A Mechanistic Study of the Stability of Triple-Stranded DNA Structures

Belinda Boehm
Supervisors: Dr David Huang & A/Prof Tara Pukala

A thesis submitted in partial fulfilment of the
requirements for the degree of
Master of Philosophy in Chemical Science

October 2017

THE UNIVERSITY
of ADELAIDE

Department of Chemistry, School of Physcial Sciences
The University of Adelaide
Australia

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

A	adenine
C	cytosine
C–H	Crick–Hoogsteen
DNA	deoxyribonucleic acid
ds	double stranded
FEP	free energy perturbation
FRDA	Friedreich's ataxia
G	guanine
GB	generalised Born
GBIS	generalised Born implicit solvent
LJ	Lennard Jones
MD	molecular dynamics
N	north
PBCs	periodic boundary conditions
PBE	Poisson-Boltzmann equation
PME	Particle-Mesh Ewald
PMF	potential of mean force
REMD	replica exchange molecular dynamics
RMSD	root-mean-square deviation
RNA	ribonucleic acid
S	south
SMD	steered molecular dynamics
ss	single stranded
T	thymine
TFO	triplex–forming oligomer
T_{m}	melting temperature
TTS	triplex target site
U	uracil
US	umbrella sampling
W–H	Watson–Hoosteen
W–C	Watson-Crick
WHAM	weighted histogram analysis method

ABSTRACT

The DNA triple helix is a non-canonical nucleic acid structure with roles in cellular regulation. Formed when a single strand of nucleic acid, a triplex–forming oligomer (TFO), binds in the duplex major groove, its formation is known to be involved in the onset of the neurodegenerative disorder Friedreich's ataxia (FRDA). Alternatively, it may have clinical applications in the development of novel antimicrobial agents. As a means to treat FRDA, methods to destabilise triplexes have been examined, with the binding of ligands in the minor groove a promising starting point. On the other hand, stabilisation of these structures may be achieved through modification of the backbone or the bases with the aim of suppressing bacterial genes as an interesting solution to the problem of multi-drug resistance. The origins of triplex stability are currently not well understood and a better comprehension of these properties is therefore an important step in the development of new medicinal technologies involving the creation or destruction of triplexes.

In examining the effect of ligand binding on triplex stability, a single minor-groove binder, netropsin, a known triplex destabiliser, is considered. Free energy calculations on the binding of a 15-base TFO to a duplex give good agreement with experiment and it is found that netropsin destabilises this triplex by approximately 15 kcal/mol. This appears to be a highly localised effect, occurring only when netropsin is bound opposite the TFO, and associated with a decrease in the width of the minor groove. Structural distortions associated with netropsin and TFO binding appear, therefore, to play a large role in the ligand's ability to destabilise triplexes.

Although netropsin is thought to bind in the minor groove, binding in the other two grooves formed when a TFO binds, has not been examined. Here it is found that binding in both the minor and W–H, being the larger of the two grooves formed by triplex binding, grooves are similarly stable, with the minor groove potentially being more stable due to strong vdW interactions.

To study the effect of TFO composition on triplex stability, the relative stabilities of purine and pyrimidine triplexes are examined. Significant distortion in the TFO backbone is observed in the case of the purine triplex, with Hoogsteen pairs between adenine bases failing to form, indicating it is likely less stable. However, pyrimidine TFO binding requires significantly more backbone rearrangement to bind to the duplex, potentially disfavouring its formation despite it likely being the more stable triplex.

Examining the effect of changing the TFO backbone composition from DNA to RNA, for a purine triplex, the two structures are found to have similar stability, despite reports of the RNA purine triplex not forming experimentally. A significant change in the sugar pucker of

the purine RNA backbone is found to be required for binding, potentially explaining this lack of formation. However, no large structural differences were found between DNA and RNA pyrimidine triplexes to explain the previously reported greater stability of the RNA structure, suggesting that conformational change may not entirely explain the relative stabilities of the triplexes.

DECLARATION

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

ACKNOWLEDGEMENTS

There are a number of people without whom the last two years would have been significantly more difficult. Firstly, to my supervisor Dr David Huang. Your willingness to take on a student with very little background knowledge, and then patience as I grappled with new techniques and theories, is greatly appreciated. Although this project has thrown up many unexpected challenges, you have remained optimistic that things will work out eventually, and provided many useful insights without which this thesis would be significantly shorter. I am incredibly grateful for the time you put in to supporting each of your students, myself included.

To the entire Huang group. I count myself very lucky to have shared an office with you over the past two years, even if it has gotten very hot on occasion! Not only your knowledge, but your willingness to share it, has been both helpful and inspirational.

Although not directly contributing to this work, there are many people who have kept me sane over the last two years and inspired me to continue, and to whom I owe considerable thanks. To my friends who have put up with me talking (probably nonsensically) about my research - thank you. And thank you for then distracting me with board games/food/good conversation. I am indebted to you all.

To my family. Mum and dad - I don't think you ever believed that I would do any kind of research (my adamance that it was the one thing I did *not* want to do probably contributed to that), but here I am. You have instilled in me a desire to do my best at everything I try, and a willingness to work hard, without which I would be lost. Thank you for your love and support (and baked goods), it has made the last couple years much easier. My brother, Martin - your willingness to attempt to help me when I throw problems at you is greatly appreciated.

Finally, to my husband, Jonathan. You have always encouraged me to continue, especially when things haven't going to plan, and helped me work through problems when I need a second opinion. I am incredibly lucky to have such a supporting and caring person in my life and I wouldn't have made it this far without you. Thank you.

DEDICATION

For Jonathan.

CHAPTER

ONE

INTRODUCTION

1.1 DNA STRUCTURE

The double helical deoxyribonucleic acid (DNA) structure, famously published by Watson and Crick in 1953,[1] is well known. At the most basic level, the structure consists of three building blocks: a nitrogenous base, a sugar, and a phosphate group linking the monomers together. In the common DNA double helix, strands are held together by complementary pairing between the bases, where adenine (A) pairs with thymine (T) (or uracil (U) in RNA) and guanine (G) with cytosine (C), forming 2 and 3 hydrogen bonds respectively. This bonding pattern is termed Watson-Crick (W–C) pairing (Fig. 1.1) but is not the only possible motif. Under certain circumstances these bases can be rotated, putting them in different relative positions with different hydrogen bond donors and acceptors interacting. In the case of Hoogsteen bonding, the same bases pair (A/T, G/C) but in a different orientation.[2] For this to occur, protonation of the cytosine residue is required. Alternatively, reverse Hoogsteen bonding can occur between two copies of the same base (A/A, G/G). These structures are shown in Fig. 1.2.

Of the nitrogenous bases, adenine and guanine contain two rings and are termed purines, while thymine/uracil and cytosine are single–ringed pyrimidine bases. Typically, the nucleic acid strands pair in an antiparallel fashion, where one strand runs 3'-5' and the other 5'-3' (Fig. 1.1), but parallel duplexes are also possible. In this case the bonding pattern is changed to reverse-W–C, where T pairs with T, G with G and so forth,[3] or Hoogsteen motif where the same pairing rules apply, but at different sites on the base.[4]

Overall, duplex DNA generally has a right handed structure; that is, it is twisted clockwise. Furthermore, the sugars are generally in an appropriate conformation such that the bases stack in what is known as a B-structure. To comprehensively define the geometry of the helix, a number of structural parameters are required. If the three axes of a base pair are defined such that x is the short axis, y is the long, and z is perpendicular to the plane of the base pair (Fig. 1.3), rotations about, and translations along, these axes can broadly describe the geometry. Rotations about the x, y, and z axes, relative to the base pair below, are known as the tilt, roll, and twist, respectively. Translations along these axes, again relative to adjacent base pairs, are likewise known as shift, slide and rise. Other parameters are also used to

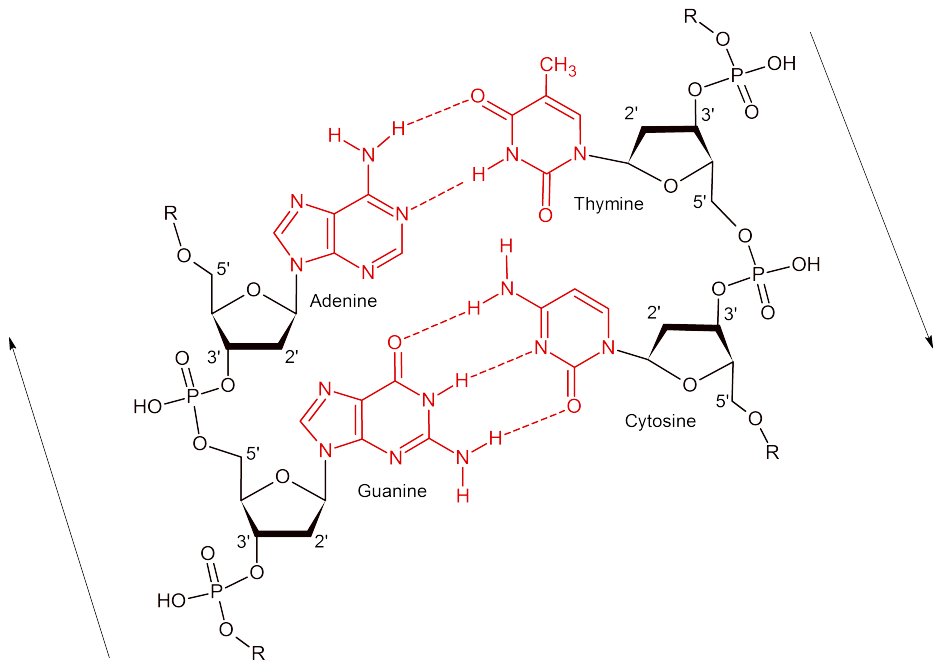


Fig. 1.1: The DNA double helix structure consists of a phosphate group, sugar (here deoxyribose) and base (red). Watson-Crick pairing (dashed lines) holds the two antiparallel strands together. Arrows represent the directionality of the strands.

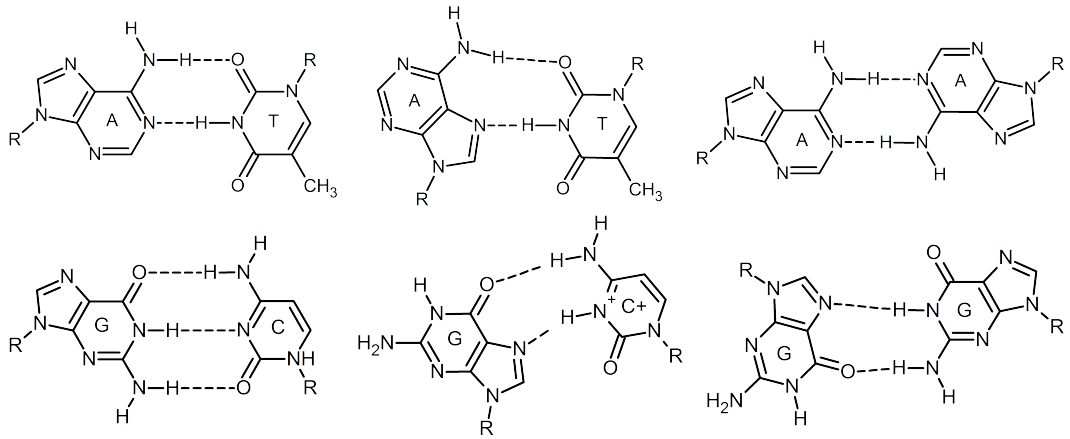


Fig. 1.2: The potential base pairing strategies based on adenine (top) and guanine (bottom). From left to right: Watson–Crick, Hoogsteen, reverse–Hoogsteen.

describe the helix geometry and are measured for individual base pairs, or the overall helical geometry, rather than relative to their neighbours. These include propeller twist, and inclination. Propeller twist is defined as the structure where, starting from a base–pair with a planar structure, one base is twisted backwards, and its pair forwards around the y-axis. Inclination, on the other hand is a slight rotation about the y-axis and is a general property of the overall helix.[5] Changes in these properties can define changes in the overall helical geometry and are hence important for quantifying the structural character of a helix.

In order to form a helix from the ladder like base pairs, the strands are twisted around a helical axis. The location of this axis is important for defining the overall geometry of the helix. In BDNA, this axis is not centrally located between the strands, but off center, giving a helix that has two asymmetric grooves. The smaller of these is known as the minor groove,

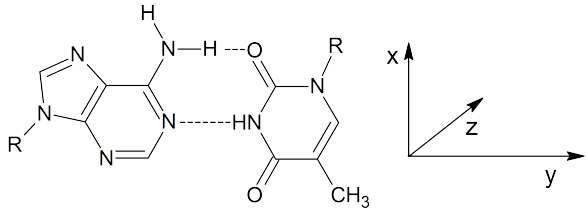


Fig. 1.3: The axes of a base pair. In the plane of the base pair, the short axis is the x and the long the y, while the z-axis is perpendicular to this plane.

and the larger the major (Fig. 1.4a), providing important recognition sites for DNA targeting molecules or proteins.

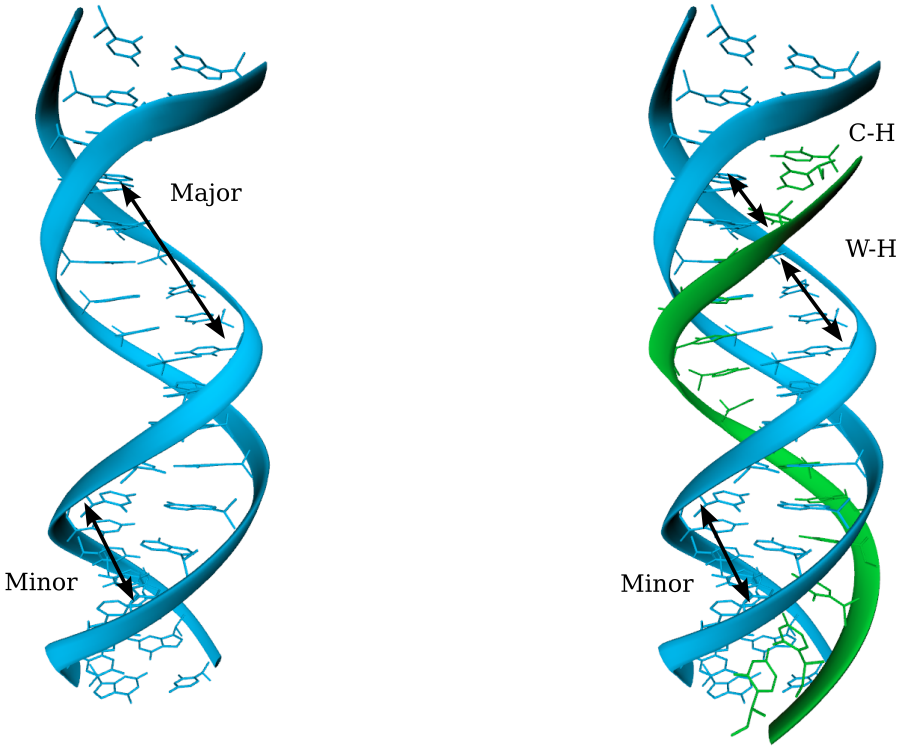
1.2 NON-CANONICAL DNA STRUCTURES IN HUMAN BIOLOGY

While the right-handed BDNA double helix is the most well known, many variations on this classic nucleic acid structure have been found to exist, either *in vivo* or synthetically. These range from slight variations in structure with retention of the W–C bonding pattern (such as A and ZDNA[6]), to changes in the strand direction to give parallel duplexes with reverse W–C[3] or Hoogsteen[4] base pairing, folded structures such as cruciforms or hairpins, and higher order structures such as the previously mentioned triplexes, and quadruplexes.[7,8] Although significantly less common than BDNA, many of these structures have been shown to form under physiological conditions and are generally accepted to have to have biological relevance, including involvement in certain diseases.[9–12]

Notably, DNA triple helices have been shown to compete with transcription factors in binding of DNA and have implications in both the genesis and prevention of disease.[13,14] These interesting structures form when a single–stranded nucleic acid fragment binds in the major groove of a polypurine/polypyrimidine B-DNA duplex. The incoming DNA strand, known as a triplex–forming oligomer (TFO), forms Hoogsteen bonds, rather than the regular W–C, with the polypurine duplex strand. In this structure, the polypurine strand is known as the 'Crick' strand, and the polypyrimidine the 'Watson' strand. Binding of the TFO splits the major groove into two asymmetric grooves. Of these, the smaller is known as the Crick–Hoogsteen (C–H), groove, and the larger the Watson–Hoosteen (W–H) groove, where the groove is between either the Watson or Crick, and Hoogsteen (TFO) strands (Fig. 1.4b). Given the right sequence, the three–stranded complex is relatively stable and has been implicated in the development of many neurodegenerative diseases including, but not limited to, Huntington disease, myotonic dystrophy, and Friedreich's ataxia.[12,15] Due to their ability to halt transcription, TFOs have also been proposed as a novel method for the treatment of bacterial infections or cancer. In this case, they are administered with the purpose of binding to known sites in the bacterial or human genome to prevent the expression of the dysfunctional or undesired gene.

1.3 TRIPLEX FORMATION IN FRIEDREICH'S ATAXIA

The neurodegenerative disease Friedreich's ataxia (FRDA) is caused by a mutation in the FXN gene, a gene involved in the expression of the iron chaperone protein frataxin, which leads to a repeat expansion of the GAA triplet.[16,17] Biologically, this manifests as a reduction in the production of the aforementioned protein, leading to an accumulation of iron in the

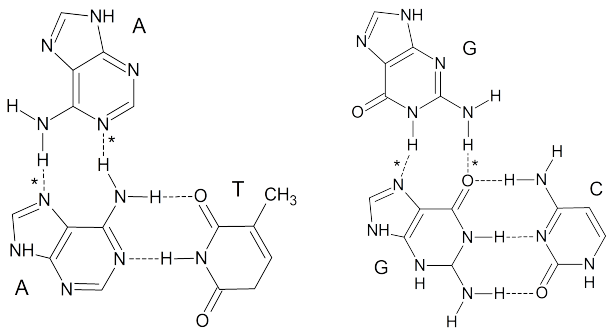


(a) Double helix (b) Triplex helix, TFO shown in green.

Fig. 1.4: The grooves of duplex and triplex DNA. To form a triplex, the TFO binds in the major groove, splitting it into two asymmetric grooves.

mitochondria.[18] Physiologically, sufferers endure severe weakening of the muscles to the extent of disability, and often have cardiac problems and diabetes, leading to an early death.[17] Despite the knowledge of its cause, current treatments address only symptoms.

The mechanism by which the GAA repeat expansion leads to the disease is accepted to be through intramolecular triplex formation. This mechanism has been well studied[16] and the purine nature of the expansion means that the structure is relatively stable under physiological conditions, involving reverse Hoogsteen base pairs between the TFO and the duplex (Fig. 1.5). Finding a method to destabilise this triplex is therefore the most obvious path to finding a more robust treatment for both this condition and other similarly caused diseases.



(a) A*AT triplet (b) G*GC triplet

Fig. 1.5: The base pairing in the FRDA triple helix. The Hoogsteen pair is indicated with *

1.3.1 PROPERTIES OF DNA TRIPLE HELICES

The ability of nucleic acids to form triple helices has been known since 1957[19] and in the time since then many studies have been undertaken to understand the structure, kinetics, and thermodynamics of this interesting molecule. The triple helix associated with FRDA involves the binding of a purine TFO antiparallel to the purine strand of the duplex purine strand and has been structurally characterised through many methods encompassing both molecular simulation[20,21] and various experimental techniques.[21,22] Such studies have shown that triplex stability has a complex dependency on a variety of factors such as pH,[23] salt concentration and cation valency,[19,23] and sequence.[24] It has also been shown that the presence of small molecules able to interact with the helix can have either a stabilising or destabilising effect on the structure.[25–28] The potential to use small molecules to modulate triplex stability has obvious therapeutic utility. Accordingly, it is important to understand the structural and thermodynamic effects underlying these stability changes in order to enhance the efficiency of development, and the potency, of any potential therapeutic agents.

1.3.2 MINOR GROOVE BINDERS

The minor groove is one of three major drug targets for DNA, the others being the major groove, and intercalation between the rungs, and binding of ligands here is generally associated with triplex destabilisation.[29] As such, it is a promising starting point for this study. These ligands typically constitute a series of hydrocarbon rings, have a curved structure complementary to that of the groove, and feature a similar binding mode. Netropsin is a typical example of this class of compounds, which also includes molecules such as distamycin, SN-6999, and Hoechst-33258.[29–31] Netropsin, having the typical curved, planar structure (Fig. 1.6), is easily incorporated into the minor groove with preference for AT–rich regions,[25,32] and can be thought of as a model compound for this class of minor groove binders. Although it has significant utility as a sequence–specific triplex destabiliser, it is also highly toxic, limiting its therapeutic use.[32] It is accordingly desirable to understand the mechanisms behind netropsin's triplex destabilising properties in order to develop molecules with similar, or improved, abilities, whilst eliminating the toxicity.

