 18, 609-15.

Ponting, C.P., Schultz, J., Copley, R.R., Andrade, M.A. & Bork, P. (2000) Evolution of domain families. *Adv Protein Chem,* 54, 185-244.

Premzl, M., Gready, J.E., Jermiin, L.S., Simonic, T. & Marshall Graves, J.A. (2004) Evolution of vertebrate genes related to prion and Shadoo proteins--clues from comparative genomic analysis. *Mol Biol Evol,* 21, 2210-31.

Premzl, M., Sangiorgio, L., Strumbo, B., Marshall Graves, J.A., Simonic, T. & Gready, J.E. (2003) Shadoo, a new protein highly conserved from fish to mammals and with similarity to prion protein. *Gene,* 314, 89-102.

Rabbitts, T.H., Forster, A., Larson, R. & Nathan, P. (1993) Fusion of the dominant negative transcription regulator CHOP with a novel gene FUS by translocation t(12;16) in malignant liposarcoma. *Nat Genet,* 4, 175-80.

Ramos, A., Hollingworth, D. & Pastore, A. (2003) G-quartet-dependent recognition between the FMRP RGG box and RNA. *Rna,* 9, 1198-207.

Rauch, U. (2004) Extracellular matrix components associated with remodeling processes in brain. *Cell Mol Life Sci,* 61, 2031-45.

Said, E.A., Krust, B., Nisole, S., Svab, J., Briand, J.P. & Hovanessian, A.G. (2002) The anti-HIV cytokine midkine binds the cell surface-expressed nucleolin as a low affinity receptor. *J Biol Chem,* 277, 37492-502.

Schaeffer, C., Bardoni, B., Mandel, J.L., Ehresmann, B., Ehresmann, C. & Moine, H. (2001) The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. *Embo J,* 20, 4803-13.

Seidenbecher, C.I., Richter, K., Rauch, U., Fassler, R., Garner, C.C. & Gundelfinger, E.D. (1995) Brevican, a chondroitin sulfate proteoglycan of rat brain, occurs as secreted and cell surface glycosylphosphatidylinositol-anchored isoforms. *J Biol Chem,* 270, 27206-12.

Siomi, H. & Dreyfuss, G. (1995) A nuclear localization domain in the hnRNP A1 protein. *J Cell Biol,* 129, 551-60.

Siomi, H., Siomi, M.C., Nussbaum, R.L. & Dreyfuss, G. (1993) The protein product of the fragile X gene, FMR1, has characteristics of an RNA-binding protein. *Cell,* 74, 291-8.

Soulard, M., Della Valle, V., Siomi, M.C., Pinol-Roma, S., Codogno, P., Bauvy, C., Bellini, M., Lacroix, J.C., Monod, G., Dreyfuss, G. & Et Al. (1993) hnRNP G: sequence and characterization of a glycosylated RNA-binding protein. *Nucleic Acids Res,* 21, 4210-7.

Spudich, A., Frigg, R., Kilic, E., Kilic, U., Oesch, B., Raeber, A., Bassetti, C.L. & Hermann, D.M. (2005) Aggravation of ischemic brain injury by prion protein deficiency: role of ERK-1/-2 and STAT-1. *Neurobiol Dis,* 20, 442-9.

Srivastava, M. & Pollard, H.B. (1999) Molecular dissection of nucleolin's role in growth and cell proliferation: new insights. *Faseb J,* 13, 1911-22.

Strasser, K., Masuda, S., Mason, P., Pfannstiel, J., Oppizzi, M., Rodriguez-Navarro, S., Rondon, A.G., Aguilera, A., Struhl, K., Reed, R. & Hurt, E. (2002) TREX is a conserved complex coupling transcription with messenger RNA export. *Nature,* 417, 304-8.

Strumbo, B., L. Sangiorgio, S. Ronchi, J.E. Gready, and T. Simonic (2006) Cloning and analysis of transcripts and genes encoding fish-specific proteins related to PrP. *Fish Physiology & Biochemistry,* 32, 339-353.

Tanaka, K.J., Ogawa, K., Takagi, M., Imamoto, N., Matsumoto, K. & Tsujimoto, M. (2006) RAP55, a cytoplasmic mRNP component, represses translation in Xenopus oocytes. *J Biol Chem,* 281, 40096-106.

Tay, N., Chan, S.H. & Ren, E.C. (1992) Identification and cloning of a novel heterogeneous nuclear ribonucleoprotein C-like protein that functions as a transcriptional activator of the hepatitis B virus enhancer II. *J Virol,* 66, 6841-8.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., Mcdowell, J.C., Maskeri, B., Hansen, N.F., Schwartz, M.S., Weber, R.J., Kent, W.J., Karolchik, D., Bruen, T.C., Bevan, R., Cutler, D.J., Schwartz, S., Elnitski, L., Idol, J.R., Prasad, A.B., Lee-Lin, S.Q., Maduro, V.V., Summers, T.J., Portnoy, M.E., Dietrich, N.L., Akhter, N., Ayele, K., Benjamin, B., Cariaga, K., Brinkley, C.P., Brooks, S.Y., Granite, S., Guan, X., Gupta, J., Haghighi, P., Ho, S.L., Huang, M.C., Karlins, E., Laric, P.L., Legaspi, R., Lim, M.J., Maduro, Q.L., Masiello, C.A., Mastrian, S.D., Mccloskey, J.C., Pearson, R., Stantripop, S., Tiongson, E.E., Tran, J.T., Tsurgeon, C., Vogt, J.L., Walker, M.A., Wetherby, K.D., Wiggins, L.S., Young, A.C., Zhang, L.H., Osoegawa, K., Zhu, B., Zhao, B., Shu, C.L., De Jong, P.J., Lawrence, C.E., Smit, A.F., Chakravarti, A., Haussler, D., Green, P., Miller, W. & Green, E.D. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature,* 424, 788-93.

Totaro, A., Grifa, A., Carella, M., Rommens, J.M., Valentino, M.A., Roetto, A., Zelante, L. & Gasparini, P. (1998) Cloning of a new gene (FB19) within HLA class I region. *Biochem Biophys Res Commun,* 250, 555-7.

Tourriere, H., Chebli, K., Zekri, L., Courselaud, B., Blanchard, J.M., Bertrand, E. & Tazi, J. (2003) The RasGAP-associated endoribonuclease G3BP assembles stress granules. *J Cell Biol,* 160, 823-31.

Tourriere, H., Gallouzi, I.E., Chebli, K., Capony, J.P., Mouaikel, J., Van Der Geer, P. & Tazi, J. (2001) RasGAP-associated endoribonuclease G3Bp: selective RNA degradation and phosphorylation-dependent localization. *Mol Cell Biol,* 21, 7747-60.

Ule, J. & Darnell, R.B. (2006) RNA binding proteins and the regulation of neuronal synaptic plasticity. *Curr Opin Neurobiol,* 16, 102-10.

Virbasius, C.M., Wagner, S. & Green, M.R. (1999) A human nuclear-localized chaperone that regulates dimerization, DNA binding, and transcriptional activity of bZIP proteins. *Mol Cell,* 4, 219-28.

Vodermaier, H.C. (2001) Cell cycle: Waiters serving the Destruction machinery. *Curr Biol,* 11, R834-7.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature,* 420, 520-62.

Watts, J.C., Drisaldi, B., Ng, V., Yang, J., Strome, B., Horne, P., Sy, M.S., Yoong, L., Young, R., Mastrangelo, P., Bergeron, C., Fraser, P.E., Carlson, G.A., Mount, H.T., Schmitt-Ulms, G. & Westaway, D. (2007) The CNS glycoprotein Shadoo has PrP(C)-like protective properties and displays reduced levels in prion infections. *Embo J,* 26, 4038-50.

Whitehead, S.E., Jones, K.W., Zhang, X., Cheng, X., Terns, R.M. & Terns, M.P. (2002) Determinants of the interaction of the spinal muscular atrophy disease protein SMN with the dimethylarginine-modified box H/ACA small nucleolar ribonucleoprotein GAR1. *J Biol Chem,* 277, 48087-93.

Yang, B., Yang, B.L., Savani, R.C. & Turley, E.A. (1994) Identification of a common hyaluronan binding motif in the hyaluronan binding proteins RHAMM, CD44 and link protein. *Embo J,* 13, 286-96.

Yu, M.C., Bachand, F., Mcbride, A.E., Komili, S., Casolari, J.M. & Silver, P.A. (2004) Arginine methyltransferase affects interactions and recruitment of mRNA processing and export factors. *Genes Dev,* 18, 2024-35.

Yugami, M., Kabe, Y., Yamaguchi, Y., Wada, T. & Handa, H. (2007) hnRNP-U enhances the expression of specific genes by stabilizing mRNA. *FEBS Lett,* 581, 1-7.

Zalfa, F., Giorgi, M., Primerano, B., Moro, A., Di Penta, A., Reis, S., Oostra, B. & Bagni, C. (2003) The fragile X syndrome protein FMRP associates with BC1 RNA and regulates the translation of specific mRNAs at synapses. *Cell,* 112, 317-27.

Zanotti, K.J., Lackey, P.E., Evans, G.L. & Mihailescu, M.R. (2006) Thermodynamics of the fragile X mental retardation protein RGG box interactions with G quartet forming RNA. *Biochemistry,* 45, 8319-30.

Zdobnov, E.M., Von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, G.M., Mueller, H.M., Dimopoulos, G., Law, J.H., Wells, M.A., Birney, E., Charlab, R., Halpern, A.L., Kokoza, E., Kraft, C.L., Lai, Z., Lewis, S., Louis, C., Barillas-Mury, C., Nusskern, D., Rubin, G.M., Salzberg, S.L., Sutton, G.G., Topalis, P., Wides, R., Wincker, P., Yandell, M., Collins, F.H., Ribeiro, J., Gelbart, W.M., Kafatos, F.C. & Bork, P. (2002) Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science,* 298, 149-59.

Zhang, A.S., Anderson, S.A., Meyers, K.R., Hernandez, C., Eisenstein, R.S. & Enns, C.A. (2007) Evidence that inhibition of hemojuvelin shedding in response to iron is mediated through neogenin. *J Biol Chem,* 282, 12547-56.

Zhang, A.S., West, A.P., Jr., Wyman, A.E., Bjorkman, P.J. & Enns, C.A. (2005) Interaction of hemojuvelin with neogenin results in iron accumulation in human embryonic kidney 293 cells. *J Biol Chem,* 280, 33885-94.

Zhang, A.S., Yang, F., Wang, J., Tsukamoto, H. & Enns, C.A. (2009) Hemojuvelin-neogenin interaction is required for bone morphogenic protein-4-induced hepcidin expression. *J Biol Chem,* 284, 22580-9.

Zhang, S. & Grosse, F. (1997) Domain structure of human nuclear DNA helicase II (RNA helicase A). *J Biol Chem,* 272, 11487-94.

CHAPTER 3

THEORETICAL AND PRACTICAL BASIS OF MOLECULAR DYNAMICS (MD)

3.1 INTRODUCTION

The first molecular dynamics (MD) studies of proteins/peptides were performed by Karplus and McCammon in the late 1970s (McCammon et al., 1976, McCammon et al., 1977). The new techniques pioneered in these studies heralded a major change in the 'world view' of proteins. Prior to this, most information on the nature of proteins came from X-ray crystallography. This naturally led to a paradigm where proteins were considered fixed and quite rigid in their native state with ligand binding occurring through a 'lock and key' type mechanism. MD provided a new tool to explore the dynamic motion of proteins around their equilibrium structures, showing the importance of structural fluctuation and flexibility to protein function (Karplus and McCammon, 1983).

The first MD simulations of nucleic acids were studies of helical DNA (Levitt, 1983, Tidor et al., 1983) and tRNA (Harvey and McCammon, 1981, Prabhakaran et al., 1983) published some 7 years after the first protein simulations were reported. Simulation of nucleic acids poses different challenges to protein simulation. In particular, the electrostatic treatment of these highly charged molecules was very difficult prior to the implementation of long-range electrostatic methods in 1995 (Cheatham et al., 1995). The implementation of particle-Mesh Ewald (PME) methods for calculation of long range electrostatics together with improvements in force fields have enabled the generation of stable trajectories for DNA (Cheatham and Young, 2001, Sponer and Spackova, 2007) and a wide variety of RNA molecules (McDowell et al., 2007) as well as nucleic acid-protein complexes. MD is now embraced enthusiastically by the scientific community as a useful tool for studying the structure and dynamics of nucleic acids (Laughton and Orozco, 2009).

MD has been used to study different forms of nucleic acid including the G-quadruplex. The first MD study of G-quadruplex DNA was performed in 1994 (Ross and Hardin, 1994). Since that time several studies have shown the reliability of the commonly used AMBER force fields (see Table 3-1) for simulating the tetrad stack of quadruplex

DNA (Spackova et al., 1999, Spackova et al., 2001, Stefl et al., 2001, Spackova et al., 2003, Stefl et al., 2003), although simulation of the single strand loops is more problematic (Fadrna et al., 2004, Fadrna et al., 2009).

The early MD studies of proteins and nucleic acids were performed frugally by today's standards, with inclusion of the minimum number of explicit atoms possible. Huge advances in computer power, the development of parallelized computing, and the implementation of more efficient computational algorithms now allows computational chemists to simulate large systems, e.g. explicitly solvated proteins and nucleic acid on biologically relevant time scales (μs).

In MD, each atom in a molecule is treated as a single particle; a van der Waals sphere centered on the nucleus of the atom. The motion of each sphere is predicted by calculating the force exerted on it by the other atoms with which it interacts, using the classical equations of motion (Newton's Second Law). This treatment is clearly 'approximate' as we know that atoms are quantum particles and that bonding occurs through the redistribution of electron density. To obtain an accurate assessment of the energy of the system it would be necessary to solve the Schrödinger wave function for the system. However, such quantum mechanical (QM) methods are very computationally expensive and generally restricted to the study of small systems.

Rather than using a wave function to derive the energy of the system, in MD the approximate potential energy of the system is calculated using a set of equations and parameters known as a 'force field.' As many different force fields have been developed and released in MD codes, the choice of an appropriate force field and recognition of its limitations is the first step taken in designing a MD simulation. Similarly, although a range of dynamics algorithms, conformational sampling strategies and solvation models have been developed and released for general use, this chapter will only cover those employed in this thesis.

For a comprehensive review of MD methods, applications and future directions I suggest the recent review published by Tamar Schlick (Schlick et al., 2011) and for a more succinct summary of methods the review by Van der Kamp (Van der Kamp et al., 2008). For deeper insight into popular MD packages; a detailed description of the AMBER suite of programs is provided by David Case (Case et al., 2005), the evolution of the CHARMM program is provided by Brooks and colleagues (Brooks et al., 2009)

and an overview of GROMOS is given by Christen and colleagues (Christen et al., 2005).

3.2 FORCE FIELDS AND MD SIMULATIONS

3.2.1 Force fields

The force field provides a set of equations used to calculate the potential energy of the system. A term is included for the most important dimensions that may be affected as a result of the motion of an atom. Typically, the total energy is approximated as a sum of independent stretching, bending, torsional and non-bonded terms (van der Waals and electrostatic components).

$$E = E_{str} + E_{bend} + E_{tors} + E_{vdW} + E_{el} + E_{cross} \quad (1)$$

Where E_{str} = the energy function for the stretching of a bond

E_{bend} = the energy function for the bending of an angle between 3 atoms

E_{tors} = the energy function for torsional rotation in a 4 atom configuration e.g. atoms C_2-N_2-C_3-O in Figure 3-1(A) where the torsional rotation is around N_2-C_3

E_{vdW} = the energy function for the repulsion or attraction of atoms that are not directly bonded

E_{el} = the energy function describing the electrostatic interactions of positively or negatively charged parts of the molecule

E_{cross} = the energy function for the coupling of any of first 3 terms

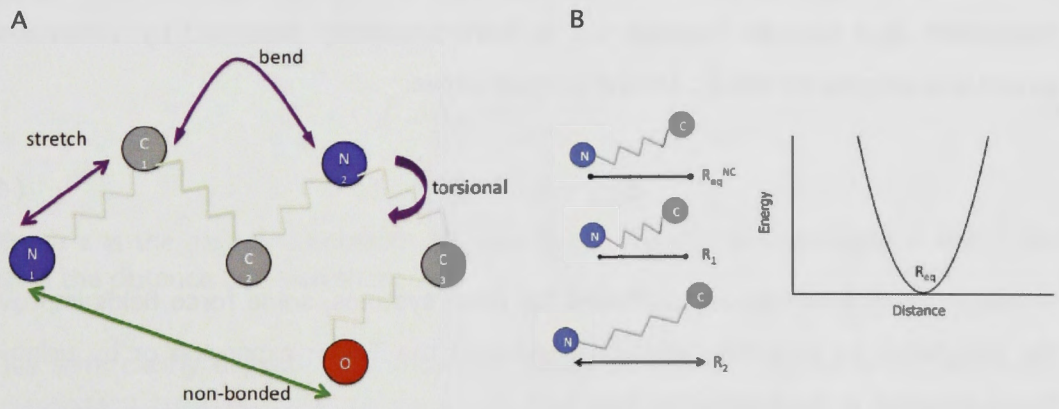


Figure 3-1. Schematic representation of MD treatment of molecules
(A) Different atom types are shown by the blue, grey and red spheres. Examples of bond stretching, angle bending and torsional rotation are depicted by purple arrows and a single non-bonded is shown in green. (B) Potential energy profile for the compression and extension of the N–C bond.

A function is associated with each of the terms in equation 1. The energy function for bond stretching, (e.g. let *R* represent the length of the bond between atoms N and C in Figure 3-1), can be expanded as a Taylor series around a chosen equilibrium bond length (R_{eq}). To second order this gives:

$$E_{str}(R-R_{eq}) = E(0) + \frac{dE}{dR}(R-R_{eq}) + \frac{1d^2E}{2dR^2}(R-R_{eq})^2 \quad (2)$$

$E(0)$ is the zero point for the energy scale and is normally set to zero. The first derivative at the equilibrium point is also zero. With the first 2 terms reduced to zero the 2nd order Taylor expansion is usually written in the form of a simple harmonic oscillator where $\frac{1}{2}\frac{d^2E}{dR^2}$ is called the force constant *k*.

$$E_{str} = k(R-R_{eq})^2 \quad (3)$$

The equilibrium bond lengths and force constants are assigned on the basis of atom connectivities e.g. all C-Cα bonds will be given the same parameters, in the AMBER force field R_e (C-Cα) = 1.409 Å, k = 469 kcal/(mol Å²) (Cornell et al., 1995). These parameters vary between force fields.

The E_{bend} function, which determines the energy required to change the angle between 3 atoms, such as that formed by C_1, C_2 and N_2 in Figure 3-1, should properly be formulated as a periodic function but is more commonly described by a harmonic potential analogous to the E_{str} term discussed above.

$$E_{bend} = k_\theta(\theta-\theta_{eq})^2 \quad (4)$$

Although these functions are sufficient for most systems, some force fields improve the calculation by providing additional terms in the Taylor expansion or by using a Morse potential in the bond stretching case.

E_{tors} must be described by a periodic function as the rotation around 360° must bring the structure back to its original geometry and energy. E_{tors} is written as a Fourier series with the form:

$$E_{tors}(\omega) = \sum_{n=1} V_n \cos(n\omega) \quad (5)$$

Where the n = 1 term describes a rotation of 360°, n = 2 describes rotation periodic by 180°, n = 3 describes rotation periodic by 120° and so on, and V_n is a constant determined by the size of the barrier to rotation.

The van der Waals energy accounts for both the attractive and repulsive forces that exist between non-polar atoms. The attractive forces between two non-polar atoms are termed 'London' or 'dispersion' forces and arise from distortions in the electron cloud surrounding each atom. The momentary distortions in electron distribution create temporary dipoles that interact with similarly induced dipoles resulting in attraction. The repulsive forces occur at short range and result from both electron repulsion (overlap of electron charge densities) and nuclear repulsion. The Lennard-Jones potential provides a functional form that fits this behaviour:

$$E_{vdW}(R_{AB}) = \frac{C_{12}}{(R_{AB})^{12}} - \frac{C_6}{(R_{AB})^6} \quad (6)$$

where R_{AB} is the distance between two non-bonded atoms, A and B, $(R_{AB})^{-12}$ is the repulsive part of the energy, $(R_{AB})^{-6}$ is the attractive part, and C_{12} and C_6 are constants that depend on atom types.

The dominant component of the total non-bonded interaction energy is the electrostatic energy. In its simplest form, this is modeled as the interaction between partial charges assigned to the atoms, interacting via the Coulomb potential:

$$E_{el}(R_{AB}) = \frac{Q_A Q_B}{\varepsilon R_{AB}} \quad (7)$$

Where ε is the dielectric constant, Q_A and Q_B are the charges on atoms A and B and R_{AB} is the distance between them.

This term clearly depends upon how the partial charges are assigned to the atoms. These are normally derived using QM methods to calculate the molecular electrostatic potential (ESP) and then using a fitting function to determine values for the partial charges. In other words, the partial charges are those values that produce the best least-squares fit of the model function to the ESP. As charges on buried atoms do not

make a large contribution to the ESP, they can vary widely with only minor improvements to the 'quality-of-fit'. This can result in the assignment of unrealistic charges for these atoms. The Restrained Electrostatic Potential (RESP) method used in the 1994 and later AMBER force fields introduces a penalty function to address this problem (Bayly et al., 1993).

The fixed charge model lacks the flexibility to describe the electrostatics of a polar molecule. Although this flaw has been acknowledged for at least the last decade, rectification involves a significant increase in both complexity and computational expense (Ponder and Case, 2003). Nevertheless, new generation "polarizable" force fields that contain a more rigorous treatment of electrostatics are currently being developed. For example, the AMOEBA force field replaces fixed charges with a multipole expansion (up to dipoles and quadrupoles), while most other force fields truncate this expansion at the dipole-dipole term.

The total potential energy of the system, which is essentially the energy difference between the observed conformation and a conformation in which all bond lengths, angles etc. are at their parameterized ideal value, is calculated by applying the functions referred to above (or variations of these) to the atoms in the system. E_{str} is calculated for each bonded pair (A–B), E_{bend} for each triplet of connected atoms (A–B–C) and E_{tor} between all pairs that are 1,4 (A–B–C–D). The non-bonded interactions are calculated for pairs that are separated by 3 or more bonds. There may be a cut-off point (e.g. ~ 20 Å) beyond which non-bonded interactions are ignored. Alternatively, in a periodic system, long-range electrostatics may be treated with a particle-mesh Ewald (PME) scheme that obviates the need for a cut-off.

The parameters used; equilibrium bond lengths, bond angles, force constants, torsional parameters and van der Waals parameters may be extracted from multiple experimental data sets or high-level QM calculations, and vary between different force fields. The functional forms noted above may also be tailored for particular systems and vary between force fields.

A good explanation of force field construction and further detailed explanation of the above derivations can be found in *Introduction to Computational Chemistry* (Jensen, 2007).

3.2.2 AMBER Force fields

The AMBER (Assisted Model Building with Energy Refinement) (Case et al., 2005) package of computational software was used for all MD simulations reported in this thesis. The ff03 protein force field (Duan et al., 2003) was used for the initial peptide studies (Chapter 4). Subsequently, ff99 with the Stony Brook modifications (ff99SB) (Hornak et al., 2006) was used in the RNA/DNA studies (Chapter 5) and the RNA-peptide simulations (Chapters 6-8). The AMBER force fields are designed for use with large systems - proteins and nucleic acids. To accommodate this, a minimalist philosophy is applied to the force field formulation. The basic form of the equations referred to above (i.e. without higher order terms) are used to calculate the potential energy. Fixed partial charges are centered on the atoms, hydrogen atoms are explicit and there is no special function for hydrogen bonding. Both ff03 and ff99SB have evolved from the 1994 'Cornell et al.' force field (Cornell et al., 1995). A chronology of the development of the AMBER force fields is set out in Table 3-1.

Table 3-1. Evolution of the AMBER force field leading up to the AMBER 9 release

Year	Designation	Description	Ref
1984	ff84	1st generation multipurpose force field for simulation of proteins and nucleic acids. Partial charges fit to electrostatic potentials (ESP) using QM (Hartree-Fock STO-3G) gas phase calculations. Bond lengths, angles and vdW parameters from crystal structures. Hydrogens bonded to carbon treated as united atoms, only polar hydrogens explicitly included.	(Weiner et al., 1984)
1986	ff86	Modifies ff84 by introducing new atom types to provide an all atom description of H-C groups.	(Weiner et al., 1986)
1994	Cornell et al. ff94	2nd generation force field for simulation of proteins and nucleic acids. Introduction of new atom type for H-C group bonded to electronegative atom. New partial charges calculated in gas phase with QM (HF/6-31G*) and RESP. New vdW parameters. Modification of the bonded parameters of ff84. Introduction of new torsional parameters for protein backbone atoms.	(Cornell et al., 1995)
1999	ff99	Extended the 1994 atom types and made changes to torsional parameters, but retained the 1994 charges.	(Wang et al., 2000)
2002	ff02	First multipurpose polarizable force field. Lone pair sites were included on donor atoms as well as inducible polarizability on all atoms. RESP charges were determined from DFT/cc-pVTZ calculations. Bond and angle parameters from ff99 were retained	(Cieplak et al., 2001, Wang et al., 2006a)
2003	ff03	3rd generation fixed point charge, all atom force field for protein modeling. Partial charges derived using RESP with the potential calculated using QM (B3LYP/cc-pVTZ) with a low-dielectric (ε=4) continuum to mimic the interior of the protein. Modification of torsional parameters for protein backbone atoms.	(Duan et al., 2003)
2006	ff99SB	The 'Stony Brook' modifications of ff99 to introduce new torsional parameters for protein backbone atoms.	(Hornak et al., 2006)

While ff99SB uses the same treatment of charges as ff94 and ff99, ff03 calculates fixed charges with a continuum treatment (instead of gas-phase), and can more properly be considered a new force field. The only polarizable force field, ff02, introduced greater accuracy but at a significant computational cost (~ 6 fold reduction in computational efficiency) (Duan et al., 2003) and is still under development.

As a result of the very wide use of ff99 for peptide simulation, it was observed that simulations appeared to favour α-helical peptide conformations (Garcia and Sanbonmatsu, 2002, Okur et al., 2003). Efforts to rectify this problem are seen in the ff03 (Duan et al., 2003) and ff99SB force fields (Hornak et al., 2006).

A good review of the history of force field development is provided by Ponder and Case (Ponder and Case, 2003).

3.2.3 Molecular Dynamics simulations

The dynamic motion of the system is computed by treating each atom as a van der Waals sphere and solving the classical equations of motion (Newton's Second Law) to determine its motion as a result of the forces exerted on it by the other atoms in the system.

Newton's Second Law of Motion (Force = Mass x Acceleration) is written in differential form as:

$$\mathbf{F} = -\frac{\partial \mathrm{V}}{\partial \mathbf{r}} = m\frac{\partial^2 \mathbf{r}}{\partial t^2} \tag{8}$$

When applied to a MD simulation V is the potential energy function which depends on the positions of the atoms, and **r** is a position matrix containing the Cartesian (x,y,z) co-ordinates of all the atoms in the system.

For a system of atoms with positions $\mathbf{r}_i$, the new positions after a very small time step (Δt) can be derived by a Taylor expansion:

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \frac{\partial \mathbf{r}}{\partial t}(\Delta t) + \frac{1}{2}\frac{\partial^2 \mathbf{r}}{\partial t^2}(\Delta t)^2 + \frac{1}{6}\frac{\partial^3 \mathbf{r}}{\partial t^3}(\Delta t)^3 + \ldots \tag{9}$$

As the first derivative of position with respect to time is velocity (v) and the second derivative of position with respect to time is acceleration (a), this can be rewritten as:

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \mathbf{v}_i(\Delta t) + \frac{1}{2}\mathbf{a}_i(\Delta t)^2 + \frac{1}{6}\mathbf{b}_i(\Delta t)^3 + \ldots \quad (10)$$

Similarly, the position of the system a small time step before r (r_{i-1}) can be derived by substituting (-Δt) into the above equations:

$$\mathbf{r}_{i-1} = \mathbf{r}_i - \mathbf{v}_i(\Delta t) + \frac{1}{2}\mathbf{a}_i(\Delta t)^2 - \frac{1}{6}\mathbf{b}_i(\Delta t)^3 + \ldots \quad (11)$$

Adding equations 12 and 13, all the odd-order terms (1st, 3rd, 5th order terms, etc.) cancel out. Truncating this expression to second order gives the Verlet algorithm (Verlet, 1967) for calculating the position of the atom a time step after the current and previous positions.

$$\mathbf{r}_{i+1} = (2\mathbf{r}_i - \mathbf{r}_{i-1}) + \mathbf{a}_i(\Delta t)^2 \quad (12)$$

This algorithm incurs errors of magnitude $(\Delta t)^4$ in solving Newton's equations of motion.

Equation 14 requires that the acceleration of the atom in each of the x,y,z directions be known at each step, in order for its new position to be determined. Acceleration is calculated as the first derivative of the potential energy with respect to position, based on x,y,z atomic co-ordinates:

$$\mathbf{a}_i = \frac{\mathbf{F}_i}{m_i} = -\frac{1}{m_i}\frac{\partial V}{\partial \mathbf{r}_i} \quad (13)$$

where $\mathbf{r}_i$ is a matrix containing the Cartesian co-ordinates of all atoms at time step *i*, $\mathbf{a}_i$ is the corresponding matrix for acceleration, m_i is the vector of all atomic masses and $\mathbf{F}_i$ is a matrix of forces acting in the x, y and z directions on each atom.

If equations 13 and 14 are written using half a time step and then subtracted the leap-frog algorithm is produced:

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \mathbf{v}_{i+\frac{1}{2}}\Delta t \quad (14)$$

From which the velocities of the atoms at each step can be derived:

$$\mathbf{v}_{i+\frac{1}{2}} = \mathbf{v}_{i-\frac{1}{2}} + \mathbf{a}_i\Delta t \quad (15)$$

Because the position and velocity updates are out of phase by half a time step, they 'leap-frog' over each other at each iteration.

The initial conditions for a MD simulation require a set of atomic Cartesian co-ordinates and velocities. At each time step the new position and velocities of the atoms is determined using an integrator such as the Verlet algorithm (or modified versions e.g. the velocity Verlet integrator and the leap-frog integrator). Following this process, a trajectory of the motion of the system over time is generated.

As the integrator provides an approximate solution to the equations of motion, the total energy of the system is not strictly conserved. However, the smaller the time step, the more accurate the trajectory (Berne). The time step must also be smaller than the fastest process observed in the simulation. The fastest vibrations are those involving the hydrogen atom which occur with frequencies ~ 10^{14} s^{-1}. The time steps required are therefore in the order of femtoseconds (10^{-15} s). The bond lengths involving hydrogen atoms can be constrained e.g. by using the SHAKE algorithm to remove this motion from the system and permit a bigger time step e.g. 2 fs is commonly used. Even using a 2 fs time step, a billion steps are necessary to produce a 2 μs trajectory.

In a simulation certain quantities such as the number of particles (N), temperature (T), volume (V), pressure (P) and total energy (E) will be fixed while others will be allowed to vary. A micro-canonical ensemble implies NVE are fixed, NVT are fixed in a canonical ensemble and fixed NPT constitute an isothermal-isobaric ensemble.

The total energy of the system being simulated has 2 components – the potential energy (which is discussed above) and the kinetic energy. This is generally described by saying that the Hamiltonian for the system has 2 components, the potential energy as a function of atomic position $\mathbf{r} = \{\mathbf{r}_1,...,\mathbf{r}_n\}$ and the kinetic energy as a function of atomic momenta $\mathbf{p} = \{\mathbf{p}_1,...,\mathbf{p}_n\}$:

$$H(\mathbf{r},\mathbf{p}) = K(\mathbf{p}) + E(\mathbf{r}) \tag{16}$$

Kinetic energies are calculated using the usual classical mechanics relationship between energy and momentum:

$$K(\mathbf{p}) = \sum_{i=1}^{n} \frac{1}{2m_i}(p_{i,x}{}^2 + p_{i,y}{}^2 + p_{i,z}{}^2) \quad (17)$$

and potential energies are calculated using a force field as described above in section 3.2.1.

The total kinetic energy, averaged over time, is related to temperature, *T*, by

$$\langle K(\mathbf{p}) \rangle = \frac{3}{2} N k_B T \quad (18)$$

Where *N* is the number of atoms, k_B is the Boltzmann constant and *T* is temperature.

Given the relationship between temperature and kinetic energy, it follows that the temperature of the simulation may be kept constant by scaling the velocities of the atoms. Similarly, the pressure of the system can be kept constant by adjusting the volume of the system through scaling the Cartesian position co-ordinates. Accordingly, although the NVE ensemble is naturally produced in a MD simulation it is possible to generate a NVT or NPT ensemble by varying the velocities or atomic positions at each time step.

3.2.4 Limitations of force fields

Constructing a balanced force field where approximation errors (e.g. small number of atom types, neglect of cross terms, incomplete potential energy functions) are evenly spread across all interactions, is not a trivial a task. Although certain parameters can be fine-tuned it is possible that these changes will lead to errors in another part of the calculation (McDowell et al., 2007, Sponer and Spackova, 2007).

For example, changes made to the Cornell et al. force field in 1999 improved certain aspects of DNA simulation (better representation of helical twist, angle of bases to the backbone and sugar pucker) but led to the unexpected loss of other desirable aspects of the force field. In the words of Cheatham and Young:

"This shows the complexity of tweaking force fields due to the interrelated and correlated effects of all the separate parameters." (Cheatham and Young, 2001)

The major limitation of many force fields, including ff03 and TIP3P used in this thesis, is lack of polarizability in the treatment of electrostatics. The use of fixed partial charges fails to take account of the change in charge distribution as the molecular environment alters. This is particularly problematic in the context of flexible charged domains, such as the sugar-phosphate backbone of nucleic acids. According to Sponer and colleagues the difficulty in applying this treatment to the nucleic acid backbone arises because:

1. the highly charged backbone is also flexible and the constant point charges can not reproduce the electrostatic potential equally well for distinct conformations and;
2. the phosphodiester is a highly polarizable anion, with complicated electronic structure that changes with solvation (McDowell et al., 2007, Sponer and Spackova, 2007).

The treatment of polarizability has been a major research target for some time now and polarization options for the AMBER (Cieplak et al., 2001, Wang et al., 2006b) , CHARMM (Patel and Brooks, 2004, Patel et al., 2004, Patel and Brooks, 2005) and OPLS (Kaminski et al., 2002, Kaminski et al., 2009) force fields are under development. The AMOEBA multipurpose polarizable force field developed by Ponder and co-workers has recently been released for general use (Ponder et al., 2010).

While the limitations of the MD method must be understood, it is also clear that MD has proved a very useful tool for studying the dynamics of biomolecules. Accuracy and efficiency have both increased over the years and this is a trend that is likely to continue.

3.3 SOLVATION

3.3.1 Explicit and implicit solvation

As biological molecules are found in an aqueous environment, it is important to account for solvation effects in MD simulations of biomolecules. There are 2 ways of doing this, illustrated in Figure 3-2. Water molecules may either be simulated explicitly or their effect approximated using a continuum model.

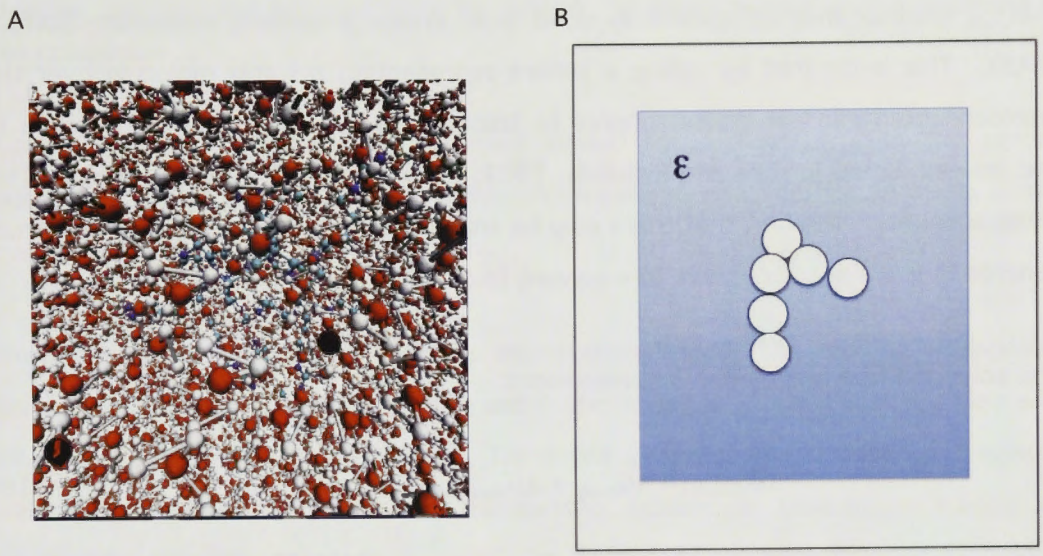


Figure 3-2. Explicit vs Implicit Solvation
(A) Snapshot of explicit water solvation with molecule (blue spheres) surrounded by explicit water molecules (B) Schematic diagram of implicit solvation in which a continuum with a dielectric constant (ε) surrounds a molecule-shaped hole.

Several different water models have been developed since the early 1980s. The TIP3P and TIP4P models were introduced in 1983 (Jorgensen et al., 1983) and, together with the simple point charge (SPC) model (Berendsen et al., 1987), remain popular (Jorgensen and Tirado-Rives, 2005). In all of these models, the geometry of the water molecule is kept fixed so there is no bond stretching or angle bending. Partial charges fixed on the atomic nuclei are used for calculating intermolecular electrostatic interactions.

Periodic boundary conditions are applied to the solvent molecules, i.e. they are placed in a box, to ensure that the outer molecules do not 'boil off' into space. The boxes are then duplicated in all directions, creating a periodic system. The incorporation of periodic boundary conditions also allows the electrostatic interactions to be evaluated using the Particle Mesh Ewald (PME) method for calculating the full electrostatic energy of cell in a periodic lattice (Darden et al., 1993, Essmann et al., 1995).

Continuum methods treat the solvent as a homogenous medium with a dielectric constant (Honig and Nicholls, 1995). A hole is made in this medium to represent the solute. There are different methods for creating this cavity, ranging from using a simple sphere to represent the molecule to creating a molecule-shaped hole. One way of creating the molecule shaped cavity is to draw interconnected spheres based on

the van der Waals radius of the atoms (van der Waals surface), as illustrated in Figure 3-2(B). Another method commonly used is to create a Solvent Accessible Surface (SAS). This is created by rolling a sphere representing a water molecule over the interconnected van der Waals spheres to trace out a surface which is accessible to the solvent molecule (Lee and Richards, 1971, Richards, 1977, Connolly, 1983). This takes account of the fact that there may be small crevices between the van der Waals spheres that are not accessible to a solvent (e.g. water) molecule.

The solvation free energy has 3 components:

$$\Delta G_{solvation} = \Delta G_{cavity} + \Delta G_{dispersion} + \Delta G_{elec} \quad (19)$$

Where ΔG_{cavity} is the destabilization of the system caused by making a hole in the continuum.

$\Delta G_{dispersion}$ represents stabilizing dispersion effects between the solvent and the solution.

ΔG_{elec} is the stabilization resulting from the electric charge distribution of the solute molecule. This charge distribution will polarize the continuum, which in turn has a stabilizing effect on the solute molecule.

The values of ΔG_{cavity} and $\Delta G_{dispersion}$ are generally considered to be proportional to the SAS area.

3.3.2 Implicit solvation models used in these simulations

AMBER provides a number of options for the treatment of implicit solvation. Here, the generalized Born (GB) method with modifications by Onufriev, Bashford and Case (GB^{OBC}) (Onufriev et al., 2004) was used for the implicit solvent simulations discussed in Chapter 4.

In AMBER, the solvation free energy ΔG_{solv} is decomposed into 2 parts; an electrostatic part G_{el} and a non-electrostatic part G_{nonel}.

$$\Delta G_{sol} = \Delta G_{el} + \Delta G_{nonel} \quad (20)$$

where ΔG_{nonel} is the energy associated with solvating a molecule that has had all its charges removed and ΔG_{el} accounts for adding the charges back in the presence of the continuum

Accordingly, ΔG_{nonel} captures the energies termed ΔG_{cavity} + $\Delta G_{dispersion}$ in the general discussion above. In Amber, ΔG_{nonel} is taken to be proportional to the SAS area of the molecule (Weiser et al., 1999).

The electrostatic component (ΔG_{el}) is approximated using the analytic generalized Born method (Born, 1920). Each atom within the molecule is treated as a sphere with effective Born radius R_i and charge q_i. The inside of the atom is treated as having a dielectric constant of 1 and the surrounding solvent is treated as having an appropriate dielectric constant, (ε_w = 80 for water). ΔG_{el} is approximated as follows (Still et al., 1990, Srinivasan et al., 1999):

$$\Delta G_{el} \approx -\frac{1}{2}\sum_{ij}\frac{q_i q_j}{f_{GB}(r_{ij},R_i,R_j)}\left[1-\frac{e^{-\kappa f_{GB}(r_{ij},R_i,R_j)}}{\varepsilon_w}\right] \qquad (21)$$

where r_{ij} is the distance between the centres of atoms *i* and *j*. f_{GB} is a smooth function of r_{ij} that takes into account the change in atomic radii as the nuclei approach one another. κ is an electostatic screening parameter that accounts for salt-induced screening using a Debye-Hückel model.

As the conformation of the molecule changes the effective Born radii of each atom will also change. Re-computing these values at every step is costly and has led to the introduction of further approximations. One such protocol, mentioned above, is the use of van der Waal spheres to map out the surface of the molecule. However, this approach can lead to underestimation of the Born radii of deeply buried atoms. This problem is addressed by the OBC modifications to GB, which essentially introduces a scaling function that re-scales the effective Born radii proportional to the degree of each atom's burial. Two different parameterizations of the OBC modification are offered in AMBER. In this study the option (igb=5) was chosen as it has been reported to provide results closely aligned with the more rigorous Poisson-Boltzman calculation of solvation free energy (Feig et al., 2004).

Although GB greatly improves the efficiency of simulations relative to explicit solvation, there is an accuracy cost. GB has been reported to overstabilize ion pair

interactions thus favouring salt bridges at the expense of hydrophobic contacts in peptide simulations (Zhou and Berne, 2002, Zhou, 2003) and to favour the formation of α-helix secondary structure (Okur et al., 2006, Okur et al., 2008).

3.4 ENHANCED SAMPLING

3.4.1 Enhancement of conformational sampling

Complex systems such as peptides and nucleic acids have been described as having a 'rugged potential energy surface' (Wolynes et al., 1995, Onuchic et al., 1997). This phrase implies that as the system moves from one potential energy basin to another it must cross barriers that are larger than k_BT. As such transitions occur rarely they may only be observed in very long MD simulations. In proteins, rotation around torsional angles presents local barriers, while the close positioning of sidechain and mainchain atoms is strongly repulsive. Accordingly, proteins may fold into conformations which are difficult to escape (Berne). It is possible, if not likely, that simulations will get stuck at one of these stable points, or 'local minima' without fully exploring all possible conformations.

This is called the problem of 'quasi-ergodicity', where it may appear that a simulation is converged in certain parameters, but large energy barriers have prevented the sampling of other important and potentially lower energy conformations. The validity of MD is based on the ergodic hypothesis, which states that observing a single particle over a significant period of time (e.g a MD trajectory) will produce the same average results as observing a large number of particles over a very short time (e.g. as occurs in experimental measurements).

To study complex systems it is important to thoroughly explore computational space within a realistic timeframe. Common strategies for achieving this involve:

1. reduction in the number of degrees of freedom in the system, e.g. by using an implicit solvent model rather than explicit solvent
2. facilitation of energy barrier crossing using an enhanced sampling technique

Several methods have been developed to enhance conformational sampling. Some of these are described in a special issue of the Journal of Molecular Graphics & Modeling

dealing with 'Conformational Sampling' Volume 22, Issue 5 (May 2004). Also for review see (Tai, 2004, Christen and van Gunsteren, 2008, Klein and Shinoda, 2008, Liwo et al., 2008, Schlick, 2009). The Replica Exchange method of enhanced sampling and the hybrid-REMD method were used in these MD studies and will be discussed in more detail.

3.4.2 Replica Exchange Molecular Dynamics (REMD) and hybrid REMD

REMD (also called parallel tempering) involves the creation of a number of replicas of the original model that are simulated simultaneously at different temperatures. After a specified number of steps, successive replicas are given the opportunity to exchange temperatures. If an exchange occurs the velocities of the atoms in the replicas are rescaled by the square root of the ratio of the 2 temperatures (Sugita and Okamoto, 1999).

In making an exchange it is necessary to maintain the Boltzmann distribution of energy states for that temperature. For example, considering Replica 1 (T_1) and Replica 2 (T_2): it will not be possible to swap Replica 2 into T_1 if there is little chance of finding a conformation with the potential energy of Replica 2 in the Boltzmann potential energy distribution for T_1. In other words, an exchange is only allowed if there is sufficient overlap in the potential energy distribution of the 2 replicas. The exchange probability is thus constructed in a manner that allows the canonical ensemble properties to be maintained in each simulation, and provides information about conformational ensembles as a function of temperature.

The exchange probability is calculated as:

$$P = \min\left(1,\ \exp\left\{\left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)\left(E_i - E_j\right)\right\}\right) \qquad (22)$$

where k_B is the Boltzmann constant, T_i and T_j are the temperatures of replicas i and j, with associated energies E_i and E_j.

Exchange decisions are made based upon a Metropolis criterion (Metropolis et al., 1953) applied to the above probability distribution.

The main advantage of the REMD algorithm is that it allows a replica that might otherwise be trapped in a local minimum to increase its kinetic energy and thereby escape into a different conformation. As a result of these features, REMD has been widely used to study peptide/protein folding (Garcia and Sanbonmatsu, 2001, Zhou et al., 2001, Sanbonmatsu and Garcia, 2002, Pitera and Swope, 2003, Kinnear et al., 2004, Roe et al., 2005, Pitera et al., 2008, Vaiana and Sanbonmatsu, 2009).

However, a drawback of REMD is its computational cost. To achieve reasonable exchange rates the replicas must be chosen so that their temperatures and potential energy distributions are not too different. Hence, a large number of replicas may be needed to span the desired temperature range. Accordingly, the cost of performing REMD simulations can in some cases be higher than performing long conventional MD, and potentially less efficient (Beck et al., 2007).

Recently, the hybrid-REMD method has been developed as one approach to addressing this issue (Okur et al., 2006). This method allows the system to be simulated surrounded by explicit solvent but removes most of the water molecules at the time when the exchange probability is calculated. It is the calculation of the exchange probability that requires overlap in the potential energy distribution of the replicas and therefore determines the number of replicas required to span the desired temperature range. As the number of replicas required to span a given temperature range increases with the square root of the number of degrees of freedom in the system (Fukunishi et al., 2002, Kofke, 2002, Rathore et al., 2005), using a subset of the explicit water molecules surrounded by a continuum representation of solvation significantly reduces computational cost. While exchange probability is calculated in this manner the dynamic simulation is performed with fully explicit solvation.

A reduction in cost generally comes at the expense of some other feature. Because the Hamiltonian used for exchange differs from that employed during dynamics, a correct canonical ensemble is not guaranteed.

Nevertheless, original testing of the hybrid-REMD method on alanine peptides produced results very similar to those obtained with REMD and explicit waters although at one-fifth the computational expense (Okur et al., 2006). Subsequently, the hybrid-REMD method was found to perform very well with a 4 residue charged peptide (Arg-Ala-Ala-Glu) and an 11 residue peptide containing charged residues (Okur et al., 2008).

3.5 PERFORMING AND ANALYSING AN AMBER SIMULATION

There are 3 main processes in an MD study; model preparation, simulation and analysis. Different pieces of computer software perform these tasks in AMBER. tLEAP is used for model preparation, dynamics are performed by SANDER and Ptraj is used for analysis.

tLEAP assembles the molecular system from a PDB file. It prepares the chosen force field terms and parameters and adds in the specified solvation state and counter ions. The program produces two files, which are essential for the running of the simulation. One is the parameter-topology file (prmtop), which contains information about the atoms in the system as well as the parameters of the chosen force field. The other file (prmcrd) contains the original Cartesian co-ordinates of all the atoms in the system (taken from a PDB file).

MD calculations are performed using the SANDER program. The 2 principle algorithms in SANDER are a conjugate-gradient method for minimizing the molecular energy and a leap-frog Verlet integrator for generating a MD trajectory. An input file is created to specify simulation variables and is read together with the prmtop and prmcrd files. Simulations are performed in parallel using the MPI programming interface to communicate between computers.

The trajectories produced from dynamics simulations may be analysed using the Ptraj program. Ptraj reads in accessed conformations from designated trajectory files and computes values such as inter-atomic distances, RMSD from a reference structure, number and location of hydrogen bonds etc.

I chose to use the ff03 force field (Duan et al., 2003) for my peptide simulations and the ff99SB force field (Hornak et al., 2006) for simulations involving nucleic acids. The ff03 force field and ff99SB were developed by the Kollman and Simmerling groups, respectively. Both groups set out to remedy some of the shortfalls of the earlier AMBER force fields (ff94 and ff99) including the problem of α-helix bias. The ff03 force field was developed specifically for protein simulations and was available when I commenced my peptide studies at the beginning of 2006. Benchmarking of the new force field by Duan et al. (2003) led to the conclusion that tests of ff03 "clearly showed that the new force field has achieved a reasonable balance in helical and extended conformations".

By the time I went on to study nucleic acids the Stony Brook modification of ff99 (ff99SB) was available. Testing of several different force fields conducted by the Simmerling group showed that ff03 and ff99SB produced similar conformational ensembles (Hornak et al. 2006). As ff03 could not be used for nucleic acid simulations, ff99SB was the most appropriate choice of force field for these simulations.

I realized that some method of enhanced sampling would be necessary to adequately explore the conformational space of the disordered peptides. This was confirmed by comparing an MD simulation of the peptides in explicit water (without enhanced sampling) to simulations conducted using REMD. The MD simulations were discontinued after 12 ns because the conformational sampling was clearly inadequate, and are not reported in this thesis. A new method of enhanced sampling hybrid-REMD (Okur et al. 2006) was released by the Simmerling group around this time. I thought this new approach provided a very promising method for simulating large systems at a realistic cost.

In proceeding with this method I relied upon benchmarking studies performed by the Simmerling group (Okur et al. 2006, Okur et al. 2008). In the first study benchmarking of 3 different sized Ala peptides including a 10 residue Ala peptide (Ala_{10}) were tested. Three different methods were tested (1) REMD with explicit solvation (reference simulations) (2) REMD with implicit solvation and (3) hybid-REMD. The ff99SB force field was used in this study and the 2008 study. The 2006 study concluded that the hybrid-REMD simulations accorded well with the reference REMD simulations, but differed from the implicit solvation REMD simulations (which showed a bias towards a-helical conformations).

The 2008 study investigated more complicated peptides, both contained charged residues: a 4 residue peptide (Ace-Arg-Ala-Ala-Glu-NH2) and a 13 residue peptide (MLSDEDFKAVFGM). The same 3 methods were tested as for the earlier study (1) REMD with explicit solvation (reference simulations) (2) REMD with implicit solvation and (3) hybid-REMD. This study concluded that hybrid-REMD significantly reduced the computational cost of REMD, while providing backbone conformational sampling, free

energy profiles, and salt bridge geometries that were in excellent agreement with the data from standard REMD in explicit solvent. They noted only a slight (0.5 - 1 kcal/mol) bias favoring salt bridge conformations in the hybrid model.

I therefore chose to use the hybrid-REMD method of enhanced sampling for the simulations of the disordered peptides. For comparison I also simulated the peptides using standard REMD, with implicit solvation to reduce the cost of the simulations.

A brief overview of the MD simulations discussed thoroughly in Chapters 4-8 are given below.

3.6 OVERVIEW OF MD SIMULATIONS IN THIS THESIS

3.6.1 Peptide simulations (Chapter 4)

Two 21 residue disordered peptides (Sho and Fx), each containing an RGG box motif, were studied. The ff03 force field was used with two different sampling methods;

1. standard REMD using an implicit solvent model (GB^{OBC})
2. Hybrid-REMD using explicit solvation (TIP3P model) for the dynamics simulation and replacing most of these waters with a continuum model (GB^{OBC}) for replica exchanges.

Ptraj was used to calculate several parameters of interest including distances between Arg residues, end-to-end distances, radius of gyration, ϕ and ψ backbone torsional angles, hydrogen bond occupancy and DSSP structural elements. Distributions of these parameters over the trajectories were compared for the different peptides, and for the different methods used.

3.6.2 RNA and DNA simulations (Chapter 5)

A model RNA G-quadruplex was created based on the known structure of a DNA G-quadruplex. The RNA parameters found in rna.ff99 were used for the RNA quadruplex together with the ff99SB force field.

Ptraj was used to calculate hydrogen bond occupancy, radius of gyration, the distance between stabilizing K^+ ions and the RMSD of the loops and tetrads comprising the G-

quadruplexes. The stability of the model RNA quadruplex was assessed and compared to the DNA quadruplex.

3.6.3 RNA-peptide docking simulations (Chapters 6-8)

The RNA model quadruplex examined in Chapter 5 was allowed to dock with randomly chosen conformations of the 2 peptides, derived from the hybrid-REMD peptide simulations. A python script was written to bring the RNA and peptide within close spatial proximity, subsequent MD simulation allowed the 2 to move towards each other and dock. The ff99SB force field was used for these simulations.

The contacts made by the RNA and peptide were analyzed, the free energy of the complexes was estimated using MM-PBSA scripts implemented in AMBER (Kollman et al., 2000). Interaction between the RNA and peptides is described in detail.

3.7 REFERENCES

Bayly, C.I., Cieplak, M., Cornell, W.D. & Kollman, P.A. (1993) A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.,* 97, 10269-10280.

Beck, D.A., White, G.W. & Daggett, V. (2007) Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. *J Struct Biol,* 157, 514-23.

Berendsen, H.J.C., Grigera, J.R. & Straatsma, T.P. (1987) The Missing Term in Effective Pair Potentials. *J. Phys. Chem.,* 91, 6269-6271.

Berne, B.J. Molecular Dynamics and Hybrid Monte Carlo in Systems with MultipleTime Scales and long-range Forces: Reference System Propagator Algorithms. *Encyclopedia of Computational Chemistry.*

Born, M. (1920) Volume and heat of hydration of ions. *Z. Phys,* 1, 45-48.

Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M. & Karplus, M. (2009) CHARMM: the biomolecular simulation program. *J Comput Chem,* 30, 1545-614.

Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B. & Woods, R.J. (2005) The Amber biomolecular simulation programs. *J Comput Chem,* 26, 1668-88.

Cheatham, T.E., Miller, J.L., Fox, T., Darden, T.A. & Kollman, P.A. (1995) Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA and Proteins. *J Am Chem Soc,* 117, 4193-4194.

Cheatham, T.E. & Young, M.A. (2001) Molecular Dynamics Simulation of Nucleic Acids: Successes, Limitations, and Promise. *Biopolymers,* 56, 232-256.

Christen, M., Hunenberger, P.H., Bakowies, D., Baron, R., Burgi, R., Geerke, D.P., Heinz, T.N., Kastenholz, M.A., Krautler, V., Oostenbrink, C., Peter, C., Trzesniak, D. & Van Gunsteren, W.F. (2005) The GROMOS software for biomolecular simulation: GROMOS05. *J Comput Chem,* 26, 1719-51.

Christen, M. & Van Gunsteren, W.F. (2008) On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *J Comput Chem,* 29, 157-66.

Cieplak, P., Caldwell, J. & Kollman, P. (2001) Molecular Mechanical Models for Organic and Biological Systems Going Beyond the Atom Centered Two Body Additive Approximation: Aqueous Soluion Free Energies of Methanol and N-Methyl Acetamide, Nucleic Acid Base, and Amide Hydrogen Bonding and Chloroform/Water Partition Coefficients of the Nucleic Acid Bases. *J. Comput. Chem,* 22, 1048-1057.

Connolly, M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science,* 221, 709-13.

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. & Kollman, P.A. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.,* 117, 5179-5197.

Darden, T., York, D. & Pedersen, L.G. (1993) Particle mesh Ewald--an Nlog(N) mehtod for Ewald sums in large systems. *J. Chem. Phys.,* 98, 10089-10092.

Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. & Kollman, P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem,* 24, 1999-2012.

Essmann, U., Perera, L., Berkowitz, M., Darden, T., Lee, H. & Pedersen, L.G. (1995) A smooth particle mesh Ewald method. J. Chem. Phys, 8577-8593.

Fadrna, E., Spackova, N., Sarzynriska, J., Koca, J., Orozco, M., Cheatham, T., Kulinski, T. & Sponer, J. (2009) Single Stranded Loops of Quadruplex DNA as Key Benchmark for Testing Nucleic Acids Force Fields. *J. Chem. Theory Comput.,* 5, 2514-2530.

Fadrna, E., Spackova, N., Stefl, R., Koca, J., Cheatham, T.E., 3rd & Sponer, J. (2004) Molecular dynamics simulations of Guanine quadruplex loops: advances and force field limitations. *Biophys J,* 87, 227-42.

Feig, M., Onufriev, A., Lee, M.S., Im, W., Case, D.A. & Brooks, C.L., 3rd (2004) Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J Comput Chem,* 25, 265-84.

Fukunishi, H., Watanabe, O. & Takada, S. (2002) On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *Journal of Chemical Physics,* 116, 9058-9067.

Garcia, A.E. & Sanbonmatsu, K.Y. (2001) Exploring the energy landscape of a beta hairpin in explicit solvent. *Proteins,* 42, 345-54.

Garcia, A.E. & Sanbonmatsu, K.Y. (2002) Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc Natl Acad Sci U S A,* 99, 2782-7.

Harvey, S.C. & Mccammon, J.A. (1981) Intramolecular flexibility in phenylalanine transfer RNA. *Nature,* 294, 286-7.

Honig, B. & Nicholls, A. (1995) Classical electrostatics in biology and chemistry. *Science,* 268, 1144-9.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins,* 65, 712-25.

Jensen, F. (2007) *Introduction to Computational Chemistry*, John Wiley & Sons, Ltd.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.,* 79, 926.

Jorgensen, W.L. & Tirado-Rives, J. (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci U S A,* 102, 6665-70.

Kaminski, G.A., Ponomarev, S.Y. & Liu, A.B. (2009) Polarizable Simulations with Second order Interaction Model - force field and software for fast polarizable calculations: Parameters for small model systems and free energy calculations. *J Chem Theory Comput,* 5, 2935-2943.

Kaminski, G.A., Stern, H.A., Berne, B.J., Friesner, R.A., Cao, Y.X., Murphy, R.B., Zhou, R. & Halgren, T.A. (2002) Development of a polarizable force field for proteins via ab initio quantum chemistry: first generation model and gas phase tests. *J Comput Chem,* 23, 1515-31.

Karplus, M. & Mccammon, J.A. (1983) Dynamics of proteins: elements and function. *Annu Rev Biochem,* 52, 263-300.

Kinnear, B.S., Jarrold, M.F. & Hansmann, U.H. (2004) All-atom generalized-ensemble simulations of small proteins. *J Mol Graph Model,* 22, 397-403.

Klein, M.L. & Shinoda, W. (2008) Large-scale molecular dynamics simulations of self-assembling systems. *Science,* 321, 798-800.

Kofke, D.A. (2002) On the acceptance probability of replica-exchange Monte Carlo trials. *Journal of Chemical Physics,* 117, 6911-6914.

Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D.A. & Cheatham, T.E., 3rd (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res,* 33, 889-97.

Laughton, C.A. & Orozco, M. (2009) Nucleic acid simulations themed issue. *Phys Chem Chem Phys,* 11, 10541-2.

Lee, B. & Richards, F. (1971) Interpretation of Protein Structures - Estimation of Static Accessibility. *J Mol Biol,* 55, 379.

Levitt, M. (1983) Computer simulation of DNA double-helix dynamics. *Cold Spring Harb Symp Quant Biol,* 47 Pt 1, 251-62.

Liwo, A., Czaplewski, C., Oldziej, S. & Scheraga, H.A. (2008) Computational techniques for efficient conformational sampling of proteins. *Curr Opin Struct Biol,* 18, 134-9.

Mccammon, J.A., Gelin, B.R. & Karplus, M. (1977) Dynamics of folded proteins. *Nature,* 267, 585-90.

Mccammon, J.A., Gelin, B.R., Karplus, M. & Wolynes, P.G. (1976) The hinge-bending mode in lysozyme. *Nature,* 262, 325-6.

Mcdowell, S.E., Spackova, N., Sponer, J. & Walter, N.G. (2007) Molecular dynamics simulations of RNA: an in silico single molecule approach. *Biopolymers,* 85, 169-84.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.,* 21, 1087-1092.

Okur, A., Strockbine, B., Hornak, V. & Simmerling, C. (2003) Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins. *J Comput Chem,* 24, 21-31.

Okur, A., Wickstrom, L., Layten, M., Geney, R., Song, K., Hornak, V. & Simmerling, C. (2006) Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model. *Journal of Chemical Theory and Computation,* 2, 420-433.

Okur, A., Wickstrom, L. & Simmerling, C. (2008) Evaluation of salt bridge structure and energetics in peptides using explicit, implicit, and hybrid solvation models. *Journal Of Chemical Theory And Computation,* 4, 488-498.

Onuchic, J.N., Luthey-Schulten, Z. & Wolynes, P.G. (1997) Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem,* 48, 545-600.

Onufriev, A., Bashford, D. & Case, D.A. (2004) Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins: Structure, Function, and Bioinformatics,* 55, 383-394.

Patel, S. & Brooks, C.L., 3rd (2004) CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J Comput Chem,* 25, 1-15.

Patel, S. & Brooks, C.L., 3rd (2005) A nonadditive methanol force field: bulk liquid and liquid-vapor interfacial properties via molecular dynamics simulations using a fluctuating charge model. *J Chem Phys,* 122, 024508.

Patel, S., Mackerell, A.D., Jr. & Brooks, C.L., 3rd (2004) CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J Comput Chem,* 25, 1504-14.

Pitera, J.W. & Swope, W. (2003) Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. *Proc Natl Acad Sci U S A,* 100, 7587-92.

Pitera, J.W., Swope, W.C. & Abraham, F.F. (2008) Observation of noncooperative folding thermodynamics in simulations of 1BBL. *Biophys J,* 94, 4837-46.

Ponder, J.W. & Case, D.A. (2003) Force fields for protein simulations. *Adv Protein Chem,* 66, 27-85.

Ponder, J.W., Wu, C., Ren, P., Pande, V.S., Chodera, J.D., Schnieders, M.J., Haque, I., Mobley, D.L., Lambrecht, D.S., Distasio, R.A., Jr., Head-Gordon, M., Clark, G.N., Johnson, M.E. & Head-Gordon, T. (2010) Current status of the AMOEBA polarizable force field. *J Phys Chem B,* 114, 2549-64.

Prabhakaran, M., Harvey, S.C., Mao, B. & Mccammon, J.A. (1983) Molecular dynamics of phenylalanine transfer RNA. *J Biomol Struct Dyn,* 1, 357-69.

Rathore, N., Chopra, M. & De Pablo, J.J. (2005) Optimal allocation of replicas in parallel tempering simulations. *Journal of Chemical Physics,* 122, 024111/1-024111/8.

Richards, F.M. (1977) Area, Volumes, Packing, and Protein Structure. *Ann. Rev. Biophys. Bioeng.* , 6, 151-176.

Roe, D.R., Hornak, V. & Simmerling, C. (2005) Folding cooperativity in a three-stranded beta-sheet model. *J Mol Biol,* 352, 370-81.

Ross, W.S. & Hardin, C.C. (1994) Ion-Induced Stabilization of the G-DNA Quadruplex - Free-energy Perturbation Studies. *J Am Chem Soc,* 116, 6070-6080.

Sanbonmatsu, K.Y. & Garcia, A.E. (2002) Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins,* 46, 225-34.

Schlick, T. (2009) Molecular dynamics-based approaches for enhanced sampling of long-time, large-scale conforational changes in biomolecules. *F1000 Biology Reports,* 8, 51.

Schlick, T., Collepardo-Guevara, R., Halvorsen, L.A., Jung, S. & Xiao, X. (2011) Biomolecular modeling and simulation: a field coming of age. *Quarterly Reviews of Biophysics,* 44, 191 - 228.

Spackova, N., Berger, I. & Sponer, J. (1999) Nanosecond Molecular Dynamics Simulations of Parallel and Antiparallel Guanine Quadruplex DNA Molecules. *J Am Chem Soc,* 121, 5519-5534.

Spackova, N., Berger, I. & Sponer, J. (2001) Structural dynamics and cation interactions of DNA quadruplex molecules containing mixed guanine/cytosine quartets revealed by large-scale MD simulations. *J Am Chem Soc,* 123, 3295-307.

Spackova, N., Cheatham, T.E., 3rd, Ryjacek, F., Lankas, F., Van Meervelt, L., Hobza, P. & Sponer, J. (2003) Molecular dynamics simulations and thermodynamics analysis of DNA-drug complexes. Minor groove binding between 4',6-diamidino-2-phenylindole and DNA duplexes in solution. *J Am Chem Soc,* 125, 1759-69.

Sponer, J. & Spackova, N. (2007) Molecular dynamics simulations and their application to four-stranded DNA. *Methods,* 43, 278-90.

Srinivasan, J., Trevathan, M.W., Beroza, P. & Case, D.A. (1999) Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor Chem Acc,* 101, 426-434.

Stefl, R., Cheatham, T.E., 3rd, Spackova, N., Fadrna, E., Berger, I., Koca, J. & Sponer, J. (2003) Formation pathways of a guanine-quadruplex DNA revealed by molecular dynamics and thermodynamic analysis of the substates. *Biophys J,* 85, 1787-804.

Stefl, R., Spackova, N., Berger, I., Koca, J. & Sponer, J. (2001) Molecular dynamics of DNA quadruplex molecules containing inosine, 6-thioguanine and 6-thiopurine. *Biophys J,* 80, 455-68.

Still, W.C., Tempczyk, A., Hawley, R.C. & Hendrickson, T. (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.,* 112, 6127-6129.

Sugita, Y. & Okamoto, Y. (1999) Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters,* 314, 141-151.

Tai, K. (2004) Conformational sampling for the impatient. *Biophys Chem,* 107, 213-20.

Tidor, B., Irikura, K.K., Brooks, B.R. & Karplus, M. (1983) Dynamics of DNA oligomers. *J Biomol Struct Dyn,* 1, 231-52.

Vaiana, A.C. & Sanbonmatsu, K.Y. (2009) Stochastic gating and drug-ribosome interactions. *J Mol Biol,* 386, 648-61.

Van Der Kamp, M., Shaw, K., Woods, C. & Mulholland, A. (2008) Biomolecular simulation and modelIng: status, progress and prospects. *Journal of the Royal Society Interface,* 5, S173-S190.

Verlet, L. (1967) Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* 159, 98-103.

Wang, J., Cieplak, P. & Kollman, P. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem,* 21, 1049.

Wang, Z.-X., Zhang, W., Wu, C., Lei, H., Cieplak, P. & Duan, Y. (2006a) Strike a Balance: Optimization of Backbone Torsion Parameters of AMBER Polarizable Force Fields for Simulations of Proteins and Peptides. *J. Comput. Chem,* 27, 781-790.

Wang, Z.X., Zhang, W., Wu, C., Lei, H., Cieplak, P. & Duan, Y. (2006b) Strike a balance: optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *J Comput Chem,* 27, 781-90.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984) A New Force Field for Molecular Mechanical Simulation fo Nucleic Acids and Proteins. *J. Am. Chem. Soc.,* 106, 765-784.

Weiner, S.J., Kollman, P.A., Nguyen, D.T. & Case, D.A. (1986) An All Atom Force Field for Simulaitons of Proteins and Nucleic Acids. *J. Comput. Chem,* 7, 230-252.

Weiser, J., Shenkin, P.S. & Still, W.C. (1999) Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem,* 217-230.

Wolynes, P.G., Onuchic, J.N. & Thirumalai, D. (1995) Navigating the folding routes. *Science,* 267, 1619-20.

Zhou, R. (2003) Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins,* 53, 148-61.

Zhou, R. & Berne, B.J. (2002) Can a continuum solvent model reproduce the free energy landscape of a beta -hairpin folding in water? *Proc Natl Acad Sci U S A,* 99, 12777-82.

Zhou, R., Berne, B.J. & Germain, R. (2001) The free energy landscape for beta hairpin folding in explicit water. *Proc Natl Acad Sci U S A,* 98, 14931-6.

CHAPTER 4

INVESTIGATING THE FLEXIBLE RGG BOX DOMAINS OF SHO AND FMRP USING MOLECULAR DYNAMICS

4.1 INTRODUCTION

4.1.1 Background

While it has long been known that the function of many protein domains relies upon the protein forming a particular 3 dimensional structure, it has only more recently been recognized that for a significant proportion of the proteome the opposite is true. This class of proteins, or protein domains, require flexibility for function (Wright and Dyson, 1999, Dunker et al., 2001, Uversky, 2002, Tompa and Csermely, 2004). These intrinsically disordered or unstructured proteins have been found to have roles in cell-cycle control, signal transduction, transcriptional and translational regulation and in macromolecular complexes such as ribosomes. As observed by Keith Dunker in a recent review, disordered proteins are typically involved in regulatory processes, in contrast to well-ordered proteins such as enzymes which are primarily associated with catalysis (Dunker et al., 2008). Like many RNA binding domains (Tompa and Csermely, 2004), it is expected that the RGG box is primarily disordered, that is, it does not have a unique 3 dimensional structure.

A biological advantage of structural disorder is that it may allow interaction with a number of binding partners. This is exemplified by the RGG box of the Fragile X Mental Retardation Protein (FMRP) which can bind to an extensive range of RNA targets (Darnell et al., 2001, Darnell et al., 2005). A second advantage arises if the disordered protein becomes ordered on binding, that is, there is a disorder-to-order transition. In this scenario a ligand may bind with high specificity as a result of the number and nature of the contacts established, but low affinity because the enthalpy and entropy cost associated with ordering the protein reduces the total free energy of the interaction. This decoupling of specificity and affinity may allow a specific interaction to be readily reversible (Dunker et al., 2001). These features are obviously beneficial in signaling and regulatory processes. Another suggested advantage is that the disordered protein provides an elegant solution to the problem of how to achieve a large molecular interface yet smaller protein, genome and cell size (Gunasekaran et al., 2003). A large interaction surface provides the opportunity for segmental

association and dissociation e.g. contacts may be lost or created in a step wise fashion as the interacting species remain tethered yet still able to form new interactions with each other or additional molecules (Dunker et al., 2008).

Previous studies have shown that disordered protein domains have a high proportion of polar, and in particular, charged residues relative to hydrophobic residues (Uversky et al., 2000). This under-representation of hydrophobic residues clearly limits the potential to form the hydrophobic core seen in globular proteins.

Several algorithms are now available to predict whether a protein sequence may be disordered. The DisEMBL (http://dis.embl.de) algorithm predicts that both Sho and Fx RGG box domains are disordered under all 3 definitions of disorder used in this tool. NMR and circular dichroism techniques have confirmed that the RGG box domain of Fx is flexible and unstructured (Ramos et al., 2003). The experimental studies presented in Chapter 9 show that the Sho peptide is also unstructured.

Molecular dynamics (MD) has been used for many years now to investigate the dynamics of biological molecules and to identify their stable conformations. However, its use in exploring disordered regions of molecules is still novel. Nonetheless, the techniques that have proven useful in understanding structured domains can also be applied to better understand behaviour in a disordered region. This study abandons any preconception that disordered peptides must adopt some preferential tertiary structure, transitory or otherwise, in order to carry out its natural function. Rather, it seems plausible that a highly flexible protein domain capable of binding several different ligands will have a potential energy surface with very many shallow potential energy wells. That is, a surface that can be easily traversed as a result of the thermal motion of the system. In this way the protein will be poised to adapt to its environment, including the presence of a potential binding partner, without having to escape a particular stable conformation first.

Given the predicted flexibility of these Gly-rich peptides, methods that maximize the sampling of conformational space within a realistic timeframe are required. The replica exchange algorithm is a well-established method for enhancing conformational sampling. At the start of this work, a modification to the replica exchange method had just been published by the Simmerling group (Okur et al., 2006). This new method, known as hybrid-REMD, significantly reduces computational time by reducing the

number of replicas required, as discussed in Chapter 3. Here, the results of hybrid-REMD simulations are compared with standard REMD runs whose computational expense has been reduced by using an implicit solvation model.

4.1.2 Aims of this chapter

The aim of this work is to explore the conformational ensemble of two RGG box peptides, in particular to quantify their flexibility and compare the Sho RGG domain with the known RGG box of the Fx peptide. An additional objective is to compare the new hybrid-REMD method to standard REMD with an implicit treatment of solvation.

4.2 METHODS

4.2.1 Peptide Models

Two peptides, each consisting of 21 residues and each containing an RGG box sequence were simulated in this study. The first peptide was extracted from Sho (residues 25-46, KGGRGGARGSARGGVRGGARG), and the second from FMRP (residues 531-551, GRRRGGGGRGQGGRGRGGGFK). These sequences contain three residues before the first RGG repeat and three residues after the last RGG repeat. The peptides are set out in Figure 4-1 together with the number by which each residue is identified in this study.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Sho	K	G	G	**R**	**G**	**G**	A	R	G	S	A	**R**	**G**	**G**	V	**R**	**G**	**G**	A	R	G
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Fx	G	R	R	**R**	**G**	**G**	G	G	R	G	Q	G	G	R	G	**R**		**G**	G	F	K

Figure 4-1. RGG box peptide sequences
The 21 residue peptides used in these simulations. Sho is highlighted in green and Fx in purple. This colour scheme will be used throughout this chapter.

Both peptides have a high proportion of polar residues relative to hydrophobic residues. The Sho peptide consists of 34 % polar residues (of which 29% are charged), 48% Gly and 19% hydrophobic residues (Ala and Val). The Fx peptide consists of 38 % polar residues (of which 33% are charged), 57 % Gly residues and has only one hydrophobic residue Phe20. The grand average of hydropathicity (GRAVY) (Kyte and Doolittle, 1982) calculated using the ProtParam program (http://au.expasy.org/cgi-bin/protparam) is -1.029 for Sho and -1.733 for Fx. These values indicate that both peptides would be fully solvated in a physical system rather than being membrane associated (Kyte and Doolittle, 1982).

The DS Modeling program (Accelyrs Inc.) was used to generate starting co-ordinates via homology modeling. As the RGG peptides are disordered, these are not true homology models, but rather a mechanism for obtaining a set of starting co-ordinates. A structurally diverse set of starting structures was generated for each peptide as shown in Figure 4-2. After preliminary simulation an extended conformation was extracted to form the fourth model for use in the implicit solvation simulations. The models are referred to hereafter as Sho1, Sho2, Sho3 and Sho4 and Fx1, Fx2, Fx3 and Fx4 (collectively referred to as the Sho models and Fx models, respectively). The implicit solvation REMD simulations may be referred to as Sho-IS and Fx-IS and the hybrid-REMD simulations may be referred to as Sho-HR and Fx-HR.

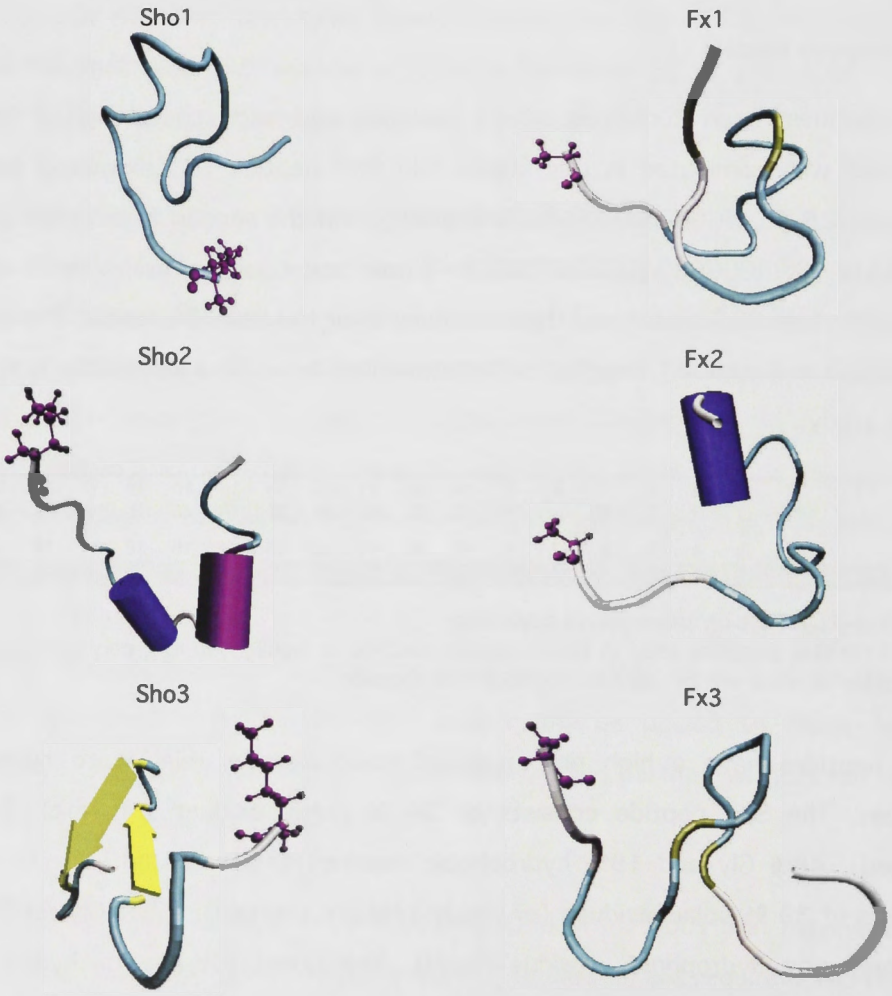


Figure 4-2. Starting models of the peptide sequences
Cartoon representations of the 21 residue peptides. LHS - Sho1, Sho2 and Sho3, RHS - Fx1, Fx2 and Fx3. The first residue of the peptide is highlighted in purple. Helical regions are represented as cylinders and β-bridges and β-strands are coloured yellow.

4.2.2 Software and force fields

All simulations were performed with the AMBER 9 (Case et al., 2005) suite of programs described in Chapter 3. The ff03 force field (Duan et al., 2003) was used for all calculations.

4.2.3 Implicit solvation REMD simulations

The replica-exchange molecular dynamics (REMD) simulation technique (Sugita and Okamoto, 1999), as implemented in AMBER, was used to sample the conformational space of the peptides. The generalized Born (GB^{OBC}) implicit solvent model was used to model the effects of solvation (Onufriev et al., 2004) using the mbondi 2 radii. Saltcon was set to 0.2 in order to simulate a 0.2 M concentration of mobile counterions (Sitkoff et al., 1994).

Following energy minimization, the MD simulations were performed using a 2 fs time step and the SHAKE algorithm was used to constrain bonds involving hydrogen (Ryckaert et al., 1977). The cutoff for non-bonded interactions was set at 20 Å. Temperature scaling was performed using Langevin dynamics with a collision frequency of 2 ps^{-1} (Loncharich et al., 1992, Izaguirre et al., 2001).

Eight replicas were created at temperatures of 300, 315, 330, 365, 405, 447, 494, and 545 K. Exchanges were attempted every 1 ps (500 steps) according to the usual Metropolis criteria as described in Chapter 3. Co-ordinates for analysis were collected every 1 ps.

4.2.4 Hybrid-REMD simulations

The hybrid-REMD method (Okur et al., 2006) as implemented in AMBER was used to perform these calculations. During the MD simulations the peptides were solvated in a 20 Å radius octahedron of TIP3P (Jorgensen et al., 1983) water molecules. At the replica exchange step only water molecules within a 1.5 Å radius of the peptides were retained. Different initial conformations resulted in a different number of water molecules in each model. Counterions (Cl^-) were added to balance the positively charge Arg residues and to create a charge neutral system.

The hybrid-REMD simulations were performed using 8 replicas, over a temperature range 300 - 533 K (300, 330, 360, 390, 420, 450, 473, 533). To equilibrate the

system, all models were first minimized and then each replica was equilibrated by slowly raising the temperature from 0.1 K to the desired temperature, at constant volume.

The simulations were run in the NPT ensemble for at least 30 ns. Temperature was kept constant using the weak coupling algorithm (Berendsen et al., 1984). The time constant for heat bath coupling was set to 0.1 ps. Pressure was kept constant using an analogous 'weak coupling' algorithm implemented in AMBER (ntp=1) and a pressure relaxation time of 2 ps. The simulations were carried out using a 2 fs time step with the SHAKE algorithm used to constrain bonds involving hydrogen (Ryckaert et al., 1977). Periodic boundary conditions were imposed. The particle-mesh Ewald (PME) procedure was used to calculate long-range electrostatic interactions, the direct space sum was limited with a cutoff of 10 Å which was also the cutoff for van der Waals interactions. Replica exchange was attempted every 1 ps. Co-ordinates for analysis were collected every 1 ps.

4.2.5 Trajectory analysis

The following nomenclature will be used in this study. The simulations started from each of the different starting models is referred to a Sho or Fx simulation; accordingly there are three Sho simulations and three Fx simulations conducted with hybrid-REMD solvation and four Sho and four Fx simulations conducted with implicit solvation. Where average data for the simulations is presented the results for the three or four simulations of Sho or Fx are averaged. At times Sho and Fx may be referred to as a system.

The implicit solvation simulations of Fx were run for approximately 240 ns per replica (total simulation time: 240 ns x 8 reps x 4 models = 7.68 μs), while the Sho simulations were allowed to run for approximately 540 ns per replica (total simulation time: 540 ns x 8 reps x 4 models = 17.28 μs). The first 40 ns of simulation time (of each model) have been excluded from analysis.

The hybrid-REMD simulations were run for at least 30 ns for Fx and Sho (total simulation time per system: 30 ns x 8 reps x 3 models = 720 ns). For all simulations, unless otherwise specified, only those co-ordinate frames produced at 300 K were used and the first 5 ns of the time (of each model) have been excluded from analysis.

Trajectories were analysed using the ptraj program in AMBER 9 and my own custom-written python scripts. Ptraj was used to calculate the distance between particular atoms and residues, the radius of gyration, hydrogen bond formation, dihedral angles and elements of secondary structure (DSSP) for all frames analysed. Hydrogen bonds were calculated using a distance cutoff of 3.5 Å between the heavy atoms and an angle cutoff of 120° (the AMBER default).

Cluster analysis was performed using the perl scripts available in the MMTSB toolset (Feig et al., 2004). For the cluster analysis every 5th frame was collected from all the hybrid-REMD simulations and clustering was performed using the kclust algorithm with a fixed cluster radius. The radius was based on the backbone RMSD of the peptides and fixed at 5 Å. Prior to cluster analysis, the collected frames were sorted into sets based on radius of gyration.

The DSSP (Define Secondary Structure of Proteins) algorithm developed by Kabsch and Sander which relies on the pattern of backbone hydrogen bonds to predict protein structure (Kabsch and Sander, 1983) was used in this study to compare the structural predictions for the implicit solvation and hybrid-REMD simulations. In this method, a turn is defined as a hydrogen bond involving the backbone carbonyl CO(i) and the backbone amide NH(i + n) where n=3, 4, 5. Two consecutive turns are required to identify a helix. A β-bridge is indicated by 2 hydrogen bonds formed between 2 non-overlapping stretches comprised of 3 residues. Where this structure is found in consecutive residues the extended β structure (β-strand) is identified.

Python scripts were created to sort and analyze data and to create the frequency distributions presented in the results section.

4.3 RESULTS AND DISCUSSION

The bulk of the work presented in this chapter focuses on exploring the conformational ensembles of the Sho and Fx peptides and investigates the functional implications of any similarities or differences between them. As a preface to the main discussion the convergence of the simulations and effect of solvation methods will be discussed briefly.

4.3.1 Convergence

Convergence was tested by evaluating and comparing the distribution of distances between the guanidinium groups (e.g. Arg4-Arg8, Arg8-Arg12, Arg12-Arg16, Arg4-Arg16 for Sho) for each of the starting models. Although the Arg residues are expected to explore a range of conformational space, resulting in a broad distribution of Arg-Arg distances, the shape of these distributions will stop changing when the simulations are converged. The degree of overlap of the distributions derived from different initial conformations is therefore an indicator of comprehensive sampling, or 'convergence'. By way of example, the distributions of the Arg4-Arg9 distance in hybrid-REMD simulations of the Fx models are shown in Figure 4-3. Distributions generated over the first nanosecond (Figure 4-3(C)) show 74% overlap between the 3 models while distributions generated from the final 25 ns (Figure 4-3(D) show an overlap of 93 %.

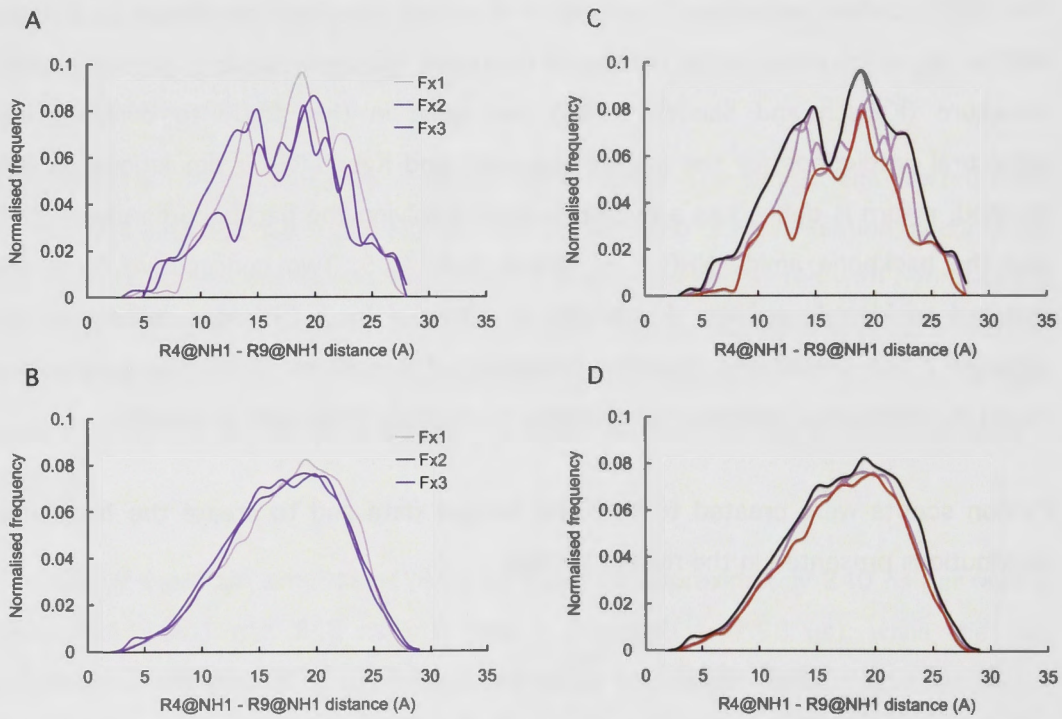


Figure 4-3. Overlap in Arg4-Arg9 distance distributions during Fx simulation
Distribution of Arg4-Arg9 distances for 3 different starting models over the (A) first 1 ns of hybrid-REMD simulation and (B) last 25 ns. Intersection of the 3 distributions ('overlap') traced by the red line at (C) 1 ns ~ 74% (D) final 25 ns ~ 93%.

Snapshots from the REMD trajectories provide a visual representation of how the Arg4-Arg9 distance fluctuates over time (Figure 4-4).

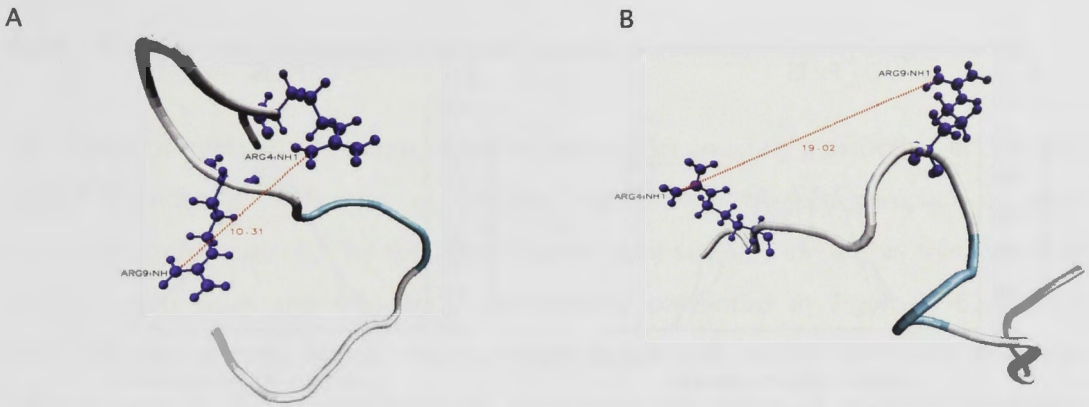


Figure 4-4. Fluctuation in the Arg4-Arg9 distance during an Fx simulation

Frequency distributions for the distance between the first and second Arg residue in the RGG box domains of each of the peptides using the different solvation methods are presented in Figure 4-5.

Comparing the 1 ns and 25 ns shows how the overlap of the distributions increases as the conformational ensemble converges. The broad distribution of the Arg-Arg distances, seen in Figure 4-5, indicates that a wide range of conformational space has been explored over the course of the simulations. The high degree of overlap seen in the distributions generated from different initial conformations provides some reassurance that the conformational space of these peptides has been comprehensively explored.

The percentage overlap of the Arg-Arg distance distributions from both the hybrid-REMD simulations and the implicit solvation simulations are presented in Table 4-1.

Table 4-1 Arg-Arg distribution overlap for 3 starting models in the hybrid-REMD simulations and 4 starting models in the implicit solvation REMD simulations

	Arg4-Arg8	**Arg8-Arg12**	**Arg12-Arg16**	**Arg4-Arg12**	**Arg4-Arg16**
Sho_HR[a]	89%	96%	84%	88%	80%
Sho_IS[b]	93%	81%	80%	89%	82%
	Arg4-Arg9	**Arg9-Arg14**	**Arg14-Arg16**	**Arg4-Arg14**	**Arg4-Arg16**
Fx_HR [a]	93%	95%	90%	90%	83%
Fx_IS [b]	91%	91%	93%	91%	91%

[a] HR = hybrid-REMD, distributions generated over the last 25 ns, [b] IS = implicit solvation, distributions generated over the last 200 ns (Fx) and 500 ns (Sho)

For all subsequent analysis, results obtained using different starting models will be amalgamated. Hybrid-REMD data will be extracted from the last 25 ns of each simulation, while Fx and Sho implicit solvent REMD will use the last 200 ns and 500 ns of each trajectory.

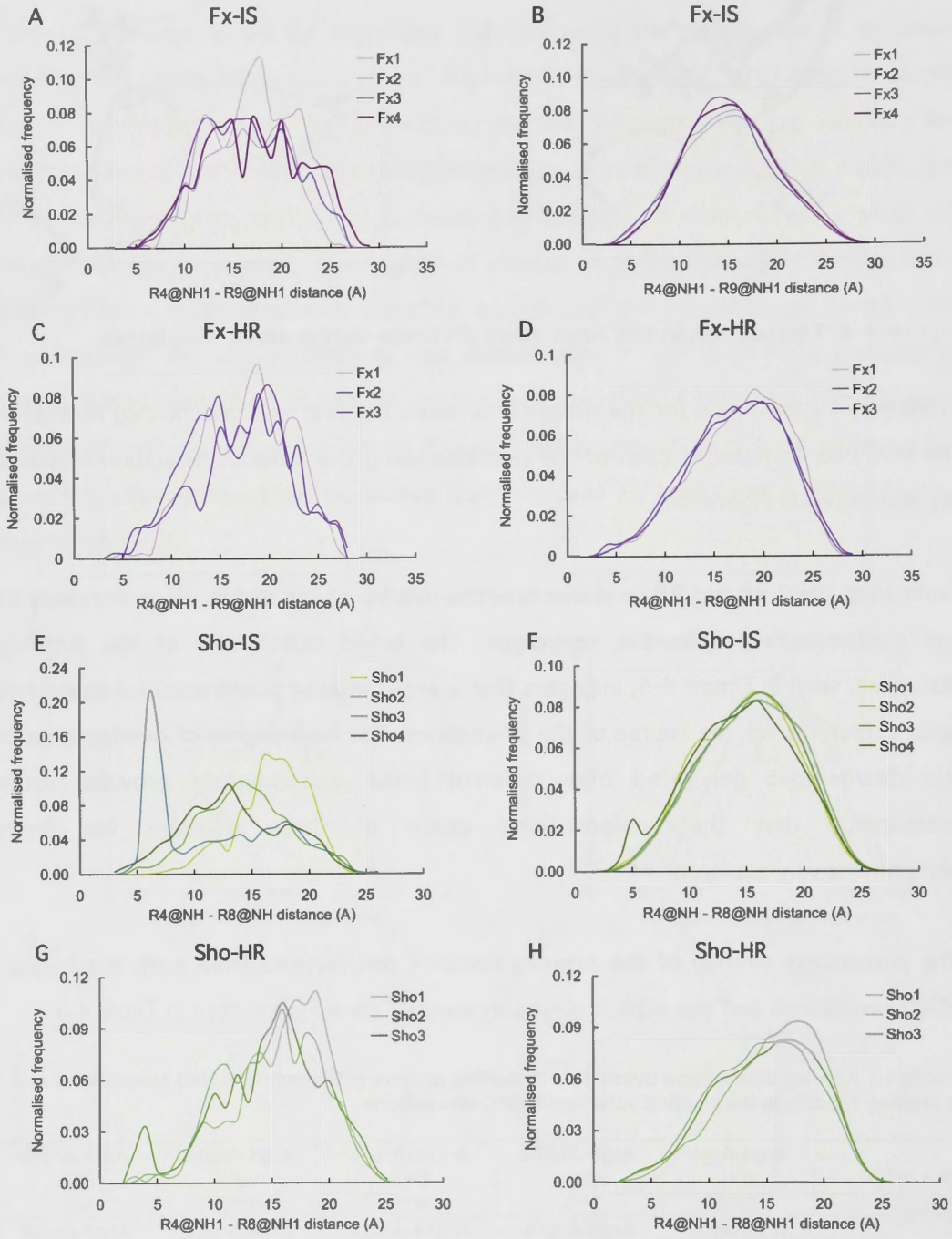


Figure 4-5. Frequency distributions of Arg4-Arg8 (Sho) and Arg4-Arg9 (Fx) distances

(A-D) Frequency distributions of the distance between the NH1 atoms of Arg4-Arg9 (Fx) and (E-H) Arg4-Arg8 (Sho). Distributions for the 1st ns of simulation (A, C, E, G) are compared with distributions over the entire simulation (B, D, F, H).

4.3.2 Comparison of solvation methods – global parameters and hydrogen bonds

The radius of gyration is a measure of the volume occupied by a structure and, in this study, provides an indication of whether the peptide favours compact or more extended conformations. The solvation method used strongly influences this metric as can be seen from the frequency distributions presented in Figure 4-6. Implicit solvation (lighter hue) favours more compact structures relative to explicit solvation (darker hue). In the implicit solvation simulations the radius of gyration converges towards a peak around 7.5 Å for both peptides. By contrast, the hybrid-REMD simulations move towards a broad distribution centered around an average of 10.4 Å for Sho and 11.5 Å for Fx.

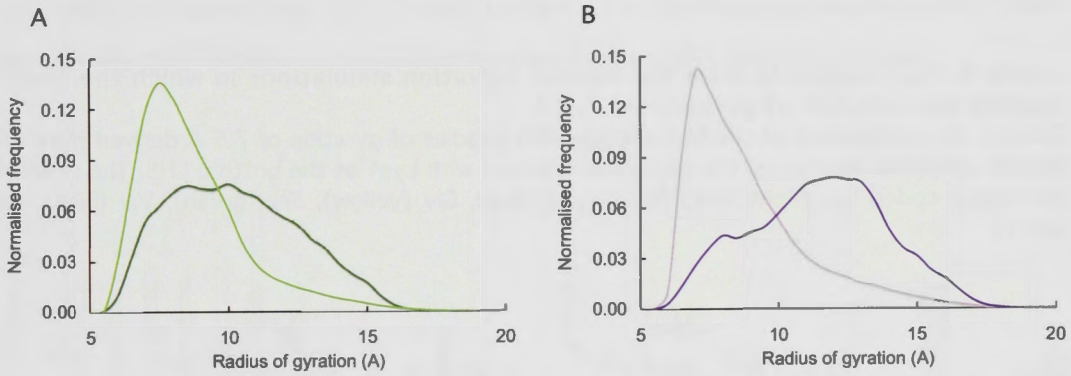


Figure 4-6. Radius of gyration of Sho and Fx with different solvation methods
Frequency distributions of the radius of gyration (A) Sho and (B) Fx, comparing implicit solvation REMD (light hue) and hybrid-REMD (dark hue).

This indicates a definite propensity for both peptides to assume more compact conformations when simulated with an implicit model of solvation, although the nature of the conformations formed is extremely variable. Comparing snapshots taken from the implicit solvation simulations where the peptide has a radius of gyration of 7.5 Å (Figure 4-7a) with snapshots taken from the hybrid-REMD simulations in which the Sho peptide has a radius of gyration in the range 10 Å to 11 Å (Figure 4-7b) illustrates the difference between the compact structures favoured by implicit solvation and the more extended structures which emerge from explicit solvation. In both cases, the flexible peptides can adopt a large range of different conformations within the general regions favoured by each solvation model.

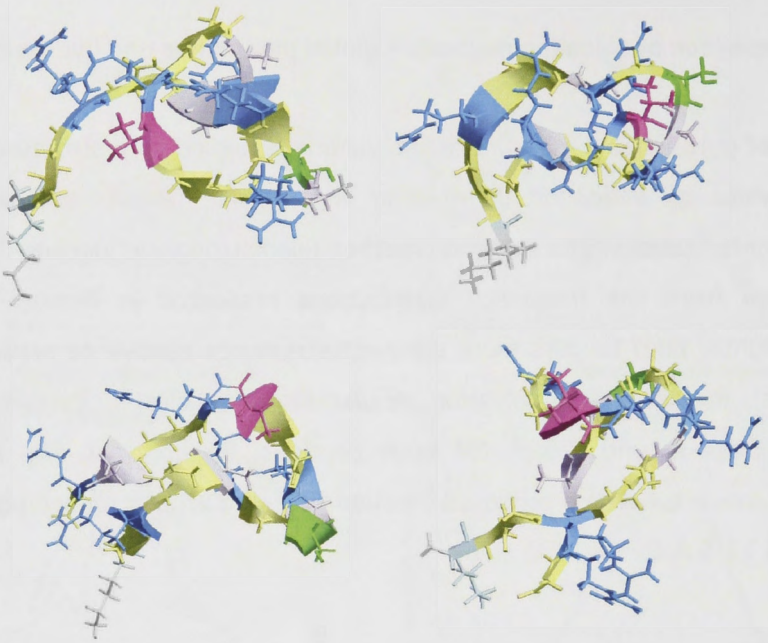


Figure 4-7a. Snapshots from the implicit solvation simulations in which the Sho peptide has a radius of gyration of 7.5 A
Cartoon representations of the Sho peptide with a radius of gyration of 7.5 A derived from the implicit solvation simulation. The peptide is oriented with Lys1 at the bottom LHS. The residues are colour coded Lys (light blue) Arg (bright blue), Gly (yellow), Ser (green), Val (pink), Ala (grey).

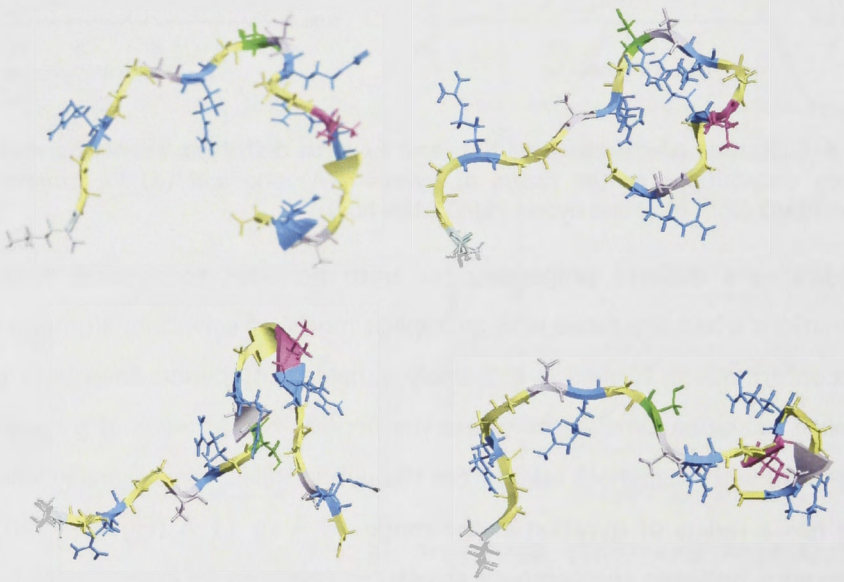
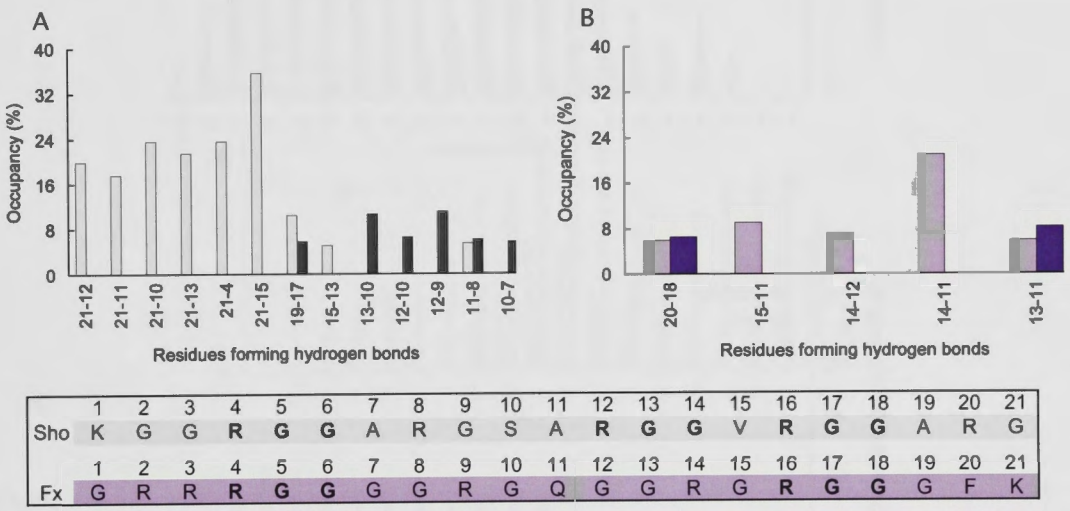


Figure 4-7b. Snapshots from the hybrid-REMD simulations in which the Sho peptide has a radius of gyration in the range 10-11 Å
Cartoon representations of the Sho peptide with a radius of gyration in the range 10-11 Å derived from the hybrid-REMD simulation. The peptide is oriented with Lys1 at the bottom LHS. The residues are colour coded Lys (light blue) Arg (bright blue), Gly (yellow), Ser (green), Val (pink), Ala (grey).

This differences in the distributions of the radius of gyration in the hybrid-REMD and implicit solvation simulations is related to a difference in the formation of backbone hydrogen bonds in the two solvation models. All the backbone hydrogen bonds formed more than 5% of the time in each of the models are set out in Table A4-1 (hybrid-REMD) and Table A4-2 (implicit solvation REMD) of Appendix 4. These data are summarized in Figure 4-8.

Overall, there is relatively low occupation of backbone hydrogen bonds under both solvation models. This does not mean that hydrogen bonds are not formed. They are formed, but only transiently. In fact, each residue forms hydrogen bonds with every other residue with the exception of those immediately next to it, for at least a small percentage of the time. However no permanent hydrogen bonds are formed. Instead, the hydrogen bonds that are formed behave in a promiscuous manner with bonds being made then broken and new bonds being formed with another partner.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Sho	K	G	G	**R**	**G**	**G**	A	R	G	S	A	**R**	**G**	**G**	V	**R**	**G**	**G**	A	R	G
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Fx	G	R	R	**R**	**G**	**G**	G	G	R	G	Q	G	G	R	G	**R**	**G**	**G**	G	F	K

Figure 4-8. Comparing backbone hydrogen bond occupancy in implicit solvation and hybrid-REMD simulations of Sho and Fx
Backbone hydrogen bond occupancy for implicit solvation simulations (lighter hue) and hybrid-REMD simulations (darker hue); (A) average of all Sho models (green), (B) average of all Fx models (purple). Only hydrogen bonds occupied at least 5% of the time are counted.

While both solvation models reveal the flexible nature of the peptides through the lack of permanent hydrogen bonds, implicit solvation tends to favour the formation and retention of hydrogen bonds relative to hybrid-REMD. During implicit solvation simulations a greater number of hydrogen bonds are formed over 5% of the time in both peptides and in Sho there is a noticeable increase in the maximum occupancy, up to approximately 35% for some bonds compared to a maximum of around 10% in the hybrid-REMD simulations.

The DSSP (Define Secondary Structure of Proteins) algorithm developed by Kabsch and Sander which relies on the pattern of backbone hydrogen bonds to predict protein structure (Kabsch and Sander, 1983) was used to compare the structural predictions for the implicit solvation and hybrid-REMD simulations. The frequencies at which each of the residues of Sho and Fx are predicted to form a turn, helix or β-bridge under each of the solvation methods are plotted in Figures 4-9 and 4-10.

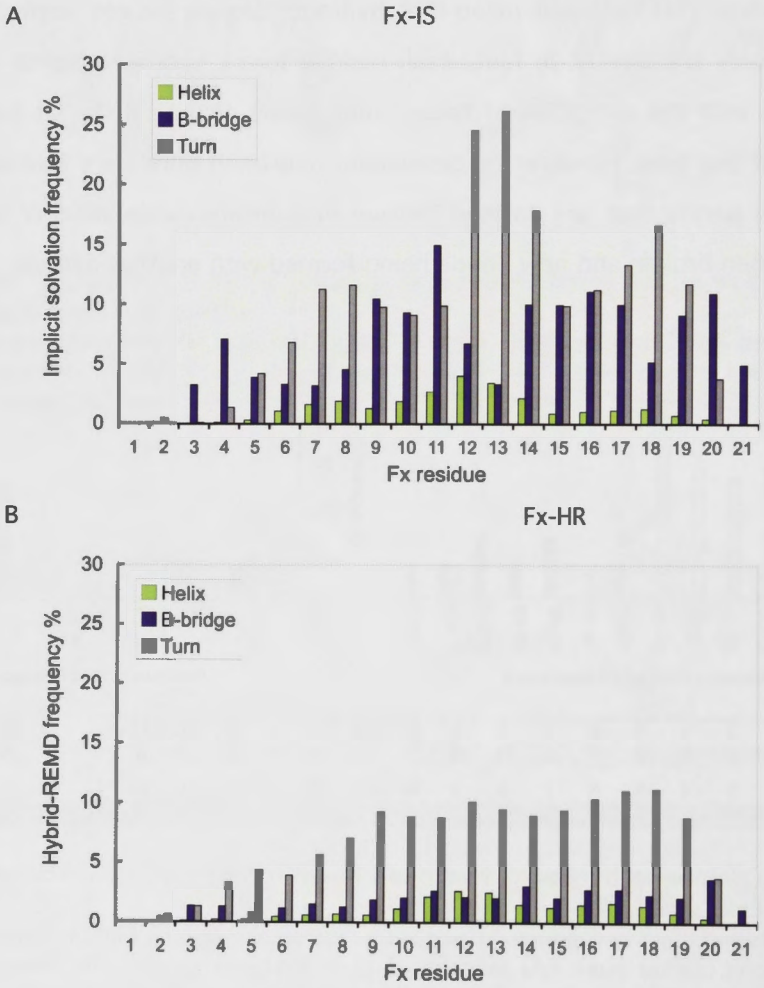


Figure 4-9. Comparison of DSSP prediction of structural elements in Fx in (A) implicit solvation REMD and (B) hybrid-REMD simulations

Overall, the turn and the β-bridge are the most common elements in the implicit solvation simulations. These elements are commonly found in unstructured coil regions of proteins (Swindells et al., 1995). The most notable difference between the implicit solvation and hybrid-REMD results is that implicit solvation predicts higher

β-bridge formation in both peptides. There is also a higher number of turns in Gly12-Gly13 in Fx.

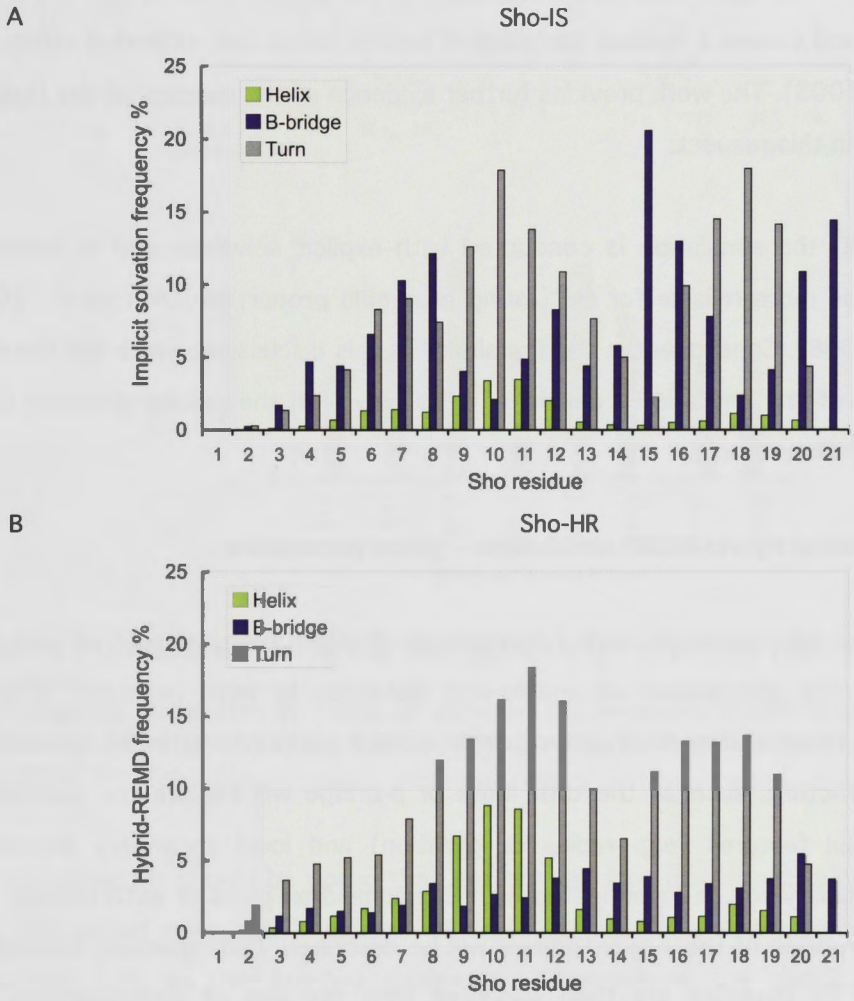


Figure 4-10. Comparison of DSSP prediction of structural elements in Sho in (A) implicit solvation REMD and (B) hybrid-REMD simulations

In summary, the generalized Born implicit solvation model favours the formation of backbone hydrogen bonds when compared with explicit solvation simulations. This finding is consistent with the conclusion reached by the Simmerling group, that backbone hydrogen bonds are overstabilized in the generalized Born model available in AMBER, including GB^{OBC} used in these simulations. Overstablization arises because the Born radii for certain backbone atoms are underestimated leading to an insufficient desolvation penalty when these atoms form hydrogen bonds. This results in a greater propensity for turns and more compact structures including α-helical structure (Roe et al., 2007). It is interesting to see that the implicit solvation simulations performed here do not produce a greater sampling of α-helical space compared to the hybrid-

REMD simulations. Zhou suggested that the known problem of exaggerated sampling of α-helical space in peptide simulations stems more from the force field parameters for backbone torsions than the implicit solvation method (Zhou, 2003). One of the principal aims of the Duan et al. force field, used in the current studies, was to remedy this problem and ensure a realistic sampling of both α-helcial and extended structures (Duan et al., 2003). The work provides further evidence of the success of the Duan et al. force field in this respect.

In hybrid-REMD the simulation is conducted with explicit solvation and is therefore expected to be more reliable for calculating ensemble properties (Okur et al., 2006, Okur et al., 2008). Consequently, the remainder of this discussion, which will focus on a comparison of the Sho and Fx peptides, deals only with the results obtained using the hybrid-REMD method.

4.3.3 Analysis of Hybrid-REMD simulations – global parameters

This discussion will commence with a comparison of global features such as radius of gyration and the distribution of end-to-end distances in both peptides. A DSSP analysis that predicts the likelihood of each residue participating in an element of secondary structure, such as the turn, helix or β-bridge will explore the correlation between global features (e.g. radius of gyration) and local secondary structure. Backbone flexibility will be investigated with Ramachandran plots of each residue and the torsional motion of the Arg sidechains will be described. Distinguishing features in the Sho and Fx peptides are then surveyed with the aim of understanding the underlying cause of any differences observed in ensemble properties. In this context, the different composition of the 2 sequences, particularly the presence of the hydrophobic residues in Sho will be discussed. The results of these investigations will then be summarized as a comparison of Sho and Fx.

From the results already presented, it is clear that both Sho and Fx have a broad distribution in the radius of gyration at physiologically relevant temperatures with few backbone hydrogen bonds occupied more than 5% of the simulation time. To assist with further discussion, hybrid-REMD results of these parameters (averaged over all 3 starting models) are presented in Figure 4-11. No data for sidechain-backbone hydrogen bonds are presented as these occur less than 5% of the time.

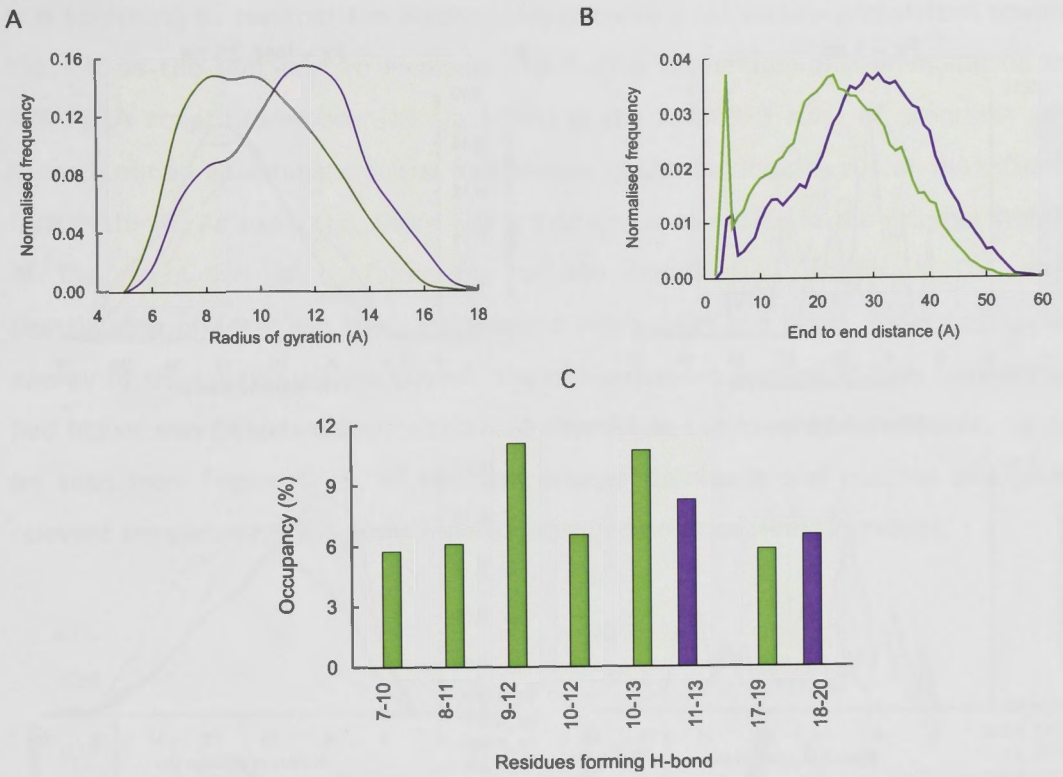


Figure 4-11. Global parameters of Sho (green) and Fx (purple) from hybrid-REMD simulations
(A) Frequency distribution of radius of gyration. (B) Frequency distribution of end-to-end distance. (C) Backbone hydrogen bond occupancy where hydrogen bonds are formed greater than 5% of the time.

At the outset, it should be noted that the distributions of the radius of gyration are less converged over the 3 models of each system than was observed for other measures such as the Arg-Arg distances. Nonetheless, the distributions of the trajectories at 300 K are clearly moving towards convergence with the overlap in the distributions increasing from 49% at 1 ns to 75% over the last 25 ns of the simulation for Fx and from 41 % to 75% for Sho as shown in Figure 4-12.

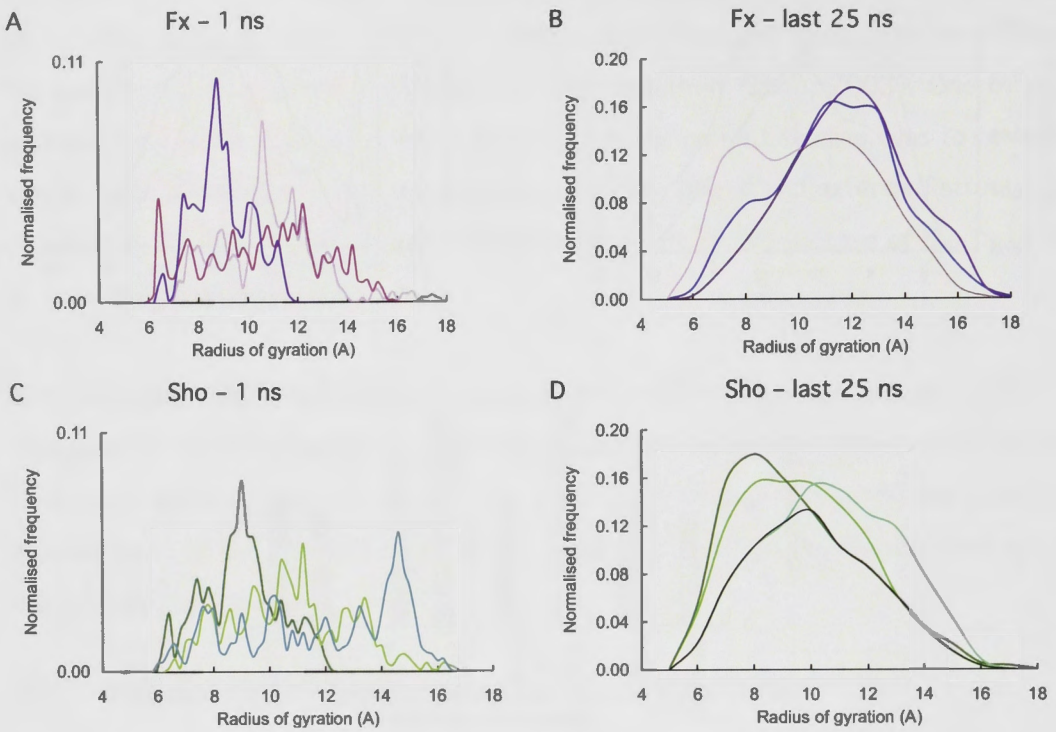


Figure 4-12. Comparing distributions of radius of gyration for the 3 starting models of Sho and Fx over the first 1 ns and over the last 25 ns of simulation
(A-B) Distributions for the 3 Fx models over 1 ns have an overlap of 36% while the distributions for the last 25 ns of simulation have an overlap of 75% (depicted by the black line). (C-D). Distributions for the 3 Sho models over 1 ns have an overlap of 41% while the 3 distributions over the last 25 ns of simulation have an overlap of 75% (depicted by the black line).

After 30 ns simulation, the range of values has converged and it appears that the Fx models are moving towards distributions centered around a mean of 11.5 Å while Sho is trending towards a distribution that is skewed towards slightly lower values with a mean of 10.5 Å (Figure 4-12(D)). It is interesting to note that the test of convergence over 3 different starting models is a good deal more stringent than the test of convergence within a single model. Distributions of the radius of gyration for the first half and second half of the simulations of each model were very similar with overlaps ranging from 88% to 93%.

The distribution of radius of gyration was calculated for each of the replicas in the expectation that the trajectories run at higher temperatures would show a greater degree of convergence, as observed in other studies (Zhang et al., 2005). The areas of overlap in the distributions generated at different temperatures are plotted in Figure 4-13. The overlap in distributions increases slightly as temperature increases but then decreases slightly at the 3 highest temperatures.

It is surprising to see that the distributions become more narrow and shifted towards the left as the temperature increases. To further investigate this phenomenon the MM_PBSA scripts (Srinivasan et al., 1998) in the AMBER 9 suite of programs were used to obtain an estimate of the free energy of the trajectories run at the different temperatures. As expected, there was a systematic decrease in the internal stability of the population of conformations as the temperature increased. The major destabilizing effect came from the internal energy, i.e. the bond, angle and torsion energy of the conformations. Overall, the conformations formed at high temperature had higher energies compared with those formed at the lower temperatures. As can be seen from Figure 4-13, as the temperature decreases and reaches biologically relevant temperatures the population of extended conformations increases.

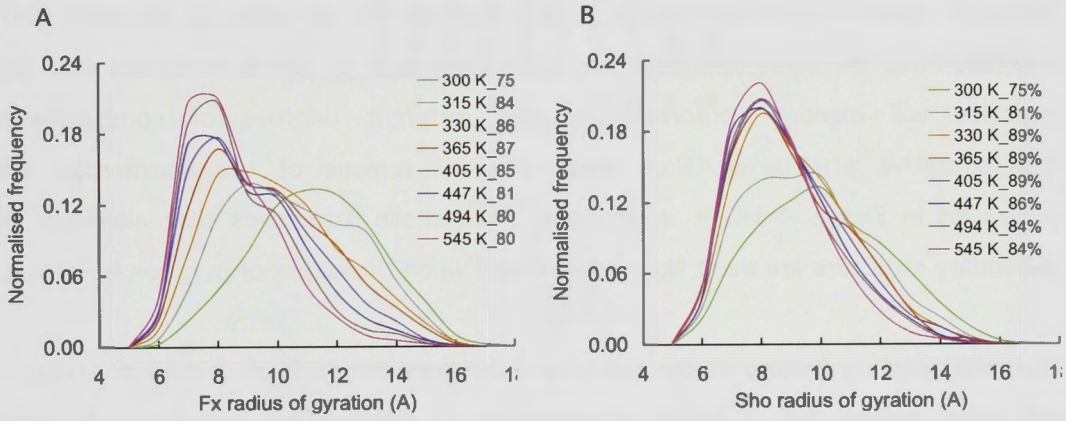


Figure 4-13. Radius of gyration distributions over replica temperature range
Only the area of overlap of the 3 starting models is plotted. The percentage overlap is specified in the legend, (A) the Fx replicas (B) the Sho replicas.

The results presented above are interesting and initially unexpected. We are accustomed to seeing a broad distribution of conformations, particularly extended conformations, at higher temperatures where the molecule of interest has sufficient kinetic energy to break out of stable compact conformations. In this case, we see the opposite, that is, a substantial increase in the population of more compact structures at higher temperatures while at the lower temperatures there is a more even distribution in the radius of gyration. In these highly flexible peptides the number of ways in which the peptide can collapse into a compact structure vastly outnumbers the ways in which an extended conformation may be formed. These results suggest that at high temperatures there is sufficient energy to allow the peptide to explore rotation through a great range of backbone dihedral angles, many of which are

energetically unfavourable, while at lower temperatures more stable structures are preferred which include compact and extended conformations.

While this trend can be seen in both Sho and Fx it is more pronounced in Fx. It is possible that the distributions at 300 K would become more similar if the simulations were run for longer, however, the results obtained at this point indicate the propensity for Fx to find more extended conformations, at biologically relevant temperatures, relative to Sho. The average distributions presented in Figure 4-12 indicate that Sho has more conformations with a radius of gyration in the range 8-10 Å while Fx has more conformations in the range 10-14 Å. The end-to-end distance plot also indicates that Fx assumes more extended conformations.

Although there is some difference in the distributions of radius of gyration both peptides have the same broad range of this parameter at 300 K indicating that the peptides will explore conformations with different degrees of compactness. Representative structures taken from different regions of the distribution are portrayed in Figure 4-14. It is apparent from these structures that elements of secondary structure are most likely to be found when the radius of gyration is < 11 Å.

The most striking feature of the end-to-end distance distributions (Figure 4-11(B)) is the spike around 3-6 Å in an otherwise broad distribution. This effect is more prominent in Sho where the spike comprises 6% of the frames (averaged over the 3 models) compared with 3% in the case of Fx, and is consistent with the higher proportion of conformations with a radius of gyration at the lower end of the scale.

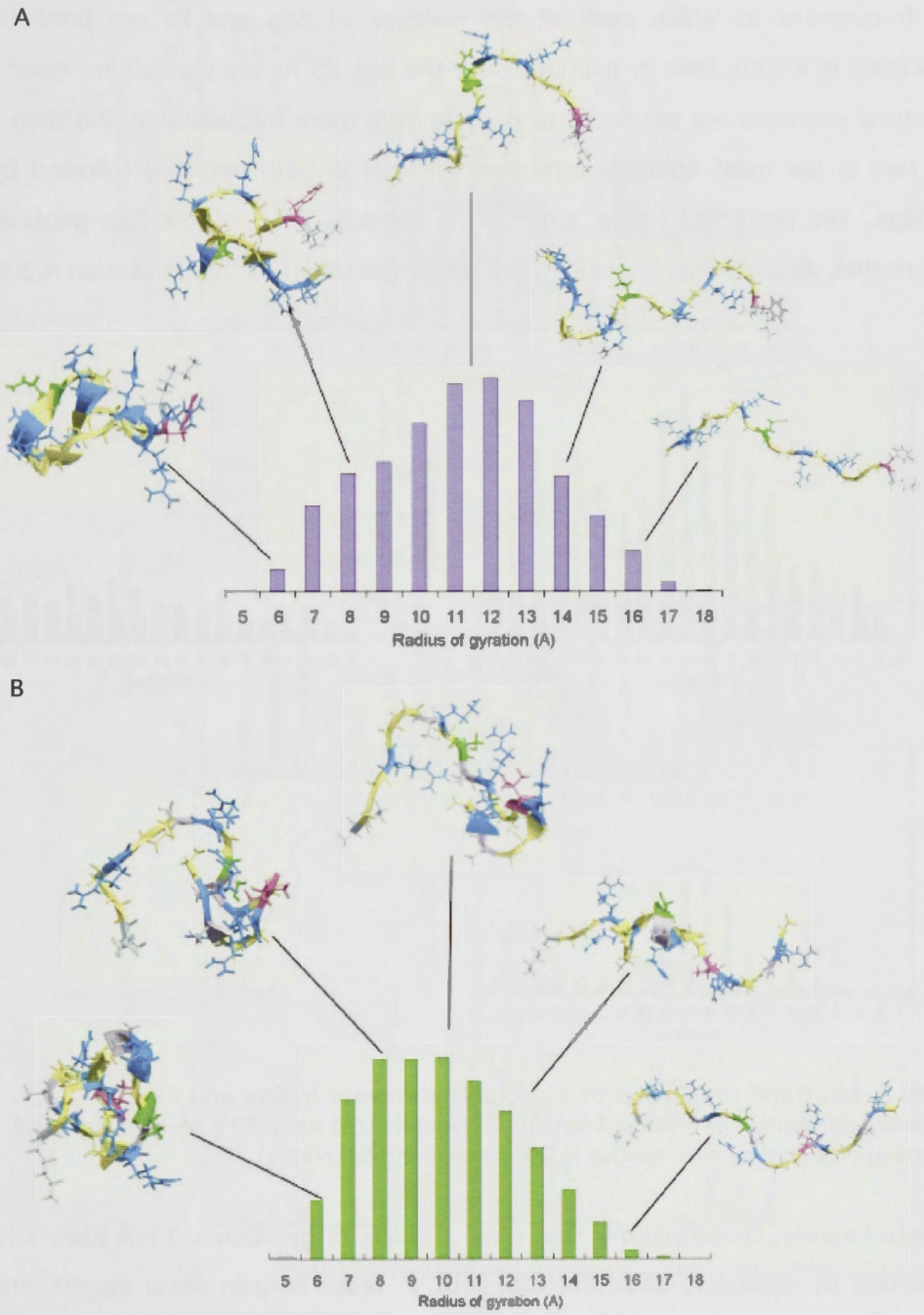


Figure 4-14. Examples of compact, folded and extended conformers for (A) Fx and (B) Sho across the range of radius of gyration
Frequency histogram of the radius of gyration with associated snapshots

4.3.4 DSSP analysis of secondary structure

The frequencies at which each of the residues of Sho and Fx are predicted to participate in a turn, helix or β–bridge over the last 25 ns are plotted in Figure 4-15. Structural elements are predicted to occur a little more frequently in Sho than in Fx. The turn is the most common structural element in both peptides followed by the β–bridge. The predicted helical structure in residues 7-13 of the Sho peptide is a feature that distinguishes Sho and Fx, as will be discussed further in section 4.3.8.

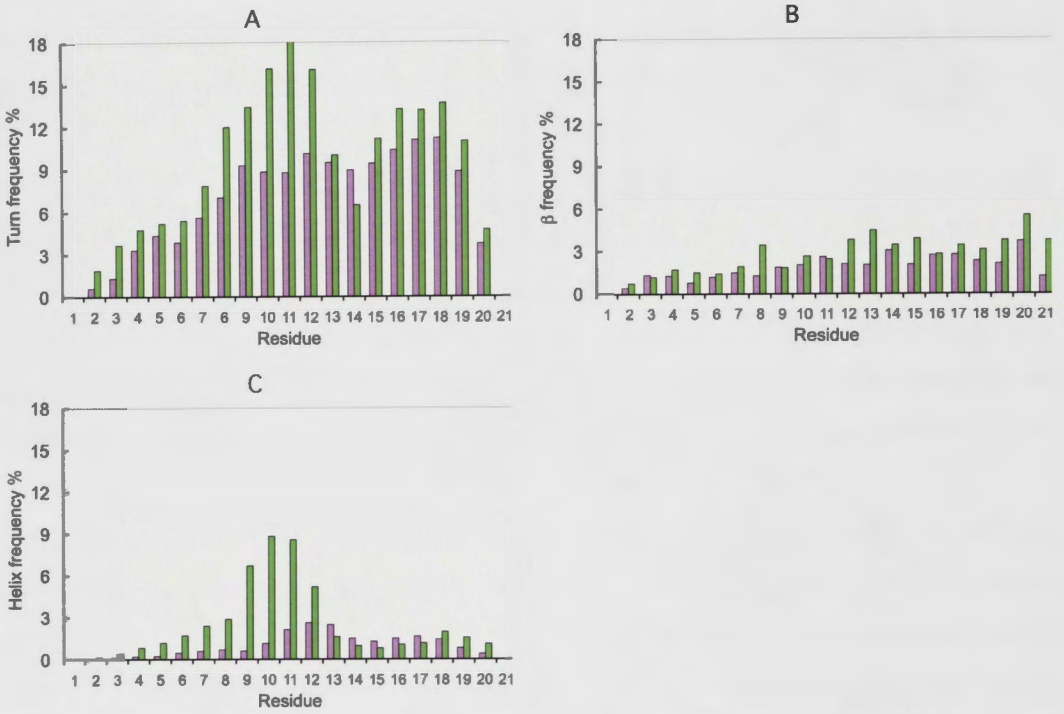


Figure 4-15. DSSP prediction of structural elements in Sho and Fx
Frequency distribution of predicted structural elements ((A) turn, (B) β-bridge, (C) helical structure) accessed by each residue in Sho (green) and Fx (purple).

As noted earlier, conformations that have a radius of gyration < 11 Å have a higher proportion of secondary structure elements. In order to gain some insight into the type of secondary structure in these conformations all the frames generated in the simulation of Sho1 were divided into groups based on radius of gyration and DSSP analysis was performed on each group. The results are plotted in Figure 4-16. The 3 plots of Figure 4-16 compare conformations with radius of gyration 5-7 Å (pink) 7-9 Å (yellow) and 9-11 Å (blue) with the entire simulation (dark green).

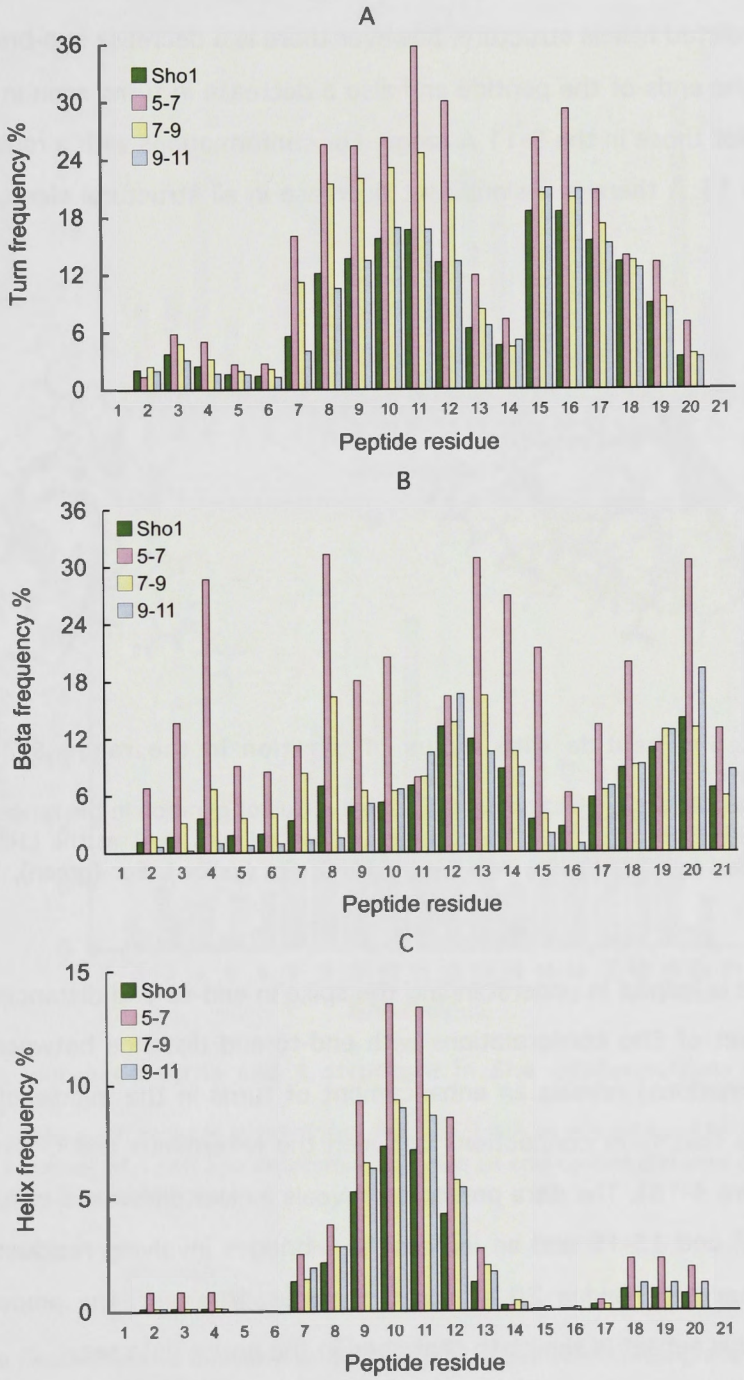


Figure 4-16. DSSP analysis based on radius of gyration for one model of Sho
Frequency of (A) turns, (B) β structure and (C) helical structure in groups of conformations with a radius of gyration 5-7 Å (pink), 7-9 Å (yellow) and 9-11 Å (blue) compared to the entire simulation (dark green).

The prominent structural feature in the most compact conformations, those where the radius of gyration is 5-7 Å (pink), is the β-bridge that forms connections between non-contiguous parts of the peptide as demonstrated by the snapshots in Figure 4-18. In addition there is some enrichment of turns and helical structure in this subset.

Conformations with a radius of gyration in the range 7-9 Å and 9-11 Å (blue) are similar in predicted helical structure, however there is a decrease in β-bridges involving residues at the ends of the peptide and also a decrease in turns seen in the middle of the peptide for those in the 9-11 Å range. For conformations with a radius of gyration greater than 11 Å there is a significant decrease in all structural elements (data not shown).

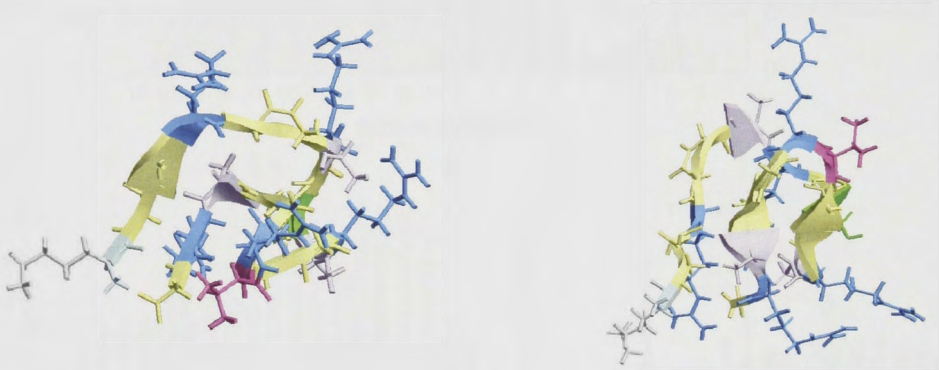


Figure 4-17. Sho peptide with radius of gyration in the range 5-7 A and beta hydrogen bonds
Cartoon representations of the Sho peptide with a radius of gyration in the range 5-7 Å derived from the hybrid-REMD simulation. The peptide is oriented with Lys1 at the LHS. The residues are colour coded Lys (light blue) Arg (bright blue), Gly (yellow), Ser (green), Val (pink), Ala (grey).

DSSP analysis is helpful in understanding the spike in end-to-end distances between 3-6 Å. A subset of Sho conformations with end-to-end distance between 3-6 Å (n = 1286 conformations) reveals an enhancement of turns in the middle of the peptide and β-bridges that form connections between the N-terminus and C-terminus of the peptide (Figure 4-18). The data presented reveals a clear difference in turn activity in residues 8-12 and 15-16 and an increase in β-bridges involving residues 4 and 8 at the N-terminus and residue 20 at the C-terminus. However, the amount of helical structure in this subset is similar to that seen in the entire data set.

Hydrogen bond analysis reveals that 2 hydrogen bonds are enriched in this subset, Ser10-Gly13 is found 21% of the time, compared with around 11% when the whole ensemble is analyzed (Figure 4-11) and Gly21-Gly2 which is found 12% of the time but < 5% in the whole ensemble. Overall, it seems that the peak in the end-to-end distance distribution is due to a combination of turns in the middle of the peptide and fluctuating β-bridges that bring the N-terminus and C-terminus of these short peptides close together.

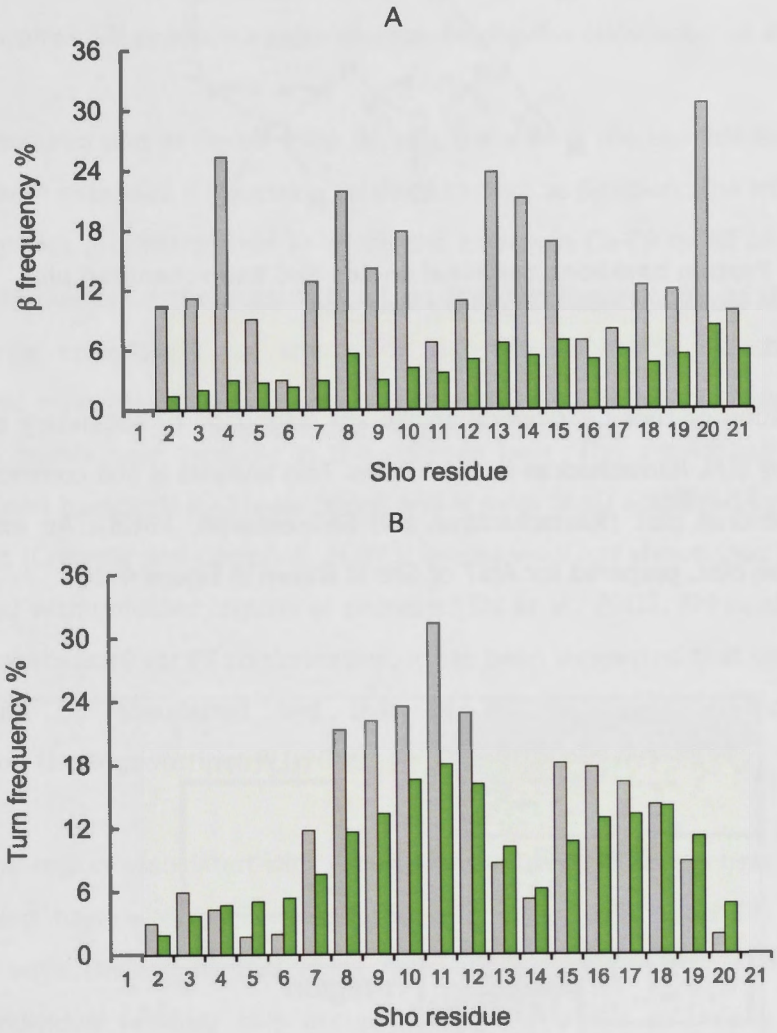


Figure 4-18. Increased turns and β structure in Sho conformations with end-to-end distance 3-6 Å
Frequency at which each residue is predicted to participate in a β-bridge (A) or turn (B) are compared for a subset of 1286 Sho conformations with an end-to-end distance of 3-6 Å (grey) and the entire Sho ensemble (green).

4.3.5 Peptide flexibility and dihedral angle analysis - Ramachandran plots

In essence, peptide flexibility arises from torsional motion about the peptide backbone, which can be assessed by calculating the distribution of backbone dihedral angles accessed by each residue. As illustrated in Figure 4-19, the ϕ angle involves rotation around the N-C_α bond which determines the range of distances between C' atoms whereas the ψ angle involves rotation around the C_α-C bond and determines the range of distances between N atoms of consecutive residues.

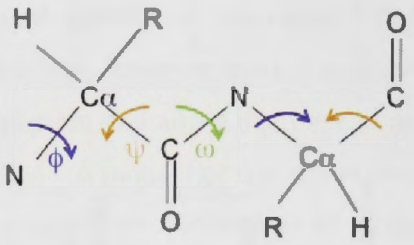
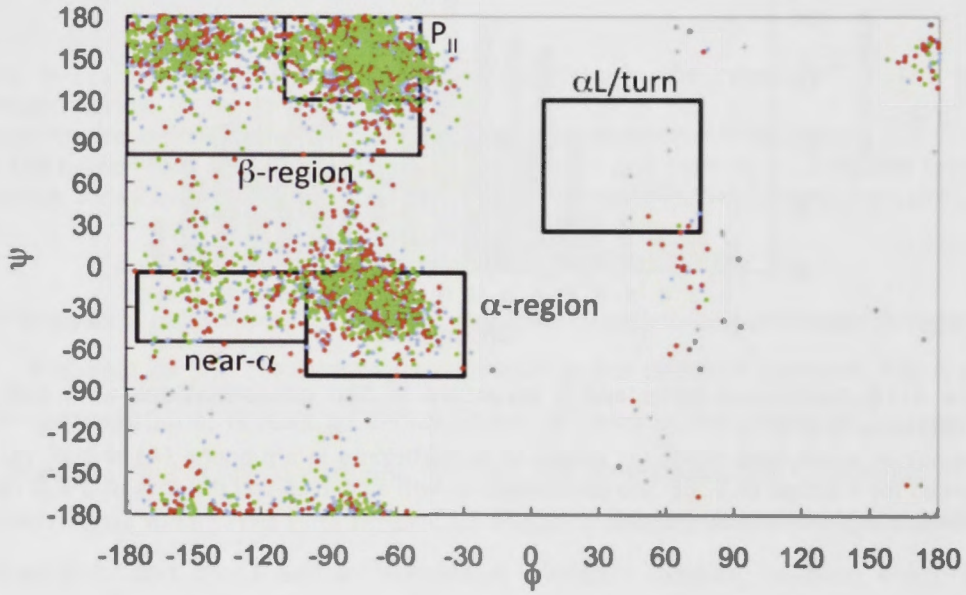


Figure 4-19 Protein backbone torsional angles and Ramachandran plot
(A) The protein backbone dihedral angles are illustrated

The combination of these 2 dihedral angles are diagnostic of secondary structure as discovered by G.N. Ramachadran in the 1960s. This analysis is still common today via the Ramachandran plot (Ramachandran and Sasisekharan, 1968). An example of a Ramachandran plot, prepared for Ala7 of Sho is shown in Figure 4-20



	β range		P_{II} range		α range		near-α range		αL range	
ϕ	-180	-50	-110	-50	-100	-30	-175	-100	5	75
ψ	80	180	120	180	-80	-5	-55	-5	25	120

Figure 4-20. Ramachandran plot for Ala7
Data was collected every 20 ps during the last 25 ns for each of the 3 starting models, Sho1 (blue), Sho2 (red), Sho3 (green).

A full rotation around these backbone bonds will be obstructed by the residue's sidechain, by steric interference of neighbouring residues and by electrostatic interactions such as the formation of a stable backbone hydrogen bond. Solvation effects will also play a part in determining favourable conformers. Stable secondary

structure will result in a peak of ϕ and ψ dihedral angles around optimal values whereas flexible structures will produce a more diverse distribution (Hovmoller et al., 2002).

The Ramachandran plot in Figure 4-20 depicts the area in the top left hand quadrant associated with extended β structure, referred to here as β-region. The left side of the β-region signifies a conformation in which the sidechain Cα-Cβ bond of a residue is almost perpendicular to the peptide bond, as found in the β-sheet. In this formation there is little opportunity for interaction between sequential sidechains or the sidechain and mainchain. The right side, the polyproline II (P_{II}) region is associated with left-handed spirals seen typically in the collagen helix. This conformation does not have stabilizing backbone hydrogen bonds and is most likely stabilized by interactions with solvent (Creamer and Campbell, 2002). Recent work has shown that the P_{II} region is associated with unfolded regions of proteins (Shi et al., 2002, Shi et al., 2006). As proline is unnecessary for PII conformation, it has been suggested that this misleading nomenclature be abandoned and that PII should stand for 'polypeptide-II" conformation (Hollingsworth et al., 2009).

Below is the region associated with α-helical structure referred to here as α-region. The adjoining near α-region is associated with turns. The upper RH quadrant is associated with the left-handed alpha helix and is referred to as the α_L-region. Although individual residues may occupy the α_L-region this pattern is not seen in repeating segments and accordingly this type of helix is not seen in protein structures (Hollingsworth et al., 2009). Rather, residues in this basin are likely to be found in γ-turns such as reverse turns seen in the β-hairpin structure.

The Ramachandran plot of a coil region is expected to have a scatter of $\phi\,\psi$ angles in the β–region, the α–region and to a lesser extent the α_L-region although the population in each of these regions will vary depending upon the residue type (Smith et al., 1996). For example, Ala and Ser are likely to have a higher propensity for $\phi\,\psi$ angles in the α-region while Val has branching at the Cβ position which restricts rotation and will have a high population of $\phi\,\psi$ angles in the β-region (Swindells et al., 1995). In addition, it is now clear that flanking residues, particularly those with a long sidechain will modulate the ϕ and ψ angles preferred by their neighbours even in a random coil (Griffiths-Jones et al., 1998).

Contour plots of the backbone dihedral angles for each of the residues of Sho and Fx were created by placing pairs of (ϕ ψ) angles into a 2 dimensional matrix consisting of 36 x 36 (10°) bins to determine the frequency that each region is populated by each residue (Figures 4-21 - 4-25). Data was collected every 20 ps during the last 25 ns and averaged over all 3 starting models. This population occupancy data is summarized in Table 4-2.

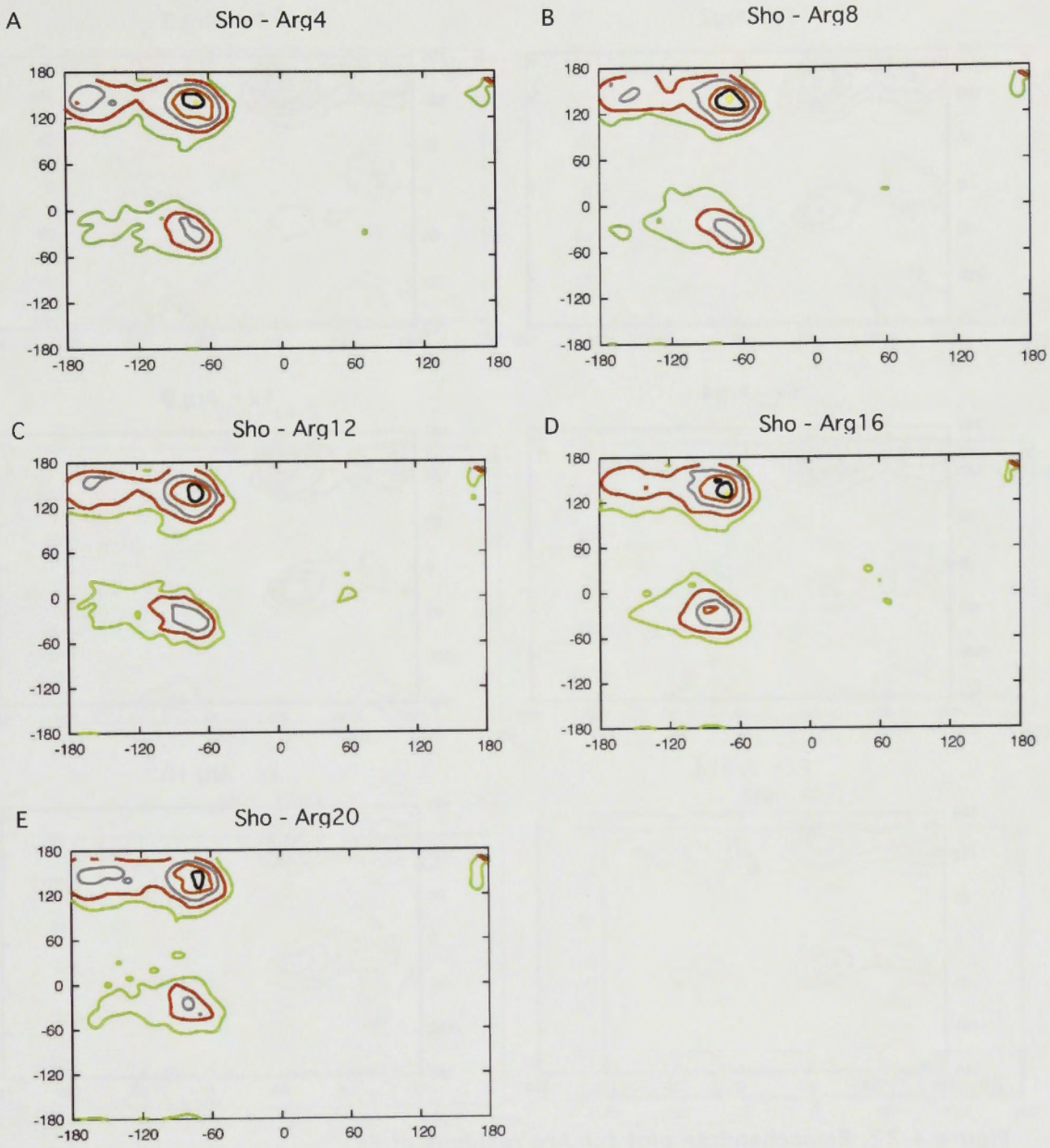


Figure 4-21. Ramachandran plot for Arg residues of Sho
Legend (total bin counts): ≥ 5 <15 (green), ≥ 15 <30 (red), ≥ 30 < 50 (grey), ≥ 50 < 70 (orange), ≥ 70 < 90 (black).

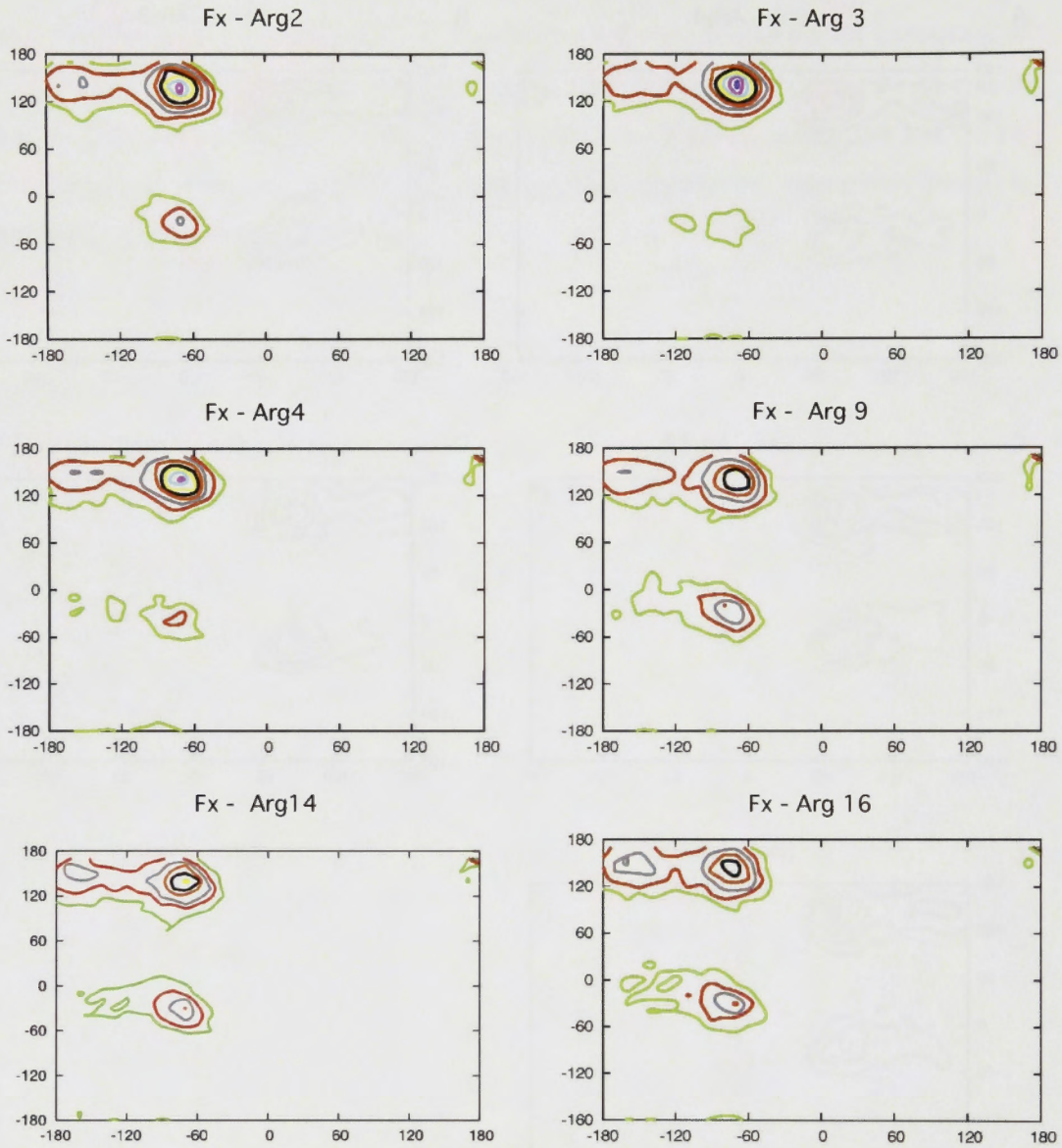


Figure 4-22. Ramachandran plot for Arg residues of Fx
Legend (total bin counts): ≥ 5 <15 (green), ≥ 15 <30 (red), ≥ 30 < 50 (grey), ≥ 50 < 70 (orange), ≥ 70 < 90 (black), ≥ 90 < 110 (yellow), ≥ 110 < 130 (blue), ≥ 130 < 150 (pink).

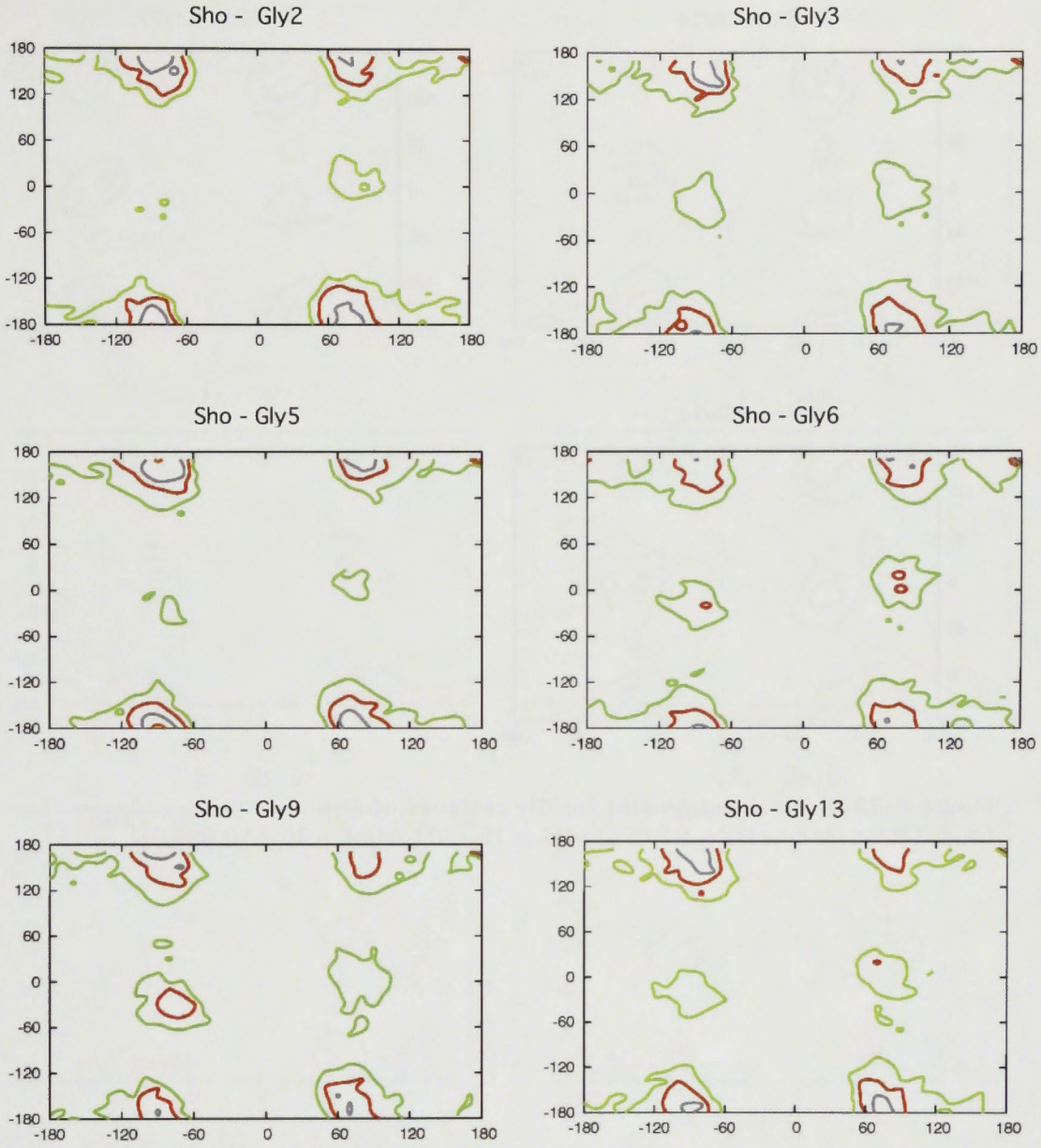


Figure 4-23 Continued on next page

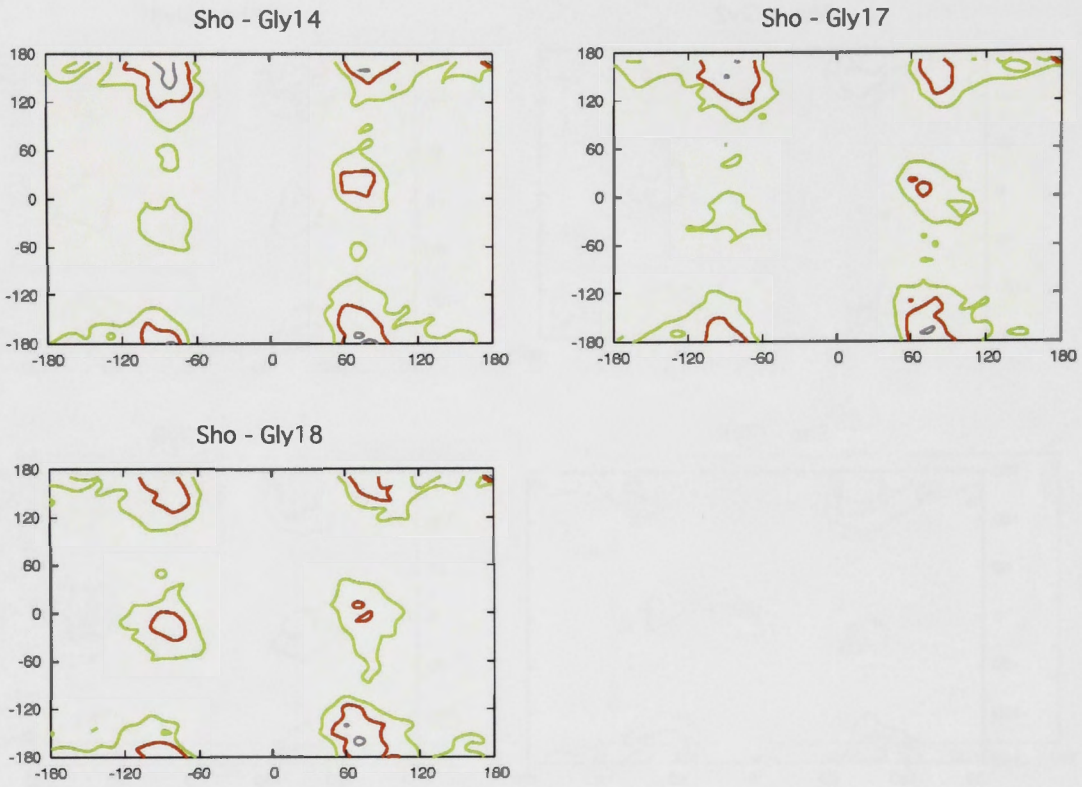


Figure 4-23. Ramachandran plot for Gly residues of Sho
Legend (total bin counts): ≥ 5 <15 (green), ≥ 15 <30 (red),), ≥ 30 < 50 (grey).

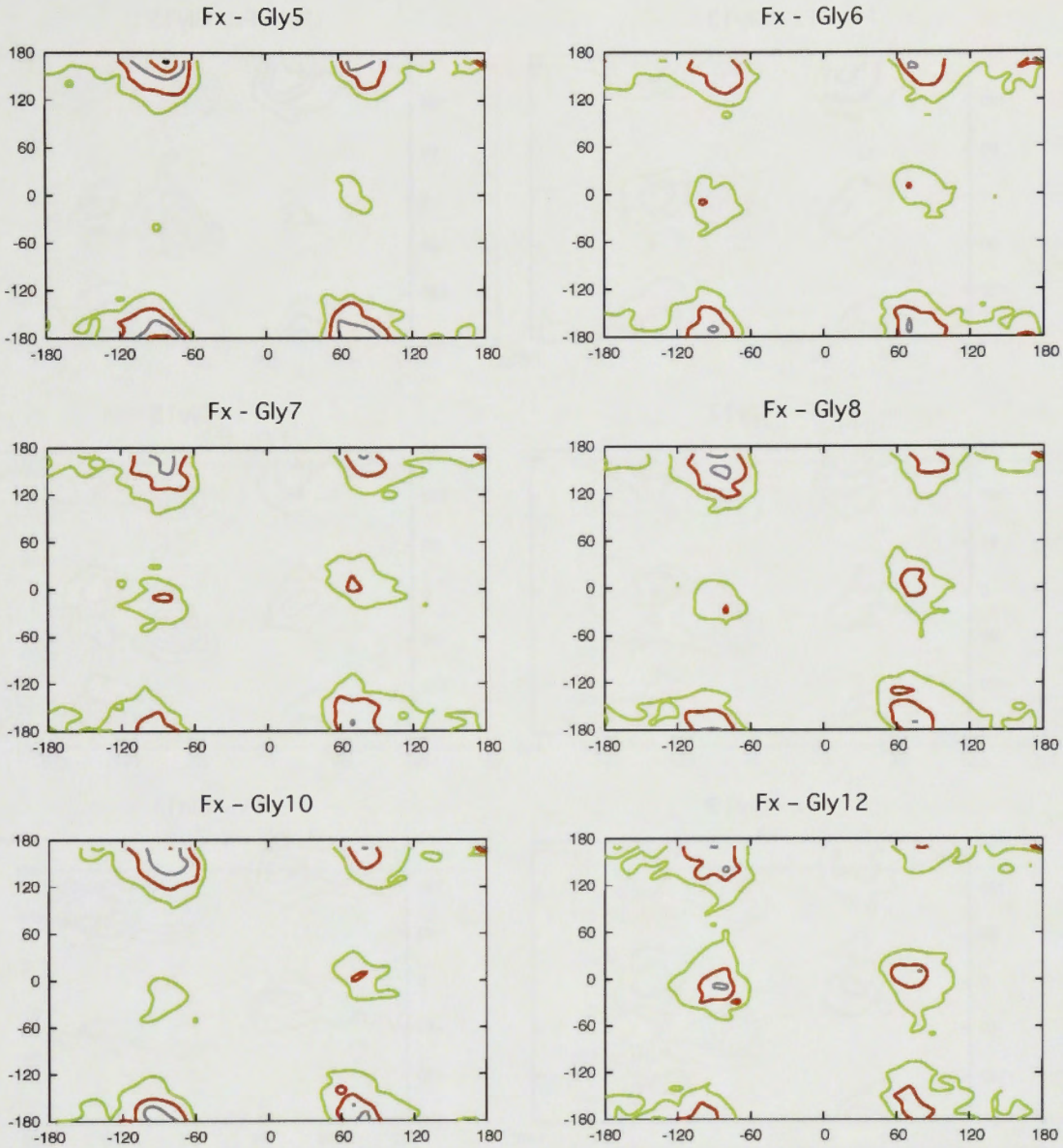


Figure 4-24 Continued on next page

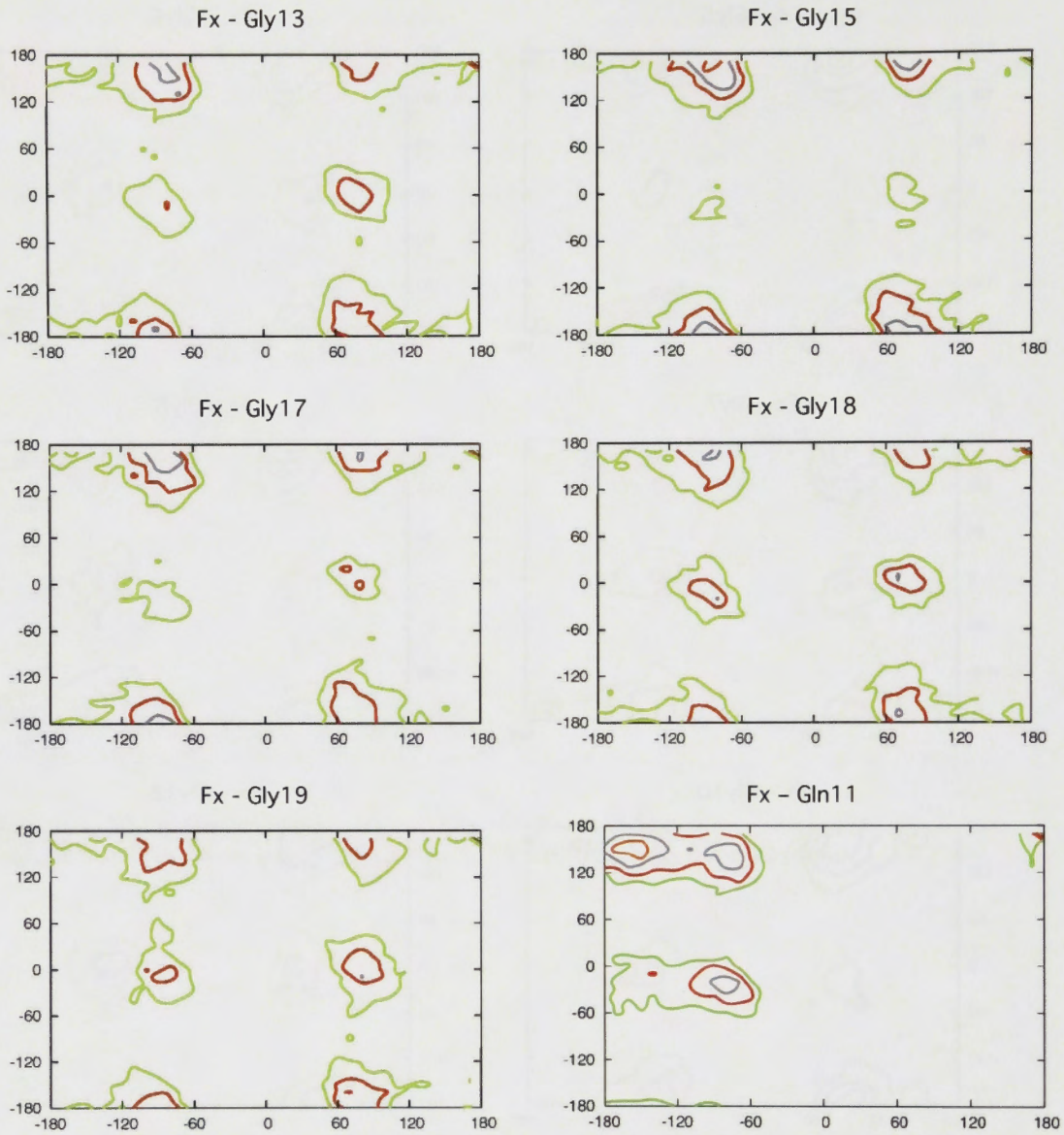


Figure 4-24. Ramachandran plot for Gly and Gln residues of Fx
Legend (total bin counts): ≥ 5 <15 (green), ≥ 15 <30 (red),), ≥ 30 < 50 (grey), ≥ 50 < 70 (orange).

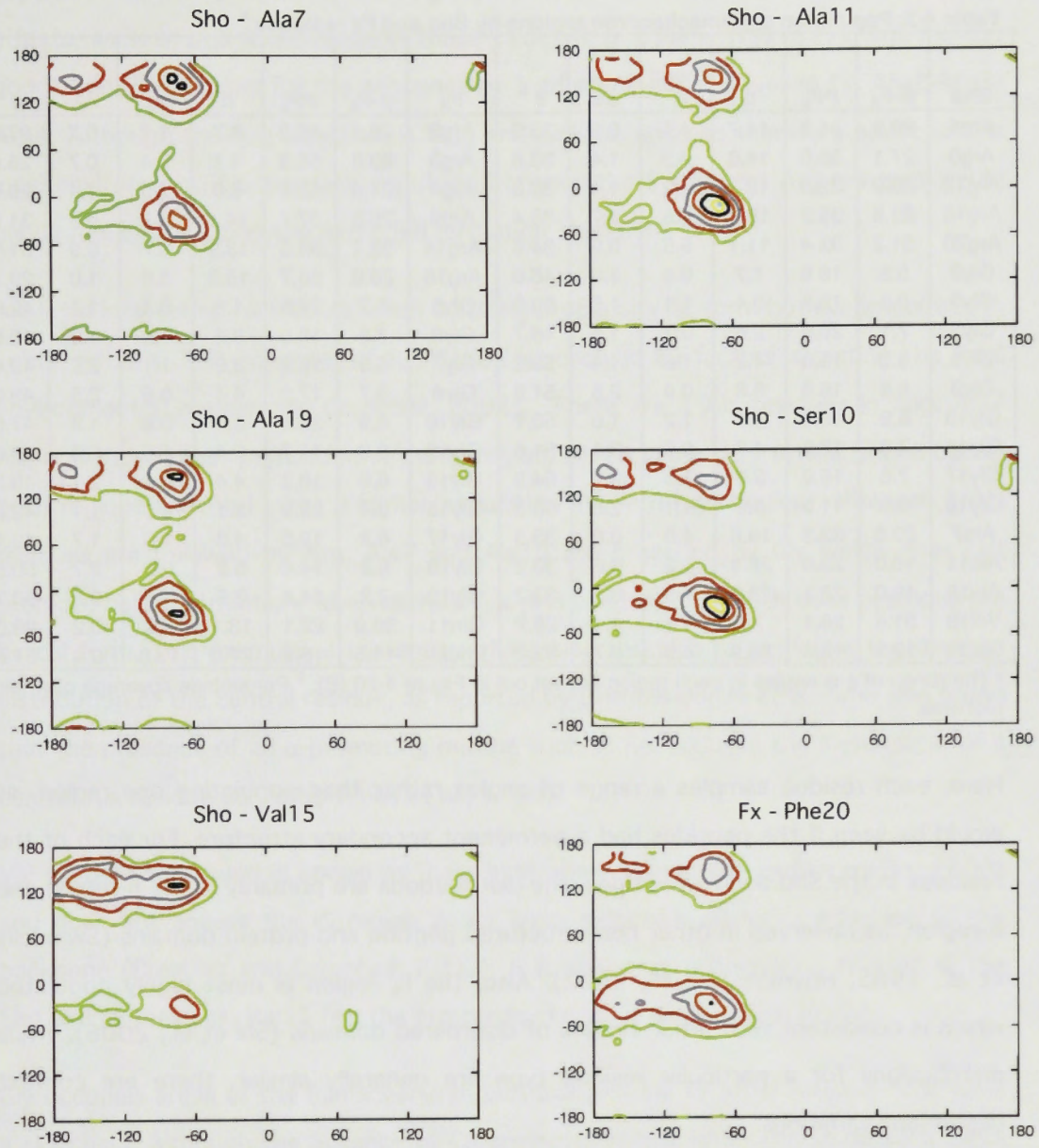


Figure 4-25. Ramachandran plot for non-Gly and non-Arg residues of Sho and Fx
Legend (total bin counts): ≥ 5 <15 (green), ≥ 15 <30 (red),), ≥ 30 < 50 (grey), ≥ 50 < 70 (orange), ≥ 70 < 90 (black), ≥ 90 < 110 (yellow), ≥ 110 < 130 (blue), ≥ 130 < 150 (pink).

Table 4-2 sets out the percentage occupancy of the Ramachandran regions by each residue and also gives the percentage coverage of the entire plot area.

Table 4-2. Population of Ramachandran regions by Sho and Fx residues [a]

Sho	β-P_{II}	PP_{II}	α	near α	αL	# [b]	Fx	β-P_{II}	PP_{II}	α	near α	αL	# [b]
Arg4	29.9	34.9	11.7	4.4	0.9	33.9	Arg2	26.1	48.2	9.7	1.7	0.7	27.8
Arg8	27.1	35.9	14.0	4.3	1.4	33.5	Arg3	23.8	56.8	4.1	2.4	0.7	25.6
Arg12	26.9	32.3	15.2	4.2	1.4	33.3	Arg4	24.4	52.1	6.0	2.5	1.1	29.6
Arg16	23.8	35.3	16.2	5.5	1.7	33.4	Arg9	26.5	37.7	14.4	3.7	0.8	31.9
Arg20	31.2	33.4	11.1	5.5	0.9	34.5	Arg14	32.1	33.3	13.9	3.7	0.9	31.0
Gly2	5.8	18.6	1.7	0.6	1.0	46.0	Arg16	28.0	36.7	15.0	3.8	1.0	29.7
Gly3	8.3	16.8	3.4	1.1	1.5	50.9	Gly5	6.7	22.6	1.9	0.4	1.1	42.0
Gly5	7.1	20.6	2.6	0.7	1.5	48.7	Gly6	7.5	13.0	3.2	0.8	1.2	48.5
Gly6	8.3	13.0	4.2	1.6	1.4	53.2	Gly7	6.9	16.9	3.9	1.1	2.2	49.8
Gly9	5.8	16.6	8.8	0.9	2.5	51.9	Gly8	6.7	17.2	4.4	0.6	2.5	49.0
Gly13	8.9	16.8	4.1	1.2	1.8	50.1	Gly10	6.9	23.5	3.6	0.8	1.8	47.5
Gly14	7.8	16.8	4.7	0.9	3.1	51.6	Gly12	8.2	14.7	5.4	1.8	3.4	50.0
Gly17	7.8	16.0	3.7	1.5	2.7	54.9	Gly13	6.6	18.2	4.4	0.8	2.7	48.2
Gly18	6.5	11.5	6.8	1.1	2.8	53.3	Gly15	6.4	22.9	2.1	0.5	1.1	43.8
Ala7	20.8	32.3	19.6	4.0	0.6	33.3	Gly17	6.2	19.5	4.3	1.0	1.7	49.1
Ala11	16.0	23.6	28.4	4.9	0.6	30.2	Gly18	6.2	14.8	5.2	0.9	2.7	50.5
Ala19	19.0	28.1	23.6	4.9	0.7	33.2	Gly19	7.2	14.4	3.8	1.3	2.2	50.7
Val15	51.4	26.1	7.1	4.4	1.1	29.2	Gln11	38.8	22.1	13.1	7.7	0.2	29.0
Ser10	19.3	15.8	28.9	8.6	0.7	33.2	Phe20	14.5	18.9	27.0	17.8	0.4	31.6

[a] The range of ϕ ψ angles in each region are set out in Figure 4-20 (B). [b] Percentage coverage of whole plot area.

Here, each residue samples a range of angles rather than populating one region, as would be seen if the peptides had a permanent secondary structure. For each of the residues in the Sho and Fx peptides, the distributions are primarily in the β-region and α-region, as observed in other non-structured peptide and protein domains (Swindells et al., 1995, Hovmoller et al., 2002). Also, the P_{II} region is most highly populated which is consistent with other studies of disordered domains (Shi et al., 2006). While distributions for a particular residue type are generally similar, there are context dependent variations.

The Arg residues of both Sho and Fx populate the extended β-region most strongly. On average, Arg populates the β-region 62% of the time in Sho and 71% of the time in Fx. This can be compared to a much lower occupation of the α-region; on average 14% in Sho and 11% in Fx. A larger contribution to this population comes from the P_{II} region (LHS). In the extended P_{II} conformation backbone amide and carbonyl groups are exposed and stabilized by interactions with the solvent. This conformation will be less likely to form where a bulky sidechain shields the backbone from solvent. Residues which lack a sidechain and those which have long flexible sidechains are therefore favoured in the P_{II} region (Creamer and Campbell, 2002). While Arg has a long sidechain, it is extremely flexible and does not block solvent molecules.

Population of the P_{II} region by the Arg residues is most sensitive to local environment effects, as seen in a wider range in these values (e.g. 33% - 57% in Fx). The highest population of P_{II} occurs for the sequence of 3 adjacent Arg residues in Fx, Arg2-Arg3-Arg4. Although the Arg sidechain is long and flexible, and can therefore adapt to the local environment, electrostatic repulsion between the 3 adjacent Arg residues would play a role in stabilizing the extended formation seen here.

Ala is an interesting residue as it has a strong propensity to form both α-helical and P_{II} conformations. This apparent contradiction is resolved by factors affecting the local environment (Creamer and Campbell, 2002). There are 3 Ala residues in Sho; Ala7 has a slightly higher population in the P_{II} region, Ala19 has a very similar population in both regions and Ala11 has a higher population in the α-helix region. While all 3 Ala residues are followed by Arg, Ala7 and Ala19 are preceded by Gly which does not promote helical formation, whereas Ala11 is preceded by Ser which does promote the helical form. This exemplifies the ability of flanking residues to modulate the dihedral distribution of the central residue, as reported by Griffiths-Jones et al., who also found that the presence of an α-promoting residue such as Ser reduces the β-character of a central Ala residue (Griffiths-Jones et al., 1998).

Val on the other hand is known as 'helix indifferent' (Mitsutake and Okamoto, 2000) and does not access the P_{II} region as its bulky sidechain occludes solvation of the backbone (Creamer and Campbell, 2002). It is therefore unsurprising that of all the Sho and Fx residues, Val15 has the strongest cluster in the β-strand region.

Gly occupies areas of the Ramachandran plot inaccessible to other residues that have a sidechain. Although the absence of Cβ promotes mirror symmetry in dihedral angle distribution around both ϕ =0 and ψ=0 axes, the plots presented show some deviation in symmetry as a result of local effects. On average Gly occupies the β-region (mainly the P_{II} area) 27% of the time in Sho and 25% of the time in Fx. This is compared to 5% and 4% for the α-region in the 2 peptides. Gly occupies the αL region associated with turns more frequently than the other residues. Gly12 in the middle of the Fx peptide occupies this region most highly (3.4%) and is likely involved in turns in the middle of the peptide which give rise to a transient hairpin structure. Gly residues with the highest population in the β-region are those preceded by an Arg residue (RG). In Fx the β-region is most populated by Gly when surrounded by Arg (RGR): Arg precedes Gly (RG) > Gly precedes Arg (GR) > Gly between Gly (GGG). The backbone flexibility of Gly has consequences for the broader peptide landscape. As Gly rotates

through its normal range of dihedral angles it changes the environment of neighbouring residues and facilitates the reorientation of the adjacent Arg residues in space.

The breadth of the distribution of $\phi\psi$ angles, or in other words the coverage of the Ramachandran plot area (occupied bins vs unoccupied bins in the contour plots), reveals the flexibility of the residues. The results presented in Table 4-2 show similar coverage for the Arg residues of Sho (33.3-34.5%) while there is slightly more variation in the Arg residues in Fx (25.6%-31.9%) consistent with a higher presence of extended structure. As expected, the most flexible residue in both peptides is Gly which has a coverage of 46-54.9% in Sho and 42-50.7% in Fx. The least flexible residues in Sho and Fx are Val (29.2%) and Gln (29%), respectively. These results indicate that the Arg and Gly residues of Sho are a little more flexible than their counterparts in Fx.

Overall, this analysis shows that each residue in both peptides most frequently accesses dihedral angles associated with the β-region, with a maximum cluster in the P_{II} region. This is most pronounced for the Arg residues where the β-region is populated between 59 and 80% of the time. The propensity for extended conformation is greater in these RGG box peptides than found in other studies of unstructured domains (Griffiths-Jones et al., 1998) and when compared to simulations of GGXGG peptides (Beck et al., 2008). The α-region is populated most frequently by Ala and Ser in Sho and to similar extent by Phe in Fx, with populations similar to those reported by Beck et al., for the pentapeptide simulations. However, for all the other residues the α-region is populated less frequently in the RGG box peptides. This is also true of the near α-region and the αL-region. The near α-region is slightly more populated in Sho compared to Fx consistent with higher frequency of turns in Sho. The relative populations of the β-region and α-region are set out in Table 4-3 and compared to results published for a survey of unstructured domains in 512 proteins.

Table 4-3. Occupation of β-region as a percentage of β-region and α-region combined [a, b]

	Arg	Gly	Ala	Val	Ser	Gln	Phe
Sho	82	84	66	92	55		
Fx	87	87				82	55
G-J [c]	63	68	63	79	55	63	78

[a] Percentage calculated as (β-region/(β-region + α-region)) * 100). [b] Average over all residues of the same type. [c] Results presented by Griffiths-Jones et al. (1998) based on unstructured regions in 512 proteins from the PDB data base.

Taken together, this analysis suggests that the high number of Arg residues in the RGG box encourages these peptides to adopt extended conformations. The significant population of the P_{II} region seen here provides further evidence for the proposition that unfolded protein domains are likely to assume this conformation (Creamer and Campbell, 2002, Shi et al., 2002, Shi et al., 2006). As the P_{II} form is not stabilized by backbone hydrogen bonds, the carbonyl and amide groups of the mainchain are exposed to solvent and may interact with other binding partners. This conformation is therefore poised for interaction, and has been found in a variety of peptide/receptor complexes (Siligardi and Drake, 1995). P_{II} formation has been found in the N-terminus of PrP (Smith et al., 1997, Gill et al., 2000) and in other disordered protein domains which aggregate to form fibrils, such as a-synuclein and tau (Syme et al., 2002) and AB peptides (Danielsson et al., 2005). It has been suggested that there is a low energy barrier for conversion between P_{II} conformations and β-sheet structure, which may facilitate amyloid formation (Danielsson et al., 2005). However, this may also be important in RNA binding, as the β formation has been recognized as a particularly advantageous RNA recognition surface structure (Draper, 1995, Rhie et al., 2003, Maris et al., 2005). As the P_{II} conformation is stabilized by interactions with water molecules, it is less likely to be observed in implicit solvation simulations. This may provide another reason for the tendency towards more compact conformations seen in the implicit solvation simulations.

In addition to the rotations around the peptide backbone there will be rotation around bonds within the sidechains of the residues (χ dihedral angles). The χ angles can form 3 low energy staggered conformations: *gauche* (-) in which the sidechain points towards the N-terminus; *gauche* (+) where the sidechain points towards the peptide backbone and *trans* where the sidechain points towards the C-terminus, as illustrated in Figure 4-26.

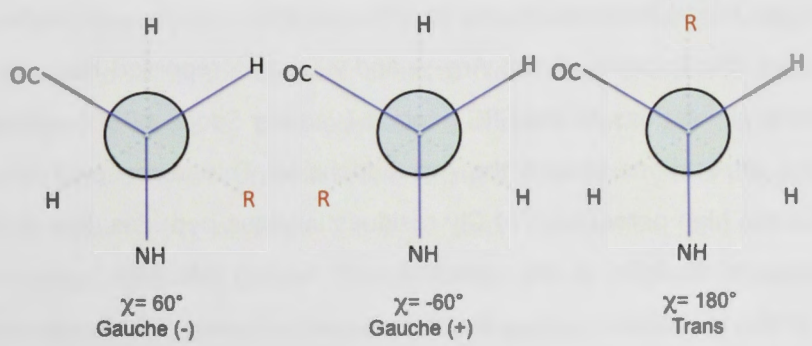


Figure 4-26. χ dihedral angles, following IUPAC-IUB Commission nomenclature

Arg with a long sidechain, has 4 χ dihedral angles, χ_1, χ_2, χ_3 and χ_4. Distributions of the 4 χ angles of each of the Arg residues in Fx and Sho are presented in Figures 4-27 and 4-28.

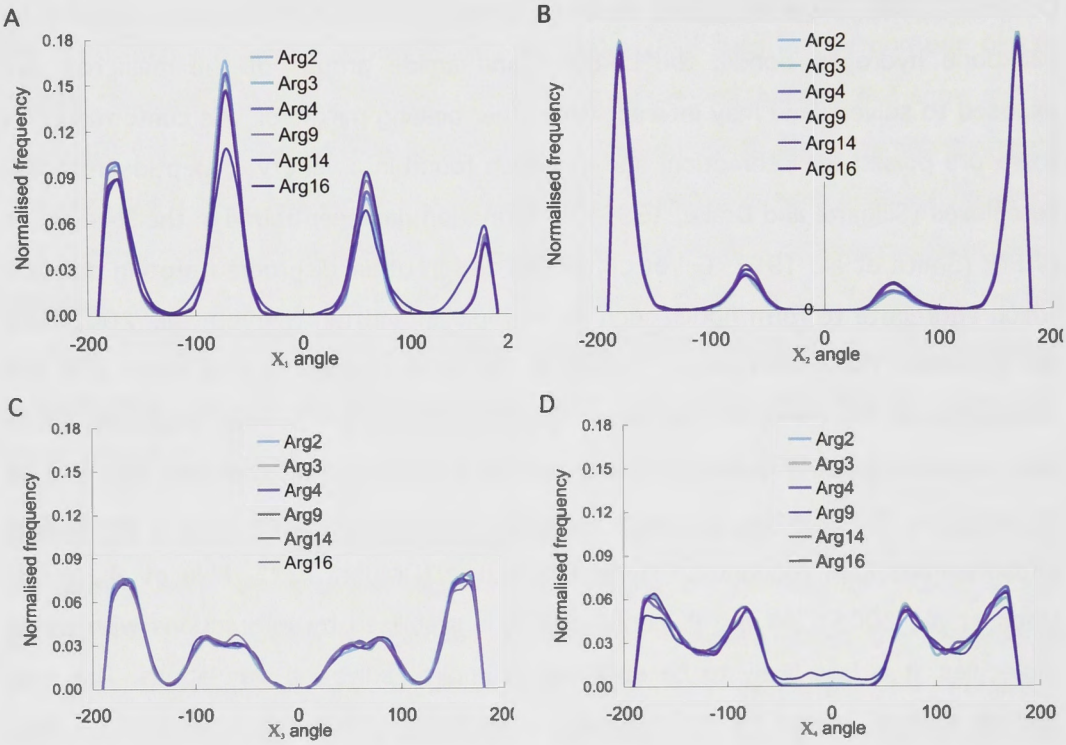


Figure 4-27. χ angle distributions of Arg residues in Fx
Frequency distributions of χ_1(:n@N :n@CA :n@CB n@CG) χ_2(:n@CA :n@CB n@CG :n@CD) χ_3(:n@CB n@CG :n@CD :n@CZ) and χ_4(n@CG :n@CD :n@CZ :n@NH1) for each of the Arg residues in Fx.

From these it can be seen that all of the Arg residues experience a similar rotation around each of the χ angles although most variation is seen in χ_1 where there is likely to be some influence on flexibility from the sidechains of neighbouring residues. The χ_1 distributions prefer the *gauche* (+) and *trans* conformations, χ_2 prefers the *trans* orientation and broad distributions are seen for rotations at the end of the sidechain, χ_3 and χ_4. The distributions of the Arg χ_1 and χ_2 angles reported here are similar to those found in simulations of the GRG peptide (van der Spoel, 1998) indicating a lack of restriction in rotations around the Arg sidechains. This absence of obstruction is likely due to the high percentage of Gly residues in these peptides. It is interesting to see the rotational flexibility at the extremities of the Arg sidechain (χ_4) which signifies the ability of the guanidinium group to orient towards a potential binding partner.

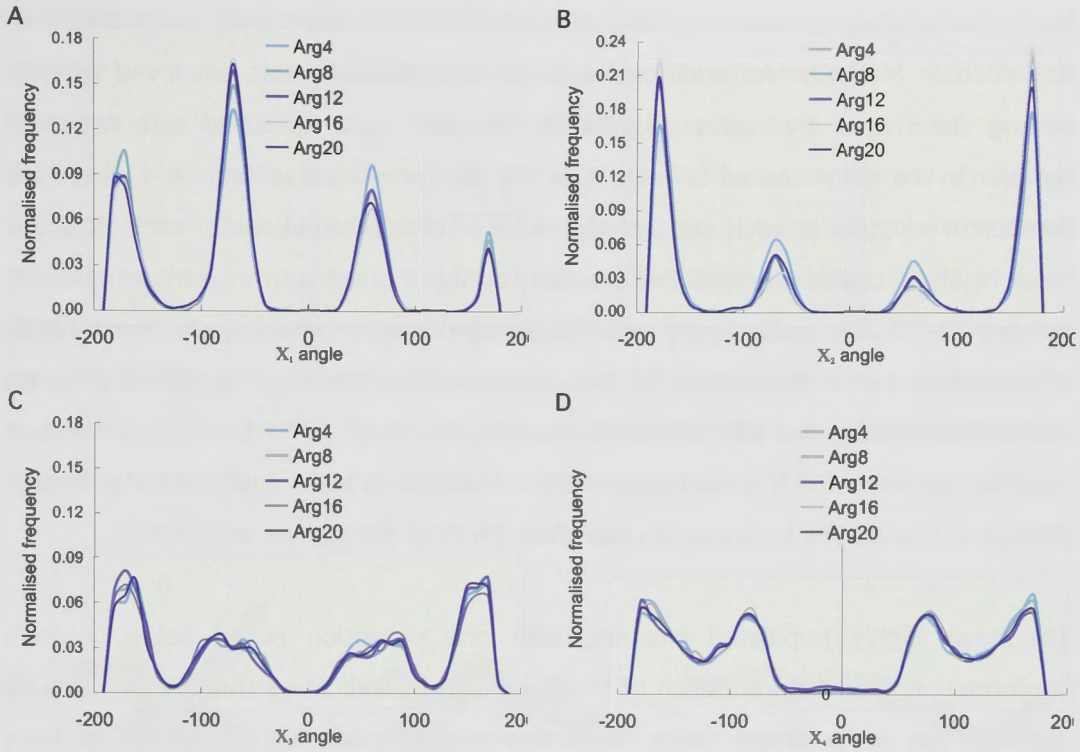


Figure 4-28. χ angle distributions of Arg residues in Sho
Frequency distributions of χ_1(:n@N :n@CA :n@CB n@CG) χ_2(::n@CA :n@CB n@CG :n@CD) χ_3(:n@CB n@CG :n@CD :n@CZ) and χ_4(n@CG :n@CD :n@CZ :n@NH1) for each of the Arg residues in Sho.

The preceding dihedral angle analysis illustrates the broad similarity between Sho and Fx in backbone flexibility and as well as the dihedral angles accessed by the Arg sidechains. However, the residues of Fx are a little more likely to occupy extended conformations, (as indicated by an increase in β-region density), consistent with the higher mean radius of gyration compared with Sho.

4.3.6 Cluster analysis

The lack of permanent structure shown by DSSP and dihedral angle analyses is likely to result in a very diverse conformational ensemble. Cluster analysis is a useful tool to gauge the degree of similarity or dissimilarity in a conformational ensemble, as well as identifying structures that provide some insight into the general conformational features that occur most often in the ensemble. The number and size of clusters reflects the diversity of the entire ensemble with a very diverse ensemble having a high number of clusters containing relatively few members.

Here, cluster analysis was performed using the MMTSB toolset (Feig et al., 2004) as described in Methods. An additional level of discrimination was employed by first sorting the frames by radius of gyration. Six sets were produced with radius of gyration in the ranges noted in Table 4-4. The data presented in Table 4-4 shows the number of clusters in each set and the number of conformations in each of the 5 most highly occupied clusters. For example, in Sho the set with a radius of gyration between 9-11 Å is most highly populated (4341 conformations), the most highly populated cluster in this set is t.92 (cluster names generated by the MMTSB program are retained) which has 230 elements, equal to 5.3 % of all frames with a radius of gyration between 9-11 Å, but only 1.5% of all Sho frames analyzed. The largest clusters in Sho and Fx both contain less than 2% of all the frames analyzed.

The most highly populated clusters, with one exception noted below, contain conformations derived from each of the 3 simulations indicating that the simulations explored the same broad range of conformational space although starting from different conformations. The exception is the Fx cluster t.3 derived from the set of conformations with a radius of gyration in the range 5-7 Å. This set contains fewer conformations than the other sets (225 conformations) and the most highly populated cluster, t.3 contained only 40 conformations, of these, most were derived from Fx1 and none were derived from Fx3.

Table 4-4. Cluster analysis of Sho and Fx in subsets based on radius of gyration

			Sho			
R_G (Å)	**5-7**	**7-9**	**9-11**	**11-13**	**13-15**	**15-18**
Conformations [a]	652	3905	4341	3567	2001	532
Clusters [b]	57	306	338	215	127	26
Top 5 clusters (conformations) [c]	t.15 (56)	t.234(156)	t.92 (230)	t.215 (159)	t.4 (271)	t.23 (103)
	t.16 (35)	t.220 (98)	t.13 (201)	t.205 (141)	t.87 (99)	t.1 (70)
	t.26 (34)	t.9 (83)	t.255(147)	t.208 (104)	t.32 (93)	t.24 (67)
	t.11 (34)	t.1 (80)	t.257 (90)	t.206 (96)	t.83 (74)	t.26 (44)
	t.45 (31)	t.49 (78)	t.269 (82)	t.158 (96)	t.27 (71)	t.13 (39)
			Fx			
R_G (Å)	**5-7**	**7-9**	**9-11**	**11-13**	**13-15**	**15-18**
Conformations	225	2035	3048	4387	3250	1338
Clusters	40	178	255	252	155	41
Top 5 clusters (conformations)	t.3 (40)	t.27 (114)	t.51 (120)	t.88 (180)	t.1 (275)	t.7 (289)
	t.8 (35)	t.35 (98)	t.12 (119)	t.114 (173)	t.20 (212)	t.1 (213)
	t.2 (15)	t.70 (54)	t.204 (105)	t.9 (124)	t.9 (191)	t.28 (91)
	t.37 (10)	t.68 (54)	t.67 (95)	t.83 (122)	t.49 (136)	t.3 (76)
	t.27 (8)	t.12 (47)	t.5 (95)	t.118 (103)	t.52 (88)	t.19 (62)

[a] Number of conformations analyzed within a given radius of gyration. [b] Number of clusters generated. [c] The five most highly occupied clusters with the number of conformations in each.

The diversity of this conformational ensemble is demonstrated by the large number of clusters produced and the relatively small number of conformations in each cluster. When considering the 5 most populated clusters in a set there is generally no one cluster that is significantly larger than the next 4. These data confirm that there is more diversity (more clusters with fewer elements) in the lower ranges of the radius of gyration than in the sets that contain the most extended conformations. In other words there are more ways in which collapsed conformations can be formed and less ways in which the extended conformations can be formed, as noted earlier in the discussion of Figure 4-13. Centroids, derived from 3 of the highly occupied clusters spanning the range of radius of gyration, are shown in Figure 4-29.

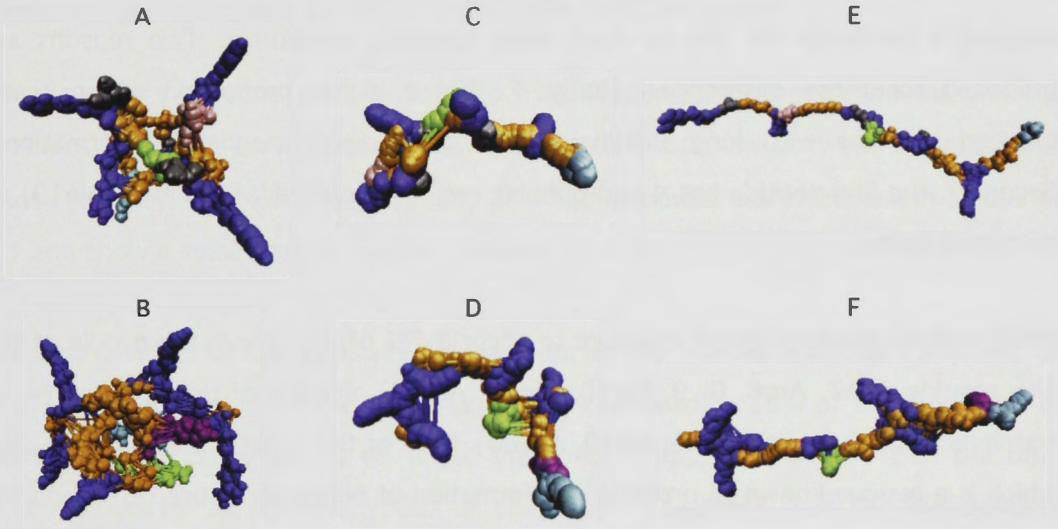


Figure 4-29. Centroids from 3 clusters of the Sho and Fx simulations
(A) Sho centroid 5-7_t.15 (56 members), (B) Fx centroid 5-7_t.8(35 members), (C) Sho centroid 9-11_t.92 (230 members), (D) Fx centroid 9-11_t.12 (119 members), (E) Sho centroid 15-18_t.23 (103 members), (F) Fx centroid 15-18_t.7 (289 members). Arg (dark blue), Gly (orange), Gln/Ser (green), Phe (purple) Lys (light blue), Ala (grey), Val (pink). In C and E the Sho peptide is oriented from residue 21-1 whereas the Fx peptide in D and F is oriented from residue 1-21.

The centroids depicted in Figure 4-29 illustrate the range of structures sampled and also the similarity in conformations explored by Sho and Fx. Given the potential for variation, the similarity in the centroids produced from the highly occupied clusters of Sho and Fx spanning the range of radius of gyration, is striking.

In summary, the cluster analysis illustrates:

- the high degree of variation in these peptides
- that the 3 trajectories for each peptide sample similar regions of conformational space

- centroids show that the Arg residues tend to extend out perpendicular to the mainchain, as seen in β-conformations and are well placed to interact with a binding partner, either alone or in concert.

4.3.7 Differences between Sho and Fx - transient helical structure and hydrophobic contacts

So far, the analysis techniques employed – distance distributions, hydrogen bond identification, DSSP, Ramachandran plots and cluster analysis – all demonstrate a strong similarity between the Sho and Fx peptides. Nonetheless, subtle differences do exist. DSSP analysis shows a slightly higher propensity for helical structure and turns in the middle of Sho, and the radius of gyration and end-to-end distance distributions suggest a tendency for Sho to form more compact structures. Two reasons are proposed for these differences. Firstly, Fx has a higher proportion of positively charged residues with long sidechains that encourage extended conformations. Secondly, the Sho peptide has 4 hydrophobic residues (Ala7, Ala11, Val15, Ala19) as discussed below.

DSSP analysis predicts helical structure to occur 3-7% of the time in the middle of the Sho peptide (Ala7, Arg8, Gly9, Ser10, Ala11, Arg12) and 3% of the time at the C-terminus of the peptide (Gly18, Ala19, Arg20). Both of these regions incorporate Ala, which is a residue known to promote the formation of helical structure (Chakrabartty et al., 1991, Rohl et al., 1999). The most prevalent helical form is the 3_{10} structure which occurs when the amide of a residue forms a hydrogen bond with the carbonyl of an amino acid 3 residues earlier (i + 3) as depicted in Figure 4-30 (A). By comparison α-helical structure (i + 4) is seen much less frequently as shown in Figure 4-30 (B). When either of these patterns is repeated at least 3 times the helical turn is formed. Generally 3_{10} structure is seen much less frequently in proteins than the α-helix and often comprises short repeats e.g. 1-2 turns found at the ends of longer stable α-helices (Hollingsworth et al., 2009).

Ala is very effective at promoting helical structure, as its short hydrophobic sidechain does not pay an entropy penalty as a result of helical formation. By contrast, the Gly backbone is more flexible and favours the random coil. As polar residues and Gly destabilize helical structure (Chakrabartty et al., 1991) the infrequent occurrence of helical structure seen here is likely due to the strong presence of Gly and Arg in the peptide.

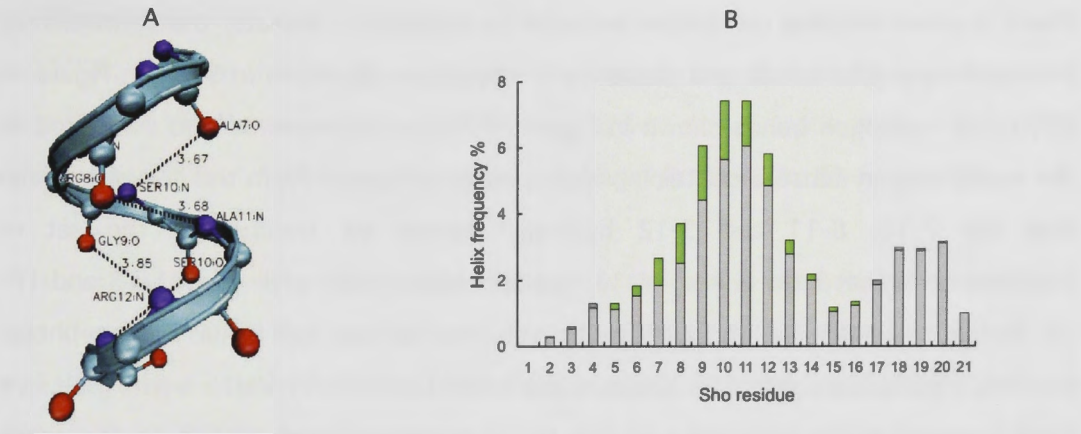


Figure 4-30. Helical structure in the Sho peptide
(A) Sho with a 3_{10}-helix depicting the hydrogen bonds with (i +3 spacing) Ala7-Ser10, Arg8-Ala11, Gly9-Arg12, Ser10-Gly13. (B) Frequency at which each residue participates in 3_{10} helix (grey) and alpha helix (olive green) in the Sho hybrid-REMD simulations.

The possibility of contacts between the hydrophobic sidechains of the Ala and Val residues has been investigated by calculating the distances between Cβ atoms of these residues. Frequency distributions of these distances are presented in Figure 4-31 and show a small peak at contact distance (3-6 Å).

Detailed analysis of the hydrophobic contacts in Sho reveals that the Cβ atoms of Ala7-Ala11 are within contact distance (3-6 Å) on average 16% of the time, Ala11-Val15, approximately 8% of the time, Val15-Ala19 approximately 11% of the time and Ala7-Ala19 approximately 5% of the time. It appears that the hydrophobic contacts are uncoupled, as pair-wise contact formation occurs no more than 0.75% of the time. All 3 contacts occur at the simultaneously only 0.07% of the time.

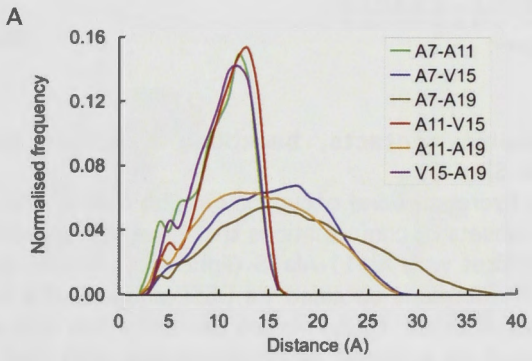


Figure 4-31. Frequency distributions of distance between hydrophobic residues in Sho, measured from Cβ atoms
Ala7-Ala11 (green), Ala7-Val15 (red), Ala7-Ala19 (orange), Ala11-Val15 (blue), Ala11-Ala19 (purple), Val15-Ala19 (brown).

There appears to be a correlation between hydrophobic contacts, the formation of particular hydrogen bonds and elements of secondary structure in Sho. In Figure 4-32(A) the hydrogen bonds shown in Figure 4-11 are re-presented and compared to the occupancy in subsets containing hydrophobic contacts. From this it can be seen that the 7-10, 8-11 and 9-12 hydrogen bonds are enriched in the set of conformations that have a Ala7-Ala11 hydrophobic contact while the 11-13 and 17-19 hydrogen bonds are enriched in the set that have a Val15-Ala19 hydrophobic contact. Hydrophobic contacts between Ala7-Ala11 and Ala11-Val15 both result in a small increase in the occupancy of the 10-12 hydrogen bond relative to the entire ensemble.

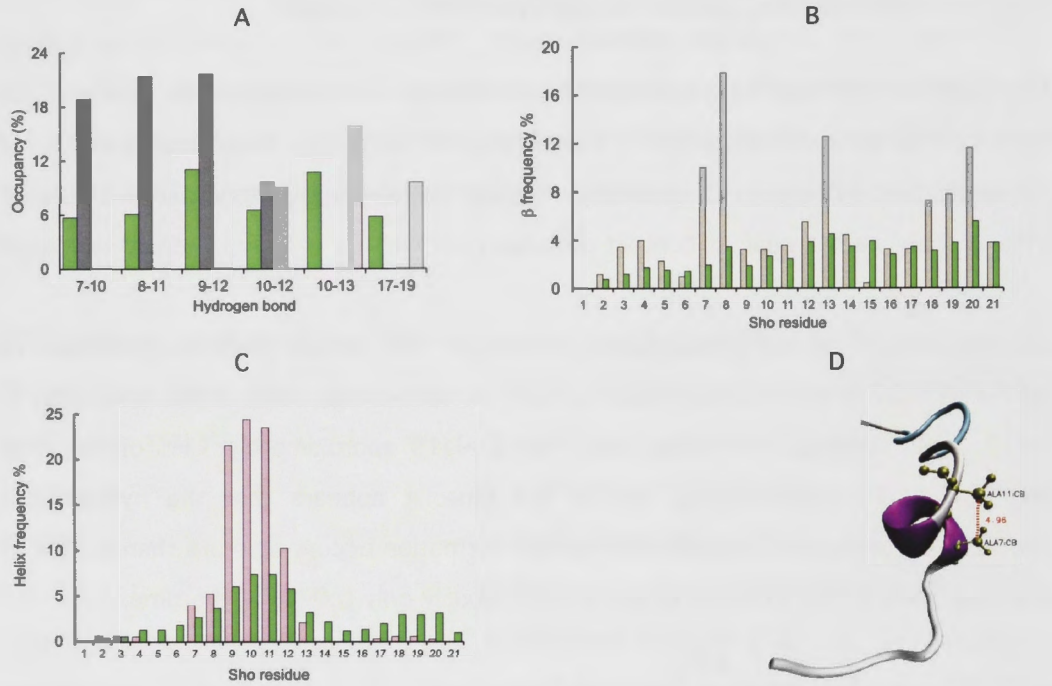


Figure 4-32. Hydrophobic contacts, backbone hydrogen bond occupancy and secondary structure in Sho
(A) Comparing backbone hydrogen bond occupancy in Sho over entire hybrid-REMD simulation (green) to occupancy in subsets of conformations that have hydrophobic contacts. Subset with Ala7-Ala11 (dark grey) subset with Ala11-Ala15 (light grey) subset with Val15-Ala19 (pink). (B) β-bridges frequency distributions obtained via DSSP analysis of a subset of conformations with Ala7-Ala19 in contact distance (grey) versus the entire Sho data set (green). (C) Results of DSSP analysis performed on a subset of conformations with Ala7 and Ala11 in contact distance showing the frequency of helical structure in this subset (pink) compared to the entire Sho data set (green). (D) One conformation which has Ala7-Ala11 contact and elements of helical structure.

DSSP analysis reveals that the subset with the Ala7-Ala19 contact is enriched for β-bridges which bring the ends of the peptide together (Figure 4-32(B)) and

conformations with the Ala7-Ala11 contact are enriched for helical structure (Figure 4-32(C)). The distribution of end-to-end distance in the Ala7-Ala19 subset reveals that 15% of the frames in this subset have an end-to-end distance in the 3-6 Å range which is approximately 3 times the proportion found in the entire Sho ensemble.

This analysis indicates a correlation between the formation of hydrophobic contacts and the creation of hydrogen bonds seen most frequently in the Sho simulations, including those responsible for the helical structure. DSSP analysis indicates that these subsets have an exaggerated presentation of those features which distinguish Sho from Fx, that is, helical structure in the middle of the peptide and β-bridges that bring the ends of the peptide into close proximity. Both of which affect global parameters such as the radius of gyration and end-to-end distance distributions.

Before leaving this topic it is interesting to note that Ser10 and Ala11, the residues most likely to be involved in a helical turn, are part of a putative common protein kinase C (PKC) phosphorylation site (SAR). In Chapter 2 attention was drawn to this site and it was suggested that Ser10 may be phosphorylated in a biological system with functional implications. In this context it is interesting to see how the helical form causes the sidechains of participating residues to be on the outside of the structure and in this case makes the hydroxyl group of Ser easily accessible for interaction, including phosphorylation. This is demonstrated by Figure 4-33 where the helical structure is shown to provide an environment clear of steric interference from neighbouring residues while exposing the hydroxyl group for potential interaction.

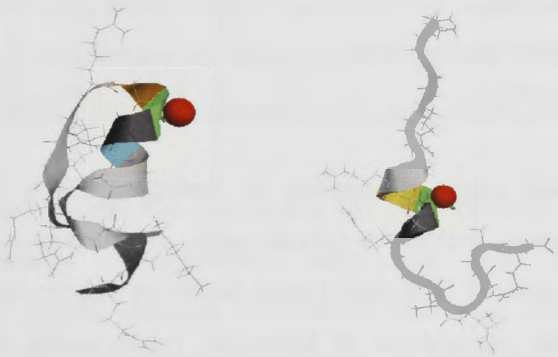


Figure 4-33. Sho conformation with helical structure involving Ser10
Cartoon representation of a conformation of Sho with Ser10 (green) and hydroxyl oxygen of Ser (red spacefill representation) in conformation with 3_{10}-helix (LHS) and conformation with elements of alpha-helix (RHS).

The fact that helical structure occurs infrequently raises the interesting possibility that this metastable region might facilitate certain types of interaction and could be advantageous in short-lived reversible processes.

4.3.8 Summary of comparisons of Sho and Fx and the nature of RGG box peptides

Analysis of the Sho and Fx peptides reveals that both peptides are disordered, with no permanent hydrogen bonds or structural elements that occur a significant amount of the time. These peptides are flexible and can explore a range of rotations around the backbone bonds rather than being stabilized in a particular conformation. Comparison of Sho and Fx reveals a similar pattern of backbone dihedral angles around the Gly and Arg residues and in the distribution of χ angles explored by the Arg sidechains. Generally, the Arg and Gly residues have a similar degree of flexibility in Sho and Fx, as demonstrated by their coverage of the Ramachandran plot areas. Global parameters and dihedral distributions, presented in this chapter for the Sho and Fx peptides are listed in Table 4-5 for ease of reference.

Table 4-5. Summary of Sho and Fx parameters (means) reported in this chapter

	Sho (SD)	**Fx** (SD)
Radius of gyration (Å)	10.5 (2.1)	11.8 (2.4)
End-to-end Distance (Å)	25.1 (9.9)	30.3 (11.1)
Arg β–population (%)	62.1 (2.8)	71 (7.1)
Arg α–population (%)	13.6 (2.2)	10.5 (4.7)
Gly β-population (%)	23.7 (2.8)	24.8 (3.5)
Gly α-population (%)	4.5 (2.2)	3.8 (1.1)
Hydrogen bonds > 5% occurrence	9 (2.6)	4 (2.6)
Max helical structure	Ser10 7.4%	Gly12 2.6%

The 2 peptides have subtle differences in that Sho has 4 hydrophobic residues, including 3 Ala residues, that make contact some of the time and are associated with the formation of transient hydrogen bonds between nearby residues. This results in Sho having a higher percentage of hydrogen bonds being formed > 5% of the simulation time. These hydrogen bonds include those that promote the formation of helical structure in residues 7-12 and those that bring the ends of the peptide together. Another point of difference is that Fx has an additional Arg residue and includes a stretch of 3 consecutive Arg residues. Electrostatic repulsion between these 3 Arg sidechains tends to stabilize extended structure. These factors appear to

explain the differences in distributions of radius of gyration and end-to-end distances seen in the 2 peptides.

Overall these differences are quite small and seem unlikely to have a significant effect on the binding potential of the peptides. The centroids generated from cluster analysis demonstrate that the Arg residues can assume a variety of orientations in both peptides, which may facilitate binding while the broad distributions of the Arg χ_3 and χ_4 angles indicate the ability of the terminal guanidinium groups to search for an optimum binding orientation.

4.4 CONCLUSIONS

The first conclusion that can be drawn from these investigations is that implicit solvation REMD and hybrid-REMD simulations produce different global characteristics as backbone hydrogen bonds are favoured in calculations performed using generalized Born implicit solvation. The hybrid-REMD simulations produce a broad distribution of the radius of gyration at a physiologically relevant temperature whereas the implicit solvation simulations find more collapsed structures, although permanent secondary structure is not indicated in either case. Another contributing factor may be that extended P_{II} structure seen to predominate in the hybrid-REMD simulations will be under-represented in simulations lacking explicit water molecules.

The hybrid-REMD simulations demonstrate that both Sho and Fx are disordered, with no particular 3D structure emerging for either peptide. All the residues access a range of backbone dihedral angles most pronounced in the case of the abundant Gly residue.

The Ramachandran plots reveal the propensity for both peptides to form extended conformations particularly conformations associated with the P_{II} region of conformational space. Similar findings have been made for other flexible protein domains and bioactive molecules. As P_{II} structure has been identified in the N-terminus of PrP, this finding constitutes another similarity between the RNA binding region of Sho and the RNA binding region of PrP. A low barrier for conversion between P_{II} conformations and β-structure (Danielsson et al., 2005) may indicate that these

peptides can readily assume β-structure which is known to form an advantageous surface for nucleic acid binding (Draper, 1995).

Some differences have emerged between Sho and Fx. Most notably, the interaction of the regularly spaced small hydrophobic residues in Sho seems to encourage the formation of hydrogen bonds and helical structure in residues 7-12. However, this effect is relatively small. Also Fx has 6 Arg residues compared to 5 in Sho, which contribute to the formation of extended conformations. As a result Sho is less likely than Fx to form extended structures.

A subsidiary functional implication arises from the way that helical structure exposes the hydroxyl of Ser10 for interaction with other binding partners. This may enable phosphorylation of Ser10 which falls within a putative PKC phosphorylation site (SAR) as observed in Chapter 2.

It appears that the RGG domain constitutes a flexible framework from which Arg residues may act as molecular beacons searching out likely binding partners. It is known that Arg is capable of multiple interactions including the formation of a salt bridge with the positively charged guanidinium sidechain, formation of up to 5 hydrogen bonds and that electron delocalization in the guanidinium group also has a pseudo aromatic character allowing Arg to participate in stacking interactions with nuclear bases. The placement of Arg residues in a Gly-rich peptide which has great torsional flexibility allows full utilization of this interaction potential as well as enhancing the likelihood that several Arg residues will participate in the binding interface. This hypothesis is explored in Chapter 6 where interactions of the Sho and Fx peptides with G-quadruplex RNA is investigated.

4.5 REFERENCES

Beck, D.A., Alonso, D.O., Inoyama, D. & Daggett, V. (2008) The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. *Proc Natl Acad Sci U S A,* 105, 12259-64.