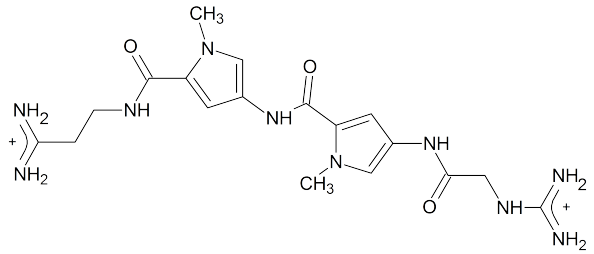


Fig. 1.6: Netropsin has a curved, planar structure, typical of many minor groove binders

The structural changes upon netropsin binding to triplex DNA have been previously studied both experimentally and computationally. It has been shown that the incorporation of netropsin into the minor groove effects the width of the groove,[33–35] as well as influencing the roll, propeller, and inclination of the helix.[36] These parameters are important in defining the overall structure of the helix and netropsin's influence on them indicates a significant structural rearrangement on binding. Interestingly, computational and experimental studies on the width of the minor groove in the presence of netropsin find conflicting results: computational studies show a slight narrowing, particularly in the region where the molecule is

bound,[35] while experimental solid state structures show a broadening.[33,34,37] Furthermore, it has been shown that the binding of a TFO in the major groove causes significant widening of the groove,[38] but whether the interplay of structural effects of netropsin and TFO binding plays a role in the destabilisation of the triplex structure remains unknown.

While it is well known that netropsin binds in the minor groove, a nucleic acid triple helix has two other grooves which may be suitable in size for netropsin to bind: the C–H and W–H grooves, being the smaller and larger grooves respectively formed between TFO and duplex when TFO binds. The reasons for netropsin's stability in the minor groove are well understood. They include a combination of displacement of ordered solvent hydrating the DNA's spine, the negative charge of the nucleic acid phosphate groups and the large electrostatic potential of A–rich regions of minor groove attracting the positive charge of netropsin, the shape complementarity, and hydrogen bonding interactions with the N3 and C2=O atoms the A/T or G/C pairs.[39–41] Despite this, it is not clear why binding would favour the minor groove over either the C–H or W–H grooves, particularly the similarly sized W–H groove.[29] As computational tools allow for a netropsin molecule to be placed in any groove and its interactions examined, this may provide useful insight into the netropsin's preference for minor groove binding.

1.4 INDUCED TRIPLEX FORMATION

As triplex formation is known to suppress gene expression, it has potential use in the development of antimicrobial compounds, whereby it effectively knocks out an undesired gene. Through the introduction of short TFOs, which bind at specific bacterial genomic sites, the expression of genes which are essential for either bacterial survival or drug resistance can be greatly reduced, if not completely turned off.

This process, known as antigene therapy involves the formation of an intermolecular triplex between the native (bacterial) duplex and a TFO designed to bind at a specific sequence. Ideally, the TFO binds specifically, and with good *in vivo* stability, at the promoter region of an essential gene, preventing its entire transcription. Evidently, this requires knowledge of the target sequence, which must also be appropriate for triplex formation. The affinity of a TFO for a duplex is maximised when the sequences are homopurine or homopyrimidine,[9] but these bacterial sequences must be sufficiently unique that the same sequence will not be found and additionally targeted in the human genome. Furthermore, the formation of Hoogsteen bonds requires cytosine to be protonated, which is generally not feasible at physiological pH due to its pK_a of 4.45. To combat this, TFOs must contain a purine motif to be useful in this regard. The search for these relatively unique, homopurine/homopyrimidine triplex target sites (TTSs) can be facilitated using the Triplexator software.[42]

Once bound, it is important that the TFO remains so. Good *in vivo* stability, including resistance to degradation by nucleases, is accordingly exceedingly important for biological activity of TFOs as antimicrobial agents. Much like destabilisation of triplexes, this can be achieved through the binding of small molecules,[27,28] or alternatively through modifications to the TFO backbone[43] or bases.[44,45]

1.4.1 BASE MODIFICATION: STABILITY OF PURINE AND PYRIMIDINE TRIPLEXES

For a given duplex sequence, there are three major classes of TFO which can bind with sequence specificity: GA (purine), CT (pyrimidine), and GT (mixed). Each will recognise the

same sequence but display different base pairing patterns. A purine TFO will bind antiparallel to the purine strand of the duplex, forming reverse Hoogsteen A*A and G*G bonds. On the other hand, pyrimidine TFOs bind parallel to the duplex purine strand forming C^+*G and T*A Hoogsteen pairs. Mixed sequences can bind either parallel or antiparallel, forming either Hoogsteen or reverse Hoogsteen pairs.[46–48] As protonation of the cytosine residue is required for the formation of the pyrimidine triplex, this is a pH dependent process and, although this is the most commonly studied triplex, due to cytosine's pK_a of approximately 4.5 its formation is generally not feasible at physiological pH. Modifications to the cytosine bases may be able to circumvent this need for protonation, making it a viable option as an antigene therapeutic and have been shown to be capable of removing the pH dependence of this triplex formation.[49]

In designing TFOs to target specific bacterial genes, it is important to be able to maximise the stability of the triplex through good choice of sequence. The parallel-pyrimidine triplex has been much more extensively studied than its antiparallel-purine analogue despite both having been shown to have fairly similar stabilities under the right conditions.[46] However, it is known that for the purine TFO, the percentage of guanine bases in the TFO is important for stability, with reports of at least 40–50 % guanine content required for stable purine triplex formation.[47,50] Too high a G content, on the other hand, can lead to the favouring of quadruplex structures over the desired triplexes.[47] The origin of this effect is not particularly clear, although it would suggest that the reverse Hoogsteen A*A pair is weaker than the corresponding G*G.

It is known that the G*G and A*A pairs are not isomorpohous, as the G*C and A*T pairs are, which may cause a structural distortion of the backbone between A and G bases,[48] potentially contributing to the dependence of stability on G content. Comparing the structures of the purine and pyrimidine triplexes, particularly the positioning of the TFO in the major groove, may therefore provide insight into the reasons for the stability of one over the other, and the dependence of the stability of the purine triplex on G content.

1.4.2 Backbone modification: stability of RNA-DNA hybrid triplexes

Possibly the simplest backbone modification is the replacement of the deoxyribose sugar in the DNA motif with a ribose sugar, transforming the TFO to ribonucleic acid (RNA). RNA may be more useful in the formation of triplexes, not only for its potentially higher stability as a TFO, which has been extensively studied,[51–53] but also as the majority of *in vivo* single–stranded nucleic acids are RNA based and therefore may have better resistance to degradation in the cellular environment.

For a duplex with two DNA strands (DD), it is known that the introduction of a parallel pyrimidine RNA–based TFO, where the pyrimidine based TFO binds parallel to the purine strand of the duplex, provides further stability to the triplex relative to the same triplex formed with a DNA TFO. Despite this, no quantitative consensus has been reached in the literature as to the extent of this effect. For this pyrimidine–based triplex, Roberts *et al.*[52] report a ΔG which is 0.9 kcal/mol per base pair more favourable for the DD·R (DNA duplex with RNA TFO) triplex than the DD·D (DNA TFO) triplex. Han *et al.*,[54] on the other hand, found only a slight change in ΔG of 1.2 kcal/mol over an entire triplex of similar sequence and length. Wang and Kool[51] report a free energy difference of 4 kcal/mol for an 11-base–pair triplex, and Asensio *et al.*[55] a similar 3 kcal/mol for a 6-base–pair triplex. Although these differences may be attributed in part to different methods or sequences, the similarities

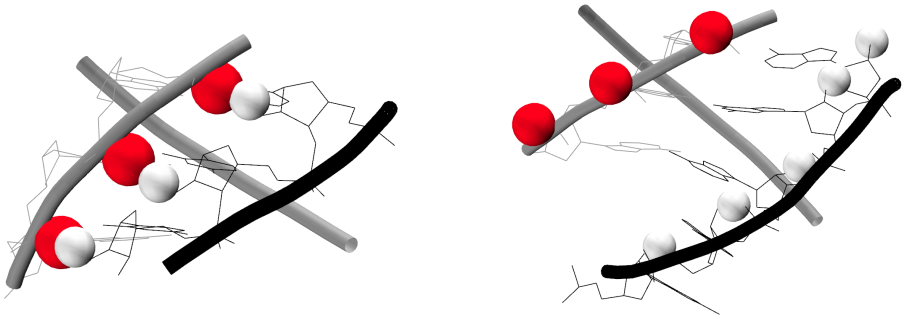
between experiments are enough that such a discrepancy is not expected.

All aforementioned studies limited their scope to the stability of the case of a parallel pyrimidine TFO. Studies on the antiparallel purine TFO, where the purine third strand binds antiparallel to the duplex purine strand, have found that it is significantly less stable than the corresponding DD·D purine triplex, not forming in many cases.[46,56,57]

Differences between DNA and RNA as TFO. There are two fundamental differences between RNA and DNA TFOs: the presence of a hydroxyl group on the sugar, and the substitution of thymine in DNA for uracil in RNA, which differ by a methyl group. Both of these changes are likely to have an effect on the overall structure of the nucleic acid, and potentially influence their stability.

As the thymine–uracil substitution is unique to the pyrimidine triplex, it is tempting to think that this may be the origin of the greater relative stability of the DD·R triplex over the DD·D. However, based on studies of the hybrid duplex, which find an increase in stability with the purine content of the ribose strand,[58] it would appear that the change from thymine to uracil does not result in a more favourable complex, and may in fact have the opposite effect.

The other difference, that of the sugar composition, may have two effects. Firstly, the 2'-hydroxyl group on ribose has been proposed to form hydrogen bonds with the phosphate group of the central triplex strand when in an appropriate orientation.[55,56] This interaction is only possible when the TFO is parallel to the central strand (Fig. 1.7), potentially explaining the opposite effects of antiparallel and parallel TFOs on the stability relative to the DD·D triplex. However, methylation at this position (OH $\rightarrow$ OMe) does not reduce the stability of the DD·R triplex, but rather increases it, despite a loss of hydrogen bonding capability. This effect is often attributed to an increase in van der Waals interactions.[55,59] Secondly, when in duplex form, the sugar pucker, defined by the positioning of the C2' and C3' ribose atoms relative to the plane of the sugar, varies significantly between DNA and RNA. Having sugars in a south (S) conformation, predominantly C2'endo (where the C2' atom is out of the plane of the ring), gives rise to B–form structures, while A–form structures, such as those preferred by double stranded (ds)RNA, characteristically present north (N), or C3'-endo (with the C3' atom out of the plane), sugars. Favouring one over the other, therefore, may lead to a large degree of structural rearrangement.



(a) Parallel triplex (b) Antiparallel triplex

Fig. 1.7: Comparison of the capability of hydrogen bonding between ribose 2'-OH (hydrogen shown in white) and phosphate group (oxygen shown in red) in antiparallel and parallel triplexes. In the parallel structure, the groups are close enough to allow for hydrogen bonding to occur. TFO shown in black.

These differences in puckering are well know for homogeneous duplexes, but it does not necessarily hold that the same differences will be observed in a hybrid structure such as the DD·R, or even the DD·D, triplex. Reports on the structure of hybrid DNA–RNA duplexes have shown an intermediate structure between the A and B forms,[60] where the RNA strand shows an ARNA structure and the DNA strand a mixture of N and S sugar conformations. Asensio *et al.*[55] noted that for the parallel pyrimidine DD·R triplex, a similar locking of the RNA TFO into an A conformation (with N, C3'-endo sugars) is observed while the underlying DNA duplex maintains its S conformation sugars typical of BDNA. If a strand of polypyrimidine single stranded (ss)RNA in solution also favours N sugars, the lower degree of structural rearrangement required for it to bind to the duplex may begin to provide a rationalisation for the higher stability of this structure, provided that the barrier between C2' and C3'-endo is sufficiently high. The structural rearrangement the duplex undergoes when TFO binds should also be considered.

In considering the degree of structural change required of the duplex in the binding of the TFO, it has been shown that the structures that require a smaller conformational rearrangement of the duplex, namely the parallel-pyrimidine DD·R and corresponding methylated TFO triplexes, show higher stability. In fact, the amount of structural reorganisation required was found to, not unexpectedly, mirror the order of stabilities.[55] Although this begins to explain the stability of the pyrimidine motif DD.R triplex, the purine motif has not been equivalently studied, nor has the effect of the structural changes occurring in the TFO. A comparison of the sugar conformations of purine and pyrimidine triplexes, with both DNA and RNA TFOs, with the conformations of the native duplex and single stranded nucleic acid may shed light on this issue.

1.5 Project Outline

The research presented here focuses on examining the stability of nucleic acid triplexes with aims of either inducing destabilisation through the binding of minor–groove binder netropsin, or forming triplexes with higher stability through alterations to the TFO sequence or backbone.

Concerning the effect of netropsin on the stability and structure of a DNA triple helix, only binding in the minor groove is considered, as this is its known mode of action. This is discussed in detail in chapter 3. Netropsin binding in the C–H and W–H grooves is then compared to regular minor groove binding to uncover the reasons behind its preference for minor groove binding, also discussed in chapter 3. Through the use of computational, molecular dynamics (MD), methods, the effect of netropsin on the dissociation free energy of a 15-base DNA triplex is calculated. As this system has been previously studied without netropsin,[61] it initially also provides a good model system to test the viability of free energy calculations conducted by means of atomistic MD simulations. Binding free energy calculations of DNA triple helices have not previously been done, nor have these types of calculations been attempted for the study of such large, flexible molecules, being more commonly found in the study of binding free energies of small ligands or relative stabilities under mutation. Netropsin's effect on the overall structure, and certain features such as the minor groove width, are examined, providing a potential understanding of the underlying reasons for its ability to destabilise triplexes.

In chapter 4, the effect of TFO sequence on triplex stability is examined, with the aim of determining an optimal TFO for antigene strategies. To elucidate the reasons behind any

potential sequence effect, purine and pyrimidine TFOs able to bind to the same duplex sequence are considered. Through examination of properties such as the number of hydrogen bonds between TFO and duplex, as well as the structure of the TFO in the groove, a better understanding of the stability, or lack thereof, of these triplexes can be obtained. Although a number of studies have experimentally looked at the stabilities of different triplexes,[46,47,62] this work presents the first analysis via computer simulations of the potential reasons behind the stability of different sequences, allowing for the development of a better understanding of triplex stability. Understanding the reasons behind the stability of certain triplex sequences paves the way for the design of potentially more stable structures with promising applications in antigene therapy.

In the examination of the effect of backbone composition on triplex structure, computational methods, by means of atomistic MD simulations, are again used. Structural parameters of DD.D and DD.R triplexes are calculated, focusing on the parallel purine structures which have not been extensively studied, and compared with their antiparallel pyrimidine equivalents. The overall structure of the triplexes with varying backbone composition is compared, as well as the effect triplex formation has on the puckering of sugars in both TFO and duplex. This qualitative structural information, combined with further free energy calculations, aims to provide new insights into the reasons underlying the different stabilities of DNA and RNA TFOs for triplex formation. Furthermore, the free energy landscape of the puckering in ssDNA and ssRNA is studied as a means of understanding the barrier to structural rearrangement required for the formation of an antiparallel purine DD·R triplex. Although the purine DD.R triplex has been shown to be very unstable experimentally,[46,56,57] no reasons have been proposed for this. It is therefore important to understand if, and why, this triplex is significantly less stable than similar structures to better understand the factors that stabilise and destabilise triplexes. Additionally, the pyrimidine triplex is known to be more stable with an RNA TFO than with DNA and a number of reasons have been hypothesised for this effect. This work attempts to determine the viability of these predictions through analysis of the structure of these triplexes. These results are outlined in chapter 5.

CHAPTER

TWO

COMPUTATIONAL METHODS

2.1 MOLECULAR DYNAMICS SIMULATION

In the computational study of biomolecular systems, one is often challenged by the scale of nature. A short deoxyribonucleic acid (DNA) triplex, too short to be biologically relevant, already contains over 1,000 atoms and when the cellular environment, full of water and ions, is also taken into account, the system size generally exceeds many tens of thousands of atoms. The methods most readily available to this scale fall under the category of classical molecular dynamics (MD). As we are generally only interested in nuclear motion on this scale, the classical treatment is appropriate and quantum effects generally do not need to be taken into account. While it is relatively simple to study these large systems experimentally, the limits of experimental techniques mean that neither atomistic detail, nor dynamic properties at a nanosecond scale can be obtained in this way. This can lead to a lack of mechanistic detail which MD simulation can often help uncover.

Despite the existence of many MD codes, the general algorithm for running a simulation remains the same: for a given configuration of all atoms, an initial velocity distribution is specified, often a Boltzmann distribution. The forces $\vec{F}(\vec{r})$, defined by the chosen force field, acting on each atom are then calculated according to Newton's classical equations of motion,

$$\vec{F}(\vec{r}) = -\nabla U(\vec{r}), \tag{2.1}$$

where $U(\vec{r})$ is the potential energy function defined by the interactions between each pair of particles. For simulations in the NVE (constant number of particles, volume, and energy), the atomic positions are then updated by integrating the above equation. In NAMD,[63] the MD software used for this work, this is achieved by means of the velocity Verlet algorithm.[64] The process is iterated, with a timestep on the order of 1–2 fs, for as long as required. An initial equilibration period is generally required, where this process is repeated until the average system properties become independent of time, that is, the system has reached thermodynamic equilibrium, prior to any data collection or measurements.

If constant temperature or pressure is required, as is generally the case when attempting to model experimental systems, algorithms to maintain these parameters at a specified average value must be applied. Average temperature can be maintained through the use of a thermostat. In this work, Langevin dynamics were used for temperature control, in which additional damping and random forces are applied to the momenta. The random forces are assigned from a Gaussian distribution, and the damping introduces friction to the system. Combined, these lead to the correct temperature distribution. Pressure control was implemented through the Nosé–Hoover Langevin piston barostat.[65,66] This algorithm introduces an additional degree of freedom, the volume of the simulation box, to which the system is coupled. This acts as a piston with a given mass and adjusts to compensate for internal and applied pressures. As with the temperature control, Langevin dynamics are used to control the piston fluctuations.

In order to accurately model the dynamics of the system, it is required that the timestep be smaller than the period of the fastest vibration. For high frequency vibrations such as those for bonds containing hydrogen, a very small timestep would accordingly be required. It is therefore often useful to constrain hydrogen–containing bonds to their equilibrium values which allows for a longer timestep to be used, increasing the speed of the calculation. Here, the SHAKE algorithm[67] is used to constrain all bonds containing hydrogen, allowing for the use of a 2 fs timestep.

2.1.1 MOLECULAR MECHANICS FORCE FIELDS

The choice of force field for MD simulation is an important one, as it defines how the atoms interact with one another. In all cases, the total potential energy, is a sum of bonded and non-bonded interactions

$$U(\vec{r}) = \sum U_{\text{bonded}}(\vec{r}) + \sum U_{\text{nonbonded}}(\vec{r}). \tag{2.2}$$

For molecular simulation of nucleic acids, the CHARMM36[68] and AMBER parm99 force field[69] with BSC0 modification[70] (parm99bsc0) are the most common currently in use, although some modifications to parm99bsc0, in BSC1[71] and OL15,[72] have recently been shown to improve structural calculations on longer timescales.[73] In this work we primarily use the AMBER parm99bsc0 force field, as it has been shown to accurately represent the structure of DNA triplexes, as well as RNA structures, on the relevant timescales.[70] For this force field, the total potential energy takes the form:[69]

$$\begin{aligned} U(\vec{r}^N) = & \sum_{i\in\text{bonds}} k_{b,i}(l_i - l_{0,i})^2 \\ & + \sum_{i\in\text{angles}} k_{a,i}(\theta_i - \theta_{0,i})^2 \\ & + \sum_{i\in\text{torsions}} \sum_n \frac{1}{2} U_n[1 + \cos(n\omega_i - \gamma_i)] \\ & + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \varepsilon_{ij} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right\} \end{aligned} \tag{2.3}$$

where the bond, angle, and torsion terms constitute the bonded interactions and the final

term the non-bonded interactions between all pairs of atoms. The non-bonded interactions comprise a Lennard Jones (LJ) (van der Waals) and a Coulombic (electrostatic) term, the first and second terms of the summation respectively. The first term in this equation, the bond length, is described as a harmonic potential with force constant $k_{b,i}$ for a bond length at distance $b_i - b_{0,i}$ from its equilibrium distance $b_{0,i}$. Similarly, the angle term has force constant $k_{a,i}$ and angle $\theta_i - \theta_{0,i}$ from equilibrium. The torsional term includes amplitude U_n, dihedral angle ω_i with multiplicity n, and phase angle γ_i. The LJ component of the non-bonded interactions is calculated between atoms i and j at distance r_{ij}. The potential energy well defined by this equation has depth ε at interatomic distance r_{0ij}. Finally, the electrostatics term is calculated between atoms i and j with charges q_i and q_j respectively and interatomic distance r_{ij}. The parameters are derived from a combination of quantum chemical calculations and experimental data so as to provide a good model for the 'real' system under study.

2.1.2 Approximating a bulk system: periodic boundary conditions

Computationally, it is currently only feasible to simulate systems up to on the order of 1,000,000 atoms, giving results dominated by surface effects. As bulk properties are desired, periodic boundary conditions (PBCs), which effectively extend the system infinitely by replicating the simulation box in three dimensions to produce a continuous medium, are usually applied. In this case, generally only the forces between the closest periodic images are calculated (Fig. 2.1). Known as the minimum image convention, this greatly reduces the number of interactions that need to be calculated. Assuming the system size is larger than the correlation length of the fluid, this can improve the efficiency of the calculation while still giving a good approximation of the bulk solvent environment.