Berendsen, H.J.C., Postma, J.P.M., Van Gunsteren, W.F., Dinola, A. & Haak, J.R. (1984) Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics,* 81, 3684-90.

Case, D.A., Cheatham, T., Darden, T., Gohlke, H., Luo, R., Mertz, K.M., Onufriev, A., Simmerling, C., Wang, B. & Woods, R.J. (2005) *J Comput. Chem.,* 26, 1668.

Chakrabartty, A., Schellman, J.A. & Baldwin, R.L. (1991) Large differences in the helix propensities of alanine and glycine. *Nature,* 351, 586-8.

Creamer, T.P. & Campbell, M.N. (2002) Determinants of the polyproline II helix from modeling studies. *Adv Protein Chem,* 62, 263-82.

Danielsson, J., Jarvet, J., Damberg, P. & Graslund, A. (2005) The Alzheimer beta-peptide shows temperature-dependent transitions between left-handed 3-helix, beta-strand and random coil secondary structures. *Febs J,* 272, 3938-49.

Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T. & Darnell, R.B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell,* 107, 489-99.

Darnell, J.C., Mostovetsky, O. & Darnell, R.B. (2005) FMRP RNA targets: identification and validation. *Genes Brain Behav,* 4, 341-9.

Draper, D.E. (1995) Protein-RNA recognition. *Annu Rev Biochem,* 64, 593-620.

Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. & Kollman, P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem,* 24, 1999-2012.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C. & Obradovic, Z. (2001) Intrinsically disordered protein. *J Mol Graph Model,* 19, 26-59.

Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z. & Uversky, V.N. (2008) The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics,* 9 Suppl 2, S1.

Feig, M., Karanicolas, J. & Brooks Iii, C.L. (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling,* 22, 377-395.

Gill, A.C., Ritchie, M.A., Hunt, L.G., Steane, S.E., Davies, K.G., Bocking, S.P., Rhie, A.G., Bennett, A.D. & Hope, J. (2000) Post-translational hydroxylation at the N-terminus of the prion protein reveals presence of PPII structure in vivo. *Embo J,* 19, 5324-31.

Griffiths-Jones, S.R., Sharman, G.J., Maynard, A.J. & Searle, M.S. (1998) Modulation of intrinsic phi,psi propensities of amino acids by neighbouring residues in the coil regions of protein structures: NMR analysis and dissection of a beta-hairpin peptide. *J Mol Biol,* 284, 1597-609.

Gunasekaran, K., Tsai, C.J., Kumar, S., Zanuy, D. & Nussinov, R. (2003) Extended disordered proteins: targeting function with less scaffold. *Trends Biochem Sci,* 28, 81-5.

Hollingsworth, S.A., Berkholz, D.S. & Karplus, P.A. (2009) On the occurrence of linear groups in proteins. *Protein Sci,* 18, 1321-5.

Hovmoller, S., Zhou, T. & Ohlson, T. (2002) Conformations of amino acids in proteins. *Acta Crystallogr D Biol Crystallogr,* 58, 768-76.

Izaguirre, J.A., Catarello, D.P., Wozniak, J.M. & Skeel, R.D. (2001) Langevin stabilization of molecular dynamics. *Journal of Chemical Physics,* 114, 2090-2098.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.,* 79, 926.

Kabsch, W. & Sander, C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers,* 22, 2577-2637.

Kyte, J. & Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol,* 157, 105-32.

Loncharich, R.J., Brooks, B.R. & Pastor, R.W. (1992) Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers,* 32, 523-35.

Maris, C., Dominguez, C. & Allain, F.H. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *Febs J,* 272, 2118-31.

Mitsutake, A. & Okamoto, Y. (2000) Helix-coil transitions of amino-acid homo-oligomers in aqueous solution studied by multicanonical simulations. *Journal of Chemical Physics,* 112, 10638-10647.

Okur, A., Wickstrom, L., Layten, M., Geney, R., Song, K., Hornak, V. & Simmerling, C. (2006) Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model. *Journal of Chemical Theory and Computation,* 2, 420-433.

Okur, A., Wickstrom, L. & Simmerling, C. (2008) Evaluation of salt bridge structure and energetics in peptides using explicit, implicit, and hybrid solvation models. *Journal Of Chemical Theory And Computation,* 4, 488-498.

Onufriev, A., Bashford, D. & Case, D.A. (2004) Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins: Structure, Function, and Bioinformatics,* 55, 383-394.

Ramachandran, G.N. & Sasisekharan, V. (1968) Conformation of polypeptides and proteins. *Adv Protein Chem,* 23, 283-438.

Ramos, A., Hollingworth, D. & Pastore, A. (2003) G-quartet-dependent recognition between the FMRP RGG box and RNA. *Rna,* 9, 1198-207.

Rhie, A., Kirby, L., Sayer, N., Wellesley, R., Disterer, P., Sylvester, I., Gill, A., Hope, J., James, W. & Tahiri-Alaoui, A. (2003) Characterization of 2'-fluoro-RNA aptamers that bind preferentially to disease-associated conformations of prion protein and inhibit conversion. *J Biol Chem,* 278, 39697-705.

Roe, D.R., Okur, A., Wickstrom, L., Hornak, V. & Simmerling, C. (2007) Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J Phys Chem B,* 111, 1846-57.

Rohl, C.A., Fiori, W. & Baldwin, R.L. (1999) Alanine is helix-stabilizing in both template-nucleated and standard peptide helices. *Proc Natl Acad Sci U S A,* 96, 3682-7.

Ryckaert, J.P., Ciccotti, G. & Berendsen, H.J.C. (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics,* 23, 327-41.

Shi, Z., Chen, K., Liu, Z. & Kallenbach, N.R. (2006) Conformation of the backbone in unfolded proteins. *Chem Rev,* 106, 1877-97.

Shi, Z., Woody, R.W. & Kallenbach, N.R. (2002) Is polyproline II a major backbone conformation in unfolded proteins? *Adv Protein Chem,* 62, 163-240.

Siligardi, G. & Drake, A.F. (1995) The importance of extended conformations and, in particular, the PII conformation for the molecular recognition of peptides. *Biopolymers,* 37, 281-92.

Sitkoff, D., Sharp, K.A. & Honig, B. (1994) Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *Journal of Physical Chemistry,* 98, 1978-88.

Smith, C.J., Drake, A.F., Banfield, B.A., Bloomberg, G.B., Palmer, M.S., Clarke, A.R. & Collinge, J. (1997) Conformational properties of the prion octa-repeat and hydrophobic sequences. *FEBS Lett,* 405, 378-84.

Smith, L.J., Fiebig, K.M., Schwalbe, H. & Dobson, C.M. (1996) The concept of a random coil. Residual structure in peptides and denatured proteins. *Fold Des,* 1, R95-106.

Srinivasan, J., Cheatham, T.E., Cieplak, P., Kollman, P.A. & Case, D.A. (1998) Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. *Journal of the American Chemical Society,* 120, 9401-9409.

Sugita, Y. & Okamoto, Y. (1999) Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters,* 314, 141-151.

Swindells, M.B., Macarthur, M.W. & Thornton, J.M. (1995) Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nat Struct Biol,* 2, 596-603.

Syme, C.D., Blanch, E.W., Holt, C., Jakes, R., Goedert, M., Hecht, L. & Barron, L.D. (2002) A Raman optical activity study of rheomorphism in caseins, synucleins and tau. New insight into the structure and behaviour of natively unfolded proteins. *Eur J Biochem,* 269, 148-56.

Tompa, P. & Csermely, P. (2004) The role of structural disorder in the function of RNA and protein chaperones. *Faseb J,* 18, 1169-75.

Uversky, V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci,* 11, 739-56.

Uversky, V.N., Gillespie, J.R. & Fink, A.L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins,* 41, 415-27.

Van Der Spoel, D. (1998) The solution conformations of amino acids from molecular dynamics simulations of Gly-X-Gly peptides: comparison with NMR parameters. *Biochem Cell Biol,* 76, 164-70.

Wright, P.E. & Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol,* 293, 321-31.

Zhang, W., Wu, C. & Duan, Y. (2005) Convergence of replica exchange molecular dynamics. *J Chem Phys,* 123, 154105.

Zhou, R. (2003) Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins,* 53, 148-61.

CHAPTER 5

DEVELOPING AND TESTING A MODEL RNA G-QUADRUPLEX

5.1 INTRODUCTION

5.1.1 Formation of G-quadruplex

The central element of the G-quadruplex is the guanine tetrad in which 4 guanine bases hydrogen bond in a cyclic formation. The hydrogen bonds formed are between the N1 and N2 (donors) of one base and the carbonyl O6 and N7 (acceptors) of another guanine base as illustrated Figure 5-1(D). This form of hydrogen bonding known as Hoogsteen bonding differs from the most common Watson-Crick base pairing in that it involves bonds between purines rather than purine-pyrimidine pairs and it utilizes a different face of the guanine base from that involved in Watson-Crick base pairing.

Within each tetrad the carbonyl O6 atoms are in square planar arrangement and form a bipyramidal structure when 2 tetrads are stacked on top of each other. The tetrad stacking is stabilized by cations. Experimental studies have shown that different cations provide varying degrees of stabilization and can dictate the type of quadruplex formed (Hardin et al., 1997). The size of the cation is a key factor in determining where it will bind in the quadruplex and also the energy penalty involved in desolvating it to allow it to enter the quadruplex. The potassium ion (K^+), with a radius of 1.33 Å, is ideally suited to occupy the position between tetrads, whereas the sodium ion (Na^+), with ionic radius of 0.95 Å, can insert within, or between, the tetrads where it coordinates with the 8 carbonyl oxygens. By contrast, the smaller lithium ion (Li^+) provides very little stabilization, most probably because it holds its solvating water more strongly and accordingly, there is a higher de-solvation penalty. NMR spectroscopy has been used to measure the dynamics of cations within G-quadruplexes. As the time during which a cation remains in the quadruplex is less than the time it takes for the stacked bases to open, it has been postulated that cations move between the quadruplex and bulk solution by passing through the axial channel of the quadruplex, similar to the movement of ions in an ion channel (Hud and Plavec, 2006).

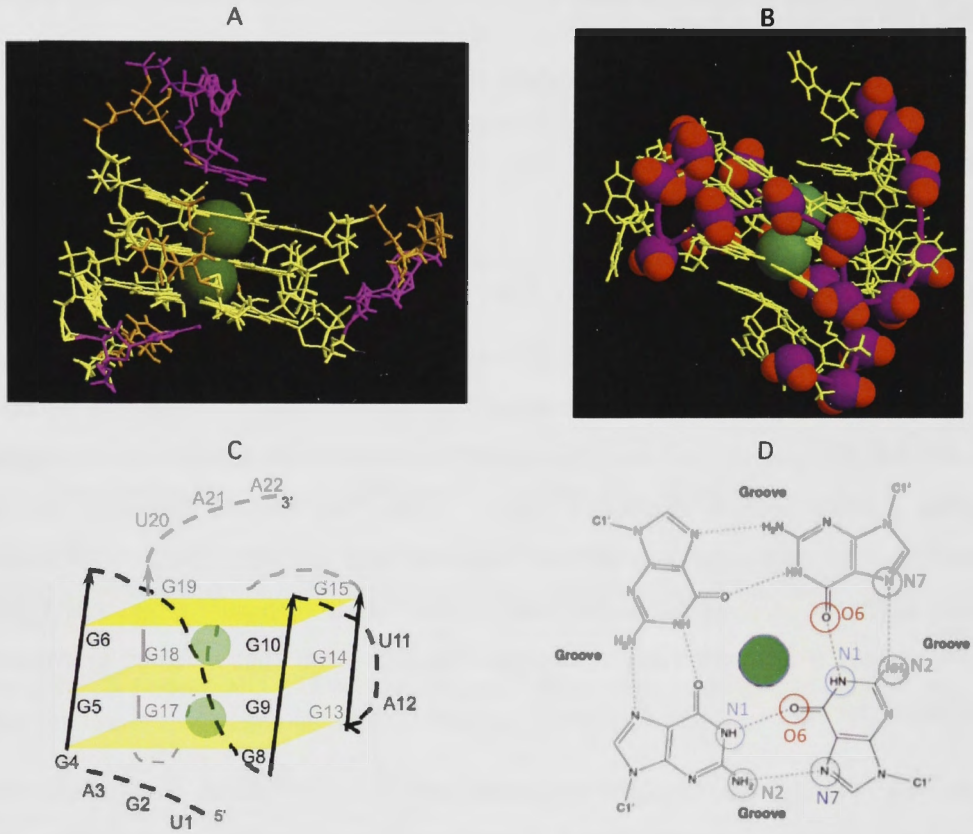


Figure 5-1. Structure of a 3 tetrad G-quadruplex
(A) Snapshot from MD simulations showing 3 planar tetrads (yellow), adenine nucleotides of the loop/termini (pink) and uracil nucleotides of the loop/termini (orange). (B) With backbone phosphate groups highlighted, phosphorous (magenta), oxygen (red). (C) Schematic diagram of the G-quadruplex. Each of the yellow planes represents a guanine tetrad; K^+ ions are green spheres. The 4 strands run in a parallel direction (black arrows) loops 1 and 3 are formed by single nucleotides and loop 2 consists of 2 nucleotides. (D) A guanine tetrad showing (G-quartet) with Hoogsteen hydrogen bonds between the 4 guanine bases marked by dotted lines. The stabilizing cation (green) is in the center. Figure modified from Burge et al. (2006)

In addition to cation-induced stability, the stacked tetrads are also stabilized by base stacking which shields the non-polar base from solvent, as well as van der Waals and electrostatic interactions. Overall, quadruplex stability results from these interactions, and also potential hydrogen bonding and base stacking within the loops.

The loops connecting the strands may vary in length and in orientation to the tetrad stack and it is possible to see a mix of lateral, diagonal and external loops. Short loops of 1 or 2 nucleotides can form lateral or external loops encompassing 3 or 4 tetrads whereas longer loops also allow for a diagonal arrangement and are necessary for larger numbers of tetrads. The loops are the most flexible elements of

the G-quadruplex which suggests that they may be well suited for adapting to a binding partner (Haider et al., 2008).

The linkage of the base to the sugar unit may be in a *syn* or *anti* arrangement as depicted in Figure 5-2. Parallel quadruplexes are formed when all are in an *anti* arrangement while an alternating pattern of *syn*, *anti*, *syn* is found in anti-parallel G-quadruplexes.

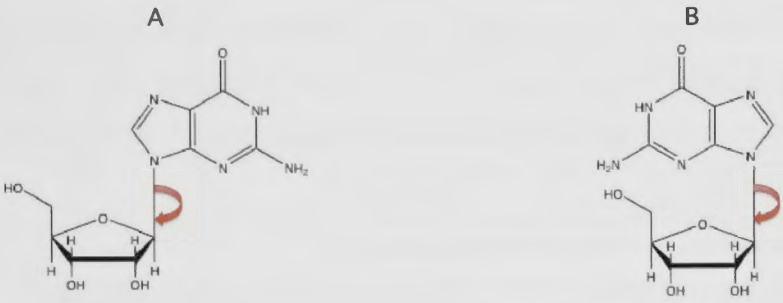


Figure 5-2. Glycosidic torsion angles of guanosine

As RNA does not easily form *syn* glycosidic angles it generally forms parallel G-quadruplexes. Recently, two groups have compared the characteristics of a series of G-rich RNA sequences with the DNA counterparts and found that in all cases the RNA formed parallel quadruplexes whereas the form of DNA quadruplex varied with sequence and buffer (Joachimi et al., 2009, Zhang et al., 2010). They also found that the parallel RNA quadruplexes were more stable than both the DNA parallel and DNA antiparallel quadruplexes. Similar results were also reported for the comparison of DNA and RNA telomeric quadruplexes (Arora and Maiti, 2009). Zhang et al. (2010) suggest that the hydroxyl group of ribose (2-OH') favours the formation of *syn* glycosidic angles and increases the stability of the RNA quadruplex by creating hydrogen bonds with water molecules which provide a supportive network around the quadruplex. MD simulations have also indicated that the ribose hydroxyl group stabilizes the RNA quadruplex by forming hydrogen bonds with the phosphate groups of the backbone (Pagano et al., 2008). Mutation of the ribose hydroxyl group to a methyl group (CH) has been found to alter the structure of an RNA quadruplex (Sacca et al., 2005). In summary, G-rich RNA is likely to form a parallel G-quadruplex of greater stability than the DNA counterpart. Furthermore, while it is

known that DNA quadruplexes are polymorphic, current evidence suggests that this may not be the case for G-rich RNA which predominantly form parallel quadruplexes.

5.1.2 Experimental studies

The structures of several G-quadruplexes have now been determined by X-ray crystallography or NMR as summarized in Table 5-1. Techniques widely used to identify but not fully characterize the G-quadruplex include recording the spectra of exchangeable imino and amino resonances by NMR, which detects the G-quadruplex through the guanine imino signals, and circular dichroism (CD) which allows differentiation between parallel and anti-parallel quadruplexes, as discussed in more detail in Chapter 9.

Table 5-1. Defined structures of DNA and RNA G-quadruplexes

	Source	Method PDB code	Comments/ Protein binding and mechanism	Ref
$G_4T_4G_4$	3 ' of *Oxytricha* telomere	X-ray structure 1D59	First crystal structure of dimeric G-Q formed by 2 hairpin sequences. 4 tetrads stabilized by K^+.	(Kang et al., 1992)
$G_4T_4G_4$ and $G_4T_4G_4T_4G_4T_4G_4$	3 ' of *Oxytricha* telomere	NMR structure	Form symmetrical bimolecular quadruplex and an intramolecular quadruplex, respectively.	(Smith and Feigon, 1992)
T_2G_4T	Tetrahymena telomeric DNA	NMR structure 139D	4 strand parallel G-Q with 4 tetrads stabilized by K^+.	(Wang and Patel, 1993a)
$AG_3(T_2AG_3)_3$	Human telomeric repeat	NMR structure 143D	Intramolecular parallel G-Q with 3 tetrads stabilized by Na^+.	(Wang and Patel, 1993b)
$G_2T_2G_2TGTG_2T_2G$ (15 mer)	Thrombin binding DNA aptamer	x-ray structure 1 HUT	Intramolecular G-Q with 2 tetrads and 3 short loops (TT, TT and TGT). Structure has G-Q aptamer bound to thrombin. Thrombin binds mainly by electrostatic interactions between NH of Lys, Arg and His and the OP groups of 3 G involved in the tetrads and Ts and G in the loops.	(Padmanabhan et al., 1993)
$G_2T_2G_2TGTG_2T_2G_2$	Thrombin binding DNA aptamer	NMR structure 148d	Intramolecular DNA with 2 tetrads and 3 short loops (TT, TT, TGT)	(Schultze et al., 1994a)
$G_4T_4G_4$	3 ' of *Oxytricha* telomere	NMR structure 156D	Dimeric anti-parallel G-Q with 4 tetrads.	(Schultze et al., 1994b)
$G_3T_4G_3$		NMR structure 1FQP	Dimeric hairpin, diagonal-looped anti-parallel G-Q with 3 tetrads stabilized by Na^+.	(Keniry et al., 1995)
$G_4(TU)_2G_4T_4G_4U_2T_2G_3X$	3 ' of *Oxytricha* telomere	NMR structure 230D	Intramolecular G-Q with diagonal central loop	(Smith et al., 1995)
GCG_2T_3GCGC	FMRP gene triple repeat	NMR structure 1A6H	Dimeric hairpin forming symmetric G-Q with 4 tetrads stabilized by Na^+	(Kettani et al., 1995)
$G_2T_2G_2TGTG_2T_2G_2$	Thrombin binding aptamer	NMR structure 1QDF	Determination of Mn^{2+} binding sites on DNA quadruplex.	(Marathias et al., 1996)
TAGG	*Bombyx mori* telomeric DNA	NMR structure 1AFF	4 strand G-Q with 2 tetrads sandwiched between novel triads	(Kettani et al., 1997)

$G_3CT_4G_3C$		NMR structure 1A8W	Dimerization of 2 hairpin DNA structures to form G-Q with different structures depending on whether the cation is K^+ or Na^+	(Bouaziz et al., 1998)
TG_2TG_2C	*Saccharomyces cerevisiae* telomere repeats	NMR structure 1EMQ	Parallel G-Q incorporating a novel T-tetrad incorporated between G-tetrads.	(Patel and Hosur, 1999)
TG_3CG_2T	SV40 viral genome repeat	NMR structure 1EVO	Novel G-Q incorporating a C-tetrad between G-tetrads.	(Patel et al., 2000)
$G_2T_2T_2TGTG_2T_2G_2$		NMR structure 1C38	Comparison of effect of K^+ concentration on formation of G-Q.	(Marathias and Bolton, 2000)
$G_3T_2CAG_2$	Designed sequence	NMR structure 1F3S	Double stranded anti-parallel G-Q in which tetrads are sandwiched between triads.	(Kettani et al., 2000)
$G_2AG_2AG_2AG_2A$		NMR structure 1MYQ	Forms intramolecular G-Q comprised of a G-tetrad and a heptad.2 G-Qs form a dimer stabilized by stacking of the heptads.	(Matsugami et al., 2001)
$G_2T_4G_2CA_3T_4G_2T$	Designed sequence	NMR structure 1I34	Intramolecular anti-parallel G-Q consisting of 2 G-tetrads stabilized by Na+ and flanked by a triad on one side.	(Kuryavyi et al., 2001)
$GAGCAG_2T$	Designed sequence	NMR structure 1JVC	Dimer of 2 hairpins forming mixed tetrad quadruplex with G-tetrad and G-C-G-C tetrad as well as A-T-A-T tetrad.	(Zhang et al., 2001a)
$G_4T_4G_4$	Telomeric DNA	X-ray structure 1JB7	Structure of dimeric G-Q complexed with binding protein onTEBP.	(Horvath and Schultz, 2001)
$G_3AG_2T_3G_3AT$		NMR structure 1JJP	Dimeric quadruplex with 2 tetrads in each monomer unit and stacked pentads at the dimeric interface.	(Zhang et al., 2001b)
XAG_3XTA_3T	Human telomeric DNA	X-ray structure 1K8P	Intramolecular parallel G-Q stabilized by K^+ with propeller like loop.	(Parkinson et al., 2002)
GGCAG		EMSA CD	Note same sequence of RNA also used	(Fukuda et al., 2002)
$(G_4T_4G_3)_2$		NMR 1lvs	Dimeric anti-parallel G-Q with 3 tetrads stabilized by Na^+.	(Crnugelj et al., 2002)
TG_4T	Telomeric DNA	NMR structure with drug bound *1O0K	4 stranded DNA G-Q bound to drug daunomycin..	(Clark et al., 2003)
$G_4T_4G_4$	Oxytricha nova telomeric DNA	X-ray structure 1L1H	Dimeric anti-parallel G-Q bound to drug molecule. Binding occurs at end of the tetrad stack within a T loop.	(Haider et al., 2003)
$(T_2AG_3)_4$	Human telomeric repeat	NMR structure 1NZM	4 stranded parallel DNA quadruplex complexed with telomerase inhibitor RHPS4. The drug molecule stacks at the ends of the tetrad stack.	(Gavathiotis et al., 2003)
GCATGCT		X-ray structure 1MF5	Complementary 4 loop quadruplex stabilized through cobalt hexamine.	(Thorpe et al., 2003)
GCGGTGGAT		NMR structure 1NYD	4 strands creating a G-Q which is a dimer of dimers.	(Webba da Silva, 2003)
$(G_2A)_8$	Common motif found in gene promoters	Structure 1OZ8	2 intramolecular parallel K^+ stabilized G-Q are formed which are packed tail to tail.	(Matsugami et al., 2003)
GCGAGAGC		NMR structure 1V3P and 1V3N	8 strands in which the G5 bases form a double tetrad in the center of the structure at high K^+ the structure is split in 2.	(Kondo et al., 2004)
$TGAG_3TG_3TAG_3TG_3TAA$	Human c-MYC promoter	NMR structure 1XAV	Intramolecular parallel G-Q with 3 tetrads stabilized by K^+ 2 loops are single nucleotide and the other a 2 nucleotide loop.	(Ambrus et al., 2005)
$GXGT_2AG_3T_2AG_3T$	Human telomeric DNA	NMR structure 2AQY	Asymmetric anti-parallel dimeric G-Q with 3 tetrads stabilized by Na^+.	(Zhang et al., 2005)
T_2AG_3	Human telomeric DNA	Structure 1X0F	Structure of single stranded DNA known to form G-Q bound with C-terminus binding domain of hRNP D which is found to unwind the G-Q.	(Enokizono et al., 2005)
$TGAG_3TG_2XGAG_3TG_4A_2G_2$	Human c-MYC promoter	Structure 2A5P	Intramolecular parallel G-Q with 3 tetrads complexed with the stabilizing molecule TMPyP4.	(Phan et al., 2005)

$TAG_3(TTAGGG)_3$	Human telomeric DNA	NMR structure	Intramolecular 3 tetrad G-Q with 3+1 structure i.e. 3 strands in same direction and the other in the opposite direction.	(Luu et al., 2006)
$TA(G_3T_2A)_3G_3$ and $TA(G_3T_2A)_3G_3T_2$	Human telomeric DNA	NMR structure	Both quadruplexes have the 3+1 structure with one double-chain-reversal loop and 2 edgewise loops but differ in loop appearance.	(Phan et al., 2007)
$(T_2AG_3)_4T_2$	Human telomeric DNA	NMR structure	Hybrid-type intramolecular quadruplex with mixed parallel/antiparallel strands.	(Dai et al., 2007)
$(G_3T_2A)_3G_3T$	Human telomeric DNA	NMR structure	A 2 tetrad antiparallel basket type quadruplex is formed in K^+ solution.	(Lim et al., 2009)
RNA	**Source Structure**	**Method PDB code**	**Comments**	**Ref**
$(UGGGGU_4$	3' terminus of Escherichia coli 5S RNA	NMR structure 1 RAU	4 strand quadruplex with 3 G-tetrads and 2 U-tetrads in the central core stabilized by K^+.	(Cheong and Moore, 1992)
$G_2AG_2U_4G_2AG_2$		NMR structure 1MY9	Dimeric G-Q with each dimer forming a G-tetrad and a hexad (when 2 Ade bind to the edges of the 2nd G-tetrad). The hexads stack to form the dimer interface.	(Liu et al., 2002)
	Human telomeric RNA	NMR structure 2KBP	Dimeric parallel propeller type G-Q with 3 tetrads stabilized by K^+.	(Martadinata and Phan, 2009)

5.1.3 Protein binding

Proteins found to bind to the quadruplex structure include:

(i) those that promote quadruplex formation such as the Lys rich β-subunit of the *Oxytricha* telomere-binding protein which acts as a chaperone and accelerates G-quadruplex formation in the *Oxytricha* telomere (Fang and Cech, 1993),

(ii) those which unwind the quadruplex structure such as the UP1 protein which is a proteolytic product of hnRNP A1 (Fukuda et al., 2002, Fan et al., 2009, Paramasivam et al., 2009),

(iii) those which stabilize the G-quadruplex e.g. the QUAD protein (Weisman-Shomer and Fry, 1994) and the unimolecular quadruplex telomere-binding protein 25 (uqTBP25) (Erlitzki and Fry, 1997),

(iv) enzymes that have G-quadruplex dependent activities including examples from the following classes: topoisomerases (Chung et al., 1992), nucleases (Liu and Gilbert, 1994) and helicases (Sun et al., 1998), and

(v) transcription factors such as MyoD (Walsh and Gualberto, 1992) and c-MYC (Siddiqui-Jain et al., 2002).

5.1.4 Aims of this chapter

Molecular dynamics has been widely used to explore the G-quadruplex (Spackova et al., 2003, Stefl et al., 2003, Fadrna et al., 2004, Sponer and Spackova, 2007, Haider et al., 2008) as reviewed by Sponer and Spackova (2007). However, to my knowledge this is the first time that the hybrid-REMD method has been used to investigate G-quadruplex nucleic acids. Here hybrid-REMD simulations are performed on a parallel DNA G-quadruplex derived from the promoter region of the human c-MYC oncogene and its RNA analogue. Experimental studies have indicated that the RGG box of the Fragile X Mental Retardation Protein binds to quadruplex RNA which has a similar structure to the c-MYC complex (Menon and Mihailescu, 2007).

Accordingly, the primary aim of this study is to compare the DNA and RNA quadruplexes, to assess the stability of the RNA quadruplex and evaluate its suitability for use as a model in docking studies with RGG box peptides.

5.2 METHODS

5.2.1 G-quaduplex models

The structure of the modified c-MYC oncogene promoter region, available in the Brookhaven Protein Data Bank (PDB) 1XAV (Ambrus et al., 2005) was used as the model DNA G-quadruplex in these simulations. This quadruplex consists of 3 tetrads and parallel strands with the chain folding back on itself between successive strands via the loop nucleotides, referred to as a 'double-chain reversal' (Phan et al., 2004, Seenisamy et al., 2004). The 22 nucleotide sequence is set out in Figure 5-3 together with the RNA analogue.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
DNA-xav1	T	G	A	G	G	G	T	G	G	G	T	A	G	G	G	T	G	G	G	T	A	A
RNA-xav1	U	G	A	G	G	G	U	G	G	G	U	A	G	G	G	U	G	G	G	U	A	A

Figure 5-3. G-quadruplex sequence
The c-MYC oncogene G-quadruplex consisting of 22 nt and its RNA analogue. The guanines that form a tetrad are colour coded, tetrad 1 (orange), tetrad 2 (yellow), tetrad 3 (brown).

The PDB file of the DNA quadruplex was manually edited and then the LEaP program within AMBER 9 was used to convert this DNA structure into the RNA equivalent.

5.2.2 Software and Force Fields

All simulations were performed with the AMBER 9 (Case et al., 2005) suite of programs. The LEaP program was used to prepare input files and the SANDER module was used for all simulations. The ff99 forcefield (Wang et al., 2000) as modified by the Simmerling group 'the Stony Brook modification ff99SB (Hornak et al., 2006) was used with the RNA parameters in the AMBER file rna.ff99 for the RNA quadruplex simulations.

5.2.3 Hybrid-REMD simulations of DNA and RNA G-quadruplexes

The hybrid-REMD method (Okur et al., 2006) as implemented in Amber 9 was used to perform these calculations. Each quadruplex (RNA and DNA) was solvated in a 10 Å radius octahedron of TIP3P (Jorgensen et al., 1983) water molecules (3283 waters for the DNA quadruplex and 2818 waters for the RNA quadruplex which was slightly more compact). Counterions (Na^+) were used to create a charge neutral system. A 1 Å radius octahedron of waters was retained in the stripped files used for calculating the exchange probability using the hybrid-REMD algorithm.

All models were energy minimized and then each replica was equilibrated by firstly raising the temperature from 0.1 K to 100 K, at constant volume. Following this, the temperature was raised from 100 K to the desired temperature for each replica under constant pressure over 120,000 steps with a time step of 1 fs. Harmonic restraints with force constants of 10 kcal mol^{-1} $Å^{-2}$ were used to restrain the nucleic acid and structural K^+ ions. Preliminary testing revealed that the quadruplex structure became unstable using the temperature scale normally used in hybrid-REMD. These simulations were therefore conducted using only a moderate increase in temperature, the temperatures of the 8 replicas, in Kelvin were 300, 310.6, 321.7, 333.2, 345.1, 357.4, 370.1, 383.3. Once the desired temperature had been reached the restraints on the nucleic acid and structural K^+ ions were relaxed from 10 kcal mol^{-1} $Å^{-2}$ to 1 kcal mol^{-1} $Å^{-2}$ over 100000 steps and then from 1 kcal mol^{-1}

$Å^{-}$ to 0.1 kcal mol^{-1} $Å^{2}$ over a further 100,000 steps. Following this 100000 steps were completed without any restraints.

The production run was continued for 20 ns at constant pressure with a pressure relaxation time of 2 ps. Temperature was kept constant using the weak coupling algorithm (Berendsen et al., 1984). The time constant for heat bath coupling was set to 0.1 ps. The simulations were carried out using a 2 fs time step with the SHAKE algorithm used to constrain bonds involving hydrogen (Ryckaert et al., 1977). Periodic boundary conditions were imposed. The particle-mesh Ewald (PME) procedure was used to calculate long-range electrostatic interactions, the direct space sum was limited with a cutoff of 10 Å, which was also the cutoff for van der Waals interactions.

5.2.4 MD simulation of RNA G-quadruplex without replica exchange

A separate 20 ns MD simulation of the RNA quadruplex was performed without replica exchange. The starting point for this simulation was the replica equilibrated to 300 K following the equilibration procedure referred to above. The MD simulation was performed using the same parameters described above for the production run.

5.2.5 Trajectory analysis

The Ptraj module of Amber 9 was used to calculate radius of gyration, RMSD, atomic fluctuations, hydrogen bonds and the distance between residues. Structures are imaged using RasMol and VMD. Unless otherwise indicated the results are presented for the trajectories generated at 300 K.

5.3 RESULTS AND DISCUSSION

5.3.1 Flexibility and behaviour of the quadruplex loops and tetrads

Time averaged structures of the RNA and DNA quadruplexes calculated in the first part of the production run (100 – 300 ps) confirm the stability of structures over the initial part of the simulation (Figure 5-4).

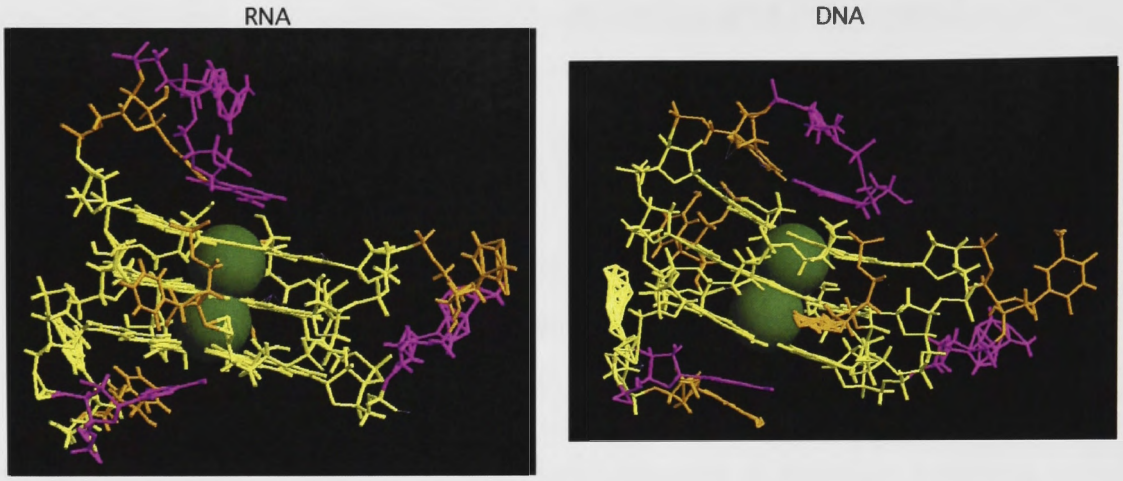


Figure 5-4. RNA quadruplex and DNA quadruplex
Time averaged structures (100 - 300 ps with frames taken every 10 ps) of (A) the RNA quadruplex and (B) the DNA quadruplex, depicting the 3 tetrads (yellow) the adenine loop/termini nucleotides (pink) and the uracil loop/termini nucleotides (orange).

The stability of the various components of the 2 quadruplexes was evaluated by calculating the root-mean-square deviation (RMSD) of different components over the 20 ns simulation (Table 5-2) as well as the atomic fluctuation (AF) of each nucleotide as plotted in Figure 5-5. The RMSD is a measure of the degree of variation from the starting structure whereas the atomic fluctuation is a measure of the variation from frame to frame and is related to the B factor (temperature factor) of crystal structures by a scaling factor of $8/3\ \pi^2$.

Table 5-2. RMSD values for RNA and DNA quadruplex components of 300K replica

Component	Nucleotides	RNA RMSD [a]	DNA RMSD [b]
5' terminus	1-3	1.58	1.39
Tetrad 1	4, 8,13,17	1.38	1.24
Tetrad 2	5, 9, 14, 18	1.01	0.59
Tetrad 3	6, 10, 15, 19	0.59	0.45
Loop 1	7	1.1	1.01
Loop 2	11, 12	0.94	1.00
Loop 3	16	0.63	1.01
3' terminus	20-22	0.95	0.77
Whole sequence	1-22	3.57	3.58
Whole sequence (first 5 ns only)	1-22	3.29	3.3

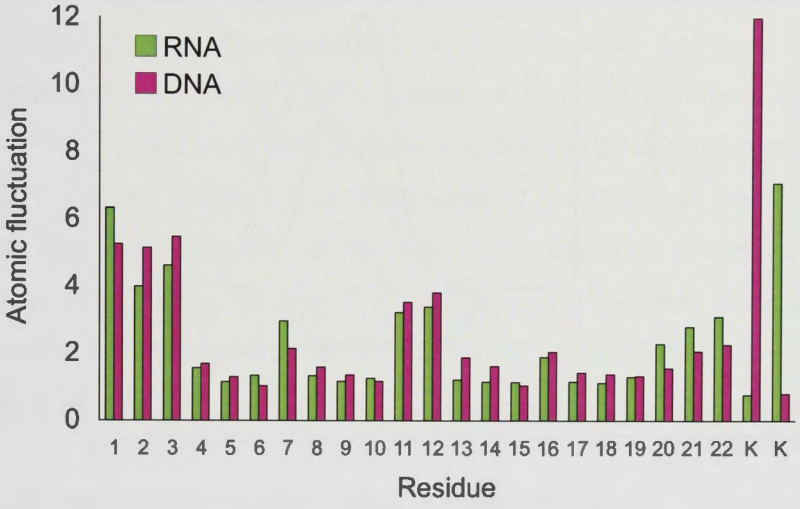


Figure 5-5. Atomic fluctuation of RNA and DNA nucleotides
Atomic fluctuation of each nucleotide and the structural K^+ ions for RNA quadruplex (green) and DNA quadruplex (pink).

The RMSD values reported in Table 5-2 show a high degree of similarity between the DNA and RNA quadruplexes with both having a whole sequence RMSD of 3.3 Å. The RMSD data, in conjunction with the atomic fluctuations, reveal that both quadruplexes experience most flexibility in the 3 nucleotides at the 5' terminus. By comparison, the 3' terminus is relatively stabilized. The RMSD values of the 3 loops in the DNA quadruplex are very similar to one another, however, the atomic fluctuations reveal that the propeller loop that extends out from the quadruplex (loop 2, nucleotides 11 and 12) is the most flexible of the loops in both the DNA and RNA. Unsurprisingly, the nucleotides comprising the tetrads are the most stable. The biggest difference in flexibility between the RNA and DNA is seen in tetrad 2 and loop 3. Tetrad 2 is more stable in the DNA whereas loop 3 is more flexible. Overall, the increased flexibility of loops 2 and 3 in the DNA may account for the slightly broader distribution of radius of gyration seen in the DNA compared to the RNA (Figure 5.6). The high atomic fluctuation seen in the structural potassium ions will be discussed further in section 5.3.5.

A frequency distribution of radius of gyration is given in Figure 5-6. Although there is not much difference in the range of these distributions (11.5-12.8 Å for the RNA and 11.5-13 Å for the DNA), the RNA quadruplex is a little more compact than the DNA which is consistent with slightly less mobility of the loop regions seen in the RMSD results.

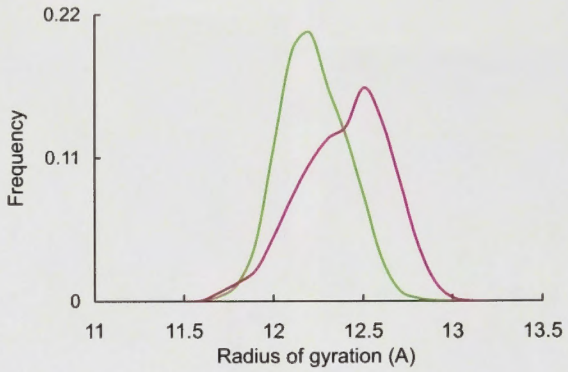


Figure 5-6. Radius of gyration frequency distributions for RNA (green) and DNA (pink) quadruplexes

It could be expected that the nucleotides comprising the loops and terminal regions might form hydrogen bonds with other parts of the quadruplex. To investigate this the hydrogen bonds formed by the bases of non-tetrad nucleotides were calculated. Those occurring at least 15% of the time are presented in Table 5-3.

Table 5-3 Hydrogen bonds involving loop and terminal nucleobases in the RNA and DNA quadruplexes

RNA Acceptor [a]	RNA Donor [b]	Occ % [c]	Distance [d]	DNA Acceptor [a]	DNA Donor [b]	Occ % [c]	Distance [d]
:22@N7	:20@N3	71.6	2.975	:22@N7	:20@N3	97.4	2.933
:20@O4	:15@N1	66.8	3.125	:20@O4	:22@N6	74.2	3.023
:20@O4	:22@N6	64	3.012	:20@O4	:15@N2	50.8	3.264
:19@O6	:20@N3	49.7	3.247	:20@O2	:19@N1	50.5	3.29
:20@O4	:15@N2	46.1	3.249	:15@O6	:22@N6	44.7	3.175
:20@O2	:19@N1	43.7	3.208	:19@O6	:20@N3	29	3.323
:15@O6	:22@N6	39.2	3.238	:2@N3	:17@N2	27.1	3.077
:2@N3	:17@N2	36.3	3.036	:2@O6	:18@N2	22.6	2.829
:21@N3	:20@N3	28.4	3.304	:8@O6	:3@N6	20	3.216
:12@N7	:10@N2	23.9	3.089	:3@N1	:2@N2	19.7	3.019
:19@N7	:20@N3	22.4	3.283	:18@N3	:2@N2	18.7	3.113
:22@N7	:19@N1	20	3.202	:22@N3	:10@N1	18.7	3.332
:7@O4	:4@N2	16.6	2.934	:21@N3	:20@N3	18.4	3.333
:8@O6	:3@N6	16.1	3.243	:15@N1	:22@N6	18.2	3.361
:22@N3	:10@N1	15.5	3.358	:12@N1	:10@N2	17.4	3.049
:17@N3	:2@N2	15	3.172	:3@N1	:17@N1	15.8	3.177
				:3@N7	:4@N1	15	3.271

[a] The nucleotide number and atom acting as hydrogen bond acceptor. [b] The nucleotide number and heavy atom associated with the donor hydrogen. [c] The percentage occupation of this hydrogen bond over the 20 ns simulation performed at 300 K. [d] The average distance (Å) between heavy atoms involved in the hydrogen bond over the course of the simulation. Atoms belonging to loops or terminal regions are highlighted in red font.

The bases at the 3' terminus of the RNA and DNA quadruplexes tend to form hydrogen bonds with each other and with the guanine bases at the top of the tetrad

stack, which stabilizes this flap over the tetrad. This behaviour contrasts with bases at the 5' terminus which do not form hydrogen bonds as frequently and which generally retain more flexibility. This is consistent with the difference in flexibility seen in the RMSD results presented in Table 5-2 and the atomic fluctuations of the nucleotides. The bases involved in all three loops of the DNA and loops 2 and 3 of the RNA do not form hydrogen bonds with other bases in the quadruplex and therefore appear to be free to engage in interactions with other binding partners. The first loop in the RNA quadruplex (U7) forms a hydrogen bond with G4 17% of the time.

5.3.2 Hydrogen bonds formed by the hydroxyl group of ribose

A key difference between DNA and RNA is the presence of the hydroxyl group (2-OH') on the ribose sugar in RNA. It is therefore of interest to examine whether this group forms hydrogen bonds that affect the structure of the G-quadruplex. The hydroxyl group can act as hydrogen bond donor or acceptor. Here, hydrogen bonds involving 2-OH' of ribose were calculated and reveal that most hydrogen bonds involve donation of hydrogen to the phosphate oxygen of the next nucleotide. The hydrogen bonds occupied > 40% of the time are presented in Table 5-4. The nucleotide and atoms involved in the hydrogen bonds are noted as well as the occupancy of the bond over the course of the simulation. For example, the first entry refers to the hydroxyl group of nucleoside 1 (Donor) making a hydrogen bond with the phosphate group of nucleotide 2 (Acceptor), 63.6% of the simulation time, and the average distance of that hydrogen bond is 2.695 Å.

Table 5-4. Hydrogen bonds formed by the ribose hydroxyl group in the RNA quadruplex

Acceptor [a]	Donor [b]	Occupancy (%) [c]	Distance (Å) [d]
:2@O1P	:1@HO'2	63.6	2.695
:3@O1P	:2@HO'2	66.7	2.796
:4@O1P	:3@HO'2	60.6	2.421
:5@O1P	:4@HO'2	51.5	2.52
:6@O1P	:5@HO'2	45.5	2.511
:7@O1P	:6@HO'2	57.6	2.827
:8@O1P	:7@HO'2	66.7	2.623
:9@O1P	:8@HO'2	51.5	2.684
:10@O1P	:9@HO'2	45.5	2.579
:11@O1P	:10@HO'2	63.6	2.73
:12@O1P	:11@HO'2	51.5	2.87
:13@O1P	:12@HO'2	60.6	2.729
:14@O1P	:13@HO'2	60.6	2.707
:15@O1P	:14@HO'2	51.5	2.493
:16@O1P	:15@HO'2	57.6	2.856
:17@O1P	:16@HO'2	57.6	2.799
:18@O1P	:17@HO'2	42.4	2.658
:19@O1P	:18@HO'2	42.4	2.536
:20@O1P	:19@HO'2	60.6	2.635
:22@O1P	:21@HO'2	72.7	2.864
:12@O2'	:9@N2	53.4	3.043

[a] The nucleotide number and atom acting as hydrogen bond acceptor. [b] The nucleotide number and hydrogen atom acting as hydrogen bond donor. [c] The percentage occupation of this hydrogen bond over the 20 ns simulation. [d] The average distance between heavy atoms involved in the hydrogen bond over the course of the simulation. The hydrogen bond in which the ribose hydroxyl group acts as acceptor is highlighted in red font.

Only one hydrogen bond involving the oxygen of the ribose hydroxyl group as acceptor (highlighted in red) was occupied > 40 % of the simulation time. The results reported in Table 5-4 show that backbone hydrogen bonds formed by ribose are likely to enhance the stability of the RNA quadruplex compared to the DNA quadruplex. In addition, the hydrogen bond formed between the O2' of nucleotide 12 (loop 2) and G9 may be a factor in the reduced flexibility of loop 2 compared to the DNA quadruplex.

5.3.3 Cation instability

Although the RNA and DNA G-quadruplex structures remained relatively stable, in 2 of the 8 replicas the K^+ ions relocated over the course of the simulation. This is illustrated by a snapshot of an arbitrarily chosen moment in the simulation. Each of the 8 replicas is presented in Figures 5-7 and 5-8. As the unstable replicas will

exchange into different temperatures over the course of the simulation it is not relevant to identify the replicas in this illustration.

Figure 5-7. Snapshots taken at 19 ns of all 8 replicas used in the simulation of the RNA quadruplex illustrating cation instability in 2 replicas
Snapshots taken at 19 ns of each of the 8 replicas involved in the hybrid-REMD simulations of the RNA quadruplex. Two of the 8 replicas (bottom line) display cation instability. The surrounding water molecules are removed for clarity of presentation.

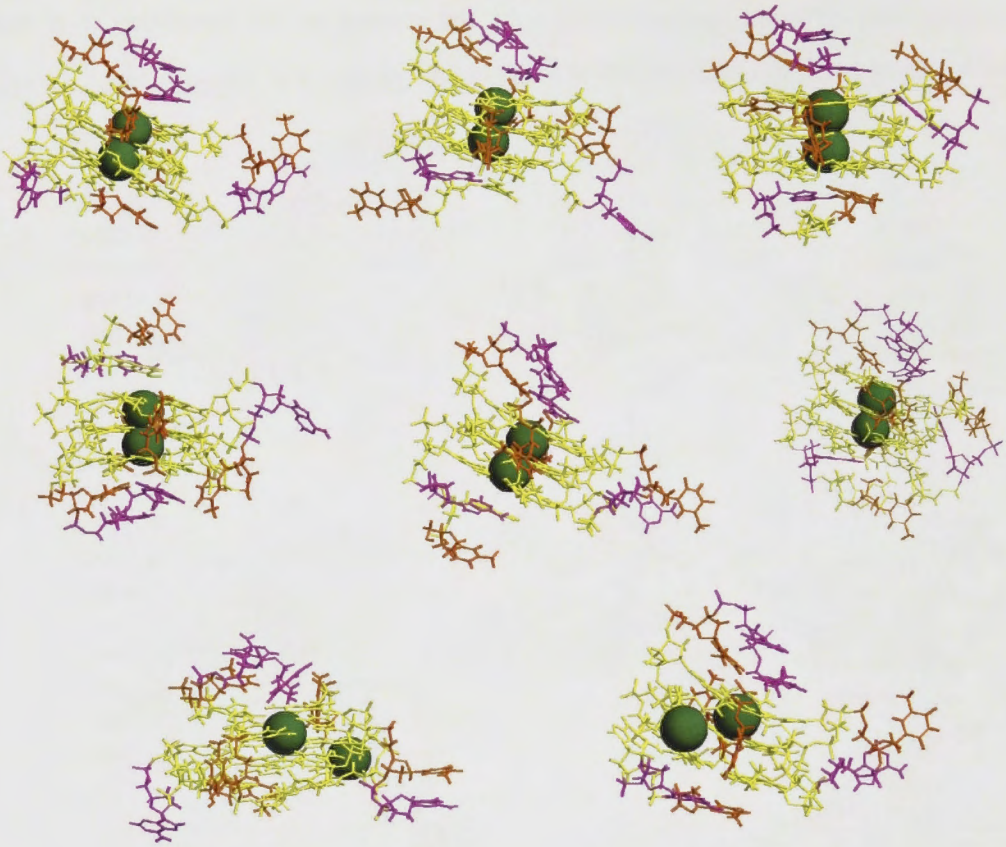


Figure 5-8. Snapshots taken at 19 ns of all 8 replicas used in the simulation of the DNA quadruplex illustrating cation instability in 2 replicas
Snapshots taken at 19 ns of each of the 8 replicas involved in the hybrid-REMD simulations of the DNA quadruplex. Two of the 8 replicas (bottom line) display cation instability. The surrounding water molecules are removed for clarity of presentation.

It is interesting that this phenomenon occurred in both the RNA and DNA quadruplex simulations after ~ 5 ns. In the RNA quadruplex the K^+ ion between the second and third tetrads changed position while in the DNA the K^+ ion between the first and second tetrads relocated. These replicas were swapped into each temperature over the course of the simulation. As a result, large variations in the K^+ - K^+ distance are seen in the simulations conducted at 300 K as illustrated in Figure 5-9. This factor does not appear to have a significant effect on the energy of the system as replica swaps still occur frequently.

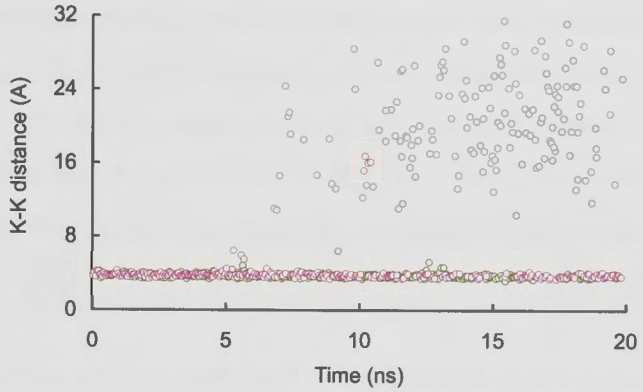


Figure 5-9. K^+-K^+ distance for RNA (green) and DNA (pink) quadruplexes simulated at 300 K

Although cations are known to move between the quadruplex and bulk solution in physical systems (Hud and Plavec, 2006), an affect which has also been simulated *in silico* (Pagano et al., 2008), the phenomenon observed in these simulations appears to be a result of the inability of the force field to capture short range polarization effects necessary to accurately describe the cation-guanine interactions, as previously noted (Spackova et al., 1999).

Spackova and colleagues (1999) conducted shorter simulations of a 4 tetrad DNA G-quadruplex using the Cornell force field (Cornell et al., 1995) and noted instability in the Na^+ cations inserted between the guanine tetrads. They note that one Na^+ cation 'suddenly jumped' from one position to another after 1.4 ns, leaving the second Na^+ cation unaffected. In a second simulation they noted that one of the 3 cations in the central channel of the quadruplex left the molecule entirely after 200 ps leaving the other 2 cations to redistribute in the channel. They attributed this phenomenon to the fact that the pair additive nature of the force field neglects short-range polarization effects whereas in reality such complexes show pronounced polarization effects. The authors note that this shortcoming should not introduce any significant error into the overall outcome of the simulations since the long range electrostatic term is properly included. They also conclude that while removal of all cations is destabilizing the quadruplex structure can sustain a reduction of the number of cations in the channel without significant perturbation (Spackova et al., 1999).

To better understand the degree of instability introduced into the present simulations as a result of cation mobility I assessed the stability of the Hoogsteen hydrogen bonds formed between the guanine bases comprising each tetrad. The hydrogen bond occupancy for the 300 K simulations is presented in Table 5-5 together with the average distance of the hydrogen bonds over the course of the 20 ns simulations.

Table 5-5. Hoogsteen hydrogen bond occupancy in the RNA and DNA quadruplexes over the course of the simulations performed at 300 K

Acceptor [a]	Donor [b]	RNA [c]	Distance [d]	DNA [e]	Distance [d]
		Tetrad 1			
:4@N7	:17@N2	99	3.00	97.1	3.00
:4@O6	:17@N1	98.8	2.95	98.6	2.94
:8@N7	:4@N2	99.6	2.98	82.0 [f]	3.02
:8@O6	:4@N1	99.7	2.94	86.5	2.95
:13@N7	:8@N2	99.3	2.99	94.2	3.00
:13@O6	:8@N1	99.4	2.96	93.9	2.96
:17@N7	:13@N2	95.7	3.02	86.5	3.01
:17@O6	:13@N1	98	2.95	87.8	2.95
		Tetrad 2			
:5@N7	:18@N2	99.6	2.95	99.9	2.95
:5@O6	:18@N1	96.1	3.09	98	3.02
:9@N7	:5@N2	98.1	3.00	99.6	2.98
:9@O6	:5@N1	95.4	3.09	93	3.13
:14@N7	:9@N2	98.5	2.98	99.3	2.99
:14@O6	:9@N1	97.2	3.08	98.1	3.01
:18@N7	:14@N2	98.2	2.98	99.5	2.98
:18@O6	:14@N1	95.9	3.08	95.9	3.09
		Tetrad 3			
:6@N7	:19@N2	85.8	3.00	99.7	2.99
:6@O6	:19@N1	86.4	2.96	99.7	2.98
:10@N7	:6@N2	99.3	2.98	99.7	2.99
:10@O6	:6@N1	98.1	2.97	99.9	2.94
:15@N7	:10@N2	85.3	3.02	99.5	3.00
:15@O6	:10@N1	85.8	2.94	99.8	2.95
:19@N7	:15@N2	99.4	2.98	99.8	2.98
:19@O6	:15@N1	92.1	2.98	99.8	2.96

[a] The nucleotide number and atom acting as hydrogen bond acceptor. [b] The nucleotide number and heavy atom associated with the hydrogen donor. [c] The percentage occupation of this hydrogen bond over the 20 ns simulation of the quadruplex performed at 300 K. [d] The average distance between heavy atoms involved in the hydrogen bond over the course of the simulation. [e] The percentage occupation of this hydrogen bond over the 20 ns simulation of the DNA quadruplex performed at 300 K. [f] Hydrogen bond occupancy < 95% is highlighted in red font.

The results presented in Table 5-5 show a reduced occupancy of up to 15% in some of the hydrogen bonds required for tetrad stability. These data indicate instability in the third tetrad of the RNA which is consistent with the cation situated between the

second and third tetrad moving to a position above the third tetrad as portrayed in Figure 5-10.

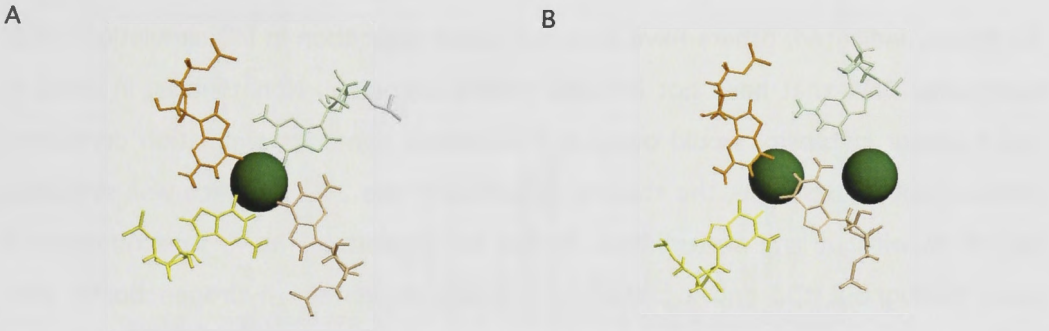


Figure 5-10. Comparing a stable and unstable replica from the RNA quadruplex simulation
Snapshots taken at 19 ns of 2 replicas involved in the hybrid-REMD simulations of the RNA quadruplex. View is looking down onto the top of tetrad 3. (A) A stable replica in which the 2 K^+ ions are in vertical alignment, only one is visible in the quadruplex channel from this perspective. (B) In this replica the K^+ ion between the second and third tetrad (the top ion) has moved out of the quadruplex channel resulting in some instability of the third tetrad.

The hydrogen bond analysis shows a similar degree of instability in the first tetrad of the DNA. This is consistent with the images presented in Figure 5-10 which depict the relocation of the cation from its position between the first and second tetrad. Visual inspection of the quadruplex shows that there is a downward tilt of bases in Tetrad 1, most pronounced in G13 as shown in Figure 5-11.

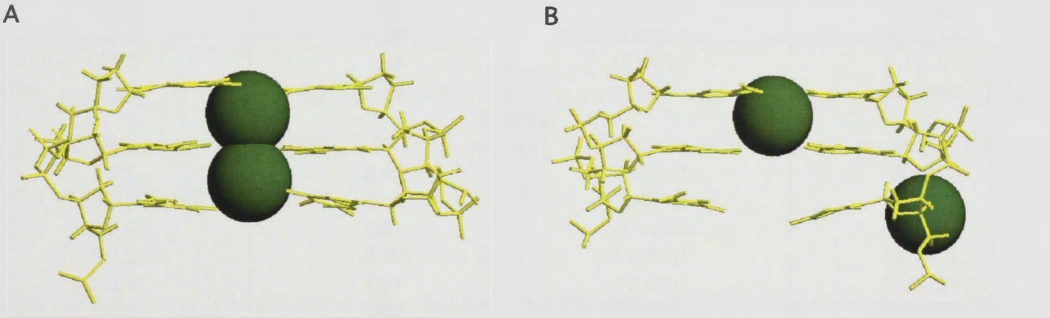


Figure 5-11. Snapshots comparing the tetrad stack of a stable and unstable replicas from the DNA quadruplex simulation
Snapshots taken at 19 ns of 2 of the replicas involved in the hybrid-REMD simulations of the DNA quadruplex. The view is from the from the front of the quadruplex. (A) A stable replica with the K^+ ions in place between the tetrads. (B) Unstable replica in which the K^+ ion between Tetrads 1 and 2 ad has relocated resulting in instability of Tetrad 1.

5.3.4 Comparing MD and hybrid-REMD simulations of the RNA quadruplex

As already indicated, others have reported cation relocation in MD simulations of G-quadruplex DNA that have not included replica exchange. Nonetheless, in order to see if similar instability would occur in a molecular dynamics simulation conducted without replica exchange, the starting structure of the 300 K replica was simulated for 20 ns without any perturbation. In this MD simulation the K^+ ions remained in place throughout the entire simulation and the Hoogsteen hydrogen bonds were occupied no less than 98% of the time (Table 5-6). The tetrad stack remained extremely stable over the course of the simulation.

Table 5-6. Hoogsteen hydrogen bond occupancy in the RNA quadruplex over a 20 ns MD simulation without replica exchange

Acceptor	Donor	Occupancy	Distance avg (A)
	Tetrad 1		
:4@N7	:17@N2	100	2.96
:4@O6	:17@N1	99.2	2.97
:8@N7	:4@N2	98.7	3.02
:8@O6	:4@N1	99.7	2.94
:13@N7	:8@N2	99.2	2.98
:13@O6	:8@N1	100	2.95
:17@N7	:13@N2	99.5	3.00
:17@O6	:13@N1	100	2.94
	Tetrad 2		
:5@N7	:18@N2	100	2.97
:5@O6	:18@N1	97.6	3.08
:9@N7	:5@N2	99.2	2.98
:9@O6	:5@N1	96.8	3.08
:14@N7	:9@N2	100	2.96
:14@O6	:9@N1	98.4	3.07
:18@N7	:14@N2	100	2.97
:18@O6	:14@N1	98.4	3.08
	Tetrad 3		
:6@N7	:19@N2	99	3.04
:6@O6	:19@N1	99.5	2.94
:10@N7	:6@N2	100	3.01
:10@O6	:6@N1	100	2.94
:15@N7	:10@N2	99.5	3.02
:15@O6	:10@N1	100	2.93
:19@N7	:15@N2	100	2.96
:19@O6	:15@N1	100	2.93

[a] The nucleotide number and atom acting as hydrogen bond acceptor. [b] The nucleotide number and heavy atom associated with the hydrogen donor. [c] The percentage occupation of this hydrogen bond over the 20 ns MD simulation of the RNA quadruplex. [d] The average distance between heavy atoms involved in the hydrogen bond over the course of the simulation.

The atomic fluctuations of each nucleotide and the K^+ ions from the 20 ns MD simulation of the RNA quadruplex are plotted in Figure 5-12.

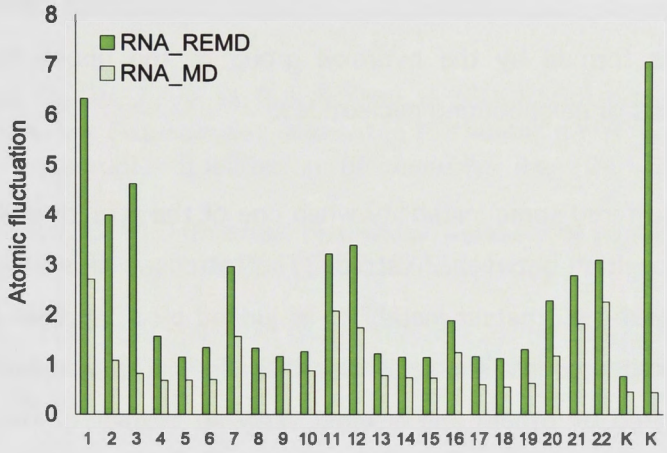


Figure 5-12. Comparison of atomic fluctuation in hybrid-REMD and MD simulations of the RNA G-quadruplex
Atomic fluctuation (Å) of each nucleotide of the RNA quadruplex and the structural K^+ ions over the course of the 20 ns hybrid-REMD (bright green) and the 20 ns MD simulation (light green). Both trajectories were generated at 300 K.

The atomic fluctuations of each nucleotide follow the same trend as occurred in the hybrid-REMD simulation although the amount of fluctuation is less, as would be expected. This is illustrated in Figure 5-12 where the RNA results presented earlier in Figure 5-5 are compared to the results of the 20 ns molecular dynamics simulation.

5.4 CONCLUSIONS

The primary conclusion to be drawn from this work is that the RNA analogue of the c-MYC DNA forms a stable 3 tetrad parallel G-quadruplex and is a sound model for use in MD simulations.

The simulations reveal that the tetrad stacks are stable except when perturbed through the relocation of the structural K^+ ions. The 5' terminus and the propeller loop are the most flexible components of both quadruplexes, as measured by RMSD and atomic fluctuation of each nucleotide, with slightly greater flexibility seen in the DNA compared to the RNA. Overall, the loops have very little engagement with the tetrad stack and retain the flexibility to interact with other molecules. The RNA is a little more compact than the DNA, as shown by the distribution of radius of gyration, with slightly reduced flexibility in the loops compared with the DNA

quadruplex. This may be explained by the extra stabilization of the RNA due to hydrogen bonds formed by the hydroxyl group of the ribose moieties and the phosphate groups of neighbouring nucleotides.

Both systems suffered some instability when one of the structural K^+ ions relocated away from its position between 2 tetrads. The tetrads in closest contact with the mobile ion showed the greatest instability as judged by a decrease in occupancy of Hoogsteen hydrogen bonds. The relocation of K^+ ions in quadruplex simulations has also been reported by others and is most likely an artifact of the non-polarizable force field used in these calculations. The total hybrid-REMD simulation time (8 replicas @ 20 ns per replica) is longer than the total time of the single MD simulation (20 ns). Combined with the enhanced sampling due to exchange of replicas at different temperatures, the hybrid-REMD simulations explore more conformational space than the single MD simulation. Therefore, problems inherent in the method are more likely to manifest in the hybrid-REMD simulations compared a single MD run.

A strong motivation for this work was to test the RNA quadruplex in anticipation of performing docking studies with the Sho and Fx peptides. This work has drawn attention to a weakness in the force field, which can lead to cation relocation and some destabilization of the quadruplex. Despite the risk of K^+ ion relocation, the RNA analogue of the c-MYC quadruplex remained stable in the simulations performed at 300 K and over a conservative timeframe of ~ 20 ns. Future simulations, including docking studies, should be monitored to ensure that they do not reach the point where the force field inadequacy is encountered, leading to K^+ ion relocation.

5.5 REFERENCES

Ambrus, A., Chen, D., Dai, J., Jones, R.A. & Yang, D. (2005) Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry,* 44, 2048-58.

Arora, A. & Maiti, S. (2009) Differential biophysical behavior of human telomericRNA and DNA quadruplex. *J Phys Chem B,* 113, 10515-20.

Berendsen, H.J.C., Postma, J.P.M., Van Gunsteren, W.F., Dinola, A. & Haak, J.R. (1984) Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics,* 81, 3684-90.

Bouaziz, S., Kettani, A. & Patel, D.J. (1998) A K cation-induced conformational switch within a loop spanning segment of a DNA quadruplex containing G-G-G-C repeats. *J Mol Biol,* 282, 637-52.

Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B. & Woods, R.J. (2005) The Amber biomolecular simulation programs. *J Comput Chem,* 26, 1668-88.

Cheong, C. & Moore, P.B. (1992) Solution structure of an unusually stable RNA tetraplex containing G- and U-quartet structures. *Biochemistry,* 31, 8406-14.

Chung, I.K., Mehta, V.B., Spitzner, J.R. & Muller, M.T. (1992) Eukaryotic topoisomerase II cleavage of parallel stranded DNA tetraplexes. *Nucleic Acids Res,* 20, 1973-7.

Clark, G.R., Pytel, P.D., Squire, C.J. & Neidle, S. (2003) Structure of the first parallel DNA quadruplex-drug complex. *J Am Chem Soc,* 125, 4066-7.

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. & Kollman, P.A. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.,* 117, 5179-5197.

Crnugelj, M., Hud, N.V. & Plavec, J. (2002) The solution structure of d(G(4)T(4)G(3))(2): a bimolecular G-quadruplex with a novel fold. *J Mol Biol,* 320, 911-24.

Dai, J., Carver, M., Punchihewa, C., Jones, R.A. & Yang, D. (2007) Structure of the Hybrid-2 type intramolecular human telomeric G-quadruplex in K+ solution: insights into structure polymorphism of the human telomeric sequence. *Nucleic Acids Res,* 35, 4927-40.

Enokizono, Y., Konishi, Y., Nagata, K., Ouhashi, K., Uesugi, S., Ishikawa, F. & Katahira, M. (2005) Structure of hnRNP D complexed with single-stranded telomere DNA and unfolding of the quadruplex by heterogeneous nuclear ribonucleoprotein D. *J Biol Chem,* 280, 18862-70.

Erlitzki, R. & Fry, M. (1997) Sequence-specific binding protein of single-stranded and unimolecular quadruplex telomeric DNA from rat hepatocytes. *J Biol Chem,* 272, 15881-90.

Fadrna, E., Spackova, N., Stefl, R., Koca, J., Cheatham, T.E., 3rd & Sponer, J. (2004) Molecular dynamics simulations of Guanine quadruplex loops: advances and force field limitations. *Biophys J,* 87, 227-42.

Fan, J.H., Bochkareva, E., Bochkarev, A. & Gray, D.M. (2009) Circular dichroism spectra and electrophoretic mobility shift assays show that human replication protein A binds and melts intramolecular G-quadruplex structures. *Biochemistry,* 48, 1099-111.

Fang, G. & Cech, T.R. (1993) The beta subunit of Oxytricha telomere-binding protein promotes G-quartet formation by telomeric DNA. *Cell,* 74, 875-85.

Fukuda, H., Katahira, M., Tsuchiya, N., Enokizono, Y., Sugimura, T., Nagao, M. & Nakagama, H. (2002) Unfolding of quadruplex structure in the G-rich strand of the minisatellite repeat by the binding protein UP1. *Proc Natl Acad Sci U S A,* 99, 12685-90.

Gavathiotis, E., Heald, R.A., Stevens, M.F. & Searle, M.S. (2003) Drug recognition and stabilisation of the parallel-stranded DNA quadruplex d(TTAGGGT)4 containing the human telomeric repeat. *J Mol Biol,* 334, 25-36.

Haider, S.M., Parkinson, G.N. & Neidle, S. (2003) Structure of a G-quadruplex-ligand complex. *J Mol Biol,* 326, 117-25.

Haider, S.M., Parkinson, G.N. & Neidle, S. (2008) Molecular Dynamics and Principal Components Analysis of Human Telomeric Quadruplex Multimers. *Biophys J.*

Hardin, C.C., Corregan, M.J., Lieberman, D.V. & Brown, B.A., 2nd (1997) Allosteric interactions between DNA strands and monovalent cations in DNA quadruplex assembly: thermodynamic evidence for three linked association pathways. *Biochemistry,* 36, 15428-50.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins,* 65, 712-25.

Horvath, M.P. & Schultz, S.C. (2001) DNA G-quartets in a 1.86 A resolution structure of an Oxytricha nova telomeric protein-DNA complex. *J Mol Biol,* 310, 367-77.

Hud, N.V. & Plavec, J. (2006) The Role of Cations in Determining Quadruplex Structure and Stability. IN NEIDLE, S. & BALASUBRAMANIAN, S. (Eds.) *Quadruplex Nucleic Acids.* London, Royal Society of Chemistry.

Joachimi, A., Benz, A. & Hartig, J.S. (2009) A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorg Med Chem,* 17, 6811-5.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.,* 79, 926.

Kang, C., Zhang, X., Ratliff, R., Moyzis, R. & Rich, A. (1992) Crystal structure of four-stranded Oxytricha telomeric DNA. *Nature,* 356, 126-31.

Keniry, M.A., Strahan, G.D., Owen, E.A. & Shafer, R.H. (1995) Solution structure of the Na+ form of the dimeric guanine quadruplex [d(G3T4G3)]2. *Eur J Biochem,* 233, 631-43.

Kettani, A., Basu, G., Gorin, A., Majumdar, A., Skripkin, E. & Patel, D.J. (2000) A two-stranded template-based approach to G.(C-A) triad formation: designing novel structural elements into an existing DNA framework. *J Mol Biol,* 301, 129-46.

Kettani, A., Bouaziz, S., Wang, W., Jones, R.A. & Patel, D.J. (1997) Bombyx mori single repeat telomeric DNA sequence forms a G-quadruplex capped by base triads. *Nat Struct Biol,* 4, 382-9.

Kettani, A., Kumar, R.A. & Patel, D.J. (1995) Solution structure of a DNA quadruplex containing the fragile X syndrome triplet repeat. *J Mol Biol,* 254, 638-56.

Kondo, J., Adachi, W., Umeda, S., Sunami, T. & Takenaka, A. (2004) Crystal structures of a DNA octaplex with I-motif of G-quartets and its splitting into two quadruplexes suggest a folding mechanism of eight tandem repeats. *Nucleic Acids Res,* 32, 2541-9.

Kuryavyi, V., Majumdar, A., Shallop, A., Chernichenko, N., Skripkin, E., Jones, R. & Patel, D.J. (2001) A double chain reversal loop and two diagonal loops define the architecture of a unimolecular DNA quadruplex containing a pair of stacked G(syn)-G(syn)-G(anti)-G(anti) tetrads flanked by a G-(T-T) Triad and a T-T-T triple. *J Mol Biol,* 310, 181-94.

Lim, K.W., Amrane, S., Bouaziz, S., Xu, W., Mu, Y., Patel, D.J., Luu, K.N. & Phan, A.T. (2009) Structure of the human telomere in K+ solution: a stable basket-type G-quadruplex with only two G-tetrad layers. *J Am Chem Soc,* 131, 4301-9.

Liu, H., Matsugami, A., Katahira, M. & Uesugi, S. (2002) A dimeric RNA quadruplex architecture comprised of two G:G(:A):G:G(:A) hexads, G:G:G:G tetrads and UUUU loops. *J Mol Biol,* 322, 955-70.

Liu, Z. & Gilbert, W. (1994) The yeast KEM1 gene encodes a nuclease specific for G4 tetraplex DNA: implication of in vivo functions for this novel DNA structure. *Cell,* 77, 1083-92.

Luu, K.N., Phan, A.T., Kuryavyi, V., Lacroix, L. & Patel, D.J. (2006) Structure of the human telomere in K+ solution: an intramolecular (3 + 1) G-quadruplex scaffold. *J Am Chem Soc,* 128, 9963-70.

Marathias, V.M. & Bolton, P.H. (2000) Structures of the potassium-saturated, 2:1, and intermediate, 1:1, forms of a quadruplex DNA. *Nucleic Acids Res,* 28, 1969-77.

Marathias, V.M., Wang, K.Y., Kumar, S., Pham, T.Q., Swaminathan, S. & Bolton, P.H. (1996) Determination of the number and location of the manganese binding sites of DNA quadruplexes in solution by EPR and NMR in the presence and absence of thrombin. *J Mol Biol,* 260, 378-94.

Martadinata, H. & Phan, A.T. (2009) Structure of propeller-type parallel-stranded RNA G-quadruplexes, formed by human telomeric RNA sequences in K+ solution. *J Am Chem Soc,* 131, 2570-8.

Matsugami, A., Okuizumi, T., Uesugi, S. & Katahira, M. (2003) Intramolecular higher order packing of parallel quadruplexes comprising a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad of GGA triplet repeat DNA. *J Biol Chem,* 278, 28147-53.

Matsugami, A., Ouhashi, K., Kanagawa, M., Liu, H., Kanagawa, S., Uesugi, S. & Katahira, M. (2001) An intramolecular quadruplex of (GGA)(4) triplet repeat DNA with a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad, and its dimeric interaction. *J Mol Biol,* 313, 255-69.

Menon, L. & Mihailescu, M.R. (2007) Interactions of the G quartet forming semaphorin 3F RNA with the RGG box domain of the fragile X protein family. *Nucleic Acids Res,* 35, 5379-92.

Okur, A., Wickstrom, L., Layten, M., Geney, R., Song, K., Hornak, V. & Simmerling, C. (2006) Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model. *Journal of Chemical Theory and Computation,* 2, 420-433.

Padmanabhan, K., Padmanabhan, K.P., Ferrara, J.D., Sadler, J.E. & Tulinsky, A. (1993) The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer. *J Biol Chem,* 268, 17651-4.

Pagano, B., Mattia, C.A., Cavallo, L., Uesugi, S., Giancola, C. & Fraternali, F. (2008) Stability and cations coordination of DNA and RNA 14-mer G-quadruplexes: a multiscale computational approach. *J Phys Chem B,* 112, 12115-23.

Paramasivam, M., Membrino, A., Cogoi, S., Fukuda, H., Nakagama, H. & Xodo, L.E. (2009) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription. *Nucleic Acids Res.*

Parkinson, G.N., Lee, M.P. & Neidle, S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature,* 417, 876-80.

Patel, P.K., Bhavesh, N.S. & Hosur, R.V. (2000) NMR observation of a novel C-tetrad in the structure of the SV40 repeat sequence GGGCGG. *Biochem Biophys Res Commun,* 270, 967-71.

Patel, P.K. & Hosur, R.V. (1999) NMR observation of T-tetrads in a parallel stranded DNA quadruplex formed by Saccharomyces cerevisiae telomere repeats. *Nucleic Acids Res,* 27, 2457-64.

Phan, A.T., Kuryavyi, V., Gaw, H.Y. & Patel, D.J. (2005) Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter. *Nat Chem Biol,* 1, 167-73.

Phan, A.T., Kuryavyi, V., Luu, K.N. & Patel, D.J. (2007) Structure of two intramolecular G-quadruplexes formed by natural human telomere sequences in K+ solution. *Nucleic Acids Res,* 35, 6517-25.

Phan, A.T., Modi, Y.S. & Patel, D.J. (2004) Propeller-type parallel-stranded G-quadruplexes in the human c-myc promoter. *J Am Chem Soc,* 126, 8710-6.

Ryckaert, J.P., Ciccotti, G. & Berendsen, H.J.C. (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics,* 23, 327-41.

Sacca, B., Lacroix, L. & Mergny, J.L. (2005) The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides. *Nucleic Acids Res,* 33, 1182-92.

Schultze, P., Macaya, R.F. & Feigon, J. (1994a) Three-dimensional solution structure of the thrombin-binding DNA aptamer d(GGTTGGTGTGGTTGG). *J Mol Biol,* 235, 1532-47.

Schultze, P., Smith, F.W. & Feigon, J. (1994b) Refined solution structure of the dimeric quadruplex formed from the Oxytricha telomeric oligonucleotide d(GGGGTTTTGGGG). *Structure,* 2, 221-33.

Seenisamy, J., Rezler, E.M., Powell, T.J., Tye, D., Gokhale, V., Joshi, C.S., Siddiqui-Jain, A. & Hurley, L.H. (2004) The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4. *J Am Chem Soc,* 126, 8702-9.

Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. & Hurley, L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A,* 99, 11593-8.

Smith, F.W. & Feigon, J. (1992) Quadruplex structure of Oxytricha telomeric DNA oligonucleotides. *Nature,* 356, 164-8.

Smith, F.W., Schultze, P. & Feigon, J. (1995) Solution structures of unimolecular quadruplexes formed by oligonucleotides containing Oxytricha telomere repeats. *Structure,* 3, 997-1008.

Spackova, N., Berger, I. & Sponer, J. (1999) Nanosecond Molecular Dynamics Simulations of Parallel and Antiparallel Guanine Quadruplex DNA Molecules. *J Am Chem Soc,* 121, 5519-5534.

Spackova, N., Cheatham, T.E., 3rd, Ryjacek, F., Lankas, F., Van Meervelt, L., Hobza, P. & Sponer, J. (2003) Molecular dynamics simulations and thermodynamics analysis of DNA-drug complexes. Minor groove binding between 4',6-diamidino-2-phenylindole and DNA duplexes in solution. *J Am Chem Soc,* 125, 1759-69.

Sponer, J. & Spackova, N. (2007) Molecular dynamics simulations and their application to four-stranded DNA. *Methods,* 43, 278-90.

Stefl, R., Cheatham, T.E., 3rd, Spackova, N., Fadrna, E., Berger, I., Koca, J. & Sponer, J. (2003) Formation pathways of a guanine-quadruplex DNA revealed by molecular dynamics and thermodynamic analysis of the substates. *Biophys J,* 85, 1787-804.

Sun, H., Karow, J.K., Hickson, I.D. & Maizels, N. (1998) The Bloom's syndrome helicase unwinds G4 DNA. *J Biol Chem,* 273, 27587-92.

Thorpe, J.H., Teixeira, S.C., Gale, B.C. & Cardin, C.J. (2003) Crystal structure of the complementary quadruplex formed by d(GCATGCT) at atomic resolution. *Nucleic Acids Res,* 31, 844-9.

Walsh, K. & Gualberto, A. (1992) MyoD binds to the guanine tetrad nucleic acid structure. *J Biol Chem,* 267, 13714-8.

Wang, J., Cieplak, P. & Kollman, P. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem,* 21, 1049.

Wang, Y. & Patel, D.J. (1993a) Solution structure of a parallel-stranded G-quadruplex DNA. *J Mol Biol,* 234, 1171-83.

Wang, Y. & Patel, D.J. (1993b) Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. *Structure,* 1, 263-82.

Webba Da Silva, M. (2003) Association of DNA quadruplexes through G:C:G:C tetrads. Solution structure of d(GCGGTGGAT). *Biochemistry,* 42, 14356-65.

Weisman-Shomer, P. & Fry, M. (1994) Stabilization of tetrahelical DNA by the quadruplex DNA binding protein QUAD. *Biochem Biophys Res Commun,* 205, 305-11.

Zhang, D.H., Fujimoto, T., Saxena, S., Yu, H.Q., Miyoshi, D. & Sugimoto, N. (2010) Monomorphic RNA G-quadruplex and polymorphic DNA G-quadruplex structures responding to cellular environmental factors. *Biochemistry,* 49, 4554-63.

Zhang, N., Gorin, A., Majumdar, A., Kettani, A., Chernichenko, N., Skripkin, E. & Patel, D.J. (2001a) Dimeric DNA quadruplex containing major groove-aligned A-T-A-T and G-C-G-C tetrads stabilized by inter-subunit Watson-Crick A-T and G-C pairs. *J Mol Biol,* 312, 1073-88.

Zhang, N., Gorin, A., Majumdar, A., Kettani, A., Chernichenko, N., Skripkin, E. & Patel, D.J. (2001b) V-shaped scaffold: a new architectural motif identified in an A x (G x G x G x G) pentad-containing dimeric DNA quadruplex involving stacked G(anti) x G(anti) x G(anti) x G(syn) tetrads. *J Mol Biol,* 311, 1063-79.

Zhang, N., Phan, A.T. & Patel, D.J. (2005) (3 + 1) Assembly of three human telomeric repeats into an asymmetric dimeric G-quadruplex. *J Am Chem Soc,* 127, 17277-85.

CHAPTER 6

RGG BOX PEPTIDES DOCKED TO AN RNA G-QUADRUPLEX - MOLECULAR DYNAMICS SIMULATIONS

6.1 INTRODUCTION

6.1.1 Background

The last decade has seen a revolution in the understanding of RNA, its roles and functions. The historical view of RNA as a somewhat passive 'middle man' between DNA and protein has been supplanted by a growing appreciation of the dynamic regulatory role played by RNA in all cellular processes. The discovery of non-protein-coding RNAs, namely microRNAs (Lau et al., 2001, Ambros, 2004), small interfering RNAs (Mello, 2007) and long non-coding RNAs [for review see (Wilusz et al., 2009)] has complemented our existing knowledge of messenger, transfer and ribosomal RNA (mRNA, tRNA, rRNA) and caused an explosion of interest in this area.

Whilst RNA may interact directly with other nucleic acids, its biogenesis and many functions are intimately tied to its interaction with proteins (Dreyfuss et al., 2002, Carlson et al., 2003). Accordingly, over recent years researchers have set out to examine the ways in which proteins interact with RNA and what features confer specificity to these interactions (De Guzman et al., 1998, Draper, 1999). As the amount of available structural information has increased it has also been possible to carry out statistical studies on the binding interface of nucleic acid-protein complexes. This has exposed some of the differences between RNA-protein and DNA-protein binding, as well as deciphering some of the features that provide specificity to the particular RNA-protein interactions (Allers and Shamoo, 2001, Treger and Westhof, 2001, Morozova et al., 2006, Ellis et al., 2007).

An extensive range of RNA-protein binding modes have been identified, revealing that a common characteristic of RNA–peptide interactions is mutual adaptation at the binding interface, a process also known as 'mutually induced fit' (Williamson, 2000). Some

generalizations can also be made for particular classes of RNA. For example, proteins which bind single stranded RNA, including those containing the RNA recognition motif (RRM), have a preponderance of hydrogen bonds with the RNA bases. Ribosomal proteins, on the other hand, have a majority of contacts with the sugar-phosphate backbone of RNA (Allers and Shamoo, 2001, Ellis et al., 2007).

The RGG box proteins are one class of RNA binding proteins. By definition they contain an Arg rich binding motif (RGG box), and possibly other RNA binding motifs, as discussed in Chapter 2 (Table 2-1). At least in some such proteins, for example, nucleolin, the RGG box contributes to binding affinity but not specificity; in nucleolin the RGG box binds non-specifically to the RNA target and destabilizes it by unstacking nucleic acid bases thus making the RNA accessible for specific binding to a second RNA binding domain (Ghisolfi et al., 1992). On the other hand, the RGG box of some proteins, including the RGG box of the Fragile X Mental Retardation Protein (FMRP) does bind specifically to certain RNA quadruplexes (Darnell et al., 2001, Schaeffer et al., 2001, Ramos et al., 2003, Zanotti et al., 2006, Menon and Mihailescu, 2007). In this context 'specific binding' means that the protein recognizes a particular sequence of the RNA and has a particular affinity for that sequence.

Although little is known about the exact mode of binding of RGG box proteins to G-quadruplex RNA, studies of Arg methylation in RGG domains indicate that the state of the Arg residues is critical in determining their functionality (Boisvert et al., 2005, Dolzhanskaya et al., 2006, Lukasiewicz et al., 2007). This implies that the guanido side chains of Arg are responsible for forming electrostatic interactions or hydrogen bonds with binding partners. Unfortunately, this observation does not significantly narrow the problem of determining where an RGG box peptide might bind to quadruplex RNA. As shown in Figure 6-1, the G-quadruplex presents a shell of phosphates, surrounding the stacked guanine bases, any of which could interact non-specifically with positively charged Arg residues of the peptide, although tighter binding might be facilitated by other factors such as the co-ordinated binding of more than one Arg residue and non-polar interactions.

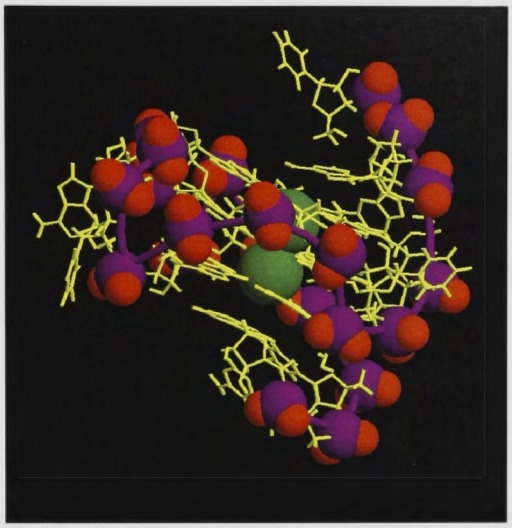


Figure 6-1. RNA quadruplex with phosphate groups of the backbone highlighted
Model RNA quadruplex used in docking simulations. The phosphate groups (phosphorous magenta, oxygen red) in space-fill representation are highlighted.

As discussed in Chapter 5, the G-quadruplex structure may be either stabilized or destabilized through interaction with different proteins. Such interactions have functional relevance in processes such as transcription and telomere maintenance. It has been suggested that the Fragile X RGG box may stabilize or destabilize RNA depending upon concentration and the particular RNA target (Menon and Mihailescu, 2007). The heterogeneity of the G-quadruplex structure, the range of binding proteins and the different outcomes of the binding process point to a wide range of protein binding modes, including preferential binding to the relatively constrained regions of the G-quadruplex (Wen and Gray, 2002) versus preferential binding to the long tails of the G-quadruplex (Oliver et al., 2000).

This chapter involves a systematic examination of the interaction of the Sho and Fx peptides with a model RNA G-quadruplex. MD simulations are used to dock the peptide to the RNA in different orientations and then a subset of the docked structures is simulated to observe the development of the binding interface. Dynamic simulation of the peptide and the RNA allows both species to adapt to each other, which is appropriate and necessary for interactions characteristically described as 'mutually induced fit'. Although the simulations conducted in this study are not long enough to capture a process such as quadruplex unfolding, this study is expected to provide some

insight into the type of interactions that may stabilize or destabilize the RNA quadruplex.

6.1.2 Aims of this chapter

The aims of the docking studies reported in this chapter and chapters 7 and 8 are: to explore ways in which the RGG box domain of 2 different peptides (the Sho and Fx peptides) may bind a model RNA G-quadruplex; to compare modes of binding of the Sho and Fx peptides with the G-quadruplex; and to explore the range of dynamical adaptivity at the binding interface and eludicate its functional relevance.

6.2 METHODS

6.2.1 Creating the RNA G-quadruplex

Before beginning any simulations, it is necessary to construct a suitable model RNA quadruplex structure. Turning to experiment, Mihailescu and co-workers (Zanotti et al., 2006, Menon and Mihailescu, 2007) note that the RNA transcripts used in their experiments appear to form an intramolecular parallel G-quadruplex similar to the parallel DNA G-quadruplex that is formed in the promoter region of the c-MYC oncogene and which has been characterized by NMR studies. In the absence of a published structure of a parallel intramolecular RNA quadruplex, the RNA analogue of the c-MYC G-quadruplex was used as a model RNA quadruplex in these simulations, as described in Chapter 5. Although the RNA analogue of a DNA quadruplex will not necessarily form the same quadruplex structure, in this instance, both MD simulations (Chapter 5) and circular dichroism experiments (Chapter 9) confirm that the RNA analogue does form a stable intramolecular parallel G-quadruplex.

The solution structure of the modified c-MYC oncogene is available in the Brookhaven Protein Data Bank (PDB) 1XAV (Ambrus et al., 2005). The 22 nucleotide sequence is set out in Fig. 6-2 together with its RNA analogue. NMR studies of the DNA quadruplex confirm that the glycosidic torsion angles of all guanines are in the *anti*-conformation,

thus forming a parallel-stranded G-quadruplex with side loops consisting of T, TA, T. This DNA quadruplex is very stable with a melting temperature over 85° C.

A

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
DNA	T	G	A	G	G	G	T	G	G	G	T	A	G	G	G	T	G	G	G	T	A	A
RNA	U	G	A	G	G	G	U	G	G	G	U	A	G	G	G	U	G	G	G	U	A	A

B

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Sho	K	G	G	**R**	**G**	**G**	A	R	G	S	A	**R**	**G**	**G**	V	**R**	**G**	**G**	A	R	G
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Fx	G	R	R	**R**	**G**	**G**	G	G	R	G	Q	G	G	R	G	**R**	**G**	**G**	G	F	K

Figure 6-2. G-quadruplex sequence and peptide sequences used in docking
(A)The c-MYC oncogene G-quadruplex sequence and its RNA analogue. The guanine tetrads are colour coded, tetrad 1 (orange), tetrad 2 (yellow), tetrad 3 (brown). (B) The Sho and Fx peptide amino acid sequences.

The PDB file of the DNA quadruplex was manually edited and then the LEaP program within AMBER 9 was used to convert this DNA structure into the RNA equivalent.

6.2.2 Docking Algorithm

Methods commonly used to explore conformational space, such as the replica-exchange algorithms employed during the peptide simulations (Chapter 4), generated instability in the G-quadruplex, as discussed in Chapter 5. A different approach was therefore required to explore potential interactions between the Sho and Fx peptides and the RNA quadruplex.

This involved the creation of many different starting models in which the peptide and RNA were oriented towards each other in a random manner. A python script was written to create the starting models, which manipulated the co-ordinates of the original RNA pdb file and pdb files of snapshots of the peptides in the following way:

1) Translate the RNA so that its first atom is located at the origin.
2) Translate the peptide by selecting one atom and placing it at the origin to ensure proximity to the RNA.
3) Translate the peptide away from the RNA by changing all x, y and z co-ordinates by a randomly selected number between 10 and 14 Å.

4) Create a new pdb file with the adjusted and merged co-ordinates of the RNA and peptide.

In order to sample different conformations of the peptides, conformations were selected randomly from the hybrid-REMD simulations of the Sho and Fx peptides (described in Chapter 4). Random selection could be expected to produce a subset which reflects the conformational distribution of the population. The model RNA G-quadruplex described in Chapter 5 was used as the RNA structure. In this way, 600 different starting structures of the RNA and the Sho peptide in close spatial proximity were created and 300 starting models of the Fx peptide and RNA were generated. The same peptide conformation was not used more than once.

6.2.3 Initial Docking Simulations

The 900 starting structures were solvated with a 10 Å radius octahedron of TIP3P water (Jorgensen and Tirado-Rives, 2005) and counter ions (Na^+) were used to create a charge neutral system using tLEaP. All simulations were performed with the AMBER 9 (Case et al., 2005) suite of programs. The ff99 force field (Wang et al., 2000) as modified by the Simmerling group 'the Stony Brook modification ff99SB (Hornak et al., 2006) was used with the RNA parameters in the AMBER file rna.ff99 for the RNA quadruplex.

All models were minimized over 1000 steps allowing only the waters to move, followed by a second minimization of 1000 steps without restraints. An equilibration simulation followed in which the temperature was raised from 0.1 K to 300 K over 20,000 steps. The simulation continued for a total of 600 ps under constant volume conditions. Harmonic restraints with force constants of 100 kcal mol^{-1} $Å^{-2}$ were used to keep the RNA and structural potassium ions fixed. Temperature scaling was performed using Langevin dynamics with a collision frequency of 2 ps^{-1} (Loncharich et al., 1992, Izaguirre et al., 2001). The simulations were carried out using a 2 fs time step and invoking the SHAKE algorithm to constrain bonds involving hydrogen (Ryckaert et al., 1977). Periodic boundary conditions were imposed. The particle-mesh Ewald (PME) procedure was used to calculate long-range electrostatic interactions, the direct space sum was limited with a cutoff of 10 Å which was also the cutoff for van der Waals

interactions.

6.2.4 Longer Simulations of Complexes

A subset of 49 Sho-RNA complexes and 52 Fx-RNA complexes were continued under similar conditions to those described except that simulations were run for 25 ns at constant pressure with a pressure relaxation time of 2 ps. The forces holding the RNA and structural potassium ions in place were relaxed to 10 kcal mol^{-1} $Å^{-2}$ for 1 ns, then to 1 kcal mol^{-1} $Å^{-2}$ for a further 1 ns, following which all restraints were removed for the remainder of the simulation. Simulations of subsets of 6 Sho-RNA and 6 Fx-RNA complexes were continued to 56 ns.

6.2.5 Analysis

Data were collected every 500 steps (1 ps). Trajectory analysis was carried out using the ptraj programs and custom written python scripts. Ptraj scripts (using the Average function) were used to produce average structures of the complexes from the last 5 ns of the 25 ns simulations. These average structures were used to identify the binding interface of the RNA and peptide by finding atoms in close contact, as discussed in section 6.3.3. Python scripts were written to calculate the distance between each of the non-hydrogen atoms of the RNA and each of the non-hydrogen atoms of the peptide in the average structures of all complexes. Ptraj was also used to calculate hydrogen bonds, RMSD values over the simulations, and K^+ - K^+ distances.

6.3 RESULTS AND DISCUSSION

6.3.1 Initial Docking Runs

The primary aim of the initial docking process was to obtain a subset of docked structures spatially distributed around the RNA that could be used as the starting point for longer MD simulations. Although there was a great deal of variability in the initial docking process, one example of the formation of a complex is shown in Figure 6-3. For clarity, the waters and counterions are omitted

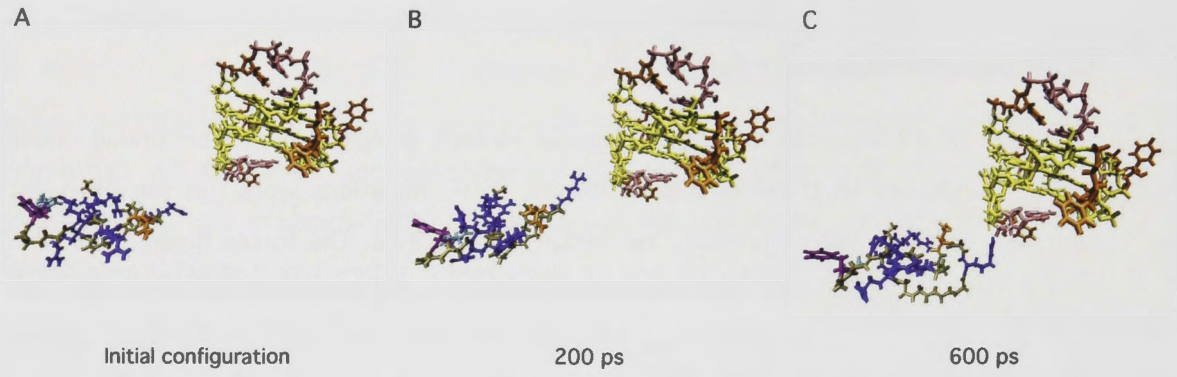


Figure 6-3 One example of the initial docking process in which the Fx peptide moves towards the RNA quadruplex and docks through an Arg residue
Three snapshots are shown of; (A) the original configuration of the Fx peptide and RNA quadruplex generated randomly as described in Methods, (B) snapshot taken after 200 ps which shows the sidechain of Arg9 extended out towards the RNA, (C) snapshot taken at 600 ps when Arg9 of Fx has made contact with the RNA.

The 600 ps simulations of the starting models were analyzed to determine the positioning of the peptide relative to the RNA quadruplex, based on the hydrogen bonds formed between the phosphate groups of the RNA and the Arg residues of the peptides. The phosphate groups were used as spatial markers as they form a shell around the quadruplex as can be seen in Figure 6-1. The hydrogen bonds formed between a particular phosphate of the RNA and each of the Arg residues of the RGG box were counted for all models and totals for the 600 Sho-RNA models are presented in Figure 6-4. As shown, each phosphate and each Arg residue were involved in several phosphate-Arg hydrogen bonds.

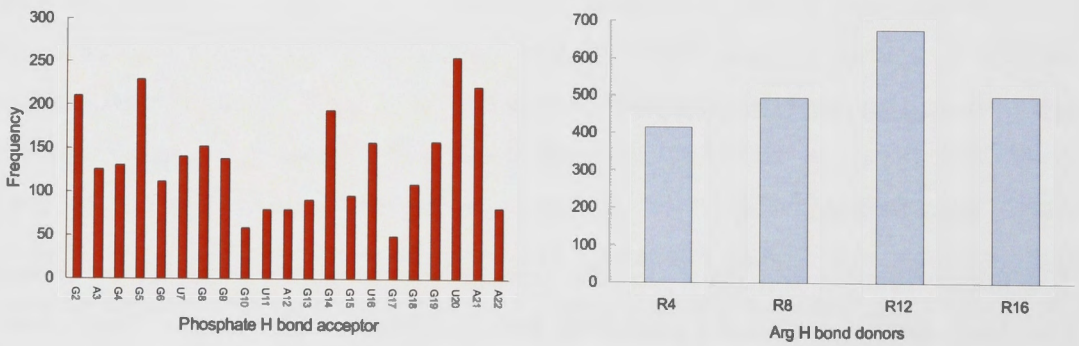


Figure 6-4. Phosphate-Arg H bonds formed in 600 docked Sho-RNA complexes
Number of hydrogen bonds between phosphate groups on the RNA backbone and arginine sidechains across all docked complexes sorted by (A) nucleotide number (RNA) and (B) residue number (peptide)

A subset of the starting models was selected for extended simulations, based on the condition that the subset was to contain an example of every possible phosphate-Arg contact. For example, the requirement that the first phosphate of the RNA interacts with each of the Arg residues of the RGG box of Sho would lead to the selection of 4 models to satisfy this condition. If there were no redundancy this process would lead to the selection of 88 (22 x 4) models. This approach is designed to ensure that, in the longer MD simulations, the peptide would have the opportunity to see all faces of the RNA. In addition, single trajectory MM-PBSA calculations were performed to check whether the complex formed from the random orientation of the peptide and the RNA was more or less stable than the monomers. Based on these considerations and, given a level of spatial redundancy, subsets consisting of 49 Sho-RNA and 52 Fx-RNA models were selected for longer MD simulations.

6.3.2 25 ns MD simulations of the complexes

The subsets of 49 Sho-RNA and 52 Fx-RNA models were simulated for 25 ns each (total simulation time 2.5 μs). The RMSD of the RNA G-quadruplex alone (GQ), the distance between the structural potassium ions (K^+ - K^+), the RMSD of the peptide (Sho or Fx) and the estimated free energy of the complex for each of the models is presented in Table 6-1. Overall, this data indicates that the structural K^+ ions remained stable in the RNA quadruplexes in the vast majority of the complexes. This contrasts with the hybrid-REMD simulations of the RNA and DNA quadruplexes discussed in Chapter 5 where the K^+ - K^+ distance increased markedly for 2 of the 8 replicas over the course of the 20 ns simulation. However, very high K^+ - K^+ distances indicative of quadruplex instability were found in 4 of the Fx-RNA complexes and one Sho-RNA complex (highlighted in red font). The instability in these models can be attributed to the original docking where the peptide was threaded through a part of the quadruplex (only later as a result of calculating the K^+-K^+ distances). This is illustrated in Figure 6-5. Whilst the starting point for these simulations seems physically unlikely it is interesting to note that some of these complexes have a relatively high estimated binding free energy, as discussed in more detail in Chapter 7.

Figure 6-5. Docking-induced quadruplex instability in a Fx-RNA complex
An example of a docked structure (xav1fx16) where the tail of the peptide (blue) threads through the tetrad stack of the quadruplex (yellow).

As reported in Chapter 5, the average RMSD of the RNA quadruplex in the hybrid-REMD simulations was 3.56 Å. Here, the RMSD of the G-quadruplex is in the range 1.63 – 4.7 Å with an average over all the Sho-RNA complexes of 2.7 ± 0.8 Å, and an average of 3.23 ± 0.6 Å for the Fx-RNA complexes. This suggests that, in most complexes, the peptide tends to stabilize the quadruplex. The average RMSD of the peptide calculated in the hybrid-REMD simulations (Chapter 4) was 9 Å for the Fx peptide and 9.7 Å for the Sho peptide. The RMSD of the peptides reported in Table 6-1 are generally lower than for the free peptides with an average of 5 ± 1.5 for Sho and 7 ± 1.5 for Fx. A decrease in the peptide RMSD is not surprising; the peptide would be expected to lose some of its flexibility upon binding to the RNA quadruplex.

Analysis of the average structures and estimated free energies of the complexes reveals that all complexes formed between the RNA quadruplex and the peptides are more stable than the monomers although the mode of binding varies greatly. Time averaged structures of the RNA-peptide complex derived from the last 5 ns of the 25ns simulations were calculated for each of the complexes and used to identify the contacts formed between the RNA and the peptide which constitute the binding interface. Two

such average structures are presented in Figure 6-6. Images of all the average structures are presented in Appendix 6A (Sho-RNA) and 6B (Fx-RNA).

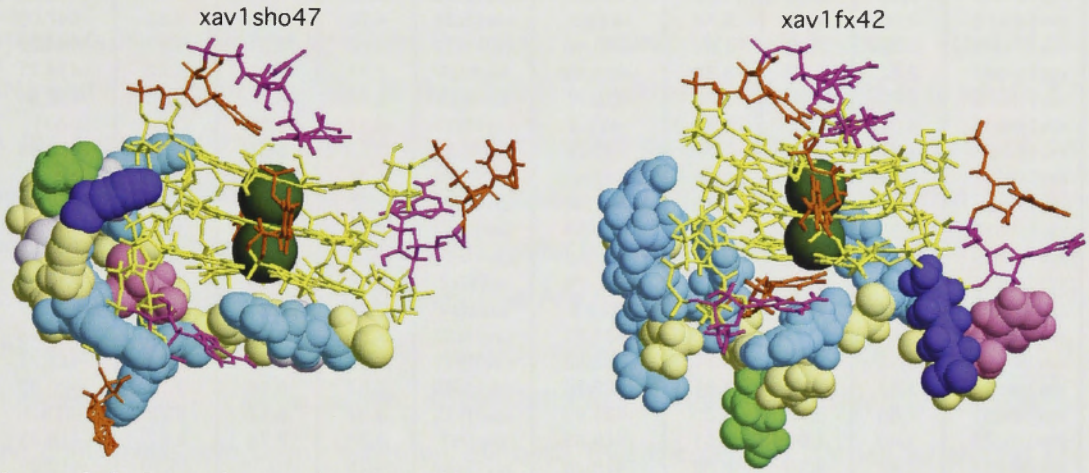


Figure 6-6. Examples of the Sho-RNA and Fx-RNA complexes
Average structures calculated from the last 5 ns of the 25 ns MD simulations. (A) Sho-RNA complex. (B) Fx-RNA complex. The RNA G-quadruplex is represented as a line drawing with colour-coded nucleobases, guanine (yellow), adenine (pink) and uracil (orange). The peptides are represented as VDW spheres, the amino acid residues are colour-coded, Arg (bright blue), Lys (dark blue), Gly (yellow), Ser/Gln (green), Val/Phe (pink), Ala (grey).

Table 6-1. RMSD, K-K distance and estimated free energy of Sho-RNA and Fx-RNA complexes [a]

Sho-RNA	RNA [b] (Å)	Sho [c] (Å)	K^+ - K^+ [d] (Å)	$\Delta H'_{bind}$ [e] (kcal/mol)	Fx-RNA	RNA (Å)	Fx (Å)	K^+ - K^+ (Å)	$\Delta H'_{bind}$ (kcal/mol)
xav1sho49	3.43	6.75	3.85	-454.07	xav1fx30	2.48	7.57	3.88	-450.29
xav1sho18	3.00	4.47	3.74	-443.9	xav1fx28	4.70	6.34	3.85	-447.09
xav1sho47	3.00	8.76	3.83	-440.49	xav1fx37	3.06	8.01	3.82	-444.97
xav1sho43	2.62	6.28	3.79	-439.99	xav1fx5	2.71	5.49	3.82	-434.71
xav1sho36	2.72	6.90	3.75	-438.71	xav1fx1	2.06	7.12	3.78	-432.94
xav1sho46	4.06	6.27	3.82	-435.4	xav1fx16	4.54	9.50	23.31	-431
xav1sho3	2.27	4.25	3.77	-435.22	xav1fx31	2.56	6.89	3.87	-425.99
xav1sho41	2.73	4.92	3.86	-432.45	xav1fx34	3.23	5.90	12.91	-425.26
xav1sho1	2.23	5.09	3.80	-431.47	xav1fx40	2.30	6.83	3.83	-425.14
xav1sho35	3.93	3.53	4.13	-429.34	xav1fx47	2.78	4.90	3.82	-424.8
xav1sho39	2.79	5.92	3.81	-427.13	xav1fx11	2.88	7.65	3.77	-424.24
xav1sho27	2.77	3.59	24.78	-426.79	xav1fx2	3.09	5.53	3.80	-422.66
xav1sho16	1.63	2.80	3.90	-425.8	xav1fx4	3.15	7.35	3.82	-422.51
xav1sho13	2.65	5.13	3.80	-424.21	xav1fx52	2.80	6.52	3.84	-422.46
xav1sho32	2.52	7.65	3.81	-422.82	xav1fx41	3.83	5.96	3.80	-421.71
xav1sho31	3.42	5.49	3.81	-422.18	xav1fx45	3.32	8.68	3.78	-420.92
xav1sho51	1.65	3.76	3.80	-421.98	xav1fx23	3.81	6.52	3.78	-419.9
xav1sho28	3.00	7.04	3.81	-419.65	xav1fx7	3.30	6.78	3.87	-419.83
xav1sho12	2.49	4.56	3.88	-419.28	xav1fx44	4.26	10.25	25.90	-419.08
xav1sho14	3.47	6.63	3.78	-415.55	xav1fx42	2.63	9.91	3.85	-418.52
xav1sho10	2.18	5.50	3.81	-414.92	xav1fx26	3.31	9.84	3.75	-416.19
xav1sho7	2.86	5.65	5.01	-414.32	xav1fx35	3.89	7.17	6.74	-415.69
xav1sho2	2.48	5.60	3.83	-412.28	xav1fx14	3.62	5.72	3.79	-414.76
xav1sho37	2.26	5.85	3.78	-412.2	xav1fx12	2.76	6.34	3.85	-413.49
xav1sho53	4.15	5.98	3.81	-411.99	xav1fx43	2.86	4.12	3.78	-411.55
xav1sho9	2.14	8.15	3.83	-411.13	xav1fx49	2.68	5.84	6.46	-410.32
xav1sho45	3.42	5.94	3.84	-410.2	xav1fx46	2.96	5.11	3.82	-409.63
xav1sho4	2.28	4.35	3.83	-410.19	xav1fx13	3.63	7.08	3.83	-408.93
xav1sho44	2.83	5.65	3.85	-409.81	xav1fx38	3.54	6.16	3.81	-407.89
xav1sho42	3.77	3.72	4.74	-409.07	xav1fx27	3.25	6.32	3.87	-406.64
xav1sho40	2.78	5.31	3.80	-407.97	xav1fx9	4.20	8.98	3.83	-406.63
xav1sho26	3.73	4.95	3.81	-407.41	xav1fx18	3.38	6.23	3.83	-406.59
xav1sho15	3.23	7.10	3.83	-406.9	xav1fx29	3.12	7.53	3.87	-406.36
xav1sho5	2.76	4.08	3.92	-406.33	xav1fx32	3.98	7.71	3.83	-403.88
xav1sho50	2.80	6.71	3.77	-405.73	xav1fx10	2.47	6.46	3.79	-403.02
xav1sho30	1.88	4.53	3.84	-404.9	xav1fx50	3.21	8.52	3.84	-402.96
xav1sho17	2.32	7.02	3.77	-401.59	xav1fx53	3.49	9.88	3.82	-399.4
xav1sho21	2.50	7.34	3.86	-398.89	xav1fx21	2.89	6.42	3.78	-399.26
xav1sho23	3.08	4.67	3.83	-398.81	xav1fx39	2.59	6.75	3.82	-398.82
xav1sho25	1.65	3.76	3.80	-397.92	xav1fx15	2.87	6.41	3.79	-398.39
xav1sho34	2.34	5.83	3.81	-397.59	xav1fx3	2.93	5.41	3.79	-398.13
xav1sho38	1.82	8.66	3.86	-396.18	xav1fx24	3.69	4.52	3.79	-397.71
xav1sho8	4.19	6.16	3.82	-396.01	xav1fx48	4.45	8.68	5.02	-397.43
xav1sho52	4.23	5.78	3.79	-395.87	xav1fx36	3.66	7.04	3.80	-395.61
xav1sho33	2.38	3.64	3.72	-392.75	xav1fx22	2.99	5.92	5.51	-394.59
xav1sho29	2.37	6.48	3.82	-392.72	xav1fx19	2.78	7.14	3.81	-393.48
xav1sho22	1.97	5.55	3.89	-391.77	xav1fx17	3.45	7.33	3.99	-391.17
xav1sho24	1.97	5.55	3.89	-391.77	xav1fx25	3.77	4.40	3.75	-387.34
xav1sho19	2.76	5.06	3.77	-388.1	xav1fx6	1.75	5.21	3.78	-387.01
xav1sho6	2.17	9.22	3.82	-387.78	xav1fx8	3.23	7.03	3.86	-386.09
				-	xav1fx33	3.89	5.69	22.64	-370.79
					xav1fx20	3.28	8.12	3.91	-368.77

[a] The complexes are ordered from most to least stabilized on the basis of enthalpy of binding ($\Delta H'_{Bind}$, defined in Chapter 7). All RMSD values calculated from snapshots taken every 50 ps over the 25 ns simulation. Complexes affected by docking-induced instability are highlighted in red font. [b] RNA G-quadruplex RMSD, standard deviations (SDs) are presented in appendix 6C. [c] Peptide RMSD, SDs are presented in appendix 6C. [d] Average distance between K^+ structural ions within the RNA quadruplex SDsI are presented in appendix 6C. [e] Enthalpy of binding $\Delta H'_{Bind}$ calculated as described in Chapter 7.

6.3.3 Analysis of RNA-peptide contacts in average structures of the complexes

The average structure of each of the complexes (calculated as described in section 6.2.5) is used to examine the contacts made between the peptides and the RNA or, in other words, the binding interface. A contact is counted if a non-hydrogen atom of the RNA and a non-hydrogen atom of the peptide are within 5 Å of each other or within 3.5 Å for a close contact. Firstly, the total number of contacts made by each of the peptide residues with the RNA is presented, in Figure 6-7. Contacts within 5 Å are presented in panels A and B, while close contacts (within 3.5 Å) are presented in panels C and D. From these plots it appears that Arg makes significantly more contacts with the RNA than any other residue, as foreshadowed by the discussion in Chapter 4. However, there is a large amount of variability in the binding interfaces (by design, as the complexes have been chosen on the basis that the peptide is oriented around the different faces of the RNA) with some complexes having a very high number of contacts to a particular residue while others may have none.

To better understand whether the Arg count could be due to a very high number of contacts in a subset of complexes or whether this is a more general phenomenon seen across all the complexes, the binding interface was examined in an alternative way. The percentage of complexes that have a particular residue at the binding interface (at least one contact) is depicted in Figure 6-7(E) for the Sho-RNA complexes and Figure 6-7(F) for the Fx-RNA complexes. The same test was then applied with the requirement that the residue must make at least 3 contacts with the RNA (Figure 6-7(G) and 6-7(H)). These data show that each of the Arg residues contributes to the binding interface in at least 50% of the complexes, which is greater than that found for the other residues. Comparison of plots 6-7(E) and 6-7(F) with plots 6-7(G) and 6-7(H) shows that Arg tends to bind to the RNA through multiple contacts rather than via a single contact point. For the other residues there is a greater difference in these plots indicating fewer contacts.

However, it is also apparent that other residues are found in the binding interface in a significant percentage of the complexes. Overall, these data show that several residues are likely to be found in the binding interface, consistent with the peptide wrapping around the RNA.

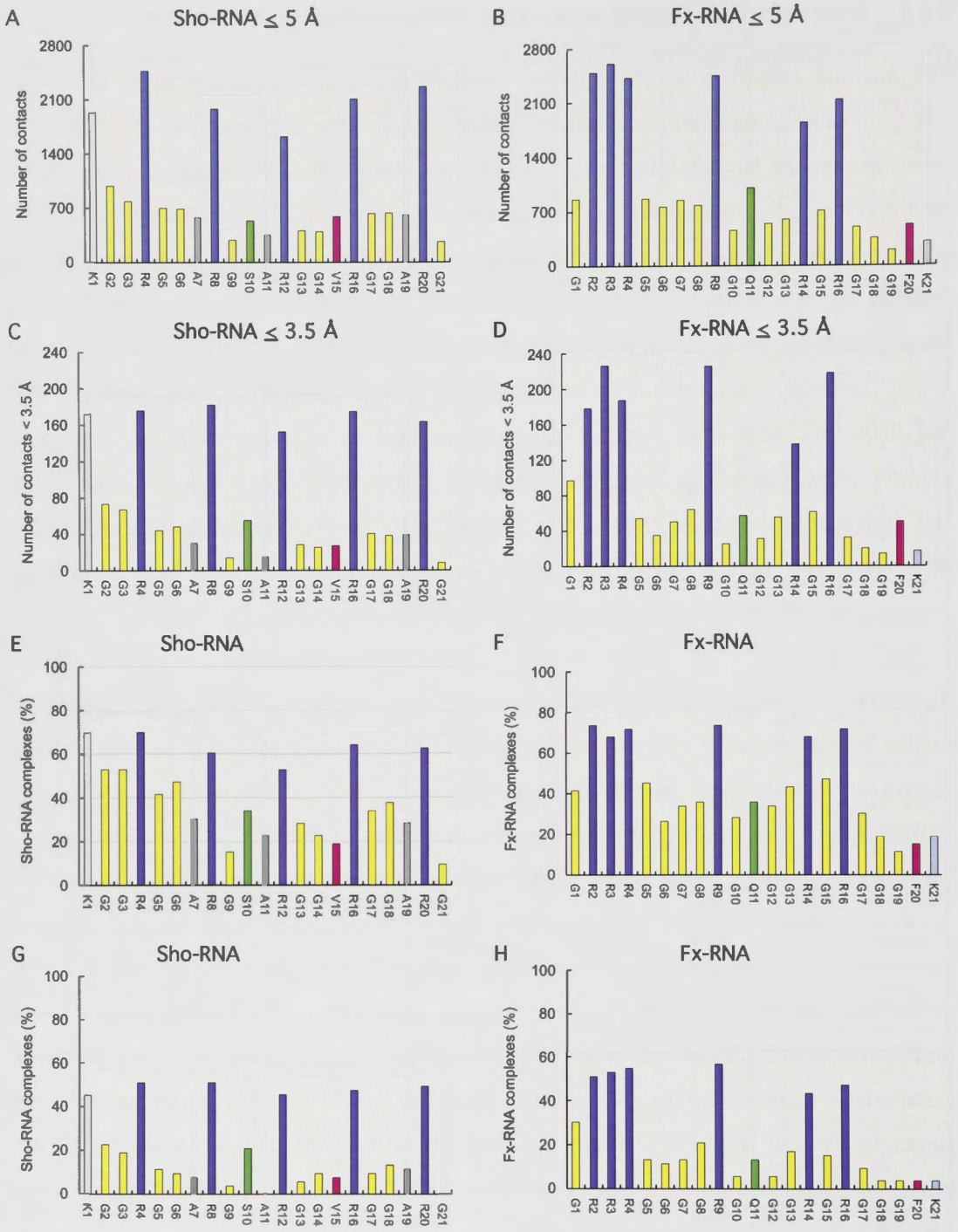


Figure 6-7. Frequency histograms of number of contacts made by each of the peptide residues and the RNA over the last 5 ns.
(A) Sho-RNA complexes with contacts ≤ 5 Å, (B) Fx-RNA complexes with contacts ≤ 5 Å, (C) Sho-RNA complexes with all contacts ≤ 3.5 Å, (D) Fx-RNA complexes with all contacts contacts ≤ 3.5 Å, (E) Percentage of Sho-RNA complexes with at least one atom contact between a given Sho residue and the RNA, (F) Fx equivalent of (E), (G) Percentage of Sho-RNA complexes that have at least 3 atom contacts between a given Sho residue and the RNA. (H) Fx equivalent of (G).

The average number of peptide atoms and residues at the binding interface (within 3.5 Å of the RNA) are presented in Table 6-2, showing that both Sho and Fx have a similar number of residues at the binding interface. In both peptides over 50% of the binding interface is comprised of Arg atoms although the percentage of Arg atoms is higher in Fx because it has an additional Arg residue.

Table 6-2. Average number of atoms and residues involved in the binding interface calculated from all Sho-RNA and Fx-RNA complexes listed in Table 6-1[a, b]

	# Atoms	SD	# Residues	SD	% Arg	SD	% Gly	SD
Sho-RNA	30	11	8.5	2.7	55	17	24	11
Fx-RNA	35	12	9	3	65	15	29	15

[a] Counts are based on contacts between the peptide and RNA within 3.5 Å. [b] All complexes are derived as the time averaged structure calculated from the last 5 ns of the 25 ns simulations.

Further analysis revealed that approximately 60% of Sho-RNA complexes are formed with more than one Arg residue binding to the RNA quadruplex. Three Arg residues are involved in the binding interface in approximately 45% of complexes. This reduces to around 30% for 4 Arg residues and 20% for all 5 Arg residues. These percentages are somewhat higher for the Fx-RNA complexes. This can be explained by the fact that Fx has 3 contiguous Arg residues with a greater chance of any 2 of these binding to the RNA target at the same time. Three Arg residues are found in the binding interface in approximately 60 % of the Fx-RNA complexes with 4 Arg residues being found in at least half the complexes. This number does not change greatly when considering the presence of 5 Arg residues in the binding interface as this is seen in 46% of complexes.

The presence of 5 Arg residues in the binding interface of half the Fx-RNA complexes and 20% of the Sho-RNA complexes validates the prediction that several Arg residues may participate in binding as the flexible RGG box peptide facilitates multiple Arg contacts.

Overall, the frequent presence of Arg in the binding interface, together with the likelihood that Arg will make several contacts with the RNA, shows that Arg residues contribute significantly to the stabilization of these complexes.

The peptide residues may contact the RNA through the backbone atoms (Cα, N, O, C) or through the sidechains. As shown in Figure 6-8, Arg, Lys and Phe primarily make contact with the RNA through the sidechains rather than the backbone. Gly lacks a side chain and so must make contact with the RNA through the backbone atoms. This is also the predominant mode of contact for the other residue with a small sidechain, Ala.

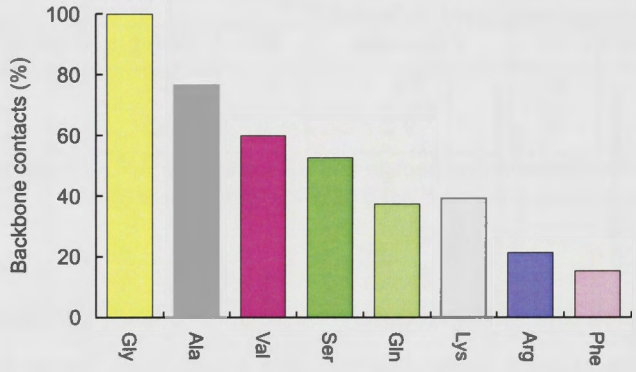


Figure 6-8. Percentage of contacts made by the peptide backbone
The percentage of contacts made by backbone atoms (Cα, N, C, O) for each residue type, combining the results of the Sho-RNA and Fx-RNA complexes

Some additional observations can be drawn from Figures 6-7 and 6-8. Firstly, the number of contacts made by an Arg residue is not greatly affected by the neighbouring residues, or whether it is part of an RGG repeat. It is also notable that Lys, which is found at the end of both peptides, is frequently involved in contacts in the Sho peptide but not in the Fx peptide. It is possible that the neighbouring Phe residue prevents Lys from exploring a full range of contacts in the Fx peptide. Although each peptide has a polar residue in the middle of the peptide, Gln, with a longer sidechain, makes more contacts than Ser. The 2 longer chain hydrophobic residues, Val in Sho and Phe in Fx, make a similar number of contacts.

Next, an analogous analysis is performed for the RNA component of the complex. The contacts established by each of the RNA nucleotides are counted to determine whether any are favoured in the binding interface (Figure 6-9).

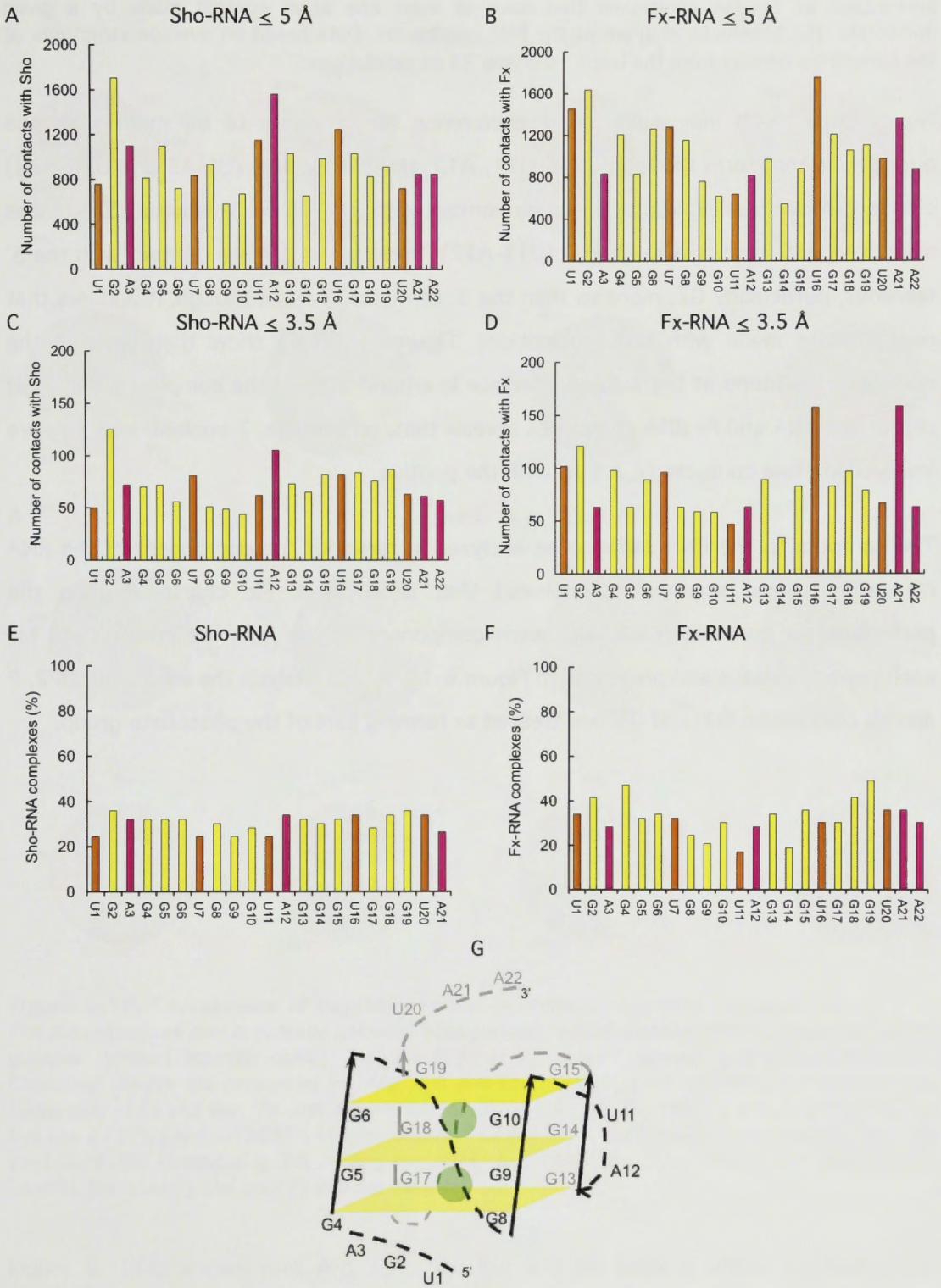


Figure 6-9. Contacts made by each of the RNA nucleotides and the peptides
Frequency histogram of the number of contacts made by each RNA nucleotide across all models (A) Sho-RNA complexes, contacts ≤ 5 Å. (B) Fx-RNA complexes, contacts ≤ 5 Å. (C) Sho-RNA complexes, contacts ≤ 3.5 Å. (D) Fx-RNA complexes, contacts ≤ 3.5 Å. (E) The percentage of

Sho-RNA complexes that have at least one atom contact made by a given nucleotide. (F) The percentage of Fx-RNA complexes that have at least one atom contact made by a given nucleotide. (G) Schematic diagram of the RNA quadruplex. Data based on average structures of the complexes derived from the last 5 ns of the 25 ns simulations.

Figures 6-9 (A-D) indicate a slight preference for contacts to be made with the nucleotides that form the loops (U7, U11, A12, U16) or termini (U1-A3 and U20-A22) of the RNA quadruplex. Fx makes more contact with U16 (Loop 3) whereas Sho makes more frequent contact with Loop 2 (U11-A12). Both peptides make contact with the 5' terminus, particularly G2, more so than the 3' terminus. Overall, though, it appears that contacts are made with RNA nucleotides. Figures 6-9(E-F) show that each of the nucleotides is found at the binding interface in around 30% of the complexes. Analysis of the Sho-RNA and Fx-RNA complexes reveals that, on average, 7 nucleotides ($\pm$ 2) are involved in close contacts ($\leq$3.5 Å) with the peptide.

The contacts to the RNA can also be analyzed in terms of the component of the RNA nucleotide (base, sugar or phosphate) that is involved. Pie charts showing the percentage of contacts made with each component of the RNA are constructed for each peptide residue and presented in Figure 6-10. In this analysis the atoms OP, OP2, P and by convention O3' and O5' are treated as forming part of the phosphate group.

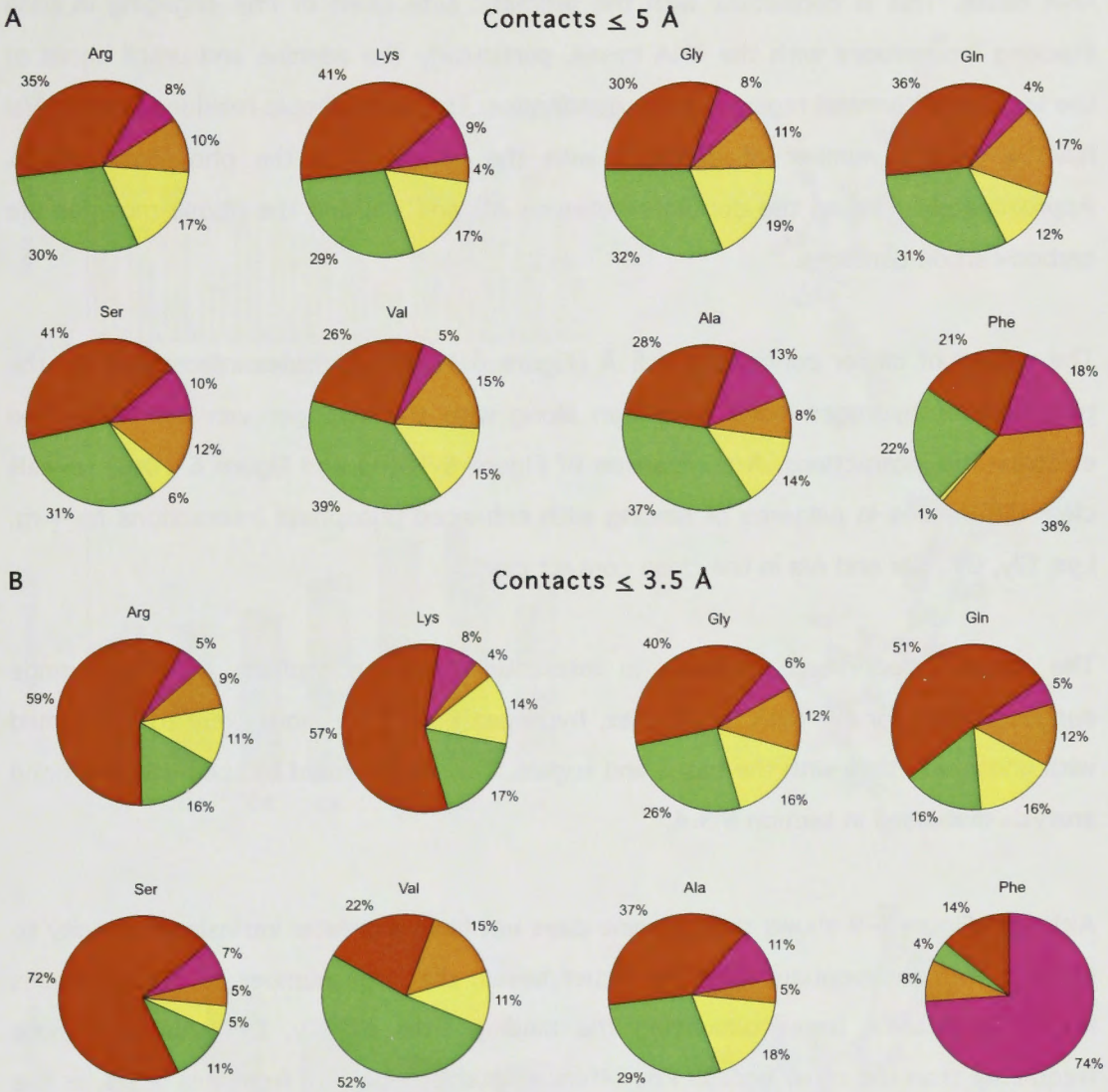


Figure 6-10. Comparison of peptide residue preference for RNA components
The percentage of contacts made between each peptide residue and the RNA components: bases, guanine (yellow) adenine (pink) and uracil (orange), sugars (green) and phosphate (red). Combined results are presented for Arg, Lys and Gly found in both peptides. Gln and Phe are found only in Fx and Ser, Val and Ala are found only in Sho. (A) Contacts ≤ 5 Å, Arg (n=24399), Lys (n= 2237), Gly (n=13246), Gln (n= 1007), Phe (n=521), Ser (n=537), Val (n=589), and Ala (n=1542). (B) Contacts ≤ 3.5 Å, Arg n= 2016, Lys (n=189), Gly (n=923), Gln (n=57), Phe (n=50), Ser (n=55), Val (n=27) and Ala (n=84).

Figure 6-10(A) shows that Arg, Lys, Gly, Ser and Gln have a similar distribution of contacts with the phosphate, sugar and bases of the RNA. Approximately one-third of the contacts are between the peptide and the RNA bases and two-thirds of the contacts are with the sugar phosphate backbone. Phe, which is found in the Fx peptide

but not in Sho, has a different profile in that its contacts are predominantly with the RNA bases. This is consistent with the aromatic side chain of Phe engaging in base stacking interactions with the RNA bases, particularly the adenine and uracil bases of the loops and terminal regions of the quadruplex. The hydrophobic residues Ala and Val have a greater number of contacts with the sugars than the phosphate atoms. Approximately 43% of the contacts between Ala and Val and the ribose moieties are carbon-carbon contacts.

The subset of closer contacts $\leq$ 3.5 Å (Figure 6-10(B)), provides information on the potential for hydrogen bond formation along with the stronger van der Waals and electrostatic interactions. A comparison of Figure 6-10(A) and Figure 6-10(B) reveals clear differences in patterns of binding with enhanced phosphate interactions for Arg, Lys, Gly, Gln, Ser and Ala in the close contact case.

The overall proportional increase in interactions with phosphate at closer range suggests that, for the RNA quadruplex, hydrogen bonds are more likely to be formed with phosphate than with the bases and sugars. This is confirmed by the hydrogen bond analysis discussed in section 6.3.4.

Although Figure 6-9 shows that guanine does not have a greater intrinsic propensity to interact with the peptides than the other bases, the large number of guanine bases results in guanine bases occupying the binding sites on Gly, Lys, Ala, Arg more frequently than the other bases. Phe offers a notable departure from this trend, as the closer contacts are predominantly with uracil or adenine. While Gly has a preference for guanine, Ser has a slight preference for adenine, results which are consistent with other studies (Treger and Westhof, 2001).

We can also investigate whether particular atoms constituting the RNA components are more likely to be involved in binding. The contribution of each atom type to the contacts ($\leq$ 5 Å) presented above are shown in Figure 6-11.

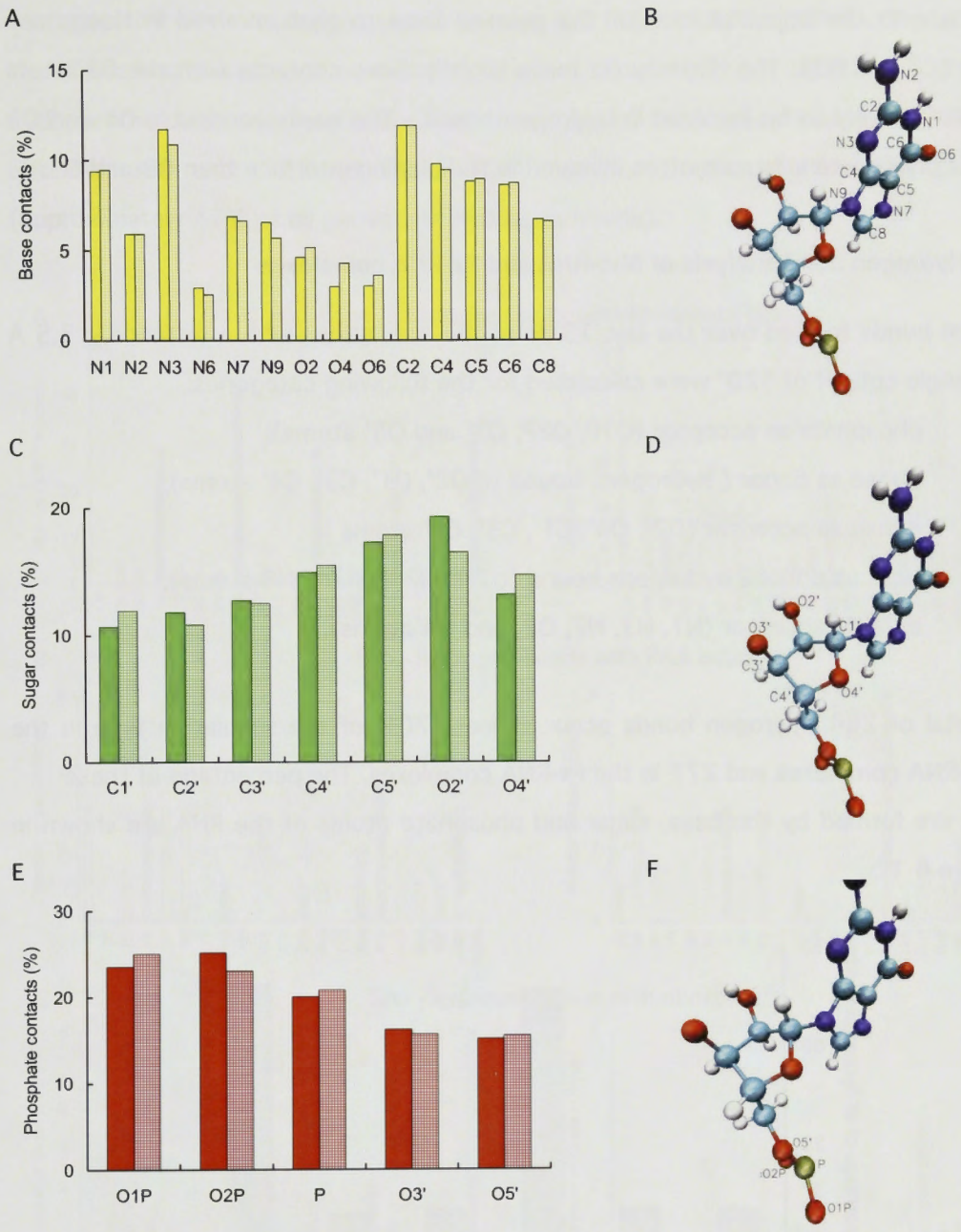


Figure 6-11. Decomposition of RNA component-peptide contacts by atom
Breakdown of all base-peptide, sugar-peptide, and phosphate-peptide contacts ≤ 5 Å in terms of percentage contributions from each non-hydrogen atom. Sho data are indicated by a solid coloured bar and Fx data are indicated by a patterned bar. (A) Contacts made by atoms involved in an RNA base (B) G6 nucleotide with base atoms labeled, clearly, not all bases have the same atom types. (C) Contacts made by atoms involved in a ribose unit. (D) G6 nucleotide with ribose atoms labeled. (E) Contacts made by atoms involved in a phosphate group. (F) G6 nucleotide with phosphate atoms labelled.

Fx and Sho make similar types of contacts with the RNA components. Most contacts are made with the opposite face of the guanine base to that involved in Hoogsteen bonding (C2 and N3). The ribose units make slightly more contacts with the O2' atom which is also likely to be involved in hydrogen bonds. The easily accessible O1 and O2 atoms of phosphate are more often involved in the binding interface than O3 and O5.

6.3.4 Hydrogen bond analysis of Sho-RNA and Fx-RNA complexes

Hydrogen bonds formed over the last 15 ns of the simulations with a distance $\leq$ 3.5 Å and an angle cut-off of 120° were calculated for the following categories:

(i) phosphate as acceptor (O1P, O2P, O3' and O5' atoms);

(ii) ribose as donor (hydrogens bound to O2', C1', C3', C4' atoms);

(iii) ribose as acceptor (O2', O4', C1', C3', C4' atoms);

(iv) base as donor (hydrogens bound to N1, N2, N3 and N6 atoms);

(v) base as acceptor (N1, N3, N7, O2, and O4 atoms).

A total of 264 hydrogen bonds occur at least 70% of the simulation time in the Sho-RNA complexes and 277 in the Fx-RNA complexes. The percentage of these that are formed by the base, sugar and phosphate atoms of the RNA are shown in Figure 6-12.

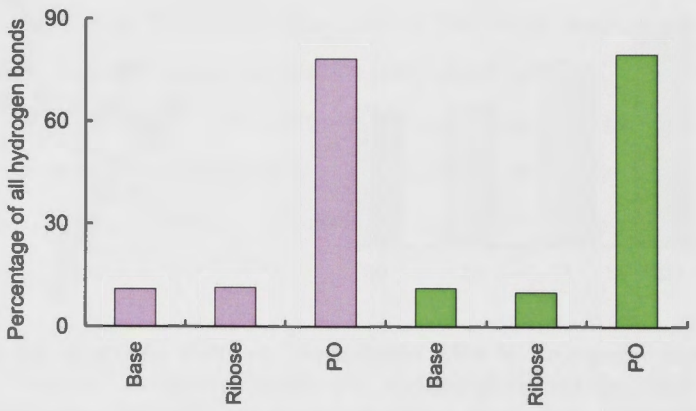


Figure 6-12. Hydrogen bonds formed by RNA components and the peptide
Classification of all hydrogen bonds formed at least 70% of the time during the last 15 ns of simulation according to the RNA component they originate from. The data are presented for all complexes of each system Fx-RNA (lavender) and Sho-RNA (green).

Figure 6-12 reveals that hydrogen bonds formed by the phosphate atoms make up the vast majority of the hydrogen bonds in both Sho (79%) and Fx (78%) complexes, with the ribose and bases each contributing about 11% of hydrogen bonds. The phosphate hydrogen bonds are formed either with the amide of the peptide backbone, 33% in the Sho-RNA set and, 32% in the Fx-RNA complexes or with the Arg sidechains (approximately 67% of all phosphate hydrogen bonds).

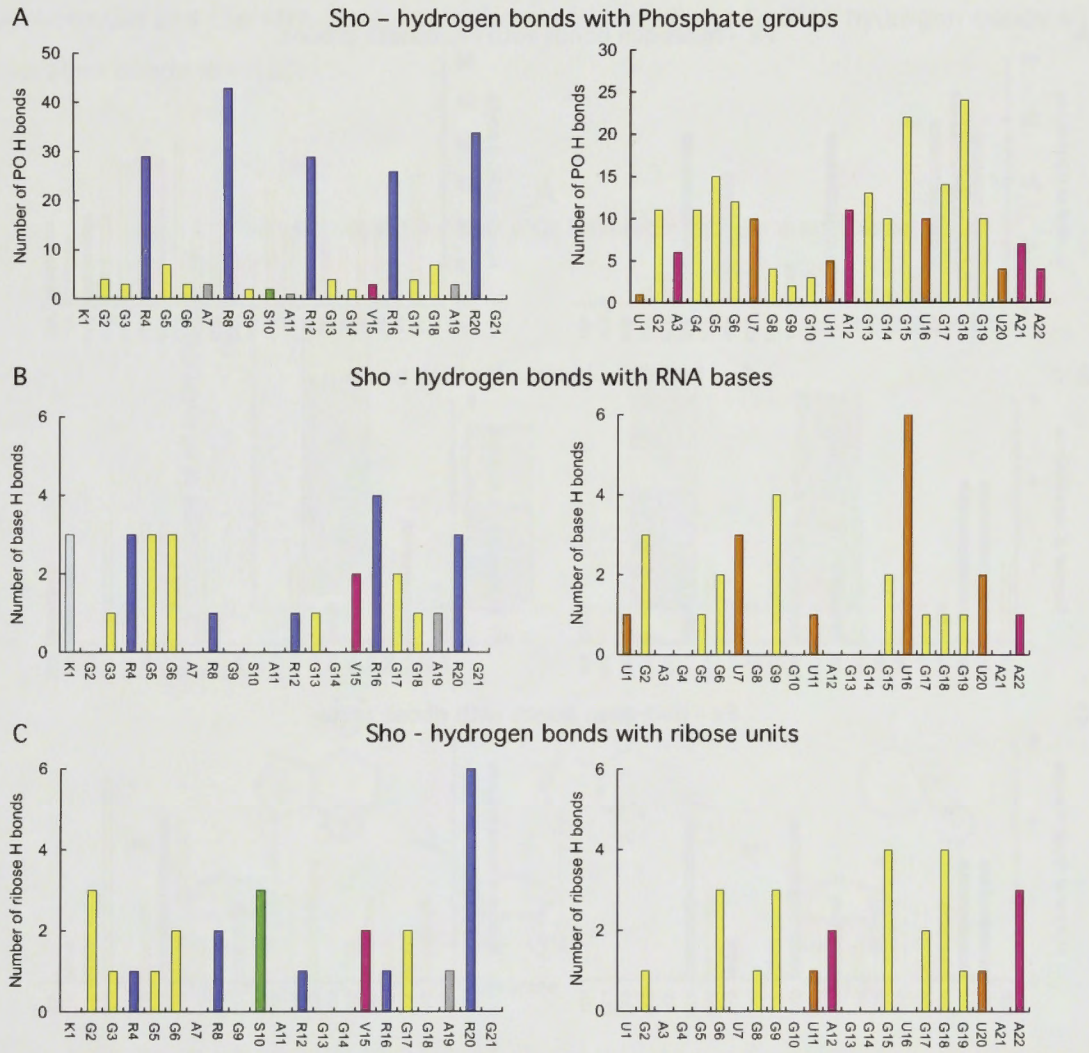


Figure 6-13. Number of hydrogen bonds involving each Sho residues and each RNA nucleotide

Hydrogen bonds occupied at least 70% of the last 15 ns of the 25 ns simulations are counted and the residue or nucleotide involved is shown. (A) Hydrogen bonds involving the phosphate atoms of the RNA. (B) Hydrogen bonds involving the bases of the RNA. (C) Hydrogen bonds involving the ribose units of the RNA.

As a whole, the Arg residues are involved in approximately 80% of phosphate hydrogen bonds (formed by the backbone or sidechain). This demonstrates that Arg participates in specific interactions with the RNA quadruplex as well as more general electrostatic interactions arising from the proximity of the oppositely charged groups of the Arg sidechain and the RNA phosphate groups. The peptide residues and RNA nucleotides involved in the 3 classes of hydrogen bonds are presented in Figures 6-13 and 6-14.

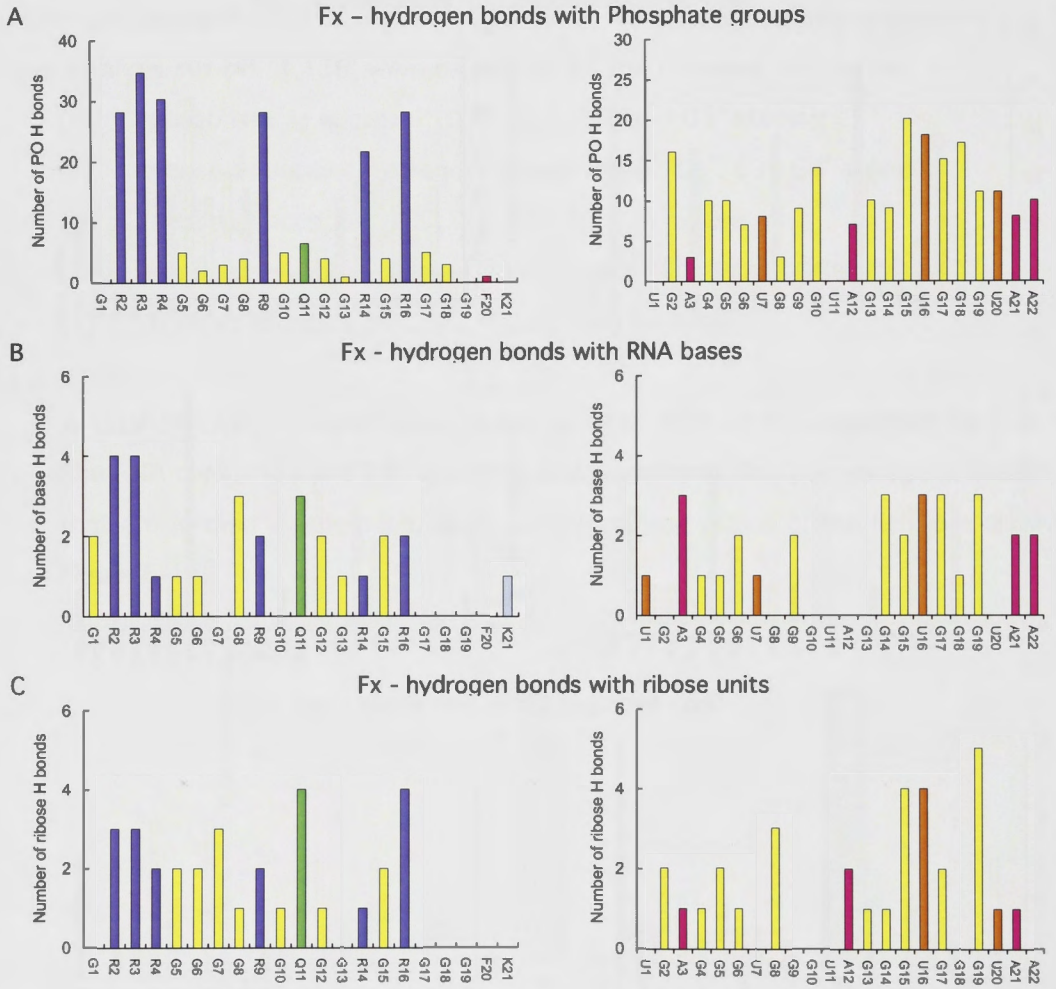


Figure 6-14. Hydrogen bonds formed by the residues of Fx and the RNA nucleotide
Hydrogen bonds occupied at least 70% of the last 15 ns of the 25 ns simulations are counted and the residue or nucleotide involved is shown. (A) Hydrogen bonds involving the phosphate atoms of the RNA. (B) Hydrogen bonds involving the bases of the RNA. (C) Hydrogen bonds involving the ribose units of the RNA.

Figures 6-13(A) and 6-14(A) demonstrate that Arg residues are strongly favoured in forming hydrogen bonds with phosphate, but that they have no strong preference for which phosphate group they bind to. Hydrogen bonds formed with the RNA bases present a different picture, as Arg is not particularly favoured and there is an increased presence of Gly in these bonds. The polar residues (Ser in Sho and Gln in Fx) appear to account for a significant proportion of hydrogen bonds to the ribose units. However, it must be borne in mind that this subset of the data is quite small e.g. hydrogen bonds between Gln and the RNA bases constitute ~ 3% of all the Fx-RNA hydrogen bonds (8 hydrogen bonds in total).

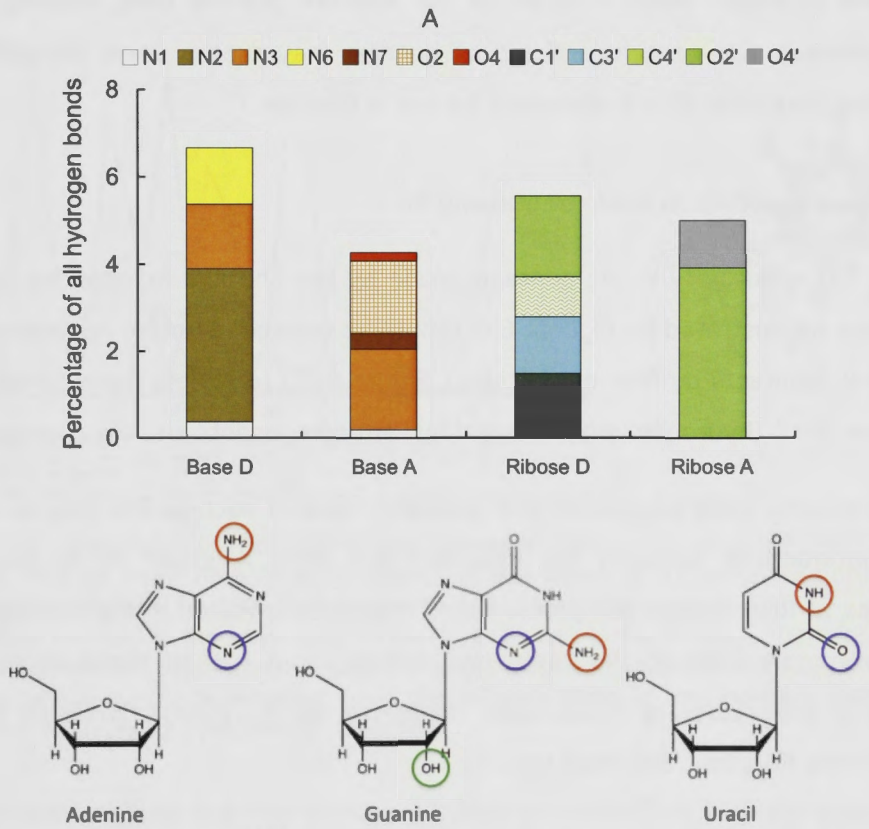


Figure 6-15. Hydrogen bonds formed by the bases and ribose of the RNA
(A) Frequency plot of RNA atoms involved in hydrogen bonds as base donor (Base D), base acceptor (Base A), ribose donor (Ribose D) and ribose acceptor (Ribose A), occupied at least 70% of the last 15 ns of the 25 ns simulations. The results for all Sho-RNA and Fx-RNA complexes are combined. (B) Adenine nucleoside with N3 circled bue and N6 circled red. (B) Guanine nucleotide with N3 circled blue and N2 circled red. (C) Uracil nucleoside with O2 circled blue and N3 circled red.

In both the Sho and Fx complexes, hydrogen bonds with the bases and ribose constitute only around 22% of all hydrogen bonds. Nevertheless, it is possible that these bonds are important in conferring specificity to the interaction (Morozova et al., 2006). It is therefore worthwhile looking more closely at the atoms involved in these hydrogen bonds. The frequency that each base and ribose hydrogen bond acceptor or donor are involved in hydrogen bonds are presented in Figure 6-15. When the base acts as hydrogen bond acceptor, it is usually through the N3 atom of the guanine or adenine base or the O2 of uracil (all circled blue in the diagram above). Donor interactions occur most frequently through the N2 of the guanine base, followed by the N3 of uracil and the N6 of adenine (all circled red). As the N2 of guanine plays a critical role in Hoogsteen hydrogen bonds with N7 of the adjacent guanine base, forming hydrogen bonds between the peptide and N2 of guanine may create some instability in the quadruplex structure. This is discussed further in Chapter 7.

6.3.5 Comparing contacts made by Sho and Fx

There is much similarity in the contacts made by the Sho and Fx peptides to the RNA quadruplex as illustrated by Figures 6-7 (contacts between peptide residues and RNA), Figure 6-9 (contacts by RNA nucleotides), Figure 6-10 (contacts made by RNA atoms) and Figure 6-12 (hydrogen bonds formed by RNA components and the peptide).

It has previously been suggested that aromatic residues such as Phe play an important role in interactions between the RGG box and RNA. A study of 41 experimental structures of RNA-protein complexes found Phe to be involved in the binding interface in approximately 80% of the complexes and in about half of these cases Phe was involved in base stacking interactions (Morozova et al., 2006). While Fx has a Phe residue in the RGG box, Sho does not.

Potential stacking interactions involving Phe and the RNA bases were investigated by calculating the distance between the centre of the aromatic ring of Phe and the centre of the 6 membered ring of each RNA base over the course of the simulation. The optimal center-to-center distance between the aromatic ring of an amino acid and nuclear base was calculated in the survey conducted by Allers and Shamoo as an average of 3.8 Å (SD 0.3 Å) (Allers and Shamoo, 2001).

The frequency distributions of the Phe-base distances were calculated for each RNA base. This analysis detected a distinct peak at distances < 5 Å in 12 Fx-RNA complexes indicating potential base stacking interactions. All of these involved interactions between Phe and the bases of the loops (U7, U11, A12, U16) and termini (U1, G2, A3 and U20, A21, A22). In the case of the guanine bases there were no distances < 5 Å. One plot of the distributions in distances between Phe and the bases of the loops and termini is illustrated in Figure 6-16. In this plot there is one distribution with a peak < 5 Å, this is for the distance between Phe and the base of U1.

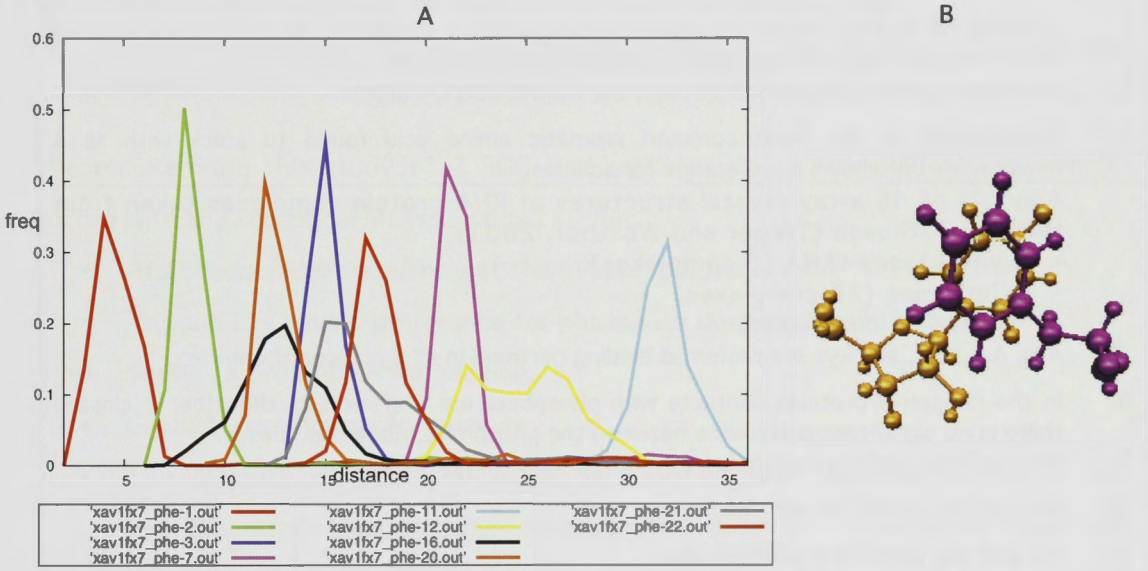


Figure 6-16. Distance (Å) distributions between aromatic ring of Phe and 6-membered ring of each RNA base involved in the loops or terminal regions of the RNA quadruplex in one Fx-RNA complex over the 25 ns simulation.
(A) RNA bases involved in the loops (U7, U11,A12, U16) and terminal regions (U1, G2, A3, and U20, A21, A22) in one Fx-RNA complex (xav1fx7). The distance distribution for Phe-U1 (red) has a peak around 3.5 Å. (B) A snapshot from the MD simulation of this complex showing the interaction between Phe (purple) and U1 (orange).

These data indicate that Phe may engage in stacking interactions with the bases of the loops and termini of the RNA quadruplex but not with the guanine bases involved in the tetrad stacks. It has previously been found that when Phe does engage in base stacking with guanine the stacking occurs above the N3, O6 and N7 atoms (Morozova et al., 2006). The participation of the guanine bases in the quadruplex formation would therefore impede stacking interactions with Phe. This accounts for the preponderance of contacts between Phe and adenine and uracil in the G-quadruplex.

6.3.5 Statistical studies of RNA-protein complexes from the PDB data base

The results obtained for the RGG box – RNA quadruplex models in this work are compared to statistical studies of RNA-protein interfaces in Table 6-3.

Table 6-3. Characteristics of RNA-protein interface obtained by statistical analysis of known structures

	Analysis of 45 RNA-protein complexes with co-ordinates taken from the PDB data base (Allers and Shamoo, 2001)
1	Distances between Arg/Lys and O1P or O2P indicate that 68 % of hydrogen bonds to the phosphate oxygens are made by Arg or Lys.
2	Hydrogen bonds to the Adenine base are mostly through N1 or N6 (73%).
3	Guanine commonly makes hydrogen bonds to Arg through the N7 and O6 atoms.
4	Cytosine O2 is most frequently used in hydrogen bond formation, the guanidium group of Arg is likely to make contact with O2 and N3 simultaneously.
5	Uracil recognition is mediated through the O2, N3 and O4 atoms.
6	Phenylalanine is the most common aromatic amino acid found to stack with all 4 nucleosides but shows a preference for adenine.
	Analysis of 45 x-ray crystal structures of RNA-protein complexes taken from the PDB database (Treger and Westhof, 2001) • **synthetases-tRNA (6 complexes)** • **ribosomes (22 complexes)** • **other (17 complexes)**
7	Arg, Asn, Ser, and Lys are preferred binding partners in all 3 classes of complex.
8	In the ribosomal proteins contacts with phosphate are preferred. In the other 2 classes there is no significant preference between the phosphate, ribose and bases.
9	Arg and Lys confer a preference for interaction with phosphate.
10	Ser, Ile, Pro prefer the adenine base.
11	Gly and Asp prefer the guanine base.
12	43% of all potential hydrogen bonds in the RNA complexes involve phosphate, compared with 60% in DNA complexes.
13	In RNA complexes ribose is involved in 20% of hydrogen bonds compared with 6% in DNA complexes, with the O2' atom being most frequently involved.
14	32% of all phosphate hydrogen bonds are made by the charged atoms of Arg and Lys.
15	32% of all hydrogen bonds are made by the protein mainchain and 68% by sidechains.
16	The O6 and N2 atoms of guanine are more often involved in hydrogen bonds and N3 of guanine is less involved in hydrogen bonds than expected.
17	O2' of ribose is more often involved and O4' is less often involved in hydrogen bonds than expected.
18	Arg (19.6% of atoms at the binding interface), Lys (14.2% of atoms at the binding interface) and Gly (6.6% of atoms at the binding interface) are more represented in RNA-protein interfaces than in protein-protein interfaces.

	Analysis of 89 RNA-protein structures taken from the PDB database, at least 5 members from each of the following classes: (Ellis et al., 2007) • ribosomal RNA • transfer RNA • viral RNA • messenger RNA • small RNAs
19	75% of hydrogen bonds involve the sugar-phosphate backbone (44% phosphate, 31% ribose)
20	65% of van der Waals interactions involve the sugar-phosphate backbone (25% phosphate, 41% ribose).
21	Protein sidechains are more likely to interact with the RNA. Sidechains comprise 72.1% of hydrogen bonds and 74.1% of van der Waals interactions.
22	Arg and Lys have a clear propensity to bind to the phosphate group of nucleotides.
23	Phe disfavours making van der Waals contacts to the phosphate group
24	The positively charged residues as well as Ser and Gly are most likely to be found at the binding interface.

The results from this study of 2 RGG box peptides and a model RNA G-quadruplex accord with the statistical analysis above in that:

- Arg occurs most frequently at the binding interface.
- Arg and Lys have a preference for phosphate over ribose and bases.
- Ser has a preference for adenine.
- Gly has a preference for guanine.
- The majority of hydrogen bonds are formed with sidechain atoms of the peptides rather than mainchain atoms (60%). However the proportion does vary depending upon the RNA component involved. When phosphate is involved 65% of hydrogen bonds are made by the sidechain, compared with only 24% when a base is involved and 53% when a ribose unit is involved.

However the findings in this study differ in the following:

- In the RGG box peptide, Arg monopolizes the hydrogen bonds made to phosphate. Here we see that approximately 80% of phosphate hydrogen bonds are formed by Arg, a significantly larger proportion than the 68% for both Arg and Lys contacts found by Allers and Shamoo (#1) and much larger than the 32% derived by Treger and Westhof (#14), again for both Lys and Arg. Two-thirds of these involve the Arg sidechain and one-third are with the backbone amide.
- Here, we find a greater proportion of hydrogen bonds are made to the sugar phosphate backbone (~89%), compared with 75% found by Ellis et al. (#19).

- When the base of the RNA G-quadruplex acts as hydrogen bond acceptor, it is usually through the N3 atom of the guanine or adenine base. When acting as donor, it is most often through N2 of the guanine base. This contrasts with the finding that O6(G) and N7(G) are more involved and N3(G) is less involved in hydrogen bonds than expected (#3, #16), but is readily explained by the tetrad structure in which Hoogsteen hydrogen bonding dominates the N7 and O6 atoms of guanine. Although N2(G) is also involved in Hoogsteen hydrogen bonding it still remains reasonably accessible as a potential hydrogen bond donor.
- Treger and Westhoff found that 26% of hydrogen bonds have the donor group on the RNA and acceptor on the protein, compared with 10% for DNA-protein complexes. Here the RNA acts as donor in approximately 12% of hydrogen bonds.
- The statistical studies report that ribose is more often involved in hydrogen bonds in RNA-protein complexes (Treger and Westhoff - 20%, Ellis et al., - 31%) compared with 6% in DNA. Here ribose is involved in only 12% of hydrogen bonds.
- Although Phe has been found to engage in base stacking interactions with all the nucleobases (#6), in the case of the RNA G-quadruplex it is most likely that stacking will occur with the bases that are not involved in the tetrad stack.

6.4 CONCLUSIONS

The calculations in this Chapter examined the binding of the RGG box motif to a model G-quadruplex. The Sho-RNA and Fx-RNA complexes bound via a large interaction surface. Approximately half of the peptide residues were involved in the binding interface, on average 30 atoms formed close contacts ($\leq$ 3.5 Å) with the RNA. Of these, at least 50% were Arg. The flexibility of the RGG box peptides allows several Arg residues to interact with the RNA through a range of contacts: electrostatic contacts, van der Waals interactions, and hydrogen bonds. The loop and terminal regions of the RNA quadruplex were only slightly favoured. Overall, it appears that the peptides wrap around or against the quadruplex, making contact with several nucleotides, primarily through the phosphate groups that are on the outside of the quadruplex structure.

Hydrogen bonds to the phosphate oxygens make up the vast majority of hydrogen bonds formed with the RGG box peptides.

The G-quadruplex structure generally remained very stable upon peptide binding. The average RMSD of the quadruplex was 2.7 + 0.8 Å in the Sho-RNA complexes and 3.2 + 0.6 in the Fx-RNA complexes, compared to an average of 3.6 Å in the hybrid-REMD simulations of the free RNA. Both peptides lost some flexibility upon binding as judged by a decrease in the RMSD from ~ 9 Å in the hybrid-REMD simulations of the free peptides to an average of ~ 6 Å in the Sho-RNA complexes and ~ 7 Å in the Fx-RNA complexes.

The G-quadruplex stack presents a very stable well-ordered array of guanine bases, which associate with each other through Hoogsteen hydrogen bonds involving the N2, N1, N7 and O6 atoms. As these atoms face towards the middle of the quadruplex structure they are less likely to be involved in hydrogen bonds with a binding partner than is generally seen in other forms of RNA. However the hydrogen atoms of N2 are more accessible than the other atoms involved in the Hoogsteen hydrogen bonds, so can participate in hydrogen bonding with the peptides.

The binding interface of G-quadruplex RNA resembles ribosomal RNA in that most contacts are made through the sugar phosphate backbone (Allers and Shamoo, 2001) and is unlike flexible single stranded RNA which is more likely to bind through its accessible nucleotide bases. In this sense the RNA quadruplex is also more similar to structured DNA than to flexible single stranded RNA.

This work reveals that Sho binds to the RNA G-quadruplex with a similar distribution of contacts and hydrogen bonds to Fx. One point of difference is that Fx has the ability to use Phe to engage in base stacking interactions with the bases forming the loops and termini of the quadruplex which may confer specificity for certain RNA targets.

The binding interface can be extremely variable – there are many ways in which contacts between the peptides and the RNA may be established. To understand whether any of these complexes are favoured it is necessary to calculate and compare the binding free energies of the complexes. This will be addressed in Chapter 7.

6.5 REFERENCES

Allers, J. & Shamoo, Y. (2001) Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J Mol Biol,* 311, 75-86.

Ambros, V. (2004) The functions of animal microRNAs. *Nature,* 431, 350-5.

Ambrus, A., Chen, D., Dai, J., Jones, R.A. & Yang, D. (2005) Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry,* 44, 2048-58.

Boisvert, F.M., Chenard, C.A. & Richard, S. (2005) Protein interfaces in signaling regulated by arginine methylation. *Sci STKE,* 2005, re2.

Carlson, C.B., Stephens, O.M. & Beal, P.A. (2003) Recognition of double-stranded RNA by proteins and small molecules. *Biopolymers,* 70, 86-102.

Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B. & Woods, R.J. (2005) The Amber biomolecular simulation programs. *J Comput Chem,* 26, 1668-88.

Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T. & Darnell, R.B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell,* 107, 489-99.

De Guzman, R.N., Turner, R.B. & Summers, M.F. (1998) Protein-RNA recognition. *Biopolymers,* 48, 181-95.

Dolzhanskaya, N., Merz, G., Aletta, J.M. & Denman, R.B. (2006) Methylation regulates the intracellular protein-protein and protein-RNA interactions of FMRP. *J Cell Sci,* 119, 1933-46.

Draper, D.E. (1999) Themes in RNA-protein recognition. *J Mol Biol,* 293, 255-70.

Dreyfuss, G., Kim, V.N. & Kataoka, N. (2002) Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol,* 3, 195-205.

Ellis, J.J., Broom, M. & Jones, S. (2007) Protein-RNA interactions: structural analysis and functional classes. *Proteins,* 66, 903-11.

Ghisolfi, L., Kharrat, A., Joseph, G., Amalric, F. & Erard, M. (1992) Concerted activities of the RNA recognition and the glycine-rich C-terminal domains of nucleolin are required for efficient complex formation with pre-ribosomal RNA. *Eur J Biochem,* 209, 541-8.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins,* 65, 712-25.

Izaguirre, J.A., Catarello, D.P., Wozniak, J.M. & Skeel, R.D. (2001) Langevin stabilization of molecular dynamics. *Journal of Chemical Physics,* 114, 2090-2098.

Jorgensen, W.L. & Tirado-Rives, J. (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci U S A,* 102, 6665-70.

Lau, N.C., Lim, L.P., Weinstein, E.G. & Bartel, D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science,* 294, 858-62.

Loncharich, R.J., Brooks, B.R. & Pastor, R.W. (1992) Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers,* 32, 523-35.

Lukasiewicz, R., Nolen, B., Adams, J.A. & Ghosh, G. (2007) The RGG domain of Npl3p recruits Sky1p through docking interactions. *J Mol Biol,* 367, 249-61.

Mello, C.C. (2007) Return to the RNAi world: rethinking gene expression and evolution (Nobel Lecture). *Angew Chem Int Ed Engl,* 46, 6985-94.

Menon, L. & Mihailescu, M.R. (2007) Interactions of the G quartet forming semaphorin 3F RNA with the RGG box domain of the fragile X protein family. *Nucleic Acids Res,* 35, 5379-92.

Morozova, N., Allers, J., Myers, J. & Shamoo, Y. (2006) Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics,* 22, 2746-52.

Oliver, A.W., Bogdarina, I., Schroeder, E., Taylor, I.A. & Kneale, G.G. (2000) Preferential binding of fd gene 5 protein to tetraplex nucleic acid structures. *J Mol Biol,* 301, 575-84.

Ramos, A., Hollingworth, D. & Pastore, A. (2003) G-quartet-dependent recognition between the FMRP RGG box and RNA. *Rna,* 9, 1198-207.

Ryckaert, J.P., Ciccotti, G. & Berendsen, H.J.C. (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics,* 23, 327-41.

Schaeffer, C., Bardoni, B., Mandel, J.L., Ehresmann, B., Ehresmann, C. & Moine, H. (2001) The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. *Embo J,* 20, 4803-13.

Treger, M. & Westhof, E. (2001) Statistical analysis of atomic contacts at RNA-protein interfaces. *J Mol Recognit,* 14, 199-214.

Wang, J.M., Cieplak, P. & Kollman, P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry,* 21, 1049-1074.

Wen, J.D. & Gray, D.M. (2002) The Ff gene 5 single-stranded DNA-binding protein binds to the transiently folded form of an intramolecular G-quadruplex. *Biochemistry,* 41, 11438-48.

Williamson, J.R. (2000) Induced fit in RNA-protein recognition. *Nat Struct Biol,* 7, 834-7.

Wilusz, J.E., Sunwoo, H. & Spector, D.L. (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev,* 23, 1494-504.

Zanotti, K.J., Lackey, P.E., Evans, G.L. & Mihailescu, M.R. (2006) Thermodynamics of the fragile X mental retardation protein RGG box interactions with G quartet forming RNA. *Biochemistry,* 45, 8319-30.

CHAPTER 7

FREE ENERGY CALCULATIONS OF THE RGG BOX PEPTIDE-RNA COMPLEXES

7.1 INTRODUCTION

7.1.1 Background

Complex formation occurs when two monomers interact favourably with one another. In other words, complex formation can only occur when the energy of the complex is lower than the sum of the energies of the independent monomers i.e. the Gibbs free energy of binding (ΔG) is a negative value. The Gibbs free energy may be determined experimentally through the relationship of ΔG to the experimentally derived dissociation constant (K_d), or it may be calculated computationally.

In this Chapter, the binding of Sho and Fx peptides to an RNA quadruplex is investigated. The computed free energies of the complexes can be used to assess the relative stabilities of various complexes. Free energy calculations also provide information about the relative stabilization or destabilization of the monomers upon complexation. This question is of particular interest here, as other studies have shown that the stability of the G-quadruplex structure may be influenced by interaction with different proteins. G-quadruplex stabilization, or destabilization (unwinding of the quadruplex), has functional relevance in processes such as transcription and in telomere maintenance (Weisman-Shomer and Fry, 1994, Erlitzki and Fry, 1997, Fry, 2007, Oganesian and Bryan, 2007, Fan et al., 2009, Paramasivam et al., 2009).

It has been suggested that the Fragile X RGG box may stabilize or destabilize RNA depending upon concentration and the particular RNA target (Menon and Mihailescu, 2007, Menon et al., 2008). Although the simulations conducted in this study are not long enough to completely describe a process such as quadruplex unfolding, the results should elucidate the number and type of interactions that affect quadruplex stability.

7.1.2 Aims of this chapter

The aims of this study are to determine whether any of the Sho or Fx complexes characterized in Chapter 6 are energetically preferred, and to examine the role of peptide and RNA stabilization or destabilization in complex formation.

7.2 METHODS

7.2.1 Molecular dynamics simulations of the RGG box peptides and a model RNA quadruplex

Molecular dynamics simulations of 49 Sho-RNA and 52 Fx-RNA complexes were conducted as described in the methods section of Chapter 6. All complexes were simulated for 25 ns, following which simulation of a subset of 6 Sho-RNA and 6 Fx-RNA complexes was continued up to 56 ns.

For ease of reference, the RNA sequence and peptide residues are set out in Figure 7-1.

A

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
DNA	T	G	A	G	G	G	T	G	G	G	T	A	G	G	G	T	G	G	G	T	A	A
RNA	U	G	A	G	G	G	U	G	G	G	U	A	G	G	G	U	G	G	G	U	A	A

B

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Sho	K	G	G	**R**	**G**	**G**	A	R	G	S	A	**R**	**G**	**G**	V	**R**	**G**	**G**	A	R	G
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Fx	G	R	R	**R**	**G**	**G**	G	G	R	G	Q	G	G	R	G	**R**	**G**	**G**	G	F	K

Figure 7-1. G-quadruplex sequence and peptide sequences used in docking
(A) The c-MYC oncogene G-quadruplex DNA sequence and its RNA analogue. The guanine tetrads are colour coded, tetrad 1 (orange), tetrad 2 (yellow), tetrad 3 (brown). (B) The Sho and Fx peptide amino acid sequences.

7.2.2 Free energy calculations

Free energy calculations were performed using the MM-PBSA scripts implemented in AMBER 9 (Case et al., 2005). MM-PBSA (Srinivasan et al., 1998) calculates the approximate free energy of complex formation from a given simulation of 2 molecules that form an association. This method is used frequently in the study of biological systems (Kollman et al., 2000), including in the study of RNA complexes (Nifosi et al.,

2000, Reyes and Kollman, 2000, Reyes et al., 2001, Tsui and Case, 2001). The MM_PBSA scripts, used for this purpose, calculate the enthalpy change (ΔH) with an implicit treatment of solvation for (i) the complex as a whole (H_{Comp}), (ii) the receptor, which in this case is the RNA (H_{RNA}), and (iii) the ligand, which in this case is the peptide ($H_{Sho/Fx}$). The enthalpy of the RNA and the peptide is subtracted from the enthalpy of the complex as a whole to derive an estimate of the enthalpy of the interaction (ΔH_{Bind}):

$$\Delta H_{Bind} = H_{Comp} - (H_{RNA} + H_{Sho/Fx}) \quad (1)$$

The enthalpy is made up of gas phase energy and solvation energy components:

$$H = E_{MM} + G_{PBSA} \quad (2)$$

E_{MM}, the gas phase energy, is determined using molecular mechanics and includes the bond (E_{bond}), angle (E_{angle}) and dihedral energies (E_{tors}), van der Waals (E_{vdw}) and electrostatic (E_{elec}) energies calculated with no nonbonded cutoff.

In this study the electrostatic component (G_{PB}) of the solvation free energy is calculated with the Poisson-Boltzmann (PB) electrostatic continuum method using the PBSA program (Luo et al., 2002), as implemented in the MM-PBSA scripts. The reaction field energy is calculated as the difference in taking the system from a vacuum dielectric (ε=1) to an aqueous dielectric (ε=80). A probe radius of 1.4 Å was used with the radii from the AMBER topology file describing the molecular system. The hydrophobic component (G_{SA}) of the solvation free energy is estimated with solvent accessible surface area dependent terms and the surface area is calculated using the molsurf program (Connolly, 1983).

To calculate the Gibbs free energy of the interaction (ΔG_{Bind}) it is also necessary to calculate the change in entropy of the interaction.

$$\Delta G_{Bind} = \Delta H - T\Delta S \quad (3)$$

$$\Delta G_{Bind} = [H_{Comp} - (H_{RNA} + H_{Sho/Fx})] - T[S_{Comp} - (S_{RNA} + S_{Sho/Fx})] \quad (4)$$

The MM-PBSA scripts were also used to run the calculation of the solute entropy term ($-T\Delta S$) by normal mode analysis (Srinivasan et al., 1998), using the NMode algorithm as

implemented in AMBER. This approach can only provide a crude estimate of entropy, as discussed by Kollman and colleagues. The main source of error arises from the harmonic approximation, which is invoked in deriving the normal mode algorithm, meaning that the entropy can only be estimated for a small number of local minimum energy conformations. Secondary sources of error arise from the possibility that the minimized structures identified might not be true local minima, and the failure to take solvent entropy into account (Donini and Kollman, 2000, Reyes and Kollman, 2000, Reyes et al., 2001). With these limitations in mind, NMode calculations were performed on a small number of snapshots from each trajectory to obtain a 'ballpark' estimate of the effect of entropy on the Gibbs free energy of the interactions.

The NMode calculations were performed on snapshots taken every 2ns starting at 13 ns and finishing at the end of the simulation. The conformers were optimized via energy minimization until the root-mean-squared (RMS) force on the atoms was less than 10^{-4} kcal/mol/Å.

ΔH_{Bind} was calculated using two different methods. In the single trajectory approach, as implemented in the MM-PBSA approach, trajectories describing the motion and energies of the peptide and RNA fragments (H_{RNA} and $H_{Sho/Fx}$) were extracted from a single simulation of the total complex. This simplification rests on the assumption that the interacting species do not undergo significant conformational change in the process of binding. This was then compared with a multiple trajectory approach in which H_{RNA} was derived from an independent 20 ns simulation of the RNA G-quadruplex (described in Chapter 5) and $H_{Sho/Fx}$ was derived using the trajectories obtained from the independent peptide simulations (described in Chapter 4). As will be recalled, the peptide simulations were conducted with the ff03 force field, not the ff99SB force field used in the RNA simulations and in the docking studies. In theory, ff03 could produce a different peptide ensemble from that which ff99SB would have produced and result in an error in quantification of peptide de/stabilization upon binding. However, this should not affect the assessment of **relative** complex stabilization, which is the primary aim of this chapter.

The superscript '*S*' will be used to denote an energy derived using the single trajectory method (as in H^S_{RNA}) and the superscript '*I*' will be used to denote the independent trajectory approach (as in H^I_{RNA}). Both methods use the same reference enthalpy for the total complex:

$$H_{Comp} = H^I_{Comp} = H^S_{Comp} \quad (5)$$

Either approach may be used to calculate the overall binding energy:

$$\Delta H^I_{Bind} = H_{Comp} - (H^I_{RNA} + H^I_{Sho/Fx}) \quad (6)$$

$$\Delta H^S_{Bind} = H_{Comp} - (H^S_{RNA} + H^S_{Sho/Fx}) \quad (7)$$

As discussed in Chapter 4, the Sho and Fx peptides are highly flexible and likely to undergo conformational change in the process of binding to RNA. It is also possible that the RNA will alter its conformation so both species can mutally adapt to one another. In these circumstances the better estimate of free energy is expected to come from the independent trajectory approach (Reyes and Kollman, 2000, Zhang and Schlick, 2006). Another benefit of the independent trajectory approach is that it provides insight into the degree to which the RNA and the peptide have adapted to each other in the binding interface Eqns (8) and (9).

$$\Delta H_{RNA} = H^S_{RNA} - H^I_{RNA} \quad (8)$$

$$\Delta H_{Sho/Fx} = H^S_{Sho/Fx} - H^I_{Sho/Fx} \quad (9)$$

$$\Delta\Delta H_{Bind} = \Delta H^S_{Bind} - \Delta H^I_{Bind} = \Delta H_{RNA} + \Delta H_{Sho/Fx} \quad (10)$$

The difference in enthalpy of binding calculated using the single trajectory approach and the independent trajectory approach (ΔH^S_{Bind} - ΔH^I_{Bind}) is equivalent to this adaptive component, that is, the difference in stabilization of the RNA and peptide in the complex compared with their stability as independent monomers. Whereas ΔH^S_{Bind} captures the interaction component of the free energy (strength of the binding interface) in similar manner to ΔH^I_{Bind}, the latter also captures the contribution of changes in stabilization in the RNA and the peptide.

In all the MM-PBSA calculations, waters and Na^+ counter ions were stripped from the trajectories but the structural K^+ ions in the RNA quadruplex were retained. A dielectric constant of 80 was used for the solvent and a value of 1 was used for the solute.

Calculations on the complexes were based on the last 20 ns of the 25 ns simulations from snapshots taken every 500 ps, and those on the independent RNA simulation were performed on snapshots taken every 500 ps over the period 500 ps - 19 ns. The independent calculations of the peptide free energy were performed by firstly capturing every 10th frame of the hybrid-REMD simulation of Sho1, Sho2 and Sho3 (or Fx1, Fx2 and Fx3) from 5ns - 25ns in a combined trajectory. Then every 5th frame of each of these trajectories was used in the MM-PBSA calculations to arrive at an average free energy for each Sho:RNA and Fx:RNA complex.

The AMBER output of the MM-PBSA calculations uses the terminology described below. These terms will also be used in the results and discussion section of this chapter. For ease of reference the enthalpy terms are also set out below.

Table 7-1. List of terms used frequently in this chapter

Term	Description
ELE	The electrostatic component of the energy of the complex or monomers.
VDW	van der Waals component of the energy of the complex or monomers.
INT	The internal energy of the complex or monomers being made up of the energies of the bonds, angles and torsional angles.
GAS	The total gas phase energy of the complex or monomers (ELE + VDW + INT)
PBSUR	The non-electrostatic component of the solvation free energy
PBCAL	The electrostatic component of the solvation free energy
PBSOL	The total solvation free energy (PBSUR + PBCAL)
PBELE	The total electrostatic energy (ELE + PBCAL)
PBTOT	The total enthalpy (GAS + PBSOL)
H_{Comp}	Enthalpy of the complex
H^S_{RNA}	Enthalpy of the RNA calculated using the single trajectory method.
H^I_{RNA}	Enthalpy of the RNA calculated using the independent trajectory method.
$H^S_{Sho/Fx}$	Enthalpy of the peptide (Sho or Fx) calculated using the single trajectory method.
$H^I_{Sho/Fx}$	Enthalpy of the peptide (Sho or Fx) calculated using the independent trajectory method.
ΔH^S_{Bind}	Enthalpy of binding (enthalpy difference upon complex formation) calculated using the single trajectory method.
ΔH^I_{Bind}	Enthalpy of binding (enthalpy difference upon complex formation) calculated using the independent trajectory method.
ΔH_{RNA}	Adaptive component of the RNA enthalpy upon binding
$\Delta H_{Sho/Fx}$	Adaptive component of the peptide enthalpy upon binding
$\Delta\Delta H^I_{Bind}$	Adaptive component of the binding enthalpy

7.3 RESULTS AND DISCUSSION

7.3.1 Estimating the free energies of the complexes

The calculated components of the free energy for each docked structure can be found in Table A7-2 (Sho-RNA complexes) and Table A7-4 (Fx-RNA complexes) of Appendix 7. The entropy calculations are set out in Table A7-3 (Sho-RNA complexes) and Table A7-5 (Fx-RNA complexes) of Appendix 7. The enthalpy of the RNA based on an independent MD simulation (described in Chapter 5) was -4522 ± 22 kcal mol^{-1}. The enthalpy of the Sho peptide based on simulations reported in Chapter 4 was -656 ± 18 kcal mol^{-1} and for the Fx peptide -860 ± 17 kcal mol^{-1}. As these values do not change, they will not be repeated in the tables below.

Grand average enthalpy across all Sho-RNA and Fx-RNA complexes and the associated standard deviations are set out in Table 7-2. Grand average entropy values for all complexes, together with the associated standard deviations, are set out in Table 7-3.

Table 7-2. Grand average stabilization energies based on all Sho-RNA and Fx-RNA complexes [a,b]

	H_{comp}		H^S_{RNA}		H^S_{Sho}		ΔH_{RNA}		ΔH_{Sho}		ΔH^S_{Bind}		ΔH^I_{Bind}	
	Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD
Sho-RNA	-5591	18	-4517	14	-643	9	5	14	13	9	-431	17	-413	**18**
Fx-RNA	-5793	17	-4521	15	-820	9	1	15	40	9	-452	21	-412	**16**

[a] All energies (kcal mol^{-1}). All reported values are the average calculated for all Sho-RNA or Fx-RNA complexes listed in Table 6-1, Chapter 6. The standard deviation (SD) of this set is also reported. [b] All enthalpy quantities are defined in the Methods section above.

The data presented in Table 7-2 show a remarkable similarity in the average enthalpy of binding (ΔH^I_{bind}) for the Sho-RNA and Fx-RNA complexes, -413 ± 18 kcal mol^{-1} and -412 ± 16 kcal mol^{-1}, respectively. On average both peptides lose internal stability upon complexation ($\Delta H_{Sho/Fx}$) = 13 ± 9 kcal mol^{-1} for Sho and 40 ± 9 kcal mol^{-1} for Fx. This indicates rearrangement of the peptide in order to establish contacts with the RNA. Over the set of complexes there is a good deal of variation as to whether the RNA quadruplex becomes relatively stabilized or destabilized upon complexation. The addition of negative and positive values produces an overall low average value and relatively high standard deviation.

There are notable differences in the average values of ΔH^{S}_{Bind} for Sho and Fx. This is the value of the binding enthalpy calculated using the single trajectory method. As such it does not take account of changes in peptide and RNA stability upon complexation, but it does describe the strength of the binding interface i.e. the number of contacts made and the strength of those contacts. The results shown here reveal that on average Fx establishes a stronger binding interface than Sho. However, Fx also incurs a greater energetic penalty as a result of losing internal stability. When both factors are offset Sho and Fx have the same average binding enthalpy (ΔH_{bind}).

The entropies of the complexes have also been calculated and used to derive the Gibbs free energy ΔG. The average results are presented in Table 7-3 and show that, upon complexation, Sho and Fx experience a very similar average change in entropy (38 kcal $mol^{-1}$and 41 kcal mol^{-1}, respectively).

Table 7-3. Entropy calculations of Sho-RNA complexes ordered by ΔG [a,b,c]

	TS_{Comp} [c]	SD	TS_{RNA} [d]	SD	TS_{Sho} [e]	SD	$-T\Delta S$ [f]	SD	ΔH^{I}_{Bind} [g]	SD	ΔG [h]	SD
Sho-RNA	801	3	609	2	230	3	38	5	-413	18	-376	17
Fx-RNA	814	4	609	2	246	5	41	6	-412	16	-369	18

[a] All energies (kcal mol^{-1}). All reported values are the average calculated for all Sho-RNA or Fx-RNA complexes listed in Table 6-1, Chapter 6. The standard deviation (SD) of this set is also reported. [b] Entropy values are averages from snapshots every 2ns over the last 12 ns of the simulation. SD = 1 standard deviation. [c] Entropy values calculated at 300 K. [d] TS_{comp} = complex entropy. [e] TS_{RNA} = RNA entropy. [f] TS_{Sho} = peptide entropy. [g] $T\Delta S$ = change in entropy upon complex formation. [h] ΔH^{I}_{Bind} = change in enthalpy upon complex formation. [i] ΔG = change in Gibbs free energy upon complex formation, calculated as $\Delta H^{I}_{Bind} - T\Delta S$.

As normal mode analysis is only a rough tool for calculation of entropy in these flexible systems, the values reported in Table 7-3 should be treated as a 'broad ballpark estimate', as suggested by Reyes and Kollman (2000). Reyes and Kollman studied the binding of a 101 residue region of the U1A protein to an RNA transcript of similar size to the G-quadruplex which could assume both a hairpin and loop structure. Their values for $T\Delta S$ are similar to the values predicted here. They found an entropy change of 51.1 kcal mol for binding to the RNA hairpin and 39.6 kcal mol for binding to the RNA loop (Reyes and Kollman, 2000).

As the error in the entropy values is likely to be high, the rest of this discussion will proceed on the basis that 'estimated free energy' refers only to the enthalpy component of the free energy. However, it should be noted that, in general, the

ordering of the complexes by their relative stability does not change if the entropy difference is taken into account.

The similarity in grand average free energies across all complexes and the relatively small associated standard deviation indicates that the Sho and Fx complexes have a similar degree of stability. The change in enthalpy was calculated using the MM-PBSA method for snapshots derived from each trajectory (as described in Methods), and the average value for each complex is plotted below. Error bars on these data represent the standard deviation in these values (Figure 7-2 (A) and (B)). A frequency distribution is presented in Figure 7-2 (C).

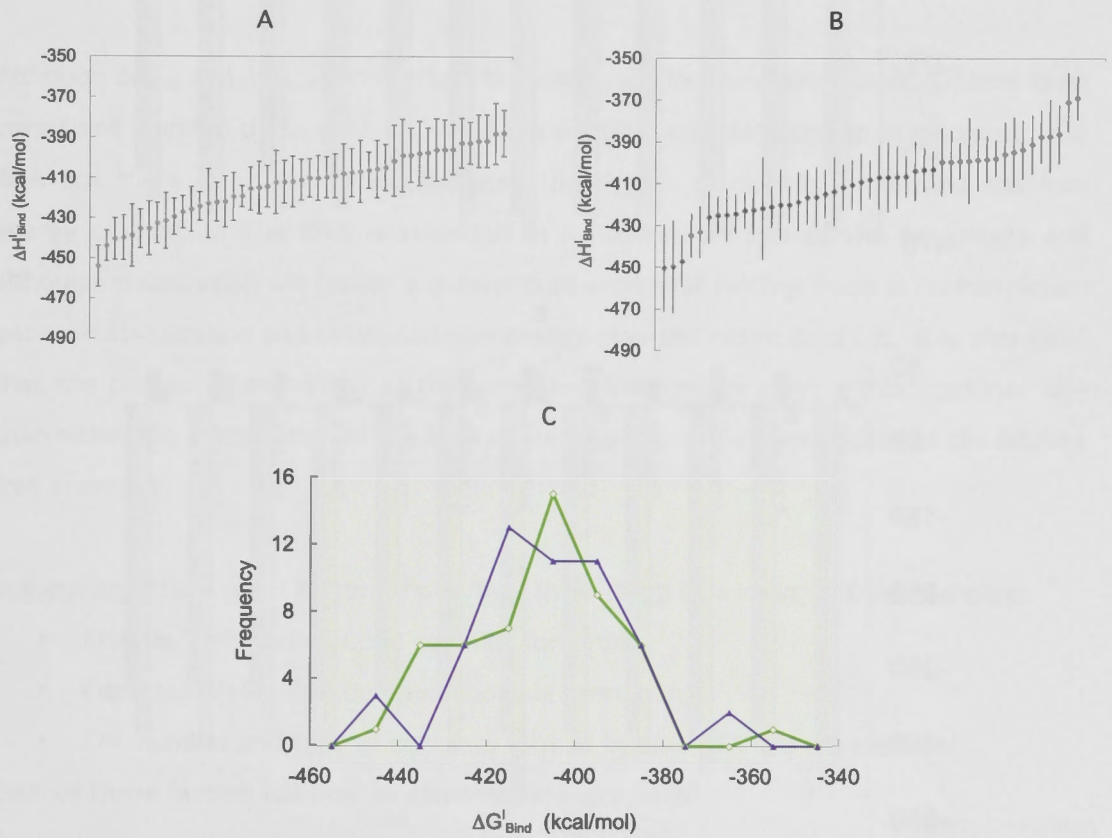


Figure 7-2. The average $\Delta H'_{Bind}$ and frequency distribution of these values
(A) Sho-RNA complexes and (B) Fx-RNA complexes. (C) Frequency distribution of the values of $\Delta H'_{Bind}$ based on the set of Sho-RNA complexes (green) and the set of Fx-RNA complexes (purple).

As shown in Figure 7-2, the complexes with the greatest estimated free energy are close in energy. When the standard deviation of each set of values is taken into account, it becomes impossible to identify a single dominant/most stabilized complex. However, the most and least stabilized complexes are clearly distinct. The factors that contribute to complex stability will now be investigated.

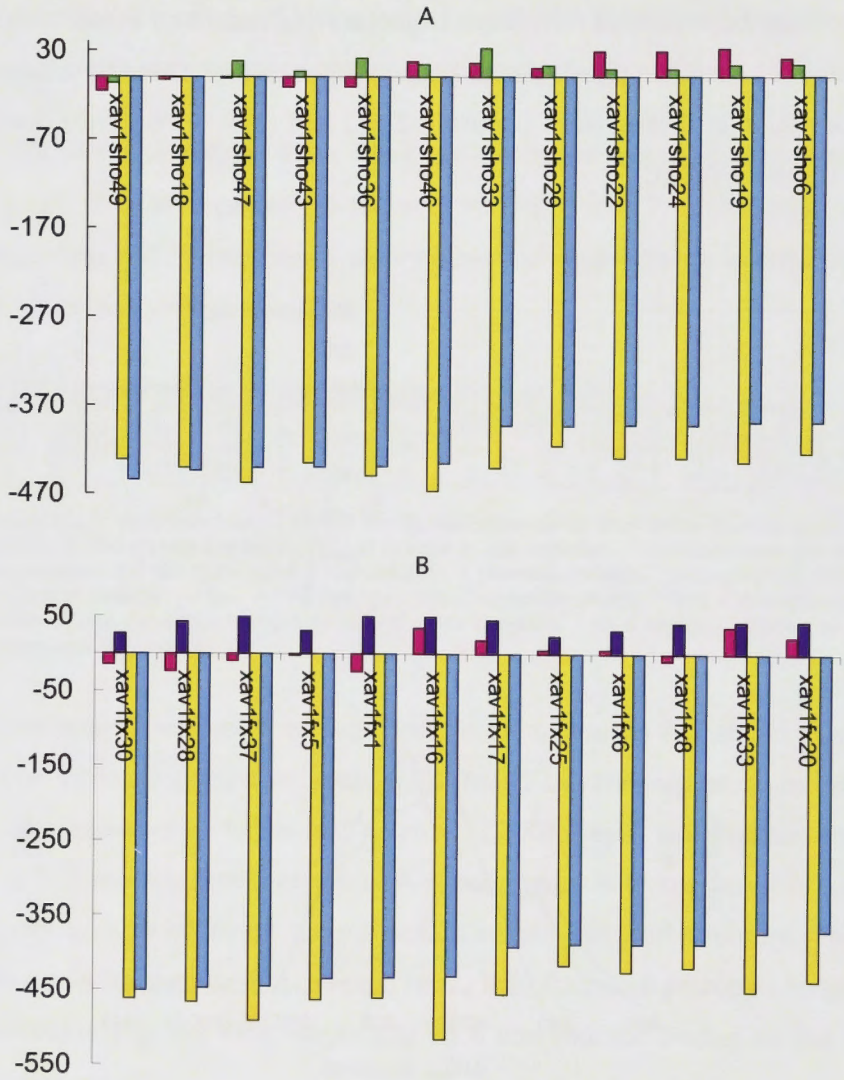


Figure 7-3. Comparison of stabilization energies and estimated free energy
The estimated free energy of binding ΔH^{I}_{Bind} (blue) is plotted with the relative stabilization energies of the RNA, ΔH_{RNA} (pink) and the peptide, $\Delta H_{Sho/Fx}$ (Sho, green) and (Fx, purple). The contacts component of the estimated free energy ($\Delta H^{I}_{Bind} - [\Delta H_{RNA} + \Delta H_{Sho/Fx}]$) (yellow) is compared to the total estimated free energy of binding (blue). All energies are in kcal mol^{-1}. (A) 6 Sho-RNA complexes with highest and 6 complexes with lowest ΔH^{I}_{Bind}. (B) 6 Fx-RNA complexes with highest and 6 complexes with lowest ΔH^{I}_{Bind}.

In Figure 7-3, the estimated free energy of binding (ΔH_{Bind}) is decomposed into RNA de/stabilization, peptide de/stabilization and interface interaction energies for 6 of the most tightly bound and 6 of the most loosely bound complexes. The complexes are arranged from greatest to least ΔH_{Bind} (blue bar). Whereas the RNA may be stabilized or destabilized upon binding, the peptide generally becomes less stable. Where both the RNA and peptide are relatively destabilized in the complex, this must be compensated by a high degree of interaction to produce a high ΔH_{Bind}. This is illustrated in the case of xav1sho46 (6^{th} data point in Figure 7-3(A)), where the RNA and peptide are both relatively destabilized in the complex, but this is offset by a high contacts component to the free energy. This makes intuitive sense, as the internally stabilizing features of the peptide and RNA may loosen in order for them to interact fully with each other.

Although ΔH_{RNA} and $\Delta H_{Sho/Fx}$ both affect the estimated free energy of binding, there is no consistent trend in these values. Overall the peptide is destabilized upon binding to the RNA but there is no correlation between the degree of destabilization and the free energy of binding. The RNA is stabilized in approximately half of the complexes and although stabilization will favour a greater free energy of binding there is no correlation between stabilization and estimated free energy over the entire data set. It is also clear that the contacts component of the estimated free energy plays a much greater role than either the stabilization of the RNA or the peptide in these estimates of the binding free energy.

In summary, there are 3 factors that affect the binding free energy of the complex:

- RNA de/stabilization upon complex formation.
- Peptide de/stabilization upon complex formation.
- The number and type of contacts formed by the RNA and the peptide.

Each of these factors will now be examined in more detail.

7.3.2 RNA stabilization or de-stabilization

The change in stabilization of the RNA upon complexation is interesting because protein-induced stabilization or destabilization of G-quadruplex nucleic acid has important functional implications, as noted in Chapter 5. The relative stabilization of the RNA is also a factor in complex formation, with increased RNA stabilization favouring

complex formation. The most stable complexes also demonstrate RNA stabilization as illustrated by the data presented in Table A7-5 of Appendix 7.

While RNA stabilization generally contributes favourably to complex stabilization (ΔH_{Bind}) it does not necessarily contribute favourably to the strength of the binding interface (ΔH^{S}_{Bind}). Figure 7-4(A) shows a weak correlation between RNA stabilization and complex stabilization (ΔH_{Bind}) while there is no correlation between RNA stabilization and ΔH^{S}_{Bind} Figures 7-4(B). The correlation between RNA stability and complex stabilization is weak as the variations in RNA stability seen here are relatively minor, both in the Sho complexes (ΔH_{RNA} ranges from -17 to 48 kcal mol^{-1}) and the Fx complexes (ΔH_{RNA} ranges from -29 to 36 kcal mol^{-1}) and are outweighed by other components of the free energy. Most of the ΔH_{RNA} values are within the range of variation seen in the independent RNA quadruplex simulations. Nevertheless, it is interesting to look more closely at factors that appear to enhance stability or instability.

A change in the stabilization energy of the quadruplex may result from disturbance of the tetrads or disturbance of hydrogen bonds between nucleotides at the 5' and 3' termini. As the G-quadruplex is largely stabilized through the stacking interactions of its bases (in addition to cation effects) it is plausible that hydrogen bonding interactions between the peptide and nucleotide bases could affect quadruplex stability. This hypothesis is investigated through identifying and classifying hydrogen bonds in terms of quadruplex stability, as shown in Figure 7-4(C).

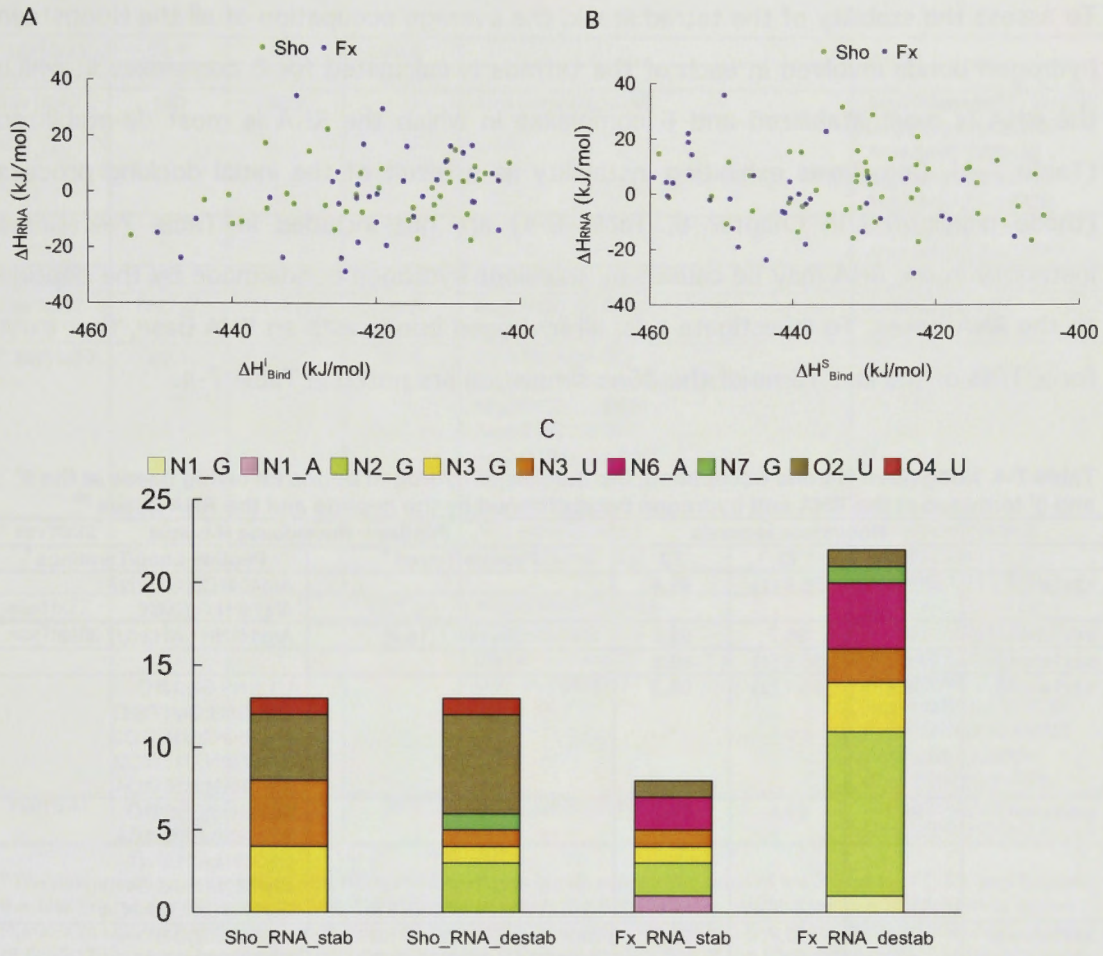


Figure 7-4. Decomposition of all Sho-RNA and Fx-RNA hydrogen bonds involving the nucleobases according to atom type and RNA stabilization/destabilization
(A) Plot of ΔH_{RNA} against ΔH^{I}_{Bind} for each of the Sho (green) and Fx (purple) complexes. (B) Plot of ΔH_{RNA} against ΔH^{S}_{Bind} for each of the Sho (green) and Fx (purple) complexes. (C)The total number of hydrogen bonds, involving given atoms ofthe RNA bases (N1 of guanine - yellow, N1 of adenine - pink, N2 of guanine - olive, N3 of guanine - gold, N3 of uracil – orange, N6 of adenine – dark pink, N7 of guanine – bright green, O2 of uracil - brown and O4 of uracil - red). The complexes are sorted according to whether the RNA is stabilized in the complex (RNA_stab) or destabilized (RNA_destab).

In both systems there is a noticeable increase in hydrogen bonds involving N2 of the guanine base (olive green) in the RNA destabilized complexes as well as the presence of a small number of hydrogen bonds involving N7 of guanine (dark green). The N2 and N7 atoms of guanine are involved in Hoogsteen hydrogen bonds, so some instability might be expected if the peptide is competing for these sites. However, hydrogen bonds involving N2 are found in some Fx complexes where the RNA is stabilized.

To assess the stability of the tetrad stack, the average occupation of all the Hoogsteen hydrogen bonds involved in each of the tetrads is calculated for 6 complexes in which the RNA is most stabilized and 6 complexes in which the RNA is most de-stabilized (Table 7-4). Complexes exhibiting instability as a result of the initial docking process (those highlighted in Chapter 6, Table 6-1) are not included in Table 7-4. Some instability in the RNA may be caused by transient hydrogen bonds made by the peptide to the RNA bases. To investigate this, all hydrogen bonds with an RNA base, that exist for ≥ 10% of the last 15 ns of the 25ns simulation are noted in Table 7-4.

Table 7-4. Hoogsteen H-bond occupation, the number of hydrogen bonds involving bases at the 5' and 3' terminus of the RNA and hydrogen bonds formed by the peptide and the RNA bases [ab]

	Hoogsteen H-bonds			Peptide – Nucleobase H-bonds	
	T1	T2	T3	Peptide-Tetrad [e]	Peptide-Loop/Terminus [f]
xav1sho7	93.3(2) [g]	95.9 (1)	99.4		Arg40@CD:G2@N7 Ala7@N:G2@O6
xav1sho16	99	98.7	99.5	G4@N2:Gly1@O (16.2)	Arg4@NH1/NH2 :U7@O2
xav1sho49	97.6 (1)	96.6 (1)	99.6		
xav1sho36	99.1	95.1 (2)	98.3		U1@N3:Gly9@O U16@N3:Gly17@O Arg16@C:U16@O2 Gly17@N:U16@O2 Ala19@N:U16@O4
xav1sho43	99.1	99.1	99.7		A21@N6:Gly3@O A22@N6:Gly5@CA Gly3@N:A21@N1 Gly6@N:U20@O4 Gly17@CA:U16@O4
xav1sho15	99.5	98.7	99.9	Arg4@NH1/NH2:G4@N3 Arg4@NH1/NH2:G5@N3 Gly5@C:G5@N3	U7@N3:Gly6@O U7@N3:Arg4@O
* xav1sho23 [h]	99.7	99.8	99.5	Arg16@NH2:G9@N3 G15@N2:Arg8@CZ Ser10@CB:G15@N7	A21@N6:Arg8@O
* xav1sho24	98.6	99.7	99.7		A12@N6:Gly5@O U11@N3:Gly2@O U11@N3:Lys1@O A12@N6:Gly5@O Ala7@N:A12@N1 Ala7@N:A12@N3
* xav1sho6	99	95.7 (2)	99.6	Arg12@NH2:G4@N3	Gly2@C:G3@N3
xav1sho21	99.9	98.9	99.8	Gly2@C:G6@N3	U11@N3:Arg12@O U11@N3:Gly13@O Arg4@NE:A21@N1 Arg4@NH2:A21@N1 Ser10@N:U11@O2 Arg16@NH1/NH2:U11@O2
xav1sho46	99.8	99.2	99.7	G17@N2:Val15@O G18@N2:Gly14@O	G2@N2: Val15@O Gly5@N :G2@O6
*xav1sho19$_a$	98.6	96.9	99.6		U7@N3:Gly5@O Arg4@CA: U7@O2 Gly5@N:U7@O2
xav1fx1	99.3	97.1	98.6	G15@N2:Gly17@N Gly17@C:G15@N3 Gly17@N:G15@N3	Gly17@C:U20@O4 Gly18@N: A21@N1 Gly19@N:A21@N1 Phe20@N:A21@N1 Gly7@N:G2@O6 Gly8@N:G2@O6

*xav1fx11[d]	99.6	98	99.9		
xav1fx28	99.5	98.9	99.8		
xav1fx42	100	100	100	Arg16@N:G13@N3	Arg8@NH1/NH2:G2@N3 U16@N3:Gly19@O Arg4@NE:G2@O6 Arg4@NH2:G2@O6 Arg2@NH2:G2@O6
xav1fx47	100	100	100		Arg3@NE:A3@N7 Arg3@NH2:A3@N7
xav1fx52	99.1	99.2	99.8	G19@N2:Gly15@O	Arg3@N:U7@O4
Xav1fx17	74.5 (2)	97.7 (1)	99.8	Arg9@CD:G18@N3 Arg9@NH2:G18@N3	A3@N6:Gly1@O
* xav1fx20	99.4	99.7	99.6	G6@N2:Arg4@O Gly1@C:G8@N3 Arg2@N:G8@N3 Arg3@CD:G8@N7 Arg3@NE:G8@N7 Arg3@NH1:G8@N7 Arg4@CA:G6@N3 G6@N2:Arg4@O	
* xav1fx22	99.9	94.7 (2)	94.6 (2)	Arg16@N:G6@N3 G5@N2:Lys21@O	U11@N3:Gly5@O Arg4@C:U11@O2 Arg5@N:U11@O2
xav1fx27	99.3	99.1	**72.7 (5)**	Arg2@NH1:G8@N3	Gly15@C:U20@O2
xav1fx48				G19@N2:Gly7@O Gly5@N:G19@N7 Gly7@C:G6@N7	A22@N6:Gly15@O Gly6@C:U20@O2 Arg14@NE:U7@O2 Arg14@NH2:U7@O2 Arg20@NH2:A21@N3 Arg9@NE:G2@O6 Arg9@NH2:G2@O6
xav1fx41	99.8	99	99.5	G13@N2:Gly7@O	U1@N3:Arg9@O U1@N3:Gly10@O

[a] The average occupancy (%) of all 8 Hoogsteen hydrogen bonds constituting each of the 3 tetrads (T1, T2, and T3) over the last 5ns of the 25ns simulations. [b] complexes in which the RNA is most destabilized are highlighted in yellow. [c] Number of U20-A22 hydrogen bonds which occur ≥ 90% of the time over the last 5ns of the 25ns simulations. [d] Number of G2-G17 hydrogen bonds which occur ≥ 90% of the time over the last 5ns of the 25ns simulations. [e] Hydrogen bonds formed between the peptide and the RNA bases involved in the tetrad stack which occur ≥ 10% of the time over the last 15ns of the 25ns simulations. [f] Hydrogen bonds formed between the peptide and the RNA bases involved in the loops and at the 5' and 3' terminus, which occur ≥ 10% of the time over the last 15ns of the 25ns simulations. [g] Complexes in which ΔH_{RNA} is > 1 SD are marked with an asterisk. [h] The number of Hoogsteen hydrogen bonds in a tetrad which are occupied < 90% of the time are indicated in red font.

The primary observation from this data is that a lower RNA stabilization energy does not necessarily indicate instability in the tetrad stack. The Hoogsteen bond occupation for all 8 bonds in a tetrad is similar in the Sho complexes, irrespective of whether the RNA is stabilized or destabilized. In the Fx complexes, there are 2 cases where there is an obvious difference in tetrad stability (xav1fx17 and xav1fx27) but otherwise there is not a great deal of difference in the average occupancy of these bonds. This suggests that the calculated decrease in RNA stabilization may arise from perturbations in the backbone, loop or terminal regions without having a significant impact on the stability of the guanine stack. Secondly, the peptides have more interactions with the loops and terminal regions of the RNA than the tetrad stack. There are a greater number of interactions with the tetrad stack in the RNA destabilized complexes compared to the

RNA stabilized complexes. It is unclear whether this difference is significant, however, it does suggest that contact with the tetrad stack may be destabilizing, while contacts with the loops and terminal regions of the RNA may be neutral or possibly stabilizing.

In the case of xav1fx17, contacts are focussed on G4 (117 ≤ 5Å) and G5 (143 ≤ 5Å) of the tetrad stack, which have a direct impact on the stability of the first tetrad. In xav1fx27 contacts are concentrated at the 3' terminus, there are also a large number of contacts with G6 in tetrad 3 (34 ≤ 5Å) with resulting instability in tetrad 3 as illustrated by the images of these complexes in Figure 7-5.

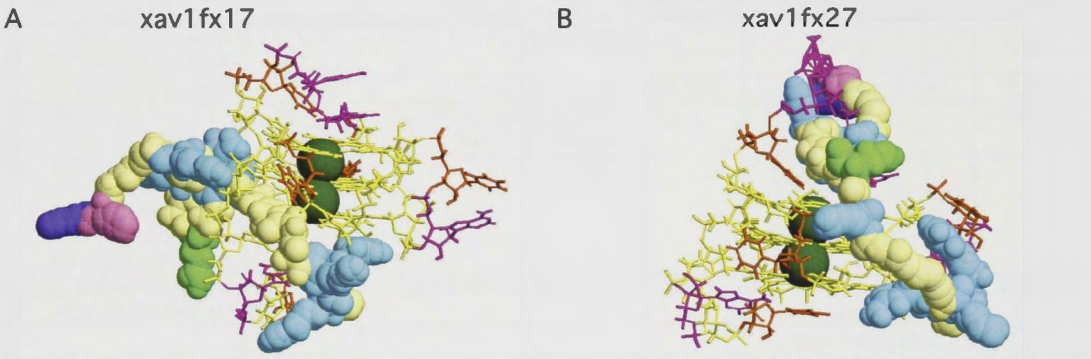


Figure 7-5. Snapshots of 2 Fx-RNA complexes in which one of the tetrads is destabilized as a result of Fx binding
(A) Complex xav1fx17 in which the Hoogsteen hydrogen bonds of Tetrad 1 are reduced to an average occupancy of 75%. (B) Complex xav1fx27 in which the Hoogsteen hydrogen bonds of Tetrad 3 are reduced to an average occupancy of 73%.

Overall, these simulations confirm the stability of G-quadruplex RNA and the likelihood that the RNA will be stabilized upon 1:1 binding with an RGG-box peptide. Minor variations in RNA stability have been scrutinized for hints as to the causes of peptide induced stability or instability. There is some indication that dense contacts with the tetrad stack tend to be destabilizing. Hydrogen bond analysis reveals that the peptides may form hydrogen bonds with the N2 and N7 atoms of guanines involved in the tetrad stack, which may also create instability over time.

7.3.3 Peptide destabilization

Unlike the RNA, the peptide is destabilized in all the Fx-RNA complexes and all but 2 of the Sho-RNA complexes. ΔH_{Sho} ranges from -7 to 33 kcal mol^{-1} and ΔH_{Fx} ranges from 21

to 56 kcal mol^{-1}. As noted earlier, the use of a different force field for the independent peptide simulations could affect the accuracy of these figures. Bearing that caveat in mind, these data point to peptide destabilization in complex formation i.e. internally stabilizing contacts are replaced by contacts with the RNA.

In Figure 7-6, the radius of gyration (R_G) of the peptide in each of the Sho and Fx complexes is plotted against the interaction energy (ΔH^S_{Bind}). These data show a weak correlation suggesting that the more extended conformations (higher R_G) are associated with a stronger binding interface.

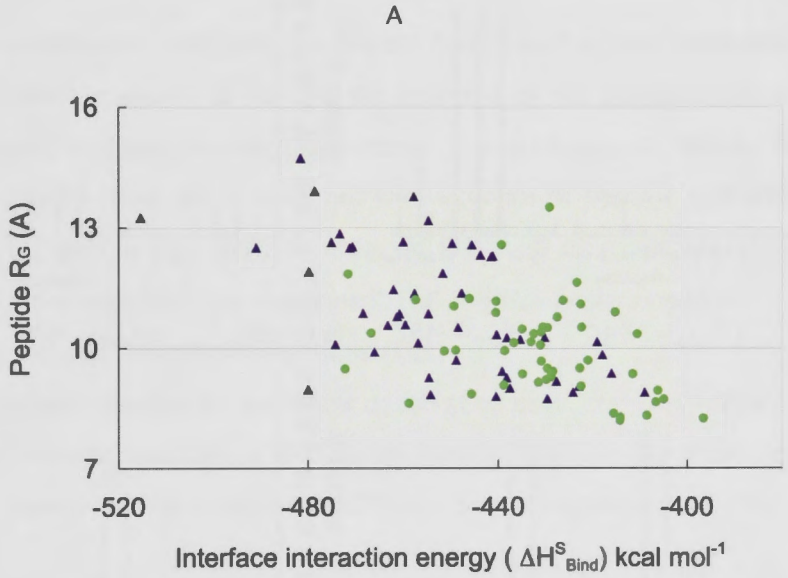


Figure 7-6. Relationship between the peptide radius of gyration and the binding interface interaction energy
R_G for Sho in each of the Sho-RNA complexes is plotted as a green dot, R_G for Fx in each of the Fx-RNA complexes is plotted as a purple triangle.

Peptide destabilization indicates that the flexible peptide has altered its conformation to adapt to the RNA, which raises the question of whether the peptide adopts a particular secondary or tertiary structure at the binding interface. DSSP results demonstrate that the Sho and Fx peptides do not prefer any particular tertiary structure at the binding interface. Most contacts are made by the peptide when it is almost devoid of any secondary structure elements, as illustrated by Figure 7-7.