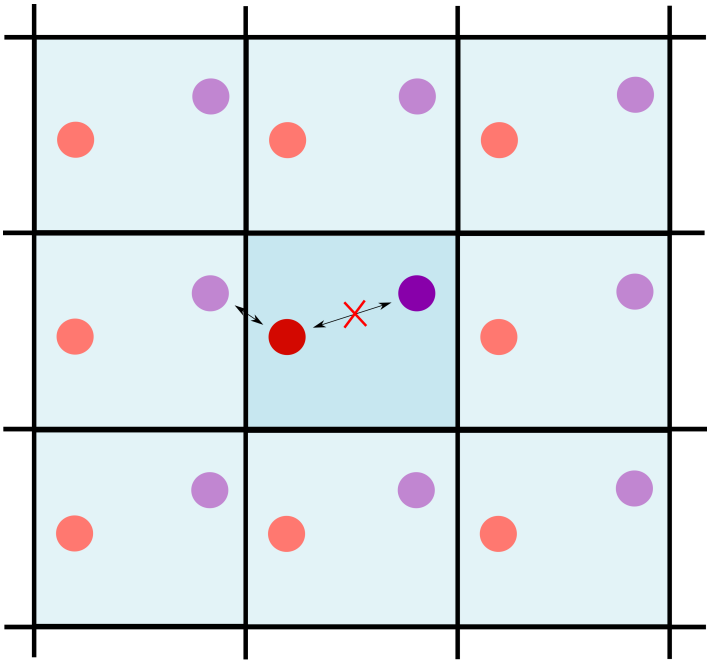


Fig. 2.1: Under PBCs, the force between two atoms is only calculated between the closest images (minimum image convention). Periodic boxes extend infinitely along all three axes.

2.1.3 COMPUTING LONG-RANGE INTERACTIONS, AND INTERACTION CUT-OFFS

As the force calculation relies on the computation of the interactions between every pair of atoms in the system, this can be a very slow process, scaling with N^2 for the number of particles, N, in the system. Instead, it is often assumed that only interactions between particles closer than a cut-off distance, r_c contribute significantly to the calculation of force and any interactions beyond this distance can be safely truncated without significant error. While the LJ interactions are short ranged, that is their strength decays at such a rate that they are negligible beyond r_c, the electrostatic interactions have a much longer range and hence cannot be so easily neglected, especially when PBCs are employed. In this case, algorithms such as Particle-Mesh Ewald (PME),[74] which is used in this work, can be applied to calculate the electrostatic interactions between all periodic images. In this method, interactions are divided into short– and long–range, being before and after the specified cut-off distance respectively. Short–range electrostatics are calculated explicitly, while the long–range electrostatics are calculated using fast Fourier transform methods.

While the use of a cut-off is attractive in that it significantly reduces the computational effort required, it introduces a discontinuity in the energy: the energy is equal to $U(\vec{r})$ below the cutoff distance, and 0 beyond it, which may produce unphysical behaviour. This can be improved in one of two ways. Firstly, the energy can be shifted such that $U(\vec{r}) = 0$ at r_c , introducing a discontinuity in the force. Alternatively, a switching function can be used, which, beginning before r_c, smoothly brings the energy to 0 at r_c, avoiding both the discontinuity in energy of the truncated method, and that of force in the cut-off and switch method. In the calculations presented here, a cut-off of 10 Å is employed, with NAMD's implementation of a switching function from 9 Å.

2.1.4 SOLVENT

It should come as no surprise that biological processes do not occur in a vacuum, nor often in the gas phase. Thus, for their accurate modelling, it is important that solvent contributions are taken into account. This can be done in one of two ways: by explicitly including the solvent and ion molecules in the simulation (explicit solvent), or by representing them as a homogeneous dielectric medium (implicit solvent).

A number of explicit solvent models exist, which differ in the number of interaction sites, the rigidity, and the polarizability of the molecule. In NAMD, three water models can be used, although many more exist. These are the TIP3P, TIP4P,[75] and SWM4-NDP.[76] TIP3P is a 3 site model, with one interaction site corresponding to each atom, while TIP4P includes a negatively charged dummy atom as a fourth site, and SWM4-NDP is a five site model, including an additional negatively charged Drude particle, which introduces polarization.[76] All three are rigid with the relative positions of all three atoms constrained to their equilibrium values. Here we use the TIP3P model as the AMBER force field used here was parametrized for this water model.

Explicitly including water molecules in the system greatly increases the number of interactions to be calculated in the course of an MD simulation, increasing the computational cost accordingly. Although it is generally more realistic physically, it is often not feasible. Instead, it is often advantageous to consider solvent as a continuum with a constant dielectric. In the implicit solvent calculations presented in this work, the generalised Born implicit solvent (GBIS) model,[77] which is an approximation of the Poisson-Boltzmann equation (PBE),

is used. While the electrostatic environment of a molecule in a dielectric medium can be calculated exactly using the PBE, it is time consuming and hence would offer little advantage over explicitly including solvent. The generalised Born (GB) model instead models the system as the solute, represented by a series of charged spheres with a lower dielectric than the medium, in a solvent of constant (higher) dielectric.

2.1.5 EQUILIBRIUM SIMULATIONS

It is often useful to be able to visualise and quantify the behaviour and properties of a system at equilibrium over the timescales available to MD. Here, for example, equilibrium structural features of a DNA triplex, such as the major and minor groove widths, are examined in the presence and absence of the minor–groove binder netropsin. Prior to data collection, the system must be equilibrated – that is simulated until the properties of the system cease to vary with time. Following this, a period of data collection can be undertaken for as long as required. In this work, simulations were conducted in the isothermal-isobaric (NPT) ensemble, where the number of particles (N), pressure (P), and temperature (T) are constant. Pressure and temperature were maintained at 1 atm and 310 K, using the Nosé-Hoover Langevin piston barostat, and Langevin thermostat respectively, to mimic the cellular environment.

2.1.6 DEFINITION OF STRUCTURAL PARAMETERS

In this work, a structural analysis of the changes occurring on netropsin binding to the triplex associated with Friedreich's ataxia (FRDA), as well as the effect of triplex–forming oligomer (TFO) backbone composition on triplex structure, is presented.

The major groove width is defined as the interstrand P_n–P_{n+5} distance, that is the distance between phosphorus atoms of base n and the base 5 bases higher of the opposite strand. The interstrand phosphorous distance has been previously used as a measure of groove–widths[78,79] and the bases were chosen to give roughly the shortest distance between the two chains of the duplex moiety on the major groove side. The minor groove width is similarly defined as the P_n–P_{n-4} distance which generally also gives the shortest interstrand distance covering the minor groove.

The number of hydrogen bonds between the TFO and the central chain of the duplex was calculated using the Hbonds plugin for VMD.[80] Hydrogen bonds are defined as having a donor–acceptor distance less than 3.0 Å and an angle less than 25°. This choice of values gave good coverage of the expected hydrogen bonds between the duplex strands without including additional intrastrand hydrogen bonds. The average number of hydrogen bonds was then calculated over the last 20 ns of a 40 ns trajectory, to be certain that the system was at equilibrium.

Large conformational changes in nucleic acid structure can be often associated with changes in the sugar puckering. As a planar 5 membered ring is subject to considerable ring strain, rather than exist in this planar conformation, 5 membered sugars, of which ribose and deoxyribose will be considered here, exhibit a conformational change whereby certain atoms exist out of the plane of the rest. This is known as the sugar pucker and is highly variable, with the ring able to adopt, and move between, any one of 10 broad classes of structure (Fig. 2.2). As both the nucleic acid backbone and the base are connected to the sugar, the conformation of this component has significant effects on the structure of the nucleic acid as a whole, being responsible for the differences in structure between, for example, A and B

form nucleic acids.

Two parameters define the sugar pucker, the puckering phase and amplitude. Broadly, the phase defines which part of the ring (ie. which of the 5 atoms) is furthest from the plane and the amplitude how far it is from planarity (Fig. 2.2). Depending on the formalism, the amplitude can be defined as the displacement of the atom from the plane, or the deviation of the ring torsion from planarity.[81] The definition used by PLUMED2[82], used in this work, is based on the Altona and Sundaralingham formalisation[83] where the puckering is defined by two Cartesian parameters, Z_x and Z_y:[84]

$$Z_x = \frac{v_1 + v_3}{2\cos(4\pi/5)} \tag{2.4a}$$

$$Z_y = \frac{v_1 - v_3}{2\sin(4\pi/5)}. \tag{2.4b}$$

Here, v_1 and v_3 are the torsions between atoms C1'–C2', and C3'–C4' respectively. Puckering parameters are generally expressed in terms of polar coordinates, P_θ for phase and A_r for amplitude. These can be readily expressed as:

$$P_\theta = \arctan(Z_y/Z_x) \tag{2.5a}$$

$$A_r = \sqrt{Z_x^2 + Z_y^2} \tag{2.5b}$$

Qualitatively, the preferred sugar pucker for each strand of the triplex can be obtained from the analysis of the average equilibrium structure of the molecule of interest. The phase of the sugar pucker was measured using the COLVAR module of PLUMED2.[82] The C3'-endo pucker typical of RNA is defined by a pseudorotational phase of between 0–36°and C2'-endo is between 144–180°. More broadly, a sugar is considered to have a north (N) pucker if it has phase between 270–90°and south (S) if between 90–270°(Fig. 2.2).

2.2 FREE ENERGY METHODS

Whilst equilibrium simulations can be useful in determining the structural properties of a system, when it comes to the calculation of the thermodynamic properties of a system, or a reaction, it is generally not so simple. Often, thermal energy is not enough to encourage a system to cross a barrier from one state of interest to another and simple equilibrium MD simulations may never sample these particular, potentially important, regions of phase space. While there are many methods for calculating, through molecular simulation, the free energy differences between two states of interest, all of which which find use in different cases, only those which are used in this work are examined here.

2.2.1 MM/GBSA

Potentially one of the simplest methods for calculating the free energy difference between two states is the use of molecular mechanics (MM) combined with the generalised Born

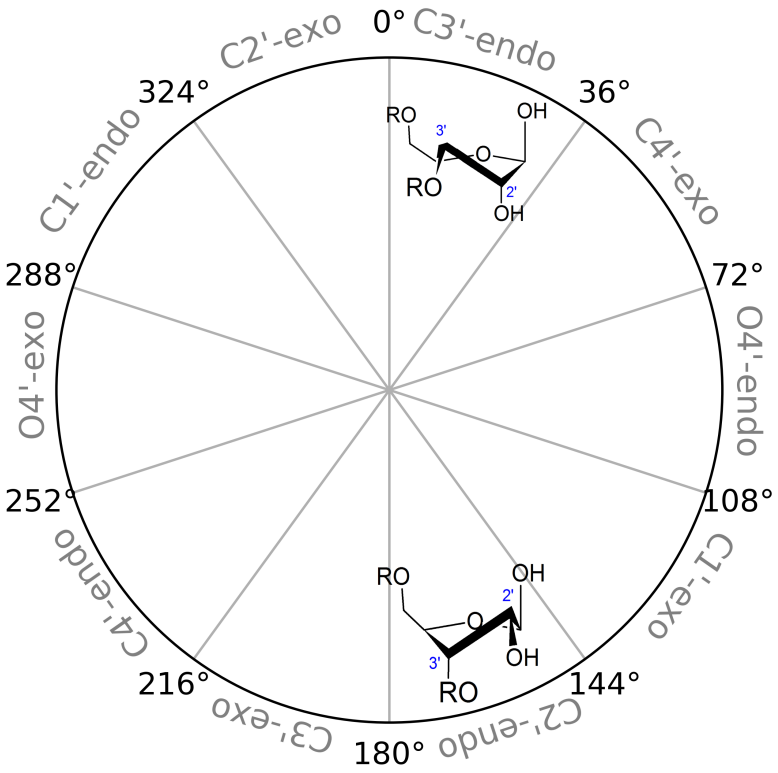


Fig. 2.2: Sugar puckering phase defines the atom positions in the 5 membered ring (here ribose shown). The two most relevant conformations for DNA/RNA are shown – C3'-endo at phase between 0–36°and C2'-endo at 144–180°. In C3'-endo, the C3' atom is above the plane of the rest. Similarly for C2'-endo

surface area implicit solvation model (GBSA). Often used for protein-ligand binding characterization, this method involves the calculation of the free energy for the protein, ligand, and protein-ligand complex individually, each as

$$G = E_{\text{bonded}} + E_{\text{elec}} + E_{\text{vdW}} + G_{\text{polar}} + G_{\text{non-polar}} - TS \tag{2.6}$$

where the E terms are the energy terms obtained from the molecular mechanics simulation, G_{polar} and $G_{\text{non-polar}}$ the polar and non-polar contributions to the solvation free energy, and TS the entropic contribution. G_{polar} is calculated through the use of the generalised Born equation and $G_{\text{non-polar}}$ using the solvent accessible surface area.[85] In this work, both are included in the value of E_{elec}.

Here, this method is used to calculate the binding free energy of netropsin to a triplex. ΔG for the binding of the ligand to the triplex can therefore be obtained as

$$\Delta G_{\text{binding}} = \left\langle G_{\text{complex}} - G_{\text{triplex}} - G_{\text{netropsin}} \right\rangle - T\Delta S \tag{2.7}$$

for the free energies of complex (netropsin + triplex), triplex, and netropsin calculated using equation 2.6, averaged over the trajectory. In this case, G_{complex}, G_{triplex}, and $G_{\text{netropsin}}$ are calculated from the simulation data as $E_{\text{bonded}} + E_{\text{elec}} + E_{\text{vdW}}$, which does not include the TΔS

term. This term is later included and is taken to be purely based on the loss of configurational entropy on ligand binding. This has been calculated previously for netropsin.[86,87]

2.2.2 Steered Molecular Dynamics

Steered molecular dynamics (SMD) is one method of speeding up molecular processes to timescales manageable by MD simulation which is particularly useful for the calculation of binding/dissociation free energies. The method functions by pulling, either with a constant force or velocity, a set of atoms along an appropriate order parameter to actively induce the process of interest. It has been commonly used to investigate the unfolding of proteins, but is equally applicable to examining the unfolding of single–stranded nucleic acids. In this case, the relevant order parameter is the end-to-end distance of the nucleic acid strand. The C5' atom at the 5' end of the chain is fixed in position, while the phosphorous atom at the 3' end is attached to a dummy atom via a harmonic spring. A force is then applied to the dummy atom, pulling the selected atom at constant velocity. Alternatively, a constant force can be applied to the spring to move the atom with variable velocity. For constant velocity SMD, which is used here, the force, F between the dummy atom and the atom attached to it (the SMD atom) is calculated as

$$\vec{F} = -\nabla U, \tag{2.8}$$

where the potential energy U is given by

$$U = \frac{1}{2}k[vt - (\vec{r} - \vec{r}_0) \cdot \vec{n}]^2 \tag{2.9}$$

for actual (at time t) and initial positions of SMD atom $\vec{r}$ and $\vec{r}_0$ respectively, with pulling velocity v in direction $\vec{n}$, given a spring constant k joining the dummy and SMD atoms.

Evidently, the enforced movement of a molecule to a desired position is not an equilibrium process, but by means of Jarzynski's equality,[88] the free energy difference can be calculated as the exponential average of the non-equilibrium work:

$$e^{-\beta \Delta F} = \langle e^{-\beta W} \rangle, \tag{2.10}$$

where β is $\frac{1}{k_B T}$ for temperature T, ΔF the Helmholtz free energy (although is equally valid for the Gibbs free energy ΔG in the NPT ensemble), and W the non-equilibrium work.

Through an analysis carried out by Park *et al.*[89] the potential of mean force (PMF), the free energy as a function of the order parameter (λ), can be calculated by

$$\Phi(\lambda_t) = \Phi(\lambda_0) - \frac{1}{\beta} \log\langle \exp(-\beta W_{0 \to t}) \rangle, \tag{2.11}$$

where Φ is the PMF, and $W_{0 \to t}$ is the non-equilibrium work from time 0 to t. The second term can be expanded in terms of cumulants as (shown to the second order)

$$\log\langle \exp(-\beta W_{0 \to t}) \rangle = -\beta \langle W \rangle + \frac{\beta}{2}(\langle W^2 \rangle - \langle W \rangle^2), \tag{2.12}$$

which may improve accuracy under certain circumstances.

2.2.3 ALCHEMICAL FREE ENERGY PERTURBATION

Alchemical free energy perturbation (FEP) is a common method for calculating ligand binding free energies, whereby the free energy difference between initial and final states of interest can be obtained via mutation of one into the other.

In this process, the free energy difference between two states, a and b, $\Delta G_{a\to b}$, can be calculated simply as the ensemble average of the energy relative to state a:

$$\Delta G_{a\to b} = -k_B T \ \ln \left\langle \exp\left[-\frac{\mathscr{H}_b(\vec{r},\vec{p}) - \mathscr{H}_a(\vec{r},\vec{p})}{k_B T} \right] \right\rangle_a \tag{2.13}$$

where k_B is the Boltzmann constant, T temperature in kelvin, and $\mathscr{H}_n$ the Hamiltonian of the initial ($n = a$) or final ($n = b$) states with position $\vec{r}$ and momentum $\vec{p}$.

It is necessary for the accuracy of this calculation that states a and b do not differ significantly, as the energy of state b should be readily obtainable from the starting point a, that is, have substantial phase space overlap (Fig. 2.3). In practice, this condition is not often met in systems of interest. However, it can be artificially achieved through the addition of n further states, each assigned a value of λ, which bridge the gap between the initial and final states, effectively having partial character of both a and b. Given that free energy is a state function, these bridging structures do not have to be physical and the free energy does not depend on how they are defined. The free energy is then calculated as the sum of the free energies for each window, that is, between λ_k and λ_{k+1}. This introduces a dependency of the Hamiltonian on λ and the ensemble average in equation 2.13 is for the system Hamiltonian specified by λ_k. Equation 2.13 can be modified to allow free energy calculation from a stratified simulation to yield

$$\Delta G_{a\to b} = -k_B T \sum_{k=1}^{N} \ln \left\langle \exp\left[-\frac{\mathscr{H}(\vec{r},\vec{p};\lambda_{k+1}) - \mathscr{H}(\vec{r},\vec{p};\lambda_k)}{k_B T} \right] \right\rangle_k \tag{2.14}$$

where k ranges from 1 to N, the number of desired windows. Within each window, a period of equilibration must be carried out, to allow the system to equilibrate with the window conditions, followed by a period of data collection from which the free energy is calculated, which must be long enough to sample all relevant conformations.

Approaching the endpoint of a simulation involves interactions of atoms with their environment becoming very small, yet still being non-zero. At these points, as repulsive interactions are minimal, atoms can easily overlap with their environment. Given that at low inter-atomic distance, the LJ potential experiences a singularity, this can introduce many problems into the calculation of the free energy. Fortunately, this is easily circumvented by the inclusion of a soft-core potential[90,91] implemented in NAMD in the form

$$U_{LJ}(\lambda, r_{ij}) = \lambda \varepsilon_{ij} \left[\left(\frac{R_{ij}^{\min 2}}{r_{ij}^2} + \delta(1-\lambda) \right)^6 - \left(\frac{R_{ij}^{\min 2}}{r_{ij}^2 + \delta(1-\lambda)} \right)^3 \right] \tag{2.15}$$

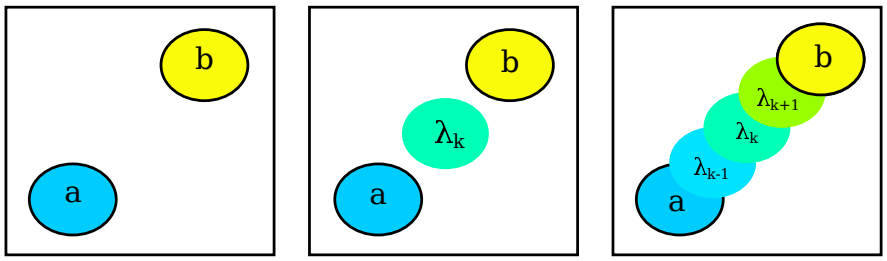


Fig. 2.3: If the overlap between states a and b is poor (left), the calculated free energy will be inaccurate. Additional windows can be added (center, right) but there must be enough to sufficiently bridge the gap between a and b (right). The better the overlap, the better the free energy estimate.

for the minimum of the potential $-\varepsilon$ at distance $R_{ij}^{\min}$. This shifts the square of the radius, r^2, of the LJ potential to $r^2 + \delta(1-\lambda)$, preventing the singularity at $r = 0$. Here, the soft core potential was implemented with a radius shifting coefficient (δ) of 6.

For FEP calculations carried out in this work, the TFO was completely decoupled from its environment while both bound to the duplex and free in solution, giving the thermodynamic cycle shown in Fig. 2.4. Here, the process of interest, ΔG_{dissoc}, was calculated as the sum of $\Delta G_{\text{annihil,1}}$ (TFO bound), $\Delta G_{\text{annihil,2}}$ (TFO free), and ΔG_{rest}, which accounts for the contribution of restraints on the TFO, introduced to keep it in its binding site when non-interacting. In both bound and free cases, the difference between the initial ($\lambda = 0$, interacting with the environment) and final ($\lambda = 1$, non-interacting) structures was such that 16 bridging windows were required ($\Delta\lambda = 0.0625$). Throughout the simulation, vdW interactions of the TFO were linearly scaled to zero from $\lambda = 0$ to 1 while its electrostatics were turned of in the same way between $\lambda = 0$ and 0.5.