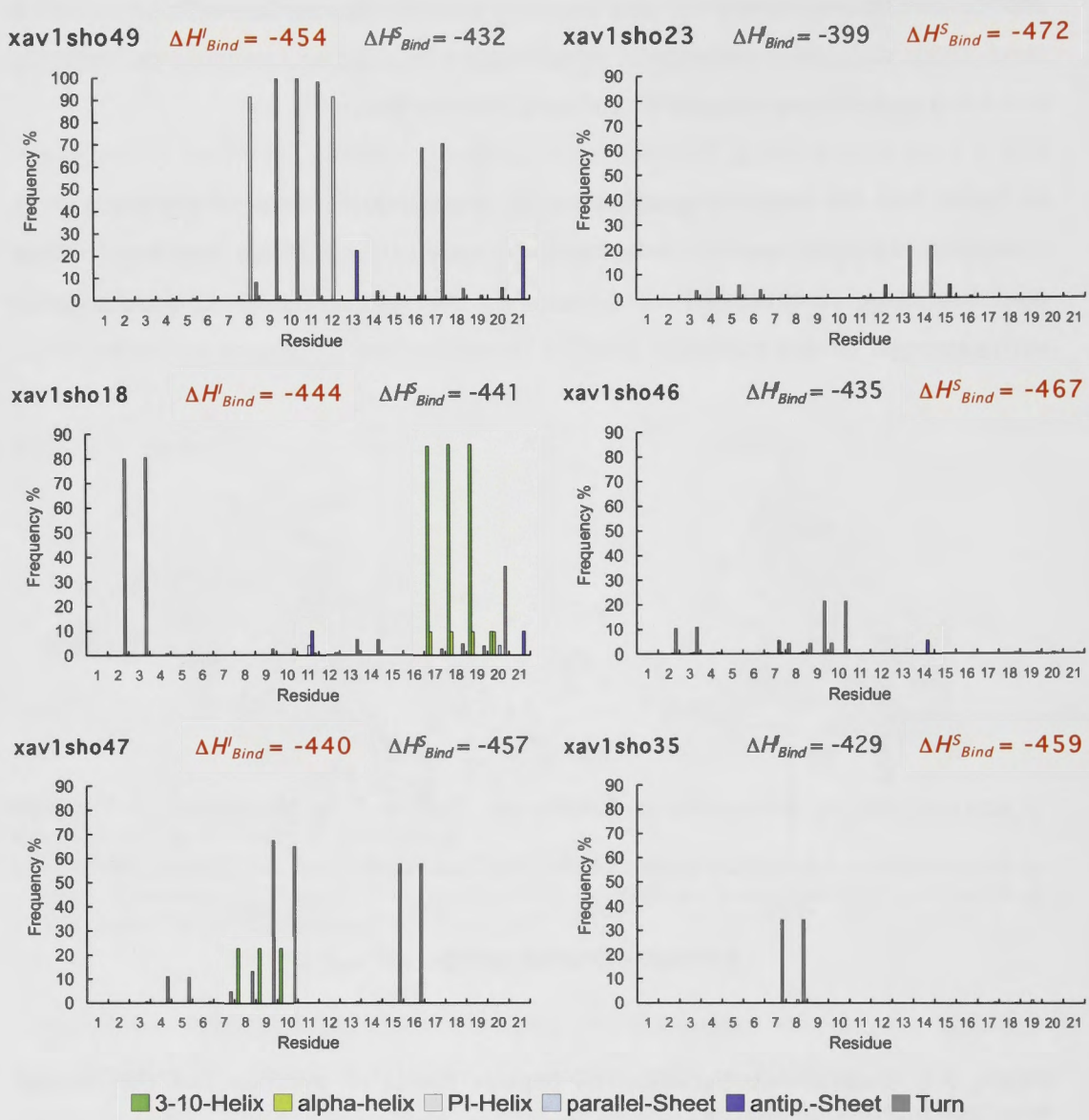


Figure 7-7. DSSP analysis of the Sho peptide comparing complexes that have a high estimated free energy of binding with complexes that have the strongest binding interface

The 3 complexes on the LHS have high estimated free energy of binding ΔH_{Bind} the 3 complexes on the RHS have the highest values of the interaction energy ΔH^S_{Bind}. The complex name and value of peptide stabilization (-ve values) or destabilization (+ve values) is given in kcal mol^{-1}.

Figure 7-7 contains DSSP plots for 3 of the Sho-RNA complexes that have the greatest estimated free energy (ΔH_{Bind}) and for 3 of the Sho-RNA complexes that have the strongest binding interface (ΔH^S_{Bind}). Comparison of these plots reveals a dynamic tension between peptide order/disorder and the strength of the binding interface.

Elements of secondary structure are associated with the peptide being more stabilized in the complex which is, in turn, related to higher free energy of binding. However, a stronger binding interface is associated with greater destabilization of the peptide. In other words, there is a trade-off between protein stabilization and the number of contacts that can be made with the RNA. The number and type of contacts must be sufficient to compensate for loss of stability in the peptide to result in a high binding free energy. This flexibility gives the Sho and Fx peptides the ability to adapt to a range of potential binding partners with variable binding interfaces.

7.3.4 Analysis of the binding interface

The binding stabilization energy is the largest component of the estimated free energy (Figure 7-3), so it is useful to explore the features of the peptide-RNA interface that contribute most strongly to this component. The enthalpy of binding based on the single trajectory method, which does not take account of peptide and RNA adaptation, (ΔH^{S}_{Bind}) will be used in this analysis. ΔH^{S}_{Bind} is made up of an electrostatic component (PBELE) and a non-electrostatic component (NE). PBELE is comprised of:

- stabilizing electrostatic interactions (ELE)
- electrostatic penalty for monomer desolvation upon complex formation (PBCAL)

The non-electrostatic component (NE) is the total of the van der Waals energy (VDW), the internal energy of the molecules (INT) and the non-polar contribution to solvation (PBSUR).

From Table 7-5, it can be seen that the electrostatic component (PBELE) contributes most towards the stabilization energy of the complex. ELE and PBCAL are large and of opposite sign and there is a large fluctuation in these quantities both during the MD simulation of each complex and across each set of complexes. However, PBELE (= ELE + PBCAL) exhibits much smaller deviations both within each run and across the set of complexes.

Table 7-5. Decomposition of the interaction component of the free energy [abcdefg]

	VDW	NE	ELE	PBCAL	PBELE	ΔH		VDW	NE	ELE	PBCAL	PBELE	ΔH
xav1sho27	-88	-131	-3400	3058	-342	-472	xav1fx16	-132	-166	-3567	3218	-350	-515
xav1sho23	-100	-143	-3346	3017	-329	-472	xav1fx33	-111	-134	-3494	3128	-316	-450
xav1sho46	-118	-162	-2860	2555	-305	-467	xav1fx35	-109	-141	-3408	3013	-333	-474
xav1sho35	-69	-111	-3104	2756	-348	-459	xav1fx31	-104	-134	-3110	2737	-345	-479
xav1sho47	-99	-143	-2868	2554	-315	-457	xav1fx48	-101	-128	-3462	3118	-343	-471
xav1sho41	-58	-100	-2998	2646	-352	-451	xav1fx37	-96	-120	-3417	3084	-366	-487
xav1sho36	-53	-94	-3017	2662	-355	-449	xav1fx49	-88	-115	-3443	3069	-347	-461
xav1sho4	-78	-118	-3063	2732	-331	-449	xav1fx46	-86	-112	-3587	3244	-359	-471
xav1sho31	-63	-103	-2972	2628	-343	-446	xav1fx17	-85	-109	-3162	2803	-346	-455
xav1sho3	-51	-91	-2958	2604	-354	-446	xav1fx22	-84	-109	-3056	2662	-346	-455
xav1sho18	-62	-103	-2826	2489	-337	-441	xav1fx34	-83	-106	-3386	2993	-373	-480
xav1sho33	-98	-141	-2973	2673	-300	-441	xav1fx27	-76	-99	-3268	2899	-375	-473
xav1sho16	-55	-96	-2985	2642	-343	-439	xav1fx20	-76	-100	-3188	2801	-335	-435
xav1sho42	-79	-121	-3104	2786	-318	-439	xav1fx7	-72	-91	-3367	3020	-367	-458
xav1sho1	-61	-98	-2911	2570	-341	-439	xav1fx23	-72	-94	-3584	3217	-366	-461
xav1sho28	-53	-91	-2759	2413	-346	-437	xav1fx5	-70	-94	-3157	2785	-369	-463
xav1sho43	-59	-100	-3132	2797	-335	-435	xav1fx19	-68	-82	-3425	3031	-364	-446
xav1sho51	-45	-86	-2838	2490	-349	-435	xav1fx45	-67	-82	-3326	2933	-367	-449
xav1sho19	-64	-106	-2837	2509	-328	-433	xav1fx1	-66	-88	-3255	2888	-372	-460
xav1sho10	-58	-99	-3014	2681	-333	-433	xav1fx2	-65	-88	-3030	2642	-363	-452
xav1sho49	-39	-76	-2918	2562	-356	-432	xav1fx41	-63	-86	-3298	2952	-395	-482
xav1sho50	-48	-87	-2965	2622	-344	-431	xav1fx15	-59	-77	-3190	2844	-364	-441
xav1sho45	-39	-74	-2771	2414	-358	-431	xav1fx42	-59	-75	-3129	2750	-373	-448
xav1sho21	-45	-81	-3126	2777	-350	-431	xav1fx40	-59	-77	-3154	2761	-394	-471
xav1sho39	-44	-84	-2903	2557	-347	-431	xav1fx21	-59	-76	-3101	2737	-365	-442
xav1sho34	-50	-89	-2916	2576	-341	-430	xav1fx47	-58	-80	-3309	2993	-364	-444
xav1sho12	-61	-99	-2640	2310	-330	-430	xav1fx13	-57	-76	-3000	2633	-373	-448
xav1sho30	-58	-96	-2842	2508	-334	-430	xav1fx24	-52	-70	-3352	2979	-359	-430
xav1sho44	-87	-127	-2713	2411	-302	-429	xav1fx30	-52	-74	-3322	2949	-388	-462
xav1sho22	-63	-100	-2517	2188	-329	-429	xav1fx32	-52	-76	-3430	3066	-379	-455
xav1sho14	-41	-78	-2910	2559	-351	-429	xav1fx3	-51	-68	-3298	2934	-370	-438
xav1sho32	-45	-85	-2861	2518	-343	-428	xav1fx10	-51	-65	-3346	2981	-393	-458
xav1sho38	-53	-93	-2764	2430	-334	-426	xav1fx50	-51	-70	-3346	2982	-360	-430
xav1sho9	-43	-82	-2707	2365	-342	-423	xav1fx9	-50	-65	-2723	2340	-394	-460
xav1sho17	-40	-80	-2951	2608	-343	-423	xav1fx36	-48	-63	-3250	2870	-356	-419
xav1sho6	-68	-109	-2861	2547	-314	-423	xav1fx29	-47	-59	-3050	2654	-379	-438
xav1sho2	-57	-95	-2909	2582	-327	-423	xav1fx28	-47	-72	-2952	2580	-394	-466
xav1sho15	-48	-84	-2858	2520	-338	-422	xav1fx52	-46	-66	-3221	2842	-373	-438
xav1sho52	-53	-93	-3015	2687	-329	-421	xav1fx14	-46	-61	-3208	2838	-366	-428
xav1sho5	-36	-72	-2574	2231	-343	-416	xav1fx26	-45	-69	-3239	2904	-388	-457
xav1sho29	-46	-84	-2801	2470	-331	-415	xav1fx38	-44	-61	-3277	2917	-393	-454
xav1sho40	-41	-80	-2623	2288	-334	-414	xav1fx12	-40	-60	-3161	2774	-380	-440
xav1sho13	-35	-72	-2753	2410	-342	-414	xav1fx43	-39	-57	-2997	2637	-383	-441
xav1sho25	-39	-77	-2616	2281	-335	-412	xav1fx18	-34	-42	-3105	2739	-387	-430
xav1sho37	-51	-89	-2818	2496	-322	-411	xav1fx6	-30	-47	-2809	2431	-377	-424
xav1sho53	-48	-84	-2533	2208	-325	-409	xav1fx25	-27	-40	-2976	2593	-377	-416
xav1sho7	-36	-71	-2778	2442	-336	-407	xav1fx11	-26	-40	-3038	2682	-383	-423
xav1sho8	-34	-71	-2573	2237	-335	-406	xav1fx39	-24	-44	-2835	2454	-395	-439
xav1sho26	-45	-82	-2872	2549	-323	-405	xav1fx8	-22	-37	-2534	2157	-381	-418
							xav1fx53	-21	-32	-2438	2060	-378	-409
AVG	**-57**	**-97**	**-2875**	**2539**	**-335**	**-432**	**AVG**	**-61**	**-82**	**-3196**	**2826**	**-369**	**-451**
SD	**19**	**21**	**191**	**189**	**14**	**16**	**SD**	**25**	**29**	**244**	**250**	**18**	**21**

[a] All energies expressed in kcal mol^{-1}. [b] VDW = Difference between the complex and individual species in Van der Waals component of free energy. [c] NE is the non-electrostatic contribution to the free energy NE = VDW + INT +SUR. [d] ELE = Difference between complex and individual species in electrostatic interactions. [e] PBCAL = electrostatic penalty involved in desolvating the peptide and RNA in complex formation . [f] $\Delta H = \Delta H^{S}_{Bind}$ = NE + PBELE. [g] SDs for each complex are presented in Appendix 7 (Tables A7-6 and A7-7).

While the electrostatic component is a major component of the estimated free energy, it does not vary consistently according to the overall estimated free energy of the complexes as demonstrated by Figure 7-8(A), in which ΔH^{S}_{Bind} is broken down into the electrostatic and non-electrostatic components. Unlike PBELE, the variation in the non-electrostatic component is consistent with the estimated free energy of the complexes. The non-electrostatic energy is decomposed into van der Waals, internal energy and the non-electrostatic component of the solvation energy in Figure 7-8(B). It is clear from this Figure that the van der Waals energy dominates and is ultimately responsible for distinguishing the strength of the binding interface in the complexes.

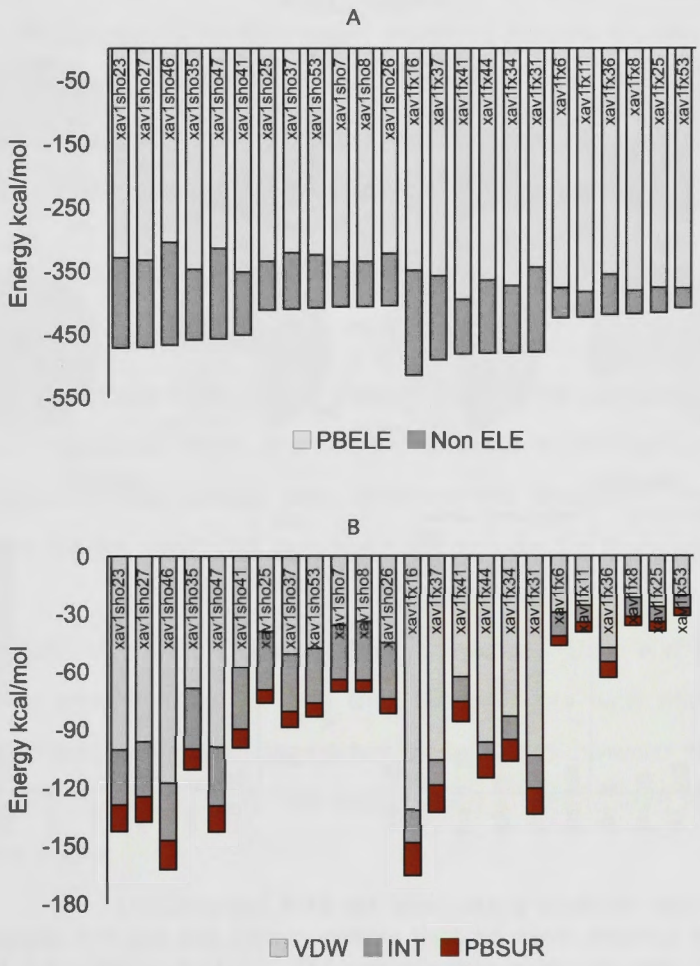


Figure 7-8. Breakdown of interaction component of free energy
(A) ΔH^{S}_{Bind} is broken down into its electrostatic (PBELE) and non-electrostatic (Non-ELE) components. (B) the non-electrostatic component is further broken down into Van der Waals energy (VDW), bond, angle and dihedral angle energies (INT) and the hydrophobic component of the solvation energy (PBSUR).

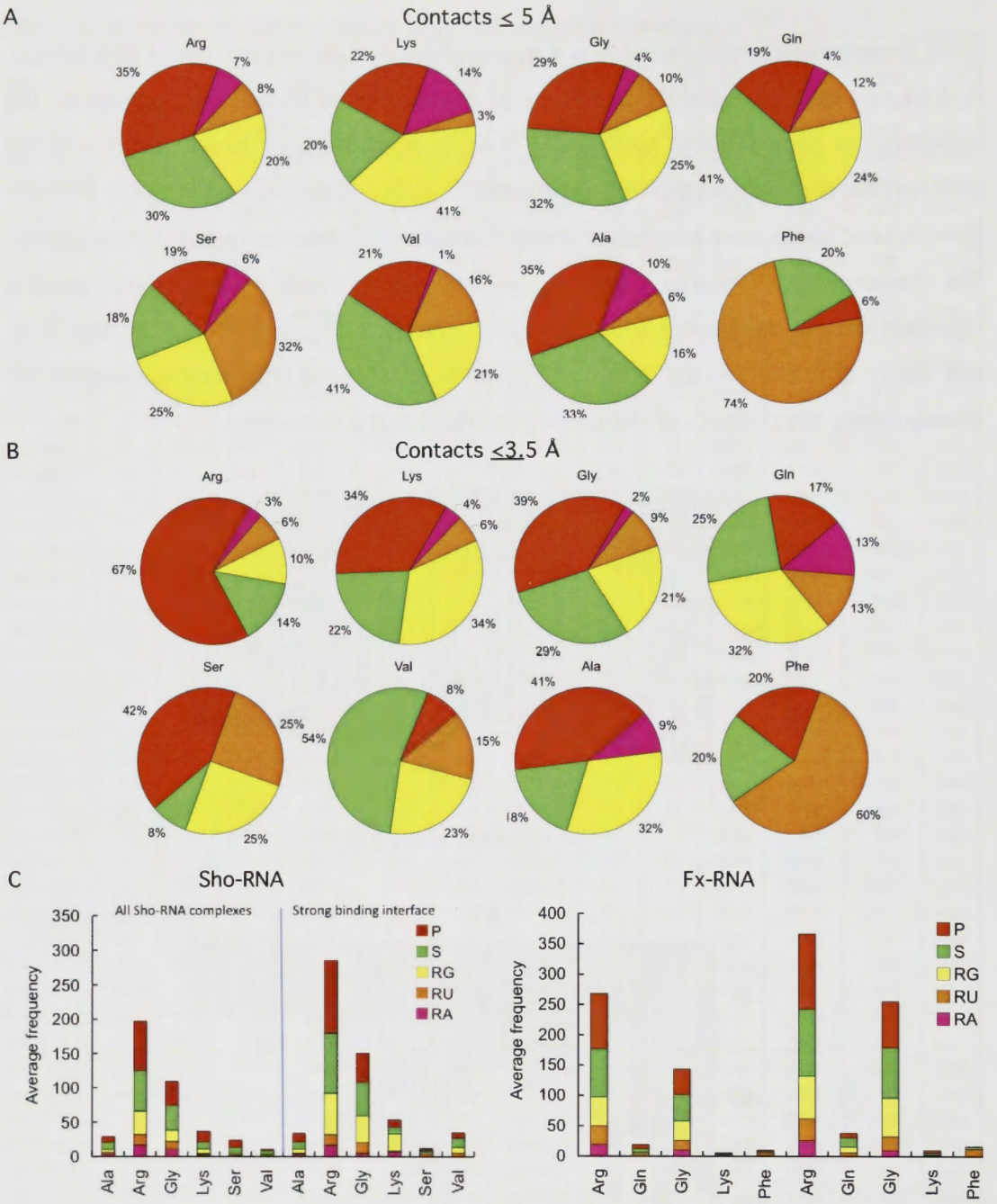


Figure 7-9. Peptide residues preference for RNA components

The percentage of contacts made by each peptide residue and the RNA components in the 10 Sho-RNA and 10 Fx-RNA complexes with greatest ΔH^{S}_{Bind}: (A) Contacts ≤ 5 Å Arg (total number of contacts (n)=6515), Lys (n=627), Gly (n=4051), Gln (n=381), Ser (n=118), Val (n=351), Ala (n=341), Phe (n=162). (B) Contacts ≤ 3.5 Å Arg (total number of contacts (n)=491), Lys (n=50), Gly (n=269), Gln (n=24), Ser (n=12), Val (n=13), Ala (n=22), Phe (n=5). Guanine (yellow) adenine (pink) and uracil (orange), sugars (green) and phosphate (red). (C)The average number of contacts per complex (≤ 3.5 Å) between each type of residue in the peptide and the RNA components are compared based on all complexes (first 6 columns) and then the average number derived from the 10 with the strongest binding interface (last 6 columns).

The MM-PBSA calculations point to the conclusion that complexes with the strongest binding interface (most -ve value of ΔH^{S}_{Bind}) will have a higher number of contacts than other complexes, rather than being distinguished by specific electrostatic contacts. Pie charts representing the number and type of contacts made at the binding interface have been derived from the complexes with the strongest binding enthalpies ΔH^{S}_{Bind} (10 Sho-RNA and 10 Fx-RNA complexes), and are presented in Figure 7-9.

The data presented in Figure 7-9 indicate that in complexes with greatest ΔH^{S}_{Bind}, the proportion of contacts between Arg and the RNA components (within 5 Å) is similar to that found for the entire data set (Figure 6-11 in Chapter 6 and Figure 7-9(C)). However, for the other peptide residue types there is a relative increase in the contacts made to the guanine bases of the RNA, compared to the entire data set. The exception to this is Phe, which primarily makes contact with uracil. It is also notable that for Val contacts with the ribose units of the RNA predominate, consistent with the results for the entire data set.

For the shorter-range contacts (Figures 7-9(B)) there is an increase in contacts between Arg and phosphate in the highly stabilized set (67% compared to 59% for the entire data set). In addition, there is a relative increase in the contacts between the other residue types and the guanine base. Although the number of contacts with Ser, Val and Phe in this set are small, this data has been included for completeness.

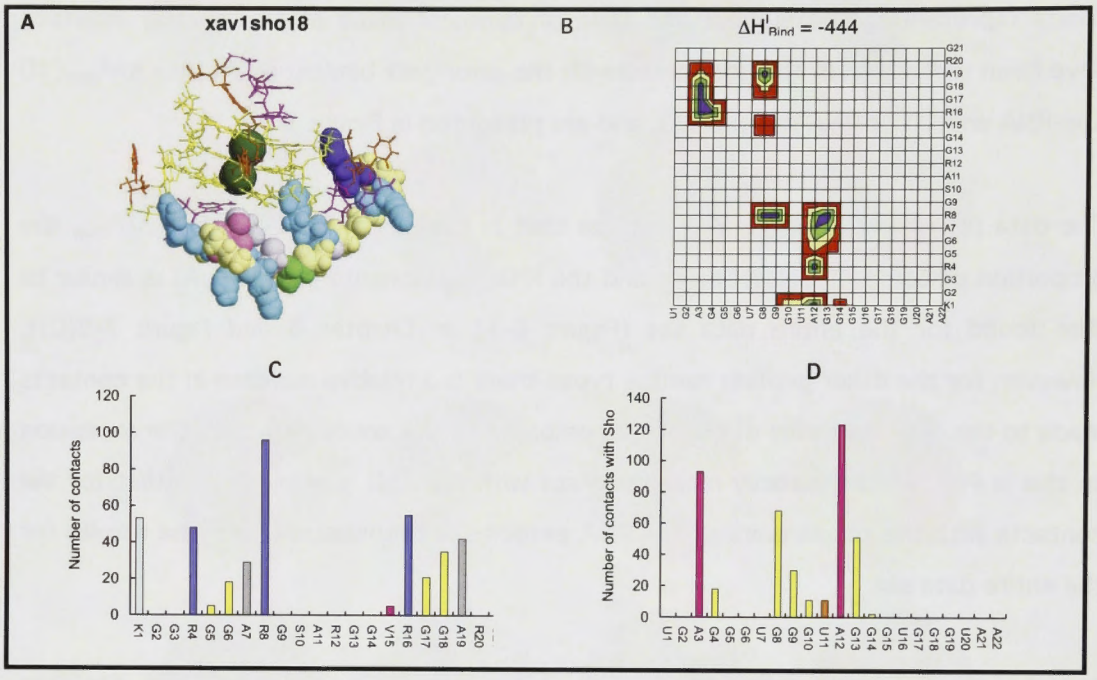
The average number of contacts is compared in Figures 7-9(C) and (D) (rather than percentages) from which it can be seen that the contacts with phosphate are also higher in the complexes with the strongest binding interface, however there is a greater proportional increase in contacts with the bases which is accentuated in the case of the abundant guanine bases.

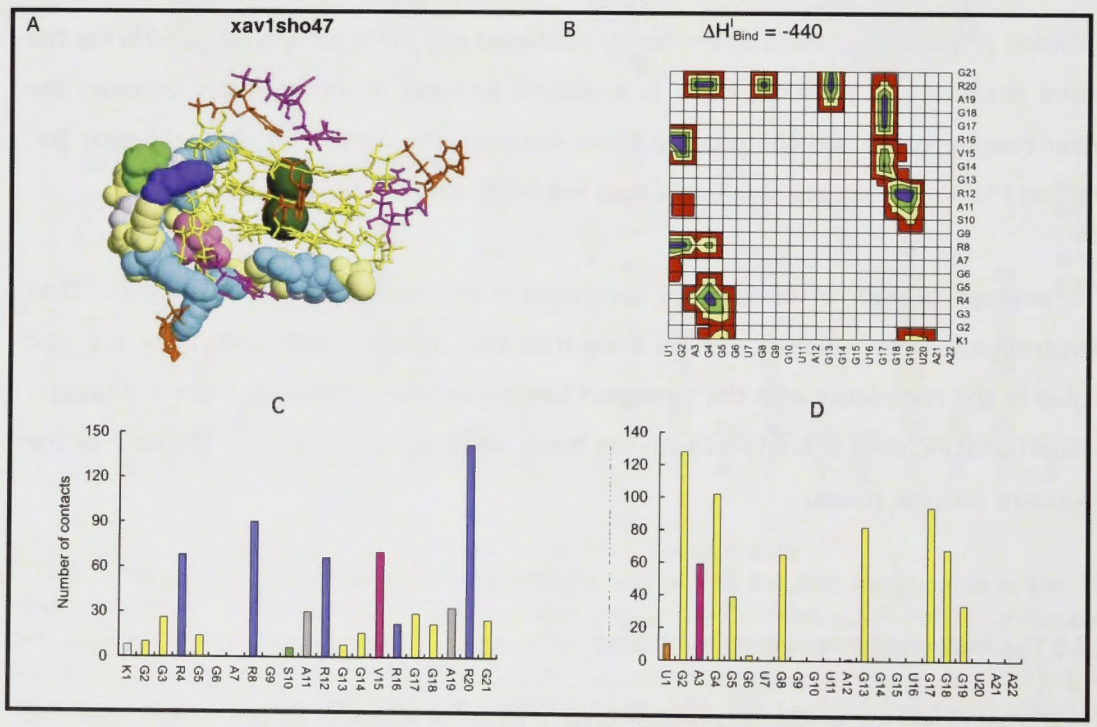
7.3.5 The estimated free energy of binding

Given the similarity in the estimated free energy of binding in the most stabilized complexes, it is informative to look at whether there are similarities in the complexes formed. A description of the binding interface of 2 of the most stabilized complexes of

each system are presented in Figure 7-10, and can be compared to the least stabilized complexes presented in Figure 7-11.

4 complexes with highest estimated free energy





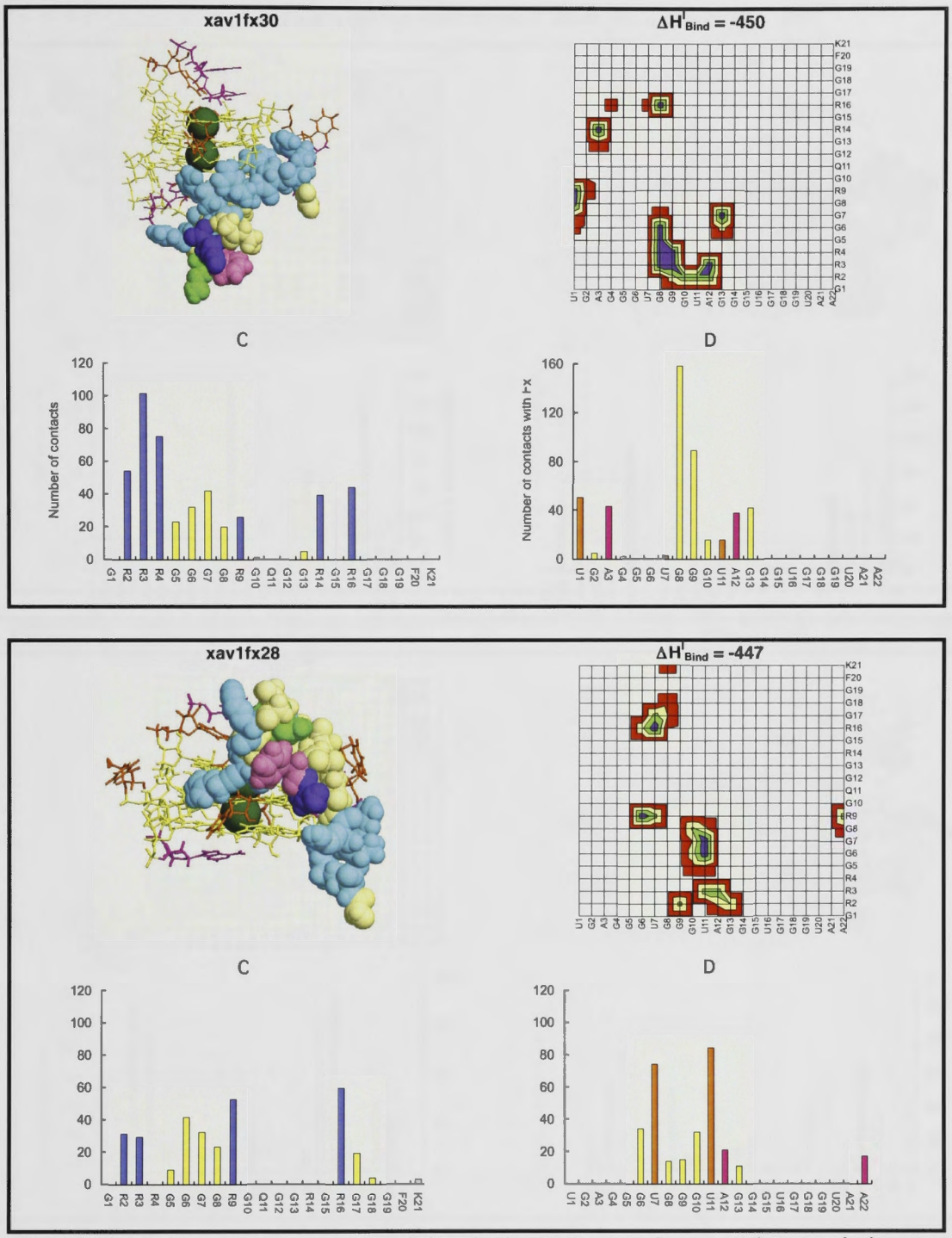
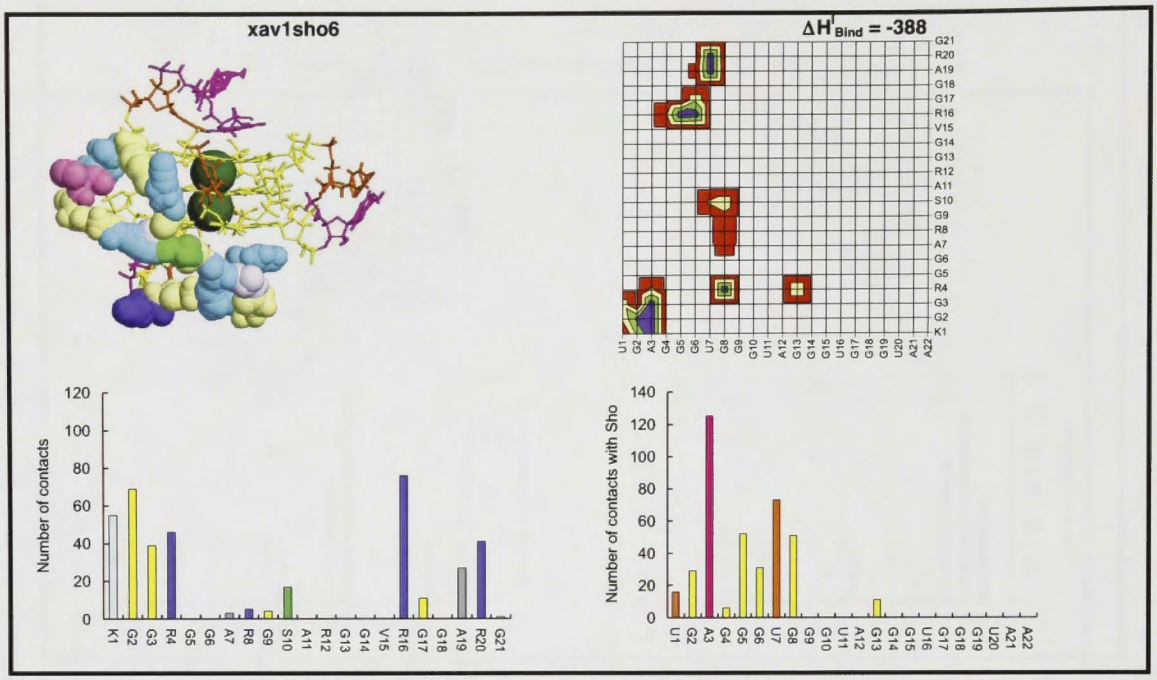
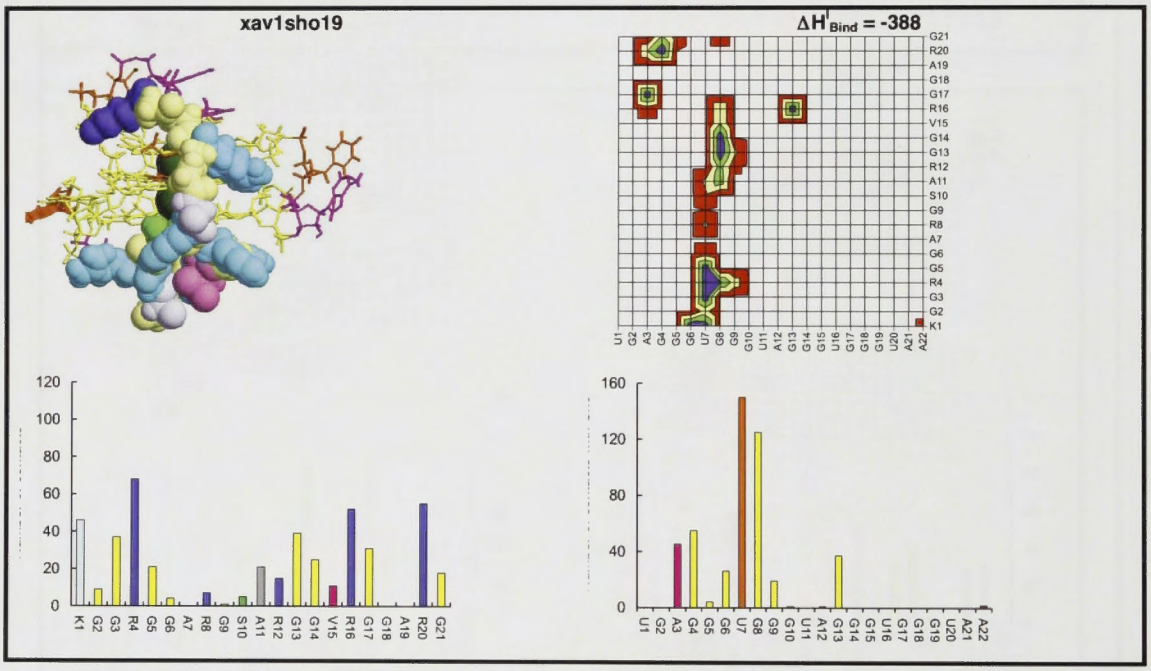


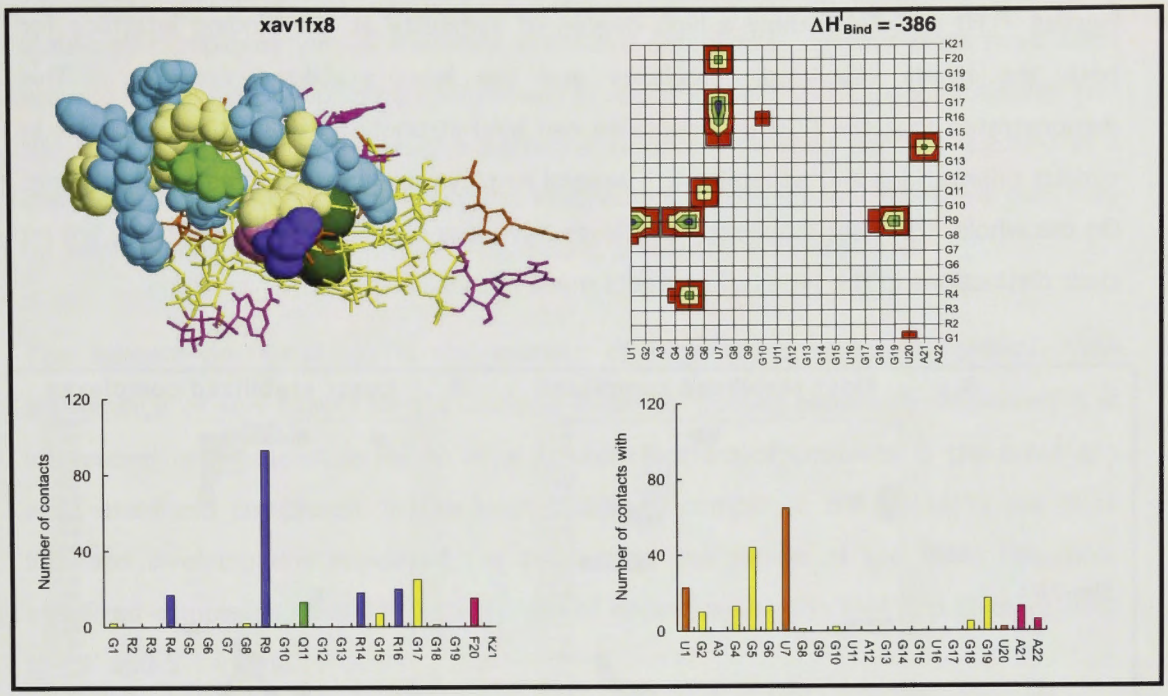
Figure 7-10. Two of the most stabilized Sho-RNA complexes and two of the most stabilized Fx-RNA complexes

Each panel has (A) an image of the average structure. (B) Contour plots showing the contacts made between RNA nucleotides identified along the x-axis and peptide residues on the y-axis, 1-5 contacts (red), 6-10 contacts (yellow), 11-15 contacts (green) >15 contacts (purple). (C) The number of contacts formed by each residue of the peptide. (D) The number of contacts (within 5 Å) formed by each nucleotide of the RNA.

4 complexes with lowest estimated free energy







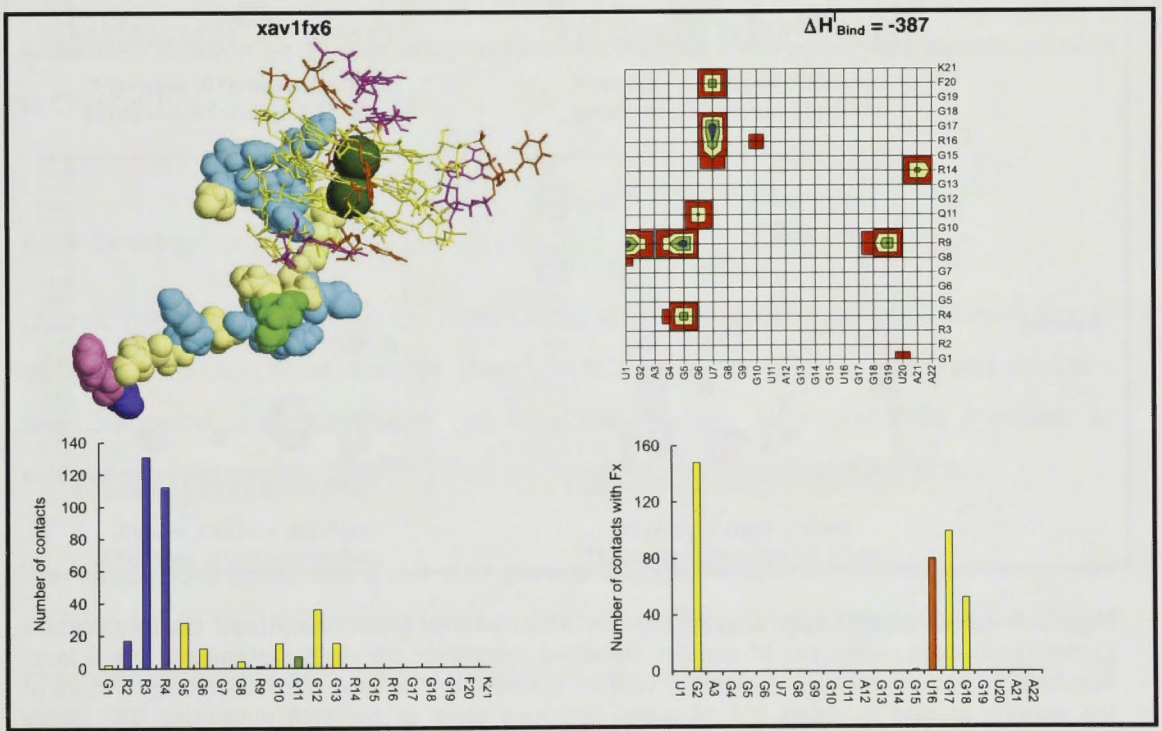


Figure 7-11. Three of the least stabilized Sho-RNA complexes and three of the least stabilized Fx-RNA complexes
See notes to Figure 7-10.

Figures 7-10 and 7-11 show a high degree of variability at the binding interface for both the highly stabilized complexes and the least stabilized complexes. This demonstrates that the RGG box peptides can bind strongly to the RNA quadruplex in various orientations and generally with several Arg/Lys residues participating in binding. On the whole, the least stabilized complexes have less contacts. However, there are no clear distinctions in the types of contacts made in these subsets.

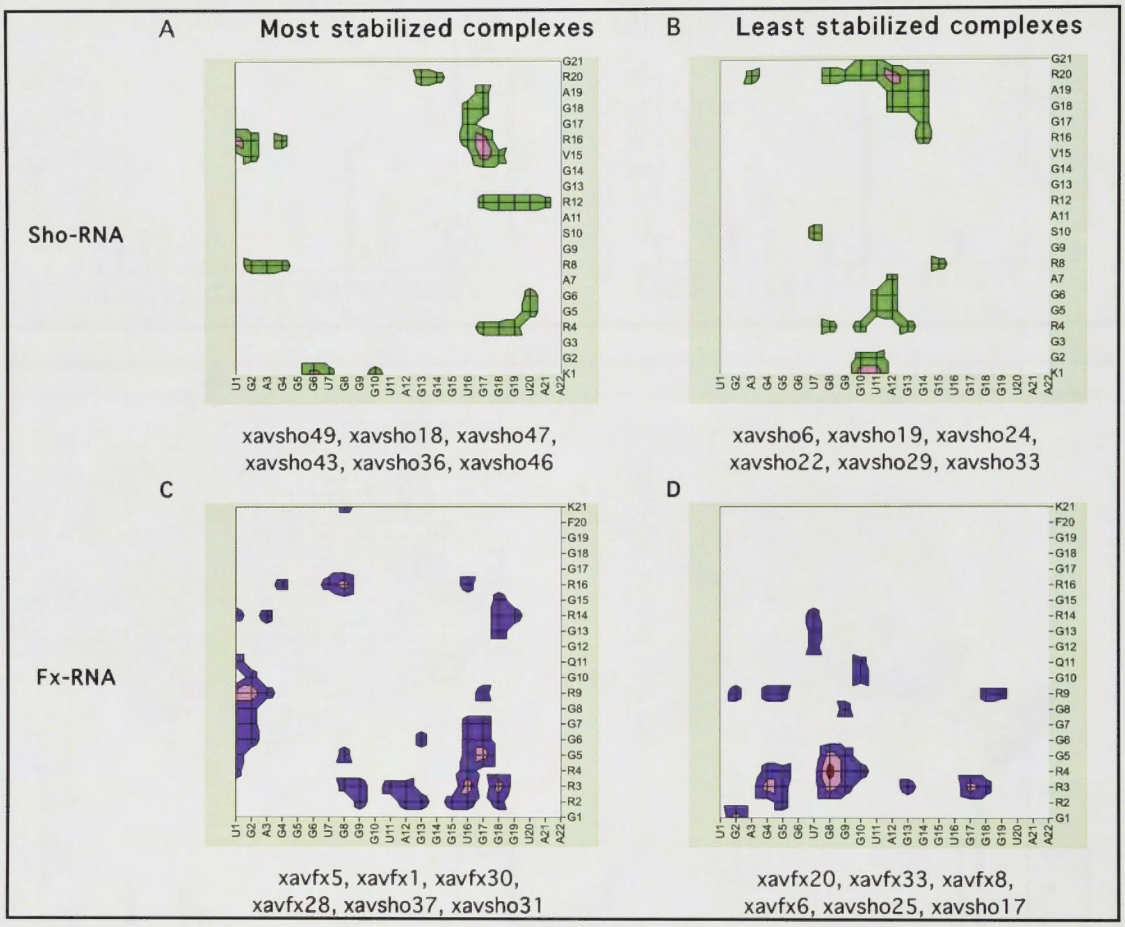


Figure 7-12. Common contacts in most stabilized and least stabilized complexes
Common contacts in the set of 6 most stabilized complexes for each system and the 6 least stabilized complexes for each system. A contact is counted if at least one non-hydrogen atom of the peptide residue is within 5 Å of a non-hydrogen atom of the RNA nucleotide. (A) Highly stabilized Sho-RNA complexes. (B) Least stabilized Sho-RNA complexes.(C) Highly stabilized Fx-RNA complexes. (D) Least stabilized Fx-RNA complexes. White = 0-1, for Sho, green =2, for Fx, purple = 2, pink = 3, red = 4.

To better understand whether there are any interactions that are common in the highly stabilized complexes versus the least stabilized complexes, contour plots have been prepared of the overlapping contacts found in subsets of the most highly stabilized and the least stabilized complexes. If a particular residue-nucleotide contact is found in a complex this is counted as one contact irrespective of the number of atoms that may be involved in making this contact. This data is presented in Figure 7-12.

The subsets are small, as is the number of common contacts, accordingly, the significance of this aspect of the analysis must be treated cautiously. However, it is interesting to see some evidence of a different pattern of contacts in the most and least stabilized complexes. In the least stabilized complexes the contacts are more localised involving one region of the peptide or one region of the RNA. The most stabilized complexes have a broader spread of contacts with the RNA and involves both the 3' and 5' termini.

A detailed description of four complexes of each system are presented as case studies in Chapter 8.

7.3.6 Docking discussion

Overall, this study has shown that there is no single Sho-RNA or Fx-RNA complex that is most likely to be formed. Instead, there are many ways in which the RGG-box peptides and the model RNA quadruplex can bind with roughly equivalent free energies, as calculated using the MM-PBSA method and an independent trajectory approach.

The simulations presented show that several Arg residues in an RGG box domain may interact, both electrostatically and in the formation of hydrogen bonds, with the phosphate backbone of G-quadruplex RNA. Although the electrostatic component of the estimated free energy appears high, there is little variation among complexes, which suggests that the number and type of ionic contacts are similar across the data set, and do not assist in differentiating the docked structures. This raises the interesting question of the role of electrostatic interactions in the binding of a protein domain, particularly a highly charged domain, with a nucleic acid.

One might expect that electrostatic interactions will be important in the binding of a polyanion such as RNA and a highly positively charged peptide such as these RGG box peptides. Clearly such interactions are typical of DNA-protein binding (Record et al., 1991, Record et al., 1998) and RNA-protein interactions (Garcia-Garcia and Draper, 2003). Electrostatic interactions are also known to be crucial in some G-quadruplex-protein binding scenarios. One such example is the 15-mer thrombin aptamer DNA G-quadruplex which is known to bind to thrombin through 4 ion pair interactions between 2 Arg residues and phosphate groups of the DNA quadruplex as well as one hydrogen bond between His and phosphate (Padmanabhan et al., 1993).

However, this analysis is not completely straightforward, as in some cases electrostatic effects contribute very little to the free energy of nucleic acid-protein complexes (Reyes and Kollman, 2000) - outcomes that might be considered counterintuitive. To understand this behaviour it is necessary to appreciate the entire environment in which a nucleic acid - protein interaction occurs. That is, the presence of counterions and co-ions, both local and diffuse, and the surrounding waters.

It has been known for some time that the ion distribution around DNA is not uniform. Rather, there is a concentration of cations ('counterion condensation') close to the polyanionic DNA and a diminution of anions from this region (Manning, 1978, Anderson and Record, 1982). The approach of a cationic binding partner and the subsequent release of cations from the binding interface is entropically favourable and has been termed 'the polyelectrolyte effect' (Anderson and Record, 1982). In a low salt environment the polyelectrolyte effect may be sufficient to drive the binding of a cationic molecule to DNA or RNA in a nonspecific manner (Anderson and Record, 1982). In fact, non-specific binding dominated by electrostatic effects manifests in a salt-dependence of binding. While helical DNA, modeled as a rod or cylinder, has been used to develop these polyelectrolyte theories, it is also know that cations are concentrated around the outside of RNA (Garcia-Garcia and Draper, 2003), and G-quadruplex nucleic acid (Federiconi et al., 2005).

Although the contribution of anions were not considered to be significant in the original work in this area, more recent research confirms that a highly basic protein interface

may attract a corresponding concentration of anions that are released upon binding (Garcia-Garcia and Draper, 2003, Hargreaves and Schleif, 2008).

The emerging picture is one of a complex interplay of enthalpic and entropic effects mediated by locally condensed ions in a more diffuse ionic environment (Wong and Pollack, 2009). The direct effect of local ions is not considered in a mean field approach to the calculation of electrostatics, such as the Poisson-Boltzmann method used in this study. Unfortunately, the deficiencies of the mean field approach are exacerbated in situations with multivalent ions or in high-surface charge densities, as occurs in the highly charged small RGG-box peptides (Wong and Pollack, 2009).

These simulations were conducted with explicit waters surrounding the RNA and peptide and only sufficient counterions to neutralize the charge of the system. Counterions and co-ions necessary to mimic a physiological buffer were not included in the simulation. Furthermore, the Poisson-Boltzmann method used to calculate electrostatic effects ignores the effects of explicit waters and ions. This has led to a systematic overestimation of PBELE, and accordingly an overestimation of the free energy of binding.

An unrealistically high PBELE may indicate that the stabilization induced by complex formation is overestimated and/or the penalty involved in desolvating the monomers is underestimated. Generally, it has been observed that the desolvation penalty is high in these scenarios, and thus offsets the stabilizing effects of the interaction. This has been observed in both RNA-peptide interactions (Reyes and Kollman, 2000) and for protein salt-bridges (Hendsch and Tidor, 1994). It seems likely that PBCAL has been underestimated in this study as the stabilization of the RNA and peptide monomers by local counterions is not taken into account.

The overestimation of PBELE has produced free energies that are much greater than would be expected in a physical system. Other studies have recognised that calculation of the electrostatic solvation free energy is a major source of error in MM-PBSA, particularly in the case of large charged systems (Zhang and Schlick, 2006, Grochowski and Trylska, 2008, Prabhu et al., 2008).

Although calculation of the absolute free energy is problematic the assessment of relative energies using these methods has proved to be reliable (Zhang and Schlick, 2006). The current investigation is essentially a comparative study in which ensembles of Sho and Fx complexes are compared with the aim of identifying likely binding modes. The validity of this approach has been confirmed by the experimental studies that will be discussed in Chapter 9.

7.4 CONCLUSIONS

Many of the RGG box complexes identified in this study are thermodynamically indistinguishable. This suggests that these inherently flexible peptides may use their plasticity to bind in different conformations to a suitable target or potentially to several different targets. This makes the RGG box motif a versatile RNA-binding platform, similar to the RNA recognition motif (Maris et al., 2005). These data also support the more general suggestion that an Arg rich motif provides a flexible framework for recognition of RNA structures (Harada et al., 1997).

The Sho and Fx peptides do not favour any particular structure at the binding interface. Sometimes the peptide binds in an extended form while in other instances there may be turns that are well maintained to form a loop in the peptide. Where elements of secondary structure, such as turns, are found in complexes they are not necessarily involved in the binding interface, as will be seen more clearly in the case studies presented in Chapter 8. This description of the RGG box motif is consistent with the idea of a 'fuzzy binding interface' advanced by Tompa and Fuxreiter to explain the way in which disordered regions of proteins may bind in more than one way to a partner thus allowing it to 'moonlight' or engage in a number of different functional roles (Tompa and Fuxreiter, 2008).

While the peptides are very flexible, the RNA is highly structured. The 8 Hoogsteen hydrogen bonds involved in the formation of each guanine tetrad are all highly occupied with little difference in occupancy regardless of the degree of RNA de/stabilization (ΔH_{RNA}). The main factor that leads to relative destabilization of the RNA is the density

of contacts around the tetrad, primarily those made with the sugar-phosphate backbone of the nucleotides involved in the tetrad stack. By contrast, contacts between the peptide and the loops tend to be stabilizing. Overall, it appears that the RNA quadruplex is likely to be stabilized in a 1:1 complex with the RGG box peptides.

Comparison of the Sho and Fx complexes reveals a similar distribution of binding free energies for both sets of complexes. Although the electrostatic contribution to the free energy is generally a little higher in the Fx complexes as a result of an additional positively charged residue, there is a similar range of non-electrostatic energies, which are the main determinant of the relative stability of the complexes. These similarities point to a likely functional role for the RGG box motif in Sho.

7.5 REFERENCES

Anderson, C.F. & Record, M.T. (1982) Poly-Electrolyte Theories and Their Applications to DNA. *Annual Review of Physical Chemistry,* 33, 191-222.

Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B. & Woods, R.J. (2005) The Amber biomolecular simulation programs. *J Comput Chem,* 26, 1668-88.

Connolly, M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science,* 221, 709-13.

Donini, O.A. & Kollman, P.A. (2000) Calculation and prediction of binding free energies for the matrix metalloproteinases. *J Med Chem,* 43, 4180-8.

Erlitzki, R. & Fry, M. (1997) Sequence-specific binding protein of single-stranded and unimolecular quadruplex telomeric DNA from rat hepatocytes. *J Biol Chem,* 272, 15881-90.

Fan, J.H., Bochkareva, E., Bochkarev, A. & Gray, D.M. (2009) Circular dichroism spectra and electrophoretic mobility shift assays show that human replication protein A binds and melts intramolecular G-quadruplex structures. *Biochemistry,* 48, 1099-111.

Federiconi, F., Ausili, P., Fragneto, G., Ferrero, C. & Mariani, P. (2005) Locating counterions in guanosine quadruplexes: a contrast-variation neutron diffraction experiment in condensed hexagonal phase. *J Phys Chem B,* 109, 11037-45.

Fry, M. (2007) Tetraplex DNA and its interacting proteins. *Front Biosci,* 12, 4336-51.

Garcia-Garcia, C. & Draper, D.E. (2003) Electrostatic interactions in a peptide--RNA complex. *J Mol Biol,* 331, 75-88.

Grochowski, P. & Trylska, J. (2008) Continuum molecular electrostatics, salt effects, and counterion binding--a review of the Poisson-Boltzmann theory and its modifications. *Biopolymers,* 89, 93-113.

Harada, K., Martin, S.S., Tan, R. & Frankel, A.D. (1997) Molding a peptide into an RNA site by in vivo peptide evolution. *Proc Natl Acad Sci U S A,* 94, 11887-92.

Hargreaves, V.V. & Schleif, R.F. (2008) The salt dependence of the interferon regulatory factor 1 DNA binding domain binding to DNA reveals ions are localized around protein and DNA. *Biochemistry,* 47, 4119-28.

Hendsch, Z.S. & Tidor, B. (1994) Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci,* 3, 211-26.

Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D.A. & Cheatham, T.E., 3rd

(2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res,* 33, 889-97.

Luo, R., David, L. & Gilson, M.K. (2002) Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J Comput Chem,* 23, 1244-53.

Manning, G.S. (1978) Molecular Theory of Polyelectrolyte Solutions with Applications to Electrostatic Properties of Polynucleotides. *Quarterly Reviews of Biophysics,* 11, 179-246.

Maris, C., Dominguez, C. & Allain, F.H. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *Febs J,* 272, 2118-31.

Menon, L., Mader, S.A. & Mihailescu, M.R. (2008) Fragile X mental retardation protein interactions with the microtubule associated protein 1B RNA. *Rna,* 14, 1644-55.

Menon, L. & Mihailescu, M.R. (2007) Interactions of the G quartet forming semaphorin 3F RNA with the RGG box domain of the fragile X protein family. *Nucleic Acids Res,* 35, 5379-92.

Nifosi, R., Reyes, C.M. & Kollman, P.A. (2000) Molecular dynamics studies of the HIV-1 TAR and its complex with argininamide. *Nucleic Acids Res,* 28, 4944-55.

Oganesian, L. & Bryan, T.M. (2007) Physiological relevance of telomeric G-quadruplex formation: a potential drug target. *Bioessays,* 29, 155-65.

Padmanabhan, K., Padmanabhan, K.P., Ferrara, J.D., Sadler, J.E. & Tulinsky, A. (1993) The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer. *J Biol Chem,* 268, 17651-4.

Paramasivam, M., Membrino, A., Cogoi, S., Fukuda, H., Nakagama, H. & Xodo, L.E. (2009) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription. *Nucleic Acids Res.*

Prabhu, N.V., Panda, M., Yang, Q. & Sharp, K.A. (2008) Explicit ion, implicit water solvation for molecular dynamics of nucleic acids and highly charged molecules. *J Comput Chem,* 29, 1113-30.

Record, M.T., Jr., Ha, J.H. & Fisher, M.A. (1991) Analysis of equilibrium and kinetic measurements to determine thermodynamic origins of stability and specificity and mechanism of formation of site-specific complexes between proteins and helical DNA. *Methods Enzymol,* 208, 291-343.

Record, M.T., Jr., Zhang, W. & Anderson, C.F. (1998) Analysis of effects of salts and uncharged solutes on protein and nucleic acid equilibria and processes: a practical guide to recognizing and interpreting polyelectrolyte effects, Hofmeister effects, and osmotic effects of salts. *Adv Protein Chem,* 51, 281-353.

Reyes, C.M. & Kollman, P.A. (2000) Structure and thermodynamics of RNA-protein binding: using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change. *J Mol Biol,* 297, 1145-58.

Reyes, C.M., Nifosi, R., Frankel, A.D. & Kollman, P.A. (2001) Molecular dynamics and binding specificity analysis of the bovine immunodeficiency virus BIV Tat-TAR complex. *Biophys J,* 80, 2833-42.

Srinivasan, J., Miller, J., Kollman, P.A. & Case, D.A. (1998) Continuum solvent studies of the stability of RNA hairpin loops and helices. *J Biomol Struct Dyn,* 16, 671-82.

Tompa, P. & Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci,* 33, 2-8.

Tsui, V. & Case, D.A. (2001) Calculations of the absolute free energies of binding between RNA and metal ions using molecular dynamics simulations and continuum electrostatics. *Journal of Physical Chemistry B,* 105, 11314-11325.

Weisman-Shomer, P. & Fry, M. (1994) Stabilization of tetrahelical DNA by the quadruplex DNA binding protein QUAD. *Biochem Biophys Res Commun,* 205, 305-11.

Wong, G.C. & Pollack, L. (2009) Electrostatics of Strongly Charged Biological Polymers: Ion-Mediated Interactions and Self-Organization in Nucleic Acids and Proteins. *Annu Rev Phys Chem.*

Zhang, Q. & Schlick, T. (2006) Stereochemistry and position-dependent effects of carcinogens on TATA/TBP binding. *Biophys J,* 90, 1865-77.

CHAPTER 8

EIGHT CASE STUDIES ILLUSTRATING THE BINDING OF RGG BOX PEPTIDES TO AN RNA G-QUADRUPLEX

8.1 INTRODUCTION

8.1.1 Background

A comprehensive discussion of the docking and MD simulation of 49 Sho-RNA and 52 Fx-RNA complexes is presented in Chapters 6 and 7. Six of the most tightly bound complexes of each system were allowed to continue simulation for a total of 56 ns per complex. A subset of 8 complexes (4 Sho and 4 Fx) are described here in detail to demonstrate the variability seen in these complexes. For each system, one of the case studies has docking induced RNA quadruplex instability. All of the models presented have a strong binding interface (high ΔH^{S}_{Bind}) and are within the group of complexes with the greatest estimated free energy of binding. These specific examples of the docked structures both expand upon and illustrate principles presented in Chapters 6 and 7.

8.2 RESULTS AND DISCUSSION

8.2.1 56 ns simulations

Overall, there is not a lot of change in the binding interface over the period 25-56 ns. After 25 ns of simulation many stabilizing contacts are established and the complex is not as adjustable as in the early stages of the simulation as illustrated by contact plots, calculated between 0-1 ns and 50-56 ns of simulation, in the Figures which follow. Nevertheless, there are changes over the second half of the simulation that are significant enough to change the ranking of the complexes in terms of their estimated free energy and the strength of the binding interface. Free energy calculations of 4 Sho-RNA and 4 Fx-RNA complexes performed after the 56ns simulations are compared with those performed after 25 ns simulations in Table 8-1. The complexes are arranged in order of the non-electrostatic component of the free energy as assessed over the last half of the 56 ns simulation.

Table 8-1 Free energy calculations from 5-25ns and 25-56 ns for 8 complexes [a,b,c]							
	ΔH^{I}_{bind}	ΔH_{RNA}	$\Delta H_{Sho/Fx}$	ΔH^{S}_{bind}	PBELE	VDW	NE
xav1sho47	-440	-1	18	-457	-315	-99	-142
xav1sho47	-425	12	28	-465	-312	-108	-154
xav1sho23	-399	48	25	-472	-329	-100	-143
xav1sho23	-417	39	20	-476	-324	-106	-152
xav1sho46	-435	17	14	-466	-305	-118	-162
xav1sho46	-464	14	12	-490	-325	-120	-136
xav1sho27	-427	23	23	-473	-342	-88	131
xav1sho27	-416	26	24	-466	-345	-80	-121
xav1fx16	-431	34	50	-515	-350	-132	-165
xav1fx16	-463	18	42	-523	-375	-120	-148
xav1fx31	-426	4	49	-479	-345	-104	-134
xav1fx31	-456	-6	56	-506	-371	-103	-135
xav1fx37	-445	-10	56	-491	-366	-96	-121
xav1fx37	-433	3	47	-483	-356	-100	-128
xav1fx41	-422	17	43	-482	-395	-63	-87
xav1fx41	-438	12	61	-511	-401	-84	-110

[a] The results of MM-PBSA calculations based on 5–25 ns data (white) compared to calculations based on 25-56 ns data (green for Sho-RNA complexes and lavender for Fx-RNA complexes), [b] all energies are in kcal mol^{-1}, [c] all energy and enthalpy measures are defined in Chapter 7 and briefly recapped here: ΔH^{I}_{bind} = enthalpy of complex formation with flexible monomers, ΔH_{RNA} = enthalpy of the RNA de/stabilization upon complex formation, $\Delta H_{Sho/Fx}$ = enthalpy of peptide stabilization/destabilization upon complex formation, ΔH^{S}_{bind} = enthalpy of complex formation with rigid monomers, PBELE = Electrostatic energy of interaction including electrostatic penalty for desolvating monomers, VDW = difference in van der Waals component of free energy upon binding, NE = non-electrostatic contribution to the free energy of binding (incl. VDW)

Comparison of free energy calculations performed on data collected from 5-25 ns and 31-56ns reveals flexibility in the complexes. This results in changes in the binding interface and changes in the de/stabilization of the monomers. In some cases the estimated free energy of the complex (ΔH^{I}_{bind}) is greater over the second half of the simulation while in others it is less. The change in ΔH^{I}_{bind} ranges from a decrease in xav1sho47 of 15 kcal mol^{-1} (5%) to an increase in xav1fx16of - 32 kcal mol^{-1} (4%). This degree of variation in estimated free energy over time is of the same order of magnitude as the differences seen across the 8 complexes (-432 ± 15 kcal mol^{-1}).

Generally, the electrostatic component of the free energies is higher in the Fx complexes compared to Sho, however, there is a similar spread of non-electrostatic energies (NE) as shown in Figure 8-1(A). The values of NE based on 25-56 ns are plotted as bars and the values of NE calculated at 5-25 ns are marked as black lines to show the difference between the first and second half of the simulation. The relative differences amongst the 8 complexes, in estimated binding free energy (ΔH^{I}_{bind} 'HI'), the interface component of the binding free energy (ΔH^{S}_{bind} 'HS)', the electrostatic component (PBELE), non-electrostatic component (NE) and Van der Waals contribution (VDW) of the free energy are plotted in Figure 8-1(C).

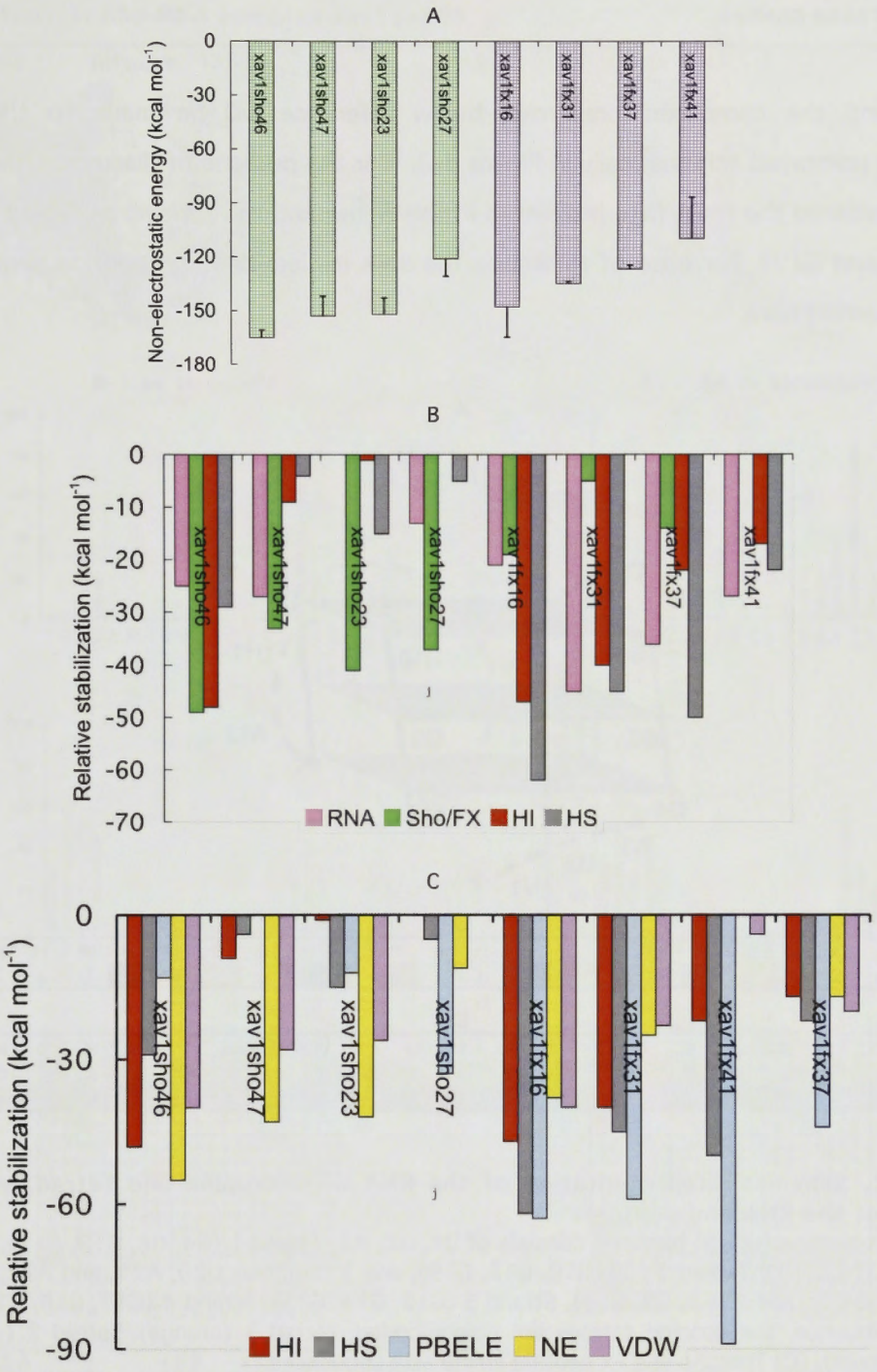
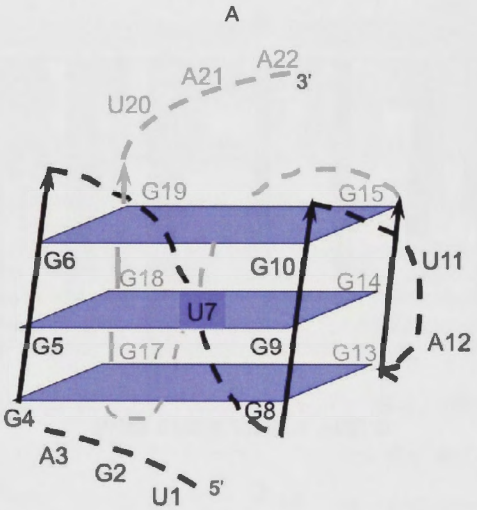


Figure 8-1. Binding interaction energy decomposition for all 8 complexes
(A) Non-electrostatic component of the interaction free energy for 4 Sho-RNA complexes (green) and 4 Fx-RNA complexes (lavender) from 25-56 ns. Values calculated from 5-25ns are indicated by black bars, (B) Differences in the estimated binding energy ΔH^{I}_{bind} (HI), and its components ΔH^{S}_{bind} (HS), ΔH_{RNA} (RNA) and $\Delta H_{Sho/Fx}$ (Sho/Fx), relative to the least negative value, (C) Differences in ΔH^{I}_{Bind} (HI) and components of ΔH^{S}_{Bind} (HS): PBELE (electrostatic), NE (non-electrostatic) and VDW (van der Waals, also included in NE).

8.2.2 Eight case studies

In discussing the complexes presented below reference will be made to the RNA quadruplex portrayed schematically in Figure 8-2. For the purpose of discussion the front will be considered the front face presented in this figure and the top will be tetrad 3 (G6, G10, G15 and G19). For ease of reference the RNA nucleotides and peptide sequences are re-presented here.



B

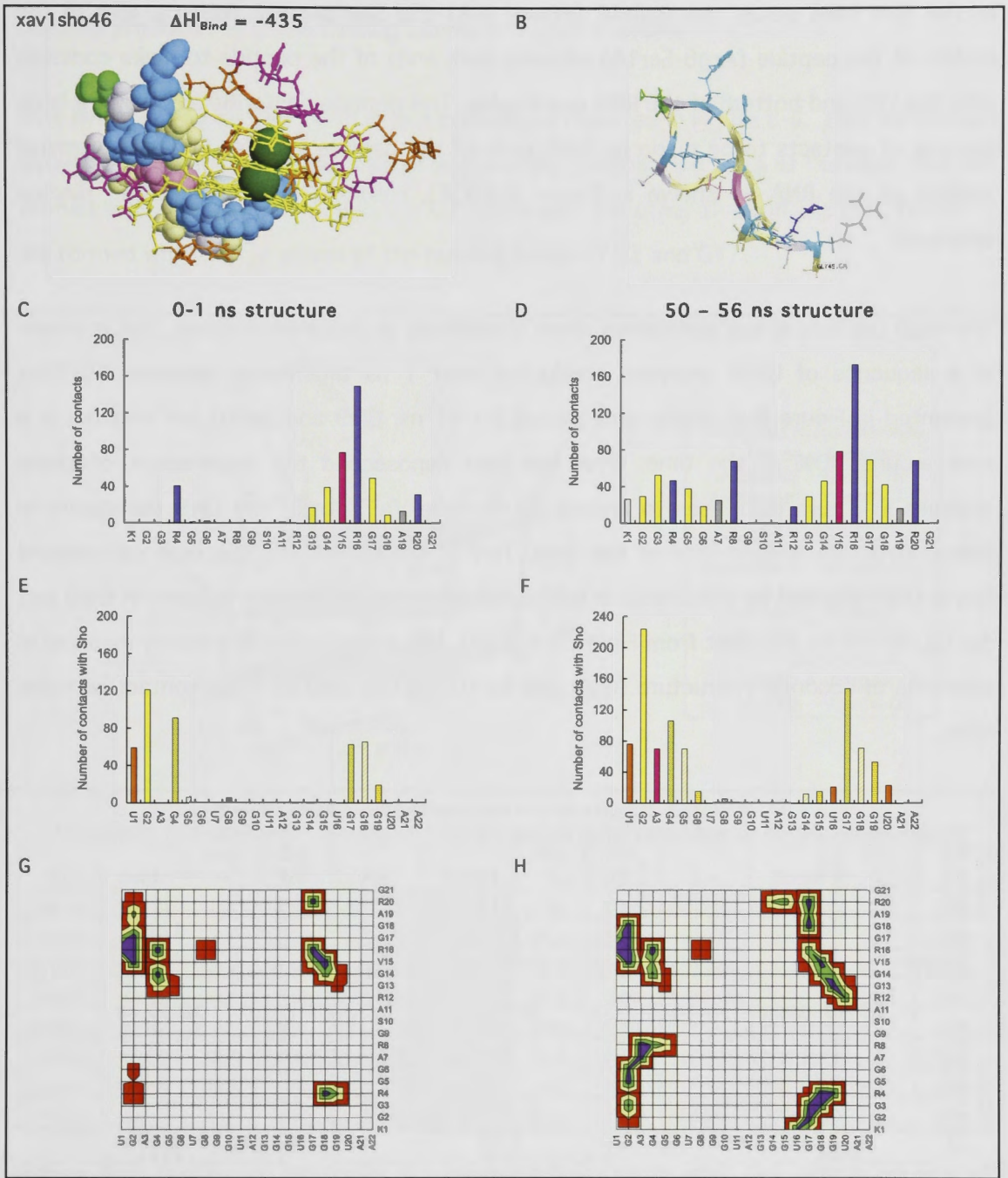
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
RNA	U	G	A	G	G	G	U	G	G	G	U	A	G	G	G	U	G	G	G	U	A	A

C

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Sho	K	G	G	R	G	G	A	R	G	S	A	R	G	G	V	R	G	G	A	R	G
Fx	G	R	R	R	G	G	G	G	R	G	Q	G	G	R	G	R	G	G	G	F	K

Figure 8-2. Schematic representation of the RNA G-quadruplex and tetrad and the sequence of the RNA and peptides
(A) The RNA quadruplex - 5' terminus consists of U1, G2, A3, Tetrad 1 (G4, G8, G13, G17), Tetrad 2 (G5, G9, G14, G18), Tetrad 3 (G6, G10, G15, G19), and 3' terminus U20, A21, and A22, Strand 1 (G4, G5, G6), Strand 2 (G8, G9, G10), Strand 3 (G13, G14, G15), Strand 4 (G17, G18, G19). (B) The RNA sequence, the guanine tetrads are colour coded, tetrad 1 (orange), tetrad 2 (yellow), tetrad 3 (brown). (C) The Sho and Fx peptide amino acid sequences.

Case study 1: Sho-RNA complex xav1sho46



≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å	≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å
20	428	2	48	838	11

Figure 8-3. Case study 1: Sho-RNA complex xav1sho46

(A) Average structure over 50-56 ns, (B) RNA hidden to show the peptide. (C-F) RNA-peptide contacts in the average structure, calculated at 0-1 ns and 50-56 ns. (C-D) Number of contacts ≤ 5 Å made by each peptide residue. (E-F) Number of contacts ≤ 5 Å made by each RNA nucleotide (G-H) Contour plots of contacts between Nucleotides (x-axis) and peptide residues (y-axis): 1-5 (red), 6-10 (yellow), 11-15 (green), >15 (purple). At the base of the figure: total number of contacts ≤ 3.5 Å, ≤ 5 Å and salt bridges (SB) formed by the charged nitrogen atoms of the Arg and Lys sidechains and the O1P and O2P atoms of the RNA phosphate groups.

In the first case study, xav1sho46 (Figure 8-3) the Sho peptide forms a turn in the middle of the peptide (Arg8-Ser10) allowing both ends of the peptide to make contacts with the LHS and bottom of the RNA quadruplex. This peptide conformation allows a large number of contacts to be made by both ends of the peptide and the 5' and 3' terminal regions of the RNA as shown in Figure 8-3(C-F), resulting in a very strong binding interface.

Although the turn is well maintained there is plasticity in this conformation. This is shown in a sequence of DSSP analyses conducted over 1 ns timeframes between 50-56ns presented in Figure 8-4. In the time period 50–51 ns, Gly9 and Ser10 are involved in a turn around 90% of the time. Over the next nanosecond the involvement of these residues in a turn decreases to around 60 % while Val7, Arg8, and Gly9 participate in helical structure around 30% of the time. This is maintained over the next nanosecond but is then followed by a decrease in helical structure and an increase in turns at Gly9 and Ser10. As will be apparent from Figure 8-3 (C-D), the residues that are mainly involved in elements of secondary structure (Gly9 and Ser10) do not tend to make contact with the RNA.

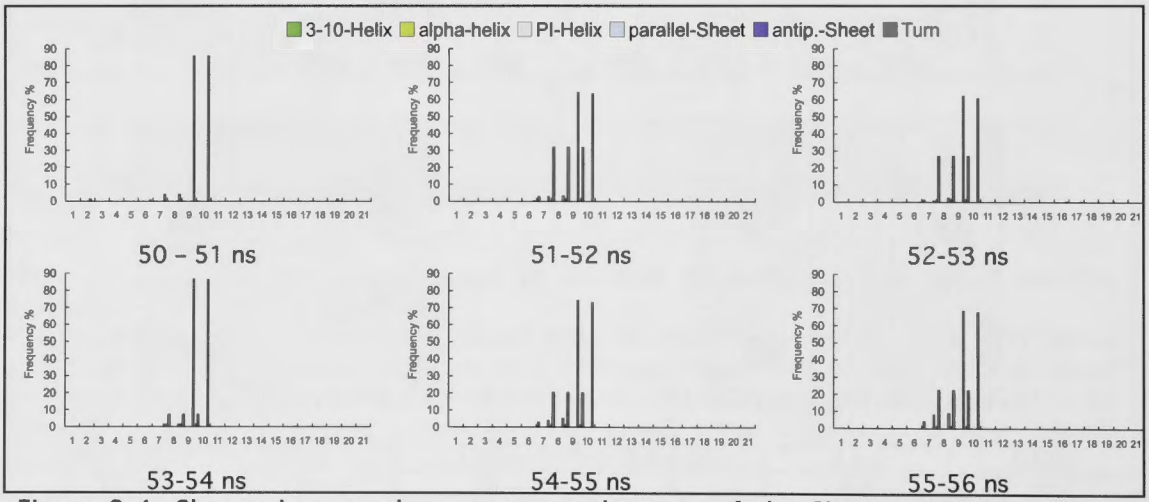
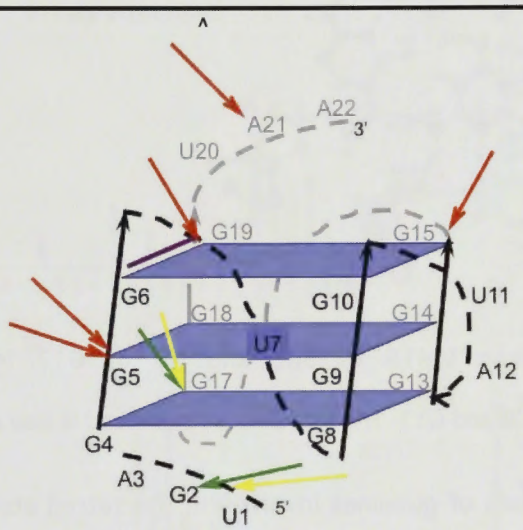


Figure 8-4. Change in secondary structure elements of the Sho peptide in complex xav1sho46 over time, according to DSSP analysis

The C-terminus end of the peptide, makes more contact with the RNA initially, primarily through Arg16 (Figure 8-3(C)). However, as the simulation progresses contacts are built

with the N-terminus end of the peptide. By the end of the simulation all Arg and Lys residues are involved in the binding interface (Figure 8-3(D)).

Nine hydrogen bonds are found in this complex as depicted in Figure 8-5. Five are formed between the guanidinium sidechain of Arg and the phosphate groups of the RNA. Two are formed with ribose, one through the O2' atom and the other through the O4'. Finally, 2 are formed with the N2 atoms of the guanine bases of G2 and G17.



Phosphate hydrogen bonds	
Donor	**Acceptor**
Arg4@NE	G19@O1P
Arg8@NH2	G5@O1P
Arg8@NH1	G5@O2P
Arg12@NH2	A21@O1P
Arg20@NH2	G15@O1P
Ribose hydrogen bonds	
G17@O2'	Gly2@O
Gly17@N	G2@O4'
Base hydrogen bonds	
G17@N2	Val15@O
G2@N2	Val15@O

Hoogsteen hydrogen bond occupancy over the last 20 ns of simulation of the 56 ns simulation

Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)
:4@O6	:17@N1	100	2.92	:5@N7	:18@N2	100	2.95	:6@N7	:19@N2	**97.3**	**3.02**
:4@N7	:17@N2	100	2.99	:5@O6	:18@N1	99.2	3.08	:6@O6	:19@N1	97.7	2.99
:8@N7	:4@N2	99.9	2.96	:9@N7	:5@N2	99.8	2.99	:10@N7	:6@N2	99.7	2.98
:8@O6	:4@N1	99.5	3.00	:9@O6	:5@N1	98.1	3.08	:10@O6	:6@N1	99.3	2.93
:13@N7	:8@N2	99.5	3.01	:14@N7	:9@N2	99.9	2.95	:15@N7	:10@N2	99.7	2.99
:13@O6	:8@N1	99.7	2.94	:14@O6	:9@N1	99.4	3.06	:15@O6	:10@N1	100	2.92
:17@N7	:13@N2	98.1	3.06	:18@N7	:14@N2	100	2.96	:19@N7	:15@N2	99.9	2.95
:17@O6	:13@N1	100	2.92	:18@O6	:14@N1	97.9	3.11	:19@O6	:15@N1	99.2	2.91

Figure 8-5. Hydrogen bonds in the xav1sho46 complex and the affect on tetrad stability

(A) Schematic representation of the peptide-RNA hydrogen bonds occupied at least 70% of the time over the last 20 ns of the 56 ns simulation. The red, green and yellow arrows represent phosphate, ribose and base hydrogen bonds, respectively. Instability in a Hoogsteen hydrogen bond (defined here as < 97.5% occupancy) is indicated by a purple line. (B) Tetrad hydrogen bond occupancy; tetrad 1 (yellow), tetrad 2 (green), tetrad 3 (orange) the occupancy (frequency, 'Occ') of each Hoogsteen hydrogen bond over the last 20 ns of the 56 ns simulation and the average hydrogen bond length (Dist) over this period are given.

It is interesting that the backbone carbonyl group of Val15 makes a permanent hydrogen bond with the N2 atom of G17 (Occ = 100%), without interfering with the Hoogsteen bonds made to the N7 atom of G4 (Figure 8-6). In fact, this interaction appears to be stabilizing, with 100 % occupancy of the G4@N7:G17@N2 bond.

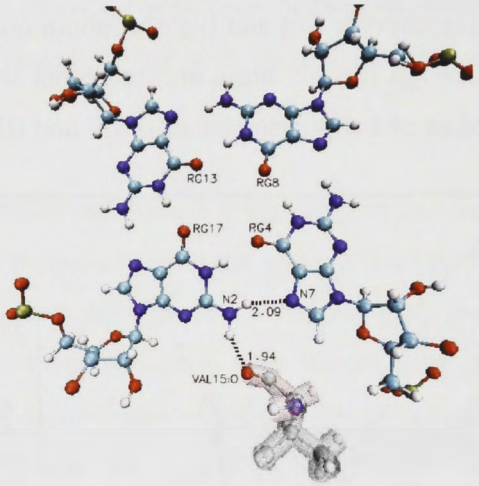
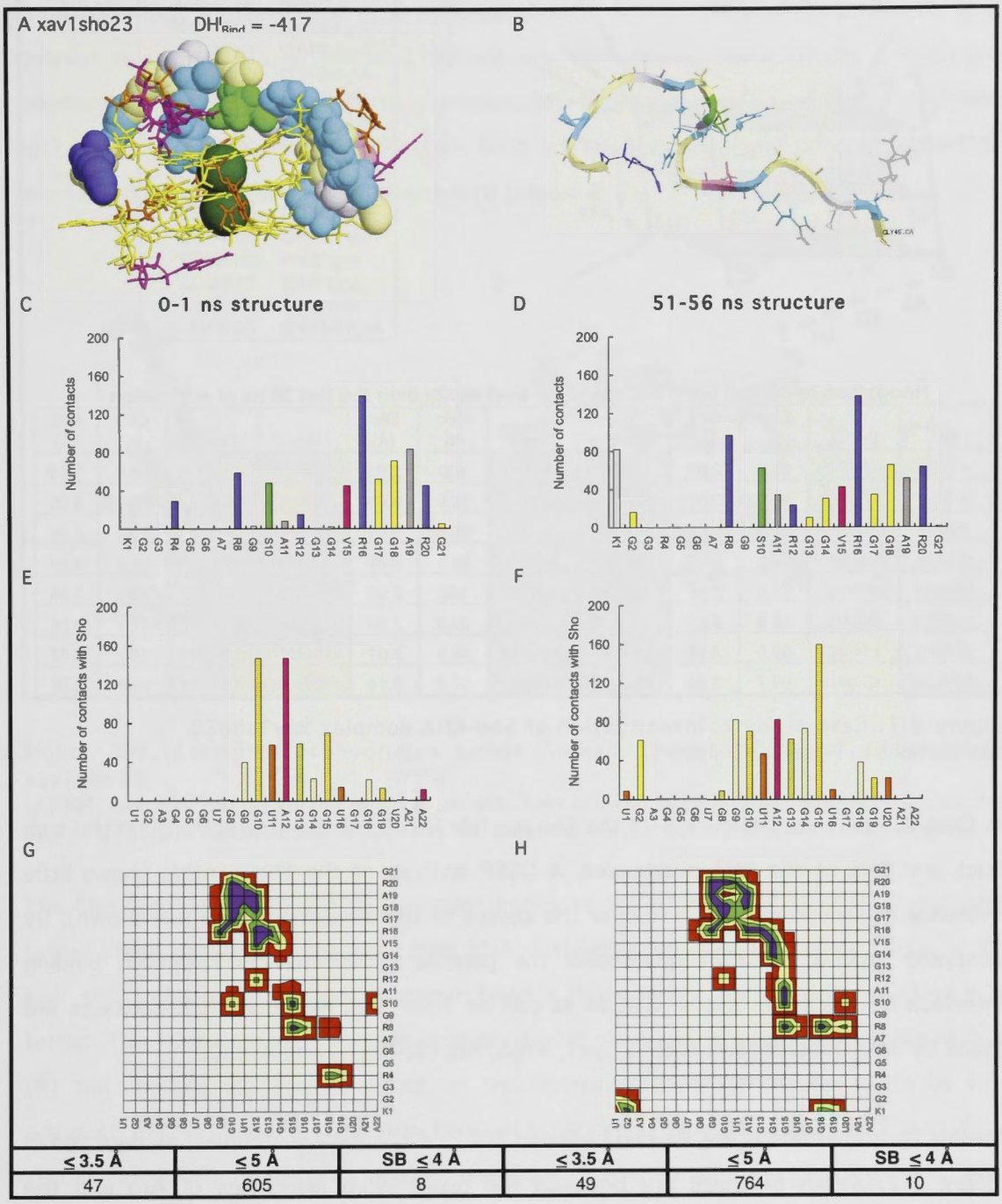


Figure 8-6. Hydrogen bond between Val15 of Sho and N2 of G17 in complex xav1sho46
G17 is in a tetrad arrangement with G4, G8 and G13. The G17@N2:G4@N7 bond is also shown.

Peptide interactions with the N2 atoms of guanines involved in the tetrad stack may be stabilizing or destabilizing, as discussed in Chapter 7. In this complex, the most unstable Hoogsteen hydrogen bond is G6@N7:G19@N2. The instability in this bond may be due to G19 (through the phosphate group) being involved in a hydrogen bond with Arg4.

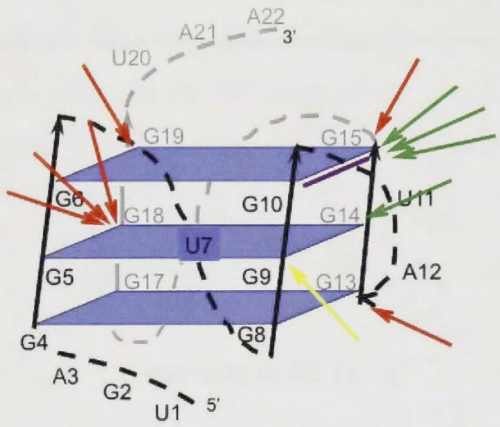
This complex has the strongest binding interface of all 4 Sho-RNA complexes. It has the greatest number of contacts $\leq$ 5 Å and the highest van der Waals contribution to the free energy. Although there are 11 salt bridges, the electrostatic contribution to the free energy (PBELE) is relatively modest compared to xav1sho27 and the Fx-RNA complexes. The RNA is moderately destabilized in this complex (ΔH_{RNA} = 14 kcal mol^{-1}). Notably, there are fewer contacts between the peptide and the sugar-phosphate groups associated with the tetrad in this complex than in other complexes that have greater RNA destabilization. For example, xav1sho23, which has greater RNA destabilization, has 12 hydrogen bonds between the peptide and the nucleotides forming the tetrad stack, compared to 6 in this complex. The Sho peptide is destabilized (ΔH_{Sho} = 12 kcal mol^{-1}) but again to a lesser extent than for other complexes. All of these factors contribute to the high binding energy predicted for this complex.

Case Study 2: Sho-RNA complex xav1sho23



≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å	≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å
47	605	8	49	764	10

Figure continued next page.



Phosphate hydrogen bonds	
Donor	**Acceptor**
Arg4@NH1	U16@O2P
Arg8@NH2	G15@O3'
Arg8@NH2	G18@O2P
Arg8@NH1	G18@O2P
Arg8@NH1	G18@O5'
Arg8@NH1	G19@O2P
Arg16@NH1	G13@O1P
Ribose hydrogen bonds	
Ser10@OG	G15@O2'
Ser10@HG	G15@O2'
Arg12@N	G14@O2'
Arg8@NE	G15@O2'
Base hydrogen bonds	
Arg16@NH2	G9@N3

Hoogsteen hydrogen bond occupancy in xav1sho23 over the last 20 ns of simulation [a]

Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)
:4@O6	:17@N1	99.8	2.92	:5@N7	:18@N2	100	2.93	:6@N7	:19@N2	98.9	3.03
:4@N7	:17@N2	99.6	3.02	:5@O6	:18@N1	99.9	3.00	:6@O6	:19@N1	98.4	3.00
:8@N7	:4@N2	97.7	3.05	:9@N7	:5@N2	99.6	3.01	:10@N7	:6@N2	99.9	2.98
:8@O6	:4@N1	100	2.88	:9@O6	:5@N1	99.1	3.07	:10@O6	:6@N1	99.3	3.00
:13@N7	:8@N2	99.5	2.99	:14@N7	:9@N2	100	2.90	:15@N7	:10@N2	98.6	3.08
:13@O6	:8@N1	99.6	2.95	:14@O6	:9@N1	99.9	2.99	:15@O6	:10@N1	97.1	3.06
:17@N7	:13@N2	99.9	2.98	:18@N7	:14@N2	99.9	3.01	:19@N7	:15@N2	100	2.92
:17@O6	:13@N1	99.7	3.00	:18@O6	:14@N1	97.8	3.14	:19@O6	:15@N1	99.8	2.98

Figure 8-7. Case study 2: investigation of Sho-RNA complex xav1sho23
See captions to Figures 8-3 and 8-5.

In Case 2, xav1sho23 (Figure 8-7) the Sho peptide is extended and wraps around the top, back and RHS of the RNA quadruplex. A DSSP analysis of the Sho peptide shows little evidence of secondary structure over the course of the simulation (data not shown). By wrapping around the RNA quadruplex the peptide establishes an extensive binding interface from early in the simulation as can be seen from Figure 8-7(C). Contacts are made by 5 of the charged residues Lys1, Arg8, Arg12, Arg16 and Arg20.

Twelve hydrogen bonds are formed between Sho and the RNA quadruplex, as depicted in Figure 8-7. Seven of these are between the guanidinium sidechains of Arg and the phosphate oxygens and thus have ionic character. Four are made to the ribose units and one is formed with a base, in this case Arg16 forms a hydrogen bond with G9 of the tetrad stack.

The interaction potential of the guanidinium sidechain is demonstrated by Arg8 in this complex, which forms a total of 6 hydrogen bonds, 5 with the phosphate groups of 3 guanine residues (G15, G18 and G19) and one with ribose. Three of the 5 hydrogen bonds are made to G18 and include a bifurcated hydrogen bond in which the NH1 and NH2 atoms of the guanidinium sidechain both act as hydrogen bond donors to the O2P atom of the phosphate associated with G18 (Figure 8-8).

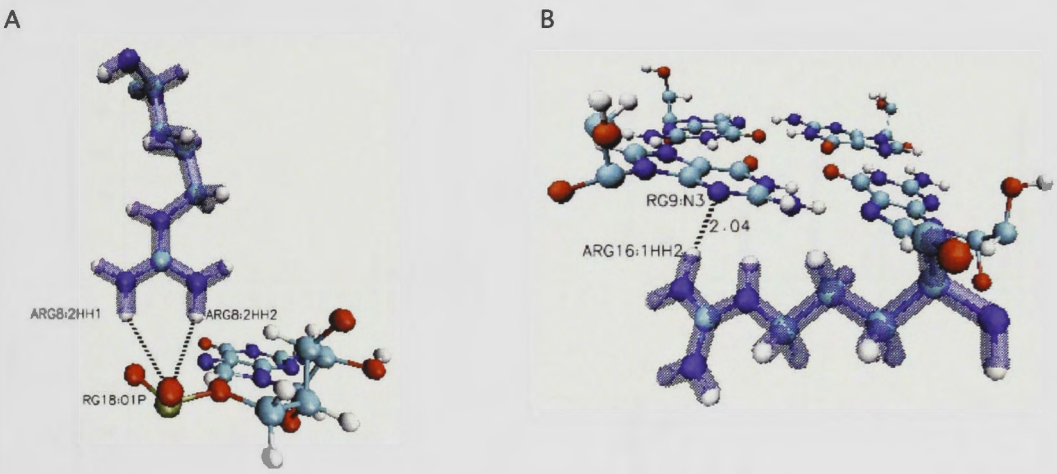
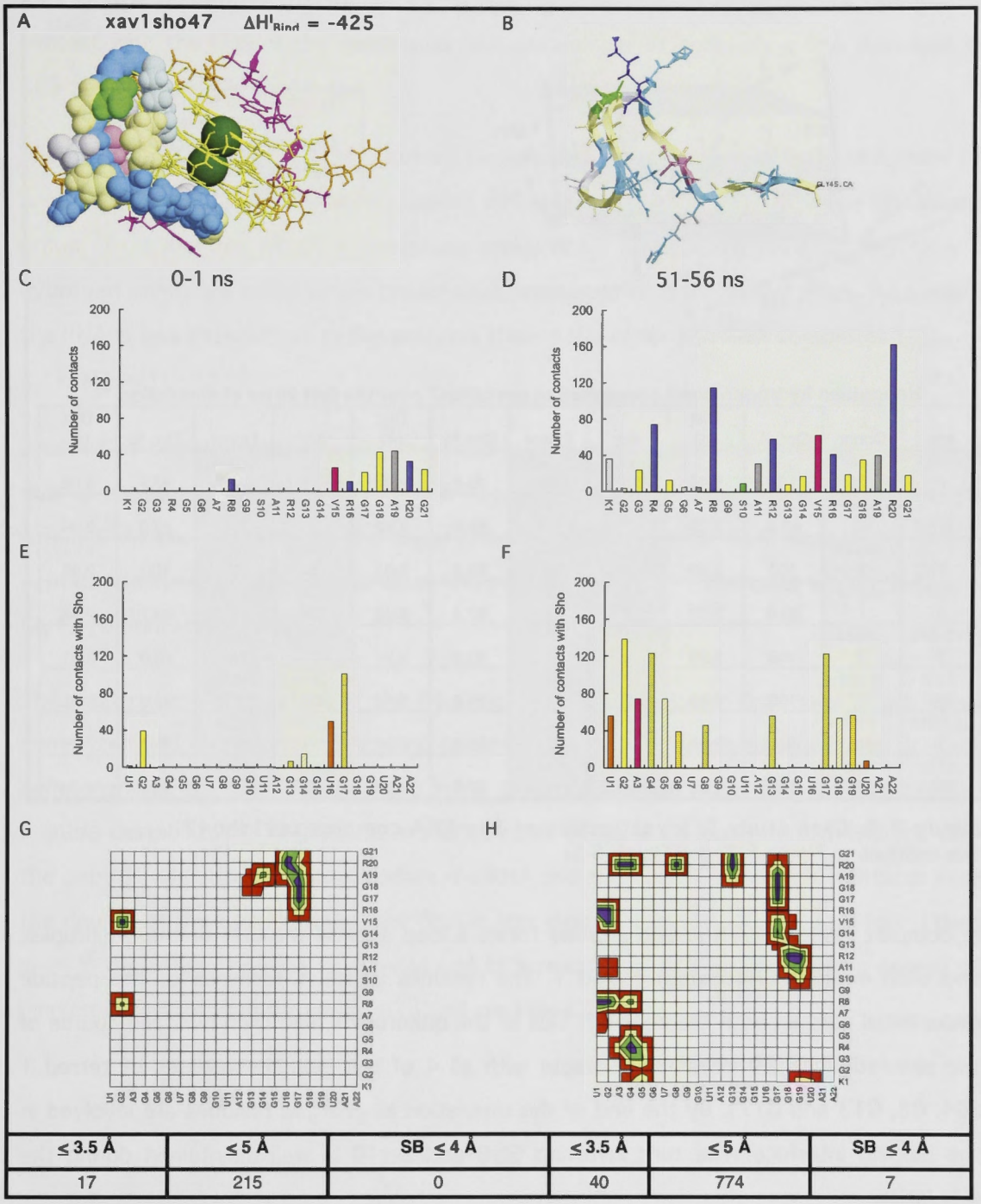


Figure 8-8. Examples of hydrogen bonds formed between Arg and the RNA in xav1sho23
(A) NH1 and NH2 atoms of the guanidinium sidechain of Arg8 both make hydrogen bonds to the O2P of G18. (B) The NH1 atom of Arg16 is the donor for a hydrogen bond to the N3 atom of G9.

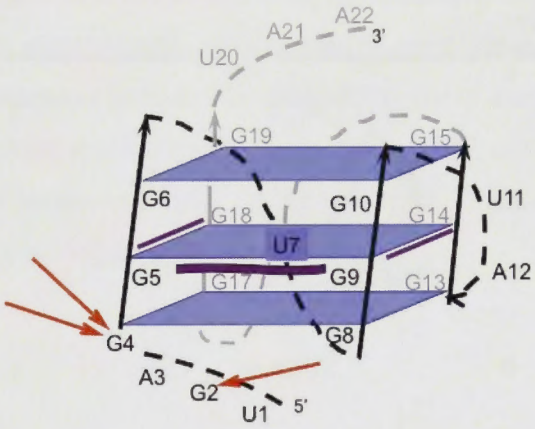
The Sho peptide makes several hydrogen bonds to the phosphates associated with the tetrad stack resulting in relatively high RNA destabilization in this complex (ΔH_{RNA} = 39 kcal mol^{-1}). However, only one hydrogen bond is made with a guanine base involved in a tetrad. This hydrogen bond is made through the N3 atom of G9, as depicted in Figure 8-8 (B) and appears to have no affect on the Hoogsteen hydrogen bonds made by G9. Although the Hoogsteen hydrogen bond occupancy is high, as shown in Figure 8-7, certain Hoogsteen bonds are occupied less frequently than others, indicating some level of instability. The most affected Hoogsteen hydrogen bond (G15@O6:G10@N2, Occ =97.1%) is found in tetrad 3. Some instability in tetrad 3 could be expected as the peptide wraps around the top of the RNA quadruplex, and particularly as the peptide makes 4 hydrogen bonds with the sugar-phosphate components of G15.

In summary, the strength of the binding interface is second only to complex xav1sho46. The electrostatic contribution is very similar to xav1sho46 but the van der Waals component is less. This extended form of the peptide has higher destabilization (ΔH_{Sho} = 20 kcal mol^{-1}) than seen in the peptide of Case 1. There is also relatively high RNA destabilization (ΔH_{RNA} = 39 kcal mol^{-1}). As a result the total enthalpy of binding is significantly less than for Case 1.

Case study 3: Sho-RNA complex xav1sho47



≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å	≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å
17	215	0	40	774	7



Phosphate hydrogen bonds	
Donor	**Acceptor**
Arg8@NH2	G4@O2P
Arg8@NH1	G4@O5'
Arg8@NE	G2@O5'

Hoogsteen hydrogen bond occupancy in xav1sho47 over the last 20 ns of simulation [ab]

Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)
:4@O 6	:17@ N1	100	2.90	:5@N7	:18@ N2	99.8	2.97	:6@N7	:19@ N2	97.7	3.03
:4@N7	:17@ N2	98.9	3.05	:5@O 6	:18@ N1	**91.0**	**3.22**	:6@O 6	:19@ N1	99.9	2.94
:8@N7	:4@N 2	100	2.93	:9@N7	:5@N 2	99.2	3.04	:10@N 7	:6@N 2	100	2.95
:8@O 6	:4@N 1	99.5	3.01	:9@O 6	:5@N 1	97.4	3.08	:10@ O6	:6@N 1	99.7	2.96
:13@N 7	:8@N 2	100	2.98	:14@N 7	:9@N 2	99.9	2.97	:15@N 7	:10@ N2	99.0	3.01
:13@ O6	:8@N 1	100	2.95	:14@ O6	:9@N 1	**95.8**	**3.14**	:15@ O6	:10@ N1	99.6	2.95
:17@N 7	:13@ N2	98.9	3.05	:18@N 7	:14@ N2	99.8	2.96	:19@N 7	:15@ N2	99.7	2.99
:17@ O6	:13@ N1	100	2.86	:18@ O6	:14@ N1	99.3	3.06	:19@ O6	:15@ N1	99.7	2.98

Figure 8-9. Case study 3: investigation of Sho-RNA complex xav1sho47
See captions to Figure 8-3- and Figure 8-5-

In complex xav1sho47, the Sho peptide forms a loop against the LHS of the quadruplex and then extends underneath tetrad 1. The residues at the C-terminus of the peptide make initial contact with the back left side of the quadruplex, and then over the course of the simulation Arg20 develops contacts with all 4 of the guanine residues of tetrad 1 (G4, G8, G13 and G17). By the end of the simulation all charged residues are involved in the binding interface. The turn involving Gly9 and Ser10 is well maintained during the simulation.

A large increase in the number of contacts made by the Sho peptide and the RNA quadruplex is seen over the course of this simulation (Figure 8-9 (C-F). At 1 nanosecond no salt bridges have been established and there are 215 contacts within 5 Å. These contacts are made by residues at the C-terminus of the peptide, which extend under the

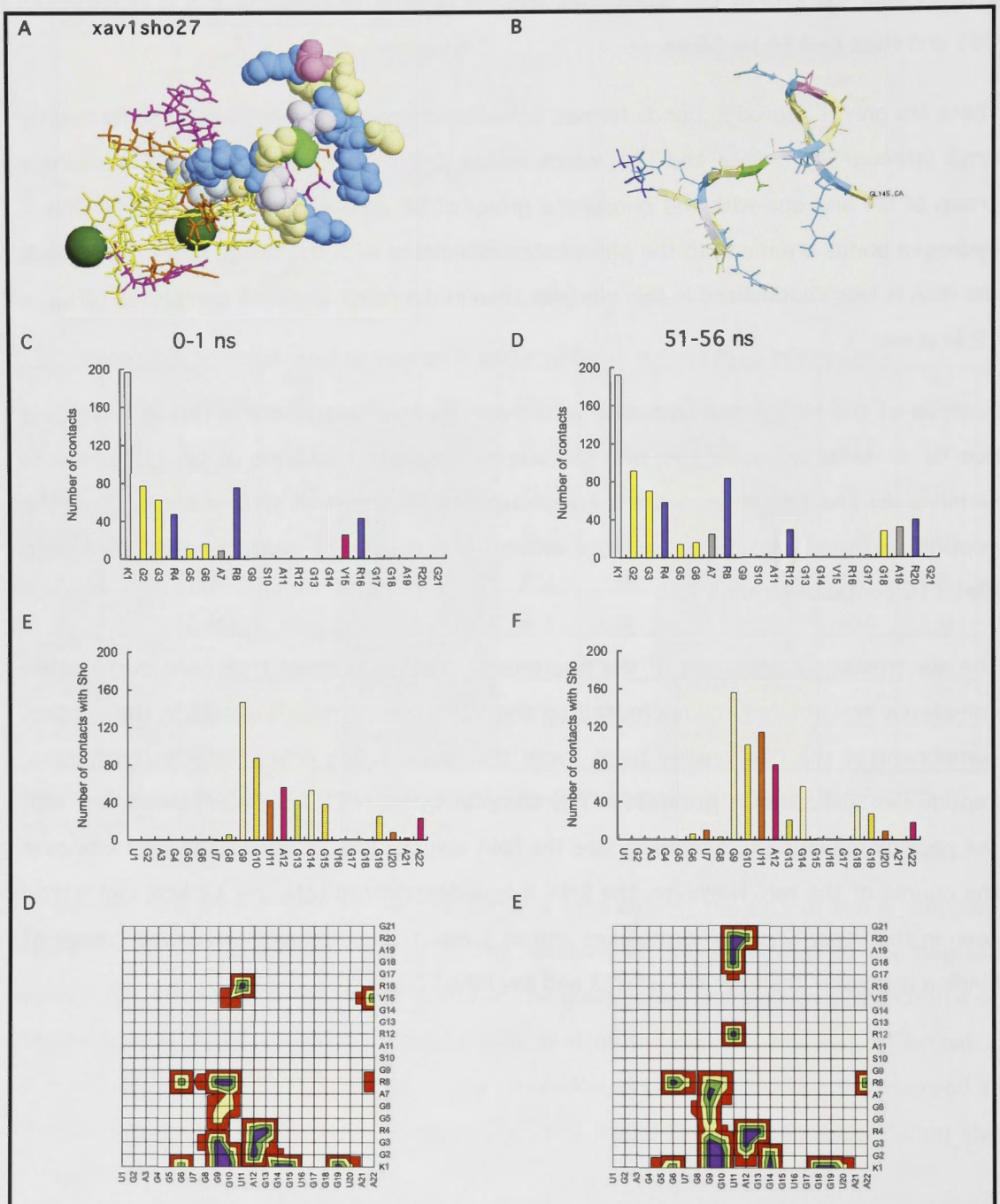
quadruplex stack, thus making contact with the quanines of tetrad 1 and the loop at the back of the quadruplex (U16). By 25 ns the other end of the peptide has moved into contact with the LHS of the quadruplex and the number of contacts $\leq$ 5 Å increases to 685 and then to 774 by 56 ns.

There are only 3 hydrogen bonds formed between the peptide and the RNA, all formed by Arg8 (through NH1, NH2 and NE), which makes 2 hydrogen bonds with the phosphate group of G4 and one with the phosphate group of G2 as depicted in Figure 8-9. Only 2 hydrogen bonds are made to the phosphates associated with the tetrad stack. As a result the RNA is less destabilized in this complex than in the other Sho-RNA complexes (ΔH_{RNA} = 12 kcal mol^{-1}).

Analysis of the Hoogsteen hydrogen bonds reveals some instability in tetrad 2 which is due to contacts made by Lys with the sugar-phosphate backbone of G5 (12 contacts within 5 Å) and contacts to the sugar-phosphate backbone of G18 made by 3 of the peptide residues: Arg12 (34 contacts within 5 Å), Ala11 (11 contacts within 5 Å) and Gly13 (6 contacts within 5 Å).

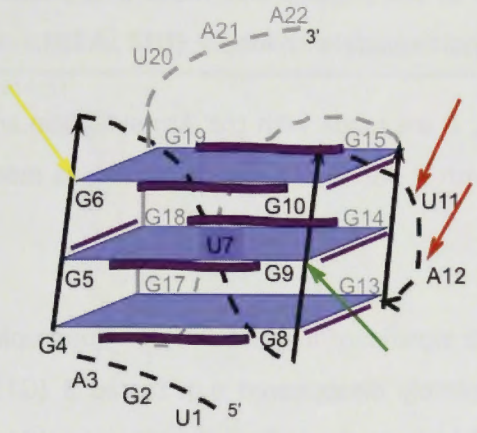
The electrostatic component of the free energy (PBELE) is lower than seen in the other complexes and not fully compensated by the VDW energy, which results in the contact component of the free energy being lower than seen in the other Sho-RNA complexes. Peptide destabilization is greatest in this complex (ΔH_{Sho} = 28 kcal mol^{-1}) consistent with the peptide adjusting to accommodate the RNA and the 260 % increase in contacts over the course of the run. However, the RNA is less destabilized (ΔH_{RNA} = 12 kcal mol^{-1}) than seen in the other Sho-RNA complexes and as a result the total estimated free energy of binding is greater than for xav1sho23 and xav1sho27.

Case study 4: Sho-RNA complex xav1sho27



≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å	≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å
55	561	4	52	678	5

Figure continued next page.



Phosphate hydrogen bonds	
Donor	**Acceptor**
Arg20@NH2	U11@OP1
Arg20@NH2	A12@OP1
Ribose hydrogen bonds	
Gly3@N	G9@O2'
Base hydrogen bonds	
G6@N1	Lys1@O

Hoogsteen hydrogen bond occupancy in xav1sho27 over the last 20 ns of the 56ns simulation [a]

Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)
:4@O6	:17@N1	98.5	2.97	:5@N7	:18@N2	99.9	2.95	:6@N7	:19@N2	99.6	2.98
:4@N7	:17@N2	99.7	3.00	:5@O6	:18@N1	89.7	3.09	:6@O6	:19@N1	99.9	2.93
:8@N7	:4@N2	89.2	3.00	:9@N7	:5@N2	98.3	3.02	:10@N7	:6@N2	99.5	2.97
:8@O6	:4@N1	88.0	2.95	:9@O6	:5@N1	**48.9**	**3.19**	:10@O6	:6@N1	95.7	3.05
:13@N7	:8@N2	**27.63**	**3.086**	:14@N7	:9@N2	**30.3**	**3.04**	:15@N7	:10@N2		
:13@O6	:8@N1	**37.17**	**2.97**	:14@O6	:9@N1	**30.9**	**3.12**	:15@O6	:10@N1		
:17@N7	:13@N2	98.2	3.00	:18@N7	:14@N2	98.8	2.96	:19@N7	:15@N2		
:17@O6	:13@N1	99.1	2.95	:18@O6	:14@N1	79.2	3.12	:19@O6	:15@N1		

Figure 8-10. Case study 4: investigation of Sho-RNA complex xav1sho27
See captions to Figures 8-3 and 8-5.

As a result of the initial random placement of the peptide and RNA, instability, resulting in migration of the structural K^+ ions was seen in a small number of complexes. Ultimately these complexes were ranked among the most stabilized complexes. One of these is the Sho-RNA complex xav1sho27.

In this complex the peptide interacts primarily with the LHS of the quadruplex. From the commencement of this simulation Lys1 and Gly2 were inserted between tetrad 2 and 3 thus making close contacts with G9, G10, G14 and G15 and causing relocation of the $K24^+$ ion from its normal position between these tetrads.

As a result of the close docking of the monomers, there are more contacts at 1 ns (561 contacts $\leq$ 5 Å) than seen in most of the other complexes at this stage of the simulation. As Lys1 is inserted between tetrads 2 and 3, the first 3 peptide residues are found to make many contacts with the RNA as shown in Figure 8-10 (C-D). Initially the N-terminus residues of the peptide make contact with the LHS of the quadruplex (strands 2 and 3

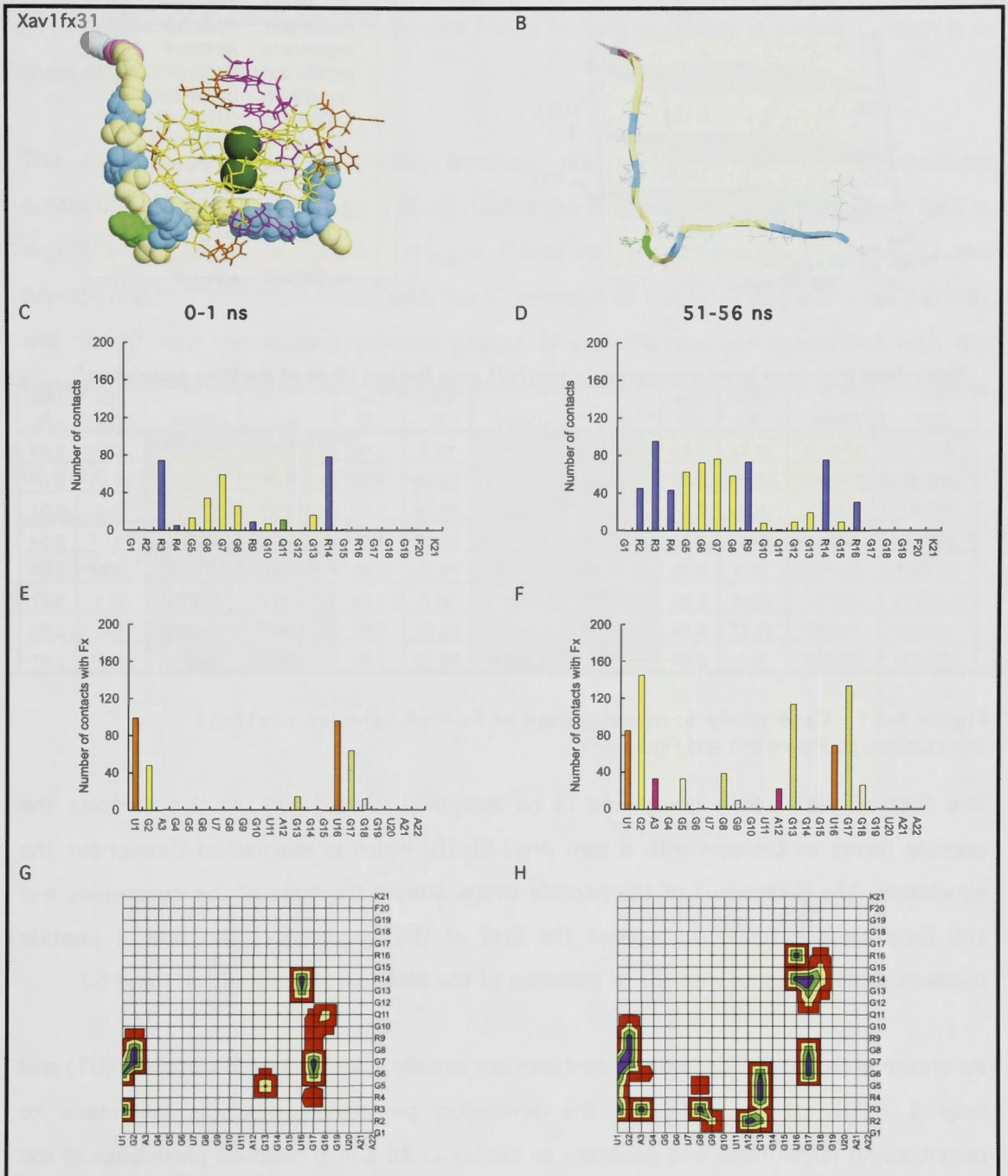
and loop 2). Over time residues at the other end of the peptide also move into position to make contact with the LHS of the quadruplex, particularly with loop 2 (U11, A12).

There are only 4 hydrogen bonds in this complex, 2 are made with the Arg sidechain and the phosphate group of the RNA, one is made with a ribose unit and the other is made with the guanine base of G6.

Analysis of the Hoogsteen hydrogen bonds reveals significant instability in the quadruplex affecting several Hoogsteen bonds, 4 have completely disappeared e.g. tetrad 3 (G15-G10 and G19-G15) while the occupancy of other bonds on the LHS of the tetraplex (G14-G9 and G13-G8) are greatly diminished. Although the tetrad stack is destabilized there is no indication of additional bonds being made to the RNA bases that are no longer involved in base stacking interactions. On the contrary, fewer peptide-RNA hydrogen bonds are seen in this complex than the other Sho-RNA case studies.

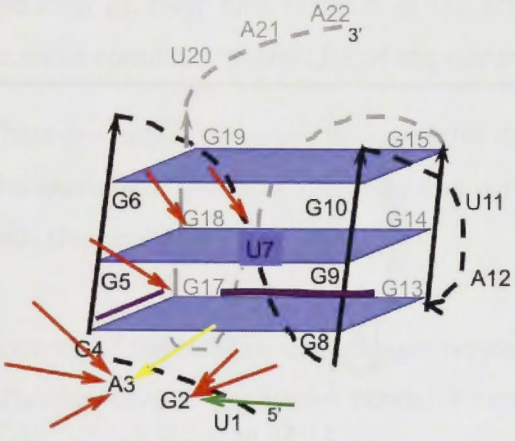
This complex has the highest electrostatic contribution to the binding interface of the Sho-RNA complexes but the lowest VDW energy. Although the Hoogsteen hydrogen bonds are significantly disturbed the value of ΔH_{RNA} (26 kcal mol^{-1}) is less than seen in xav1sho23, which has more stable tetrad bonds but also has more interaction with the sugar-phosphate groups of the guanine bases involved in the tetrad stack (12 hydrogen bonds, compared to 2 here). This again indicates that ΔH_{RNA} is affected more by the spread and number of interactions between the peptide and the RNA quadruplex rather than interactions than result in loss of particular Hoogsteen hydrogen bonds. The low VDW energy and the moderately high peptide (ΔH_{Sho}= 24 kcal mol^{-1}) and RNA destabilization results in the total estimated free energy of this complex being less than for the other complexes.

Case study 5: Fx-RNA complex xav1fx31



≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å	≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å
51	333	1	42	675	9

Figure continued next page.



Phosphate hydrogen bonds	
Donor	**Acceptor**
Arg3@NH1	A3@O1P
Arg3@NH2	A3@O1P
Arg3@NH2	A3@O2P
Arg16@NE	U16@O2P
Arg14@NH	G17@O2P
Gly15@N	G18@O1P
Arg9@N	G2@O5'
Gly7@O	G2@O5'
Ribose hydrogen bonds	
Gly8@N	G2@O4'
Base hydrogen bonds	
Arg3@NH1	A3@N7

Hoogsteen hydrogen bond occupancy in xav1fx31 over the last 20 ns of the 56ns simulation [a]

Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)
:4@O6	:17@N1	99.93	2.87	:6@N7	:19@N2	99.4	2.98	:5@N7	:18@N2	100	2.96
:4@N7	:17@N2	**87.67**	**3.17**	:6@O6	:19@N1	98.66	2.96	:5@O6	:18@N1	99.23	3.07
:8@N7	:4@N2	100	2.94	:10@N7	:6@N2	99.54	2.97	:9@N7	:5@N2	99.7	3.01
:8@O6	:4@N1	99.5	2.97	:10@O6	:6@N1	99.19	2.95	:9@O6	:5@N1	99.07	3.04
:13@N7	:8@N2	99.3	3.02	:15@N7	:10@N2	99.05	3.02	:14@N7	:9@N2	99.97	2.94
:13@O6	:8@N1	99.9	2.93	:15@O6	:10@N1	98.8	2.89	:14@O6	:9@N1	99.1	3.07
:17@N7	:13@N2	**91.13**	**3.18**	:19@N7	:15@N2	99.89	2.97	:18@N7	:14@N2	100	2.96
:17@O6	:13@N1	100	2.85	:19@O6	:15@N1	99.72	2.94	:18@O6	:14@N1	99.37	3.07

Figure 8-11. Case study 5: investigation of Fx-RNA complex xav1fx31
See captions to Figure 8-3 and Figure 8-5.

The first of the Fx-RNA complexes to be examined is xav1fx31. In this complex the peptide forms an L-shape with a turn Arg9-Gly10, which is maintained throughout the simulation. The N-terminus of the peptide wraps around the base of the quadruplex and the C-terminus extends up against the LHS of the quadruplex. The first 6 peptide residues make contacts with the 5' terminus of the RNA, particularly with U1 and G2.

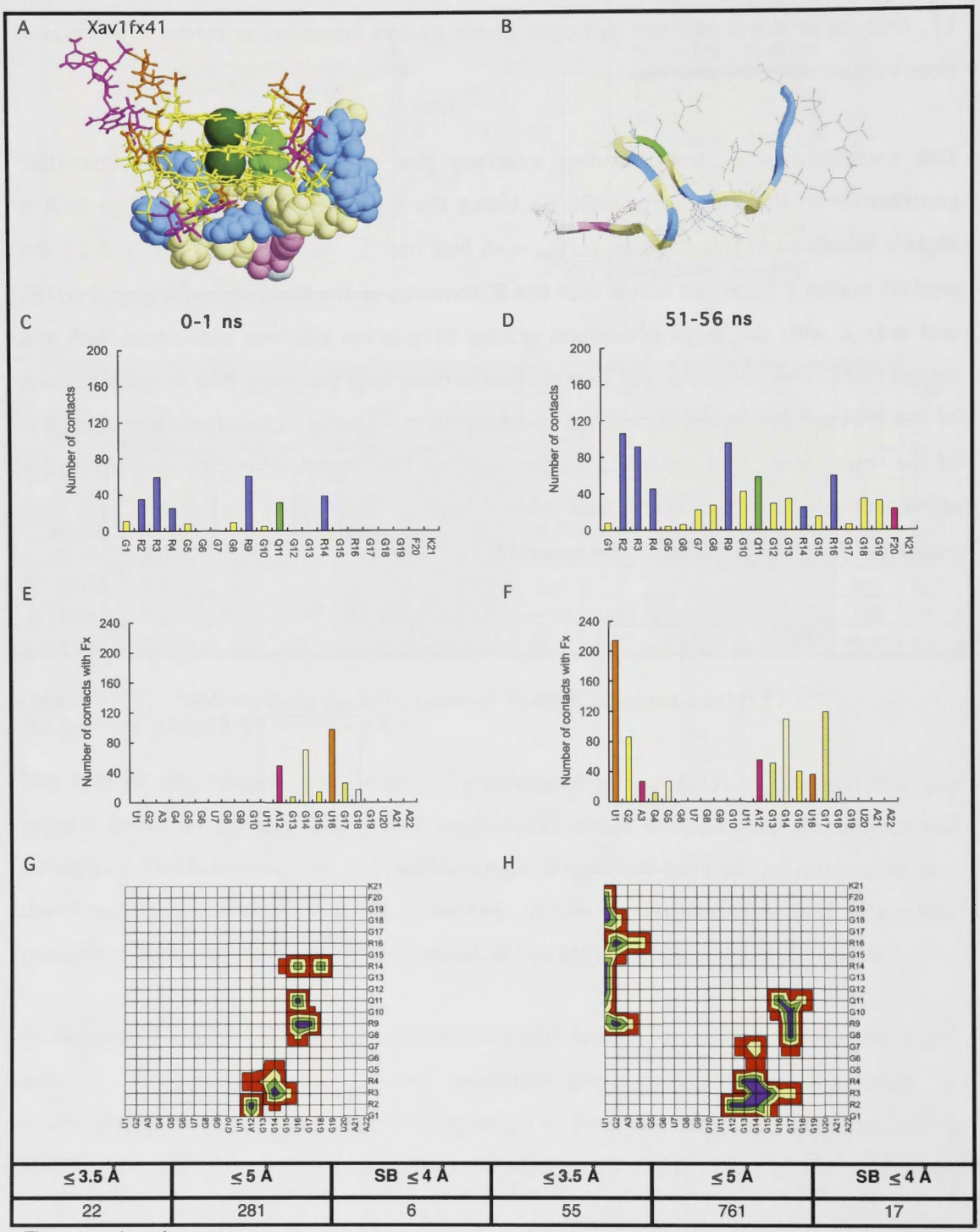
As shown in Figure 8-11 (C) most contacts are initially made to the 5' terminus (U1) and loop 3 (U16) of the RNA. As the simulation progresses contacts are made to neighbouring nucleotides and generally to tetrad 1. All 6 Arg residues participate in the binding interface with most contacts made between Arg3 and the base of the quadruplex (5' terminus, G8 and G13).

Ten hydrogen bonds are found in this complex. Eight of these are made with the phosphate groups of the RNA, 6 of which are formed by the Arg sidechain and the other

2 with the backbone of Gly17 and Gly15 respectively. In addition, one hydrogen bond is made to the ribose unit of G2 and one is made to the base of A3, as depicted in Figure 8-11. Analysis of the Hoogsteen hydrogen bonds reveals instability in tetrad 1 which is in close contact with the peptide.

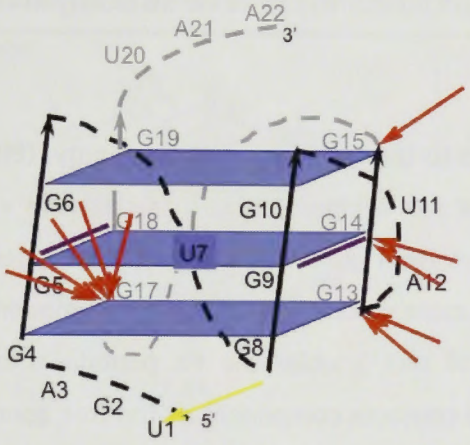
This complex has a strong binding interface due mainly to the high electrostatic contribution to the free energy (PBELE). Unlike the other Fx-RNA complexes, the RNA is slightly stabilized in this complex (ΔH_{RNA} = -6 kcal mol^{-1}). As seen from Figure 8-11 the peptide makes 7 hydrogen bonds with the 5' terminus of the RNA, one with loop 3 (U16) and only 2 with the sugar-phosphate groups of guanine residues associated with the tetrad stack. This suggests that peptide interactions with the loops and terminal regions of the RNA are stabilizing, in contrast to interactions with the sugar-phosphate backbone of the tetrad stack that appear to be destabilizing. RNA stabilization can compensate for significant peptide stabilization seen here (ΔH_{Fx}= 56 kcal mol^{-1}) and results in a high estimated free energy, second only to xav1fx16.

Case study 6: Fx-RNA complex xav1fx41



≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å	≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å
22	281	6	55	761	17

Figure continued next page.



Phosphate hydrogen bonds	
Donor	**Acceptor**
Arg2@NH1	G13@O2P
Arg2@NH2	G13@O2P
Arg3@NH1	G15@O1P
Arg3@NH1	G17@O1P
Arg3@NH2	G17@O1P
Arg4@N	G14@O1P
Gly5@N	G14@O1P
Arg9@N	G17@O5'
Gly10@N	G17@O2P
Gly10@N	G17@O5'
Base hydrogen bonds	
Arg9@O	U1@N3

Hoogsteen hydrogen bond occupancy in xav1fx41 over the last 20 ns of the 56ns simulation [a]

Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)
:4@O6	:17@N1	99.51	2.97	:5@N7	:18@N2	100	2.91	:6@N7	:19@N2	99.4	3.01
:4@N7	:17@N2	99.93	2.97	:5@O6	:18@N1	**96.41**	**3.12**	:6@O6	:19@N1	98.66	2.99
:8@N7	:4@N2	99.19	3.01	:9@N7	:5@N2	99.4	3.01	:10@N7	:6@N2	99.54	2.98
:8@O6	:4@N1	99.68	2.95	:9@O6	:5@N1	99.09	3.06	:10@O6	:6@N1	99.19	2.97
:13@N7	:8@N2	99.89	2.94	:14@N7	:9@N2	100	2.96	:15@N7	:10@N2	99.05	3.04
:13@O6	:8@N1	99.54	3.03	:14@O6	:9@N1	**96.83**	**3.11**	:15@O6	:10@N1	98.8	3.00
:17@N7	:13@N2	99.96	2.92	:18@N7	:14@N2	100	2.96	:19@N7	:15@N2	99.89	3.00
:17@O6	:13@N1	99.93	2.96	:18@O6	:14@N1	98.49	3.11	:19@O6	:15@N1	99.72	2.96

Figure 8-12. Case study 6: investigation of Fx-RNA complex xav1fx41
See caption of Figures 8-3 and 8-5.

In complex xav1fx41 the Fx peptide wraps around the RHS of the quadruplex, loops around the back of the quadruplex and extends under tetrad 1. As shown in Figure 8-12(E), contacts are initially concentrated on loops 2 and 3 of the quadruplex (A12 and U16, respectively). Contacts to A12 are made primarily by Arg2 and neighbouring Gly1, while contacts to U16 are made by Arg9 and its surrounding residues (Gly8, Gly10 and Gln11). After simulations have progressed to 56 ns, Arg9 no longer makes contact with U16 but instead makes contact with U1 together with 9 other residues.

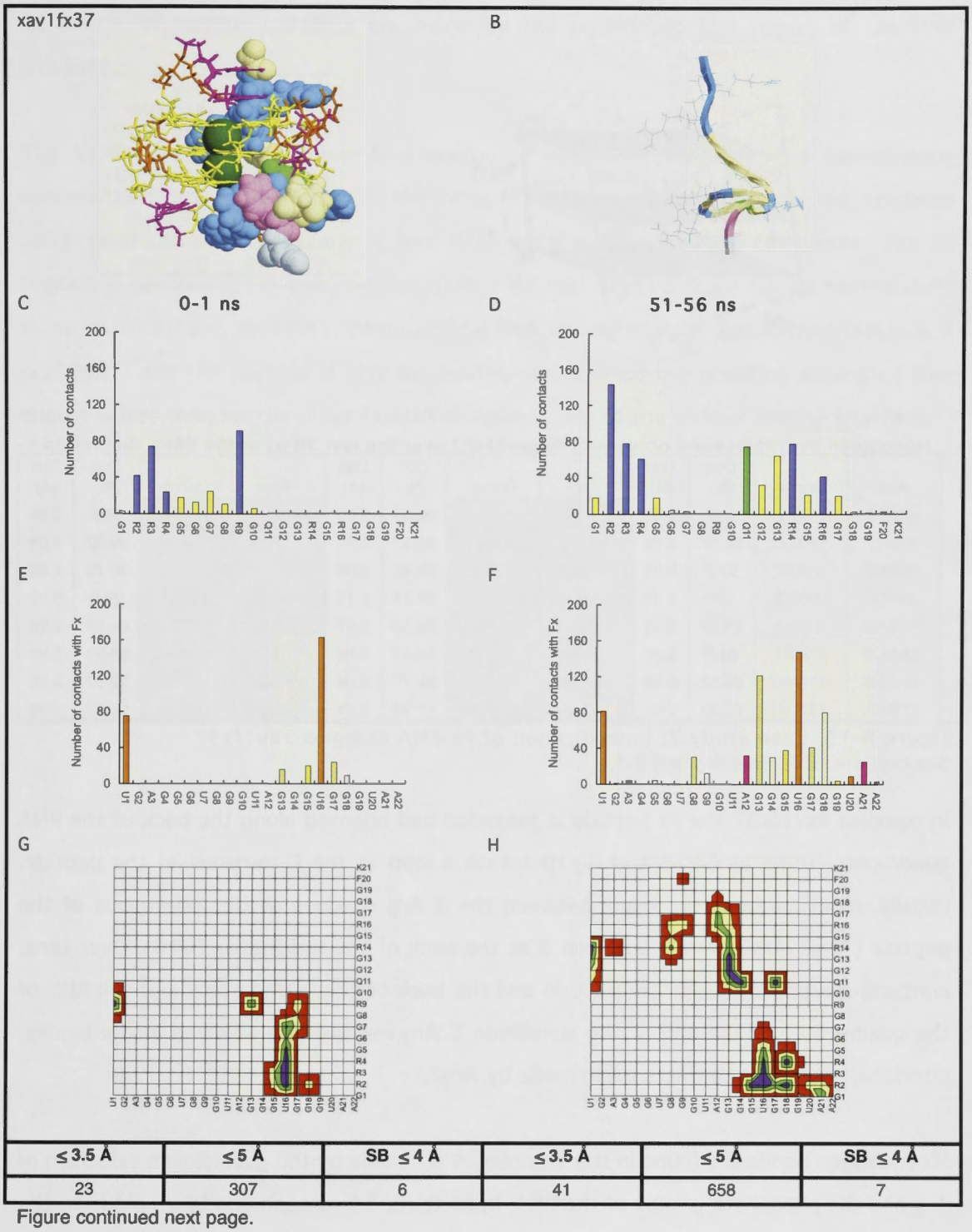
All 6 Arg residues are involved in the binding interface with most contacts being made by Arg2 to the LHS of the quadruplex (loop 2) and the back of the quadruplex (G13, G14, G15).

Eleven hydrogen bonds are found in this complex. Ten are made by phosphate groups of the RNA with 5 being formed by the Arg sidechain and 5 formed with the amide of the peptide backbone. In addition, 1 hydrogen bond is made to the base of U1 as shown in

Figure 8-12. Analysis of the Hoogsteen hydrogen bonds reveals some instability in tetrad 2 (G5-G18 and G14- G9 bonds).

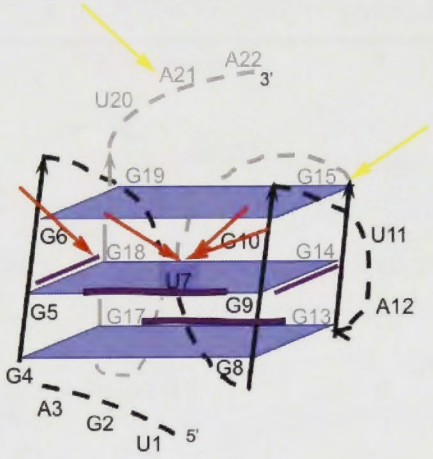
This complex has a strong binding interface due to the high electrostatic energy, (PBELE) which is partly due to the very high number of salt bridges, defined as contacts < 4 Å between the charged nitrogen atoms of Arg/Lys (NH1, NH2 and NZ) and the oxygens of the phosphate group. The VDW contribution is more modest. The RNA is only moderately destabilized in this complex (ΔH_{RNA} = 12 kcal mol^{-1}) while the Fx peptide is highly destabilized (ΔH_{Fx}= 61 kcal mol^{-1}). Although the contacts component of the free energy is high in this complex, the combination of high peptide destabilization and moderate RNA destabilization results in the overall estimated free energy of this complex being less than for complex xav1fx31 discussed previously.

Case study 7: Fx-RNA complex xav1fx37



≤3.5 Å	≤5 Å	SB ≤4 Å	≤3.5 Å	≤5 Å	SB ≤4 Å
23	307	6	41	658	7

Figure continued next page.



Phosphate hydrogen	
Donor	**Acceptor**
Arg2@NH2	G18@O2P
Arg3@NE	U16@O2P
Arg3@NE	U16@O1P
Arg3@NE	U16@O2P
Base hydrogen bonds	
A21@N6	Gly25@O
Arg2@NE	G15@N3

Hoogsteen hydrogen bond occupancy in xav1fx37 over the last 20 ns of the 56ns simulation

Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)
:4@O6	:17@N1	100	2.91	:5@N7	:18@N2	99.53	2.99	:6@N7	:19@N2	99.8	2.98
:4@N7	:17@N2	98.37	3.05	:5@O6	:18@N1	**89.57**	**3.20**	:6@O6	:19@N1	99.67	2.98
:8@N7	:4@N2	97.7	3.05	:9@N7	:5@N2	99.47	3.00	:10@N7	:6@N2	99.73	2.96
:8@O6	:4@N1	100	2.91	:9@O6	:5@N1	**96.77**	**3.14**	:10@O6	:6@N1	99.6	3.00
:13@N7	:8@N2	99.33	2.99	:14@N7	:9@N2	99.53	2.97	:15@N7	:10@N2	99.97	2.95
:13@O6	:8@N1	99.6	2.95	:14@O6	:9@N1	94.47	3.16	:15@O6	:10@N1	99.97	2.92
:17@N7	:13@N2	**90.37**	**3.08**	:18@N7	:14@N2	99.77	2.94	:19@N7	:15@N2	99.97	2.96
:17@O6	:13@N1	99.43	2.90	:18@O6	:14@N1	97.93	3.11	:19@O6	:15@N1	99.87	2.96

Figure 8-13. Case study 7: investigation of Fx-RNA complex xav1fx37
See captions of Figures 8-3 and 8-5.

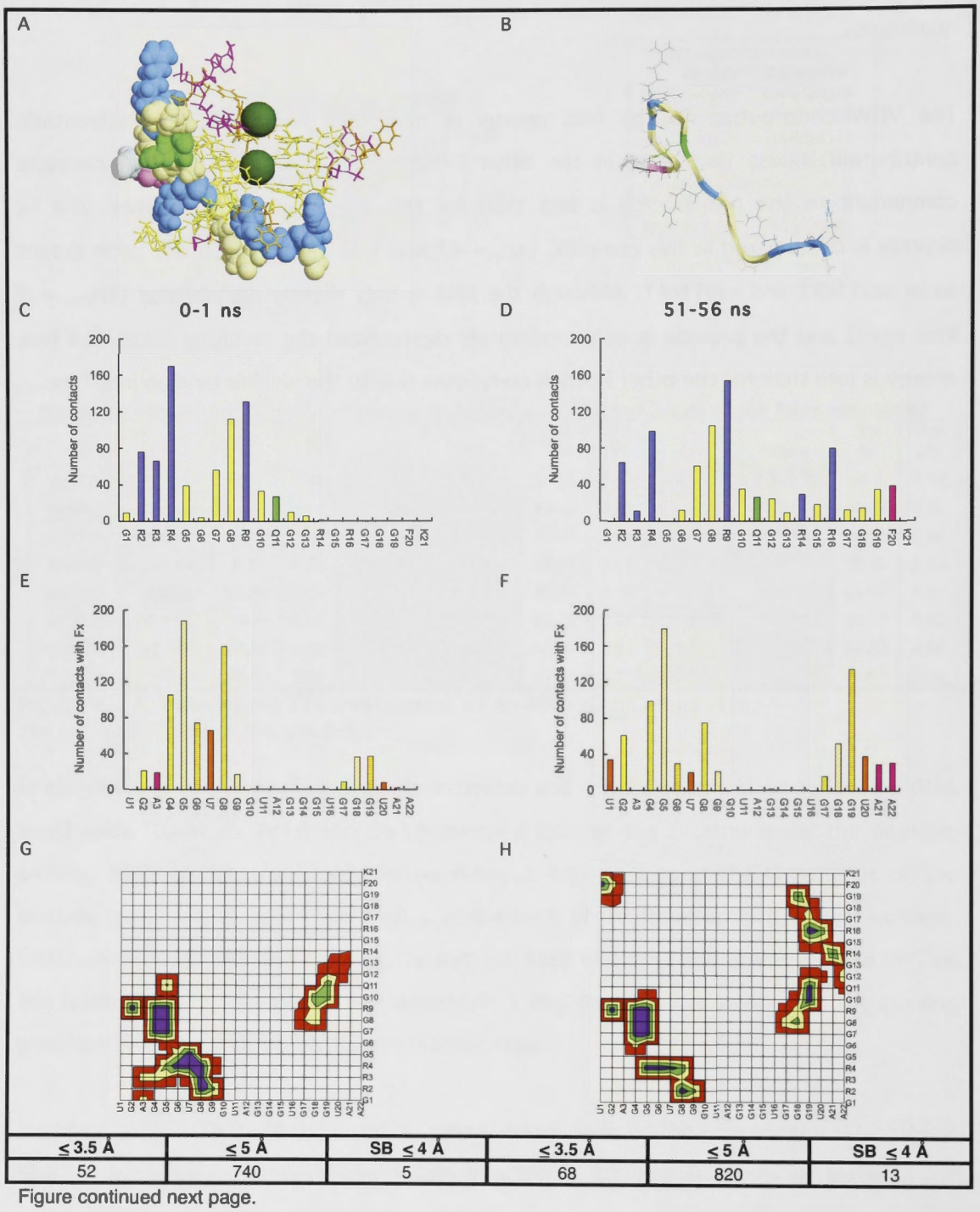
In complex xav1fx37 the Fx peptide is extended and oriented along the back of the RNA quadruplex. Turns at Gly18 and Gly19 create a loop at the C-terminus of the peptide. Initially, most contacts are made between the 3 Arg residues at the N-terminus of the peptide (Arg2, Arg3, Arg4) and loop 3 at the back of the quadruplex (U16). Over time, contacts develop between the peptide and the back of the tetrad stack and the RHS of the quadruplex. By the end of the simulation 5 Arg residues are involved in the binding interface, with most contacts being made by Arg2.

Six hydrogen bonds are found in this complex, 4 are made by the guanidinium sidechain of Arg and the phosphate groups of the RNA. In addition, 2 hydrogen bonds are made to the bases of G15 and A21 as shown in Figure 8-13.

Analysis of the Hoogsteen hydrogen bonds reveals instability in the G17-G13 and G18-G5 bonds. Numerous contacts are made by the peptide to this region of the RNA quadruplex.

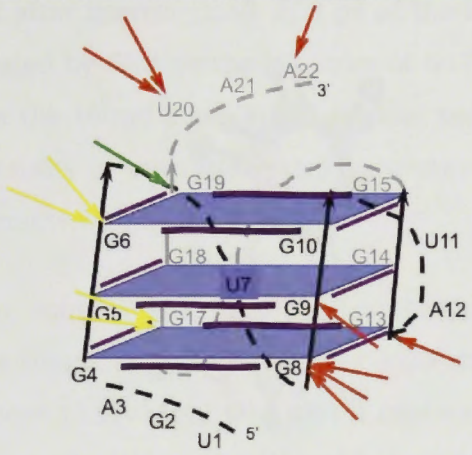
The VDW contribution to the free energy is moderate, however, the electrostatic contribution is less than seen in the other Fx-RNA complexes. Overall, the contacts component to the free energy is less than for the other Fx-RNA complexes. The Fx peptide is destabilized in this complex, (ΔH_{Fx}= 47 kcal mol^{-1}) but not to the same extent as in xav1fx31 and xav1fx41. Although the RNA is only slightly destabilized (ΔH_{RNA} = 3 kcal mol^{-1}) and the peptide is only moderately destabilized the resulting estimated free energy is less than for the other Fx-RNA complexes due to the weaker binding interface. .

Case study 8: Fx-RNA complex xav1fx16



≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å	≤ 3.5 Å	≤ 5 Å	SB ≤ 4 Å
52	740	5	68	820	13

Figure continued next page.



Phosphate hydrogen bonds	
Arg2@NH2	G9@O2P
Arg3@NE	G13@O1P
Arg4@NH2	G8@O2P
Arg4@NE	G8@O2P
Arg4@NE	G8@O5'
Gly13@N	A22@O1P
Arg16@NH2	U20@O2P
Arg16@NE	U20@O1P
Ribose hydrogen bonds	
Gly12@N	G19@O2'
Base hydrogen bonds	
Arg4@NH1	G6@N3
Arg4@NH2	G6@N3
Gly8@O	G17@N2
Gly8@O	G17@N1

Hoogsteen hydrogen bond occupancy in xav1fx16 over the last 20 ns of the 56ns simulation

Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)	Acc	Donor	Occ %	Dist (Å)
:4@O6	:17@N1	**0.03**	**3.49**	:5@N7	:18@N2	**93.6**	**3.12**	:6@N7	:19@N2	**30.2**	**3.27**
:4@N7	:17@N2	**0.1**	**3.42**	:5@O6	:18@N1	99.0	2.97	:6@O6	:19@N1	97.8	2.88
:8@N7	:4@N2			:9@N7	:5@N2			:10@N7	:6@N2		
:8@O6	:4@N1			:9@O6	:5@N1			:10@O6	:6@N1	**0.03**	**2.86**
:13@N7	:8@N2	86.8	2.97	:14@N7	:9@N2	**68.5**	**3.13**	:15@N7	:10@N2	**93.5**	**3.02**
:13@O6	:8@N1	59.3	3.12	:14@O6	:9@N1	**68.2**	**3.15**	:15@O6	:10@N1	**88.0**	**3.03**
:17@N7	:13@N2	84.6	2.98	:18@N7	:14@N2	97.8	3.04	:19@N7	:15@N2	99.9	2.92
:17@O6	:13@N1	76.2	3.11	:18@O6	:14@N1	99.8	2.91	:19@O6	:15@N1	**76.6**	**3.11**

Figure 8-14. Case study 8: investigation of Fx-RNA complex xav1fx16
See caption of Figures 8-3 and 8-5.

Complex xav1fx16 is another example of a complex with docking induced instability. The peptide is oriented towards the front and LHS of the quadruplex, and Gly7, Gly8 and Arg9 thread into the tetrad stack with the dislodgement of $K24^+$ from its normal position between tetrads 2 and 3.

As a result of the closeness of the peptide and RNA from commencement of the simulation, a much greater number of contacts are already formed at 1nanosecond (Figure 8-14 (C)) than seen in the other complexes.

Initially most contacts are between Arg9 and G4 and G5 of strand 1 of the quadruplex and Arg4 and the front face of the quadruplex (G5, G6, U7, G8 and G9). Residues in the middle of the peptide (Gly10 and Gln11) make contact with the back of tetrad 3 (G19). During the simulation contacts between the peptide and the back of the quadruplex and the 3' terminus increase. Also, Phe20 moves into a position where it has base stacking interactions with U1 as shown in Figure 8-15.

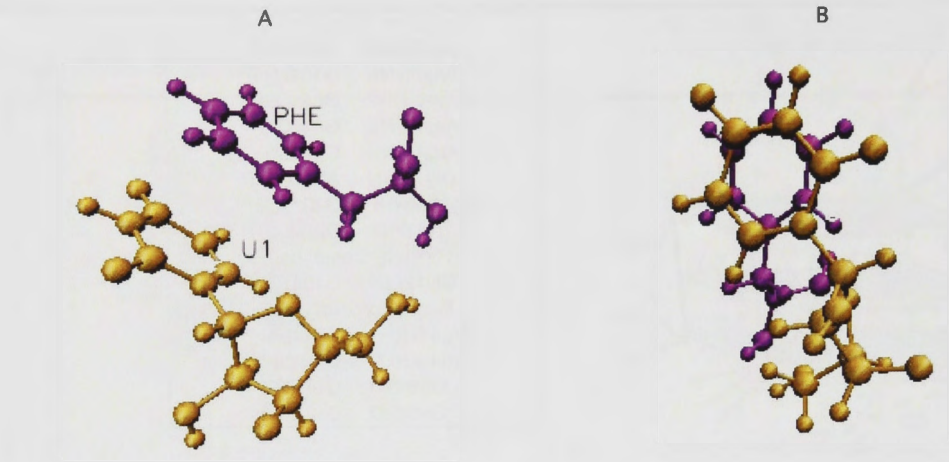


Figure 8-15. Base stacking interactions between Phe20 and U1
(A-B) Snapshot of base stacking interactions between the aromatic ring of Phe20 (purple) and the base of U1 (orange) viewed in different orientations.

All 6 Arg residues are found at the binding interface, with most contacts made by Arg9 and the 5' terminus end of the quadruplex. Ultimately there is a concentration of contacts with both the 5' and 3' ends of the RNA, as seen in Figure 8-14(C-D). It is interesting to note that this complex and xav1sho46, which has the highest ΔH^{I}_{Bind} of the Sho complexes, are both able to make abundant contacts with the 5' and 3' ends of the RNA.

Thirteen hydrogen bonds are found in this complex. Eight are made to phosphate groups with 7 made by the guanidinium sidechain of Arg and 1 made by the peptide backbone. In addition, 1 hydrogen bond is made to the ribose unit of G19. Arg4 makes a bifurcated hydrogen bond to the N3 atom of G6 and the backbone of Gly8 makes hydrogen bonds to the N1 and N2 atoms of G17. Analysis of the Hoogsteen hydrogen bonds reveals significant instability in the quadruplex with the Hoogsteen bonds between guanines at the front of the quadruplex (G4-G8, G5-G9 and G6-G10) disappearing and the occupancy of most of the other bonds being greatly diminished. This is due to the fact that Gly8, Arg9 and Gly10 insert into the tetrad stack and make contact with the N1, N2 and N7 atoms of G5, G17 and G18, thus disturbing the pattern of hydrogen bonding.

Although the Hoogsteen hydrogen bond between the N2 atom of G17 and G4 is in place following the initial docking it is quickly destabilized by the intrusion of Arg9 in the tetrad

and after approximately 200 ps of the MD simulation it is replaced by a hydrogen bond created by Gly8 to the N2 atom of G17. It seems unlikely that a peptide would intrude into the tetrad stack in the manner seen here. However, this example does show how instability in the tetrad stack generates the opportunity for the peptide to increase its contacts with the RNA.

This complex has the strongest binding interface of all the peptide-RNA complexes simulations. It has the greatest VDW energy and the electrostatic contribution is only second to xav1fx41. The RNA is moderately destabilized (ΔH_{RNA} = 18 kcal mol^{-1}) as is the Fx peptide (ΔH_{Fx}= 42 kcal mol^{-1}). As a result this peptide has the highest estimated free energy of all 8 case studies.

8.3 CONCLUSIONS

These 8 case studies demonstrate the high variability in the binding interface that is displayed by a set of highly stabilized complexes. They illustrate the ways in which peptide stabilization, RNA stabilization and the strength of the binding interface all contribute to produce an overall estimated free energy of binding. Although the stabilization due to the binding interface is greater than the RNA or peptide stabilization energy the relative difference in these quantities over the 8 complexes are of similar magnitude. As such, any of these factors can affect the ultimate ranking of the complexes in terms of their free energy of binding. This was seen in Case study 3 where there is a less extensive binding interface than in the other Sho-RNA complexes yet low RNA destabilization results in this complex being ranked second in binding free energy. Conversely, in Case study 6 high interaction energy was offset by high peptide destabilization and moderate RNA destabilization which together lowered the binding free energy ranking of this complex.

As each of these factors is important in determining the stability of each complex it is worthwhile summarizing the attributes that affect them. Firstly, the binding interface is most affected by the number of contacts established between the peptide and the RNA. From Case study 1 it can be seen that a hairpin bend allows both ends of the peptide to make contact with the same side of the RNA quadruplex, including both the 5' and 3' ends. This conformation resulted in the greatest number of contacts seen in the Sho-RNA

complexes. In the Fx-RNA complexes most contacts are found in Case study 8 where the peptide is extended and wraps around the RNA. The total number of contacts is only slightly less than seen in Case study 1. However, there is an energetic penalty involved in the peptide being fully extended.

The flexible RGG box peptides can bind to the RNA in very different conformations. In some instances the peptide binds in an extended form (Case studies 2 and 8) while in other cases there may be turns that are well maintained to form a loop or hairpin conformation in the peptide (Case study 1). However, the greatest contribution to the binding free energy occurs when the peptide has some internal stabilization rather than being fully extended. This is seen by comparing $\Delta H_{Sho/Fx}$ in Case studies 1 and 8. In the hairpin conformation (Case study 1) the peptide has greater stabilization than in the extended form and thus contributes more strongly to the total binding free energy of the complex.

The third factor to consider is RNA de/stabilization. The case studies reveal that dense contacts between the peptide and the sugar-phosphate backbone surrounding the tetrad stack result in relatively high RNA instability (ΔH_{RNA}). This has a greater impact than the loss of particular Hoogsteen hydrogen bonds, as shown by Case study 4. By contrast, contacts between the peptide and the loops appear to be stabilizing (Case study 5). Energetically, interaction with the RNA loops and terminal regions is therefore favoured. Notably, in the complexes with greatest $\Delta H'_{Bind}$ (Case studies 1 and 8) the peptide contacts are concentrated at the terminal regions of the RNA. In the most stabilized complex (Case study 8) base stacking interactions between Phe and U1 contribute to this phenomenon.

Together these results suggests that the strongest binding between an RGG box domain and an RNA quadruplex will be seen in complexes that:

- maximize the number of contacts between the peptide and the RNA,
- disfavour a completely extended peptide conformation, and
- favour binding to the RNA loops or terminal regions.

But is strong binding always biologically desirable? The answer to this is clearly, 'not necessarily'. Experimental work reveals that RGG box domains are frequently involved in processes that require dissociation from the protein target as well as binding, including processes such as mRNA transport and silencing. These functions are established for the RGG box of FMRP (Zalfa et al., 2003, Bagni and Greenough, 2005). In reversible processes low binding affinity accompanied by a large interaction surface can be advantageous. The decoupling of affinity and specificity is one of the advantages cited for disordered protein domains (Dunker et al., 2008).

While this study provides insight into the factors that can produce tight binding between an RGG box domain and an RNA quadruplex, it has also produced a number of complexes that are energetically indistinguishable as a result of the variation or flexibility in the system. This reinforces the ideas that the RGG box is both capable of interacting with different RNA targets and is capable of tuning the strength of binding as a result of its inherent flexibility. Binding interactions are also likely to be affected by other regions of the protein outside the RGG box domains modeled in this study. However the results produced here do indicate that RGG box proteins may well 'moonlight' or engage in a number of different functional roles (Tompa and Fuxreiter, 2008).

It is already known that the RGG box of FMRP is a multifunctional RNA binding domain (Darnell et al., 2005, Bechara et al., 2009, Corbin et al., 2009). This is the first study to show that the RGG box of Sho binds in similar ways to an RNA quadruplex and may therefore have similar RNA-binding functions.

8.4 REFERENCES

Bagni, C. & Greenough, W.T. (2005) From mRNP trafficking to spine dysmorphogenesis: the roots of fragile X syndrome. *Nat Rev Neurosci,* 6, 376-87.

Bechara, E.G., Didiot, M.C., Melko, M., Davidovic, L., Bensaid, M., Martin, P., Castets, M., Pognonec, P., Khandjian, E.W., Moine, H. & Bardoni, B. (2009) A novel function for fragile X mental retardation protein in translational activation. *PLoS Biol,* 7, e16.

Corbin, R., Olsson-Carter, K. & Slack, F. (2009) The role of microRNAs in synaptic development and function. *BMB Rep,* 42, 131-5.

Darnell, J.C., Mostovetsky, O. & Darnell, R.B. (2005) FMRP RNA targets: identification and validation. *Genes Brain Behav,* 4, 341-9.

Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z. & Uversky, V.N. (2008) The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics,* 9 Suppl 2, S1.

Tompa, P. & Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci,* 33, 2-8.

Zalfa, F., Giorgi, M., Primerano, B., Moro, A., Di Penta, A., Reis, S., Oostra, B. & Bagni, C. (2003) The fragile X syndrome protein FMRP associates with BC1 RNA and regulates the translation of specific mRNAs at synapses. *Cell,* 112, 317-27.

CHAPTER 9

EXPERIMENTAL STUDIES OF G-QUADRUPLEX RECOGNITION BY RGG BOX PEPTIDES

9.1 INTRODUCTION

The interaction of nucleic acid with protein is an extensively studied phenomenon, indicative of the huge range of such interactions that are of biological importance and scientific interest. Several studies have investigated the binding of the RGG box domain of FMRP to RNA (Siomi et al., 1993, Darnell et al., 2001, Schaeffer et al., 2001, Ramos et al., 2003, Darnell et al., 2004, Darnell et al., 2005, Zanotti et al., 2006, Menon and Mihailescu, 2007, Menon et al., 2008). These studies reveal that the RGG box of FMRP is capable of binding with high affinity and specificity to certain forms of G-quadruplex RNA. Understanding the way FMRP binds to RNA is important because it is known that FMRP binds to certain mRNAs found in the brain and regulates translation of the messenger RNA into proteins required in neural plasticity. Disruption of this function, as seen when the gene coding for this protein is mutated, results in the excessive growth of spindly dendrites associated with mental retardation 'Fragile X Mental Retardation' (Irwin et al., 2001, Bagni and Greenough, 2005).

The study of PrP binding to a randomized RNA library also produced evidence that the N-terminus of PrP (residues 23-52) is capable of binding with high affinity to a 3 tetrad RNA G-quadruplex (Weiss et al., 1997). Subsequently, it was found that the N-terminus of PrP (residues 23-90) is capable of binding a wide range of RNA oligonucleotides, irrespective of sequence or secondary structure, with considerable affinity (K_d < 100 nM) while a few of the RNA transcripts tested demonstrated specific binding. In this context, specific binding is defined as binding that is unaffected by competitor RNA. This led to the suggestion that the N-terminus of PrP (residues 23 - 90) contains both a non-specific and a specific RNA binding site (Zeiler et al., 2003).

A nucleic acid binding function has not yet been discovered for PrP although several studies have established that PrP binds RNA and DNA with high affinity and that binding

results in structural changes to both the nucleic acid and PrP (Grossman et al., 2003, Silva et al., 2008), as discussed in Chapter 1.

G-quadruplex nucleic acid appears to be a target for both FMRP and PrP. This form of nucleic acid has been found to occur in the telomere regions of DNA, in the promoter regions of the genome and in mRNA. It is likely that the quadruplex structure, which can be induced by the local environment (e.g. ionic environment and pH) and modified by protein binding, acts as a control mechanism in transcription of DNA and translation of protein.

While Sho has an RGG domain, similar to FMRP, and is a member of the PrP family, there is currently no published evidence that it also binds RNA.

I used three experimental tools, circular dichroism (CD), surface plasmon resonance (SPR) and fluorescence spectroscopy to investigate whether the RGG domain of Sho is capable of binding RNA and to compare the binding affinity with that of the FMRP RGG domain and a peptide from the N-terminus of PrP (residues 23-52). Five different RNA oligonucleotide targets were used in this study, two of which have previously been found to form a G-quadruplex (Zanotti et al., 2006, Menon and Mihailescu, 2007).

CD studies were performed first to test whether the RNA oligonucleotides form a G-quadruplex and to determine whether the peptides have secondary structural elements. CD was also used to monitor how the RNA quadruplex changed upon peptide binding. SPR studies were performed to compare the binding affinity of the peptides for the different RNA oligonucleotides. Fluorescence methods were used to obtain a more detailed understanding of the binding of the peptides to the RNA G-quadruplex formed in the 3'-UTR of semaphorin mRNA.

Two of the RNA oligonucleotides used in this study, 'SC' the Sc1 aptamer (a 58 mer RNA sequence) and 'SF' were originally identified from *in vitro* RNA selection experiments as having high affinity for the RGG box of FMRP (Darnell et al., 2001). SF has a G-quadruplex structure and is found in the 3' UTR of the mRNA coding for the Semaphorin 3A protein (SEMA 3F). SEMA 3F plays a role in neuronal development through repulsive axon

guidance and by facilitating the collapse of the growing tip of axons (growth cones) to stop their extension (Yazdani and Terman, 2006). Translation of the mRNA is likely to be regulated by interactions with the quadruplex structure in the 3' UTR.

Darnell et al., found that the Sc1 aptamer binds to both full length FMRP and the RGG box of FMRP with a K_d of 10 nM while the SF transcript bound with a K_d of 75 nM. Mihailescu and co-workers subsequently studied the binding of these oligonucleotides to the RGG box of FMRP. Their results for the Sc1 aptamer (K_d 7.9 $\pm$ 4 nM) (Zanotti et al., 2006) are in accord with the earlier studies, but those for the modified SF transcript indicates much tighter binding (K_d = 0.7 nM) than found by Darnell et al. (Menon and Mihailescu, 2007). The authors suggest this discrepancy may be explained by modifications they introduced to the SF transcript in order to reduce the possibility of it forming more than one type of G-quadruplex. The modified form of SF was also used in this study. Mihailescu and colleagues propose that the Sc1 aptamer and SF both form a parallel unimolecular G-quadruplex, similar in form to the c-myc DNA G-quadruplex (see Figure 9-9).

In addition to SC and SF, three other RNA oligonucleotides have been tested in this study; a transcript of the RNA model used in the MD simulations (Xav), a polyguanine transcript (PG), and a variant SF transcript (SM) produced by mutating four of the guanines involved in the tetrad stack to uridine with the intention of disrupting the tetrad stack.

9.1.1 Background – Circular dichroism

Circular dichroism is observed through the interaction of circularly polarized light with an asymmetrical molecule. In circularly polarized light the electric field vector rotates in a circular manner as illustrated in Figure 9-1. An optically active molecule will absorb right-handed circularly polarized light and left-handed circularly polarized light to a different extent and with a different refractive index.

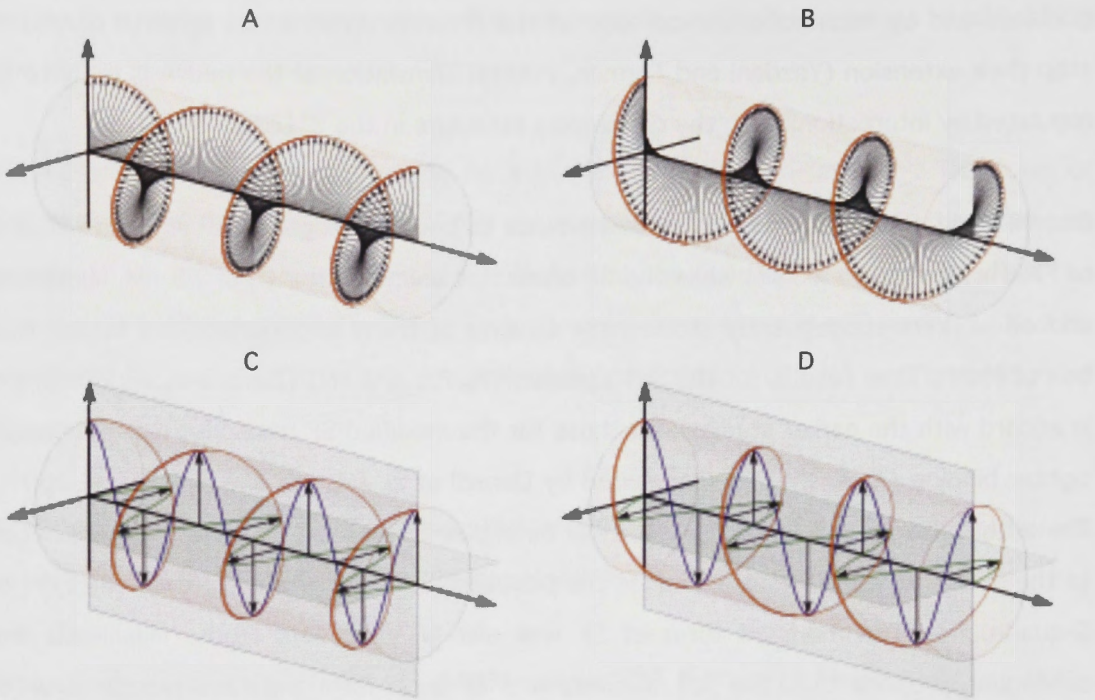


Figure 9-1. The x and y components of the electric field vector in right and left circularly polarized light
(A) Electric field vector of right-handed circularly polarized light rotating in a clockwise direction. (B) Electric field vector of left-handed circularly polarized light rotating in an anti-clockwise direction. (C and D) Circularly polarized light is often explained by splitting the electric field vector into 2 components that are perpendicular to each other, as illustrated by the horizontal green and vertical blue curves in the illustration. One component (the green line in C and the blue line in D) leads the other by a quarter of a wavelength. Stated another way they are out of phase by a quarter of a cycle or 90° and as a result the electric field vector rotates creating circularly polarized light.
Source of images - http://en.wikipedia.org/wiki/Circular_polarization

In a CD experiment, equivalent intensitites of left and right circularly polarized light of a particular wavelength is shone onto a sample. The difference in absorption (ΔA) is measured and yields the CD spectrum.

$$\Delta A = A_L - A_R \tag{1}$$

Eqn 1 can also be expressed in terms of Beer's law, as:

$$\Delta A = \varepsilon_L Cl - \varepsilon_R Cl \tag{2}$$

and

$$\Delta\varepsilon = (\varepsilon_L - \varepsilon_R) = \frac{\Delta A}{Cl} \quad (3)$$

where

- ε_L and ε_R are the molar extinction coefficients of the molecule for left and right circularly polarized light respectively
- C is molar concentration of the sample
- l is the path length of the cell holding the sample

The molar circular dichroism of a substance ($\Delta\varepsilon$) is usually expressed in mdeg M^{-1} cm^{-1} and peaks are generally in the range 0-10. However, for historical reasons, CD is often expressed in degrees ellipticity θ. Although the use of different units of measurement can be confusing, conversion between molar circular dichroism and molar ellipticity is straightforward as θ (in deg cm^2/dmol) = 3298 $\Delta\varepsilon$.

The far-UV spectrum (190 - 250 nm) of proteins provides information about the secondary structure of the protein. The amide bond of the peptide, specifically, the π - π^* absorption, gives rise to the CD spectrum in this region with characteristic features for particular secondary structure such as the α-helix, β-sheet or random coil as shown in Figure 9-2. The helical structure has a negative band (n-π^* transition) at ~ 222 nm and a π-π^* transition which is split into a negative band at ~208 nm and a positive band at ~192 nm. The anti-parallel β-sheet has a positive band at 195 nm and a negative band at 218 nm (Greenfield and Fasman, 1969). A disordered peptide has similar CD spectra to a simple amide although the magnitude of the signal will differ. Traditionally the CD spectra of a random coil has been characterized by a large negative band (π-π^* transition) at ~195 nm and a small positive band (n-π^* transition) at ~230 nm. More recently it has been suggested that a true random coil does not have a positive maximum and that evidence of such a band is indicative of P_{II} conformation (Shi et al., 2002, Chellgren and Creamer, 2004, Danielsson et al., 2005, Shi et al., 2006). The CD spectrum of poly(Pro) has a maximum positive band at~ 225 nM and and minimum at ~ 206 nM. As other residues are exchanged for proline the maximum is seen to move to lower wavelengths, so that a poly(Glu) peptide is seen to have both peaks shifted by ~ 10 nm (max ~215 nm, min ~196 nm) (Woody, 2009).

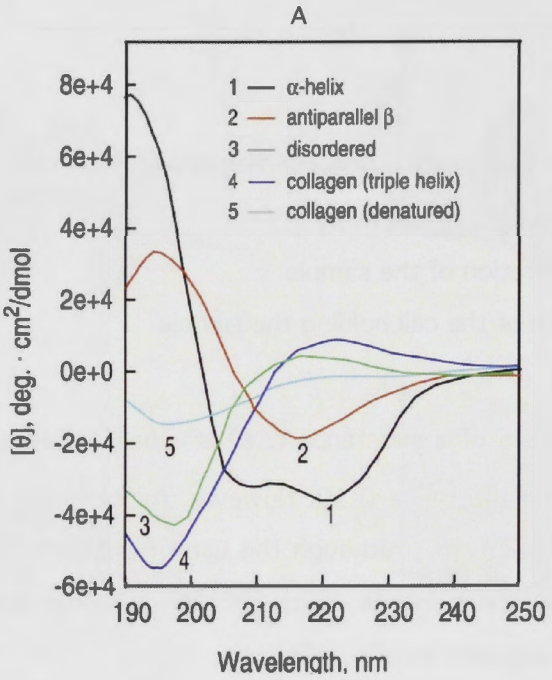


Figure 9-2. Circular dichroism spectra
(A) Characteristic CD spectra of α-helix (black), β-sheet (red) and disordered region (green), collagen helix (P_{II})(dark blue), denatured collagen helix (light blue).
Source of image: (Greenfield, 2006)

The near-UV spectrum (> 250 nm) provides information about the tertiary structure of a protein. The signals obtained in this region arise from the aromatic amino acid residues and Cys-Cys bonds.

When a biomolecule changes structure as a result of ligand binding its CD spectra will also change. CD can therefore be used to monitor binding interactions and to see whether the tertiary or secondary structure of a molecule changes upon ligand binding.

Nucleic acids also have characteristic absorbance bands and CD is commonly used to study the formation, structure and ligand binding properties of G-quadruplex DNA (Paramasivan et al., 2007). The CD spectra of 4 common types of G-quadruplex are illustrated in Figure 9-3. The form of the parallel unimolecular G-quadruplex, most relevant to this study, is shown in the lower RHS of the Figure; a positive peak at ~ 264 nm and a negative peak at ~ 240 nm is indicative of this structure. These characteristic signals arise as a result of the *syn* or *anti* orientation of the glycosidic bonds involving the quanine bases forming the G-quadruplex tetrads.

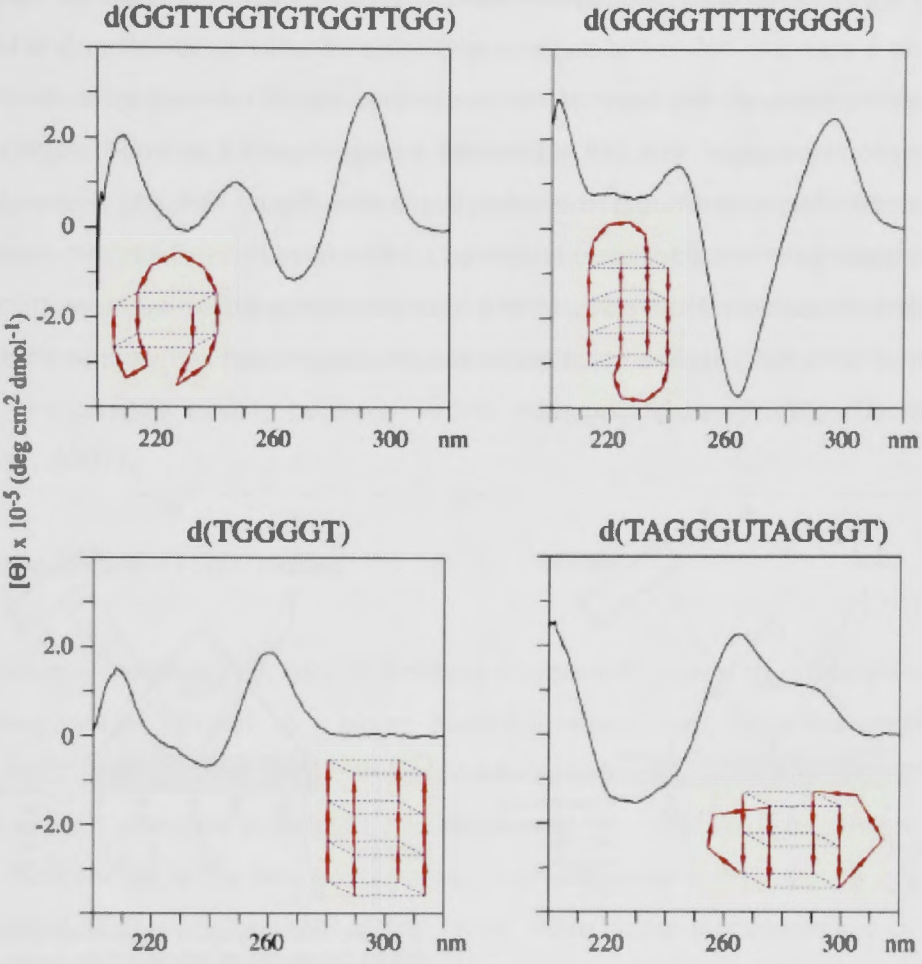


Figure 9-3. Circular dichroism spectra of 4 different forms of G-quadruplex DNA
Source (Paramasivan et al., 2007) From upper left clockwise, antiparallel basket, antiparallel, parallel propeller loops, 4-stranded parallel.

9.1.2 Background - Surface Plasmon Resonance

Surface plasmon resonance (Rich and Myszka, 2000, Van Der Merwe, 2001) is a phenomenon that occurs in thin conducting films at an interface between media of different refractive index. When light is shone onto a thin metal film, such as gold (used in the BIAcore chips) at a particular angle of incidence and wavelength, a fraction of the light can interact with delocalized electrons in the metal film ('plasmon'). When the wave vector of the incident light matches the wavelength of the oscillating mobile electrons the electrons are said to 'resonate' and hence the term surface plasmon 'resonance'. The

coupling of the incident light with the plasmon causes a reduction in the intensity of the reflected light and at a certain angle of incidence the intensity of reflected light will be reduced to a minimum value. The angle at which this occurs is predominantly affected by the refractive index at the back of the metal film and is referred to as the 'surface plasmon resonance angle'. In a SPR experiment a target ligand is immobilized on the back of the metal film thus altering the refractive index (Figure 9-4(A)). Subsequently a potential binding partner ('analyte') is allowed to flow over the surface and upon binding to the immobilized ligand there are further changes to the refractive index that can be measured (Figure 9-4(B)). The value of the response expressed as response units (RU) is thus proportional to the analyte bound.

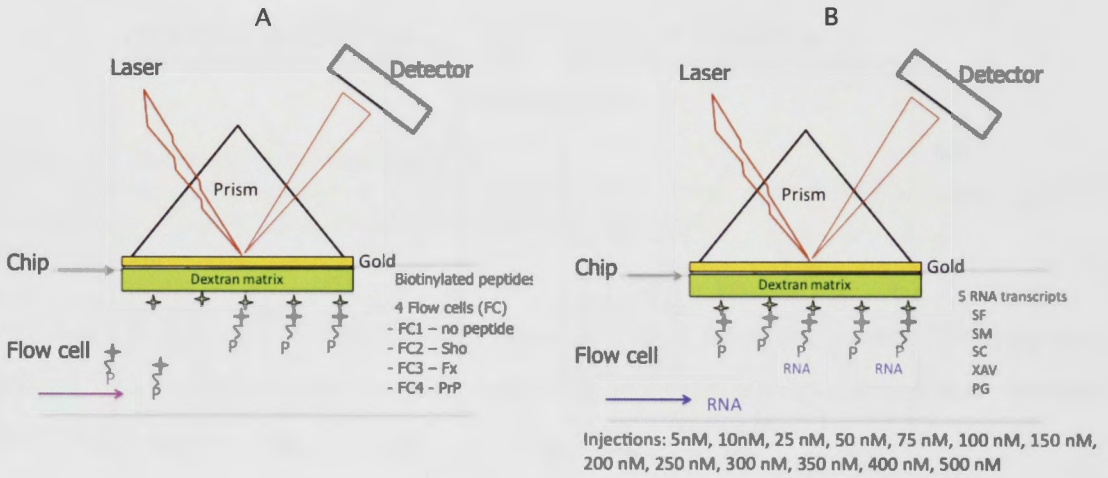


Figure 9-4. Surface Plasmon Resonance
Schematic representation of surface plasmon resonance set up used in this study (A) Biotinylated peptides were immobilized on the cell. Flow cell 1 (FC 1) was left blank, FC2 – Sho, FC3 – Fx, FC4 – PrP. (B) 5 different RNA oligonucleotides were used as analytes in this study, SF, SM, SC, Xav, PG as described in Methods. These diagrams are based on similar representations appearing in several publications dealing with SPR.

SPR is widely used to study the binding of ligands to nucleic acid. Generally the nucleic acid is immobilized on the chip surface. This is the case for G-quadruplex nucleic acid as well as other types of nucleic acid, although it has been reported that immobilization can destabilize the quadruplex structure (insert Zeng et al, J. Mol. Recognit. 2005). In this study the peptides were immobilized on the chip surface and the nucleic acids were flowed over the surface.

In contrast to spectroscopic methods, incident light is not absorbed by the solution. This means that the scan is not affected by issues such as light scattering. Some problems that may arise are non-specific binding to the cell surface, as occurred in this study, and mass transport limitations. The latter refers to the situation in which the analyte binds more quickly to the ligand than it can be delivered to the chip surface. The experimental protocol can be adjusted to mitigate these sources of potential error; for example, an increased flow rate together with reduced ligand concentration may alleviate mass transport concerns. SPR is a particularly useful technique for the study of 1:1 binding as it provides real time information regarding association and dissociation of the analyte and ligand. More complex binding behaviour is less amenable to study using this technique (Roy et al., 2007).

9.1.3 Background - Fluorescence

In fluorescence spectroscopy, light of a chosen wavelength is used to excite the molecule or chromophore of interest to a higher electronic state, then, upon relaxation to the ground state there is an emission of detectable light. More precisely, a photon of light having a specific energy is absorbed by an electron of the molecule, promoting it from its ground state energy to the first excited state. Each electronic energy state has a number of associated accessible vibrational energy levels. While in this state some energy may be lost as the electron moves into a different vibrational energy state, for example, some vibrational energy may be transferred to the surrounding environment as heat. Accordingly, when the electron returns to its ground state energy it releases a photon, which generally has lower energy (thus greater wavelength) than the photon originally absorbed. This is portrayed schematically in Figure 9-5. The light emitted in this process is detected in an emission scan and provides information about the conformation or environment of the fluorophore.

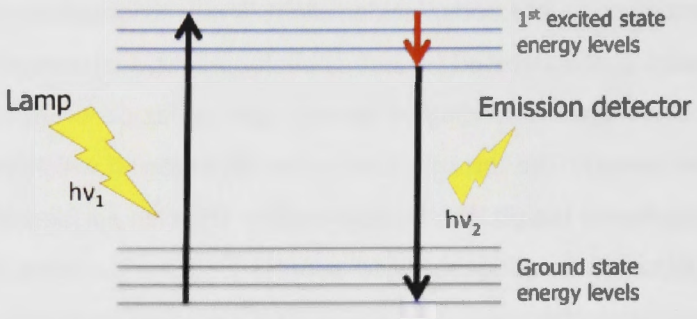


Figure 9-5. Fluorescence energy diagram
A Jablonski diagram illustrating fluorescence. An electron in the molecule of interest absorbs a photon of light with a frequency ν_1 and is excited from its ground state energy level into its first excited state. Vibrational energy levels are depicted as blue lines. In this example, the electron changes its vibrational level (loses some energy – red arrow) and then releases a photon of light, which has a lower energy and frequency (ν_2) than that originally absorbed.

As nucleic acid does not have fluorescent properties, the synthetic adenine analogue, 2-aminopurine ('2-AP') depicted in Figure 9-6, which does have fluorescent properties, can be incorporated into a nucleic acid sequence and used as a fluorescent probe.

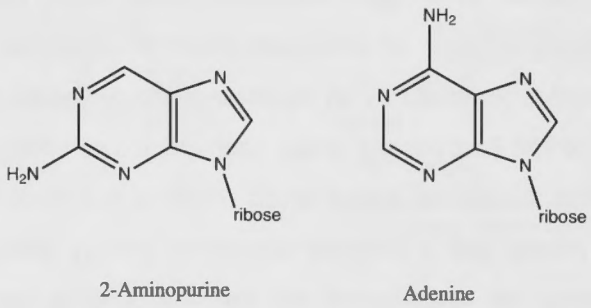


Figure 9-6. 2-aminopurine
The synthetic adenine analogue 2-aminopurine (2-AP) on the left compared with adenine on the right.

2-AP is excited by UV light with a wavelength ~ 310 nm and emits light with a wavelength of ~ 370 nm. The fluorescence of 2-AP is quenched when it stacks against nucleic acid bases or aromatic residues in proteins (Jean and Hall, 2001). An increase in fluorescence indicates that a greater population of molecules has 2-AP exposed to solvent, while a decrease in fluorescence indicates that 2-AP is stacking against other nucleic bases or aromatic residues of a binding partner. Accordingly, a change in

fluorescence implies a change in the environment of 2-AP from which it can be inferred that the nucleic acid has undergone some conformational change.

9.1.4 Binding Constants

The primary aim of binding studies is to describe the binding of the 2 species of interest, including the number of binding sites involved (stoichiometry), the affinity which the binding partners have for a particular site (Association/Dissociation constants), and where there is more than one binding site the affect which binding at that site has on binding at the other site/s (positive or negative co-operativity). The theoretical basis of this analysis will now be described. The terms used in this discussion are defined in Table 9-1.

Table 9-1. Definitions and mass balance equations used in binding analysis

$[RP]$	Concentration of the bound complex (concentration of occupied binding sites)
$[R_T]$	Total concentration of RNA $[R_T] = [R_F] + [R_B]$
$[R_F]$	Concentration of free RNA $[R_F] = [R_T] - [RP]$
$[R_B]$	Concentration of bound RNA $[R_B] = [R_T] - [R_F] = [RP]$
$[P_T]$	Total concentration of Peptide $[P_T] = [P_F] + [P_B]$
$[P_F]$	Concentration of free Peptide $[P_F] = [P_T] - [RP]$
$[P_B]$	Concentration of bound Peptide $[P_B] = [P_T] - [P_F] = [RP]$
K_a	Association constant
K_d	Dissociation constant
ϕ	Ratio of occupied binding sites to potential binding sites $[RP]/[R_T]$

In the current studies there will be 3 species at equilibrium; free RNA, free peptide and the bound complex. Free RNA and free peptide combine to form a complex at a rate described as the forward/on rate constant (k_{on}) while the complex will dissociate into free RNA and free peptide at a rate described as the off rate constant (k_{off}). The stoichiometry of the reaction is given by n_1, n_2 and n_3 - Eqn 4.

$$n_1[R] + n_2[P] \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} n_3[RP] \qquad (4)$$

where n_1 is the molar ratio of the RNA
n_2 is the molar ratio of peptide

n_3 is the molar ratio of the complex
$[R]$ is the concentration of RNA
$[P]$ is the concentration of peptide
$[RP]$ is the concentration to RNA-peptide complex

At equilibrium k_{on} and k_{off} are equal and the average concentrations of reactants and products do not change. The association constant is then given by the ratio of the concentration of product to the concentration of reactants - Eqn 5. The dissociation constant K_d is simply the inverse of the association constant (Eqn 6) but is more convenient as its units are moles/L (molarity 'M') rather than the inverse (M^{-1}).

$$K_a = \frac{[RP]^{n_3}}{[R]^{n_1}[P]^{n_2}} \tag{5}$$

$$K_d = \frac{[R]^{n_1}[P]^{n_2}}{[RP]^{n_3}} \tag{6}$$

At equilibrium, under conditions of constant pressure, the Gibbs free energy of the reaction is at a minimum and is related to the association constant as shown by Eqn 7, where R is the universal gas constant and T is the temperature.

$$\Delta G = -RT\ln K_a = RT\ln K_d \tag{7}$$

While the concentrations of total RNA and total peptide are known, the concentration of the complex and accordingly the concentrations of free RNA and free peptide are not known. These must be derived from an experimental technique such as fluorescence spectroscopy, SPR or CD that allows comparison of the bound and unbound states of the species.

Applying the definitions and mass balance equations from Table 9-1 it can be seen that Eqn 6 can also be expressed as:

$$K_d = \frac{[R_F][P_F]}{[RP]} = \frac{([R_T]-[RP])([P_T]-[RP])}{[RP]} \tag{8}$$

The common quadratic equation linking changes in observed binding behaviour to the dissociation constant (Eqn 9) can be derived from Eqn 8. Eqn 9 corrects the free ligand (titrant) concentration by taking account of the ligand which is bound. This model was used in analysis of some of the fluorescence experiments, discussed in Section 9.3.3.

$$F = 1 + \left(\frac{F_b}{F_f} - 1\right)\left(\frac{(K_d + [P_T] + [\mathrm{RNA}]) - \sqrt{(K_d + [P_T] + [\mathrm{RNA}])^2 - 4[\mathrm{RNA}][P_T]}}{2[\mathrm{RNA}]}\right) \quad (9)$$

where F = observed fluorescence at 370 nm
F_b = fluorescence at saturation of peptide
F_f = fluorescence of the RNA alone
$[P_t]$ = the total concentration of added peptide
[RNA] = the concentration of RNA

In situations where the ligand is largely in excess of its binding partner the assumption can be made that the concentration of free ligand is approximately the same as the concentration of total ligand. If this assumption can be made the binding equation can be greatly simplified to the Langmuir binding isotherm:

$$B_P = \frac{B_{\max}[P_T]}{K_d + [P_T]} \quad (10)$$

where B_P = the amount of ligand (peptide) bound at any peptide concentration
B_{max} = maximum ligand binding which occurs when all RNA binding sites are occupied
P_T = Total ligand (peptide) concentration

In the SPR experiments different concentrations of RNA were titrated into a fixed concentration of peptide and accordingly the Langmuir binding isotherm for this interaction has the form:

$$R_{eq} = \frac{R_{\max}[\mathrm{RNA}]}{K_d + [\mathrm{RNA}]} \quad (11)$$

where R_{eq} = the Response Units (RU) at equilibrium
R_{max} = Maximum Response
[RNA] = concentration of RNA

If there is more than one binding site and where binding of a ligand at one site has some affect on binding at the other site, the Hill binding equation which introduces a co-operativity coefficient 'n' may be used. In the case of positive co-operativity (n > 1) the binding on one ligand facilitates the binding of the next ligand. If there is negative co-operativity (n < 1) the binding of one ligand hinders the binding of the next ligand. The basic Hill equation is essentially a generalized version of the Langmuir binding isotherm with an additional co-operativity parameter:

$$B_P = \frac{B_{\max}[P_T]^n}{K_d{}^n + [P_T]^n} \tag{12}$$

In the case of sequential binding sites the Hill equation may be modified to introduce additional terms. The biphasic binding isotherm produced in some of the fluorescence experiments were analysed using the following modified Hill equation:

$$F_f = \frac{F_{\max 1}[P_T]^{n1}}{K_{d1}{}^{n1} + [P_T]^{n1}} + \frac{F_{\max 2}[P_T]^{n2}}{K_{d2}{}^{n2} + [P_T]^{n2}} \tag{13}$$

where F_f = fluorescence signal of SF RNA in the absence of peptide

F_{max1} = maximum increase in the fluorescence signal as a result of peptide addition
F_{max2} = maximum decrease in the fluorescence signal as a result of peptide addition (-ve number)
$[P_T]$ = total concentration of the peptide
n_1 = Hill coefficient 1
n_2 = Hill coefficient 2
K_{d1}= Dissociation constant 1
K_{d2}= Dissociation constant 2

The literature in this area reveals that the quadratic form of the 1:1 binding model (Eqn 9) is widely used, including in binding studies involving the RGG box of FMRP (Zanotti et al., 2006, Menon and Mihailescu, 2007, Menon et al., 2008). The Langmuir binding model (Eqn 11) is generally used in the analysis of SPR results, as it is assumed the analyte (titrant) concentration doesn't change as a result of binding i.e. analyte is being constantly replenished as it flows over the immobilized species during the injection phase. When these models are used to plot proportional saturation against the concentration of

the titrant a hyperbolic curve is formed. These models are clearly inappropriate for generating a K_d in the situation where the experimental data plotted in this manner, does not form a hyperbolic curve. This may be found if the binding assay is not performed under equilibrium conditions (e.g. if all ligand molecules bind immediately and there is no free ligand in solution) or if the mode of binding is more complex than a simple 1:1 association with equivalent binding sites.

Obviously, the 1:1 model does not, without modification, take account of the presence of more than one binding site, the affect of co-operativity or the possibility of competition between specific and non-specific binding. However, it is known that for many protein-nucleic acid systems binding tends to be complicated. There may be 2 or more binding sites, specific and non-specific binding and attractive or repulsive interactions between ligands. Significant effort has been devoted to developing a better understanding of the complexities of nucleic acid-protein binding and models for interpreting binding data (Mcghee and Von Hippel, 1974, Epstein, 1978, Sackett and Saroff, 1996, Tsodikov et al., 2001). Although the methods that have evolved are more complicated than the application of a simple 1:1 binding model (so often seen in the literature) the use of an appropriate binding model is essential to arrive at a meaningful description of the interaction and calculation of binding constants.

A Scatchard plot (Scatchard, 1949) which plots ϕ (ratio of occupied sites)/free ligand on the y-axis against ϕ on the x-axis will indicate whether a simple 1:1 binding model is appropriate. Deviations from linearity reveal more than one class of binding site and/or ligand-ligand interactions (Mcghee and Von Hippel, 1974). The Scatchard plot can also be used to obtain a rough estimate of B_{max} (the x-intercept) and K_d (the negative reciprocal of the slope).

Least squares regression is used to determine the value of K_d that provides the best fit of the model to the experimental data. A number of graphing/curve fitting packages are available for performing least squares regression, including Kaleidagraph (Synergy Software), Prism (Graphpad Software Inc.) and Microsoft Excel, while the Biaevaluation software and Scrubber2 are tailored for analysis of SPR binding data. Although the binding model may appear to fit the experimental data well, (e.g. the squared correlation

coefficient (R^2) value is close to 1) it is also necessary to determine the confidence interval of the derived parameters. The confidence interval defines the degree of variation that may be tolerated in a parameter value without changing the binding curve. A K_d value could potentially vary over a broad range while still producing a close fitting curve. Although the confidence interval of derived parameters such as K_d is infrequently reported in the literature, and error is not reported at all in some cases, these omissions do create some doubt in the reader's mind as to the reliability of reported results.

9.1.5 Aims of this chapter

The aim of the work in this chapter is to determine whether Sho's RGG box can bind G-quadruplex RNA and to determine its relative affinity compared to the FRMP RGG box and a region of the N-terminus of PrP that is capable of RNA binding.

9.2 METHODS

9.2.1 RNA and Peptides

Five RNA oligonucleotides (Table 9-2) were synthesized by University Core DNA Services (UCDNA Services) at the University of Calgary, AB Canada. The SF oligonucleotide was produced both with and without substitution of A15 with 2-aminopurine (SF_AP). The peptides (Table 9-2) were synthesized at the Biomolecular Resource Facility at the John Curtin School of Medical Research. Three peptides (Sho, Fx and PrP) were produced both with and without a biotin label at the N-terminus and spacer consisting of SGSG (Biotin-SGSG-peptide) and were used in the SPR experiments.

Table 9-2. RNA oligonucleotides and Peptides

RNA	
SF	**GGCUGGUGAUUGGAAGGGAGGGAGGUGGCCAGCC**
	A region of the 3' UTR of the Semaphorin mRNA discovered to form a G-quadruplex and to have high affinity for the FMRP RGG box (Darnell et al., 2001) modified to reduce the possibility of more than one G-quadruplex conformer (Menon and Mihailescu, 2007)
SF_AP	**GGCUGGUGAUUGGAAPGGGAGGGAGGUGGCCAGCC**
	The SF RNA sequence modified by replacing A15 with 2-aminopurine, which has fluorescence properties.
SM	**GGCUGGUGAUUUGAAGUGAUGGAGUUGGCCAGCC**
	The SF transcript modified by changing 4 guanines to uridine, intended to disrupt the tetrad stack.
SC	**GGCUGCGGUGUGGAAGGUGAGGCUGGGUUGCGCAGCU**

	SELEX selected RNA aptamer (Sc1) discovered to form a G-quadruplex and to have high affinity for the FMRP RGG box (Darnell et al., 2001).
Xav	**UGAGGGUGGGUAGGGUGGGUAA**
	The RNA transcript used as a model of a parallel G-quadruplex in the MD studies.
PolyG	**UUGGGGGGGGGGGGGGGGGUU**
	A polyguanine (PolyG) transcript which could be expected to form more than one type of G-quadruplex.
Peptide	
Sho	**KGGRGGARGSARGGVRGGARG**
	Sho residues 25-56 containing the RGG box.
Lys	**KGGKGGAKGSAKGGVKGGAKG**
	The Sho peptide is mutated by replacing the Arg residues with Lys.
Neg	**KGGEGGAEGSAEGGVEGGAEG**
	The Sho peptide is mutated by replacing the Arg residues with Glu.
Scr	**GGKGRGAGGRAGRVGGRARSG**
	This peptide contains all of the Sho residues in a scrambled order.
Fx	**RRGDGRRRGGGGRGQGGRGRGGGFKGNDDHSR**
	The Fx peptide (res 527-558) containing the RGG box motif.
PrP	**KKRPKPGGWNTGGSRYPGQGSPGGNRYPPQ**
	PrP (res 23-52) previously found to bind G-quadruplex RNA (Weiss et al., 1997).

The RNA oligonucleotides were solvated in a buffer of 10 mM Tris HCl pH 7.4 and unless otherwise stated either 100 or 150 mM KCl. The RNA buffer solutions were heated at 95° C for 15 minutes and then cooled to room temperature, over at least 30 minutes. The RNA stock solution was then stored in 4 μL aliquots at -70° C. The concentration of the RNA solutions was confirmed by absorbance spectroscopy scans using a Nanodrop Spectrophotometer.

The lyopholized peptides were solvated in nuclease free water to a concentration of ~ 1 mM and stored at -20° C. Samples of the peptide stock solution were sent for amino acid analysis at the Australian Proteomics Facility, to confirm stock concentration.

9.2.2 Circular dichroism

The CD spectra were obtained using a Chirascan CD spectrometer (Applied Photophysics Limited) at the Research School of Chemistry (ANU). A 1 mm path quartz cuvette was used and data were collected over the range 220-320 nm for the RNA oligonucleotides and 190-260 nm for the peptides, at 1 nm intervals. The spectral bandwidth was set at 1nm and data were obtained at a rate of 0.5 s per point. The CD spectra were averaged

over either 2 or 3 scans and a scan of the buffer (100mM Tris HCl, pH 7.4, 100 mM KCl) was subtracted. The curves were smoothed using the inbuilt Savitsky-Golay smoothing algorithm. The CD measurements were carried out at room temperature unless otherwise specified. An absorbance spectrum was taken at the same time as the CD scan. The RNA concentrations were in the range 20-30 μM which produced a maximum absorbance of ~ 1 absorbance unit. The peptide concentrations were ~ 50 μM for the individual scans and ranged from a molar ratio of 1:1 to 1:11 for the RNA-peptide scans. The CD results are expressed in ellipticity units, θ(mdeg). The final spectra were normalized to have zero ellipticity at 320 nm.

9.2.3 Fluorescence

Fluorescence emission spectra of the SF_AP RNA (2-AP at the 15th position) in a buffer of cacodylic acid (10 mM) or Tris HCl (10 mM) pH 7.4, with KCl (100 or 150 mM) and varying concentrations of the peptide were acquired at room temperature (~20 C) with a Perkin Elmer LSB50 spectrophotometer, at the JCSMR. Samples were excited at 310 nm and the emission spectra were collected from 330 nm to 430 nm. After each peptide addition the sample was allowed to equilibrate at room temperature for 15 minutes prior to being scanned. The buffer signal was subtracted from the scans.

Some experiments were performed in tandem in order to have the closest possible comparison of the binding of 2 different peptides. This involved two experiments being performed at the same time, using 2 identical cuvettes. The RNA stock was divided into the 2 cuvettes and the same volume of the 2 different peptides was added to each of the cuvettes at the same time, both were allowed to equilibrate in the same environment and scans were performed as close in time as possible.

Three different binding models were tested with the experimental data. Where the binding isotherm clearly demonstrated biphasic behaviour the modified Hill equation (Eqn 13) was used. In other cases the simple Hill equation (Eqn 12) was used and compared to the 1:1 binding (Eqn 9).

Least squares regression analysis to derive values of K_d and Hill co-efficients was performed initially using an Excel spread sheet. The results of these calculations were

then used as first guesses for input to Kaleidagraph (Synergy Software) which calculated values of K_d and associated errors presented in the Results section.

9.2.4 Surface Plasmon Resonance

SPR experiments were performed on a BIAcore T100 instrument, in the laboratory of Professor Nick Dixon, at the School of Chemistry, University of Wollongong over the course of a 10 day visit. The 3 biotinylated peptides, Sho, Fx and PrP described in Table 9-2 ('the ligands') were immobilized on the surface of a streptavadin coated chip (SA chip, BIAcore). The peptides were produced with a spacer between the peptide and the biotin tag (Biotin-SGSG-peptide) in order to minimise any effect on peptide flexibility. The 4 flow cells of the SA Chip were utilized as follows: Flow Cell 1 no peptide (control), Flow cell 2 – Sho peptide 955 RU, Flow cell 3 – FMR peptide 1060 RU, Flow cell 4 – PrP peptide 898 RU. The 5 RNA oligonucleotides ('the analytes') in a range of concentrations (5 – 500 nM) were prepared in buffer (10 mM Tris HCl (pH 7.4), 150 mM KCl and 0.005% P20) and injected at 5 μL/min for 500 s. Buffer of the same composition was used as the running buffer. Dissociation was monitored initially for 500 s. The sensor chip was regenerated by injection of 5 μL of 50 mM NaOH and 1M NaCl and after 120 s an injection of 5 μL of 1 M $MgCl_2$. The background signal measured in flow cell 1 was subtracted from the signals read in the other flow cells.

BIAevaluation software supplied with the BIAcore system as well as Scrubber2 software (supplied courtesy of Dr Tom Morton of Biologic Software Pty Ltd) was used to process the data.

A significant change in signal was recorded in the blank cell as well as in FC 2-4. Generally this is an indication that the analyte is binding to the surface of the SA chip in addition to binding to the immobilized peptides. This was not anticipated in this experiment as the dextran matrix of the chip is negatively charged at pH 7 as is the RNA, accordingly, one might expect some degree of repulsion between the surface and the analyte. The surface of the chip was washed with biotin in an attempt to ameliorate this affect, but this had little impact. Another possible explanation for this effect is that a refractive index artifact can result from the analytes having a higher refractive index than the running buffer. As a high level of immobilized ligand in Flow cells 2-4 replaces the volume found in Flow cell 1

which has no immobilized ligand, the refractive index effect can be more pronounced in the blank cell (Van Der Merwe, 2001). In analyzing these experiments the signal recorded in the blank cell is subtracted from the signal recorded in the other flow cells. As flow cells 2-4 were saturated with peptide it could be expected that the level of non-specific binding or refractive index artefact would be less than that on the blank cell. Subtracting the signal from the blank cell may therefore lead to underestimation of peptide-RNA binding. The significant level of chip binding and the very long dissociation phase has limited the quantitative interpretation of these results.

As the SPR technique provides real time information regarding the association and dissociation of the species it is possible to determine the on and off rates of the interaction. It is therefore possible to calculate the affinity (K_d) directly from k_{on} and k_{off} ($K_d = k_{off}/k_{on}$). Alternatively, the affinity of the interactions can be assessed through data collected at equilibrium. In this study K_d is estimated by assessing the responses and concentrations of analyte at equilibrium.

The RU values obtained at equilibrium were fitted to the Langmuir binding isotherm (Eqn 11) The R_{eq} and R_{max} values were obtained using Scrubber2. Although Scrubber2 could be used for regression analysis, for consistency Kaleidagraph was used for this process with initial guesses being those values obtained from Scrubber2.

9.3 RESULTS AND DISCUSSION

9.3.1 Circular dichroism

The peptides and RNA nucleotides were tested for structural elements using circular dichroism. The spectra obtained for the peptides are presented in Figure 9-7 and the spectra for the RNA oligonucleotides are shown in Figure 9-8.

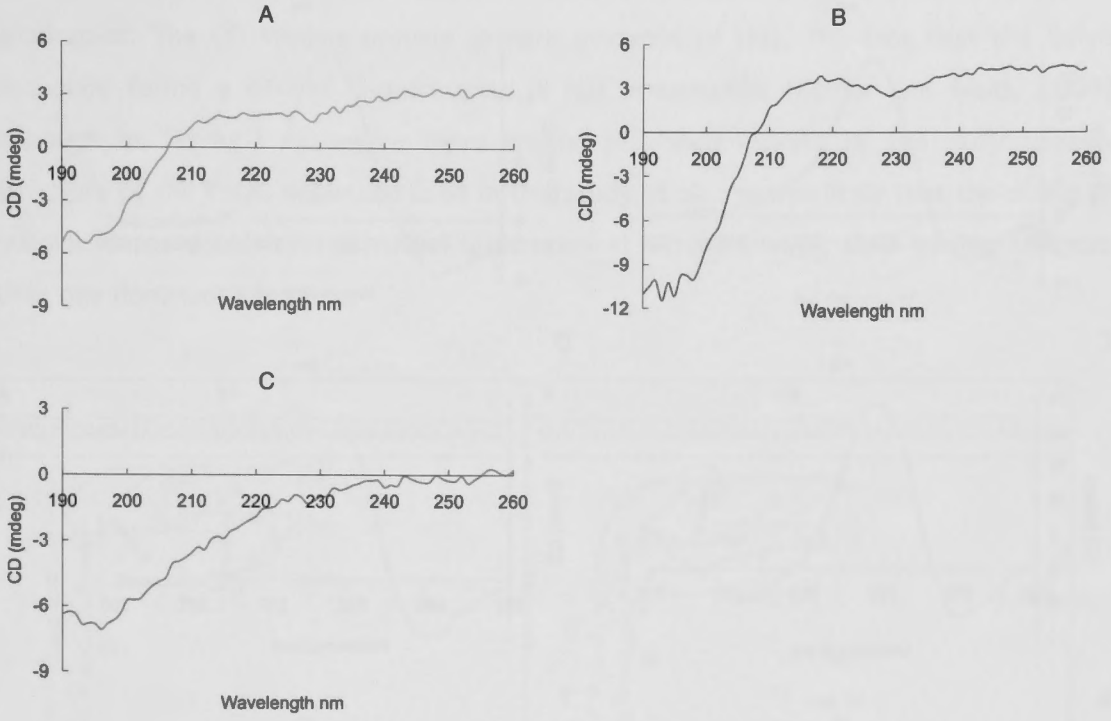


Figure 9-7. Circular dichroism studies of the Sho, Fx and PrP peptides
CD spectra of the 3 peptides in an aqueous solution (A) Sho 46 μM (B) Fx 57 μM (C) PrP 66 μM.

The peptide spectra show no evidence of dominant α-helix or β-sheet secondary structure. Rather, all are consistent with a disordered domain that interconverts between different conformations. While all 3 spectra have a negative peak at ~ 195 nm there are differences in the spectra between 205 nn and 260 nm. In this region Fx and Sho are similar in that both rise into a shoulder around 210 nm, Fx has a weak maximum ~ 218 nm and both have slight dip at 229 nm. These features are indicative of turns (Sreerama and Woody, 2004) and suggest the presence of P_{II} conformations (also referred to as β_{II} structure) which has a similar spectra to disordered regions. The PrP spectra is characteristic of disordered protein domains and also bears similarity to the CD spectra of proteins known to have β_{II} structure (Sreerama and Woody, 2003).

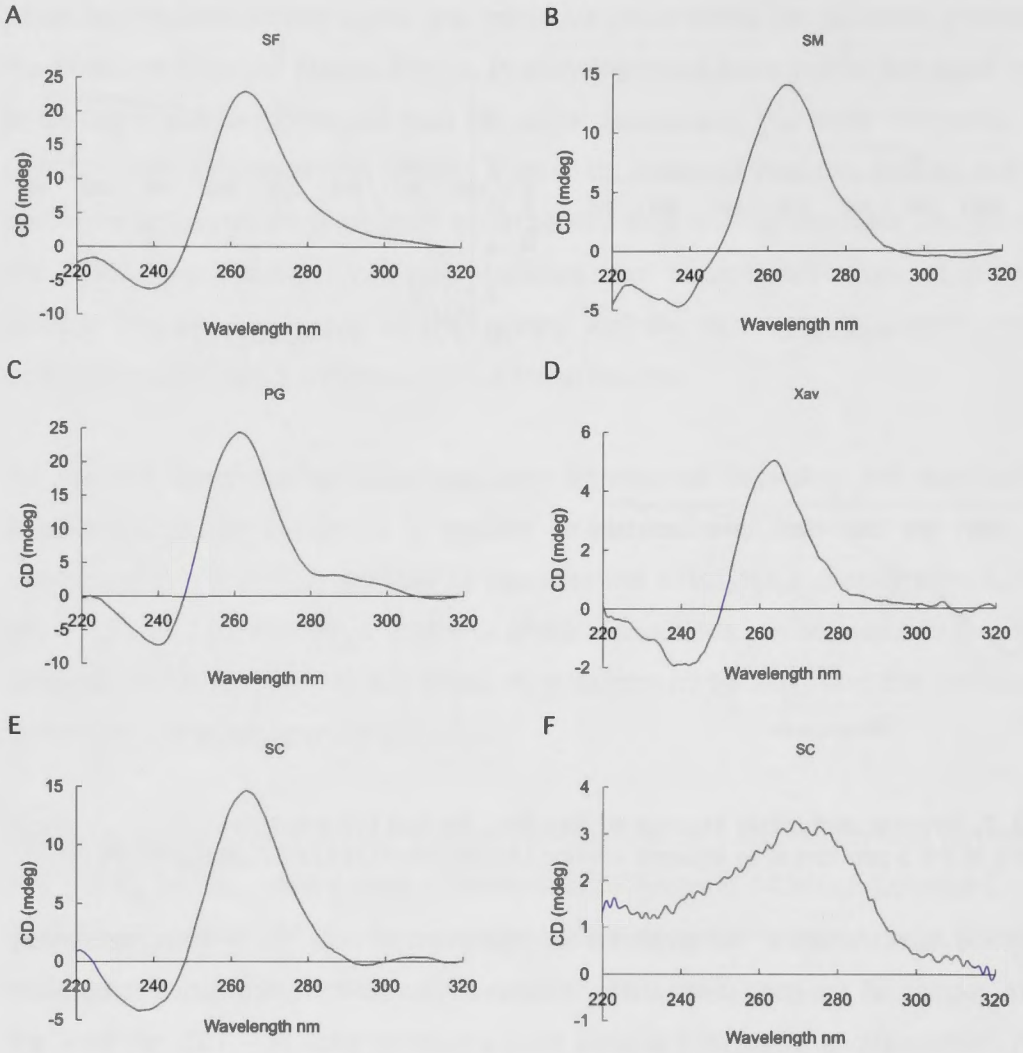


Figure 9-8. Circular dichroism studies of the RNA oligonucleotides
(A-E) CD spectra of the 5 RNA oligonucleotides, SF (29 uM), SM (26 μM), PolyG (26 μM), Xav (10 μM), and SC (25 μM) in 10 mM Tris HCl buffer pH 7.4 and 100 mM KCl. (F) CD spectra of SC RNA (1.4 μM) in a buffer of 10 mM Tris HCl without KCl.

All of the RNA oligonucleotides, in a buffer containing K^+ have a positive peak at ~264 nm and a weak negative peak at ~240 nm consistent with parallel G-quadruplex formation (Dapic et al., 2003). For comparison the spectra for SC RNA in the absence of K^+ is also shown and reveals a band at 280 nm characteristic of unstructured RNA and the absence of a prominent minimum at ~240 nm. The formation of a parallel G-quadruplex in SF and SC has been predicted previously (see Figure 9-9). However, it is surprising to find that the mutated form of SF, which I originally designed with the intention of disrupting the quadruplex structure, also produced a similar CD spectra. MD studies of the model RNA

quadrulex Xav provided *in silico* confirmation that this sequence forms a stable parallel G-quadruplex. The CD studies provide *in vitro* evidence of this. The fact that the PolyG sequence forms a parallel G-quadruplex is not unexpected (Kumar and Maiti, 2008) although as far as I am aware there are no published reports of the likely tertiary structure of the PolyG sequence used in this study. It also seems likely that the string of guanine bases could form a parallel quadruplex in different ways, thus leading to more than one dominant structure.

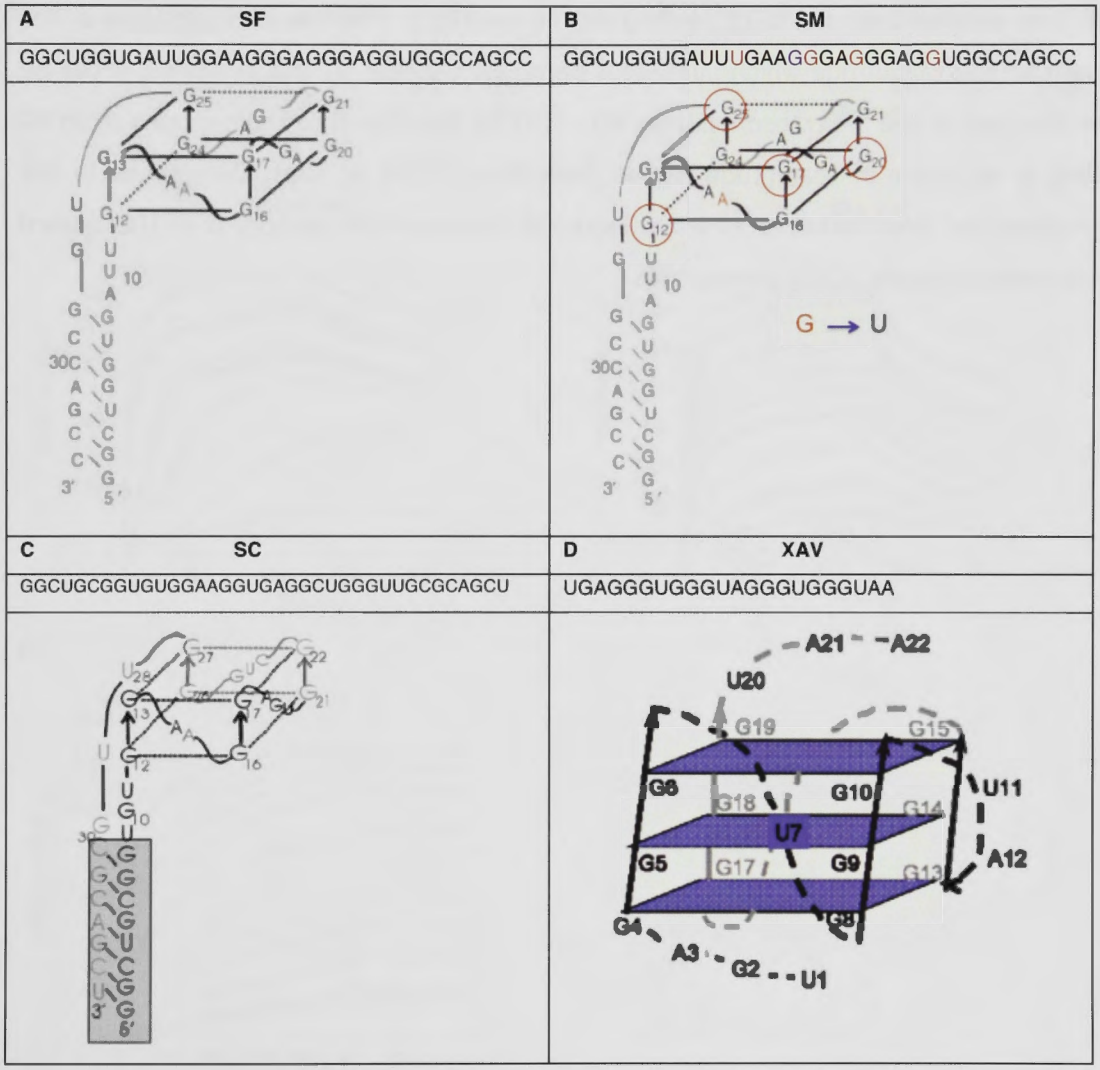


Figure 9-9. Proposed structures of the RNA oligonucleotides
(A) The structure of the SF oligonucleotide proposed by the Mihailescu group (Menon and Mihailescu, 2007) (B) I designed a SF mutant by replacing the 4 circled guanines with uridine with the intention of disrupting the G-quadruplex. (C) The structure of the SC oligonucleotide proposed by by the Mihailescu group (Zanotti et al., 2006). (D) The Xav structure I designed for use in the *in silico* studies discussed in Chapters 4-8.

9.3.2 Binding assays – SPR results

Binding of the RNA oligonucleotides to the Sho, Fx and PrP peptides was investigated using SPR. The sensograms presented in Figures 9-10 to 9-12 reveal that all 3 peptides bind the RNA oligonucleotides in a concentration dependent manner. The association rate increases as the concentration of the analyte is increased, as expected. This is evidenced by a steepening of the upward slope of the sensogram with increasing analyte concentration. All 5 RNA oligonucleotides have differently shaped sensograms indicating that they vary in their mode of binding to the peptides. There is also variation in the amount of oligonucleotide which binds to a particular peptide. At the extremes, it can be seen that less of the SM transcript (max RU ~120 for Sho) binds to the peptides while PG binding is an order of magnitude higher (max RU ~ 1200 in Sho). Differences in the association and dissociation rates are reflected in changes in the steepness of the upward and downward slopes of the sensograms.

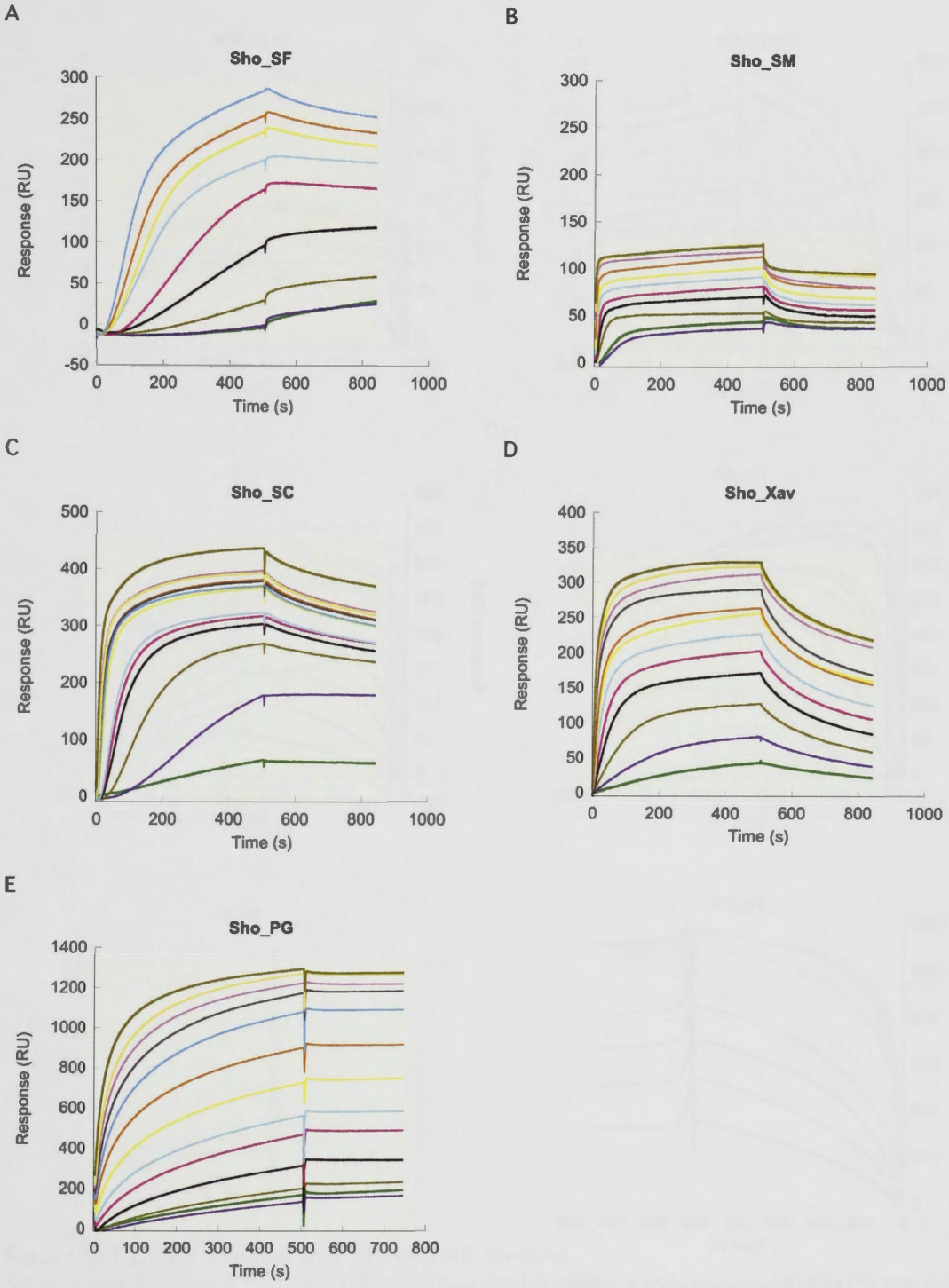


Figure 9-10. SPR Sensograms of Sho-RNA binding
Sensograms for binding of 5 RNA oligonucleotides to the immobilized Sho peptide. (A) SF, (B) SM, (C) SC, (D) Xav, (E) PolyG. Baseline time and response are detrmined at the commencement of RNA injection. The RNA oligonucleotides were injected in concentrations of 5 nM, 10 nM, 25 nM, 50 nM, 75 nM, 100 nM, 150 nM, 200 nM, 250 nM, 300 nM, 350 nM, 400 nM, and 500 nM.