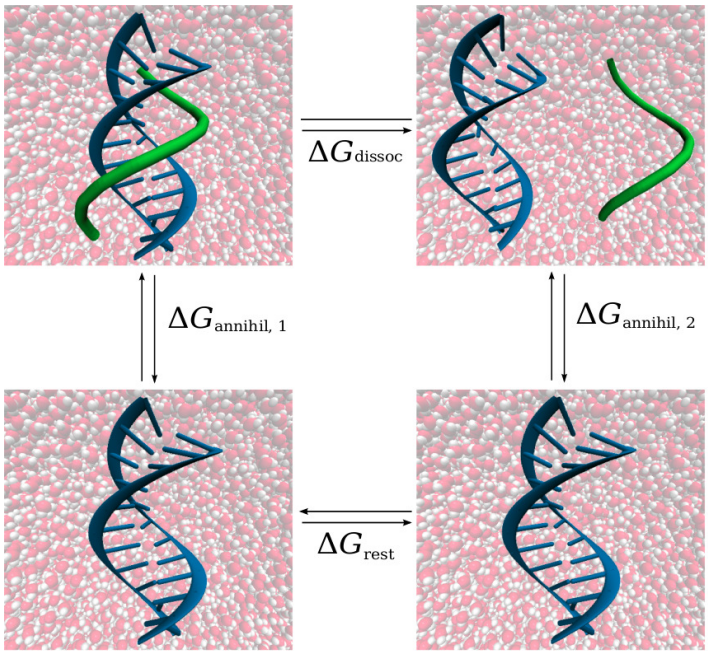


Fig. 2.4: Thermodynamic cycle for the calculation of TFO binding free energy. The top process, ΔG_{dissoc}, is the process of interest

As the conversion from state a to state b is an equilibrium process, performing the simulation

in the forward ($a \rightarrow b$) and reverse ($b \rightarrow a$) directions should give the same result. It is therefore generally thought of as best practice to, where possible, sample from both ends of the simulation. A good degree of overlap between the forward and reverse probability distributions is indicative of a well converged, and accurate, simulation.[92] Unfortunately, as the 15-base TFO studied here is very flexible, it is almost impossible to 'create' it in the appropriate binding site, even with the application of constraints. Hence, in the study of the dissociation of FRDA triplex, only the reverse (decoupling) process is considered.

RELATIVE BINDING FREE ENERGY CALCULATIONS

While FEP is useful for the calculation of absolute binding free energies, as presented above, it can also be used for the calculation of relative binding free energies through the mutation of one system to another. The principles of the method are identical, with the only difference being that the start and end points are defined simultaneously in the structure but interactions calculated such that the initial and final structures never interact with one another. Over the course of a simulation, the interactions of the the atoms representative of the starting structure are turned off, while those of the atoms representative of the final state turned on.[93] This effectively results in one chemical structure being mutated into another. As described above, this can be conducted in both forwards and reverse directions to improve accuracy.

In this work, this method is applied to the study of the relative stabilities of purine DNA and ribonucleic acid (RNA) TFOs. In one cycle, the deoxyribose sugars in the DNA TFO were mutated into ribose sugars (Fig. 2.5). Again, this is conducted in the bound and free states, so that the relative binding free energies can be measured. For the mutations outlined here, the changes result in a slight modification of every charge in the molecule, not just the atoms which are of a different type. The most rigorous method would therefore be to mutate every single atom in the TFO to match either the DNA or RNA charges as appropriate, but this is computationally impractical and the majority of charges only change by less than 0.1. Instead, only the atoms whose charges which display significant deviation from the DNA charges are mutated. In this work, the C2' and C3' atoms were mutated, along with their associated hydrogens.

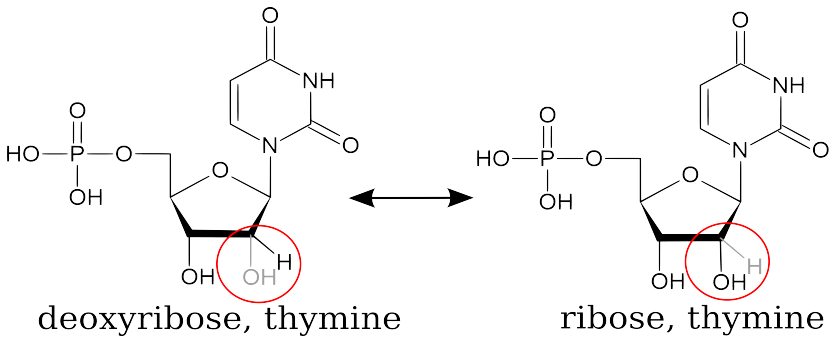


Fig. 2.5: Structures for mutation from RNA to DNA within the FEP algorithm are defined with the atoms to appear (grey) and disappear (same position, black) all present, but non-interacting.

Overall, 5 atoms per deoxyribose base were 'turned off' while 6 atoms per ribose base were 'turned on', giving only a change of 11 atoms per base (or 55 atoms for a 15-base TFO), significantly fewer than the previously described calculations. Accordingly, less simulation time was required for convergence. Using 64 windows ($\Delta\lambda = 0.015625$), 375,000 steps (0.75 ns) of simulation were conducted at each window, of which 125,000 (0.25 ns) were

equilibration. This gave a total simulation time of 48 ns and gave reasonable overlap between the forwards and reverse transformations as well as little variation with time indicating convergence.

2.2.4 Replica Exchange Molecular Dynamics

Replica exchange molecular dynamics (REMD) is an accelerated sampling technique whereby N non-interacting copies (replicas) of the system, each at a different temperature, are simulated in parallel. In order to enhance sampling the replica configurations are swapped between temperatures with an acceptance probability p_{ij} for a swap between states i and j, in accordance with the metropolis algorithm

$$p_{ij} = \min\left(1, \frac{\exp\left(-\frac{E_j}{kT_i} - \frac{E_i}{kT_j}\right)}{\exp\left(-\frac{E_i}{kT_i} - \frac{E_j}{kT_j}\right)}\right). \tag{2.16}$$

This allows low energy, or low temperature, configurations to access the normally inaccessible higher energy configurations, greatly improving the sampling. This method is commonly used to accelerate protein folding[94] as it prevents the structure becoming trapped in a local minimum, but is equally useful for the study of DNA denaturation.[95]

Here, this method was used to obtain the hydrogen bond fraction between TFO and duplex as a function of temperature, from which the melting temperature (T_m) can be determined as the temperature with hydrogen bond fraction of 0.5.

2.2.5 Umbrella Sampling

The previous methods have outlined ways to calculate the free energy difference between two states, or the melting temperature, of a system. An additional method is also used in this work, which is particularly useful for the calculation of PMFs when the two states of interest are separated by a high energy barrier (Fig. 2.6). Umbrella sampling (US) functions by introducing a so-called 'umbrella potential', which effectively biases the coordinate of interest to stay within a specific region of the free energy landscape. Many of these biasing potentials can be applied at different values of the coordinate to sample the entire free energy landscape. These potentials usually have a harmonic form and are defined by

$$U(q) = \frac{1}{2}k(q - q_0)^2 \tag{2.17}$$

where $q - q_0$ is the displacement of the system's coordinate q from the origin q_0 of the harmonic biasing potential, and k is the spring constant. Clearly, as q moves further from q_0, the energy, $U(q)$, will dramatically increase, disfavouring the system visiting these regions. This method is easily applied to the analysis of sugar puckering in nucleic acids, where the pseudorotational phase and amplitude can both be biased to obtain a free energy landscape of this process.

Assuming enough different positions are biased to give good overlap between regions and cover the entire range of order parameters between initial and final states, a PMF can be calculated. As the results obtained are biased, corrected (unbiased) values must first be obtained. This is generally achieved through the use of the weighted histogram analysis

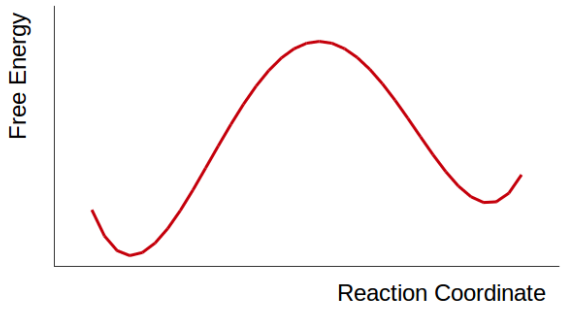


Fig. 2.6: Free energy as a function of reaction coordinate. There are two low energy wells separated by a high energy barrier, which is unlikely to be overcome through regular MD

method (WHAM).[96] From here, it is a simple matter of plotting the free energy, calculated by this process, as a function of reaction coordinate. It should be noted that this process is not limited to one-dimensional PMFs, but rather if two coordinates are biased, as is the case in the work presented here, a free energy surface can be calculated.

In the umbrella sampling simulations of the DNA/RNA sugar puckers, pseudorotational phase and amplitude were biased. The phase, P_θ was biased at every 0.2 rad between 0 and π, with a force constant of 335 kcal/mol/rad^2, so as to sample only the more relevant eastern hemisphere, that being the right half of the pseudorotational phase diagram in Fig. 2.2. Amplitude, A_r was biased every 0.1 rad between 0.6 and 1 with force constant 191.4 kcal/mol/rad^2.

The Weighted Histogram Analysis Method

In this work, US was used to calculate the free energy difference between the C2'-endo and C3'-endo sugar puckers of single stranded DNA and RNA through the biasing of the pseudorotational phase and amplitude. In order to account for the effect of the biasing potential, WHAM was used,[96] in this case, the implementation of Grossfield.[97]

When simulations are biased, such as in a process like US, the probability distribution shifts to be centered around the bias potential. That is, a biasing potential centered on a certain value will result in that area being sampled more heavily than anywhere else. In order for a free energy surface to then be constructed from a series of simulations sampling at specifically chosen regions at which a biasing potential is centered, it is first necessary to remove the effect of the bias, giving an estimate for the unbiased probability distribution.

In WHAM, a biased probability distribution at each biasing position, $P'(x)$ is calculated. The free energy, $A(x)$ can then be estimated as

$$A(x) = -k_B T \ln(P'(x)) - U'(x) + F \tag{2.18}$$

for umbrella potential $U'(x)$ and unknown constant F. By combining multiple simulations with umbrella potentials at different positions, the optimal F values can be calculated, and a good estimate of the unbiased probability distribution, $P(x)$, can be obtained. This is done

by iteratively solving the following equations until self consistent:

$$P(x) = \frac{\sum_{i=0}^{N}(x)}{\sum_{i=1}^{N} N_i \exp\left(\left[F_i - -U_i'(x)\right]/k_B T\right)} \tag{2.19a}$$

$$F_i = -k_B T \ln \sum_{X_{bins}} P(x) \exp[-U_i'(x)/k_B T]. \tag{2.19b}$$

for N_{sims} simulations and $n_i(x)$ counts in each histogram bin.

For this process to be effective, good overlap of the histograms is required to ensure adequate sampling of the entire range of order parameters. Better overlap can be achieved by increasing the number of biasing potentials used to cover the region of interest as it will give more numerous histograms. This will, however, increase the number of simulations required.

2.2.6 Error estimates

Steered Molecular Dynamics

The SMD simulation results presented in this work are based on the average value of 10 trajectories. Reported errors were approximated from the spread of the ΔG values calculated for each trajectory.

Free energy perturbation

The free energy was calculated in the stratified FEP methodology as the sum of the free energy changes of each window (Eq. 2.14). Within each window, the error in ΔG can be approximated using a block averaging method, which accounts for correlations in the data in estimating the error. The free energy was calculated (according to Eq. 2.14) for different sized blocks of data. The standard deviation of the block averages (free energies) for the different sized blocks was calculated, from which the standard error was determined. With increasing block size, the calculated standard deviation should increase before plateauing to its 'true' value. All error estimates are here reported as 2 standard errors. Further details on the block averaging method used here can be found in a number of standard texts.[64,98]

CHAPTER

THREE

TRIPLEX STABILITY AND DESTABILISATION BY NETROPSIN

Molecular dynamics simulations are used to study the structure and thermodynamics of deoxyribonucleic acid (DNA) triplexes associated with the neurodegenerative disease Friedreich's ataxia, as well as complexes of these triplexes with the small molecule netropsin, which is known to destabilise triplexes while stabilising duplexes. We find that binding of a third DNA strand (triplex–forming oligomer (TFO)) to a DNA duplex widens the major groove, while netropsin binding narrows the minor groove in the region where it is bound, an effect that has previously been under contention. The dissociation free energy, ΔG_{dissoc}, of a 15–base antiparallel-purine TFO is found to be 27.9 $\pm$ 9.5 kcal/mol, slightly higher than previously reported experimentally but remarkably close given that the simulation model was parametrised for double–stranded nucleic acid structure, not thermodynamics of triplexes. In the presence of netropsin, the calculated ΔG_{dissoc} is approximately 15 kcal/mol lower than without it, indicating that netropsin destabilises this triplex, as expected. Interestingly, destabilisation is found to be a highly localised phenomenon, occurring only when netropsin is bound directly opposite the TFO. Netropsin binding in the Crick–Hoogsteen (C–H) or Watson–Hoosteen (W–H) grooves is shown to be disfavoured relative to in the minor groove due primarily to differences in desolvation and groove width.

3.1 INTRODUCTION

As outlined in chapter 1, non-canonical DNA structures, such as the DNA triple helix, can act as cellular regulators.[9–11] Their formation outside of normal circumstances can therefore lead to disease, as typified by neurodegenerative disorder Friedreich's ataxia (FRDA). It is generally accepted that this disease has its roots in the formation of an intramolecular triple helix, which results in significant reduction in gene expression.[16] Accordingly, the ability to destabilise such structures is therapeutically promising as a means to treat this disorder.

A potential means to achieve this destabilisation is through the binding of ligands such as netropsin which bind in the minor groove (Fig. 3.1). Minor–groove binding ligands, such as netropsin, have been shown to destabilise triplexes,[25,29] and present a promising starting point for the study of triplex destabilisation mechanisms, which currently remain unknown. The work presented here provides the first mechanistic understanding of the dissociation free energies of triplexes in the presence of such molecules, as a basis for further development of triplex destabilisers in applications of gene regulation.

Previous studies on netropsin binding in the minor groove have reported that it affects the width of this groove, though whether it increases or decreases it appears to depend on the method used. Previous computational studies on the binding of netropsin to a double helix have shown a decrease in minor–groove width,[35] while solid–state experimental structures have indicated that it increase the width.[33,34] Further study in this area is therefore warranted and this work should aid in elucidating, and understanding, the structural changes that occur on netropsin binding.

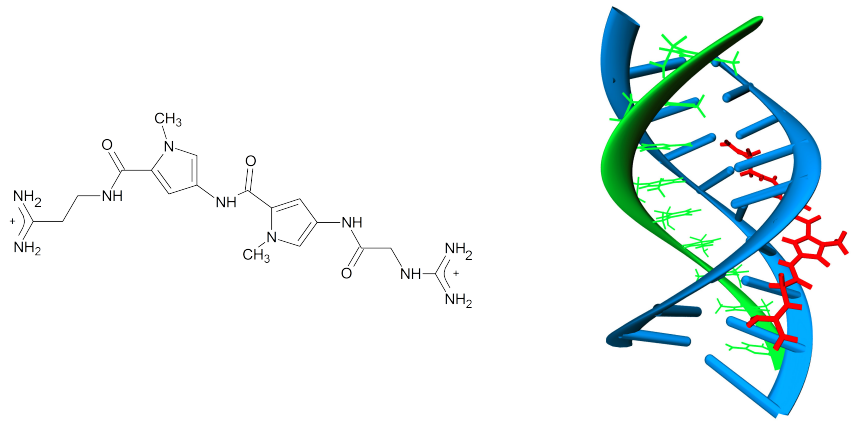


Fig. 3.1: Netropsin (red, stick representation) binds, in its +2 protonation state, in the minor groove of DNA (blue, ribbon representation). The TFO (green, ribbon + stick representation) binds in the major groove.

The choice of groove for netropsin binding is also worth considering. While it is known that netropsin binds in the minor groove of DNA, and the reasons for its stability here have been long understood,[39–41] the reasons for its total lack of binding in the other two grooves, the C–H and W–H, are unclear. A number of factors relating to the stability in either of the three grooves can be examined. Firstly, the negatively charged phosphates create a large electrostatic potential, accentuated by the narrowness of the minor groove for AT-rich regions, attracting the positively charged ligand, which then fits tightly in the appropriately sized minor groove.[40] Furthermore, hydrogen bonding interactions with A and T pairs stabilise the structure,[39] and the displacement of water from the nucleic acid's spine of hydration

introduces an entropic effect which favours binding.[41] The extent to which these properties change on binding of netropsin in either the C–H or W–H grooves should therefore provide some insight into the molecule's selectivity, with results potentially generalisable to other structures.

Molecular simulation provides a means of gaining an atomistic understanding of the processes that occur on the binding of such ligands, typified by netropsin, in the minor groove. For the theoretical study of DNA structure and thermodynamics, CHARMM36[68] and AMBER parm99bsc0[69,70] are the most common force fields currently in use. Both have been parametrised using quantum chemical data to accurately reproduce the experimental structure of the BDNA duplex. The AMBER force field in particular has been shown to accurately represent the structure of ZDNA, DNA triplexes and quadruplexes, and DNA:RNA hybrids on timescales up to 10 ns.[70] Few studies have looked at the accuracy of thermodynamics calculated using this force field. Additionally, methods such as free energy perturbation (FEP) are more commonly used for calculation of the binding free energy of small ligands or upon small mutations and may not be feasible for larger changes required for the study of triplex thermodynamics. It is therefore an important step to determine the feasibility of these calculations for the large, flexible systems presented here.

The structure of the antiparallel polypurine triplex of interest in Friedreich's ataxia has been well characterised through both simulation[20,21] and experiment.[22,24] Quantitatively, the thermodynamic stability of this triplex has been experimentally determined for a 15-base TFO under physiological conditions.[61] No computational studies, which would give an atomistic understanding of the binding or unbinding processes, have quantified the stability of the FRDA triplex, either in the absence or presence of netropsin. Whilst previous experimental studies have shown that netropsin binds with sequence specificity to DNA triple helices, destabilising them,[25,26,32] a clear mechanism of how it functions in doing so has not been determined and the extent to which destabilisation is a local, or global, effect is also unknown. Furthermore, as previous studies on netropsin have focused primarily on the pyrimidine triplex, it is not clear whether netropsin is a useful starting point for development of destabilisers of the specific GAA repeat (purine) triplex of interest in Friedreich's ataxia.[25,26]

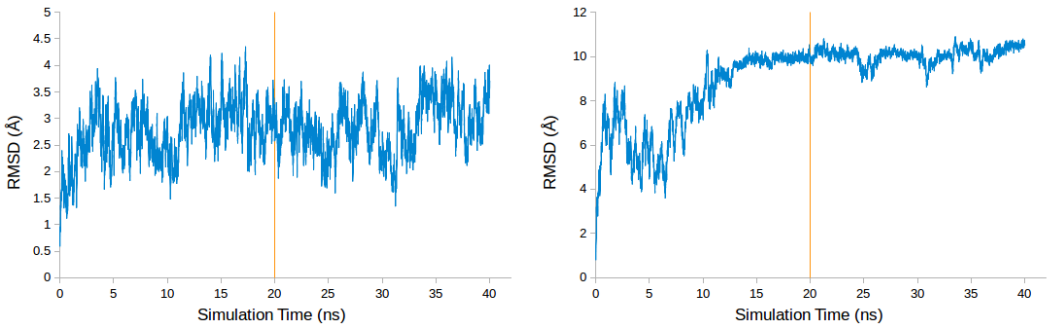
Here we examine the ability of free energy methods implemented through atomistic molecular dynamics to accurately predict thermodynamic and structural properties of the DNA triple helix and to provide molecular-scale insight into the mechanism of triplex destabilisation by a model ligand. Through comparison of experimental and theoretical results, we aim to determine whether the currently available methods are sufficient for the quantitative analysis of the structure and thermodynamics of triple helices on reasonable timescales. In addition, the structural and thermodynamic changes on netropsin binding are qualitatively examined and the reasons for its binding preference of the minor groove are investigated.

3.2 COMPUTATIONAL METHODS

Molecular dynamics (MD) simulations were conducted on a 19 base-pair $d(GC(GAA)_5GC)_2$ duplex with and without a $(GAA)_5$ TFO bound antiparallel to the central purine strand. The resultant DNA triplex represents a short segment of the triplex sequence responsible for Friedreich's ataxia. For each TFO length, two different systems were constructed: one with one molecule of netropsin bound and the other with no netropsin. In the absence of X-ray structures for the desired compounds, the simulation was started from approxi-

mate structures, with the TFO placed in close proximity to the appropriate bases of the duplex. It was then simulated until such a time as the structure ceased to change significantly, prior to further equilibration as described below. All simulations were carried out using NAMD 2.11[63] using the AMBER parm99 force field[69] with the BSC0 modification (parm99bsc0).[70] Netropsin parameters were calculated using the antechamber program in AmberTools,[99,100] using the AM1 Hamiltonian[101] then converted to the CHARMM format. The parameters are available in Appendix A.

For explicit solvent simulations, the system was solvated with approximately 15,000 TIP3P water molecules to give a (80 Å)^3 cubic solvation box, and neutralised with Na^+. 48 Na^+ ions were required to neutralise the triplex in the presence of netropsin, and 50 without it. These were added using VMD's cionize plugin, which places ions at the points of minimum energy around the nucleic acid.[80] The NaCl concentration was set to 0.15 M, requiring an additional 43 of both Na^+ and Cl^- ions. Long-range interactions were evaluated by means of the Particle-Mesh Ewald (PME) method. Bonds involving hydrogen were constrained using SHAKE,[67] allowing for a timestep of 2 fs. Simulations were conducted in the NPT ensemble, with temperature and pressure maintained at 310 K and 1 atm, respectively, using a Langevin thermostat and barostat. The positions of all triplex and ligand atoms were fixed while positions of ions and water were minimised, then allowed to equilibrate around the restrained solute for 20 ps. Restraints on the triplex were released and the system equilibrated for 40 ns. The structures appeared stable in terms of root-mean-square deviation (RMSD), temperature, and energies after approximately 10–15 ns (Fig. 3.2), and the final 20 ns of these simulations was used for structural calculations of the triplexes.