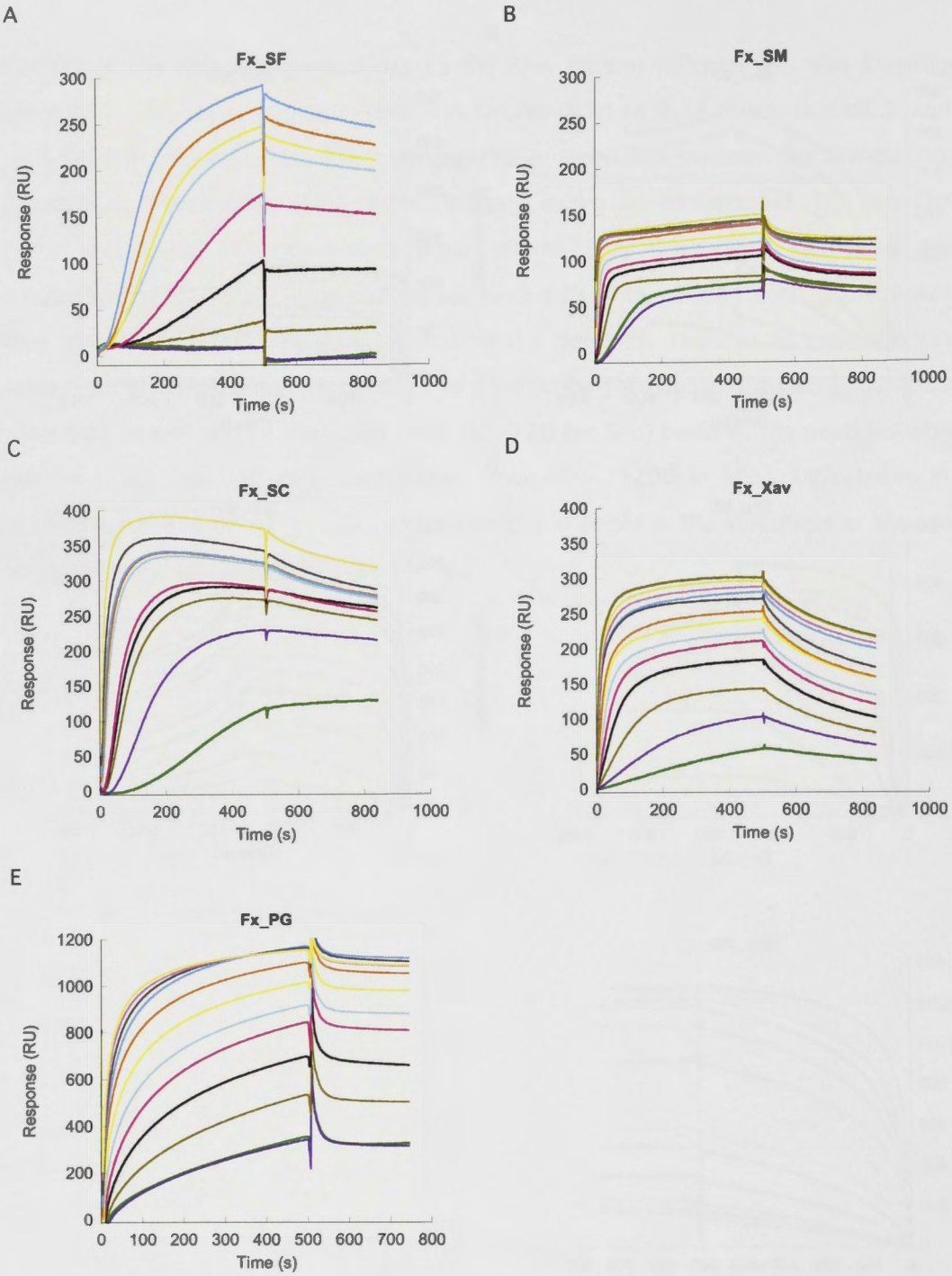


Figure 9-11. SPR Sensograms of Fx-RNA binding
Sensograms for binding of 5 RNA transcripts to the immobilized Fx peptide. (A) SF, (B) SM, (C) SC, (D) Xav, (E) PolyG. Baseline time and response are detrmined at the commencement of RNA injection. The RNA oligonucleotides were injected in concentrations of 5 nM, 10 nM, 25 nM, 50 nM, 75 nM, 100 nM, 150 nM, 200 nM, 250 nM, 300 nM, 350 nM, 400 nM, and 500 nM.

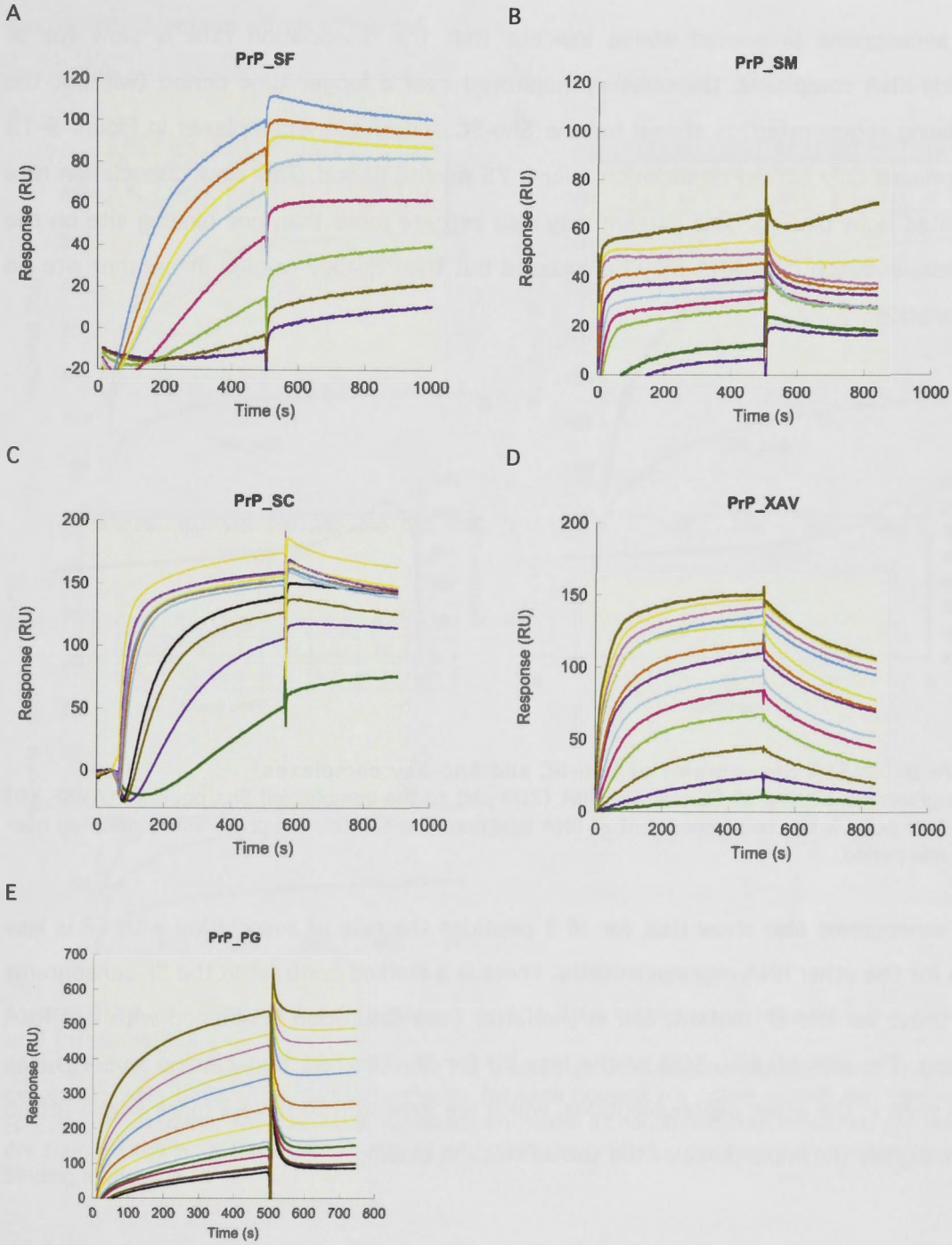


Figure 9-12. SPR Sensograms of PrP-RNA binding

Sensograms for binding of 5 RNA transcripts to the immobilized PrP peptide. (A) SF, (B) SM, (C) SC, (D) Xav, (E) PolyG. Baseline time and response are detrmined at the commencement of RNA injection. The RNA oligonucleotides were injected in concentrations of 5 nM, 10 nM, 25 nM, 50 nM, 75 nM, 100 nM, 150 nM, 200 nM, 250 nM, 300 nM, 350 nM, 400 nM, and 500 nM.

The sensograms presented above indicate that the dissociation rate is slow for all peptide-RNA complexes. Dissociation monitored over a longer time period (without the cell being regenerated) is shown for the Sho-SC and Sho-Xav complexes in Figure 9-13 and reveals only limited dissociation over a 75 minute period. This slow dissociation rate indicates tight binding. This pattern may also indicate more than one binding site on the peptide, in which case RNA may be released but then quickly re-bind at another site on the peptide.

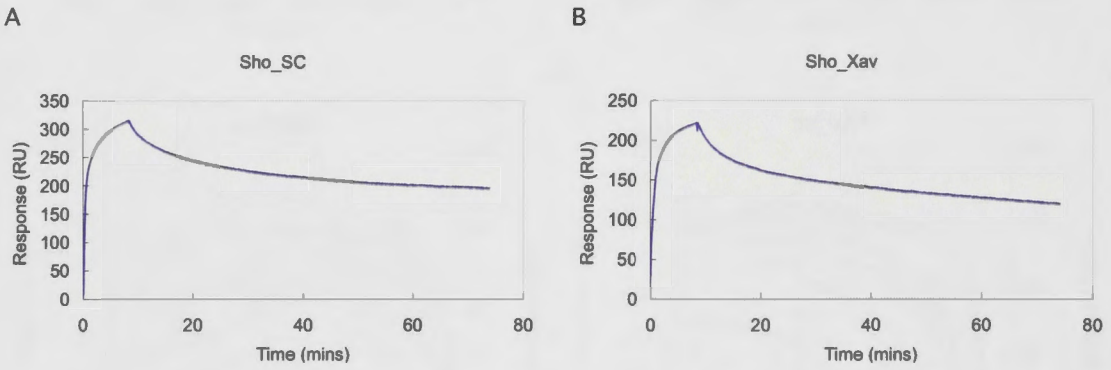


Figure 9-13. SPR Sensograms of Sho-SC and Sho-Xav complexes
Sensograms for binding of SC and Xav RNA (200 μM) to the immobilized Sho peptide. (A) SF, (B) Xav. Zero point is the commencement of RNA injection. The dissociation phase was monitored over a 75 min period.

The sensograms also show that for all 3 peptides the rate of association with SF is less than for the other RNA oligonucleotides. There is a marked contrast in the SF sensograms and those for the SF mutant, SM. In the latter case saturation is reached with less RNA binding. The max RU is ~ 30% of the max RU for SF. The fact that binding to SM is less than seen in the other oligonucleotides, which are expected to have a more stable tetrad stack signals the importance of the tetrad stack in binding.

Binding affinities were estimated using the last 50 s of the injection phase where it was most likely that equilibrium had been established. Curve fitting using the Langmuir binding model (Eqn 11) was appropriate for SC, Xav and SM and these curves are depicted in Figure 9-14. Given the different scale of PolyG binding and for ease of discussion, the binding curves will be presented in a format that allows comparison of

binding of the peptides to a particular RNA oligonucleotide rather than comparing the binding which occurs within a flow cell.

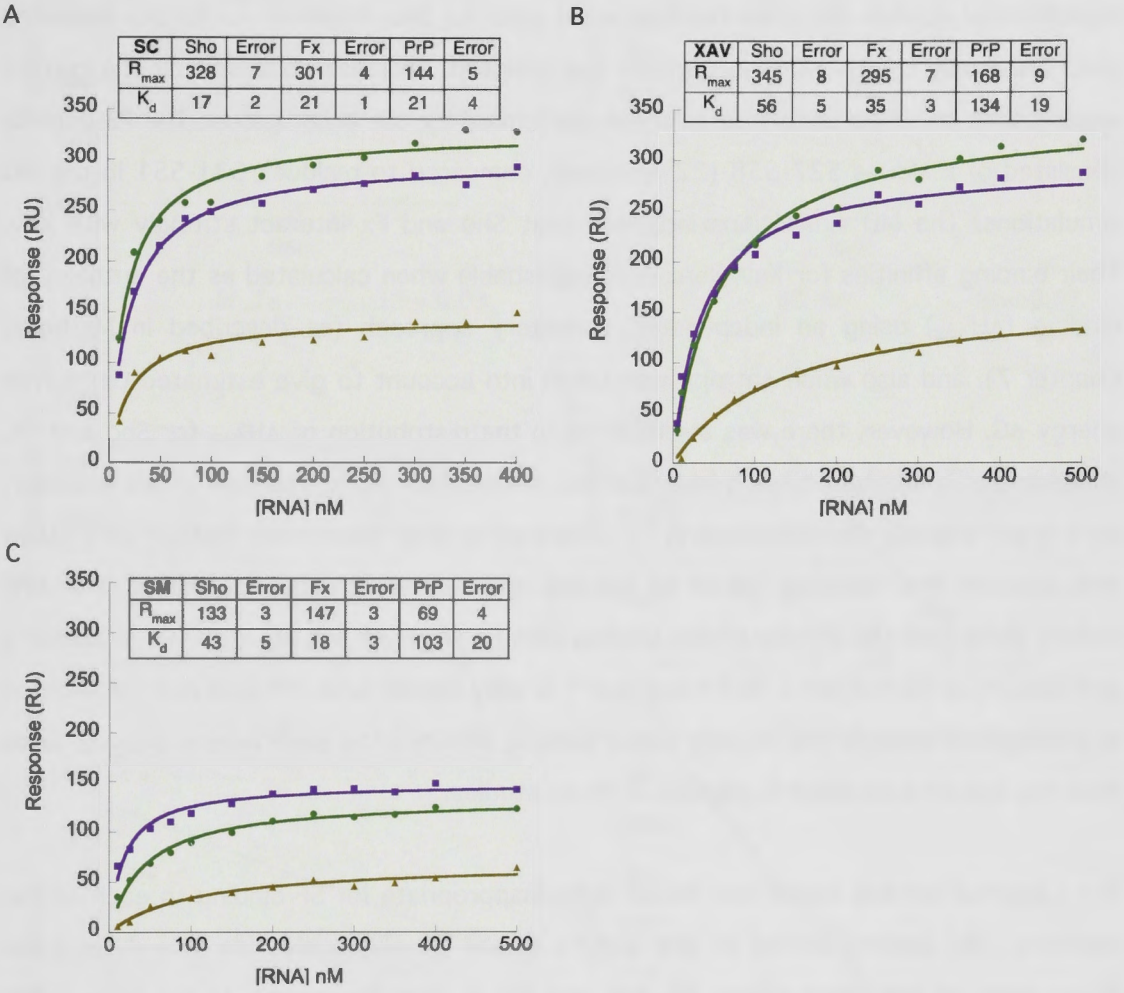


Figure 9-14. Equilibrium binding curves of the RNA oligonucleotides and the Sho, Fx and PrP peptides

Binding curves are collected and presented for a particular RNA oligonucleotide, noted in the first cell of the inset table. The binding isotherms for each peptide are colour coded; Sho (green), Fx (purple), PrP (brown). The calculated K_d values, expressed as nanomolar concentrations and max RU are found in the inset table. The solid lines represent the binding curve based on the Langmuir 1:1 binding model.

All 3 peptides bind strongly to SC with a K_d ~ 20 nM. Fx binds to SM with similar affinity ~ 18 nM and with less affinity to Xav (~35 nM). Sho does not bind as strongly to SM or Xav, but binding affinity is still considerable (43 nM and 56 nM, respectively). PrP binding to SM and Xav is weaker (103 nM and 134 nM, respectively) indicating that in these cases the RGG box is preferred to the RNA binding domain of PrP.

It is interesting to compare the binding of Sho and Fx to Xav as these interactions were also studied *in silico* (Chapters 6-8). In the MD simulations the Sho and Fx RGG box peptides both consisted of 21 residues (3 residues either side of the RGG box). In the experimental studies the same residues were used for Sho, however for Fx the sequence used previously by the Mihailescu group was adopted. This was done so that the current experiments could be compared to those performed by the other group. The Fx peptide consisted of residues 527-558 (32 residues), compared to residues 531-551 in the MD simulations. The MD simulations indicated that Sho and Fx interact strongly with Xav. Their binding affinities for Xav were indistinguishable when calculated as the enthalpy of binding (ΔH^{I}_{Bind}) using an independent trajectory approach (as described in Methods, Chapter 7), and also when entropy was taken into account to give estimated Gibbs free energy ΔG. However, there was a difference in the distribution of ΔH^{S}_{Bind} for Sho and Fx, showing that a stronger binding interface was established by Fx. This was offset however, by a larger peptide destabilization in Fx compared to Sho. When both factors were taken into account the resulting values of binding enthalpy were indistinguishable. The SPR results show that the affinity of Xav binding with Fx (K_d = 35 nM, ΔG = -10.01 kcal mol^{-1}) and Sho (K_d = 56 nM, ΔG = -9.73 kcal mol^{-1}) is very similar with < 1 kcal mol^{-1} difference in binding free energy. The slightly higher binding affinity of Fx seen here is likely to arise from the use of a modified Fx peptide in these studies.

The Langmuir binding model was found to be inappropriate for SF binding to each of the peptides. The binding curves of Sho and Fx to the SF oligonucleotide also differ from those seen for the other oligos; SC, Xav and SM, in that the curves do not have a flat region, that is, binding has not reached saturation. The unsuitability of a simple 1:1 binding model for the SF oligonucleotide is further demonstrated by Scatchard plots shown in Figure 9-15. For comparison, Scatchard plots of the SC binding isotherms are also presented. The linearity of the SC plots indicates that a 1:1 binding model is appropriate in the case of SC whereas the opposite is true for SF.

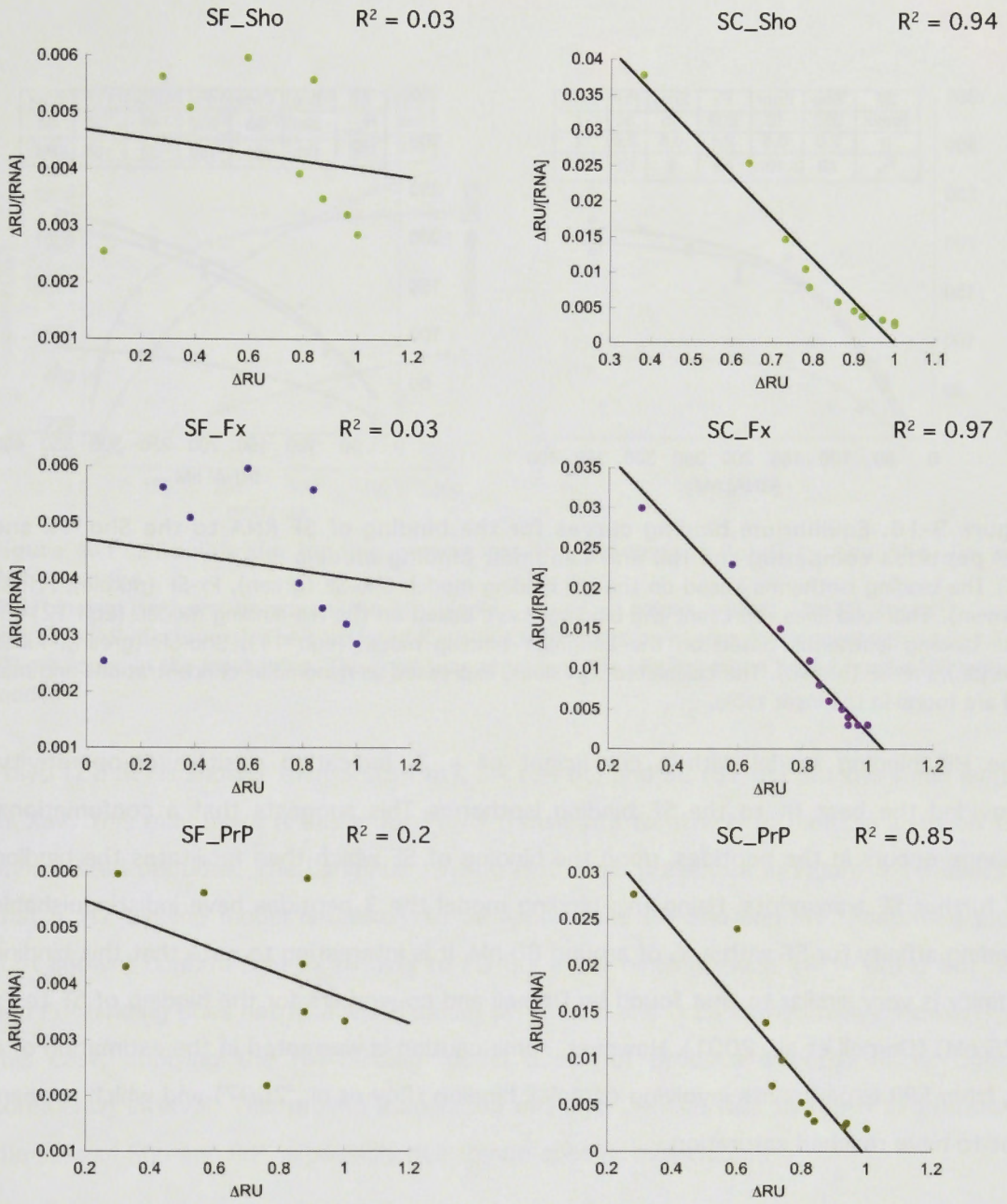


Figure 9-15. Scatchard plots for the binding of SF and SC RNA oligonucleotides and the Sho, Fx and PrP peptides

The Scatchard plots for binding of SF (LHS) and SC (RHS) to Sho, Fx and PrP. The least squares correlation co-efficient (R^2) for a linear fit is presented above the respective plot.

Given that the Langmuir binding model is inappropriate for SF binding, the Hill binding models for equivalent binding sites and for 2 sequential binding sites were also used to analyze the experimental data (Figure 9-16).

A

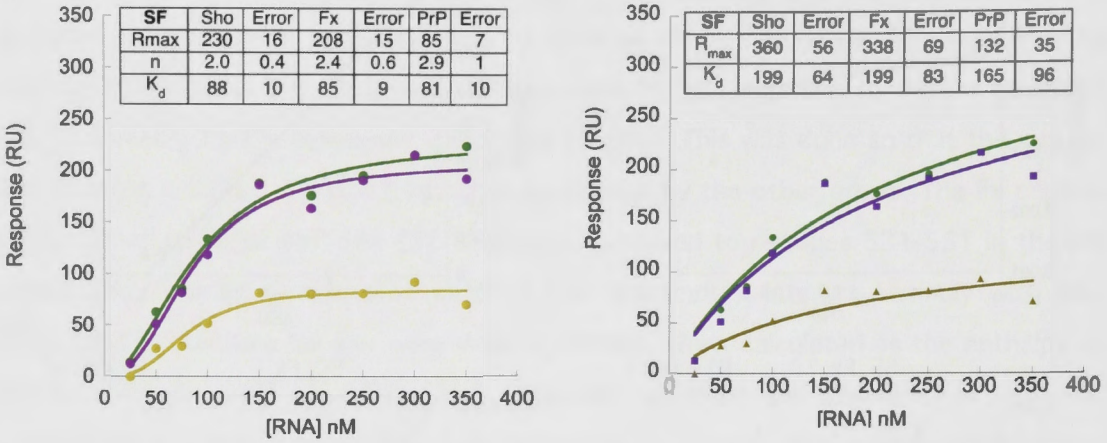


Figure 9-16. Equilibrium binding curves for the binding of SF RNA to the Sho, Fx and PrP peptides comparing the Hill and Langmuir binding models
(A) The binding isotherms based on the Hill binding model; Sho-SF (green), Fx-SF (purple), PrP-SF (brown). The solid lines represent the binding curve based on the Hill binding model. (Eqn 12) (B) The binding isotherms based on the Langmuir binding model (Eqn 11); Sho-SF (green), Fx-SF (purple), PrP-SF (brown). The calculated K_d values, expressed as nanomolar concentrations and max RU are found in the inset table.

The Hill binding model with a coefficient of ~ 2, indicating positive co-operativity, provided the best fit to the SF binding isotherms This suggests that a confomational change occurs in the peptides, upon the binding of SF which then facilitates the binding of further SF transcripts. Using this binding model the 3 peptides have indistinguishable binding affinity for SF with a K_d of around 80 nM. It is interesting to note that this binding affinity is very similar to that found by Darnell and co-workers for the binding of SF to Fx (75 nM) (Darnell et al., 2001). However, some caution is warranted in the estimation of a K_d from SPR experiments involving complex binding (Roy et al., 2007) and which appears not to have reached saturation.

As noted above the binding of PolyG is on a different scale to the binding of the other oligonucleotides. The binding isotherms presented in Figure 9-17 show that Fx alone reaches saturation.

A

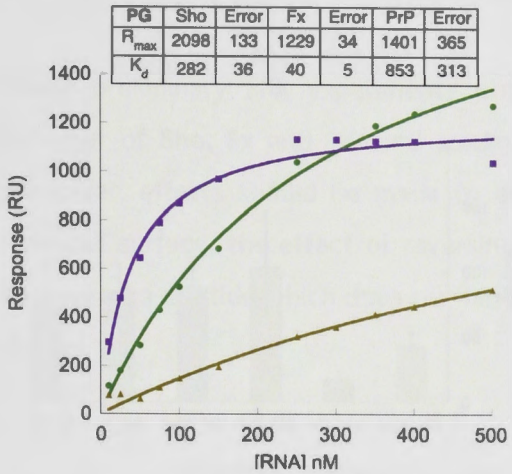


Figure 9-17. Equilibrium binding curves for the binding of PolyG to the Sho, Fx and PrP peptides

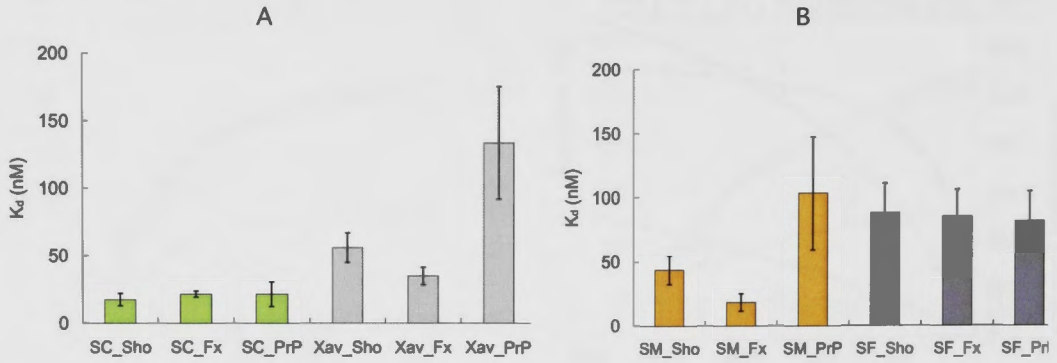
The binding isotherm based on the Hill binding model is colour coded; Sho-PG (green), Fx-PG (purple), PrP-PG (brown). The calculated K_d values, expressed as nanomolar concentrations and max RU are found in the inset table. The solid lines represent the binding curve based on the Hill binding model.

PolyG (22 nt) is shorter than SF (34 nt), SM (34 nt) and SC (37 nt) but the same length as Xav. This may make it easier for PolyG transcripts to attach to multiple sites on the immobilized peptides. The Langmuir binding isotherms presented in Figure 9-16 indicate that a 1:1 binding model is unlikely to be appropriate for Sho and PrP. Scatchard plots (not shown) confirm that Fx binding to PG fits a 1:1 binding model (R^2 = 0.94) but Sho and PrP binding does not fit a linear model (R^2 = 0.71 and 0.23 respectively). However, in this case, applying the Hill binding model does not produce a better-fit or tighter confidence interval. This prolific association of PolyG , which fails to reach saturation in the case of Sho and PrP, is possibly due to non-specific association.

The complex binding seen in the case of SF and PolyG RNA may be attributable to multiple binding sites on either the RNA or the peptide. Another possibility is that these RNA sequences form multiple structures. Future work should clarify this ambiguity.

Dissociation constants determined on the basis of the Langmuir binding isotherm, and the Hill model for SF binding, are plotted in Figure 9-18 together with the 95% confidence interval for the derived K_d values. The K_d for PG-Fx binding is noted, however, as the

binding models tested were not appropriate for PG-Sho and PG-PrP these values are not plotted.



	Sho K_d (nM)	95% CI (nM)	Fx K_d (nM)	95% C (nM)	PrP K_d (nM)	95% C (nM)
SC	17	2	21	1	21	4
Xav	56	5	35	3	134	19
SF	88	10	85	9	81	10
SM	43	5	18	3	69	4
PG	-	-	40	5	-	-

Figure 9-18. Dissociation constants for binding of the RNA oligonucleotides and the Sho, Fx and PrP peptides
Plots showing the calculated K_d, the error bars indicate the 95% confidence interval. (A) SC (green) and Xav (light grey) binding to the 3 peptides (Sho, Fx and PrP) calculated using the Langmuir binding isotherm. (B) SM (orange) binding to the 3 peptides (Sho, Fx and PrP) calculated using the Langmuir binding isotherm and SF (grey) binding using the Hill binding model.

Overall, the SPR results indicate that the Sho peptide is capable of binding G-quadruplex RNA with high affinity, the K_d for the binding of Sho with SC is calculated as 17± 2 nM, and the K_d for binding with Xav and SM is < 60 nM. The results also show that binding affinity for the different RNA oligonucleotides varies. Hence, the peptides are capable of specificity. Although all the oligonucleotides form a G-quadruplex structure, the loop nucleotides and number of tetrads clearly differentiate the quadruplex structures. Comparing the peptides, there are some differences in the order of binding affinity (Sho: SC>SM>Xav>SF>PG), (Fx: SM>SC>Xav>PG>SF), (PrP: SC>SF>SM>Xav>PG), and the binding strength for a particular nucleotide may vary considerably. For example, Xav ranks 3rd or 4th in binding affinity to each of the peptides but the calculated values, Sho, 56 ± 5

nM, Fx; 35 ± 3 nM and PrP; 134 ± 19 nM vary over a 2-4 fold range. This shows that the variations in the peptides are also meaningful and contribute to specificity of binding.

These preliminary SPR experiments provide interesting insights into the RNA binding behavior of Sho, Fx and PrP and confirmation that further experiments are justified. In particular, efforts should be made to decrease the level of non-specific binding to the flow cell surface, the effect of reversing the analyte and ligand should be examined and the use of a peptide which does not bind RNA should be introduced into the experimental design.

9.3.3 SF and Fluorescence results

The SPR experiments reveal that the complexation of the SF oligonucleotide to the peptides does not fit a simple 1:1 binding model. This result is at odds with that reported by Menon et al. (2001) who found strong 1:1 binding between SF and Fx. Binding with SF was therefore explored further using fluorescence techniques. The SF oligonucleotide was modified by replacing A15 with 2-aminopurine (2-AP), to make the system amenable to fluorescence studies. As can be seen from Figure 9-19(A) the SF oligonucleotide does not fluoresce without the 2-AP modification.

The fluorescence emission of SF_AP at 370 nm was found to change upon quadruplex formation and also upon addition of peptide. G-quadruplex formation is expected to be favoured in the presence of K^+ compared to Li^+ (Hardin et al., 1992). Upon G-quadruplex formation 2-AP will be found in the first loop of the quadruplex (see Figure 9-9(A)) while neighbouring guanines participate in the tetrad stack. As such, upon quadruplex formation 2-AP would be solvent exposed rather than stacking with the guanine bases. This is seen in the emission scan presented in Figure 9-19(B), which reveals a cation dependent difference in the 2-AP signal. The signal is much stronger in a potassium buffer, which favours quadruplex formation, indicating that the 2-AP is more solvent exposed in this medium.

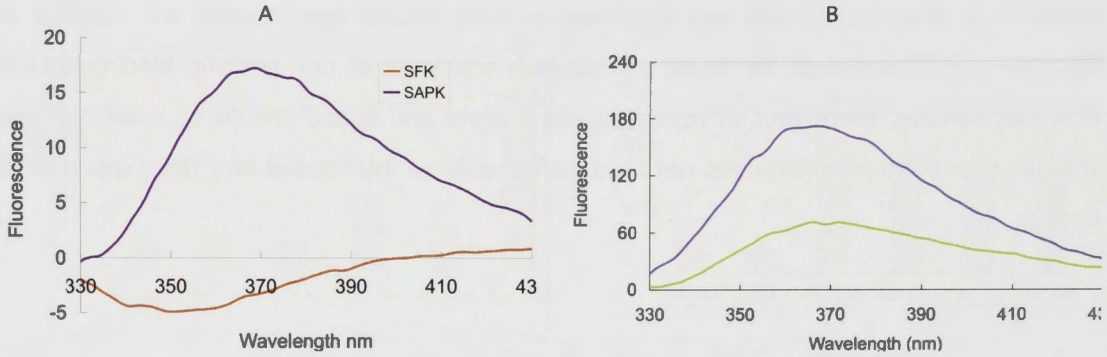


Figure 9-19. Comparison of fluorescence signal of SF RNA in potassium buffer and lithium buffer
(A) The fluorescence of the unmodified SF oligonucleotide (orange) is compared with the 2-AP modified oligonucleotide (purple) both in a buffer of 10 mM Tris HCl, pH 7.4 and 150 mM KCl. Subtraction of the emission scan of the buffer caused the apparent negative values. (B) The fluorescence emission (330-430 nm) is compared for the same concentration of RNA in a buffer of 10 mM Tris HCl pH 7.4 and 150 mM KCl (grey) and in a buffer of 10 mM Tris HCl, pH 7.4 and 150 mM LiCl (olive green).

9.3.3.1 Peptide binding analysis

Titration of SF_AP with the Sho, Fx and PrP peptides, over a wide concentration range, produced different types of binding isotherm. A biphasic isotherm was generally seen in the Fx titrations. This involved an initial increase in emission at 370 nm followed by a larger decrease in emission ('Type 1'). By contrast, titration with PrP produced an immediate decrease in emission at 370 nm with this trend continuing until saturation ('Type 2'). The results of the Sho titrations were varied. Some experiments produced a Type 1 binding isotherm, similar to Fx, while others produced Type 2 binding behaviour. In a small number of experiments there was only a minor initial decrease with the binding isoform forming a plateau followed by a decrease in emission at 370 nm ('Type 3').

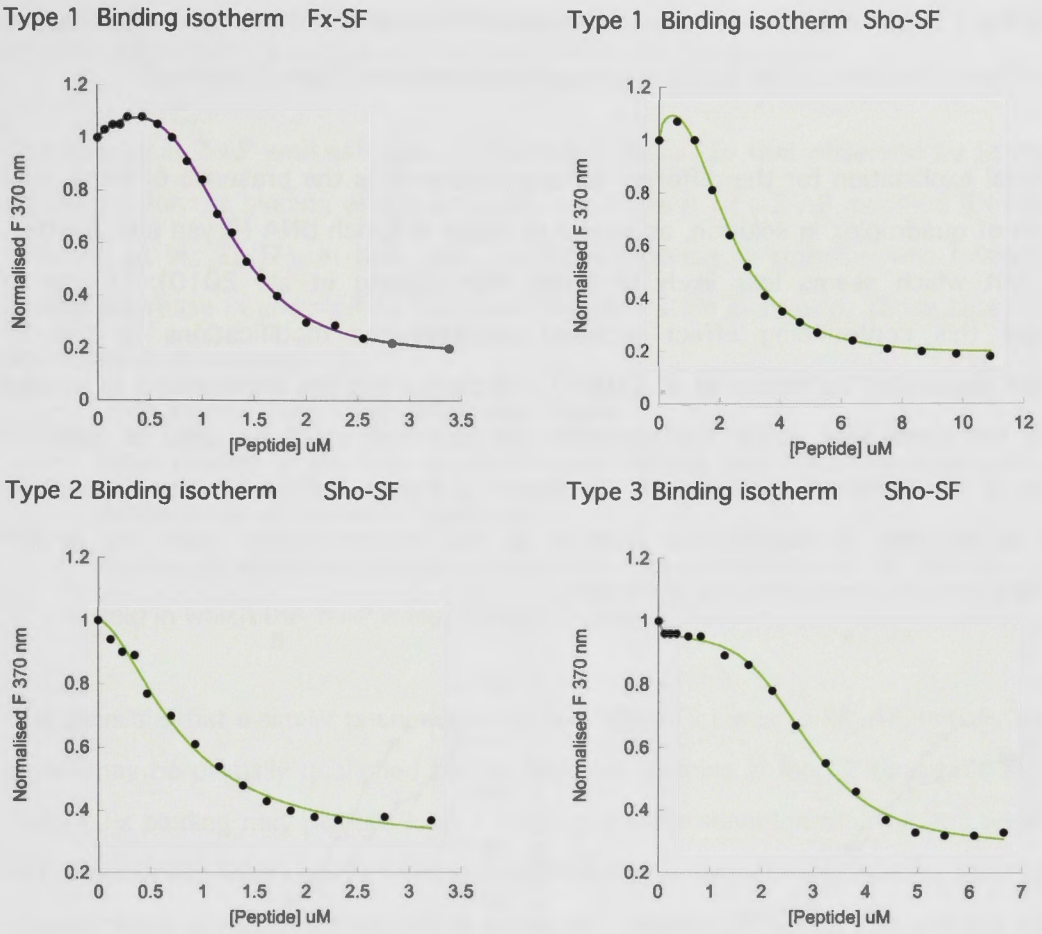


Figure 9-20. Three types of binding isotherm seen in fluorescence studies involving Sho, Fx and PrP peptide titration againsts SF_AP
(A-B) Type 1 binding isotherm demonstrates biphasic behaviour, with an initial small increase in the fluorescence signal at 370 nm, upon addition of Fx (A) or Sho (B), followed by a decrease in signal as the peptide concentration is increased. (C) Type 2 binding isotherm in which the fluorescence signal at 370 nm decreases upon addition of Sho. (D) Type 3 binding isotherm in which the fluorescence signal at 370 nm, initially plateaus with addition of Sho but then decreases as the concentration of Sho increases.

To control for variability in experimental conditions and obtain a more reliable comparison of the peptides, several experiments were done in tandem (as described in the Methods section). A total of 38 titration experiments were performed, with a constant RNA concentration around 270 nM. This subset of experiments is now used to describe binding behaviour and estimate binding affinity of the peptides. Of 5 titration experiments involving Fx, 4 (80%) produced Type 1 binding isotherms. The same behaviour was seen in 12% of the Sho titrations but none of the PrP titrations. The majority of Sho titrations (82%) resulted in a Type 2 binding isotherm. In one case the binding isotherm had a sigmoidal shape as the initial decrease remained small for some time and was then

followed by a larger decrease to saturation. In all PrP titrations, there was a decrease in emission from the start of the titration through to saturation (Type 2 binding).

A potential explanation for the different binding isotherms is the presence of more than one type of quadruplex in solution, as known to occur in G-rich DNA (Bryan and Jarstfer, 2007) but which seems less likely in G-rich RNA (Zhang et al., 2010). Efforts to minimised this confounding effect included adopting the modifications to the SF transcript suggested by Menon et al. (2007) and conducting the experiments in parallel so that the same RNA stock and operating environment could be used at least to compare 2 simultaneous experiments. As shown in Figure 9-21(A-B), the Fx peptide caused an increase in fluorescence emission at low concentrations while Sho or PrP interacting with the same RNA stock did not.

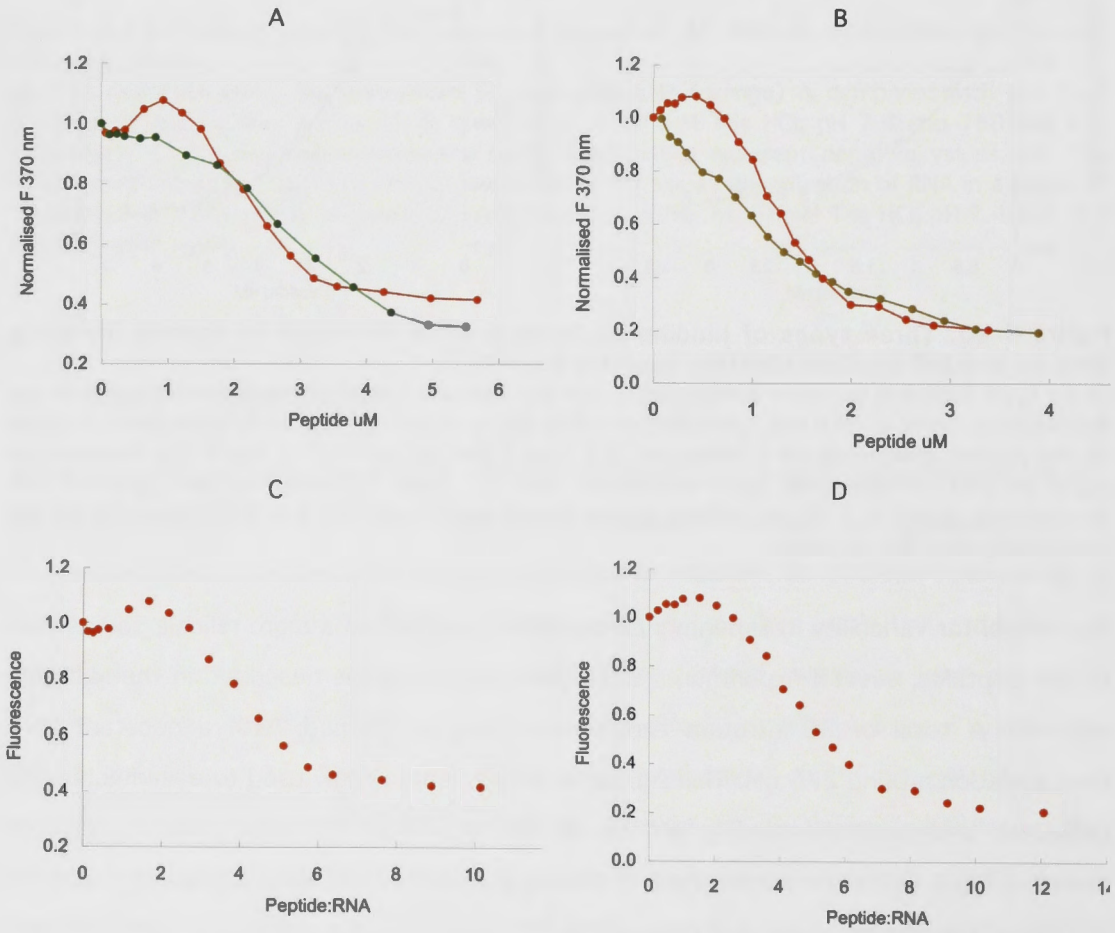


Figure 9-21. Comparison of binding for SF_AP and Fx, Sho and PrP
(A) The fluorescence emission at 370 nm is compared for Fx (red) and Sho (green) with both experiments being performed in tandem. (B) The fluorescence emission at 370 nm is compared for Fx (red) and PrP (brown) with both experiments being performed in tandem. (C) The Fx binding

isotherm shown in (A) re-plotted with the peptide:RNA ratio on the x-axis. (D) The Fx binding isotherm shown in (B) re-plotted with the peptide:RNA ratio on the x-axis.

The change in 2-AP emission upon Fx binding is similar to that observed by Stampfl and colleagues for the binding of the antibiotic neomycin B to a 2-AP modified RNA aptamer (Stampfl et al., 2007). In that case, an initial increase in emission was followed by a smaller decrease in emission as the ligand concentration increased. The authors propose an explanation involving:

- 2-AP initially stacks with other loop bases.
- Upon binding of the first neomycin molecule this loop structure opens and the 2-AP becomes more solvent exposed.
- Binding of additional neomycin molecules causes the aptamer to assume a tighter fold in which the 2-AP is less solvent exposed.

It is plausible that a similar phenomenon occurs when Fx binds to SF_AP. Initially the 2-AP signal may be partially quenched by the adjacent guanine in loop 2 (Bandwar and Patel, 2001). Fx binding may perturb loop 2 causing greater solvation of 2-AP and an increase in fluorescence. Additional Fx binding could induce conformational changes allowing 2-AP to participate in stacking interactions either with neighbouring bases or possibly with the aromatic ring of Phe. As the 2-AP signal increases upon G-quadruplex formation (Figure 9-18(B)) the behaviour seen here could also be interpreted as quadruplex stabilization at low peptide concentration followed by destabilization as the peptide concentration is increased. Changes in the quadruplex structure upon peptide binding are investigated using CD and are discussed in the next section.

Although the Type 1 binding isotherm is sometimes seen in Sho-SF binding it is more usual to see Type 2 binding for Sho and PrP. Quenching of the fluorescence signal implies that 2-AP stacks with a neighbouring base or an aromatic peptide residue. Notably, PrP has more aromatic amino acids than the other peptides and the stacking of these aromatic residues with 2-AP might explain the immediate decrease in the 2-AP signal seen in all PrP titrations. Sho binding on the other hand, must promote stacking of 2-AP with neighbouring bases, as it lacks aromatic residues.

SF-AP binding is complex as indicated by Scatchard plots of the binding isotherms presented in Figure 9-22.

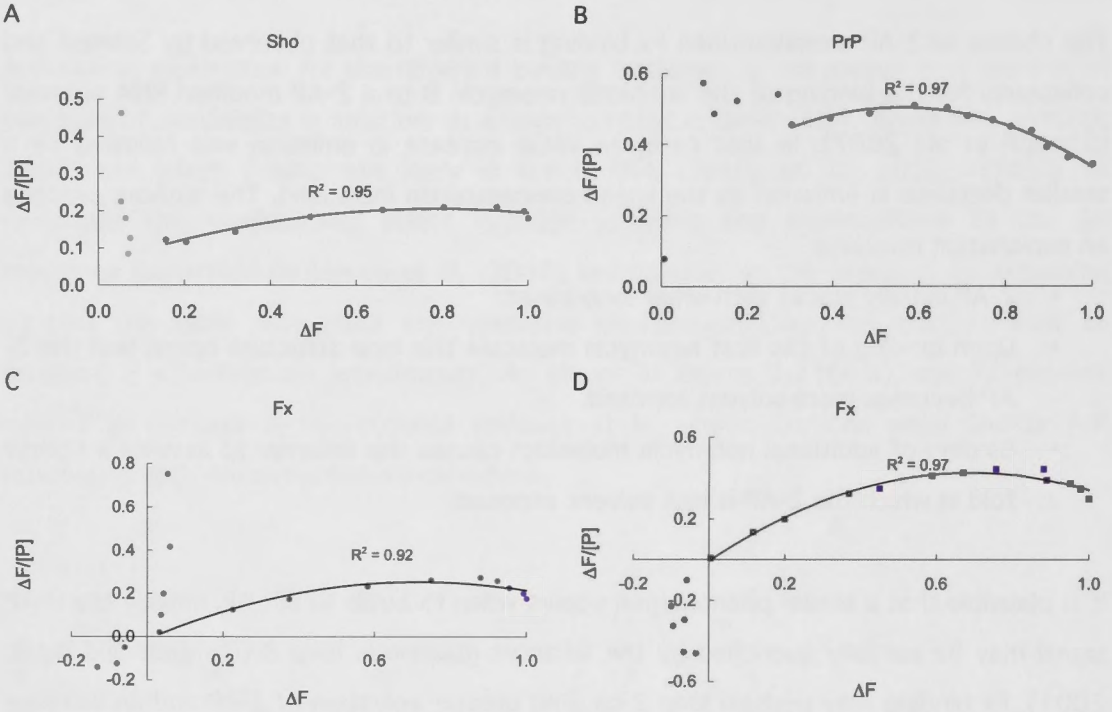
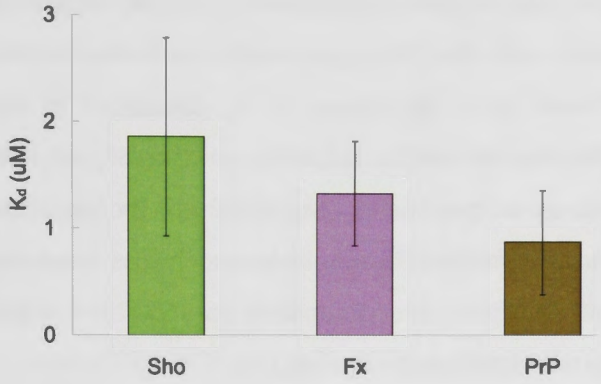


Figure 9-22. Scatchard plots for the binding of SF_AP and the Sho, Fx and PrP peptides

The binding isotherms presented in Figure 9-20 are re-presented here as Scatchard plots. ΔF is the change in fluorescence upon peptide titration and [P] is the concentration of peptide. The least squares correlation co-efficient (R^2) for a polynomial fit of data for the higher peptide concentration range is shown.

Starting from a peptide:RNA ratio of ~ 2.5:1 the Scatchard plots show a 'concave down' pattern which may be indicative of positive co-operativity between binding sites or self-association of the peptides (Van Der Merwe, 2001).

The modified Hill model (Eqn 13) was used to estimate binding affinity. Generally there were too few points in the initial upward phase of the isotherm to obtain a reliable estimate of the K_d for this process. Accordingly, it is only the K_d for the decrease in fluorescence that has been determined. The average K_d of the 3 peptides, for the process involving a decrease in fluorescence is plotted in Figure 9-23.



	Sho	Fx	PrP
K_d avg	1.9	1.3	0.9
SD	0.9	0.5	0.5
n avg	2.0	3.4	1.4
SD	0.9	0.3	0.4

Figure 9-23. Affinity of the Sho, Fx and PrP peptides for SF RNA
The average K_d (μM) based on at least 4 experiments in which the RNA concentration is ~270 nM is plotted together with error bars representing the standard deviation of the sample. The table on the right sets out the average K_d (μM), the average Hill co-efficient (n) for the decrease in fluorescence at 370 nm and the standard deviation in these values.

The data presented in Figure 9-23 reveals that the peptides bind with positive co-operativity and moderate affinity to SF, with a K_d around 1 μM. RNA regulatory proteins commonly bind to their RNA targets with greater affinity (10s – 100s nM) (Carmel et al., 2010, Dujardin et al., 2010, Ishimaru et al., 2010, Ruan et al., 2010). Although the affinity values estimated by fluorescence spectroscopy are different to those obtained by SPR, it is interesting to see that all 3 peptides demonstrate a similar degree of affinity for SF, as was also seen in the SPR experiments.

The binding of Fx to SF_AP observed here differs from that reported by Mihailescu and colleagues (Menon and Mihailescu, 2007). Instead of biphasic binding isotherms they observed a strong immediate decrease in the signal at 370 nm, and a 1:1 equilibrium binding model (Eqn 13) was used to estimate a K_d ~ 0.7 nM, indicating very tight binding. Earlier studies found the K_d for this binding event to be 100 times greater at 75 nM (Darnell et al., 2001). The authors suggest this discrepancy might be explained by the presence of more than one RNA conformer in the earlier experiments (Menon and Mihailescu, 2007). Another possibility is that the later experiments may not have been conducted under equilibrium conditions. The binding isotherm presented in the later study is not hyperbolic. Instead the peptide appears to bind immediately to the RNA, leaving no free peptide in solution. Accordingly, an equilibrium binding model may not be appropriate.

Even if the true dissociation constant for this interaction is closer to the K_d of 75 nM derived by Darnell et al. (2001), consistent with the SPR experiments performed in this study (~ 80 nM), it is still quite different from the values of K_d calculated in the fluorescence experiments. This could be due to some instability in the SF_AP, the presence of different quadruplex structures or peptide aggregation. If the peptides aggregated the actual concentration available for binding could be much less than the apparent concentrations. By contrast, protein aggregation is unlikely to occur in the SPR experiments as the peptides are bound to the flow cells as ligands.

The fluorescence experiments are based on the assumption that a change in 2-AP environment results from a structural change induced by peptide binding. However, perturbation in 2-AP environment could arise in a number of different ways. As such, this technique provides a less direct indication of binding than SPR, which responds to a change in mass on the cell surface, caused by the association of ligand and analyte.

9.3.3.2 Mutagenesis studies

Three mutants of the Sho peptide were designed to better understand the mechanism by which Sho binds SF. In the first mutant the positively charged Arg residues were replaced with negatively charged glutamic acid ('Neg'). In the second mutant the Arg residues were replaced by positively charged Lys residues ('Lys'). All the residues in Sho were retained in the third mutant, however their order was scrambled ('Scr'). Fluorescence experiments comparing SF_AP binding to a mutant with binding to Sho were performed in tandem.

A titration of SF with Neg produces results comparable with a water titration, thus indicating that the positively charged Arg residues of Sho are necessary for binding (Figure 9-24 (A)). Binding occurs with the Lys mutant, but with lower affinity than seen with Sho. Lys and Arg have the same charge, and while the ε-amino group of Lys is capable of forming hydrogen bonds, the Arg residue is capable of forming multiple hydrogen bonds. Comparison of the Lys mutant with Sho leads to the conclusion that the positive charge of the Arg residue is required for binding but there is an additional element that complements electrostatic interactions, and which increases binding

strength. This is the superior ability of Arg to form hydrogen bonds with a binding partner.

This is confirmed by considering the binding isotherm of the scrambled peptide. Given the small size of the Sho peptide and the abundance of Gly and Arg residues it is difficult to create a truly scrambled peptide. Although the 'RGG' sequences are disturbed in Scr, there are still 5 Arg residues interspaced with Gly, which gives the peptide a good deal of flexibility. Figure 9-24(C) shows that Scr produces a binding isotherm very similar to Sho.

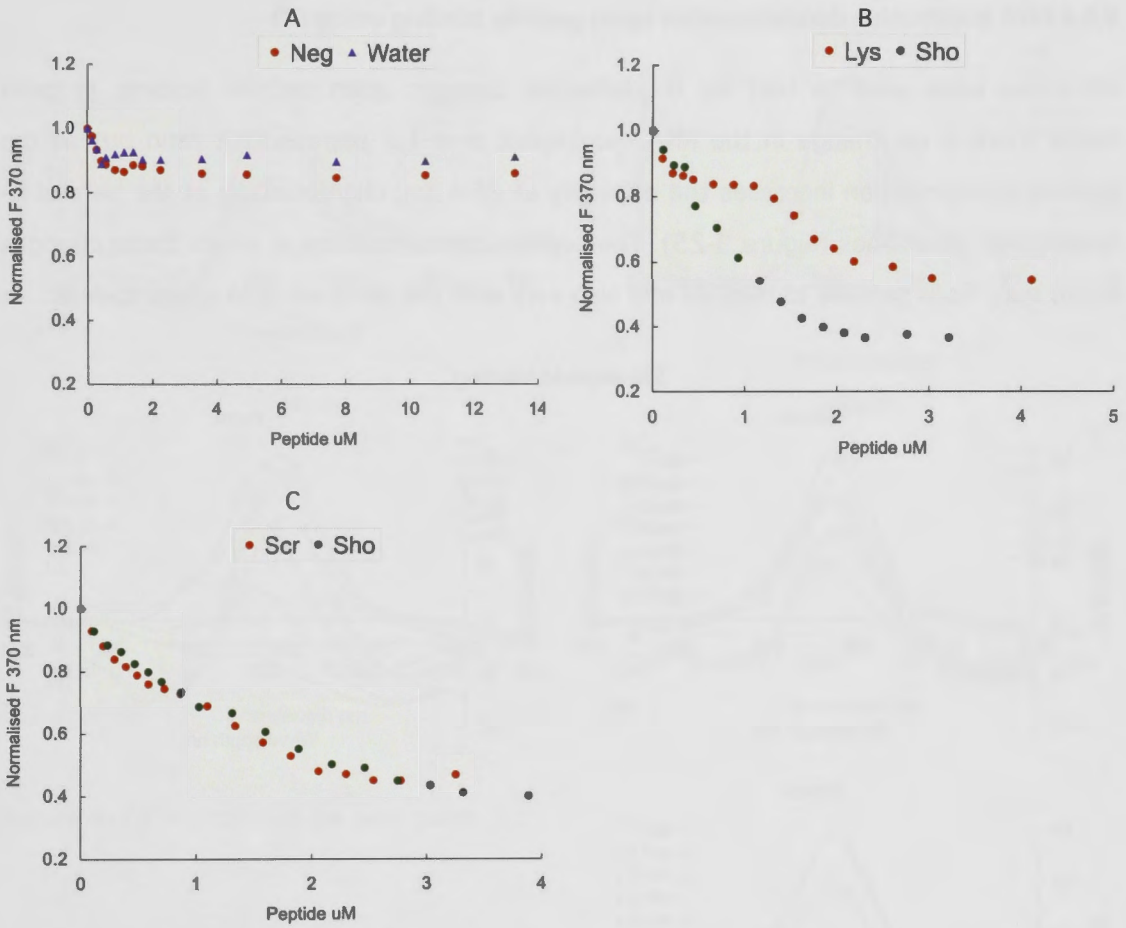


Figure 9-24. Binding isotherm seen in fluorescence studies of SF_AP binding to the Sho mutant peptides
The relative decrease in fluorescence signal of SF_AP (270 nM) at 370 nm (A) upon titration with the Neg mutant in an aqueous solution, compared to titration with the same volume of water. (B) Upon titration with the Lys mutant compared to titration with Sho. (C) Upon titration with the Scr mutant compared to titration with Sho

The dissociation constant for Scr and Lys binding to SF_AP was calculated on the basis of at least 3 experiments and produced a result of 5.8 ± 1.8 μM for Lys and 1.63 ±0.58 μM

for Scr. The K_d of Sho (1.86 ± 0.93 μM) and Scr are thus in the same range. The fact that the K_d for the Lys mutant is some 3 times greater than for Sho and Scr demonstrates that the Arg residues impart binding affinity not only through electrostatic effects, but also through the superior ability of Arg to hydrogen bond with RNA.

In Chapters 7 and 8, the quadruplex was seen to be a stable structure that can withstand peptide-loop interactions occurring at a 1:1 ratio of peptide to RNA. The stability of the quadruplex structure is explored further using CD.

9.3.4 RNA quadruplex de/stabilization upon peptide binding using CD

CD scans were used to test for G-quadruplex changes upon peptide binding. In most cases there is no change in the RNA quadruplex at a 1:1 peptide:RNA ratio but as the peptide concentration increases the ellipticity at 264 nm, characteristic of the parallel G-quadruplex, diminishes (Figure 9-25). The peptide concentrations at which these changes occur vary from peptide to peptide and also vary with the different RNA quadruplexes.

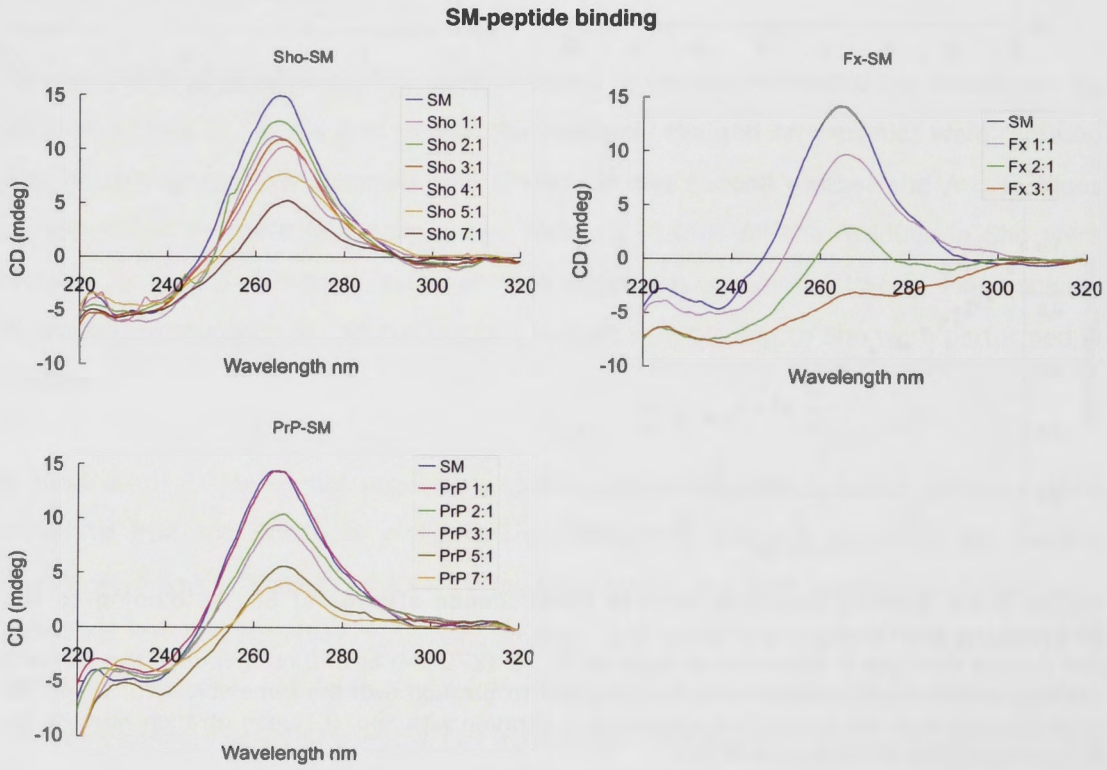


Figure 9-25 Continued on next page

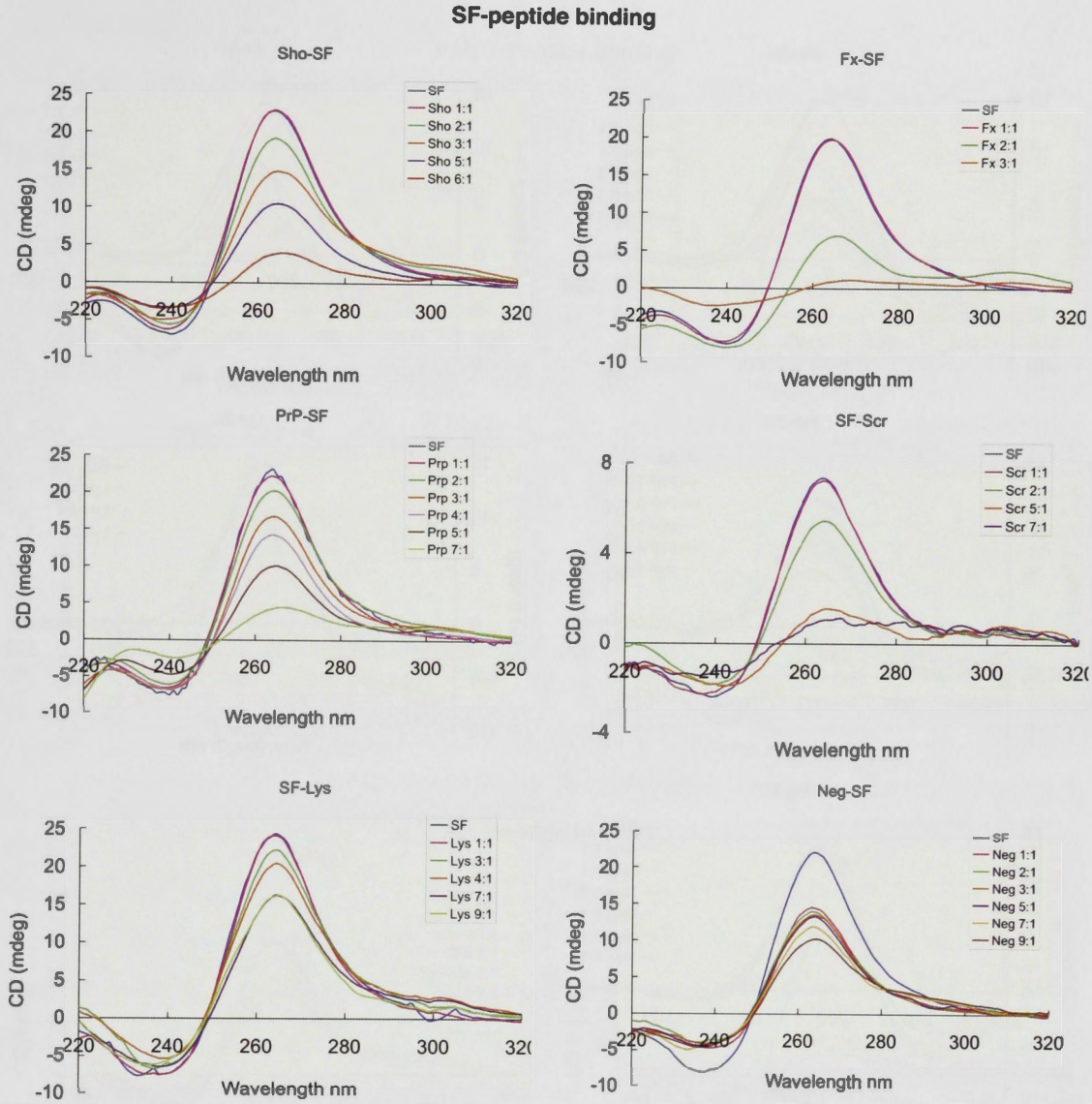


Figure 9-25 Continued on next page

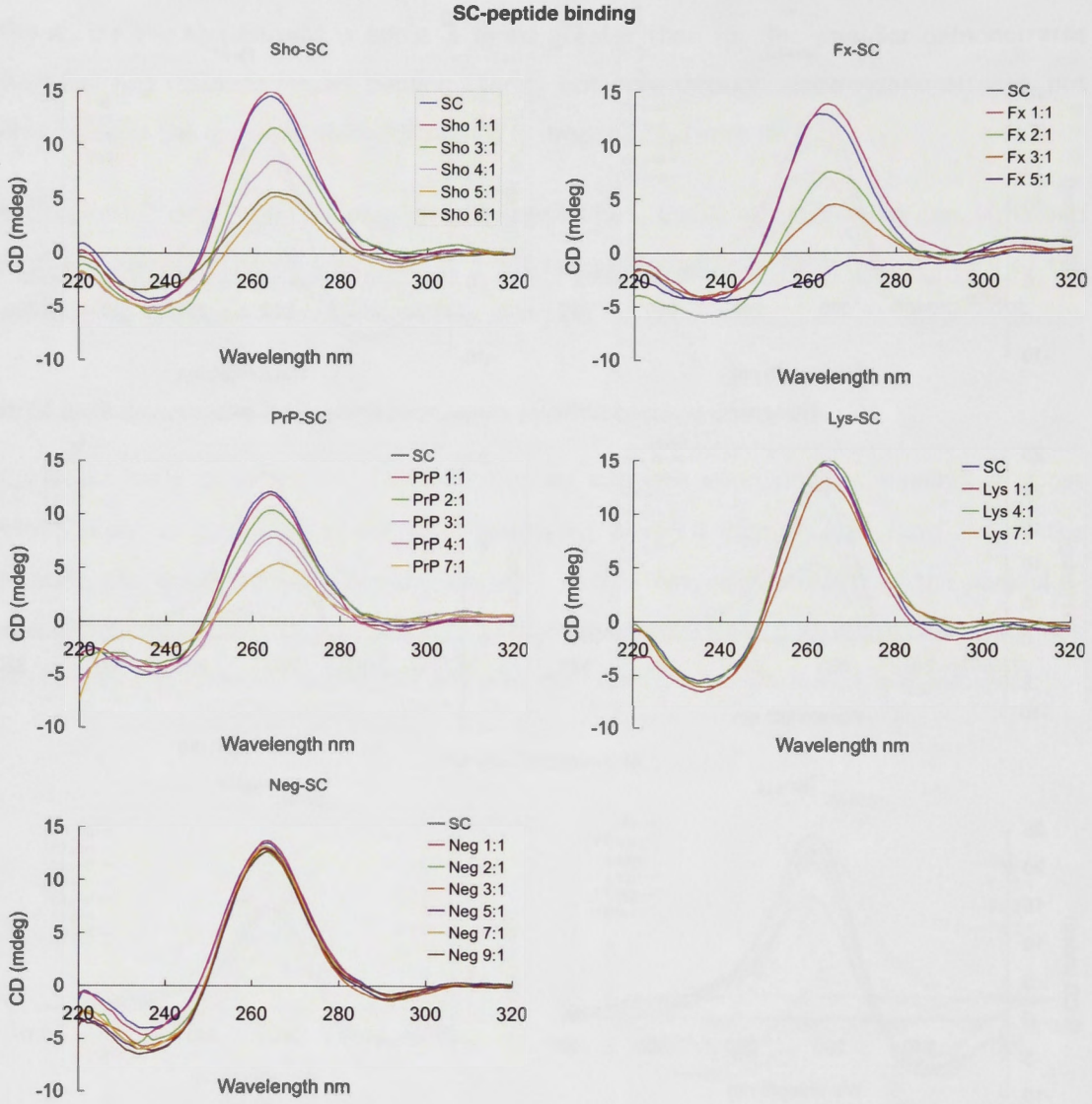


Figure 9-25 Continued on next page

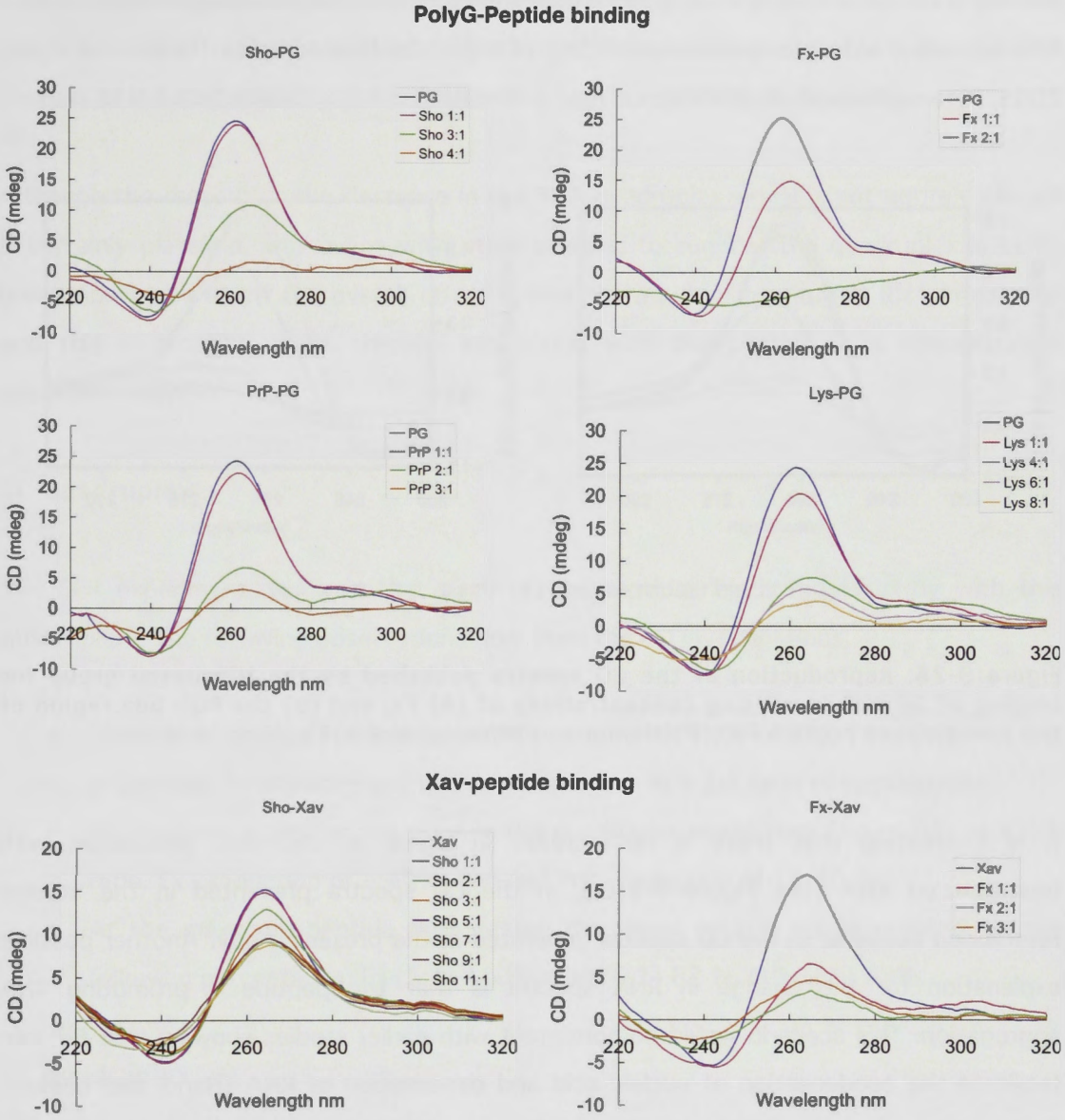


Figure 9-25. CD spectra for binding of the RNA G-quadruplex and Sho, Fx and PrP
CD spectra of the RNA oligonucleotides to which the peptide is added in increasing molar ratios as indicated by the legend in each plot.

The Mihailescu group has published CD spectra for the binding of Fx to SF RNA (Menon and Mihailescu, 2007) and MAP1B RNA (Menon et al., 2008). The CD spectra (reproduced in Figure 9-26) are similar to the spectra obtained in the present study (Figure 9-25) although the peptide:RNA ratios corresponding to spectral changes may differ. The authors interpret their spectra as showing a stabilization of the G-quadruplex RNA at a

1:1 peptide:RNA ratio followed by an unwinding of the quadruplex once the peptide reaches a certain concentration. Other studies have reported a decrease in ellipticity at ~ 265 nm observed upon protein unwinding of a parallel G-quadruplex (Enokizono et al., 2005, Paramasivam et al., 2009a).

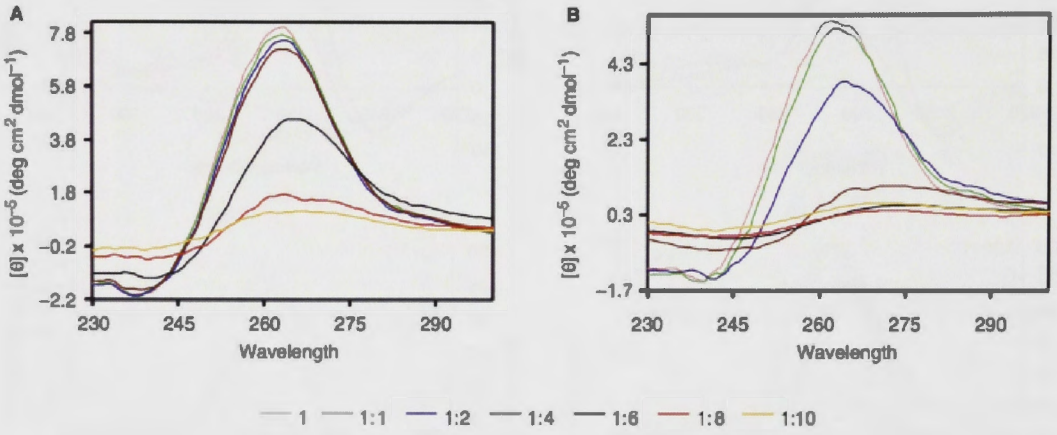


Figure 9-26. Reproduction of the CD spectra published by the Mihailescu group for binding of SF with increasing concentrations of (A) Fx, and (B) the RGG box region of the homologous protein FXR1P (Menon and Mihailescu, 2007)

It is interesting that there is no increase in signal at 280 nm, associated with unstructured RNA (see Figure 9-9(E)), in the CD spectra presented in the studies referenced above or in the CD spectra generated in the present study. Another possible explanation for the change in RNA spectra is that the peptide is promoting RNA aggregation. This scenario would be consistent with earlier studies showing that PrP can facilitate the condensation of nucleic acid and dimerisation of RNA (Nandi and Leclerc, 1999, Gabus et al., 2001, Nandi and Sizaret, 2001). However, in some of the abovementioned studies, additional experimental techniques have been used to establish quadruplex unwinding. For example, Paramasivam and colleagues used FRET-melting experiments to confirm CD results showing that hnRNP A1 protein and its derivative Up1 do unwind the quadruplex structure within the KRAS promoter (Paramasivam et al., 2009a).

If the effect seen in the CD experiments was simply aggregation, facilitated through peptide screening of the negatively charged RNAs, then the Lys mutuant should have the

same effect on the RNA as the Sho peptide, but this was not observed. Further, as addition of the Lys and Neg mutants to the SC RNA has very little effect on the RNA spectra it seems unlikely that the diminution of the quadruplex signal is a spontaneous process or induced solely by the presence of a high concentration of peptide.

Although the reason for the decrease in the RNA quadruplex signal is not entirely clear it is certainly plausible, and in line with other studies, to suggest the quadruplex is being unwound. In any event the overall effect is that there is less quadruplex RNA in solution and this is brought about through interaction with the peptide in a concentration dependent manner.

SC quadruplex

The SPR experiments showed that each of the peptides binds most tightly with this quadruplex. The following observations flow from the CD investigations:

- There is no change in ellipticity at 264 nm at a 1:1 ratio of Sho, Fx or PrP.
- A decrease in ellipticity at 264 nm commences at a 2:1 ratio of peptide:RNA.
- Fx produces a greater reduction in the quadruplex signal than Sho or PrP; at a 2:1 ratio. Fx - reduction of ~ 40%, Sho and PrP - reduction of 15-20 % .
- At the maximum peptide ratio tested the signal at 264 nm is reduced by the following percentages; Fx (5:1) 100%, Sho (6:1) 62 %, PrP (7:1) 57%.

SF quadruplex

- There is no change in ellipticity at 264 nm at a 1:1 ratio of Sho, Fx or PrP.
- The decrease in quadruplex signal commences at a 2:1 ratio of peptide:RNA.
- Fx produces a greater reduction in the quadruplex signal at a 2:1 ratio than was seen for SC. At a 2:1 ratio Fx produces a reduction of ~ 70%, Sho and PrP both produce a reduction of 15-20 %.
- At the maximum peptide ratio tested the signal at 264 nm is reduced by the following percentages; Fx (3:1) 100%, Sho (6:1) 83 %, PrP (7:1) 81%.

SM quadruplex

- Sho and Fx do not stabilize SM at a 1:1 peptide:RNA ratio instead there is a 20-30% decrease in signal at 264 nm upon complexation with Sho or Fx.
- PrP starts to cause a decrease in quadruplex signal at a 2:1 peptide:RNA ratio.
- At the maximum peptide ratio tested the signal at 264 nm is reduced by the following percentages; Fx (3:1) 100%, Sho (7:1) 66 %, PrP (7:1) 74%.

SM was originally designed as an SF mutant with the intenion of disrupting the quadruplex structure. It was therefore surprising to find that SM did in fact form a G-quadruplex. However, instability in this structure could be expected and is consistent with these results showing that SM is unwound at lower peptide:RNA ratios than seen for SF.

PolyG quadruplex

- Fx induces a large decrease in signal at a 1:1 ratio (39% reduction).
- At the maximum peptide ratio tested the signal at 264 nm is reduced by the following percentages; Fx (2:1) 100%, Sho (4:1) 94 %, PrP (3:1) 87%.

Xav quadruplex

A shortage of the Xav RNA meant that only 2 peptides could be tested. Very different results emerged for Sho and Fx.

- Sho stabilized Xav at a 1:1 ratio.
- Fx produced a large decrease in the RNA signal at a 1:1 ratio (62% reduction).
- Sho did not promote reduction of the RNA signal to the same extent as seen in the other RNAs. At an 11:1 ratio the signal at 264 nm had only decreased by 40%.
- At the maximum peptide ratio tested the signal at 264 nm is reduced by the following percentages; Fx (3:1) 73%, Sho (11:1) 40 %.

The effect of the Sho mutants, ('Lys', and 'Scr') on the SF quadruplex was also probed using CD. Analogous to the fluorescence mutagenesis the Scr mutant had a similar effect to Sho and the Lys mutant had a lesser effect. The CD experiments together with the

fluorescence experiments reveal that the Sho and Scr peptides interact with SF in a similar manner and that this differs from the complexation that occurs if the Arg residues of Sho are replaced by Lys.

The binding of the Lys peptide to the SC and PG quadruplexes was also investigated. As can be seen from the spectra presented in Figure 9-23, there was no change in the SC quadruplex signal irrespective of the concentration of the Lys peptide. However, the PG quadruplex did respond to this peptide, although to a lesser extent than for Sho.

Finally, the negatively charged mutant was tested with the SC and SF quadruplexes. No change was seen upon addition of this peptide to the SC quadruplex. However, in the case of the SF quadruplex the quadruplex signal initially decreased (reduced by 34%) but then remained reasonable stable, only a small decrease being seen as the concentration of the Neg peptide was increased. At a 9:1 ratio the signal had decreased by 53%.

Taken together these results show a good deal of variation in the effect which each of the peptides has on the different RNA quadruplex structures. In some cases the qudruplex structure is more resilient to change than in others. For the purpose of this discussion quadruplex stability is judged by the extent to which the CD spectra characteristic of the parallel G-quadruplex is preserved following complexation with the peptide. It is interesting to see that binding affinity (derived from the SPR experiments) is not positively correlated with the ability of the peptide to unwind the complex. For example, Fx at a 2:1 ratio causes the greatest diminution in ellipticity at 264 nm in the following order (PG>Xav~SM>SF>SC). That is, the least effect is seen in SC, however, the SPR experiments showed that Fx binding affinity follows the order (SM>SC>Xav>PG>SF). These results suggest a weak negative correlation - high binding affinity being associated with the more stable quadruplex structures.

This would be consistent with the observation in Chapter 7 that the stability of the RNA is a factor influencing the free energy (i.e. binding affinity) of the complex. However, RNA stability is only one factor affecting binding affinity, and ultimately may be outweighed by the number and type of interactions formed between the RNA and the peptide as discussed in Chapter 7.

Based on the assumption that preservation of the RNA spectra is associated with RNA stability, 3 effects can be seen in the CD spectra;

- The RNA quadruplexes have a different level of inherent stability.
- Some peptides are better able than others to destabilize the quadruplex structure.
- This affect is mediated through changes in peptide concentration.

These results are generally consistent with those found by the Mihailescu group in their study of Fx binding to SF RNA (Menon and Mihailescu, 2007) and MAP1B RNA (Menon et al., 2008). In both cases they found that peptide binding at a 1:1 ratio had a stabilizing effect on the quadruplex structure, as confirmed by UV melting experiments, whereas at higher peptide:RNA ratios the RNA quadruplex was unwound. The fact that the Sho RGG box and the PrP RNA binding domain also stabilize SF and SC at a 1:1 ratio yet unwinding is seen at higher peptide ratios indicates that this could be a common device by which quadruplex structure is affected and biological function may be influenced. It is plausible that increasing peptide concentration allows binding at an additional site on the RNA which may then trigger conformational change which in turn faciliates the binding of additional peptide molecules as the RNA unwinds. Clearly, the mechanism by which this might occur requires further examination.

9.4 FUTURE WORK

It is well known that G-rich DNA may form diverse quadruplex structures and that the formation of these different structures is influenced by the surrounding environment. Much less is known about G-rich RNA. However, the latest research indicates that G-rich RNA forms a stable parallel G-quadruplex rather than a number of different types of quadruplex, that is, RNA quadruplexes tend to be 'monomorphic' in contrast to polymorphic DNA (Joachimi et al., 2009, Zhang et al., 2010). This suggests a lower risk that the RNA sequences tested in this thesis are forming more than one type of quadruplex. Nevertheless, experiments should be performed to confirm this. As a first step, gel electrophoresis could be used to see whether the RNA nucleotides produce different bands on a nondenaturing gel. While the presence of multiple bands would indicate different structures, an absence would not completely rule out the formation of different structues as different RNA quadruplexes could possibly migrate at the same

rate. UV thermal difference spectroscopy (TDS) (Mergny et al., 2005) could also be used to complement the CD studies and confirm quadruplex structure.

More extensive testing is especially important for the SF and PG oligonucleotides given that SF and PG binding assays did not produce simple 1:1 binding behaviour. It is possible that these sequences are involved in complex multi-site binding or alternatively the results might be explained by the presence of different RNA structures. It would also be interesting to assess SM with different techniques given that the sequence was designed to avoid quadruplex formation yet CD spectra is consistent with parallel quadruplex structure.

As noted earlier, the SPR experiments should be treated as preliminary results. More work is required to optimise this technique for the species being tested and these experiments should be repeated, including under conditions in which the analytes and ligands are swapped.

The CD spectra of the RNA-peptide complexes show a decrease in quadruplex signal which is consistent with quadruplex unwinding. Further investigations to confirm quadruplex unwinding should incorporate complementary techniques such as UV melting experiments (Paramasivam et al., 2009b) and NMR studies (Enokizono et al., 2005). These preliminary steps should be followed by studies to elucidate the mechanism by which an increase in peptide concentration leads to destabilization of the quadruplex structure.

The RNA sequences tested are limited in number and form. Future experiments will no doubt test a comprehensive range of RNA sequences for binding affinity with Sho.

9.5 CONCLUSIONS

The preliminary experimental studies reported in this thesis indicate that the RGG box domain of Sho is competent to bind G-quadruplex RNA. Of the small number of RNAs tested, Sho bound most strongly to SC, the Sc1 aptamer, previously found to bind strongly to Fx (Darnell et al., 2001, Zanotti et al., 2006). In fact all 3 peptides (Sho, Fx

and PrP) bound most tightly and with similar predicted affinity to SC (K_d~ 20 nM). This level of binding affinity is consistent with other studies of PrP binding e.g. Mercey and colleagues identified RNA aptamers with high affinity for full length PrP and concluded that the best RNA ligand bound with a K_d of 20 nM (Mercey et al., 2006).

Although SC, Xav and SM are all synthetic RNA quadruplexes, the binding affinity of the peptides for these nucleotides: Sho (17 - 56 nM equating to -ΔG 10.43 - 9.73 kcal mol^{-1}), Fx (21 - 35 nM equating to -ΔG 10.3 - 10.01 kcal mol^{-1}) and PrP (21 - 134 nM equating to -ΔG 10.3 - 9.22 kcal mol^{-1}) was generally comparable with those for RNA regulatory protein domains which have K_ds in the 10 -100s nM range (Carmel et al., 2010, Dujardin et al., 2010, Ishimaru et al., 2010, Ruan et al., 2010). PolyG bound prolifically to each of the 3 peptides.

Binding to the biologically relevant target, the SF G-quadruplex, was complex and could not be described by a simple 1:1 model in the SPR or fluorescence experiments. In the SPR experiments the RNA was titrated into a fixed concentration of peptide, whereas in the fluorescence experiments the peptides were titrated into a fixed concentration of RNA. The binding isotherms derived using these different techniques indicate positive co-operativity. The fluorescence studies also suggested distinctions in the mode of binding for the 3 peptides. Binding of Fx to SF adopted a distintly biphasic isotherm with an initial increase in fluorescence followed by a more significant decrease. The isotherms for Sho and PrP binding were more similar, although the PrP isotherms were generally steeper suggesting higher affinity for SF. The complexity of SF binding has made it difficult to obtain a reliable estimate of K_d for these interactions. Although the estimates of K_d for SF binding should be treated cautiously, it was interesting to see that the estimates for all 3 peptides were in the same range, and that the SPR results are similar to those found by Darnell et al. (2001) but differ from those reported by Menon et al. (2007).

Finding complex binding in a regulatory region of mRNA is not particularly surprising. Although early theories of protein binding arose out of the highly specific, strong affinity enzyme-substrate model, the study of protein-nucleic acid binding has opened the door to a much more complicated and fuzzy world. In this world it is common to find the formation of nucleic acid - protein complexes involving several molecules, such as

ribonucleo-protein complexes (RNPs) that are found in all stages of RNA biogenesis. Recent discoveries indicate that the story may be yet more complicated with the involvement of micro RNAs that bind to mRNA control elements within a protein-RNA complex (Rodriguez et al., 2008).

The affinity of each of the peptides for the 5 RNA oligonucleotides differed by over an order of magnitude. This demonstrates that Sho as well as each of the other peptides is capable of distinguishing between different RNA targets, which is clearly important in a biological setting. Differences were also seen in the affect which each of the peptides had on the G-quadruplex structures.

As discussed in Chapter 5, the G-quadruplex structure is found in telomeres, promoter regions of the genome and in mRNA. This structural form might be considered a roadblock to particular biological functions, such as the transcription of a gene or the translation of a protein. Protein binding can influence these processes by mediating the stability of the quadruplex, i.e. by reinforcing or removing the roadblock. CD experiments reported in this chapter showed the effect which peptide binding has on the RNA quadruplex. It was common to find no interference with the quadruplex structure at a 1:1 peptide to RNA ratio. However, at higher peptide:RNA ratios the quadruplex structure is diminished, either through unwinding or the formation of RNA aggregates. The observation of this effect over the 3 peptides and the 5 RNA sequences suggests this may be a common regulatory device. In summary these studies show that a binding protein may stabilize or destabilize quadruplex nucleic acid depending upon the protein's concentration in the local environment.

Mutagenesis studies revealed that binding is mediated predominantly through the Arg residues of the Sho RGG box. Comparison of the native peptide with a Lys substituted mutant show that the electrostatic interactions fostered by the positive charges on Arg only account for part of the overall affinity. The unique hydrogen bond network formed by the Arg residues is important in complexation and also in promoting destabilization of the quadruplex structure. The similarity of binding in Sho and Scr indicates that closely positioned Arg and Gly residues may constitute a sufficient binding platform without necessarily being ordered into RGG repeats. This is also consistent with the *in silico*

studies reported in Chapter 7, which found that Arg residues which were part of an RGG box repeat were not favoured over other Arg residues at the RNA binding interface.

The work presented in this chapter also highlights similarities between Sho and PrP. Both peptides are capable of binding quadruplex RNA in some cases with very similar affinity as shown for SC binding. This raises the possibility that Sho and PrP could potentially share one or more RNA targets.

Finally, this work provides proof-of-principle that Sho is capable of binding quadruplex RNA. In this study the binding affinity of Sho for the RNAs was generally of a similar order of magnitude as for Fx, which is known to have a functional role in the transport and translational control of mRNAs in the brain (Bagni and Greenough, 2005). Sho is also found primarily in the brain and future studies should investigate the affinity of Sho for neuronal RNA. This future work will provide important insights into the function of Sho and elucidate further similarities or differences with Fx and PrP.

9.5 REFERENCES

Bagni, C. & Greenough, W.T. (2005) From mRNP trafficking to spine dysmorphogenesis: the roots of fragile X syndrome. *Nat Rev Neurosci,* 6, 376-87.

Bandwar, R.P. & Patel, S.S. (2001) Peculiar 2-aminopurine fluorescence monitors the dynamics of open complex formation by bacteriophage T7 RNA polymerase. *J Biol Chem,* 276, 14075-82.

Bryan, T.M. & Jarstfer, M.B. (2007) Interrogation of G-quadruplex-protein interactions. *Methods,* 43, 332-9.

Carmel, A.B., Wu, J., Lehmann-Blount, K.A. & Williamson, J.R. (2010) High-affinity consensus binding of target RNAs by the STAR/GSG proteins GLD-1, STAR-2 and Quaking. *BMC Mol Biol,* 11, 48.

Chellgren, B.W. & Creamer, T.P. (2004) Short sequences of non-proline residues can adopt the polyproline II helical conformation. *Biochemistry,* 43, 5864-9.

Danielsson, J., Jarvet, J., Damberg, P. & Graslund, A. (2005) The Alzheimer beta-peptide shows temperature-dependent transitions between left-handed 3-helix, beta-strand and random coil secondary structures. *Febs J,* 272, 3938-49.

Dapic, V., Abdomerovic, V., Marrington, R., Peberdy, J., Rodger, A., Trent, J.O. & Bates, P.J. (2003) Biophysical and biological properties of quadruplex oligodeoxyribonucleotides. *Nucleic Acids Res,* 31, 2097-107.

Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T. & Darnell, R.B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell,* 107, 489-99.

Darnell, J.C., Mostovetsky, O. & Darnell, R.B. (2005) FMRP RNA targets: identification and validation. *Genes Brain Behav,* 4, 341-9.

Darnell, J.C., Warren, S.T. & Darnell, R.B. (2004) The fragile X mental retardation protein, FMRP, recognizes G-quartets. *Ment Retard Dev Disabil Res Rev,* 10, 49-52.

Dujardin, G., Buratti, E., Charlet-Berguerand, N., Martins De Araujo, M., Mbopda, A., Le Jossic-Corcos, C., Pagani, F., Ferec, C. & Corcos, L. (2010) CELF proteins regulate CFTR pre-mRNA splicing: essential role of the divergent domain of ETR-3. *Nucleic Acids Res.*

Enokizono, Y., Konishi, Y., Nagata, K., Ouhashi, K., Uesugi, S., Ishikawa, F. & Katahira, M. (2005) Structure of hnRNP D complexed with single-stranded telomere DNA and unfolding of the quadruplex by heterogeneous nuclear ribonucleoprotein D. *J Biol Chem,* 280, 18862-70.

Epstein, I.R. (1978) Cooperative and non-cooperative binding of large ligands to a finite one-dimensional lattice. A model for ligand-oligonucleotide interactions. *Biophys Chem,* 8, 327-39.

Gabus, C., Auxilien, S., Pechoux, C., Dormont, D., Swietnicki, W., Morillas, M., Surewicz, W., Nandi, P. & Darlix, J.L. (2001) The prion protein has DNA strand transfer properties similar to retroviral nucleocapsid protein. *J Mol Biol,* 307, 1011-21.

Greenfield, N. (2006) Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols,* 1, 2876-2890.

Greenfield, N. & Fasman, G.D. (1969) Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry,* 8, 4108-16.

Grossman, A., Zeiler, B. & Sapirstein, V. (2003) Prion protein interactions with nucleic acid: possible models for prion disease and prion function. *Neurochem Res,* 28, 955-63.

Hardin, C.C., Watson, T., Corregan, M. & Bailey, C. (1992) Cation-dependent transition between the quadruplex and Watson-Crick hairpin forms of d(CGCG3GCG). *Biochemistry,* 31, 833-41.

Irwin, S.A., Patel, B., Idupulapati, M., Harris, J.B., Crisostomo, R.A., Larsen, B.P., Kooy, F., Willems, P.J., Cras, P., Kozlowski, P.B., Swain, R.A., Weiler, I.J. & Greenough, W.T. (2001) Abnormal dendritic spine characteristics in the temporal and visual cortices of patients with fragile-X syndrome: a quantitative examination. *Am J Med Genet,* 98, 161-7.

Ishimaru, D., Zuraw, L., Ramalingam, S., Sengupta, T.K., Bandyopadhyay, S., Reuben, A., Fernandes, D.J. & Spicer, E.K. (2010) Mechanism of regulation of Bcl-2 mRNA by nucleolin and A+U rich element binding factor 1 (AUF1). *J Biol Chem.*

Jean, J.M. & Hall, K.B. (2001) 2-Aminopurine fluorescence quenching and lifetimes: role of base stacking. *Proc Natl Acad Sci U S A,* 98, 37-41.

Joachimi, A., Benz, A. & Hartig, J.S. (2009) A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorg Med Chem,* 17, 6811-5.

Kumar, N. & Maiti, S. (2008) A thermodynamic overview of naturally occurring intramolecular DNA quadruplexes. *Nucleic Acids Res,* 36, 5610-22.

Mcghee, J.D. & Von Hippel, P.H. (1974) Theoretical aspects of DNA-protein interactions: co-operative and non-co-operative binding of large ligands to a one-dimensional homogeneous lattice. *J Mol Biol,* 86, 469-89.

Menon, L., Mader, S.A. & Mihailescu, M.R. (2008) Fragile X mental retardation protein interactions with the microtubule associated protein 1B RNA. *Rna,* 14, 1644-55.

Menon, L. & Mihailescu, M.R. (2007) Interactions of the G quartet forming semaphorin 3F RNA with the RGG box domain of the fragile X protein family. *Nucleic Acids Res,* 35, 5379-92.

Mercey, R., Lantier, I., Maurel, M.C., Grosclaude, J., Lantier, F. & Marc, D. (2006) Fast, reversible interaction of prion protein with RNA aptamers containing specific sequence patterns. *Arch Virol,* 151, 2197-214.

Mergny, J.L., Li, J., Lacroix, L., Amrane, S. & Chaires, J.B. (2005) Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res,* 33, e138.

Nandi, P.K. & Leclerc, E. (1999) Polymerization of murine recombinant prion protein in nucleic acid solution. *Arch Virol,* 144, 1751-63.

Nandi, P.K. & Sizaret, P.Y. (2001) Murine recombinant prion protein induces ordered aggregation of linear nucleic acids to condensed globular structures. *Arch Virol,* 146, 327-45.

Paramasivam, M., Membrino, A., Cogoi, S., Fukuda, H., Nakagama, H. & Xodo, L.E. (2009a) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription. *Nucleic Acids Res,* 37, 2841-53.

Paramasivam, M., Membrino, A., Cogoi, S., Fukuda, H., Nakagama, H. & Xodo, L.E. (2009b) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription. *Nucleic Acids Res.*

Paramasivan, S., Rujan, I. & Bolton, P.H. (2007) Circular dichroism of quadruplex DNAs: applications to structure, cation effects and ligand binding. *Methods,* 43, 324-31.

Ramos, A., Hollingworth, D. & Pastore, A. (2003) G-quartet-dependent recognition between the FMRP RGG box and RNA. *Rna,* 9, 1198-207.

Rich, R.L. & Myszka, D.G. (2000) Survey of the 1999 surface plasmon resonance biosensor literature. *J Mol Recognit,* 13, 388-407.

Rodriguez, A.J., Czaplinski, K., Condeelis, J.S. & Singer, R.H. (2008) Mechanisms and cellular roles of local protein synthesis in mammalian cells. *Curr Opin Cell Biol,* 20, 144-9.

Roy, S., Tanious, F.A., Wilson, W.D., Ly, D.H. & Armitage, B.A. (2007) High-affinity homologous peptide nucleic acid probes for targeting a quadruplex-forming sequence from a MYC promoter element. *Biochemistry,* 46, 10433-43.

Ruan, L., Osawa, M., Hosoda, N., Imai, S., Machiyama, A., Katada, T., Hoshino, S.I. & Shimada, I. (2010) Quantitative characterization of Tob interactions provides the thermodynamic basis for translation termination-coupled deadenylase regulation. *J Biol Chem.*

Sackett, D.L. & Saroff, H.A. (1996) The multiple origins of cooperativity in binding to multi-site lattices. *FEBS Lett,* 397, 1-6.

Scatchard, G. (1949) The Attractions of Proteins for Small Molecules and Ions. *Annals of the New York Academy of Sciences,* 51, 660-672.

Schaeffer, C., Bardoni, B., Mandel, J.L., Ehresmann, B., Ehresmann, C. & Moine, H. (2001) The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. *Embo J,* 20, 4803-13.

Shi, Z., Chen, K., Liu, Z. & Kallenbach, N.R. (2006) Conformation of the backbone in unfolded proteins. *Chem Rev,* 106, 1877-97.

Shi, Z., Woody, R.W. & Kallenbach, N.R. (2002) Is polyproline II a major backbone conformation in unfolded proteins? *Adv Protein Chem,* 62, 163-240.

Silva, J.L., Lima, L.M., Foguel, D. & Cordeiro, Y. (2008) Intriguing nucleic-acid-binding features of mammalian prion protein. *Trends Biochem Sci,* 33, 132-40.

Siomi, H., Siomi, M.C., Nussbaum, R.L. & Dreyfuss, G. (1993) The protein product of the fragile X gene, FMR1, has characteristics of an RNA-binding protein. *Cell,* 74, 291-8.

Sreerama, N. & Woody, R.W. (2003) Structural composition of betal- and betall-proteins. *Protein Sci,* 12, 384-8.

Sreerama, N. & Woody, R.W. (2004) Computation and analysis of protein circular dichroism spectra. *Methods Enzymol,* 383, 318-51.

Stampfl, S., Lempradl, A., Koehler, G. & Schroeder, R. (2007) Monovalent ion dependence of neomycin B binding to an RNA aptamer characterized by spectroscopic methods. *Chembiochem,* 8, 1137-45.

Tsodikov, O.V., Holbrook, J.A., Shkel, I.A. & Record, M.T., Jr. (2001) Analytic binding isotherms describing competitive interactions of a protein ligand with specific and nonspecific sites on the same DNA oligomer. *Biophys J,* 81, 1960-9.

Van Der Merwe, P. (2001) Surface Plasmon Resonance. IN HARDING, S. & CHOWDRY, B. (Eds.) *Protein - ligand interactions: hydrodynamics and calorimetry: a practical approach.* Oxford, Oxford University Press.

Weiss, S., Proske, D., Neumann, M., Groschup, M.H., Kretzschmar, H.A., Famulok, M. & Winnacker, E.L. (1997) RNA aptamers specifically interact with the prion protein PrP. *J Virol,* 71, 8790-7.

Woody, R.W. (2009) Circular dichroism spectrum of peptides in the poly(Pro)II conformation. *J Am Chem Soc,* 131, 8234-45.

Yazdani, U. & Terman, J.R. (2006) The semaphorins. *Genome Biol,* 7, 211.

Zanotti, K.J., Lackey, P.E., Evans, G.L. & Mihailescu, M.R. (2006) Thermodynamics of the fragile X mental retardation protein RGG box interactions with G quartet forming RNA. *Biochemistry,* 45, 8319-30.

Zeiler, B., Adler, V., Kryukov, V. & Grossman, A. (2003) Concentration and removal of prion proteins from biological solutions. *Biotechnol Appl Biochem,* 37, 173-82.

Zhang, D.H., Fujimoto, T., Saxena, S., Yu, H.Q., Miyoshi, D. & Sugimoto, N. (2010) Monomorphic RNA G-quadruplex and polymorphic DNA G-quadruplex structures responding to cellular environmental factors. *Biochemistry,* 49, 4554-63.

CHAPTER 10

CONCLUSIONS

The primary aim of this work was to investigate the nature of a highly conserved region of the N-terminus of the Shadoo protein. A survey of the available literature led to the idea that this region may be an RGG box and therefore an RNA binding domain. This observation forms the major hypothesis investigated in this thesis. A series of studies were designed to test this hypothesis and establish the plausibility of Sho playing a functional role as an RNA-binding protein.

Initial conceptual work involved surveying other RGG box proteins from the literature and comparing their sequences and properties to Sho. This work, presented in Chapter 2, supported the idea that the highly conserved N-terminal region of Sho does contain an RGG box. In addition, this analysis demonstrated the highly variable functional roles and cellular locations ascribed to proteins containing this motif. Insights from this analysis were used to formulate functional hypotheses for Sho. It seemed most likely that Sho could bind extracellular RNA in a signaling context, but also that it may lose its GPI anchor through cleavage and, in this truncated form, carry out RNA binding functions in the cytoplasm and/or nucleus of neurons. During this analysis, the Fragile-X mental Retardation Protein (FMRP) was identified as a well-studied system with similarities to Sho. Both have a similar biological expression pattern and there is high sequence homology in the RGG box regions of both. Therefore, subsequent work has centred around comparing the RGG boxes of Sho and FMRP.

Chapter 4 contains detailed structural comparisons between the Sho and FMRP RGG boxes (referred to as 'Sho' and 'Fx' respectively), using molecular dynamics (MD) simulations. As both the Sho and Fx RGG boxes occur in disordered regions, this analysis represents a departure from the normal use of modeling techniques, where the aim is to characterize a single stable structure or a small number of low energy states. For disordered peptides, it is necessary to consider the entire set of energetically accessible conformers as a whole. This ensemble-based analysis provided insights, at the molecular level, into the similarities and differences between the Sho and Fx peptides.

Overall, Sho and Fx were very similar in the range of conformational space they accessed. Both were extremely flexible with no permanent backbone hydrogen bonds. The most prominent structural element in both was the turn, followed by the β-bridge, consistent with flexible coil behaviour. However, Ramachandran plots revealed the propensity for both peptides to form extended conformations especially in the P_{II} region of conformational space associated with flexible protein domains.

Some differences between the peptides were noted. In particular, Fx has a slightly higher tendency to form extended conformations than Sho. Two explanations are suggested for this. Firstly, closely spaced Arg residues in Fx are likely to increase electrostatic repulsion and favour the extended structure. Second, Sho has 3 Ala residues which make hydrophobic contacts some of the time, resulting in more condensed local structure. These contacts also appear to be associated with transient 3^{10} helical structure in the middle of the Sho peptide.

On a broader scale, this work suggested ways in which the RGG box domain may bind with a partner. The abundant Gly residues confer a high degree of flexibility to the peptide backbone, which increases the conformational space accessible to other residues. In the case of the RGG box, this is particularly noticeable for the Arg residues. This led to the conclusion that the torsional flexibility of a Gly rich peptide, together with the inherent capacity of the Arg residues to participate in multiple binding interactions, creates a system poised to adopt numerous binding modes and partners.

This study also presented the opportunity to test two different solvation methods for the simulation of peptides. The two methods were found to produce different ensemble properties. Backbone hydrogen bonds were seen more frequently in implicit solvation simulations than in hybrid-REMD simulations, leading to different global properties. This finding agreed with other studies (Okur et al., 2006, Roe et al., 2007, Okur et al., 2008) and provides further evidence that implicit solvation is not yet reliable for investigating ensemble properties of highly charged peptides.

Next, simulations were performed to better understand the way in which these RGG box peptides bind to G-quadruplex RNA, known to have an affinity with RGG box

proteins. Before this could be done, a model RNA quadruplex was designed and tested for stability. The work presented in Chapter 5 demonstrated the suitability of the model RNA quadruplex for use in further modeling studies.

Simulations of the Sho-RNA and Fx-RNA complexes (Chapters 6-8) revealed a great deal of variability in the way the peptides may bind to quadruplex RNA through an array of salt bridges, hydrogen bonds and van der Waals interactions. Despite this high degree of variation there were some common features found in both Sho and Fx complexes. The most prominent common binding characteristic is the role played by the Arg residues. Multiple Arg residues participate in binding and have a preference for interacting with RNA phosphate groups both through the formation of hydrogen bonds and more general electrostatic interactions. The obvious importance of Arg interactions with the RNA explains why post-translational modifications such as Arg methylation affect RGG box binding, as discussed in Chapter 2. The placement of the Arg residues in the peptide did not appear to influence the number of contacts made with the RNA. For example, for the Fx peptide, the 3 adjacent Arg residues at the N-terminus made a similar number of contacts with the RNA as Arg4 and Arg 16 which are involved in RGG repeats.

As well as its obvious role in facilitating conformational flexibility of the peptide backbone, Gly, the other highly represented residue, contributed to the binding interface by making close contacts with the nucleotide bases including the guanine bases involved in the tetrad stack.

Analysis of the binding interface between the model G-quadruplex RNA and the RGG box peptides revealed that the types of contacts made by G-quadruplex RNA differ from that normally seen with flexible single-stranded RNA. Around 75% of contacts are made through the sugar–phosphate backbone, with fewer contacts to the nucleic acid bases than would be seen with a flexible single-stranded RNA. In this sense, the RNA quadruplex is also similar to structured DNA. These differences are explained by the quadruplex structure in which the guanine bases associate with each other in Hoogsteen hydrogen bonds involving the N1, N2, N7 and O6 atoms. In other forms of RNA, guanine commonly forms hydrogen bonds through the N7 and O6 atoms. In addition the sugar–phosphate backbone shields the guanine bases as well as providing an accessible binding surface.

On a global scale, i.e. looking at the total numbers of contacts made by each of the Sho and Fx residues and the total number of contacts made to each of the RNA nucleotides, a very similar picture of the potential for binding emerged for both peptides. Viewed from the finer perspective of individual complexes (Chapter 8), the variability in the binding interface of highly stable complexes is obvious. At the same time certain features were identified as being associated with high binding affinity.

Free energy calculations confirmed that many of the RGG box complexes studied are thermodynamically indistinguishable using the methods employed in this study. The overall stability of the complex, or free energy, was influenced by de/stabilization of both the peptide and the RNA upon complexation, as well as the number and type of contacts that were formed. Each of these factors contributed to the total free energy ranking of the complexes.

It appears that the conformational flexibility of the RGG box peptide allows it to interact with the stable quadruplex structure to produce an ensemble of low energy states with diverse structures. Accordingly, it is fitting to describe the RGG box motif as a highly variable RNA-binding platform, similar to the RNA recognition motif (Maris et al., 2005). The current data also support the more general suggestion that an Arg-rich motif provides a flexible framework for recognition of RNA structures (Harada et al., 1997).

Overall, the MD simulations support the proposition that the RGG box of Sho is competent to bind quadruplex RNA, with a high degree of plasticity. In other words, the Sho RGG box is likely to bind to multiple RNA targets in a variety of different conformations and orientations. Finally, the simulations suggest that quadruplex RNA will not be destabilized by 1:1 peptide binding.

Biophysical experiments were conducted to test these predictions (Chapter 9). In these experiments, 3 peptides were tested; the Sho and Fx RGG box domains as well as a sequence from the N-terminus of PrP (res 23-52) known to bind RNA (Weiss et al., 1997, Proske et al., 2002, Mercey et al., 2006). Five RNA sequences were used as binding targets. Two of these sequences (SC and SF) were known from earlier studies (Darnell et al., 2001, Zanotti et al., 2006, Menon and Mihailescu, 2007), to form a G-quadruplex structure and to bind with high affinity to Fx. The sequence designed for

the simulation studies (Xav) was also tested, as was a polyguanine sequence (PG) and a mutated form of SF (SM). CD spectra of the peptides demonstrated that Sho and Fx have a combination of random coil and PP_{II} structure, as predicted by the MD simulations. The CD spectra of the 5 RNA oligonucleotides were consistent with G-quadruplex structure although this characterization should be confirmed by complementary techniques, as noted in Chapter 9.

Surface Plasmon Resonance (SPR) experiments were conducted with all 3 peptides and all 5 RNA oligonucleotides. These experiments produced interesting preliminary results and warrant further refinement. The initial SPR results indicate that the 3 peptides bind in a concentration dependent manner and with different affinities to the 5 RNA targets. It was interesting to find that all 3 peptides bound most tightly and with similar predicted affinity to the SC RNA (K_d~ 20 nM), as it has previously been shown that this sequence is a favoured target of Fx (Darnell et al., 2001). Binding to the other RNA transcripts was weaker and distinguished by the extent of association. These results provide evidence that the peptides are capable of distinguishing between different RNA targets, which in turn have different modes of binding to the peptides. The SPR experiments also reveal that binding of the SF RNA to each of the peptides is complex rather than simple 1:1 binding. Additional SPR experiments should be conducted to confirm these results. The protocol used in these experiments should attempt to minimize cell surface binding and should include switching the analyte and ligands. Different types of RNA, including sequences demonstrated not to form a G-quadruplex structure, should also be tested.