(a) Bound TFO, no netropsin

(b) Unbound TFO

Fig. 3.2: RMSD of nucleic acid atoms over the 40 ns equilibration period for a triplex, and unbound single stranded nucleic acid. Orange line indicates the 20 ns equilibration period after which the data was used for structural analysis.

For implicit solvent calculations, the Generalised Born implicit solvent (GBIS) model was used with a solvent dielectric of 78.5 and ion concentration of 0.15 M. All structures were deemed to have reached equilibrium, through analysis of RMSD and energies, by 10 ns of simulation (Fig. 3.3). All other parameters were the same as for the explicit solvent.

Molecular visualisation was carried out with VMD.[80]

3.2.1 Free energy calculations

Free Energy Perturbation (FEP)

FEP calculations were carried out in explicit solvent for triplexes for the 15-base TFO, with

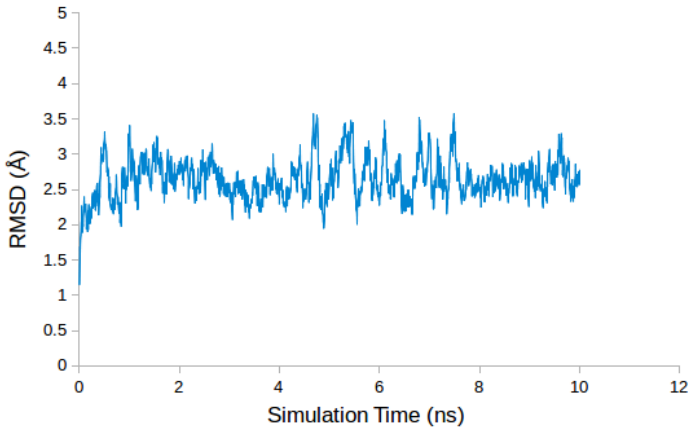


Fig. 3.3: RMSD of a DNA triplex in implicit solvent over a 10 ns equilibration period.

netropsin bound in the minor groove in the center of the triplex, or absent. Simulations were set up as described above. The simulation was divided into 16 windows with 5 ns of equilibration in each window prior to 7 ns of data collection. Windows were spaced unevenly to give more sampling over the first half of the simulation where greater free energy changes occur. Values of λ chosen were 0.04, 0.08, 0.12, 0.16, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.60, 0.70, 0.80, 0.90, and 1. Electrostatics were decoupled linearly over the λ range 0–0.5, and vdW from 0 to 1, where $\lambda = 0$ is fully interacting and $\lambda = 1$ is completely decoupled. This gave a total simulation time of 192 ns per netropsin position.

The thermodynamic cycle used is outlined in Fig. 3.4. The flexible, helical nature of the TFO means that the coupling transformations in which the TFOs interactions with the environment are gradually introduced are not possible on a reasonable timescale. As these begin with an unwound TFO, the time required for it to spontaneously achieve the required conformation for binding is prohibitive. Accordingly, only the forward (decoupling) transformations were considered. Restraints were applied in order to maintain the position of the TFO in the binding site as it was decoupled from the environment. The center-of-mass distance between the TFO and duplex was constrained to its average equilibrium value with a force constant of 10 kcal/mol/Å^2. The effect of these constraints on the binding free energy was accounted for as ΔG_{rest}, calculated according to the method outlined by Wang *et al.* with the coupling parameter taking 11 evenly spaced values between 0 and 1.[102]

Shorter (25 ns) total FEP simulations were also conducted for the 3-base TFO with netropsin bound in two different positions relative to the TFO, either opposite or below it. These used 50 windows ($\Delta\lambda = 0.02$) with 100,000 steps (200 ps) of equilibration followed by 150,000 steps (300 ps) of data collection.

Replica Exchange Molecular Dynamics

Replica exchange MD simulations were conducted in implicit solvent to examine the melting behaviour of the 3-base and 15-base TFOs. For the 3-base TFO, 12 replicas at 10 K intervals between 250 and 350 K were used. Exchange acceptance ratios ranged between 0.21 and 0.32. Systems were equilibrated for 2.5 ns prior to attempting exchanges, after which each replica was at the desired temperature. This was followed by 10 ns of replica exchange. After approximately 5 ns, the number of hydrogen bonds between the TFO and duplex had plateaued at each temperature, suggesting the system had reached equilibrium. The melting

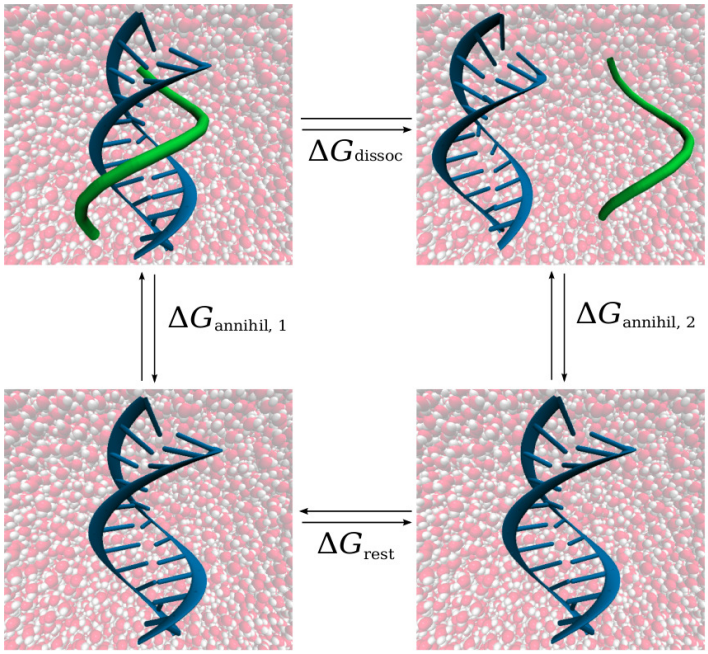


Fig. 3.4: Thermodynamic cycle for the binding free energy calculations of TFO (green, shown in tube representation) to duplex (blue, shown in ribbon representation).

curve was constructed from the final 4 ns of data. The 15-base TFO was studied with 8 replicas at 5 K intervals between 315 and 350 K, giving exchange acceptance ratios between 0.2 and 0.5. It was simulated for 28 ns. After approximately 25 ns the number of hydrogen bonds at each temperature had plateaued. The melting curve was constructed from the final 1.5 ns of data. In both cases, melting curves were calculated as the average number of hydrogen bonds over a 5 K temperature range. T_{m} was determined as the temperature where the hydrogen bond fraction was 0.5.

3.2.2 STRUCTURAL CALCULATIONS

To investigate the structural changes occurring upon netropsin and TFO binding, the final 20 ns of the 40 ns simulations of the 15 base triplex were examined. The minor groove width was measure as the average interstrand distance between the two phosphorus atoms at residues n_A of strand A (the purine duplex strand) and $n_B - 4$ of strand B (the pyrimidine duplex strand) (P_{n_A}–P_{n_B-4}). The major groove width was similarly defined as the average interstrand P_{n_A}–P_{n_B+5} distance.

The average number of hydrogen bonds over the same 20 ns period of simulation between the TFO and duplex was measured using the Hbonds plugin for VMD.[80] A donor-acceptor distance cut-off of 3.0 Å and an angle cut-off of 25° were employed.

The average structures of the two 15-base triplexes, calculated from the final 20 ns of the 40 ns simulation, were aligned and compared. The structures were positioned such that the long axis was aligned with the z axis and looking at the xy plane was looking down the center of the triplex cylinder. The coordinates of the aligned backbone phosphorous atoms were compared to determine whether netropsin binding had any effect on the overall width of the helix.

3.2.3 NETROPSIN GROOVE SELECTION

MD simulations were carried out with netropsin positioned in the C–H and W–H grooves of the 15-base triplex. 40 ns of simulation was carried out with the first 20 ns discarded as equilibration. Systems were set up as previously described for structural calculations of triplex with netropsin in the minor groove. The number of solvent molecules released on binding was calculated through comparison of the triplex with no netropsin, and with netropsin in the specified groove. Water molecules within 4 Å of the seven residues on two nucleic acid strands making up netropsin's binding site, but not closer than 4 Å to the third strand, were considered to be solvating the triplex. Each water molecule lost was assumed to contribute 1.6 cal/K/mol to the binding entropy of netropsin.[41] The binding of netropsin to a duplex has been shown to be associated with a change in configurational entropy of -30.4 cal/K/mol[86] which was also considered when calculating the contribution of desolvation to the overall stability of the netropsin–triplex complex.

Contact maps were generated by colouring triplex atoms by their proximity to netrospin to understand the interactions of the ligand with the groove. Any possible hydrogen bonding contacts were determined as potential hydrogen bonding triplex-netropsin atoms within 2 Å in the average structure of the triplex, calculated from the final 20 ns of simulation.

The interaction energies between netropsin and the triplex were calculated using MM/GBSA (molecular mechanics combined with generalised Born surface area).[103] From the 40 ns explicit solvent trajectory of netropsin in all three grooves, the last 20 ns was split into 4 ps snapshots and solvent and ions removed. Interaction energies were obtained between the triplex and netropsin (TN) for each of the netropsin positions. Similar calculations were conducted for a triplex with no netropsin (T) and netropsin with no triplex (N). The total energy, split into electrostatic and van der Waals components, was then calculated as the average of $\Delta E_{\mathrm{TN}} - (\Delta E_{\mathrm{T}} + \Delta E_{\mathrm{N}})$ over the 5000 trajectory snapshots. Calculations were conducted using NAMD 2.12[63] and the Generalised Born implicit solvent model[77] with solvent dielectric 78.5.

3.3 RESULTS AND DISCUSSION

3.3.1 FREE ENERGY CALCULATIONS

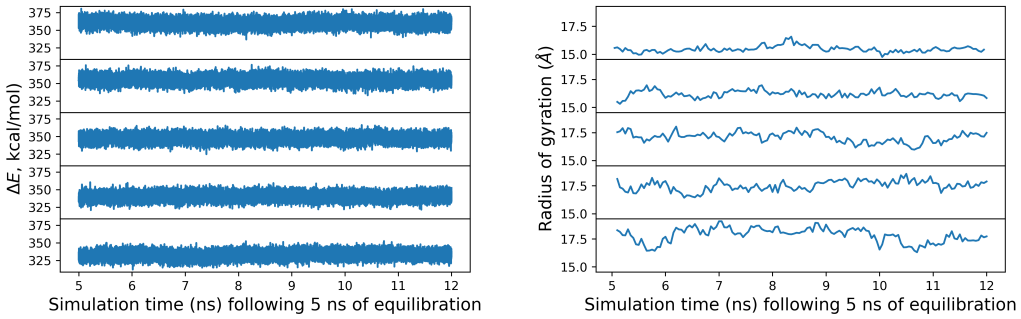
The FRDA triplex has been previously studied experimentally, and has been shown to have a $\Delta G_{\mathrm{dissoc}}$ of 12.3 kcal/mol for a sequence with the same length and under similar conditions as studied here.[61] This gives a good comparison for the FEP results obtained here, as a means of assessing the accuracy of the force field for calculating DNA binding thermodynamics. Calculations of binding free energies of the 15-base GAA repeat TFO associated with FRDA were conducted in explicit solvent without netropsin. A comparison of the literature results with FEP calculations is presented in Table 3.1. The result reported here is found to be in reasonable agreement with the literature value, although the simulated triplex does appear to be slightly overstabilised compared with experiment. However, even with long simulations of almost 200 ns in length (requiring almost 20,000 CPU hours for each of the bound and unbound simulations) error bars are still large, being approximately 40 % of the calculated free energy. The long simulation time was required as the system equilibrated slowly, requiring almost 5 ns equilibration time in each window. Additionally, structural fluctuations of the free TFOs appear to be on the order of 1–2 ns (Fig. 3.5b), indicating that to effectively sample these within each window, long simulation time is required. As such, further

increasing the simulation time may aid in improving the precision of the calculations. Despite this, as current force-fields are parametrised for structure rather than thermodynamics, the agreement found here between computational and experimental results is remarkable. It should also be noted that the conformation the unbound TFO adopts heavily influences the calculated binding free energy. The structure used in this case appears to be the equilibrium structure of the unbound TFO, so should give appropriate results, although conformational transitions seem to be very slow (occurring on a timescale longer than 80 ns). This will be discussed further in chapter 5.

Table 3.1: Comparison of ΔG_{dissoc} calculated by FEP with the literature (experimental) value.

Method	ΔG_{dissoc} (kcal/mol)
Literature[61]	12.3
FEP	27.9 ± 9.5

Examining the convergence of the simulations, within each window structural parameters and energies have generally plateaued (Fig. 3.5), indicating that the 5 ns equilibration time allowed for each window is sufficient to reach equilibrium, however structural fluctuations are still observed, occurring on the order of a couple nanoseconds. Additionally, as the free energy is calculated as an exponential average, this weights configurations with small energy differences between states (Eq. 2.14) much more heavily than large energy differences, causing drops in ΔG when a rare configuration with a small energy difference is sampled. These drops in ΔG take some time to be corrected, but can be monitored by examining the calculated ΔG with increasing simulation time. As can be seen in Fig. 3.6a, by the end of the 12 ns these small energy differences are not significantly affecting the overall free energy. This is a good indication that the system has been sampled for sufficient time within each window that a reasonably accurate result is obtainable. Despite this, increasing the simulation time to better sample these configurations with small energy differences, or increasing the number of windows to reduce the energy differences being sampled, will increase the precision of the result, and may also slightly improve the accuracy.



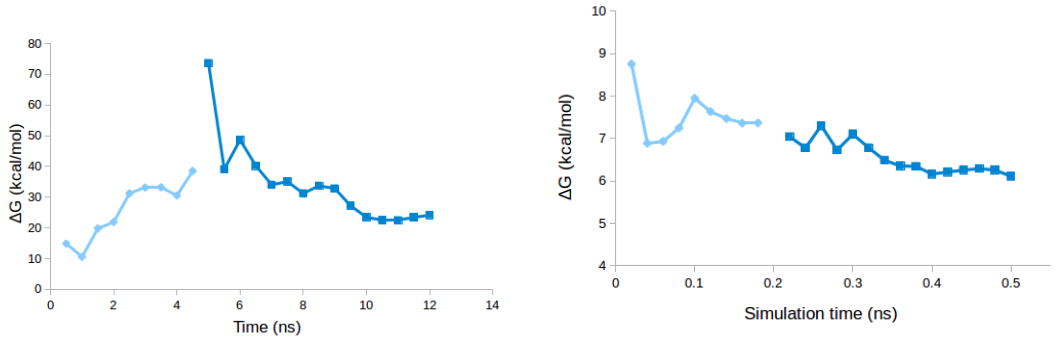
(a) ΔE (b) Radius of gyration

.

Fig. 3.5: Variation in ΔE, the energy difference between two adjacent FEP windows, and the radius of gyration, representative of the overall TFO structure, over the first 5 windows (top: first window, bottom: 5th window) of a FEP simulation. These are typical of all windows for all three DNA simulations (bound with netropsin, bound with no netropsin, and unbound).

In the case of the 3-base TFO, shorter simulations were possible as the free energy changes required to annihilate a shorter TFO are smaller, due to the smaller mutation required (of 3

bases rather than 15). Additionally, due to is significantly shorter length, the 3-base TFO does not undergo significant structural fluctuation like the 15-base TFO, again allowing for a shorter simulation time. Accordingly, these calculations converged much faster and the time-frame of 25 ns was deemed adequate to give accurate results (Fig. 3.6b).



(a) 192 ns simulation of the 15-base triplex (b) 25 ns simulation of the 3-base triplex.

Fig. 3.6: Convergence of FEP simulations of the 15- and 3-base triplexes. Each point is the value of the free energy calculated using data from each window up to the specified time. Light blue indicates the equilibration period, which is not included in the calculation of ΔG_{dissoc}.

While longer sampling time may increase the accuracy of these results relative to the literature, it has previously been shown that, in the study of DNA thermodynamics, base stacking free energies can depend on the choice of water model, with stability being overestimated in certain cases.[104] Using the same force field as in this work, it was found that TIP3P in particular resulted in overestimated stacking free energies, despite the force field being parametrized for this water model, which may contribute to the slight overstabilisation observed here. Furthermore, preliminary studies using recent modifications to the parm99 force field in BSC1[71] have shown a significant decrease in ΔG_{dissoc} relative to short calculations carried out with the BSC0 force field. The force fields differ in the description of certain dihedrals, which are modified in the BSC1 implementation to provide better structural agreement with experiment on longer time scales. Although further work is needed here, it appears that these slight structural modifications may improve estimates of DNA triplex thermodynamics.

As a comparison, melting curves calculated in implicit solvent by replica exchange molecular dynamics (REMD), where melting temperature (T_{m}) is the temperature with hydrogen bond fraction 0.5, found a melting temperature of approximately 332 K for the 15-base triplex, slightly higher than the 325.6 K experimental value from the literature.[61] Previous studies on short DNA duplexes have shown reasonable agreement with experiment in calculating melting temperatures in implicit solvent with the BSC0 force field with an average error of 9.9 K, with melting temperatures predominantly overestimated.[105] Although calculated using different MD methods (MM/GBSA rather than REMD) and different structures, these results indicate the accuracy achievable for melting temperature calculations with this force field and is comparable to the error found for the triplex in this work. To assess that the melting behaviour was following the expected trends, the melting curve of a 3-base TFO triplex was compared. It displayed a broader transition and a lower melting temperature than the 15-base TFO as expected (Fig. 3.7), although experimental data on the melting of this structure is not available.

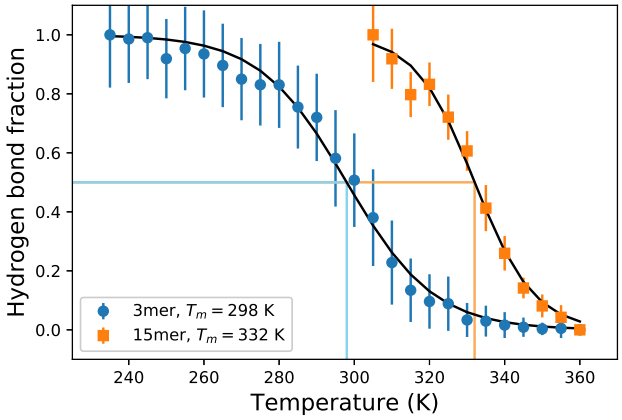


Fig. 3.7: Melting curves calculated from replica exchange MD simulations for the 3-base and 15-base TFOs. Simulations were carried out in implicit solvent. Straight lines indicate the melting temperature. Error bars indicate 2 standard errors in the H-bond fraction at each temperature.

3.3.2 Effects of netropsin on DNA triplex stability

Although slightly higher than the literature values without netropsin, the dissociation free energies for a 15-base TFO with and without netropsin can be compared. The presence of netropsin is shown to destabilise the triplex by approximately 15 $\pm$ 12 kcal/mol (Table 3.2). This is in line with previous reports of netropsin destabilising the triplex, although is the first reported value of ΔG, which although likely to be sequence dependent, helps to better understand the functionality of this molecule.

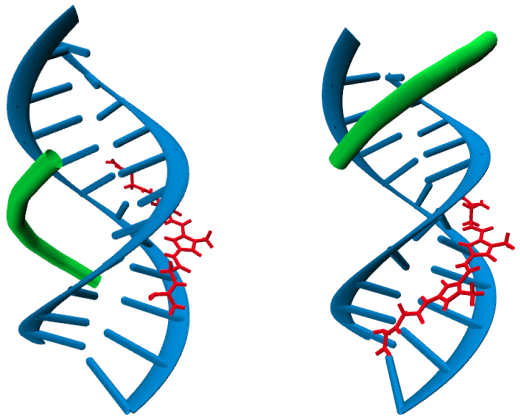
Table 3.2: ΔG_{dissoc} calculated by FEP for the dissociation of a 15-base TFO in the presence or absence of netropsin.

	ΔG_{dissoc} (kcal/mol)
No netropsin	27.9 $\pm$ 9.5
Netropsin	13.8 $\pm$ 8.5

Although clearly acting to destabilise the triplex as expected, netropsin's approximately 15 kcal/mol destabilisation is not sufficient to reduce the stability enough to prevent triplex formation (ie. reduce ΔG_{dissoc} to 0 or lower), and is likely to be indicative of a localised mode of action. Assuming that the binding of multiple netropsin molecules does not have a cooperative effect, approximately 2–3 ligands would be required for spontaneous decomposition of the 15-base triplex studied here into its corresponding duplex and single strands. Given netropsin's reported 7-base binding site in a triple helix,[25] saturation of the minor groove would be required to approach this, highlighting the need for better triplex destabilisers.

As it appears from the free energy calculations that netropsin exerts is destabilisation effects only on the local region where it is bound, the effect of the ligand's position relative to the TFO is of some interest. Free energy calculations, described above, have been carried out on a 3-base TFO bound to the same 19-base duplex with netropsin in two different positions in the minor groove. With a short TFO there are two broad descriptions of where netropsin can be located relative to the TFO: directly opposite it, or underneath it (Fig. 3.8). Free energy

calculations of the 3-base triplex, show very little difference in stability when netropsin is bound underneath the TFO, while significant destabilisation is reported when bound opposite it (Table 3.3).