The complexity of peptide binding to the SF quadruplex made it difficult to obtain a reliable estimate of K_d for these interactions. Dissociation constants derived from the SPR experiments and those derived from fluorescence experiments varied by an order of magnitude. Both experimental techniques produced results indicating that binding was complex, and that the estimated binding affinity for all 3 peptides was in a similar range. Although the estimates of K_d for SF binding derived from the SPR experiments should be treated cautiously, it was interesting to see that the result for Fx binding was similar to that reported previously (Darnell et al., 2001). Future studies should investigate this complex binding more thoroughly. Initially, experiments should be conducted to determine whether the SF sequence is forming multiple tertiary structures. The possibility that the peptide is aggregating should also be assessed.

Peptide mutagenesis studies confirmed that binding is mediated through the Arg residues of the Sho RGG box. Comparison of the native peptide with a Lys substituted mutant demonstrated that the electrostatic interactions created by the positive charges on Arg only account for part of the overall affinity. The unique hydrogen bond network formed by the Arg residues is important in complexation. It was also interesting to see that the scrambled peptide produced a very similar binding isotherm and estimated binding affinity to that of the Sho peptide. This indicates that closely positioned Arg and Gly residues may constitute a sufficient binding platform without necessarily being ordered into RGG repeats.

CD experiments indicated that peptide binding has very little effect on the RNA quadruplex at near equimolar ratios but that the characteristic quadruplex spectrum decreased at higher peptide:RNA ratios. Similar spectra have previously been interpreted as quadruplex unwinding (Menon and Mihailescu, 2007, Menon et al., 2008, Paramasivam et al., 2009). However, it is noted that there is no evidence of an increase in the peak at 280 nm associated with unstructured RNA. An alternative explanation is that the RNA may aggregate in the presence of the peptides. Irrespective of the interpretation, the CD experiments show that the interaction of the peptide and the RNA quadruplex has the effect of altering the RNA and most likely decreasing its availability for further interaction, either through unwinding or the formation of RNA aggregates. The fact that this occurs for all 3 peptides and all 5 RNA sequences suggests this may be a common regulatory mechanism. Given the potential biological importance of this observation, future work should be conducted to confirm quadruplex unwinding. Such work should include complementary techniques such as UV melting experiments and NMR to confirm that quadruplex unwinding is occurring in the presence of these peptides. The mechanism by which peptide binding might stabilize the quadruplex structure at a 1:1 ratio but then lead to destabilization at higher peptide ratios requires further study.

Overall, the experimental studies supported the initial hypothesis that the RGG box of Sho can function as an RNA-binding domain, and agreed with the observations on peptide, RNA and complex stabilities drawn from the molecular dynamics simulations. Sho and Fx both bound to the 5 RNAs tested, revealing some plasticity in the way in which the peptide can adapt to a different RNA target. The binding affinity of Sho for the RNAs tested was of a similar order of magnitude as that found for Fx, which is

known to have a functional role in the transport and translational control of mRNA (Bagni and Greenough, 2005) and PrP which has a well-documented capacity for RNA binding (Weiss et al., 1997, Proske et al., 2002, Adler et al., 2003, Zeiler et al., 2003, Mercey et al., 2006).

Accordingly, this work provides proof-of-principle that the RGG box region of Sho is capable of binding quadruplex RNA with biologically significant affinity. Future work should examine whether the Sho RGG box domain also binds other types of RNA (i.e. non-quadruplex forming RNAs) and identify whether certain RNA sequences have particular affinity for the Sho RGG box domain, possibly using similar experimental methods to those used by Mercey and colleagues to identify RNA aptamers with specific affinity for PrP (Mercey et al., 2006). An ultimate aim of future work will be to discover RNA sequences that bind Sho *in vivo*. This knowledge will lead to significant advances in understanding the natural function of Sho.

The functions performed by PrP and Sho will depend on their cellular location, i.e. their local environment. Although PrP has been found in the cytoplasm and nucleus it is generally suggested that this occurs as a result of some defect in the normal process by which GPI anchored proteins are trafficked to the surface of the cell and then recycled e.g. a defect in the proteasomal degradation pathway (Lee et al., 2001, Ma et al., 2002, Mironov et al., 2003, Mange et al., 2004, Campana et al., 2005). This type of information is not yet available for Sho, however, it is plausible that its signal sequence for GPI anchor attachment and subsequent GPI anchor processing could cause it to follow a similar path to PrP. Although there has been some speculation that PrP may have an intracellular RNA binding function, for example; binding mRNA (Radulescu and Korth, 1996), nucleic acid metabolism (Gabus et al., 2001), or transport of nucleic acid across the cell milieu (Lima et al., 2006), there is currently no evidence of such a function.

PrP and Sho may have intracellular roles. However, the fact that PrP and Sho are genetically coded to have an attached GPI anchor and thus be directed to the outer cell membrane suggests more strongly that PrP and Sho are meant to act as mediators between the extracellular and intracellular environments. It is plausible that PrP and Sho both have a role in binding extracellular RNA. It is interesting to note that the RGG box of Sho is at the same position as the RNA binding site of PrP, i.e. at the start of the N-terminus. When attached to the cell surface by a GPI anchor at the

C-terminus, the N-terminus would have a good deal of flexibility to bind a range of extracellular RNA. Evidence is emerging that extracellular RNA may comprise a new class of signaling molecule mediating cell-to-cell communication (Dinger et al., 2008). At the same time, as noted in Chapter 2, the number of identified RNA-binding proteins with dual roles as cell surface receptors is growing. Viewed in this context, PrP and Sho could interact with extracellular RNA and assist in its transport across the cell membrane, or RNA binding could initiate a signaling pathway with PrP/Sho as the protein mediator in this process.

Extracellular RNA and DNA are also associated with pathological conditions e.g. stroke, head injury and tumor burden (Fischer et al., 2007). It has previously been shown that PrP has a neuroprotective role in stroke, most likely mediated through a signaling mechanism (McLennan et al., 2004, Spudich et al., 2005). It is plausible that this function may be influenced by the recognition of extracellular nucleic acids released as a result of trauma.

If RNA binding is a normal function for GPI-anchored PrP it is feasible that binding to an aberrant RNA sequence may play a role in PrP conversion to the abnormally folded and disease producing isoform PrP^{Sc}. This proposition is consistent with recent findings in relation to the conversion of PrP^{C} to PrP^{Sc} (Wang et al., 2010) and the propagation of prion disease.

In summary, the work undertaken in this thesis supports the hypothesis that Sho has an RGG box that is competent to bind RNA. There is growing evidence that PrP and Sho have similar functions and may compensate for loss of one another (Watts et al., 2007, Young et al., 2009). These roles are likely to be essential for normal neural development and activity. A comprehensive understanding of the mechanisms by which these proteins work is therefore important. The work presented in this thesis contributes to this field by revealing that Sho has an RNA binding domain and therefore is likely to function, in some capacity, as an RNA-binding protein. This work also draws new comparisons between Sho and PrP and adds weight to the suggestion that PrP has a function in RNA binding. Future *in vitro* and *in vivo* studies to identify likely RNA binding partners for Sho provide a promising avenue for elucidating the natural function of this protein as well as shedding light on the elusive function of PrP.

10.1 REFERENCES

Adler, V., Zeiler, B., Kryukov, V., Kascsak, R., Rubenstein, R. & Grossman, A. (2003) Small, highly structured RNAs participate in the conversion of human recombinant PrP(Sen) to PrP(Res) in vitro. *J Mol Biol,* 332, 47-57.

Bagni, C. & Greenough, W.T. (2005) From mRNP trafficking to spine dysmorphogenesis: the roots of fragile X syndrome. *Nat Rev Neurosci,* 6, 376-87.

Bajenova, O., Stolper, E., Gapon, S., Sundina, N., Zimmer, R. & Thomas, P. (2003) Surface expression of heterogeneous nuclear RNA binding protein M4 on Kupffer cell relates to its function as a carcinoembryonic antigen receptor. *Exp Cell Res,* 291, 228-41.

Belyanskaya, L.L., Delattre, O. & Gehring, H. (2003) Expression and subcellular localization of Ewing sarcoma (EWS) protein is affected by the methylation process. *Exp Cell Res,* 288, 374-81.

Benimetskaya, L., Loike, J.D., Khaled, Z., Loike, G., Silverstein, S.C., Cao, L., El Khoury, J., Cai, T.Q. & Stein, C.A. (1997) Mac-1 (CD11b/CD18) is an oligodeoxynucleotide-binding protein. *Nat Med,* 3, 414-20.

Campana, V., Sarnataro, D. & Zurzolo, C. (2005) The highways and byways of prion protein trafficking. *Trends Cell Biol,* 15, 102-11.

Chang, L.Y., Mohd Ali, A.R., Hassan, S.S. & Abubakar, S. (2007) Human neuronal cell protein responses to Nipah virus infection. *Virol J,* 4, 54.

Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T. & Darnell, R.B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell,* 107, 489-99.

Dinger, M.E., Mercer, T.R. & Mattick, J.S. (2008) RNAs as extracellular signaling molecules. *J Mol Endocrinol,* 40, 151-9.

Fischer, S., Gerriets, T., Wessels, C., Walberer, M., Kostin, S., Stolz, E., Zheleva, K., Hocke, A., Hippenstiel, S. & Preissner, K.T. (2007) Extracellular RNA mediates endothelial-cell permeability via vascular endothelial growth factor. *Blood,* 110, 2457-65.

Gabus, C., Auxilien, S., Pechoux, C., Dormont, D., Swietnicki, W., Morillas, M., Surewicz, W., Nandi, P. & Darlix, J.L. (2001) The prion protein has DNA strand transfer properties similar to retroviral nucleocapsid protein. *J Mol Biol,* 307, 1011-21.

Harada, K., Martin, S.S., Tan, R. & Frankel, A.D. (1997) Molding a peptide into an RNA site by in vivo peptide evolution. *Proc Natl Acad Sci U S A,* 94, 11887-92.

Hirano, K., Miki, Y., Hirai, Y., Sato, R., Itoh, T., Hayashi, A., Yamanaka, M., Eda, S. & Beppu, M. (2005) A multifunctional shuttling protein nucleolin is a macrophage receptor for apoptotic cells. *J Biol Chem,* 280, 39284-93.

Lee, K.S., Magalhaes, A.C., Zanata, S.M., Brentani, R.R., Martins, V.R. & Prado, M.A. (2001) Internalization of mammalian fluorescent cellular prion protein and N-terminal

deletion mutants in living cells. *J Neurochem,* 79, 79-87.

Lima, L.M., Cordeiro, Y., Tinoco, L.W., Marques, A.F., Oliveira, C.L., Sampath, S., Kodali, R., Choi, G., Foguel, D., Torriani, I., Caughey, B. & Silva, J.L. (2006) Structural insights into the interaction between prion protein and nucleic acid. *Biochemistry,* 45, 9180-7.

Ma, J., Wollmann, R. & Lindquist, S. (2002) Neurotoxicity and neurodegeneration when PrP accumulates in the cytosol. *Science,* 298, 1781-5.

Mange, A., Crozet, C., Lehmann, S. & Beranger, F. (2004) Scrapie-like prion protein is translocated to the nuclei of infected cells independently of proteasome inhibition and interacts with chromatin. *J Cell Sci,* 117, 2411-6.

Maris, C., Dominguez, C. & Allain, F.H. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *Febs J,* 272, 2118-31.

Mclennan, N.F., Brennan, P.M., Mcneill, A., Davies, I., Fotheringham, A., Rennison, K.A., Ritchie, D., Brannan, F., Head, M.W., Ironside, J.W., Williams, A. & Bell, J.E. (2004) Prion protein accumulation and neuroprotection in hypoxic brain damage. *Am J Pathol,* 165, 227-35.

Menon, L., Mader, S.A. & Mihailescu, M.R. (2008) Fragile X mental retardation protein interactions with the microtubule associated protein 1B RNA. *Rna,* 14, 1644-55.

Menon, L. & Mihailescu, M.R. (2007) Interactions of the G quartet forming semaphorin 3F RNA with the RGG box domain of the fragile X protein family. *Nucleic Acids Res,* 35, 5379-92.

Mercey, R., Lantier, I., Maurel, M.C., Grosclaude, J., Lantier, F. & Marc, D. (2006) Fast, reversible interaction of prion protein with RNA aptamers containing specific sequence patterns. *Arch Virol,* 151, 2197-214.

Mironov, A., Jr., Latawiec, D., Wille, H., Bouzamondo-Bernstein, E., Legname, G., Williamson, R.A., Burton, D., Dearmond, S.J., Prusiner, S.B. & Peters, P.J. (2003) Cytosolic prion protein in neurons. *J Neurosci,* 23, 7183-93.

Okur, A., Wickstrom, L., Layten, M., Geney, R., Song, K., Hornak, V. & Simmerling, C. (2006) Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model. *Journal of Chemical Theory and Computation,* 2, 420-433.

Okur, A., Wickstrom, L. & Simmerling, C. (2008) Evaluation of salt bridge structure and energetics in peptides using explicit, implicit, and hybrid solvation models. *Journal Of Chemical Theory And Computation,* 4, 488-498.

Paramasivam, M., Membrino, A., Cogoi, S., Fukuda, H., Nakagama, H. & Xodo, L.E. (2009) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription. *Nucleic Acids Res.*

Proske, D., Gilch, S., Wopfner, F., Schatzl, H.M., Winnacker, E.L. & Famulok, M. (2002) Prion-protein-specific aptamer reduces PrPSc formation. *Chembiochem,* 3, 717-25.

Radulescu, R.T. & Korth, C. (1996) Prion function and dysfunction: a structure-based

scenario. *Med Hypotheses,* 46, 225-8.

Roe, D.R., Okur, A., Wickstrom, L., Hornak, V. & Simmerling, C. (2007) Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J Phys Chem B,* 111, 1846-57.

Said, E.A., Krust, B., Nisole, S., Svab, J., Briand, J.P. & Hovanessian, A.G. (2002) The anti-HIV cytokine midkine binds the cell surface-expressed nucleolin as a low affinity receptor. *J Biol Chem,* 277, 37492-502.

Spudich, A., Frigg, R., Kilic, E., Kilic, U., Oesch, B., Raeber, A., Bassetti, C.L. & Hermann, D.M. (2005) Aggravation of ischemic brain injury by prion protein deficiency: role of ERK-1/-2 and STAT-1. *Neurobiol Dis,* 20, 442-9.

Wang, F., Wang, X., Yuan, C.G. & Ma, J. (2010) Generating a prion with bacterially expressed recombinant prion protein. *Science,* 327, 1132-5.

Watts, J.C., Drisaldi, B., Ng, V., Yang, J., Strome, B., Horne, P., Sy, M.S., Yoong, L., Young, R., Mastrangelo, P., Bergeron, C., Fraser, P.E., Carlson, G.A., Mount, H.T., Schmitt-Ulms, G. & Westaway, D. (2007) The CNS glycoprotein Shadoo has PrP(C)-like protective properties and displays reduced levels in prion infections. *Embo J,* 26, 4038-50.

Weiss, S., Proske, D., Neumann, M., Groschup, M.H., Kretzschmar, H.A., Famulok, M. & Winnacker, E.L. (1997) RNA aptamers specifically interact with the prion protein PrP. *J Virol,* 71, 8790-7.

Young, R., Passet, B., Vilotte, M., Cribiu, E.P., Beringue, V., Le Provost, F., Laude, H. & Vilotte, J.L. (2009) The prion or the related Shadoo protein is required for early mouse embryogenesis. *FEBS Lett,* 583, 3296-300.

Zanotti, K.J., Lackey, P.E., Evans, G.L. & Mihailescu, M.R. (2006) Thermodynamics of the fragile X mental retardation protein RGG box interactions with G quartet forming RNA. *Biochemistry,* 45, 8319-30.

Zeiler, B., Adler, V., Kryukov, V. & Grossman, A. (2003) Concentration and removal of prion proteins from biological solutions. *Biotechnol Appl Biochem,* 37, 173-82.

APPENDIX 4

Acceptor	Donor	%	Distance
Sho1			
:10@O	:13@N	10.33	3.084
:9@O	:12@N	8.96	3.174
:7@O	:10@N	6.29	3.052
:14@O	:18@N	6.19	3.025
:10@O	:12@N	5.95	3.02
:17@O	:19@N	5.18	3.056
Sho2			
:10@O	:13@N	12.62	3.096
:9@O	:12@N	11.75	3.165
:10@O	:12@N	7.49	3.022
:17@O	:19@N	6.7	3.047
:8@O	:11@N	6.66	3.141
:6@O	:9@N	6.2	3.096
:11@O	:14@N	6.08	3.081
:7@O	:10@N	5.97	3.069
:8@O	:10@N	5.48	3.022
:13@O	:15@N	5.2	2.906
:21@O	:15@N	5.09	2.824
Sho3			
:9@O	:12@N	11.19	3.167
:10@O	:13@N	9.82	3.104
:8@O	:11@N	6.78	3.154
:21@O	:15@N	6.37	2.823
:17@O	:19@N	5.73	3.052
:10@O	:12@N	5.68	3.033
:17@O	:20@N	5.48	3.166
:7@O	:10@N	5.42	3.069
Fx1			
:11@O	:13@N	8.22	3.09
:18@O	:20@N	5.77	3.116
:8@O	:11@N	5.02	3.11
Fx2			
:11@O	:13@N	8.91	3.101
:18@O	:20@N	7.56	3.114
:15@O	:18@N	6.72	3.106
:10@O	:13@N	5.84	3.132
:12@O	:14@N	5.57	3.093
:11@O	:14@N	5.49	3.158
:7@O	:9@N	5.28	3.079
Fx3			
:11@O	:13@N	7.01	3.093
:18@O	:20@N	6	3.136
:15@O	:18@N	5.01	3.118

Acceptor	Donor	% Occupancy	Distance
Sho1			
:21@O	:10@N	30.18	2.855
:21@O	:11@N	25.05	2.898
:21@O	:12@N	23.08	2.877
:21@O	:13@N	20.87	2.961
:21@O	:14@N	18.62	3.03
:21@O	:15@N	15.32	2.787
:17@O	:19@N	11.3	3.115
:15@O	:21@N	7.82	3.114
:13@O	:15@N	7.8	3.016
:9@O	:12@N	7.09	3.194
:17@O	:20@N	6.9	3.236
:20@O	:17@N	6.11	3.083
:21@O	:9@N	6.09	3.06
:21@O	:8@N	5.25	3.02
Sho2			
:21@O	:10@N	38.02	2.887
:21@O	:13@N	34.57	2.955
:21@O	:15@N	32.55	2.783
:21@O	:12@N	30.77	2.875
:21@O	:11@N	27.8	2.867
:21@O	:14@N	23.59	3.028
:15@O	:21@N	22.47	3.112
:17@O	:19@N	11.82	3.129
:6@O	:17@N	10.39	3.057
Fx1			
:11@O	:14@N	12.9	3.20
:20@O	:4@N	11.7	3.00
:14@O	:11@N	8.6	3.09
:11@O	:13@N	7.6	3.14
:20@O	:17@N	7.1	3.04
:20@O	:5@N	7.0	3.07
:12@O	:14@N	6.7	3.20
:11@O	:15@N	6.6	3.02
:12@O	:20@N	6.5	3.01
:12@O	:21@N	5.8	3.12
:18@O	:20@N	5.6	3.19
:21@O	:11@N	5.4	2.84
:21@O	:12@N	5.2	2.99
:7@O	:9@N	5.0	3.18
Fx2			
:11@O	:14@N	14.0	3.20
:14@O	:11@N	9.8	3.08
:11@O	:13@N	8.8	3.12
:12@O	:14@N	7.3	3.18
:11@O	:15@N	7.3	3.01
:20@O	:17@N	6.8	3.04
:18@O	:20@N	6.2	3.18
:17@O	:19@N	5.7	3.17
:15@O	:11@N	5.2	3.11

:21@O	:16@N	6.2	2.894
:17@O	:6@N	5.64	3.114
:17@O	:20@N	5.38	3.214

Sho3			
:21@O	:15@N	34.17	2.796
:21@O	:14@N	19.16	2.96
:21@O	:10@N	14.85	2.925
:15@O	:21@N	14.14	3.118
:17@O	:19@N	12.86	3.132
:21@O	:13@N	10.22	3.059
:11@O	:8@N	6.92	3.092
:21@O	:11@N	6.69	2.923
:13@O	:15@N	6.51	3.016
:9@O	:12@N	6.18	3.215
:6@O	:13@N	6.07	3.077
:21@O	:12@N	6	2.895
:8@O	:11@N	5.08	3.209

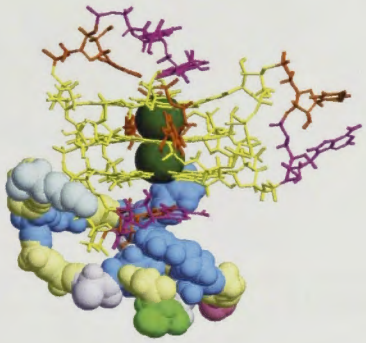
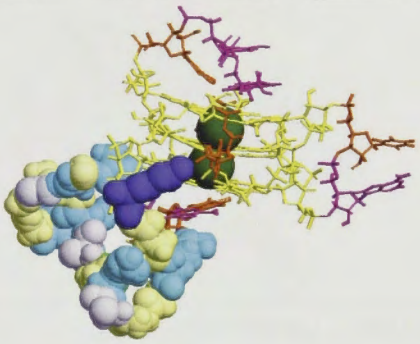
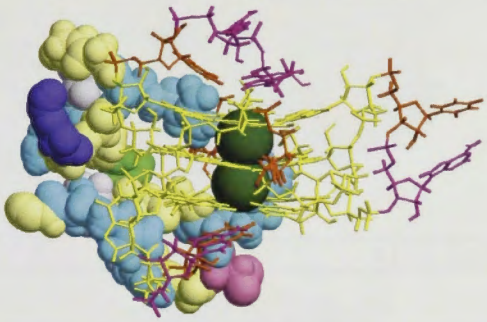
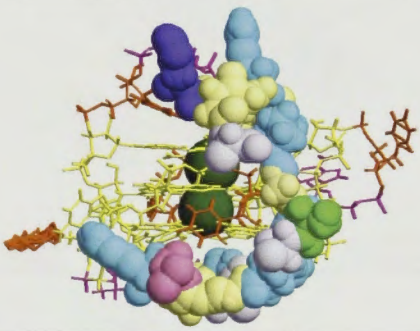
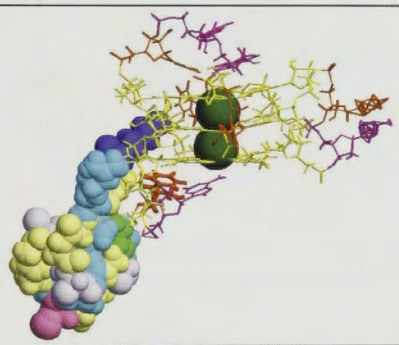
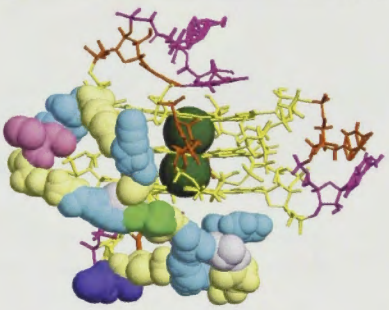
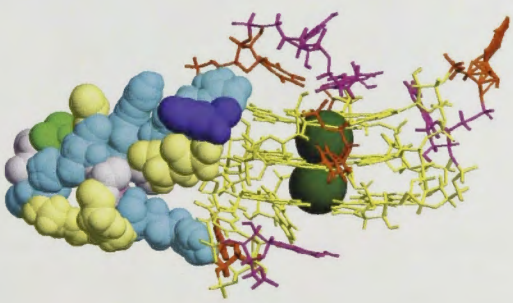
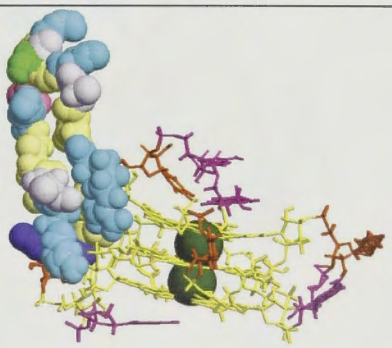
Sho4			
:21@O	:10@N	19.4	2.896
:21@O	:14@N	17.95	3.009
:21@O	:15@N	14.61	2.792
:21@O	:11@N	14.38	2.908
:11@O	:8@N	13.96	3.023
:21@O	:12@N	12.3	2.858
:6@O	:13@N	12.07	3.062
:21@O	:13@N	11.95	2.945
:17@O	:19@N	9.5	3.129
:13@O	:15@N	9.14	3.009
:21@O	:9@N	8.65	3.034
:20@O	:17@N	6.73	3.06
:7@O	:18@N	6.57	3.093
:9@O	:12@N	6.46	3.202
:15@O	:21@N	5.99	3.126
:8@O	:10@N	5.87	3.15
:8@O	:11@N	5.36	3.209

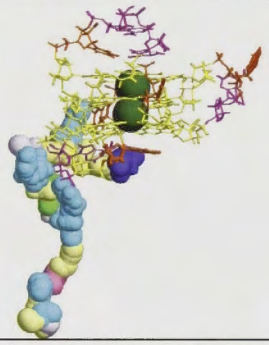
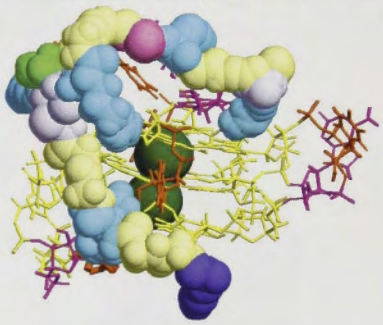
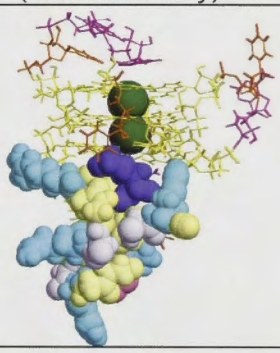
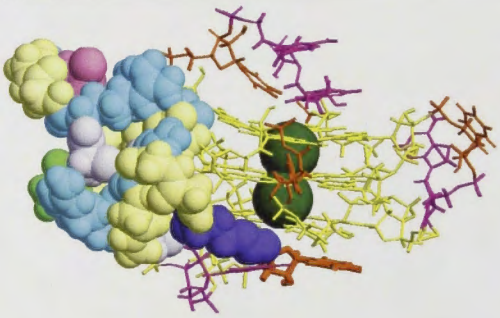
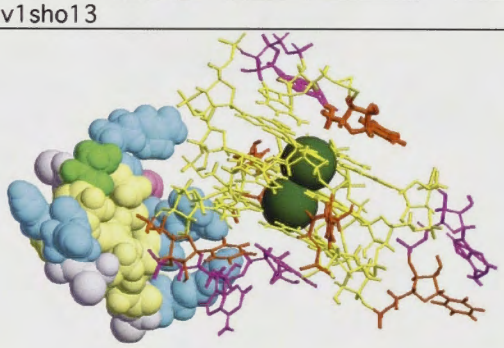
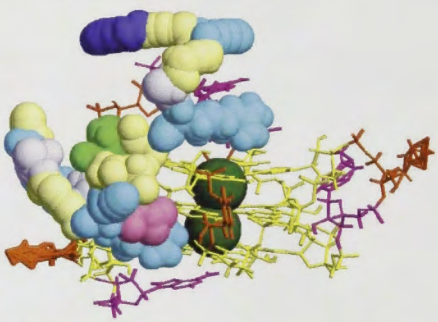
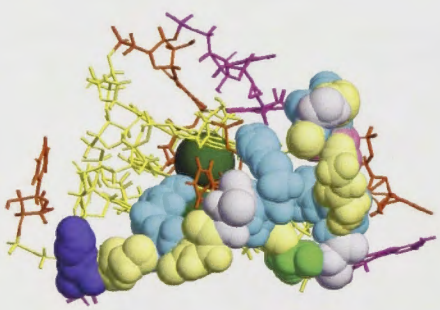
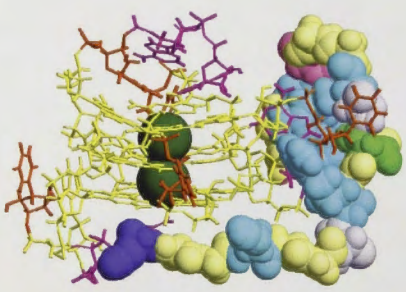
:7@O	:9@N	5.0	3.18

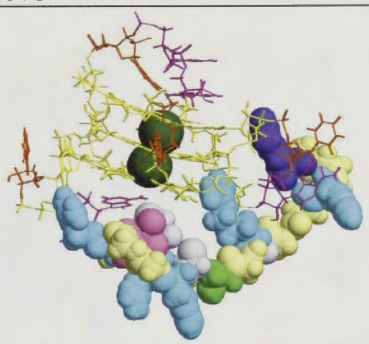
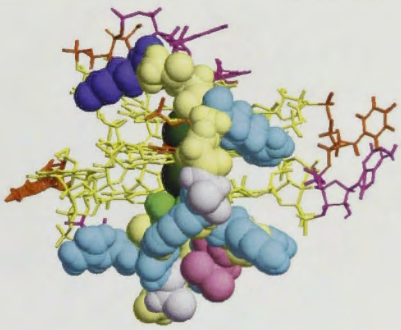
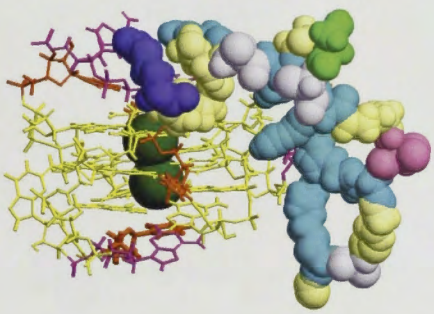
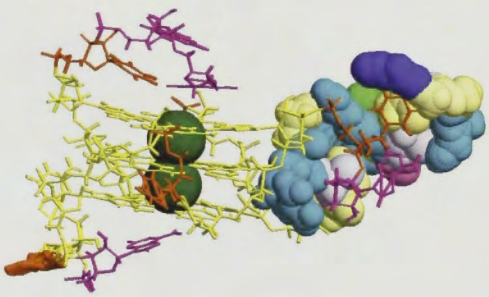
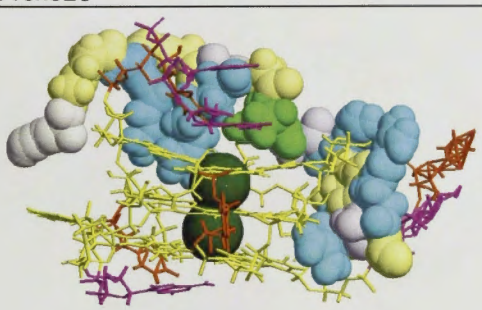
Fx3			
:11@O	:14@N	12.6	3.19
:11@O	:13@N	8.9	3.12
:12@O	:14@N	7.0	3.17
:18@O	:20@N	6.6	3.18
:11@O	:15@N	6.2	3.01
:10@O	:13@N	5.7	3.15
:14@O	:11@N	5.6	3.09
:7@O	:9@N	5.4	3.17
:17@O	:19@N	5.1	3.17
:16@O	:19@N	5.1	3.19

Fx4			
:11@O	:14@N	12.8	3.21
:11@O	:13@N	9.4	3.13
:14@O	:11@N	8.7	3.08
:20@O	:17@N	8.6	3.04
:12@O	:14@N	7.8	3.18
:21@O	:11@N	6.0	2.84
:18@O	:20@N	5.9	3.19
:17@O	:19@N	5.9	3.17
:21@O	:12@N	5.5	2.93
:7@O	:9@N	5.3	3.18
:11@O	:15@N	5.0	3.03

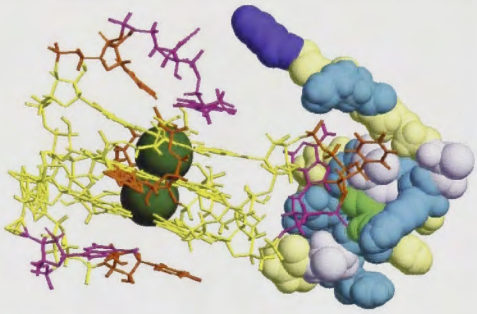
ANNEXURE 6A Sho-RNA Complexes

xav1sho1	Xav1sho2
	
Xav1sho3	Xav1sho4
	
Xav1sho5	Xav1sho6
	
Xav1sho7	Xav1sho8
	

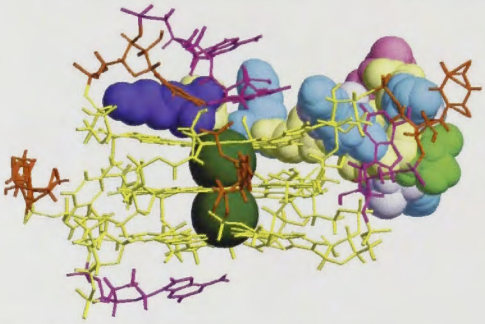
Xav1sho9	Xav1sho10
	
Xav1sho11 (terminated early)	Xav1sho12
	
Xav1sho13	Xav1sho14
	
Xav1sho15	Xav1sho16
	

Xav1sho17	Xav1sho18
	
Xav1sho19	Xav1sho20
	
Xav1sho21	Xav1sho22
	
Xav1sho23	
	

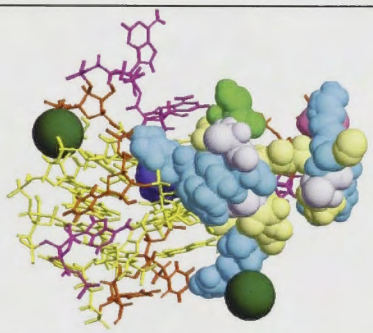
Xav1sho25



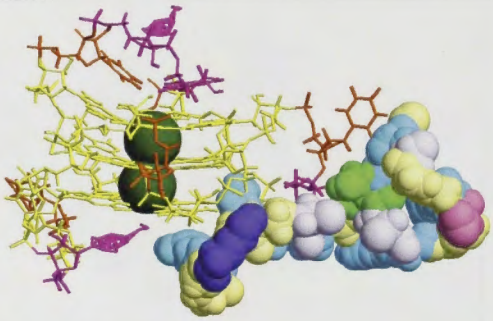
Xav1sho26



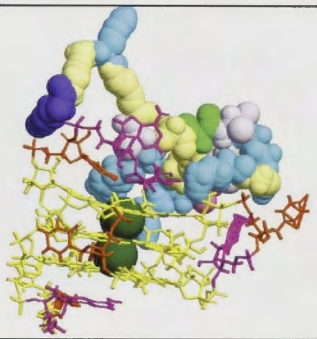
Xav1sho27



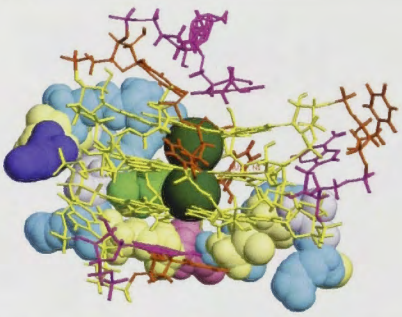
Xav1sho28



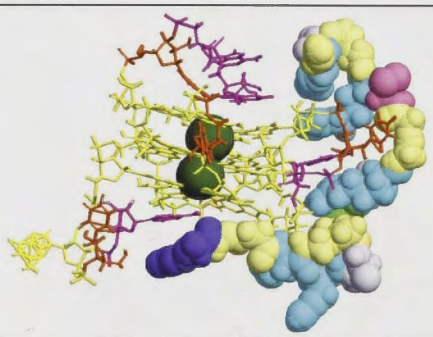
Xav1sho29



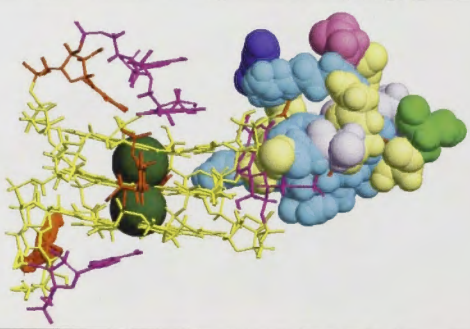
Xav1sho30

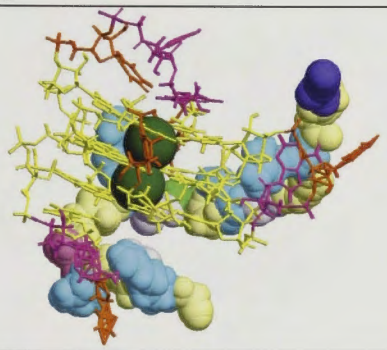
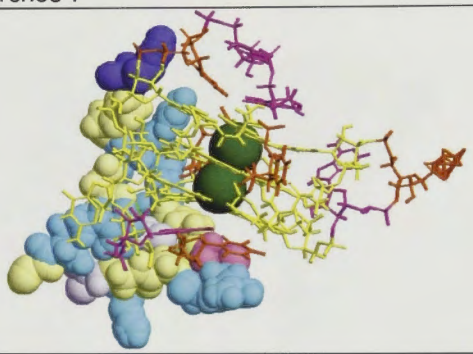
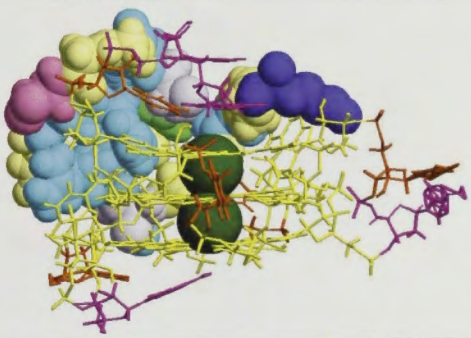
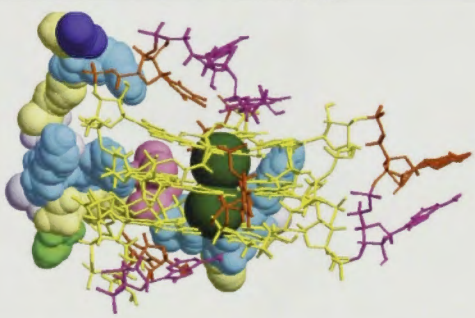
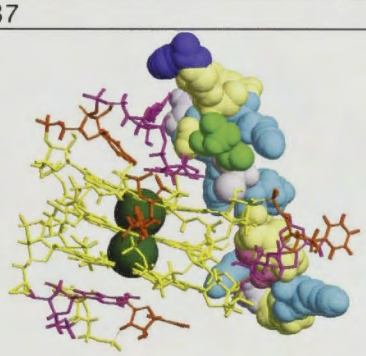
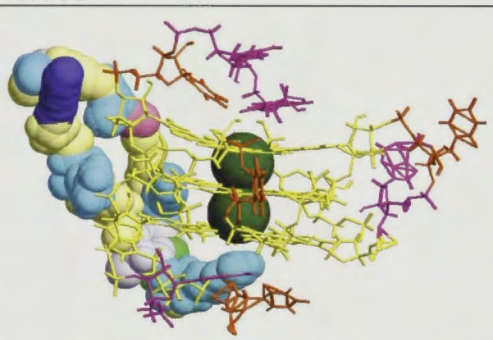
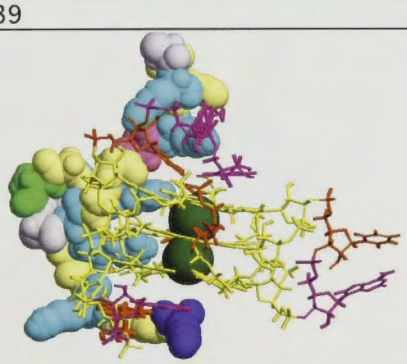
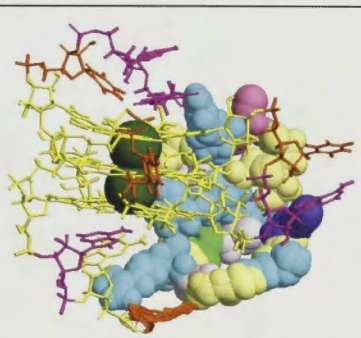


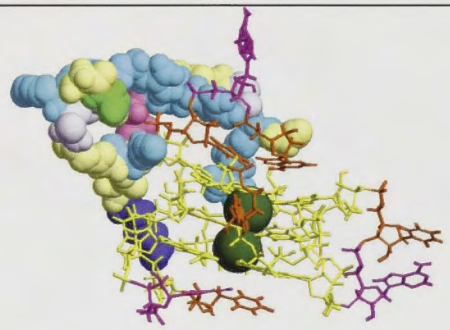
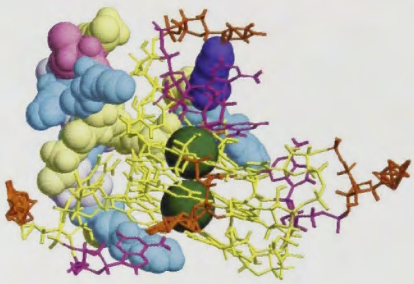
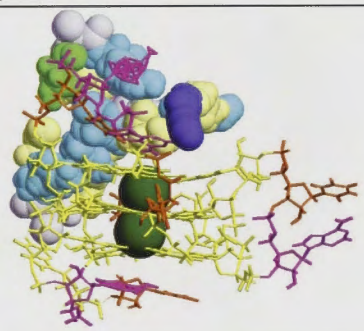
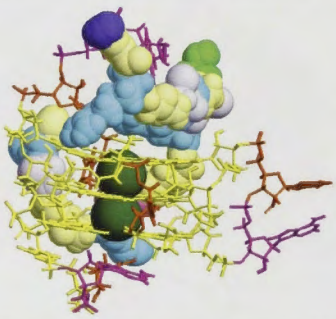
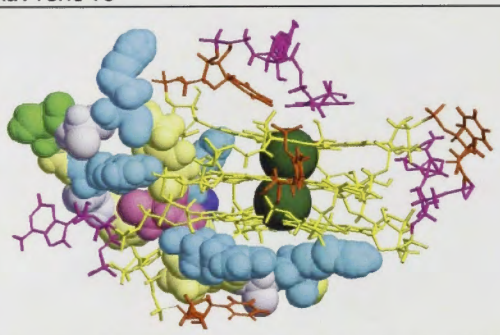
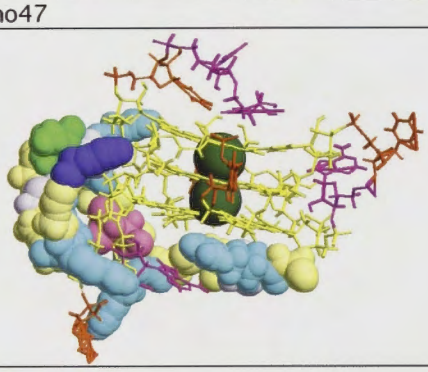
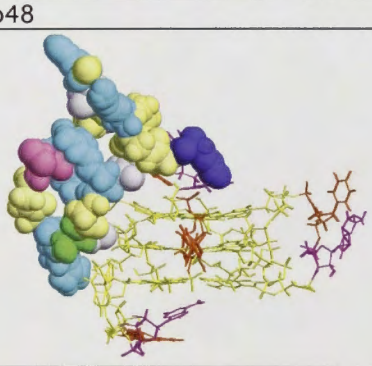
Xav1sho31

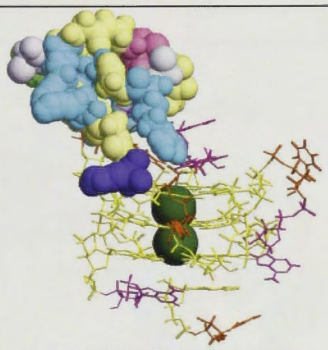
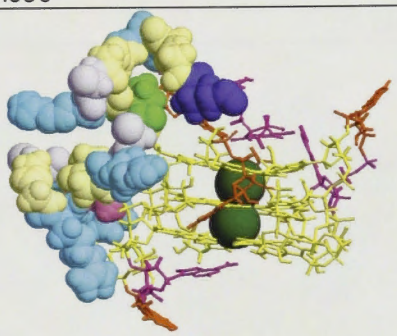
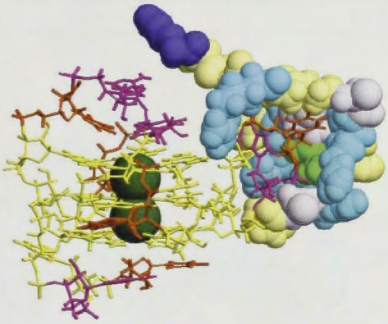
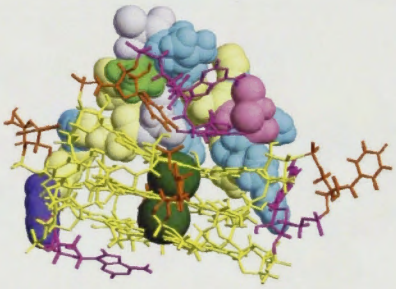
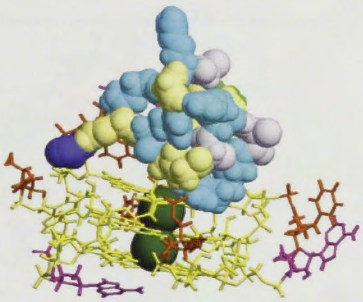


Xav1sho32

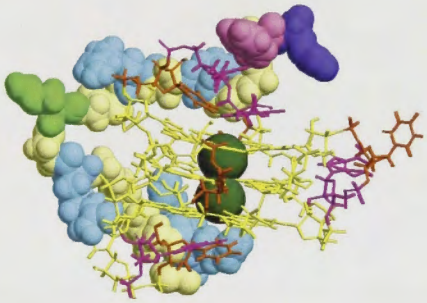
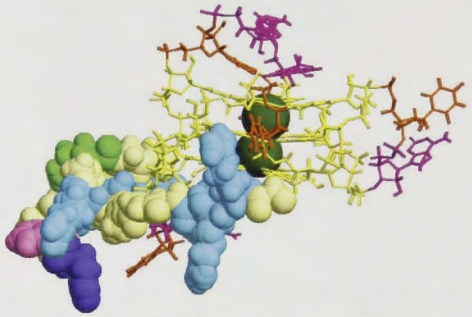
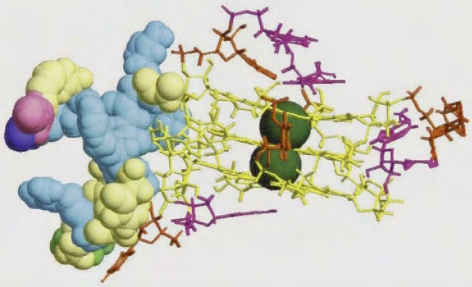
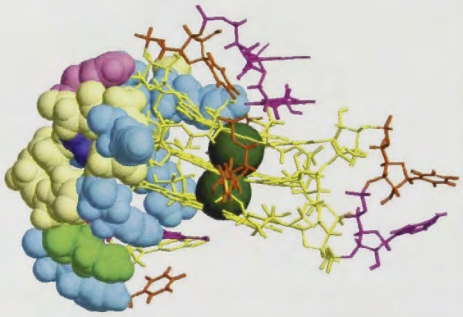
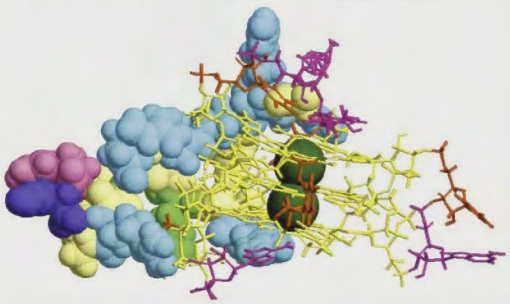
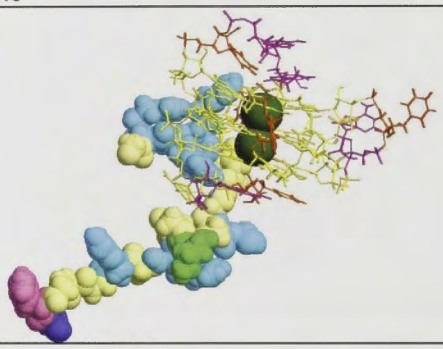
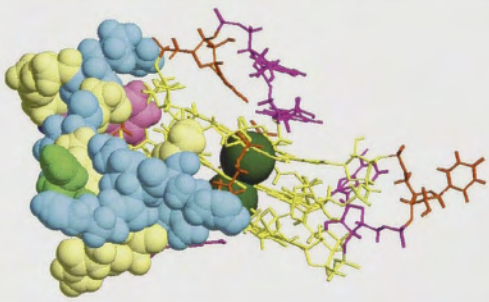
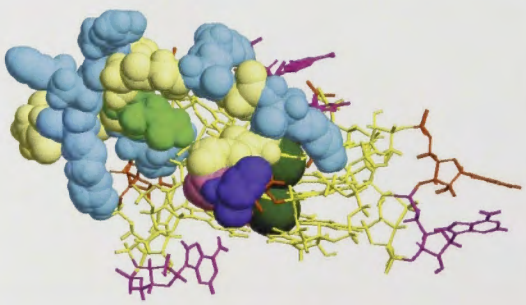


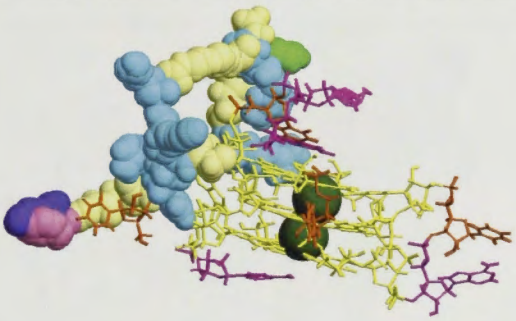
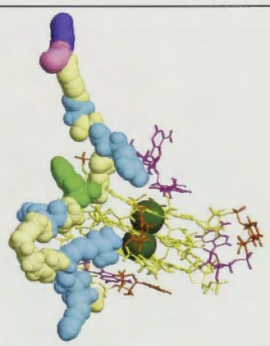
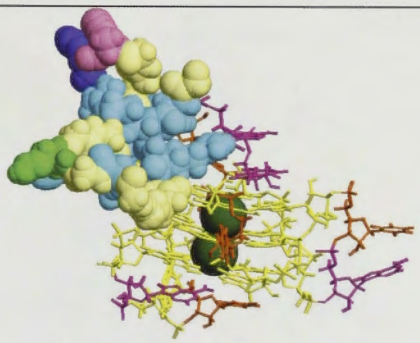
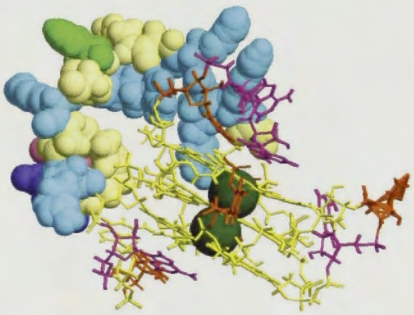
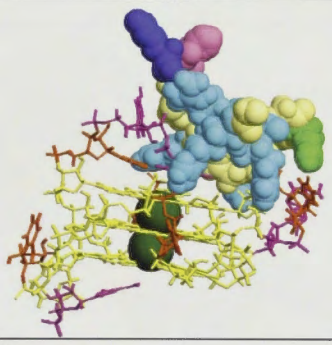
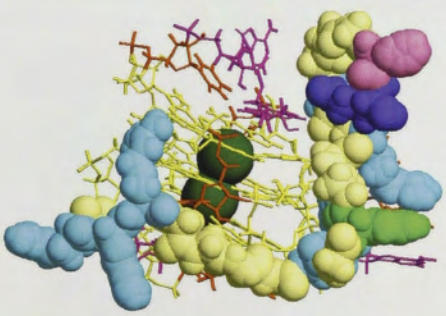
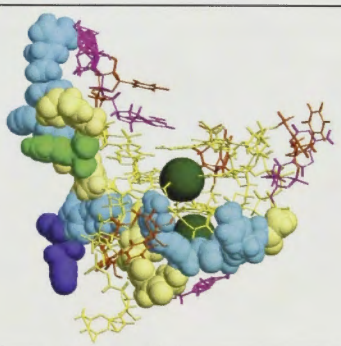
Xav1sho33	Xav1sho34
	
Xav1sho35	Xav1sho36
	
Xav1sho37	Xav1sho38
	
Xav1sho39	Xav1sho40
	

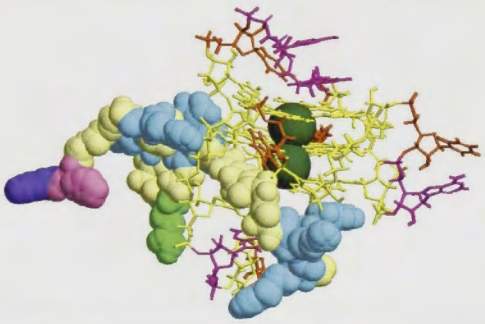
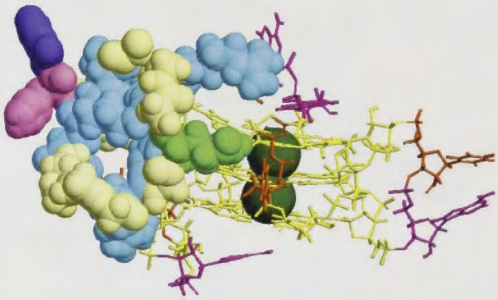
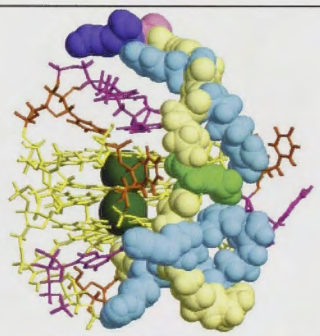
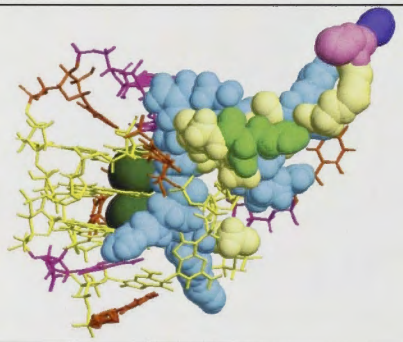
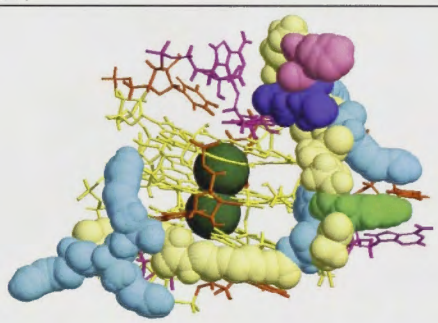
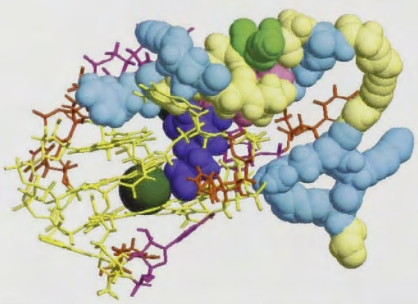
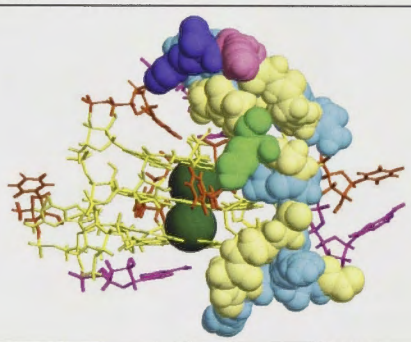
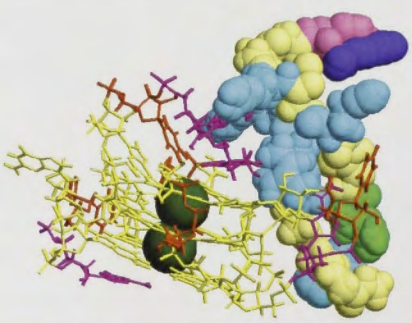
Xav1sho41	Xav1sho42
	
Xav1sho43	Xav1sho44
	
Xav1sho45	Xav1sho46
	
Xav1sho47	Xav1sho48
	

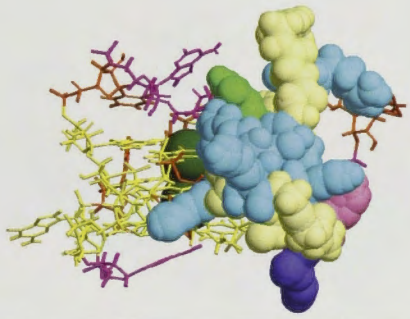
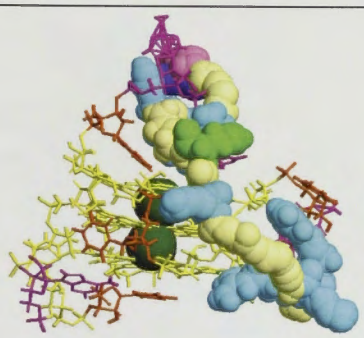
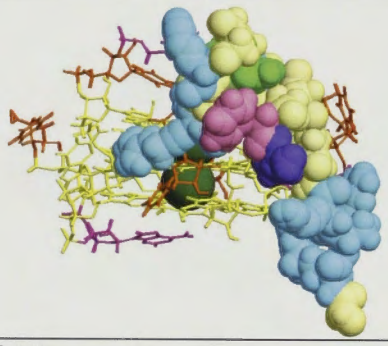
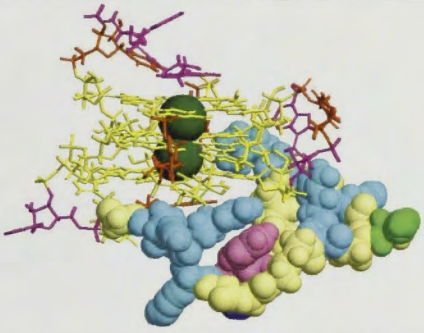
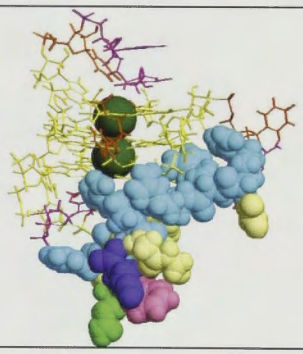
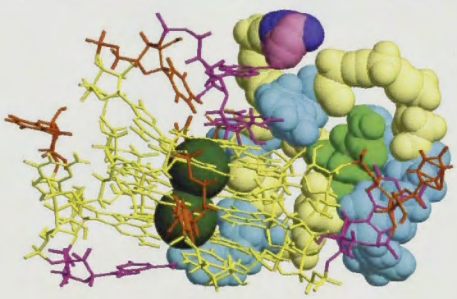
Xav1sho49	Xav1sho50
	
Xav1sho51	Xav1sho52
	
Xav1sho53	
	

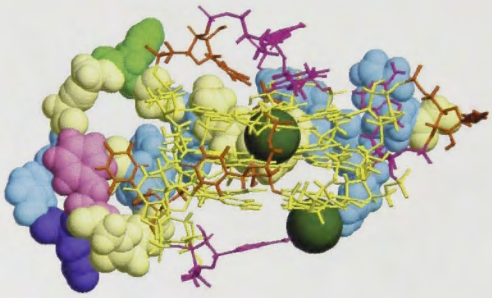
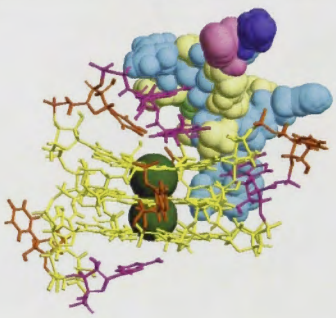
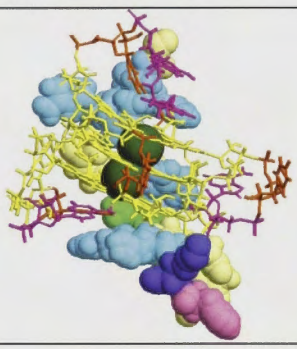
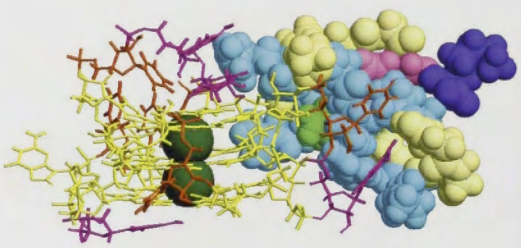
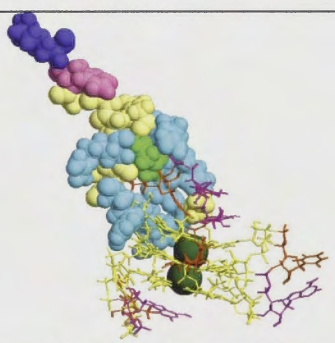
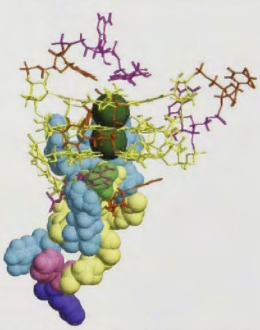
ANNEXURE 6B Fx-RNA Complexes

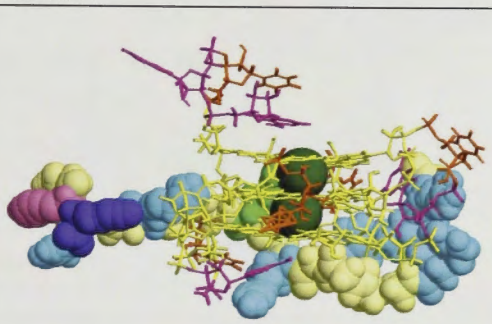
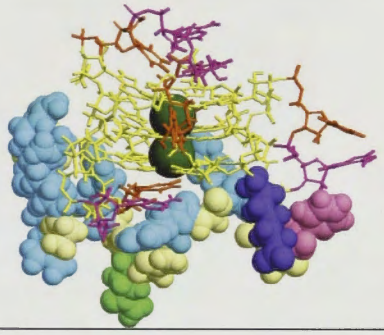
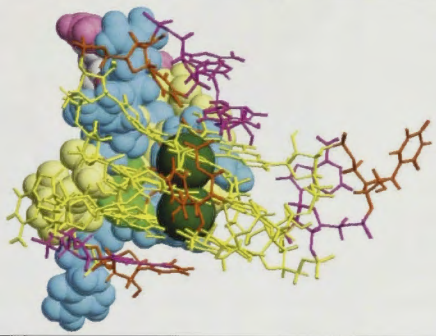
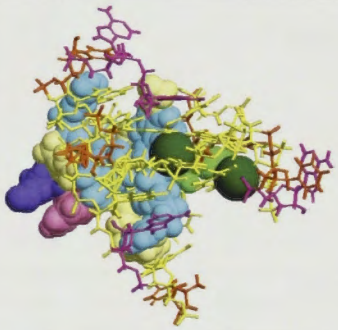
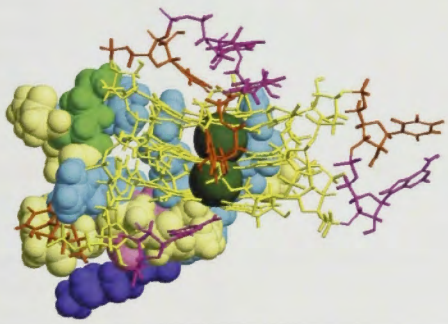
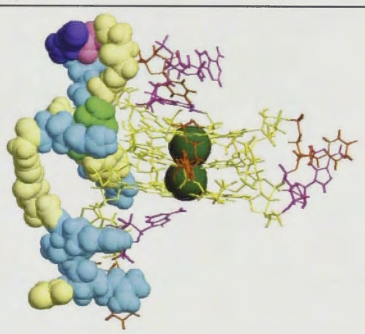
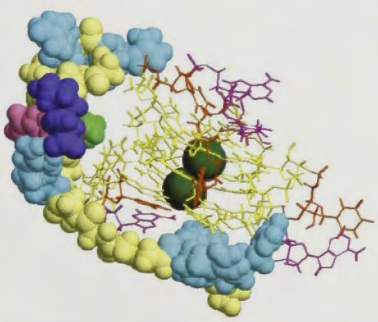
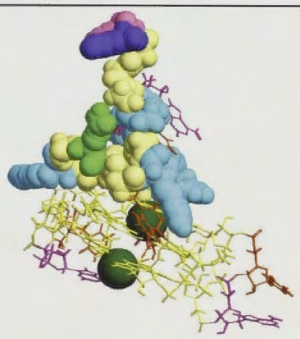
Xav1fx1	Xav1fx2
	
Xav1fx3	Xav1fx4
	
Xav1fx5	Xav1fx6
	
Xav1fx7	Xav1fx8
	

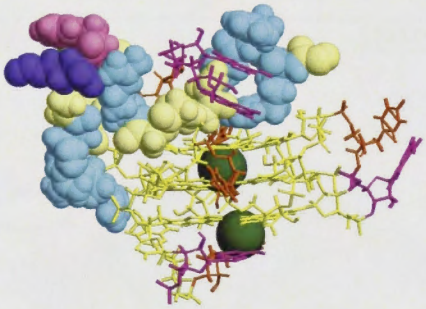
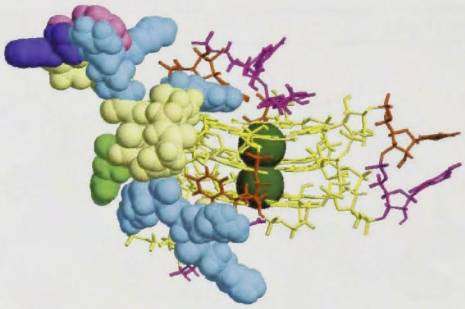
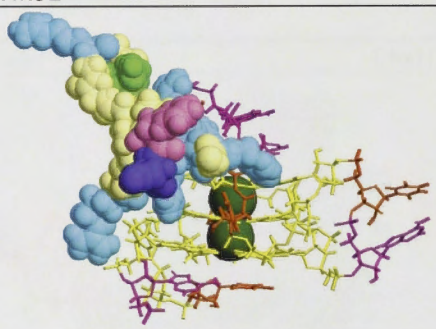
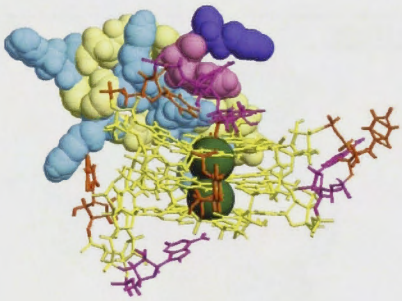
Xav1fx9	Xav1fx10
	
Xav1fx11	Xav1fx12
	
Xav1fx13	Xav1fx14
	
Xav1fx15	Xav1fx16
	

Xav1fx17	Xav1fx18
	
Xav1fx19	Xav1fx20
	
Xav1fx21	Xav1fx22
	
Xav1fx23	Xav1fx24
	

Xav1fx25	Xav1fx26
	
Xav1fx27	Xav1fx28
	
Xav1fx29	Xav1fx30
	
Xav1fx31	Xav1fx32
	

Xav1fx33	Xav1fx34
	
Xav1fx35	Xav1fx36
	
Xav1fx37	Xav1fx38
	
Xav1fx39	Xav1fx40
	

Xav1fx41	Xav1fx42
	
Xav1fx43	Xav1fx44
	
Xav1fx45	Xav1fx46
	
Xav1fx47	Xav1fx48
	
Xav1fx49	Xav1fx50

	
Xav1fx51	Xav1fx52
Run failed	
Xav1fx53	
	

APPENDIX 6C

Table A6-1. RMSD, K-K distance and estimated free energy 0f Sho-RNA complexes [a]

Sho	GQ	SD	Sho	SD	K-K	SD
xav1sho11	0.5	0.1	3.7	1.1	3.8	0.1
xav1sho16	1.6	0.4	2.8	0.6	3.9	0.2
xav1sho25	1.6	1.0	3.8	0.9	3.8	0.1
xav1sho51	1.6	1.0	3.8	0.9	3.8	0.1
xav1sho38	1.8	1.0	8.7	1.3	3.9	0.2
xav1sho30	1.9	1.0	4.5	1.2	3.8	0.2
xav1sho22	2.0	1.3	5.6	1.3	3.9	0.2
xav1sho24	2.0	1.3	5.6	1.3	3.9	0.2
xav1sho9	2.1	1.1	8.1	1.5	3.8	0.2
xav1sho6	2.2	1.3	9.2	2.2	3.8	0.2
xav1sho10	2.2	0.9	5.5	1.1	3.8	0.2
xav1sho1	2.2	0.7	5.1	1.0	3.8	0.2
xav1sho37	2.3	0.9	5.8	1.3	3.8	0.2
xav1sho3	2.3	0.9	4.2	0.8	3.8	0.1
xav1sho4	2.3	0.6	4.4	0.6	3.8	0.2
xav1sho17	2.3	1.0	7.0	1.7	3.8	0.2
xav1sho34	2.3	0.8	5.8	1.0	3.8	0.2
xav1sho29	2.4	0.7	6.5	0.7	3.8	0.2
xav1sho33	2.4	0.3	3.6	0.7	3.7	0.2
xav1sho2	2.5	0.7	5.6	1.0	3.8	0.1
xav1sho12	2.5	0.7	4.6	0.7	3.9	0.2
xav1sho21	2.5	0.7	7.3	0.9	3.9	0.2
xav1sho32	2.5	1.1	7.7	1.0	3.8	0.2
xav1sho43	2.6	0.7	6.3	1.6	3.8	0.2
xav1sho13	2.6	0.7	5.1	1.4	3.8	0.2
xav1sho36	2.7	0.8	6.9	0.8	3.8	0.2
xav1sho41	2.7	0.7	4.9	0.7	3.9	0.2
xav1sho19	2.8	1.6	5.1	1.3	3.8	0.1
xav1sho5	2.8	0.8	4.1	1.4	3.9	0.2
xav1sho27	2.8	0.4	3.6	0.6	24.8	9.4
xav1sho40	2.8	0.8	5.3	1.3	3.8	0.2
xav1sho39	2.8	0.3	5.9	0.8	3.8	0.2
xav1sho50	2.8	1.2	6.7	1.0	3.8	0.2
xav1sho44	2.8	1.1	5.6	1.6	3.9	0.2
xav1sho7	2.9	0.9	5.6	1.3	5.0	4.3
xav1sho28	3.0	0.8	7.0	1.4	3.8	0.2
xav1sho18	3.0	0.9	4.5	0.6	3.7	0.1
xav1sho47	3.0	0.8	8.8	1.0	3.8	0.2
xav1sho23	3.1	1.3	4.7	0.7	3.8	0.2
xav1sho15	3.2	0.8	7.1	1.8	3.8	0.2
xav1sho45	3.4	1.2	5.9	1.1	3.8	0.2
xav1sho31	3.4	1.0	5.5	1.0	3.8	0.2
xav1sho49	3.4	1.0	6.8	1.6	3.9	0.2
xav1sho14	3.5	1.1	6.6	1.1	3.8	0.2
xav1sho26	3.7	1.6	4.9	1.1	3.8	0.2
xav1sho42	3.8	0.9	3.7	0.9	4.7	0.5
xav1sho35	3.9	0.9	3.5	0.4	4.1	0.4
xav1sho46	4.1	0.5	6.3	0.5	3.8	0.2
xav1sho53	4.2	1.4	6.0	1.3	3.8	0.2
xav1sho8	4.2	0.5	6.2	2.4	3.8	0.2
xav1sho52	4.2	1.3	5.8	1.2	3.8	0.2

See footnotes to Table 6-1 (Chapter 6)

Fx	RNA	SD	Fx	SD	K-K	SD
xav1fx6	1.8	0.2	5.2	2.3	3.8	0.2
xav1fx1	2.1	0.2	7.1	1.2	3.8	0.2
xav1fx40	2.3	0.5	6.8	1.0	3.8	0.2
xav1fx10	2.5	0.2	6.5	1.3	3.8	0.2
xav1fx30	2.5	0.2	7.6	1.3	3.9	0.2
xav1fx31	2.6	0.2	6.9	0.9	3.9	0.2
xav1fx39	2.6	0.3	6.7	2.6	3.8	0.2
xav1fx42	2.6	0.2	9.9	1.3	3.8	0.2
xav1fx49	2.7	0.2	5.8	0.8	6.5	0.5
xav1fx5	2.7	0.3	5.5	0.6	3.8	0.2
xav1fx12	2.8	0.2	6.3	0.8	3.9	0.2
xav1fx47	2.8	0.2	4.9	0.8	3.8	0.2
xav1fx19	2.8	0.2	7.1	1.4	3.8	0.2
xav1fx52	2.8	0.2	6.5	0.9	3.8	0.2
xav1fx43	2.9	0.3	4.1	0.4	3.8	0.2
xav1fx15	2.9	0.3	6.4	0.8	3.8	0.2
xav1fx11	2.9	0.3	7.7	1.9	3.8	0.2
xav1fx21	2.9	0.3	6.4	0.8	3.8	0.2
xav1fx3	2.9	0.3	5.4	0.8	3.8	0.2
xav1fx46	3.0	0.2	5.1	0.6	3.8	0.2
xav1fx22	3.0	0.3	5.9	0.5	5.5	2.1
xav1fx37	3.1	0.3	8.0	1.0	3.8	0.2
xav1fx2	3.1	0.5	5.5	0.7	3.8	0.2
xav1fx29	3.1	0.4	7.5	1.7	3.9	0.2
xav1fx4	3.2	0.3	7.3	1.1	3.8	0.2
xav1fx50	3.2	0.5	8.5	1.1	3.8	0.2
xav1fx8	3.2	0.6	7.0	1.5	3.9	0.2
xav1fx34	3.2	0.3	5.9	0.6	12.9	10.3
xav1fx20	3.3	0.3	8.1	2.2	3.9	0.2
xav1fx7	3.3	0.3	6.8	0.9	3.9	0.2
xav1fx26	3.3	0.3	9.8	2.4	3.8	0.2
xav1fx45	3.3	0.3	8.7	1.7	3.8	0.2
xav1fx18	3.4	0.4	6.2	1.0	3.8	0.2
xav1fx17	3.4	0.4	7.3	1.2	4.0	0.2
xav1fx53	3.5	0.2	9.9	2.4	3.8	0.1
xav1fx38	3.5	0.3	6.2	0.8	3.8	0.2
xav1fx14	3.6	0.3	5.7	1.4	3.8	0.2
xav1fx13	3.6	0.3	7.1	1.9	3.8	0.2
xav1fx36	3.7	0.4	7.0	1.0	3.8	0.2
xav1fx24	3.7	0.3	4.5	0.8	3.8	0.2
xav1fx25	3.8	0.5	4.4	0.8	3.7	0.2
xav1fx23	3.8	0.5	6.5	0.7	3.8	0.2
xav1fx41	3.8	0.6	6.0	1.3	3.8	0.2
xav1fx35	3.9	0.5	7.2	1.0	6.7	7.9
xav1fx33	3.9	0.4	5.7	1.1	22.6	6.8
xav1fx32	4.0	0.3	7.7	1.5	3.8	0.2
xav1fx9	4.2	0.4	9.0	1.2	3.8	0.2
xav1fx44	4.3	0.4	10.2	1.2	25.9	9.9
xav1fx48	4.5	0.5	8.7	1.2	5.0	1.3
xav1fx16	4.5	0.4	9.5	1.1	23.3	9.8
xav1fx28	4.7	0.7	6.3	1.1	3.9	0.2

See footnotes to Table 6-1 (Chapter 6)

APPENDIX 7

Table A7-1 Enthalpy calculations of the Sho-RNA complexes ordered by ΔH^{I}_{Bind} [a,b]

	H_{comp} [c]	SD	H^{S}_{RNA} [d]	SD	H^{S}_{Sho} [e]	SD	ΔH_{RNA} [f]	ΔH_{Sho} [g]	ΔH^{S}_{Bind} [h]	SD	ΔH^{I}_{Bind} [i]
xav1sho49	-5632	33	-4538	25	-663	15	-16	-7	-432	10	-454
xav1sho18	-5622	25	-4525	17	-656	13	-3	0	-441	8	-444
xav1sho47	-5618	30	-4523	21	-638	11	-1	18	-457	10	-440
xav1sho43	-5618	24	-4533	13	-649	13	-11	7	-435	11	-440
xav1sho36	-5617	28	-4533	24	-635	13	-11	21	-449	16	-439
xav1sho46	-5613	25	-4505	23	-642	10	17	14	-467	10	-435
xav1sho3	-5613	23	-4528	19	-640	14	-6	16	-446	13	-435
xav1sho41	-5610	26	-4523	20	-636	12	-1	20	-451	10	-432
xav1sho1	-5609	26	-4526	22	-644	13	-4	12	-439	14	-431
xav1sho35	-5607	34	-4507	25	-641	18	15	15	-459	9	-429
xav1sho39	-5605	24	-4527	18	-648	13	-5	8	-431	8	-427
xav1sho27	-5605	96	-4499	94	-633	15	23	23	-472	12	-427
xav1sho16	-5604	20	-4536	18	-629	11	-14	27	-439	11	-426
xav1sho13	-5602	21	-4529	21	-659	11	-7	-3	-414	7	-424
xav1sho32	-5601	30	-4523	15	-650	13	-1	6	-428	17	-423
xav1sho31	-5600	27	-4515	20	-639	12	7	17	-446	10	-422
xav1sho51	-5600	30	-4529	23	-636	14	-7	20	-435	13	-422
xav1sho28	-5598	26	-4520	19	-640	13	2	16	-437	12	-420
xav1sho12	-5597	19	-4514	19	-654	11	8	2	-430	5	-419
xav1sho14	-5594	26	-4532	21	-633	14	-10	23	-429	11	-416
xav1sho10	-5593	28	-4528	20	-633	12	-6	23	-433	13	-415
xav1sho7	-5592	30	-4539	26	-647	10	-17	9	-407	10	-414
xav1sho2	-5590	24	-4520	21	-647	13	2	9	-423	16	-412
xav1sho37	-5590	27	-4531	16	-648	14	-9	8	-411	11	-412
xav1sho53	-5590	34	-4526	25	-655	16	-4	1	-409	13	-412
xav1sho9	-5589	38	-4523	30	-642	14	-1	14	-423	10	-411
xav1sho45	-5588	26	-4515	22	-642	15	7	14	-431	10	-410
xav1sho4	-5588	24	-4510	21	-629	11	12	27	-449	11	-410
xav1sho44	-5588	31	-4518	32	-640	10	4	16	-429	10	-410
xav1sho42	-5587	25	-4507	21	-642	9	15	14	-439	11	-409
xav1sho40	-5586	25	-4518	20	-654	12	4	2	-414	15	-408
xav1sho26	-5585	29	-4530	23	-651	16	-8	5	-405	12	-407
xav1sho15	-5585	34	-4539	23	-623	15	-17	33	-422	15	-407
xav1sho5	-5584	26	-4513	17	-656	15	9	0	-416	9	-406
xav1sho50	-5584	29	-4517	25	-635	13	5	21	-431	14	-406
xav1sho30	-5583	35	-4515	32	-638	11	7	18	-430	11	-405
xav1sho17	-5580	29	-4511	20	-645	12	11	11	-423	18	-402
xav1sho21	-5577	39	-4506	28	-640	13	16	16	-431	18	-399
xav1sho23	-5577	28	-4474	22	-631	13	48	25	-472	13	-399
xav1sho25	-5576	37	-4510	27	-655	11	12	1	-412	16	-398
xav1sho34	-5576	45	-4510	38	-635	12	12	21	-430	12	-398
xav1sho38	-5574	32	-4509	27	-639	11	13	17	-426	16	-396
xav1sho8	-5574	29	-4515	23	-653	13	7	3	-406	15	-396
xav1sho52	-5574	25	-4513	23	-640	13	9	16	-421	13	-396
xav1sho33	-5571	23	-4506	19	-624	11	16	32	-441	12	-393
xav1sho29	-5571	27	-4512	15	-643	15	10	13	-415	10	-393
xav1sho22	-5570	29	-4493	23	-648	15	29	8	-429	13	-392
xav1sho19	-5566	31	-4490	23	-643	13	32	13	-433	15	-388
xav1sho6	-5566	40	-4501	28	-642	18	21	14	-423	12	-388
AVG	**-5591**		**-4517**		**-643**		**5**	**13**	**-431**		**-413**
SD	**18**		**14**		**9**		**14**	**9**	**17**		**18**

[a] All energies (kcal mol^{-1}) [b] Enthalpy values are averages derived from snapshots taken every 500 ps between 5 - 25ns of the simulation. SD = 1 standard deviation. [c] H_{comp} = complex enthalpy. [d] H^{S}_{RNA} = RNA enthalpy (single trajectory method). [e] H^{S}_{Sho} = peptide enthalpy (single trajectory method). [f] $\Delta H_{RNA} = H^{S}_{RNA} - H^{I}_{RNa}$ (H^{I}_{RNA} = -4522 kcal mol^{-1}), a measure of the RNA stabilization upon complex formation. [g] $\Delta H_{Sho} = H^{S}_{Sho} - H^{I}_{Sho}$ (H^{I}_{Sho} = -656 kcal/mol), a measure of the peptide stabilization upon complex formation. [h] ΔH^{S}_{Bind} = the enthalpy difference upon complex formation using the single trajectory approach. [i] ΔH^{I}_{Bind} = tenthalpy difference upon complex formation (independent trajectory method).

Table A7-2 Entropy calculations of Sho-RNA complexes ordered by ΔG [a,b,c]

	TS_{Comp} [c]	SD	TS_{RNA} [d]	SD	TS_{Sho} [e]	SD	$T\Delta S$ [f]	SD	ΔH^{l}_{Bind} [g]	ΔG [h]
xav1sho49	806	5	608	3	228	5	-31	7	-454	-423
xav1sho18	804	5	609	1	227	5	-32	7	-444	-412
xav1sho3	802	4	607	2	228	4	-32	4	-435	-403
xav1sho36	802	5	611	2	229	7	-37	5	-439	-401
xav1sho47	802	3	613	2	231	5	-42	5	-440	-399
xav1sho41	800	4	611	2	227	3	-38	2	-432	-395
xav1sho46	800	3	609	1	233	4	-42	2	-435	-393
xav1sho43	794	3	610	2	231	4	-48	3	-440	-392
Xav1sho27	797	4	603	4	229	7	-35	9	-427	-391
xav1sho35	802	3	612	4	228	5	-38	7	-429	-391
xav1sho13	803	4	608	3	228	2	-33	5	-424	-391
xav1sho1	796	5	605	3	233	2	-42	7	-431	-389
xav1sho16	805	2	612	3	230	4	-37	3	-426	-389
xav1sho12	801	2	608	2	226	5	-33	6	-419	-387
xav1sho39	799	2	607	1	233	10	-42	7	-427	-386
xav1sho32	801	3	610	2	228	4	-37	3	-423	-385
xav1sho51	809	4	612	1	234	4	-37	5	-422	-385
xav1sho28	802	5	610	4	228	3	-36	9	-420	-384
xav1sho45	809	4	612	5	226	3	-29	2	-410	-381
xav1sho31	803	3	614	3	230	3	-41	4	-422	-381
xav1sho5	806	3	609	2	224	2	-28	2	-406	-379
xav1sho10	799	3	610	1	226	3	-36	4	-415	-378
xav1sho14	800	6	608	2	230	4	-38	8	-416	-378
xav1sho7	802	4	611	2	228	3	-36	4	-414	-378
xav1sho2	801	5	607	2	228	3	-35	6	-412	-377
xav1sho40	800	4	605	5	227	1	-32	9	-408	-376
xav1sho15	801	3	605	5	228	3	-32	4	-407	-375
xav1sho53	797	5	607	2	228	2	-38	6	-412	-374
xav1sho4	800	1	608	3	229	4	-37	6	-410	-373
xav1sho37	797	3	609	2	228	3	-40	3	-412	-372
xav1sho50	798	6	605	2	228	3	-35	8	-406	-371
xav1sho42	796	3	608	4	228	5	-40	5	-409	-369
xav1sho26	797	3	608	2	229	4	-40	7	-407	-368
xav1sho30	803	4	609	3	233	5	-38	5	-405	-366
xav1sho44	803	5	608	2	240	4	-45	7	-410	-365
xav1sho21	804	4	611	4	228	6	-36	6	-399	-363
xav1sho8	804	4	608	2	229	2	-33	5	-396	-363
xav1sho9	796	6	609	3	236	5	-49	7	-411	-362
xav1sho17	799	3	611	2	227	5	-40	6	-402	-362
xav1sho25	799	3	607	2	229	3	-37	4	-398	-361
xav1sho34	803	5	610	2	231	5	-38	6	-398	-359
xav1sho22	801	4	609	2	227	3	-35	3	-392	-357
xav1sho38	801	6	609	2	232	6	-40	10	-396	-357
xav1sho23	794	3	607	3	230	5	-44	6	-399	-355
xav1sho52	797	5	608	3	233	4	-44	6	-396	-352
xav1sho33	803	2	611	3	234	8	-42	8	-393	-351
xav1sho29	798	6	609	4	233	3	-44	7	-393	-349
xav1sho19	803	4	612	3	232	4	-40	7	-388	-348
xav1sho6	804	4	609	4	236	8	-41	11	-388	-347
AVG	**801**		**609**		**230**		**-38**		**-413**	**-376**
SD	**3**		**2**		**3**		**5**		**18**	**17**

[a] All energies (kcal mol^{-1}) [b] Entropy values are averages from snapshots every 2ns over the last 12 ns of the simulation. SD = 1 standard deviation. [c] Entropy values calculated at 300 K. [d] TS_{comp} = complex entropy. [e] TS_{RNA} = RNA entropy. [f] TS_{Sho} = peptide entropy. [g] $T\Delta S$ = change in entropy upon complex formation. [h] ΔH^{l}_{Bind} = change in enthalpy upon complex formation. [i] ΔG = change in Gibbs free energy upon complex formation, calculated as $\Delta H^{l}_{Bind} - T\Delta S$.

	H_{Comp} [c]	SD	H^{S}_{RNA} [d]	SD	H^{S}_{Fx} [e]	SD	ΔH_{RNA} [f]	ΔH_{Fx} [g]	ΔH^{S}_{Bind} [h]	SD	ΔH^{I}_{Bind} [i]
xav1fx30	-5832	33	-4537	25	-834	12	-15	26	-462	21	-450
Xav1fx37	-5832	36	-4532	25	-813	12	-10	47	-487	22	-450
xav1fx28	-5829	23	-4546	18	-817	14	-24	43	-466	12	-447
xav1fx5	-5817	25	-4524	20	-829	16	-2	31	-463	11	-435
xav1fx1	-5815	26	-4546	21	-809	15	-24	51	-460	12	-433
xav1fx16	-5813	25	-4488	22	-810	16	34	50	-515	11	-431
Xav1fx31	-5808	33	-4518	24	-811	15	4	49	-479	12	-426
Xav1fx34	-5807	30	-4515	22	-813	11	7	47	-480	14	-425
xav1fx40	-5807	22	-4526	19	-810	16	-4	50	-471	9	-425
xav1fx47	-5807	23	-4546	21	-817	13	-24	43	-444	13	-425
xav1fx11	-5806	32	-4551	26	-833	13	-29	27	-423	10	-424
xav1fx2	-5805	23	-4524	19	-829	12	-2	31	-452	9	-423
xav1fx52	-5805	32	-4540	22	-826	17	-18	34	-438	12	-423
xav1fx41	-5804	38	-4505	24	-817	13	17	43	-482	25	-422
xav1fx45	-5803	32	-4524	22	-831	16	-2	29	-449	10	-421
Xav1fx23	-5802	77	-4532	73	-808	15	-10	52	-461	17	-420
xav1fx7	-5802	26	-4523	22	-822	10	-1	38	-458	9	-420
Xav1fx44	-5801	21	-4492	20	-829	12	30	31	-480	11	-419
xav1fx42	-5801	23	-4541	21	-812	12	-19	48	-448	14	-419
xav1fx46	-5799	32	-4509	21	-818	13	13	42	-471	13	-417
xav1fx26	-5798	34	-4518	22	-823	14	4	37	-457	14	-416
xav1fx35	-5798	28	-4505	21	-818	13	17	42	-474	12	-416
xav1fx14	-5797	27	-4531	23	-838	12	-9	22	-428	12	-415
xav1fx12	-5796	30	-4523	16	-832	12	-1	28	-440	16	-414
xav1fx43	-5794	25	-4517	22	-836	17	5	24	-441	11	-412
xav1fx49	-5792	29	-4511	16	-820	14	11	40	-461	12	-410
xav1fx13	-5791	26	-4534	23	-808	14	-12	52	-448	14	-409
xav1fx38	-5790	22	-4508	20	-828	14	14	32	-454	9	-408
xav1fx27	-5789	29	-4505	18	-810	14	17	50	-473	15	-407
xav1fx9	-5789	24	-4516	18	-813	15	6	47	-460	17	-407
xav1fx18	-5789	33	-4525	19	-834	17	-3	26	-430	15	-407
xav1fx29	-5788	24	-4525	21	-825	10	-3	35	-438	10	-406
xav1fx32	-5786	33	-4515	25	-816	14	7	44	-455	14	-404
xav1fx10	-5785	26	-4518	21	-810	12	4	50	-458	10	-403
xav1fx50	-5785	23	-4536	19	-819	13	-14	41	-430	9	-403
xav1fx53	-5782	25	-4537	19	-835	13	-15	25	-409	16	-400
xav1fx21	-5781	25	-4528	22	-812	13	-6	48	-442	15	-399
xav1fx48	-5781	30	-4500	22	-810	12	22	50	-471	8	-399
xav1fx39	-5781	28	-4522	19	-820	16	0	40	-439	20	-399
xav1fx15	-5780	26	-4529	22	-811	13	-7	49	-441	15	-398
xav1fx3	-5780	23	-4524	16	-819	13	-2	41	-438	11	-398
xav1fx24	-5780	24	-4516	21	-834	14	6	26	-430	9	-398
xav1fx36	-5778	25	-4530	16	-829	12	-8	31	-419	15	-396
xav1fx22	-5777	24	-4500	17	-822	9	22	38	-455	17	-395
xav1fx19	-5776	26	-4515	22	-815	12	7	45	-446	14	-394
xav1fx17	-5773	27	-4503	19	-815	12	19	45	-455	14	-391
xav1fx25	-5769	29	-4517	23	-836	14	5	24	-416	13	-387
xav1fx6	-5769	26	-4516	17	-829	17	6	31	-424	17	-387
xav1fx8	-5768	20	-4531	17	-819	11	-9	41	-418	14	-386
xav1fx33	-5753	25	-4486	18	-817	12	36	43	-450	14	-371
xav1fx20	-5751	31	-4499	21	-816	14	23	44	-435	10	-369
AVG	**-5793**		**-4521**		**-820**		**1**	**40**	**-452**		**-412**
SD	**17**		**15**		**9**		**15**	**9**	**21**		**16**

[a] See notes to Table A7-1.

Table A7-4 Entropy calculations of the Fx-RNA complexes [ab]

	TS_{Comp} [c]	SD	TS_{RNA} [d]	SD	TS_{Fx} [e]	SD	$T\Delta S$ [f]	SD	ΔH_{Bind} [g]	ΔG [h]
xav1fx30	814	3	610	2	246	2	-41	5	-450	-409
xav1fx28	810	2	608	2	243	4	-41	5	-447	-406
xav1fx37	810	3	607	3	243	5	-41	6	-445	-404
xav1fx5	814	1	610	2	246	4	-41	7	-435	-394
xav1fx1	819	5	611	1	249	6	-41	11	-433	-392
xav1fx2	821	5	613	1	242	8	-33	14	-423	-390
xav1fx44	816	4	607	2	239	3	-30	4	-419	-389
xav1fx16	819	4	611	4	252	10	-44	12	-431	-387
xav1fx11	816	5	609	2	245	2	-38	5	-424	-387
xav1fx43	826	1	610	3	244	4	-29	5	-412	-383
xav1fx34	812	4	607	4	247	8	-42	8	-425	-383
xav1fx7	814	2	608	2	244	7	-38	8	-420	-382
xav1fx45	813	3	612	1	241	3	-40	4	-421	-381
xav1fx52	815	5	611	1	246	5	-41	11	-423	-381
xav1fx40	816	2	609	2	251	4	-44	5	-425	-381
xav1fx4	807	3	607	2	243	4	-43	6	-423	-379
xav1fx26	816	7	611	3	242	4	-37	9	-416	-379
xav1fx42	814	6	609	3	245	3	-40	6	-419	-379
xav1fx31	814	6	610	1	253	8	-49	8	-426	-377
xav1fx23	809	4	606	2	247	7	-44	9	-420	-376
xav1fx14	814	8	611	2	243	4	-40	6	-415	-375
xav1fx13	815	4	606	1	243	3	-35	5	-409	-374
xav1fx38	813	6	606	3	242	4	-35	9	-408	-373
xav1fx12	814	3	610	2	245	3	-41	4	-414	-372
xav1fx41	816	5	613	1	254	5	-51	6	-422	-371
xav1fx18	810	4	607	3	239	3	-36	3	-407	-371
xav1fx35	813	5	611	5	247	2	-45	7	-416	-371
xav1fx46	819	5	613	2	246	2	-40	9	-410	-370
xav1fx39	817	9	608	1	239	6	-30	8	-399	-369
xav1fx10	822	6	608	2	249	6	-34	10	-403	-369
xav1fx47	812	5	609	2	260	2	-56	6	-425	-368
xav1fx32	813	4	609	2	243	2	-38	6	-404	-366
xav1fx3	822	4	613	2	242	5	-33	5	-398	-365
xav1fx53	818	4	611	2	241	2	-35	5	-400	-364
xav1fx9	817	4	614	1	246	11	-43	14	-407	-364
xav1fx27	807	3	604	3	247	5	-45	8	-407	-362
xav1fx49	812	1	608	6	253	1	-48	8	-410	-362
xav1fx29	812	3	610	2	249	4	-46	7	-406	-360
xav1fx24	813	5	607	5	245	3	-40	7	-398	-358
xav1fx50	811	6	608	1	250	1	-47	5	-403	-356
xav1fx25	821	6	609	2	244	2	-32	7	-387	-355
xav1fx21	813	3	605	2	253	7	-45	7	-399	-354
xav1fx6	820	9	608	1	246	5	-34	12	-387	-353
xav1fx36	810	5	609	2	245	2	-44	5	-396	-352
xav1fx15	813	3	605	2	255	8	-47	6	-398	-352
xav1fx48	807	7	609	4	246	3	-48	7	-398	-350
xav1fx22	813	4	609	3	250	8	-46	9	-395	-349
xav1fx17	812	6	607	2	251	7	-46	9	-391	-345
xav1fx19	811	7	609	3	255	10	-54	7	-394	-340
xav1fx8	810	6	611	3	248	9	-48	15	-386	-338
xav1fx20	816	9	609	3	245	5	-39	8	-369	-330
xav1fx33	813	4	613	4	247	7	-47	9	-371	-324
AVG	**814**		**609**		**246**		**-41**		**-412**	**-369**
SD	**4**		**2**		**5**		**6**		**16**	**18**

[a] See notes to Table A7-2.

Table A7-5 RNA stabilization or destabilization in the complexes [a,b]

Complex	ΔH_{RNA} [c]	ΔH^{I}_{Bind} [d]	ΔH^{S}_{Bind} [e]	Complex	ΔH_{RNA}	ΔH^{I}_{Bind}	ΔH^{S}_{Bind}
xav1sho49	-16	-454	-432	xav1fx30	-15	-450	-462
xav1sho18	-3	-444	-441	xav1fx28	-24	-447	-466
xav1sho47	-1	-440	-457	xav1fx37	-10	-445	-491
xav1sho43	-11	-440	-435	xav1fx5	-2	-435	-463
xav1sho36	-11	-439	-449	xav1fx1	-24	-433	-460
xav1sho46	17	-435	-467	xav1fx16	34	-431	-515
xav1sho3	-6	-435	-446	xav1fx31	4	-426	-479
xav1sho41	-1	-432	-451	xav1fx34	7	-425	-480
xav1sho1	-4	-431	-439	xav1fx40	-4	-425	-471
xav1sho35	15	-429	-459	xav1fx47	-24	-425	-444
xav1sho39	-5	-427	-431	xav1fx11	**-29**	-424	-423
xav1sho27	23	-427	-472	xav1fx2	-2	-423	-452
xav1sho16	-14	-426	-439	xav1fx4	3	-423	-475
xav1sho13	-7	-424	-414	xav1fx52	-18	-423	-438
xav1sho32	-1	-423	-428	xav1fx41	17	-422	-482
xav1sho31	7	-422	-446	xav1fx45	-2	-421	-449
xav1sho51	-7	-422	-434	xav1fx23	-11	-420	-461
xav1sho28	2	-420	-437	xav1fx7	-1	-420	-458
xav1sho12	8	-419	-430	xav1fx44	30	-419	-480
xav1sho14	-10	-416	-429	xav1fx42	-19	-419	-448
xav1sho10	-6	-415	-433	xav1fx26	4	-416	-457
xav1sho7	-17	-414	-407	xav1fx35	17	-416	-474
xav1sho2	2	-412	-423	xav1fx14	-9	-415	-428
xav1sho37	-9	-412	-411	xav1fx12	-1	-414	-440
xav1sho53	-4	-412	-409	xav1fx43	5	-412	-441
xav1sho9	-1	-411	-423	xav1fx49	11	-410	-461
xav1sho45	7	-410	-431	xav1fx46	17	-410	-468
xav1sho4	12	-410	-449	xav1fx13	-12	-409	-449
xav1sho44	4	-410	-429	xav1fx38	14	-408	-454
xav1sho42	15	-409	-439	xav1fx27	17	-407	-473
xav1sho40	4	-408	-414	xav1fx9	6	-407	-460
xav1sho26	-8	-407	-405	xav1fx18	-3	-407	-430
xav1sho15	-17	-407	-422	xav1fx29	-3	-406	-438
xav1sho5	9	-406	-416	xav1fx32	7	-404	-455
xav1sho50	5	-406	-431	xav1fx10	4	-403	-458
xav1sho30	7	-405	-430	xav1fx50	-14	-403	-430
xav1sho17	11	-402	-423	xav1fx53	-15	-400	-409
xav1sho21	16	-399	-431	xav1fx21	-6	-399	-442
xav1sho23	**48**	-399	-472	xav1fx39	0	-399	-439
xav1sho25	12	-398	-412	xav1fx15	-7	-398	-441
xav1sho34	12	-398	-430	xav1fx3	-2	-398	-438
xav1sho38	13	-39	-426	xav1fx24	6	-398	-430
xav1sho8	7	-396	-406	xav1fx48	**23**	-398	-471
xav1sho52	9	-396	-421	xav1fx36	-8	-396	-419
xav1sho33	16	-393	-441	xav1fx22	**22**	-395	-455
xav1sho29	10	-393	-415	xav1fx19	7	-394	-446
xav1sho24	29	-392	-429	xav1fx17	19	-391	-455
xav1sho19	32	-388	-433	xav1fx25	5	-387	-416
xav1sho6	21	-388	-423	xav1fx6	6	-387	-424
				xav1fx8	-9	-386	-418
				xav1fx33	**36**	-371	-450
				xav1fx20	**23**	-369	-435
AVG	**5**	**-413**	**-431**		**1**	**-412**	**-452**
SD	**14**	**18**	**17**		**15**	**16**	**21**

[a] All energies (kcal mol^{-1}). [b] Complexes in which the RNA is relatively stabilized in the complex are highlighted in yellow. [c] $\Delta H_{RNA} = H^{S}_{RNA} - H^{I}_{RNa}$ (where H^{I}_{RNA} = -4522 kcal mol^{-1}) [d] ΔH^{I}_{Bind} = enthalpy difference upon complex formation, the complexes are arranged from most to least negative value of this quantity. [e] ΔH^{S}_{Bind} is the difference in the interaction component of the enthalpy upon complex formatio

Table A7-6. Decomposition of interaction component of free energy in Sho-RNA [a b c d e f]

	VDW	SD	NE	SD	PbCal	SD	Ele	SD	PBELE	SD
xav1sho27	-88	7	-131	8	3058	63	-3400	64	-342	13
xav1sho23	-100	6	-143	7	3017	130	-3346	137	-329	13
xav1sho46	-118	7	-162	7	2555	90	-2860	92	-305	10
xav1sho35	-69	7	-111	8	2756	83	-3104	83	-348	10
xav1sho47	-99	10	-143	10	2554	131	-2868	130	-315	11
xav1sho41	-58	7	-100	8	2646	110	-2998	113	-352	11
xav1sho36	-53	9	-94	10	2662	135	-3017	137	-355	11
xav1sho4	-78	6	-118	6	2732	63	-3063	65	-331	11
xav1sho31	-63	11	-103	11	2628	95	-2972	94	-343	9
xav1sho3	-51	7	-91	7	2604	73	-2958	77	-354	14
xav1sho18	-62	5	-103	6	2489	56	-2826	52	-337	9
xav1sho33	-98	11	-141	11	2673	118	-2973	118	-300	18
xav1sho16	-55	7	-96	7	2642	77	-2985	78	-343	11
xav1sho42	-79	10	-121	10	2786	121	-3104	114	-318	15
xav1sho1	-61	7	-98	8	2570	78	-2911	85	-341	12
xav1sho28	-53	9	-91	9	2413	200	-2759	202	-346	9
xav1sho43	-59	7	-100	7	2797	78	-3132	83	-335	12
xav1sho51	-45	6	-86	6	2490	139	-2838	145	-349	10
xav1sho19	-64	13	-106	14	2509	133	-2837	137	-328	12
xav1sho10	-58	7	-99	8	2681	117	-3014	123	-333	16
xav1sho49	-39	7	-76	7	2562	110	-2918	111	-356	10
xav1sho50	-48	8	-87	9	2622	111	-2965	114	-344	13
xav1sho45	-39	6	-74	6	2414	83	-2771	87	-358	10
xav1sho21	-45	13	-81	13	2777	215	-3126	218	-350	11
xav1sho39	-44	7	-84	7	2557	118	-2903	117	-347	7
xav1sho34	-50	9	-89	9	2576	201	-2916	202	-341	10
xav1sho12	-61	3	-99	4	2310	52	-2640	52	-330	6
xav1sho30	-58	10	-96	10	2508	121	-2842	119	-334	13
xav1sho44	-87	7	-127	8	2411	113	-2713	112	-302	13
xav1sho22	-63	6	-100	7	2188	181	-2517	185	-329	14
xav1sho24	-63	6	-100	7	2188	181	-2517	185	-329	14
xav1sho14	-41	7	-78	8	2559	198	-2910	196	-351	12
xav1sho32	-45	9	-85	9	2518	154	-2861	153	-343	16
xav1sho38	-53	7	-93	7	2430	149	-2764	158	-334	19
xav1sho17	-40	7	-80	7	2608	108	-2951	114	-343	15
xav1sho6	-68	8	-109	9	2547	79	-2861	77	-314	10
xav1sho2	-57	10	-95	10	2582	125	-2909	130	-327	16
xav1sho15	-48	12	-84	12	2520	240	-2858	242	-338	12
xav1sho52	-53	8	-93	9	2687	146	-3015	150	-329	10
xav1sho5	-36	4	-72	5	2231	88	-2574	86	-343	10
xav1sho29	-46	8	-84	8	2470	117	-2801	120	-331	13
xav1sho40	-41	8	-80	9	2288	149	-2623	153	-334	15
xav1sho13	-35	4	-72	5	2410	76	-2753	78	-342	7
xav1sho25	-39	10	-77	10	2281	138	-2616	141	-335	13
xav1sho37	-51	6	-89	6	2496	113	-2818	117	-322	11
xav1sho53	-48	5	-84	6	2208	141	-2533	141	-325	14
xav1sho7	-36	10	-71	10	2442	108	-2778	111	-336	12
xav1sho8	-34	5	-71	6	2237	132	-2573	140	-335	13
xav1sho26	-45	7	-82	7	2549	126	-2872	121	-323	15
xav1sho11	-36	4	-71	5	2215	37	-2541	37	-326	7

[a] All energies expressed in kcal mol^{-1}. [b] VDW = Difference between the complex and individual species in Van der Waals component of free energy. [c] NE is the non-electrostatic contribution to the free energy NE = VDW + INT +SUR. [d] ELE = Difference between complex and individual species in electrostatic interactions. [e] PBCAL = electrostatic penalty involved in desolvating the peptide and RNA in complex formation . [f] $\Delta H = \Delta H^{S}_{Bind}$ = NE + PBELE.

Table A7-7. Decomposition of interaction component of free energy in Fx-RNA [a b c d e f]										
	VDW	SD	NE	SD	ELE	SD	PBCAL	SD	PBELE	SD
xav1fx16	-132	9	-166	10	-3567	101	3218	98	-350	12
xav1fx37	-96	15	-120	16	-3494	173	3128	168	-366	13
xav1fx41	-63	13	-86	13	-3408	172	3013	163	-395	15
xav1fx34	-83	10	-106	10	-3110	115	2737	115	-373	11
xav1fx31	-104	7	-134	8	-3462	101	3118	101	-345	13
xav1fx35	-109	9	-141	9	-3417	68	3084	67	-333	10
xav1fx27	-76	9	-99	10	-3443	108	3069	100	-375	13
xav1fx48	-101	7	-128	8	-3587	89	3244	92	-343	13
xav1fx46	-86	9	-112	10	-3162	74	2803	71	-359	14
xav1fx40	-59	7	-77	9	-3056	91	2662	88	-394	12
xav1fx28	-47	7	-72	8	-3386	103	2993	97	-394	17
xav1fx5	-70	7	-94	8	-3268	91	2899	92	-369	10
xav1fx30	-52	12	-74	13	-3188	253	2801	248	-388	10
xav1fx49	-88	3	-115	5	-3367	83	3020	81	-347	10
xav1fx23	-72	8	-94	9	-3584	90	3217	83	-366	14
xav1fx1	-66	12	-88	13	-3157	102	2785	104	-372	12
xav1fx9	-50	10	-65	11	-3425	138	3031	138	-394	16
xav1fx10	-51	6	-65	8	-3326	89	2933	85	-393	10
xav1fx7	-72	6	-91	7	-3255	95	2888	95	-367	8
xav1fx26	-45	11	-69	12	-3030	145	2642	146	-388	10
xav1fx22	-84	6	-109	7	-3298	68	2952	69	-346	17
xav1fx17	-85	7	-109	9	-3190	122	2844	123	-346	11
xav1fx32	-52	7	-76	8	-3129	95	2750	92	-379	12
xav1fx38	-44	6	-61	6	-3154	88	2761	90	-393	11
xav1fx2	-65	5	-88	7	-3101	63	2737	66	-363	11
AVG	-61	8	-82	9	-3196	124	2826	123	-369	12
xav1fx33	-111	8	-134	9	-3309	108	2993	106	-316	13
xav1fx45	-67	5	-82	6	-3000	91	2633	92	-367	8
xav1fx13	-57	12	-76	13	-3352	97	2979	96	-373	11
xav1fx42	-59	17	-75	17	-3322	133	2949	142	-373	17
xav1fx19	-68	9	-82	10	-3430	151	3066	148	-364	13
xav1fx47	-58	4	-80	6	-3298	96	2934	95	-364	12
xav1fx21	-59	9	-76	10	-3346	82	2981	82	-365	21
xav1fx15	-59	9	-77	10	-3346	85	2982	85	-364	21
xav1fx43	-39	5	-57	6	-2723	76	2340	76	-383	9
xav1fx12	-40	8	-60	9	-3250	164	2870	159	-380	17
xav1fx39	-24	9	-44	9	-3050	208	2654	201	-395	16
xav1fx52	-46	7	-66	8	-2952	78	2580	78	-373	12
xav1fx29	-47	5	-59	6	-3221	72	2842	68	-379	12
xav1fx3	-51	7	-68	8	-3208	73	2838	69	-370	11
xav1fx20	-76	5	-100	7	-3239	82	2904	80	-335	11
xav1fx50	-51	7	-70	8	-3277	101	2917	101	-360	10
xav1fx18	-34	8	-42	9	-3161	176	2774	172	-387	9
xav1fx24	-52	6	-70	6	-2997	93	2637	93	-359	7
xav1fx14	-46	7	-61	8	-3105	118	2739	117	-366	9
xav1fx6	-30	12	-47	13	-2809	163	2431	163	-377	13
xav1fx11	-26	7	-40	8	-2976	84	2593	84	-383	8
xav1fx36	-48	6	-63	7	-3038	60	2682	59	-356	11
xav1fx8	-22	10	-37	12	-2835	411	2454	412	-381	9
xav1fx25	-27	7	-40	8	-2534	160	2157	155	-377	13
xav1fx53	-21	14	-32	14	-2438	523	2060	522	-378	6

See footnotes to Table A7-6