(a) Opposite (b) Underneath

Fig. 3.8: Netropsin (red) bound opposite or underneath a 3-base TFO (green)

Table 3.3: ΔG_{dissoc} calculated by FEP for the dissociation of a 3-base TFO in the presence of netropsin in one of two positions, or absent. $\Delta\Delta G_{\text{dissoc}} = \Delta G_{\text{dissoc, no net}} - \Delta G_{\text{dissoc, net}}$

Netropsin position	$\Delta\Delta G_{\text{dissoc}}$ (kcal/mol)
Underneath TFO	3.8 ± 3.2
Opposite TFO	12.1 ± 3.1

The significant effect of netropsin's position on the stability of the triplex indicates that netropsin likely has a highly localised mode of action, suggesting that high concentrations of the molecule would be required for destabilisation of longer, biologically relevant, triplexes. This is qualitatively consistent with the destabilisation free energy reported above of a single netropsin molecule. The destabilisation of 15 kcal/mol for a 15-base triplex indicates that near saturation of the minor groove is required to completely destabilise a triplex of this length. Previous studies on minor groove binders, including netropsin, have shown that, for an 18-base triplex at a triplex:ligand ratio of 1:0.2, triplex can still be detected. Increasing the ratio to 1:1 eliminates the triplex, with only duplex and single strands remaining.[22] Given netropsin's 7 base binding site, this is again consistent with a local mode of action that would require near saturation of the minor groove to achieve significant stabilisation.

3.3.3 Structural effects of netropsin/TFO binding

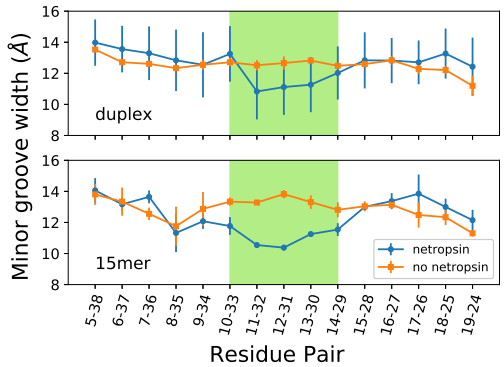
As it appears, based on the thermodynamics, that netropsin has a localised mode of action, the effect it has on structure is also potentially of note. An understanding of the interplay between the structure and thermodynamics is important for developing understanding of the mechanism of the drug's mode of action. The number of hydrogen bonds between the TFO and corresponding duplex strand, as well as the widths of the major and minor groove with and without netropsin, have been examined.

GROOVE WIDTHS

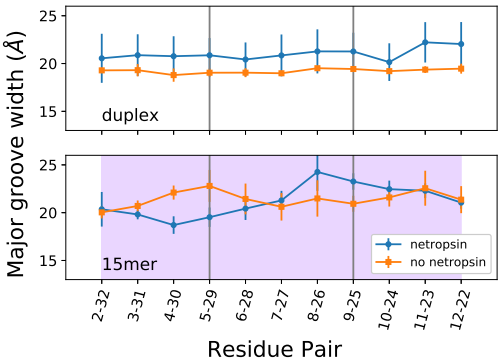
Experimentally, the binding of netropsin to a DNA duplex has been studied extensively through crystallography and been shown to widen the minor groove on the order of 0.5–2 Å.[33,34,37] On the other hand, computational studies of the netropsin–duplex complex in solution have shown that the binding of netropsin decreases the minor-groove width by 1–2 Å, particularly over the base pairs where the charged groups of netropsin are found.[35] It has previously been shown that the solvent environment influences the minor groove width[21,106] and, accordingly, it is possible that the solid-state structure measured in crystallographic studies differs from the aqueous structure studied by MD. It is hoped that the further MD studies presented here can clarify the effect of netropsin on the minor groove structure. The aforementioned studies focused on the structure of the netropsin–duplex complex, while the structure of the netropsin-triplex complex has not previously been examined in detail.

Here we find that, for a DNA triplex, the minor groove narrows by 2–4 Å in the region where netropsin is bound (Fig. 3.9a), in contrast to the reported solid–state experimental results for the duplex.[33,34,37] As a comparison, the minor groove of the duplex–netropsin complex was examined by the same method and found to narrow by a similar amount. Hence it is unlikely to be the presence of the TFO that induces this deviation from experiment. It has been shown through simulation that the interaction of atomic ions or charged molecular species with the minor groove can cause significant narrowing[35,107] which may explain the observed effect for positively charged netropsin. It therefore appears that the structure of DNA triplexes is strongly dependent on the environment and solid–state structures may not accurately describe all facets of the solution structure.

Examining the effect of the TFO on the minor groove, if is found that it causes narrowing in some regions, and widening in others (Fig. 3.9a). These changes do not appear to follow any particular trend although it is possible that this general distortion of the minor groove, combined with the narrowing caused by ligand binding, may play a role in the destabilisation of these structures by netropsin.



(a) Minor groove. Netropsin is present in the shaded region of the minor groove.



(b) Major groove. TFO is present in the shaded region of the major groove. Netropsin is opposite the region of major groove between grey lines.

Fig. 3.9: Groove widths of a DNA duplex (top) or triplex with 15-base TFO (bottom) with (blue, circles), or without (orange, squares) netropsin.

The major groove is known to widen in the region where the TFO is bound[38,108] and indeed this is observed (Fig. 3.9b). Relative to the duplex, the width of the major groove when a 15 base TFO is bound increases from approximately 19 Å to 21–24 Å. The binding of netropsin is not seen to have a systematic effect on the major groove width, although it does appear to distort it, removing some regularity. Comparing the major groove of the duplex and triplex, the triplex major groove is significantly less predictable, with changes in width over the whole structure. It appears, therefore, that TFO binding distorts the nucleic acid structure, both major and minor grooves, relative to the duplex. As netrospin binding seems to further distort the structure, it is possible that this is an important factor in its ability to destabilise triplexes.

As evidenced in Fig. 3.9a, the narrowing of the minor groove is highly localised to the 4–5 bases where netropsin is bound. It is therefore likely that this local structural effect underlies the localised effect of destabilisation described above. The combined effect of the TFO widening the major groove and distorting the minor while netropsin narrows the minor groove distorts the triplex, possibly decreasing the Hoogsteen bond contacts between TFO and duplex and resulting in its reported destabilisation.

Given that triplex destabilisation by structurally similar minor groove binders such as berenil and distamycin shows a similar dependence on triplex:ligand ratio to netropsin,[22] it is also possible that these ligands have similar modes of action, although further study of these and other similar ligands would be required to confirm this.

HYDROGEN BONDS

The number of hydrogen bonds between the TFO and central duplex strand with and without netropsin was compared for the 15-base triplex. In an ideal DNA triplex structure, each Hoogsteen base pair should contain two hydrogen bonds, giving a total of $2n = 30$ hydrogen bonds for the 15-base TFO. The results shown in table 3.4 indicate approximately a third of this number. Structurally, this is associated with a misalignment of adenine bases as the TFOs show some distortion in the base stacking, particularly at the ends. However, very little change is observed to occur in the number of hydrogen bonds between the central purine strand of the duplex and the TFO (Table 3.4). It is therefore unlikely that the narrowing of the minor groove caused by netropsin binding significantly reduces the hydrogen bonding of the triplex. A more likely explanation is that the binding of netropsin induces structural changes that disfavour the binding of the TFO, due to distortions of the major groove.

Table 3.4: Number of hydrogen bonds between duplex and triplex in the presence or absence of netropsin

	Number of hydrogen bonds
Netropsin	9.03 ± 0.49
No netropsin	8.56 ± 0.24

OVERALL HELICAL WIDTH

The overall helical width was compared for the two structures of the 15-base TFO, with and without netropsin. It was found that on netropsin binding, the TFO is pushed slightly out of the groove in the region opposite the end of the netropsin molecule (Fig. 3.10). Potentially, the local narrowing of the minor groove and the changes in major groove width disfavour

the TFO binding here, effectively squeezing it out. As this effect is observed solely opposite the ligand, it is in agreement with the free energy calculations that showed little to no destabilisation of the triplex when netropsin was not bound directly opposite it.

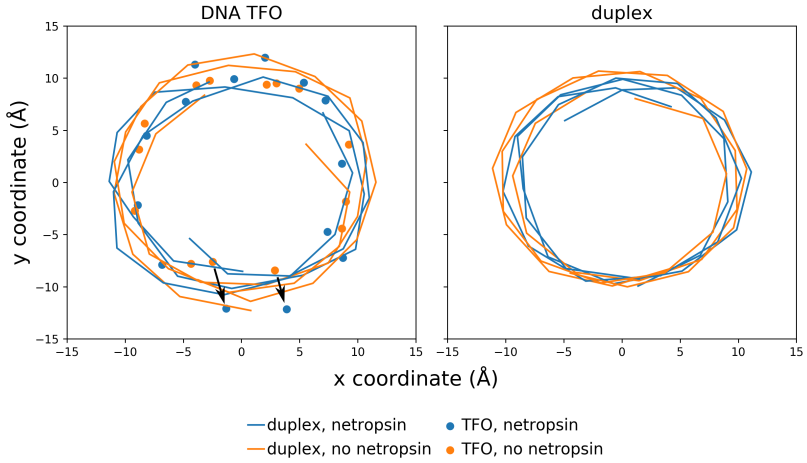


Fig. 3.10: The backbone structure of a triplex and duplex looking down the z axis. Binding of netropsin locally pushes the TFO out of the major groove. Backbone atoms of the duplex moiety are shown as a line, while TFO phosphorus atoms are shown as points. Arrows indicate the TFO bases where the greatest change is observed.

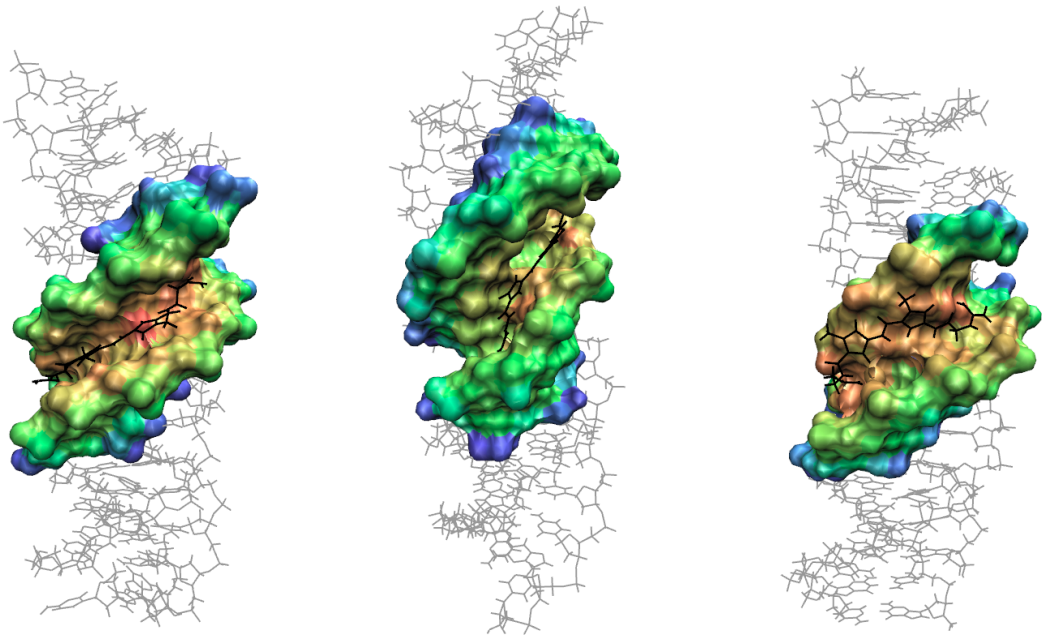
As it appears that sections of the TFO backbone moves 2–3 Å out of the groove when netropsin binds, it is surprising that no reduction in the number of hydrogen bonds between the duplex and TFO is observed. However, the bases that are furthest from their initial positions correspond to adenines, which already have very poor hydrogen bonding contacts (see chapter 4). Likely, although in a closer position previously, they were already out of the range used to define hydrogen bonds. It is also possible that, though some movement of guanine bases is likely, the distance moved did not put them out of the 3 Å range specified for hydrogen bonds.

3.3.4 NETROPSIN GROOVE SELECTION

As netropsin is well known to bind in the minor groove, only free energy calculations with it in this groove were conducted above. However, there are two other possible binding sites, the C–H and W–H grooves, and although the stability in the minor groove is known and understood,[39–41] it is not particularly clear why binding in the other grooves is so disfavoured that it has never been reported.

Examination of the contacts between the triplex and the netropsin molecule indicates that netropsin binds within the minor groove, and similarly within the larger W–H groove, but when placed in the smaller C–H groove it immediately moves to be positioned almost outside the groove, nearer to one chain (Fig. 3.11). From this analysis it is clear that netropsin binds deepest within the minor groove, interacting with the center of the duplex, likely the bases. When bound in the W–H groove, however, interactions are more localised to one chain and further towards the backbone of the triplex. This is likely due to the wider groove, which does not accommodate netropsin as well. In this case, interactions between triplex and netrospin can only readily occur with one triplex strand, rather than both, likely reducing its stability relative to binding in the minor groove. In the C–H groove, netropsin does not slide into

the groove, but rather is rotated such that it almost bridges the two chains at the edge of the groove, potentially due to unfavourable interactions with the closer walls of the narrower groove, effectively ejecting it.

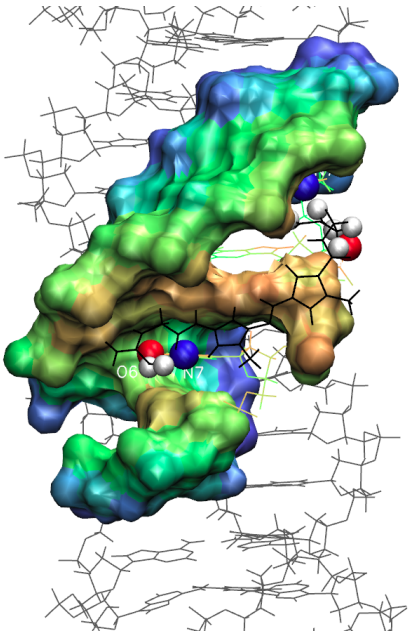


(a) minor groove (b) W–H groove (c) C–H groove

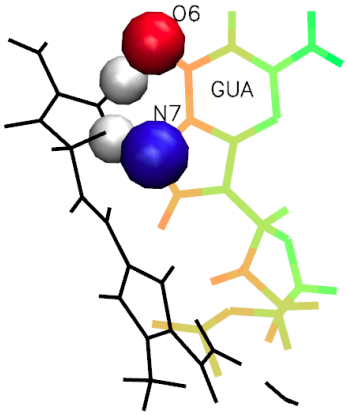
Fig. 3.11: Netrospin interaction within the three possible triplex grooves. Red indicates closer contacts while blue furthest away. Coloured surface shows all atoms within 10 Å of the netropsin molecule and are coloured to the same scale.

Despite this poor binding in the C–H groove, three possible hydrogen bond contacts are observed for the C–H–bound netropsin with the guanine residues located at either end of the ligand (Fig. 3.12). Such obvious contacts are not observed for the other two binding site. It is likely, therefore, that hydrogen bonding is not the main reason for netropsin preference for the minor groove.

Examining the energetic contributions of the electrostatic and vdW interactions between netropsin and the two strands of the groove it binds between, calculated in implicit solvent using the MM/GBSA method, the following trends are observed (Table 3.5). As these energies represent a net difference between the free and bound ligand, the positive electrostatic interactions, observed for binding in the minor and W–H grooves, indicate that it is more favourable electrostatically for the ligand to be in solution than bound to the triplex. This electrostatic component also includes the polar and non-polar contributions to the solvation free energy obtained from the MM/GBSA calculations. This unfavourable electrostatic contribution is therefore likely due to significant desolvation of the ligand required for binding. Nevertheless, purely based on electrostatic interactions, binding in the C–H groove should be most favourable, likely due to the narrowness of the groove, which has been shown to accentuate the electrostatic potential generated by the negatively charged backbone.[40] As the C–H is the narrowest groove, it is therefore not unexpected that it induces the strongest electrostatic attraction. Furthermore, the positively charged groups of the minor–groove binder are positioned closer to the negatively charged nucleic acid backbone when bound on the face of the triplex as observed for binding in the C–H groove. The interactions between



(a) Atoms within 2 Å of each other between netropsin and triplex in the C–H groove (vdW representation). Red: oxygen, blue: nitrogen, white: hydrogen.



(b) Interactions of netropsin with O6 and N7 of guanine residue at one terminus.

Fig. 3.12: Potential hydrogen bonding interactions between netropsin in the C–H groove and the triplex

these entities are therefore likely stronger than in either of the other cases where the ligand binds within the groove. A number of hydrogen bonds are also observed between netropsin in the C–H groove and the triplex itself (Fig. 3.12), which may also contribute to the more favourable electrostatics. The vdW interactions, on the other hand, decrease in strength from minor to W–H to C–H. In examining the positioning of netropsin in these grooves, it is able to form close vdW contacts with both of the walls of the groove when bound in the minor groove (Fig. 3.11). In the wider W–H groove it sits on one wall of the groove, interacting only with this, giving vdW interactions of only slightly greater than half that of in the minor groove. As previously noted, netropsin does not bind in the C–H groove, preferring to sit on the face of the triplex, thus further reducing the vdW interactions in this groove relative to the other two. Based purely on energetic differences between the triplexes, it appears that binding in the minor groove should be most favoured as expected (Table 3.5). However, as netropsin binding to the triplex is likely to result in the release of water, contributing a large entropic effect, this should also be considered.

As the ligand binds in a groove of the triplex, it should displace any solvent that binds there. In the binding of netropsin in the minor and W–H grooves, a significant change in solvation is observed, with netropsin displacing all the water molecules in the region where it is bound. On the other hand, binding in the C–H groove does not displace as many water molecules with solvent being observed between netropsin, on the face of the groove, and the interior of the triplex. Quantitatively, the number of solvent molecules released upon netropsin binding can be calculated, and is presented in table 3.5. This is calculated as the difference between the number of solvent molecules within 4 Å of the two chains between which netropsin is bound, considering only the seven bases of netrospin's binding site, and the same location in the native triplex. Additionally, water molecules within 4 Å of the third strand are excluded as these do not sit in the region where netropsin binds. As can be seen in Fig. 3.13 this gives good coverage of the desired region. Previous experimental studies have considered the desolvation of the entire triplex on netropsin binding and found that approximately 53 water molecules are displaced on ligand binding in the minor groove of a poly(dT)·poly(dA)·poly(dT) triplex.[41] This value assumes more than one netropsin molecule binding in the minor groove, where each netrospin molecule should be expected to cause the release of a number of water molecules so is not directly comparable to the case presented here. Here, only the water molecules displaced for the region of the 7 bases encompassing the single netropsin's binding site are considered. This leads to a loss of between 5–8 water molecules on netropsin binding in either of the three grooves (Table 3.5). The contribution to binding entropy is calculated assuming each solvent molecule lost contributes 1.6 cal.K^{-1}.mol^{-1}.[41] The effect of the loss of configurational entropy on netropsin binding (-30.4 cal/K/mol) previously calculated by MD simulation using the covariance matrix method is also included.[86,87] As previously observed qualitatively, the C–H groove is found to retain the most solvent, while the other two grooves experience more desolvation. The overall effect of the desolvation is that binding in the minor groove appears to be the most stable, although binding in the W–H groove is also quite favourable (Table 3.5).

The overall stabilities of netropsin in each of the three possible grooves is outlined in table 3.5 with an order of minor > W–H > C–H. However, large error bars are associated with these values and binding in either the minor and W–H grooves can be considered to be similarly stable. More statistics would therefore be required to increase the precision of these results to obtain a definitive answer for this question of stability. It should also be noted that the bonded interactions are not included in this case. As netropsin's interaction with the triplex

(a) Netropsin in the minor groove (b) No netropsin

Fig. 3.13: Water molecules solvating the minor groove according to the described definition. The seven residues which describe netropsin's binding site are shown in silver. Netropsin is shown in blue and water molecules in the vdW representation.

is purely non-bonded, the bonded interactions are not as important in exploring the stability difference between binding in the three grooves. Additionally, larger energy fluctuations are observed for the bonded interactions than the non-bonded and they therefore did not offer any additional insight. Longer simulations could potentially remedy this, although as it appears that the stability of binding in the W–H groove approaches that of the minor groove, more rigorous free energy calculations, using, for example, free energy perturbation methods, may be of more use to conclusively determine the preference of netropsin for binding in the three triplex grooves.

Overall, although less favoured electrostatically, binding in the minor groove is stabilised by strong vdW interactions that are possible with both walls of the groove. These interactions are not possible when bound in the C–H groove as the narrowness of this groove excludes netropsin, and only to a small extend with one wall of the W–H groove. A larger degree of desolvation in the minor and W–H grooves coupled with the more favourable vdW interactions therefore leads to binding being favoured in these grooves.

Only the purine motif triplex was studied here as this is the intramolecular triplex associated with Friedreich's ataxia, but it is possible that binding in the C–H and W–H grooves separated by a pyrimidine TFO may present different results, especially considering the reported potential for structural distortion of a purine third strand, which would influence the shapes of the grooves.[48] This may additionally explain the similar stabilities of netropsin in the minor and W–H grooves. If the W–H groove is narrowed in sections (see chapter 4), this would likely increase the strength of both the vdW and electrostatic interactions. For a better comprehension of the binding of netropsin to a triplex, different TFOs sequences should also be examined.

Table 3.5: Electrostatics and vdW contributions to the interaction energy between netropsin and its binding groove, combined with the entropic effect due to loss of solvent from netropsin's binding site. All values are in kcal/mol.

	minor	W–H	C–H
Electrostatic interaction energy	18.5 ± 9.9	6.25 ± 10.0	-4.2 ± 9.5
Van der Waals interaction energy	-81.2 ± 6.6	-52.1 ± 6.5	-33.5 ± 6.2
Elec + vdW	-62.7 ± 11.9	-45.8 ± 11.9	-37.7 ± 11.3
Water molecules lost	6.9 ± 0.6	8.0 ± 1.0	5.2 ± 0.6
$-T\Delta S$	6.0 ± 0.3	5.5 ± 0.5	6.9 ± 0.3
$\Delta G_{\text{non-bonded}}$	-56.7 ± 11.9	-40.3 ± 11.9	-30.8 ± 11.3

3.4 CONCLUSIONS

Dissociation free energy calculations were carried out on a 15-base triplex representative of that associated with Friedreich's ataxia. Calculated free energies, obtained using the FEP methodology, slightly overestimate experimental values from the literature, but are in reasonable agreement. It was found, however, that the DNA triplex structure is very slow to relax, and that simulation time has a significant effect on calculated free energies. Other methods of calculating the free energy, such as thermodynamic integration should be considered and give better accuracy in less time.

Binding of netropsin to the same triplex was associated with an 15 kcal/mol decrease in the dissociation free energy of the triplex, in line with previous reports of it destabilising the triplex, being the first quantitative report of this for the purine triplex. It was also shown that the position of the netropsin molecule relative to the TFO significantly affected the stability of the triplex, with netropsin binding below the TFO inducing effectively no destabilisation. Both of these results indicate a localised mode of action and imply that near saturation of the minor groove with ligand is required for destabilisation. As biologically relevant triplexes are expected to be on the order of hundreds of bases long, high concentrations of netropsin are likely to be necessary for effective destabilisation of these structures.

Structurally, the binding of netropsin induced a 2–3 Å decrease in the width of the minor groove, contrary to previous crystallographic results although in agreement with previous simulation work that focused only on binding to a DNA duplex. It is therefore likely that the triplex environment has a significant effect on its structure. This effect was highly localised to the 4–5 bases directly surrounding netropsin, providing a rationale for the similarly localised positional dependence of the free energy. The major groove was broadened around the TFO as expected and slight narrowing of the minor groove opposite the TFO was observed. The simulation data suggest that the structural distortion associated with netropsin binding, combined with that induced by the TFO, is at least partially responsible for its destabilisation ability.

Considering netropsin's preference for binding in the minor groove over the other grooves available to it, it was found that in all cases, structures were stabilised predominantly by vdW interactions and desolvation of the triplex. Binding in the minor groove was found to be the most stable, followed by W–H and C–H grooves. Binding in the C–H groove was shown to not occur, with the ligand instead attaching to the face of the groove, allowing it to remain solvated. The lower contribution to the binding entropy from a lower degree of desolvation

appears to have the most influence on the preference of netropsin for the minor over the other two grooves, although vdW interactions also play a large role here.

As minor groove binders have generally been found to destabilise DNA triplexes, the results presented here, namely the local narrowing of the minor groove, are likely generalisable to other structures of the same class. Although this requires further study of other structures, it is an important first step in understanding the mechanism by which of DNA triplexes are destabilised by minor groove binders and how these molecules may then act as gene regulators.

CHAPTER

FOUR

RELATIVE STABILITY OF PURINE AND PYRIMIDINE TRIPLEXES

The structures of antiparallel purine and parallel pyrimidine oligomers binding to the same sequence of a DNA duplex are examined through atomistic MD simulation. It is found that A*A reverse Hoogsteen pairs are likely significantly weaker than both their G*G counterparts, and the equivalent Hoogsteen pairs of A*T and G*C^+. This weakness is observed as a lack of hydrogen bonds forming between the triplex–forming oligomer (TFO) and duplex and a distortion of the ladder-like rung structure of the TFO. This causes significant distortion of the TFO backbone. While good contacts between guanine residues are maintained, adenine–containing regions of the purine TFO are pushed towards the centre of the major groove, giving the purine backbone a zig-zag appearance. No such distortions are observed for the pyrimidine equivalent. However, significant rearrangement of the pyrimidine TFO sugars is required for it to bind to the duplex, potentially adding a barrier to its formation. Additionally, a concerted rearrangement of these sugars is likely required, potentially inhibiting the formation of the pyrimidine triplex. Overall, assuming facile protonation of cytosines, or modifications to these bases to remove the necessity for this, it seems pyrimidine TFOs should form the more stable triplex, although the kinetics of its formation may be unfavourable.

4.1 INTRODUCTION

As the formation of non-canonical DNA structures, such as the DNA triple helix, is known to be important in cellular function, they have potential use in the development of novel therapeutics. Due to their ability to impede transcription,[11,109] they are often associated with disease, and understanding how to destabilise them is important (see chapter 3). However, if triplex formation can be controlled such that they form preferentially in important regions of the bacterial genome, they may provide a novel means to deal with problems such as antimicrobial resistance by spawning a new class of antibiotic agents. TFOs used for these purposes are required to have good *in vivo* stability and the choice of sequence is therefore an important one.

As two different hydrogen-bonding motifs can form between TFOs and duplex, being either Hoogsteen or reverse Hoogsteen, a variety of different sequences can be used. These include homopurine or homopyrimidine sequences (the GA or CT motifs respectively), or mixed sequences such as GT which differ in the types of Hoogsteen bonds formed, and their relative orientation to the purine strand. Clearly, as multiple different sequences can target the same duplex stretch, when considering the choice of TFO for therapeutic applications it is important to know whether one will be more stable than another. Furthermore, understanding the reasons for stability of one over another may provide valuable insights into the sequence dependence of triplex stability, which could be generalisable for the development of novel, synthetic TFOs.

While triplexes formed from both homopurine and homopyrimidine TFOs have been studied and found to have similar stabilities for certain sequences under the right conditions,[46] it is known that purine triplex stability is highly dependent on the guanine content of the third strand, with a threshold content of 40–50 % reported to be required for triplex formation.[47,50] The origins of this effect are unknown, but differences in structure between A*A and G*G base pairs have been proposed to cause backbone structural distortion between A and G bases.[48]

In this work, by means of atomistic molecular dynamics (MD) simulation, the backbone structures of purine and pyrimidine triplexes are compared for a given duplex sequence and quantified through the Watson–Hoosteen (W–H) and Crick–Hoogsteen (C–H) groove widths in an attempt to determine whether the structure the bound TFO adopts is likely to have an effect on stability. Although not a definitive measure, the number of hydrogen bonds between TFO and duplex are determined, as an approximation for the relative stability of the sequences. Additionally, the change in sugar puckering for both the duplex moiety and the TFO on triplex formation are examined for purine and pyrimidine triplexes, as this may give insight into any structural rearrangements required for the triplex to form, potentially favouring one over the other.

4.2 COMPUTATIONAL METHODS

Atomistic MD simulations were conducted on a 19-base $d(GC(GAA)_5CG)_2$ duplex with a 15-base TFO. The purine TFO, with sequence $(GAA)_5$ was bound antiparallel to the purine strand of the duplex, and pyrimidine TFO, with sequence $(C^+TT)_5$ bound parallel. Purine structures were build with NAMOT[110] and pyrimidine with 3DNA, using the fiber31 model.[111,112] For the pyrimidine triplex, TFO cytosines were protonated at the N3 position. Although not protonated at physiological pH, the non-protonated base triplet, CTT, does not

form a stable triplex. Accordingly, as this work aims to compare stable triplexes to elucidate underlying reasons for their stability, the protonated structure was used. All simulations used the AMBER parm99 force field[69] with BSC0[70] modification and parameters for protonated cytosines included for the pyrimidine TFO.[113] Simulations were carried out using NAMD 2.12.[63]

Systems were solvated with approximately 15,000 TIP3P water molecules to give an (80 Å)^3 cubic simulation box to which periodic boundary conditions were applied. Systems were neutralised with 50 (purine) or 45 (pyrimidine) Na^+ ions using VMD's cionize plugin.[80] NaCl concentration was set to 0.15 M to mimic physiological conditions by the addition of 43 further Na^+ and Cl^- ions, positioned randomly. Long-range interactions were evaluated by means of the Particle-Mesh Ewald (PME) method and hydrogen containing bonds constrained with SHAKE[67], allowing for a 2 fs timestep. Simulations were conducted in the NPT ensemble, with temperature and pressure maintained at 310 K and 1 atm using a Langevin thermostat and barostat respectively.

Systems were initially simulated for 20 ps with triplexes constrained, to allow the solvent to equilibrate around the structure. Constraints were then released and the system run for 40 ns. Equilibration was reached, as evidenced by a flattening of the RMSD and energies, after approximately 10-15 ns, and the final 20 ns of simulation was therefore used for structural analyses.

The number of hydrogen bonds between TFO and duplex was measured using the hbonds plugin available in VMD,[80] with a cutoff distance of 3.0 Å and angle 25°, over the equilibrium portion of the trajectory.The C–H groove distance was measured as the interstrand P_{n_A}–P_{n_C+2}, where P_{n_A} is the nth phosphorus on chain A (purine duplex strand) and P_{n_C+2} is the $(n+2)$th phosphorus on chain C (the TFO) distance. The W–H groove-width was likewise defined as the interstrand P_{n_B}–P_{n_C+4} distance.

The sugar pucker phase and amplitude was calculated using the implementation in PLUMED2.[82] Average values were obtained over the 20 ns of equilibrium simulation trajectory. Values of phase between -90 and 90°were considered to be north and between 90 and -90°south.

4.3 RESULTS AND DISCUSSION

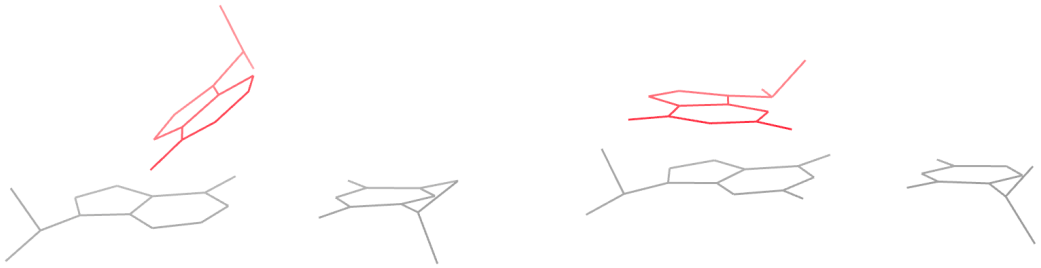
Atomistic MD simulations were conducted on two different DNA triplexes, having either an antiparallel purine, or parallel pyrimidine third strand. Examining the number of hydrogen bonds between each TFO and the purine strand of the duplex, it was found that the purine strand formed only just over half the number of hydrogen bonds of the pyrimidine strand (Table 4.1).

Table 4.1: Average number of hydrogen bonds between TFO and purine strand of duplex.

	Number of hydrogen bonds
purine	8.6 ± 0.6
pyrimidine	15.1 ± 0.5

A closer examination of the structure reveals that the A*A pairs do not sit in the plane of the other bases they pair with, but are twisted relative to the plane of the duplex rungs (Fig. 4.1). As these distorted bases are further from the duplex, this accounts for the measured reduction in hydrogen bond counts. While this may be explained by the reported non-isomorphism of

A*A and G*G base pairs, which requires some deformation of the backbone,[48] it also appears that reverse Hoogsteen A*A bonding interactions are much weaker than their G*G counterparts, being not even strong enough to hold the TFO in position. This is in agreement with previous quantum chemical calculations which examined the binding energy of the third base to the existing Watson-Crick (W–C) pair. These calculations indicated that A*A-T reverse-Hoogsteen trimer is significantly weaker than the G*G-C equivalent.[114] This provides an explanation for the requirement of a threshold fraction of guanine bases for a purine TFO to form as the poor interactions between adenines would not be sufficient to hold the structure together in the face of repulsions from the negatively charged backbones on either side. Indeed, when comparing the backbone structure of the purine and pyrimidine triplexes, a significant distortion is observed immediately following every A-G junction (Fig. 4.2). In the locations where the A bases fail to bind, the backbone of the triplex is pushed towards the center of the major groove, dividing it in these regions into two effectively symmetric grooves, rather than an asymmetric C–H and W–H groove as observed for the pyrimidine triplex. This shift towards the center of the groove is indicative of repulsive interactions between the charged backbones of the TFO and duplex which cannot be overcome by the poor bonding between adenine bases.

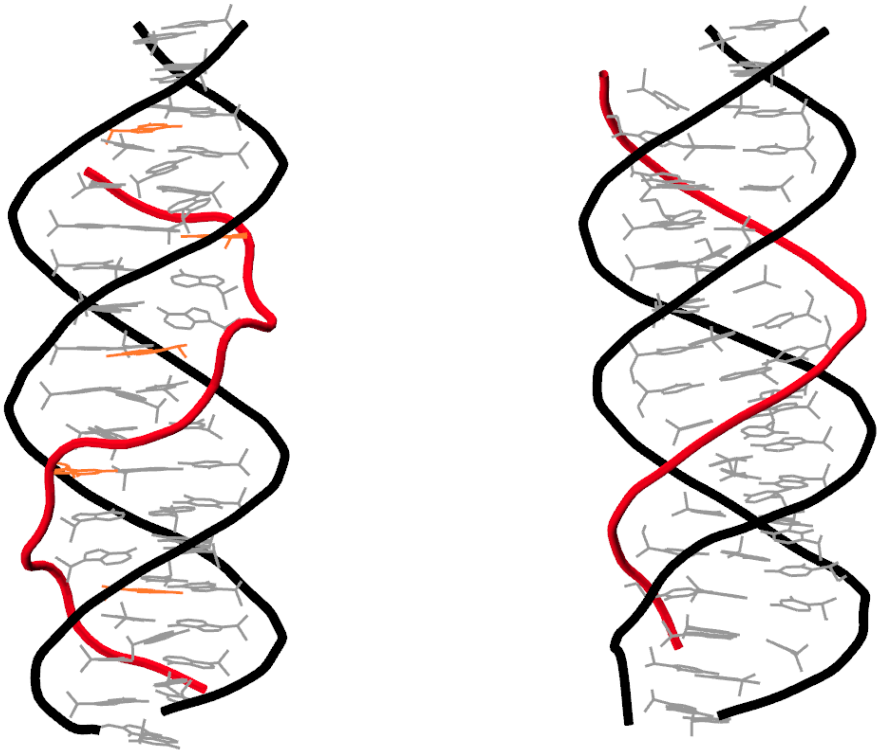


(a) T.A*A (b) C.G*G

Fig. 4.1: Structure of a T.A*A and C.G*G base triplet in a d(GC(GAA)$_5$CG) duplex with (GAA)$_5$ or (C$^+$TT)$_5$ TFO. Bases belonging to the third strand are shown in red. Hydrogens have been removed for clarity.

This effect can be further examined through measurements of the C–H and W–H grooves, defined by the backbone of the TFO and the DNA strands on either side of it. It can be seen that the width of the purine C–H groove, while generally approximately 4 Å wider than the pyrimidine groove, drops back to the width of the pyrimidine groove every 3 bases, corresponding to the locations of the guanine bases (Fig. 4.3). The W–H groove follows the same, but opposite, pattern. It is therefore clear that the lack of hydrogen bonding between adenines distorts the structure of the TFO, likely destabilising it. For the pyrimidine structure, all bases pair well, similar to the G*G structure, giving a regular backbone and higher number of hydrogen bonds.

An analysis of the sugar puckering of both TFOs in going from free to bound states also sheds some light on the differences between these structures. In both cases, very little, if any, rearrangement of the underlying duplex is observed on binding. However, changes are observed to occur in the sugar pucker of the TFO on binding to the duplex. In the case of the purine triplex, many of the bases display no preference for any sugar pucker, while a couple



(a) Purine (b) Pyrimidine

Fig. 4.2: Structures of purine and pyrimidine $(GAA)_n$ triplexes. TFO shown in red. For the purine triplex, guanines are shown in orange.

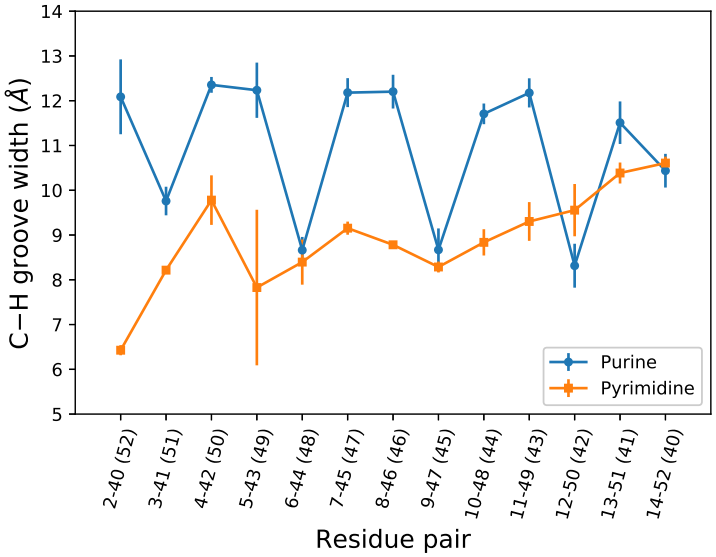
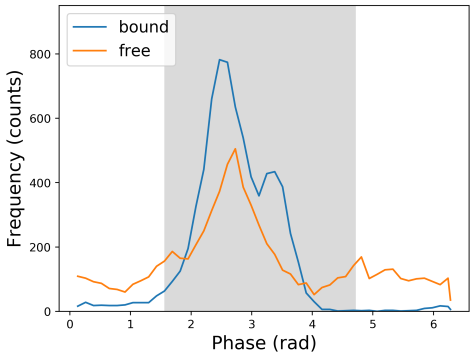


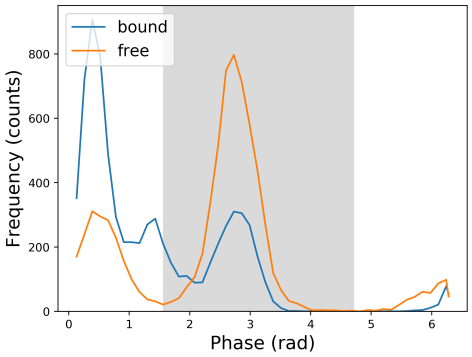
Fig. 4.3: Width of the C–H groove for a triplex with purine (blue, circles) or pyrimidine (orange, squares) TFO. X axis labels in brackets indicate the residue numbering for the purine (antiparallel) TFO.

favour south conformations (Fig. 4.4a). On binding, all bases favour south conformations. On the other hand, examining the change in sugar pucker preference for the pyrimidine TFO

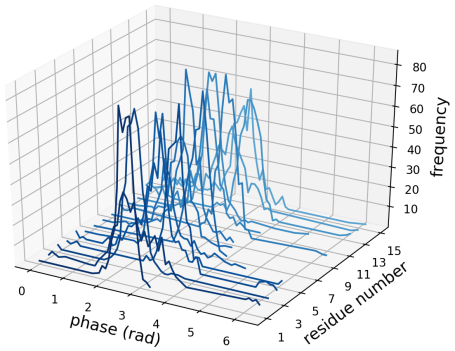
(Fig. 4.4b), a clear change in preference can be observed between the bound and free states. While free, the pyrimidine TFO favours south puckers, but when bound this changes to favour north. Clearly, significant rearrangement is required of the pyrimidine sugars to bind, which may make the binding event less likely to occur, especially if there is a large energetic barrier to pseudorotation.



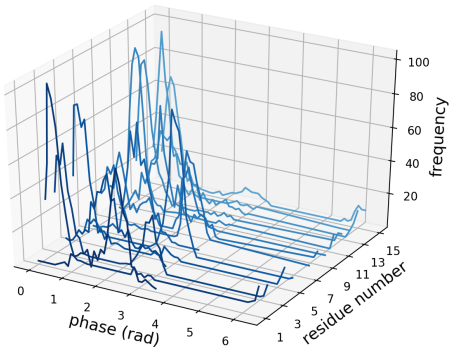
(a) Purine



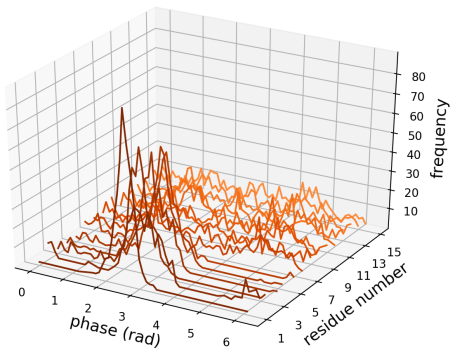
(b) Pyrimidine



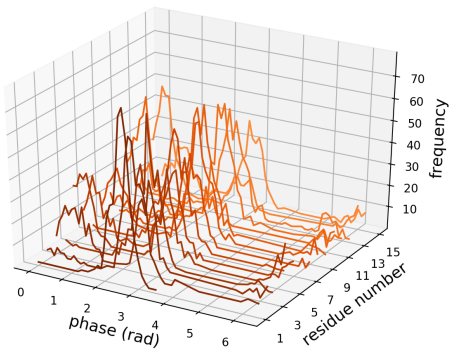
(c) Purine, bound



(d) Pyrimidine, bound



(e) Purine, free



(f) Pyrimidine, free

Fig. 4.4: Change in sugar pucker preference for purine and pyrimidine TFOs between the free (blue) and bound (orange) states. The region shaded in grey cover the range of phases considered 'south'. (a) and (b): overall change in sugar pucker. (c) – (f): the favoured sugar pucker of each individual base of the 15-base TFO in bound and free states.

4.3.1 SELECTION OF TFO FOR ANTIGENE THERAPY

When considering a TFO for antigene therapy, as has previously been mentioned, high stability in the cellular environment is desired. Although likely to be very sequence dependent, from the results presented here, it would appear that a pyrimidine structure is the best choice for stability. The large backbone distortions and the poor A*A hydrogen bonding likely make the purine TFO much less stable than the pyrimidine equivalent. As the G*G pairs did not show the same instability, purine strands with high G content may also be acceptable if the target sequence allows, although at sufficiently high G content, formation of G-quadruplexes competes with triplex formation.[47]

Despite the claimed requirement of 40–50 % G-content for triplex formation,[47] the triplex studied here has a guanine content of only 33 %. In the work of Vekhoff *et al.*, triplexes with G-content only as low as 43 % were shown to form stable triplexes, while triplex were shown to not form in some cases even with G-content as high as 45 %. This is therefore a property that strongly depends on sequence, and it is noted that A-stretches seem to disfavour triplex formation. It is possible that the sequence studied in the work presented here, where the A-stretches are interrupted with guanines every 3 bases, allows for a slightly lower guanine percentage required for stable triplex formation.

Although apparently more stable, at least for the duplex sequence studied in this work, the pyrimidine structures suffer from the requirement of protonated cytosines, which, given cytosines pK_a of 4.45, is generally not possible at physiological pH. This requirement could potentially be circumvented by the modification of these residues such that they do not require this protonation to form stable structures. Among these modifications, probably the most well known and common is the methylation of cytosine to give 5-methylcytosine.[115] This gives a triplex which forms independently of pH. It has also been shown that the replacement of cytosine with either 2'-deoxyisoguanosine or 9-deaza-9-propynylguanine N^7-(2'-deoxyribonucleoside), both modified guanine residues, results in pH independent intramolecular triplex formation. This gives triplexes that are stable at neutral pH while still having the same selectivity.[49] Alternatively, replacing the carbonyl oxygen in cytosine with a sulfur gives thio-pseudoisocytosine which is also able to bond with the guanine residue without the need for protonation.[116,117]

While it appears that pyrimidine triplexes are likely to be more stable, the large rearrangement required for their formation may reduce their usefulness therapeutically as they may not form on appropriate timescales. In looking at the potential triplex target sites within the genome of the bacterium Staphylococcus Aureus, generated with the triplexator software,[42] approximately 50 % of these were found to have a G content in excess of 50 %. Almost 85 % of sequences had a G content greater than 40 %. It therefore appears that a high G content is common in potential target sequences, meaning that the requirement of 40–50 % G content to form stable purine triplexes will be met in many cases.

4.4 CONCLUSIONS

Through atomistic simulations of purine and pyrimidine triplexes, binding to the same duplex, it was found that while G*G, A*T, and G*C^+ Hoogsteen pairs all pair well, the A*A pair does not. Although it is known that A*A and G*G are non-isomorphous, and A*A has been predicted to form weaker bonds, this almost complete lack of base pairing has not been previously observed. This corresponded to a loss of approximately half the hydrogen bonds

in going from a pyrimidine to purine strand with 33 % G content. Significant deviation from the regular TFO backbone structure was observed when adenines were present in the TFO, pushing the TFO towards the center of the major groove.

While it is known that high G content of the third strand is important for the stability of purine triplexes, the reasons why have not been made clear. The observed base pairing deficiency provides a reasonable explanation for this phenomenon, although the triplex studied, having only 33 % G content, is stable, despite the claimed requirement of at least 40 %. This is likely due to the particular sequence of this structure, in which the A-stretches are interrupted by guanines at every three bases, which differs from the sequences investigated in the experimental study that proposed this 40 % lower bound.

In examining the required structural changes, in terms of sugar puckering, between a TFO free in solution and bound in a triplex, significantly greater rearrangement was required of the pyrimidine TFO. As significant change for all the bases in the TFO is required for binding, it is likely that there is a large kinetic barrier to pyrimidine triplex formation, which may make it unlikely to form spontaneously as required for use as an antigene agent.

Overall, it appears that, not unexpectedly, good hydrogen bond contacts between the TFO and duplex are important for stability and play a large role in determining the structure of the TFO backbone. It was shown that adenine residues pair poorly while the other three bases give regular structures with good hydrogen bonding patterns, suggesting that the pyrimidine triplex is likely to be more stable. Furthermore, the requirement of high, but not too high, G content, may limit the usefulness of purine TFOs as antigene agents. However, as sequences with high G content do not appear to be particularly rare in biology, the use of purine TFOs may be generally possible in an antigene sense. This, coupled with the large rearrangement of the pyrimidine TFO required for triplex formation, suggests that in general the purine sequence may be preferred, despite its potentially lower stability.

CHAPTER

FIVE

STABILITY OF RNA–DNA HYBRID TRIPLEXES

Through the use of molecular dynamics simulations, the relative stability of hybrid DNA triplexes, where the third strand, binding to a DNA duplex, is RNA based, is examined. It has been previously reported that pyrimidine triplexes with RNA triplex–forming oligomers (TFOs) form more stable triplexes than their DNA counterparts, while purine triplexes are more stable with DNA, generally not forming when the TFO contains ribose. For the purine motif TFO, which binds antiparallel, it is found that the hybrid triplex, with RNA third strand, is approximately 2 kcal/mol less stable than the DNA counterpart. Although slightly less stable, this should not preclude its formation, as suggested in the literature. However, this value is shown to be strongly dependent on the conformation of the unbound TFOs in solution. Taking the sugar pucker to be a reasonable approximation for the general structure of the backbone, it is found that for a purine sequence, significantly greater rearrangement is required of the RNA TFO on going from its free to bound state than is required for the equivalent DNA TFO. The large changes required for the purine RNA TFO indicate that a significant kinetic barrier may prevent the experimental formation of this triplex, despite its stability. Two mechanisms have been previously proposed to explain the additional stability of the pyrimidine RNA triplex: the potential for hydrogen bonding between the H2' atom on the TFO ribose and the phosphate oxygens of the adjacent chain (which can only occur in the parallel pyrimidine motif), and the lesser degree of backbone rearrangement required to bind. Neither were found to satisfactorily describe its enhanced stability. The interstrand O–H distance, for the parallel pyrimidine triplex was found to be approximately 7 Å, too far for any significant hydrogen bonding interactions and similar backbone rearrangements were required for both DNA and RNA TFOs.

5.1 INTRODUCTION

While changing the TFO is one important, and facile, change that can improve the stability of triple helices for antigene therapy (chapter 4), the choice of backbone is also able to modulate stability.[46,51,52,54–57] Possibly the simplest change is replacing the deoxyribose sugars in the DNA TFO with ribose, giving an RNA-based TFO. For the pyrimidine triplex, it is known that the DD.R triplex – that is a deoxyribonucleic acid (DNA) duplex with an ribonucleic acid (RNA) TFO – is more stable than the equivalent DD.D triplex possessing a DNA TFO, although the value of ΔG for its binding is under some contention.[51,52,54,55] Equivalent studies of the antiparallel purine structure have shown the DD.R triplex to be less stable than the DD.D, and it generally does not form.[46,56,57] It is therefore useful to be able to study such a structure using computational methods, where the formation of the triplex is less of a challenge and the reasons for its proposed instability can be examined.

While the reasons for the lack of formation of the intermolecular DD.R purine triplex have not been previously examined, a number of reasons for the additional stability of the DD.R pyrimidine triplex have been proposed. Firstly, the triplex has been shown to adopt either an intermediate structure, in between that of A and BDNA,[52] or a structure in which the underlying duplex remains in a B conformation while the RNA strand adopts a conformation with a mix of C2' and C3'-endo sugars.[53] While significant structural rearrangement of either the duplex or TFO on binding would likely disfavour formation of the associated triplex, it is not clear whether these rearrangements are necessary for triplex formation, nor whether the barriers between different conformers are sufficiently high that the rearrangements become energetically difficult. Additionally, parallel pyrimidine triplexes with an RNA TFO have the potential to form hydrogen bonds between the 2'-OH group on the ribose sugar and one of the phosphate oxygen of the adjacent chain. This can only occur when the TFO binds parallel to the purine strand of the duplex and hence may explain the reported lower stability reported for the purine analogue, which binds antiparallel.[56] As methylation at this position has not been shown to decrease stability, but rather increases it,[55] taking a molecular modelling approach to determine the accuracy of this hypothesis may give some insight into the process.

To determine whether these are viable reasons for the additional stability of the DD.R pyrimidine triplex, the structural changes, including the sugar conformation of both the duplex and TFO, upon TFO binding are examined computationally. The relative stability of the DD.D and DD.R antiparallel purine triplexes are also calculated and compared, as well as a similar structural analysis. This work aims to provide the basis for the understanding of some of the properties affecting triplex stability, for both purine and pyrimidine TFOs, allowing for rational design of new oligomers for antigene therapy.

5.2 COMPUTATIONAL METHODOLOGY

In order to compare the relative stabilities and structural properties of DD.D and DD.R triplexes, four different structures were examined by molecular dynamics (MD) simulation: DD.D and DD.R with purine TFO, and their equivalent pyrimidine complexes. All constituted a 19-base duplex, with sequence $d(GC(GAA)_5CG)_2$, and a 15-base TFO. Antiparallel purine triplex structures had TFO sequence $(GAA)_5$ while the parallel pyrimidine had $(C^+TT)_5$, with the thymines in DNA replaced with uracil for the DD.R triplex. Parallel structures were built with 3DNA based on the fiber31 model[111,112] and TFO cytosines protonated by the addition of a hydrogen atom at the N3 position. Antiparallel structures were built with

NAMOT.[110] Simulations were carried out with NAMD 2.12[63] using the AMBER parm99 force field[69] with BSC0 modification (parm99bsc0).[70] Parameters for protonated cytosines, calculated following the same protocol as for the BSC0 force field, were also included for the pyrimidine triplexes.[113]

All simulations were carried out in explicit solvent using TIP3P water to give an approximately (80 Å)^3 cubic water box. Approximately 15,000 water molecules were required. The systems were neutralised with Na^+ using VMD's cionize plugin.[80] 50 ions were required to neutralise the purine triplexes, and 45 for the pyrimidine analogue as 5 cytosines were protonated. In all cases, 43 additional Na^+ and Cl^- ions were added to bring the ion concentration to 0.15 M, representative of physiological conditions. All simulations were conducted in the NPT ensemble at physiologically relevant conditions of 310 K, maintained using a Langevin thermostat, and 1 atm, maintained with a Langevin barostat. Hydrogen containing bonds were constrained using SHAKE,[67] and long-range interactions evaluated by the Particle-Mesh Ewald (PME) method. A timestep of 2 fs was used.

Prior to the unrestrained MD simulations, triplexes were constrained, the energy of the system minimized, and water and ions allowed to equilibrate around the triplex for 20 ps. Restraints were released and the system run for 40 ns. The final 20 ns of these simulations was analysed for structural data.

5.2.1 FREE ENERGY CALCULATIONS

Two different types of Free energy perturbation (FEP) calculations were carried out to determine the relative stability of the DD.D and DD.R purine triplexes described above. Rather than annihilating the entire TFO, the first class of simulations involved the mutation of only the 5 atoms in the sugar which changed significantly between deoxyribose and ribose. The C2', H2', H2", C3', and H3' atoms were annihilated from deoxyribose, while new C2', O2', H2', H2", C3', and H3' atoms were simultaneously created. Charges of the other sugar atoms, and the rest of the DNA structure did not change significantly and therefore, for simplicity, their mutation was not included.

In the AMBER forcefield every atom of DNA and RNA has a slightly different partial charge so the mutation of only 5 bases leads to a slight difference in net charge of the triplex, from $-50e$ at the DNA end to $-49.93e$ at the RNA end of the simulation. Although this results in a non-physical net charge of $+0.07e$ for the entire TFO at the RNA end, the effect of this slight change can be considered negligible. The overall effect of this change in charge can be accounted for through the inclusion of a Born solvation free energy for creating a net charge in a spherical cavity and a finite size correction to the free energy, calculated using the analytical equations presented by Hummer et al.[118] Given a change in charge of $0.07e/15 = 0.0047e$ per base (or $0.07e$ for the entire TFO), and solute (nucleotide base) radius of 0.3 nm, the Born solvation free energy is estimated to contribute 0.02 kcal/mol, and the finite size correction 0.1 kcal/mol, which combined are smaller than the error bars on the simulated free energies.

Furthermore, although the presence of the non-physical net charge would have a slight effect on the calculated free energy, the mutation was conducted twice, in the bound and unbound states, with the free energy calculated as the difference between the two. The effect of the net charge should therefore approximately cancel out in the overall free energy reported, giving an overall effect of the charge difference which will be even smaller than the individual contributions and thus likely to be negligible.

With the exception of the substitution of H2' in deoxyribose with O2'-H2' in ribose, only the charges of the atoms differed between the two states. The mutation was carried out starting from the end of a 20 ns equilibration simulation with all atoms defined, but only deoxyribose atoms interacting. This was conducted in the forwards (deoxyribose $\rightarrow$ ribose) and reverse (ribose $\rightarrow$ deoxyribose), directions, where the reverse simulations were started from the end point of the forwards. Each simulation was split into 64 evenly spaced windows ($\Delta\lambda = 0.015625$), with electrostatics being turned off for the annihilated particles between $\lambda = 0$ and 0.5, and vdW between $\lambda = 1$ and 1. 375,000 steps (0.7 ns) of simulation were conducted at each window, including 125,000 (0.25 ns) of equilibration, giving total simulation time of 48 ns. For the created particles, electrostatics were turned on between $\lambda = 0.5$ and 1, and vdW between $\lambda = 0$ and 1. Simulations were conducted for both bound and unbound TFOs, starting from an extended conformation.

Further FEP calculations were carried out examining the annihilation of the entire TFO, both DNA and RNA. All systems were set up as previously described in chapter 3 and the simulations started from the end point of the 40 ns equilibrium simulations. Four simulations were required: annihilation of either the DNA or RNA TFO from states bound to the triplex, and free in solution. Only the annihilation was considered as the slow relaxation of the DNA structures makes the timescales required for it to spontaneously bind to the triplex when created, prohibitive. The thermodynamic cycle associated with the process is outlined in Fig. 5.1.

Each simulation was divided into 16 unequally spaced windows, taking on values of $\lambda = 0.04, 0.08, 0.12, 0.16, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.60, 0.70, 0.80, 0.90, 1$ with electrostatic interactions scaled to zero linearly over the range $\lambda = 0 - 0.5$ and vdW interactions similarly over $\lambda = 0 - 1$. The total simulation time was 192 ns for each of the four simulations. In order to keep the TFO in the binding site while interactions were being turned off, center-of-mass restraints were applied to the distance between the TFO and duplex. This was constrained to its average equilibrium distance with a force constant of 10 kcal/mol/Å^2 using the colvars module of NAMD.[119] ΔG_{rest}, the effect of these constraints was calculated according to Wang *et al.* over 11 evenly spaced windows between 0 and 1.[102]

Additionally, as the DNA and RNA TFOs were found to adopt different structures after 40 ns of simulation, with DNA being found in a globular conformation and RNA relatively extended, steered molecular dynamics (SMD) simulations were conducted to determine the free energy change on going from the globular to extended conformation in each case. To calculate the free energy change, the work to pull the triplex from an end-to-end distance of approximately 20 Å, corresponding to a globular nucleic acid, to approximately 70 Å, the fully extended conformation, was calculated and averaged over 10 trajectories. A pulling speed of 1 Å/ns was employed. A spring with spring constant 10 kcal/mol/Å^2 was attached to the phosphorus atom of the 3'-terminus which was pulled away from the fixed atom (C5' of 5'-terminus) along the vector joining the two atoms.

5.2.2 SUGAR PUCKERING

The preferred sugar pucker of each of the four different structures was determined through examination of the equilibrium trajectories of the structures. Phase and amplitude was calculated using PLUMED2[82] over the final 20 ns of the trajectory. A phase of between -90 and 90 ° was considered to be north, and between 90 and -90 ° south. The pseudorotational phase was compared for the duplex moieties of each of the four triplexes, the bound TFOs

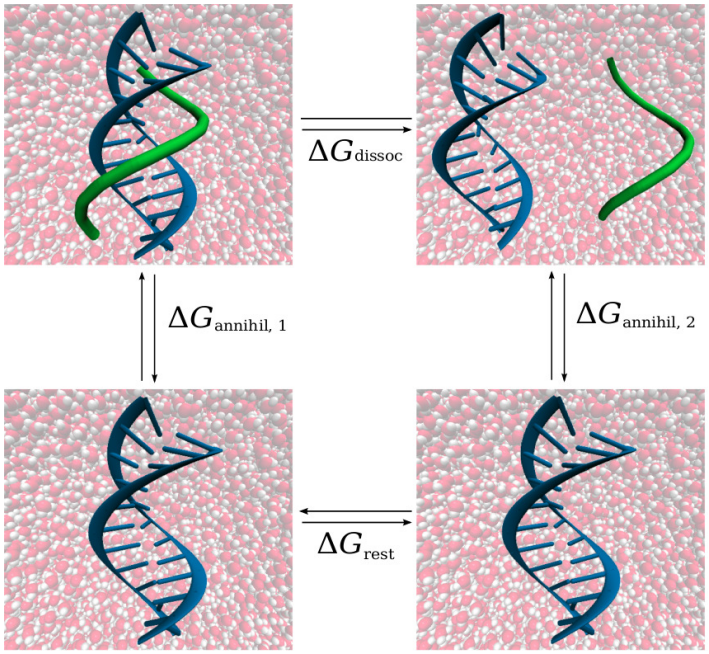


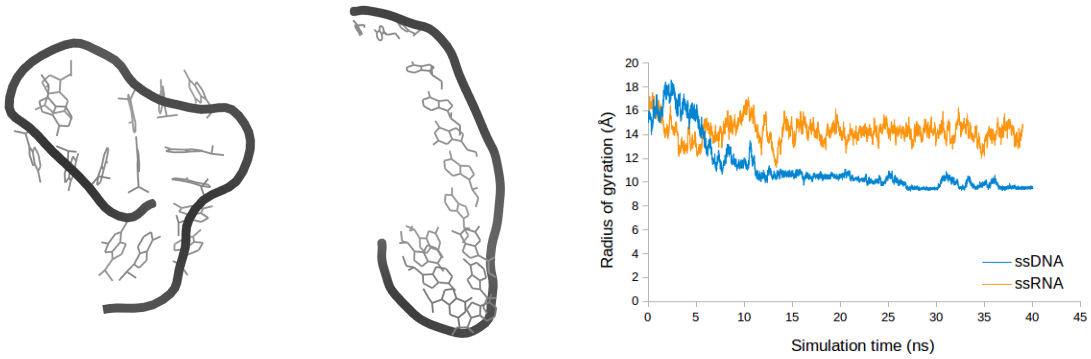
Fig. 5.1: Thermodynamic cycle for the dissociation free energy calculations of TFO (green, tube representation, either DNA or RNA) to duplex (blue, ribbon representation). ΔG_{dissoc} is calculated as $\Delta G_{\text{annihil,1}} - \Delta G_{\text{annihil,2}}$

and the free single–stranded nucleic acids.

To quantitatively determine the limitations any required rearrangements put on the formation of a triplex, the free energy landscape, as a function of pseudorotational phase and amplitude, was calculated for the two single–stranded nucleic acids, ssRNA and ssDNA with the purine sequence, as this showed the largest difference of sugar pucker between DNA and RNA on binding to the duplex. The landscape was calculated using plumed's implementation of umbrella sampling.[82] These simulations were conducted on a 6-base TFO, set up as previously described for the 15-base triplexes, although in a smaller $(40\ \text{Å})^3$ cubic water box to reduce the number of atoms and improve computational efficiency. The pseudorotational phase and the amplitude, which define the sugar pucker, of the fourth (guanine) residue of the 3'-GAAGAA-5' sequence were biased. The phase was biased with a force constant of 335 kcal/mol at 0.2 rad intervals between 0 and π. The amplitude was biased with force constant 191.4 kcal/mol at 0.1 rad intervals between 0.6 and 1. This selection of biasing potentials and positions gave good overlap between adjacent histograms. At each pair of biasing potentials the system was simulated for 2 ns. This was repeated 5 times with the data from the first 0.5 ns of each simulation discarded as equilibration. Free energy landscapes were constructed as a function of phase and amplitude using 2D-WHAM.[97]

5.2.3 Overall structure

The overall structure of the triplex was examined by looking at the average structure of the equilibrated triplexes, calculated from the last 20 ns of the 40 ns simulations. The structures were aligned with the z-axis, such that looking at the structure in the xy plane looks down the center of the cylindrical molecule. The coordinates of backbone phosphorus atoms could then be compared between the native duplex and the four triplexes to determine any significant structural changes occurring on TFO binding.



(a) Equilibrium structures of ssDNA (left) and ssRNA (right). The DNA structure is significantly more globular.

(b) Radius of gyration of ssDNA and ssRNA. A smaller radius of gyration indicates a more compact structure.

Fig. 5.2: Variation in structure of ssDNA and ssRNA over the course of a 40 ns simulation.

5.3 Results and Discussion

5.3.1 Equilibrium structures of ssDNA and ssRNA

As starting points for the FEP simulations in the unbound state, 15-base sequences of ssDNA and ssRNA were simulated under the conditions previously described. 40 ns of simulation was conducted for each, which was deemed to be enough to reach equilibrium, through analysis of energies and structural properties such as the radius of gyration and root-mean-square deviation (RMSD). The structures of ssDNA and ssRNA were found to differ significantly. While ssRNA maintained the relatively extended conformation expected of single-stranded nucleic acids,[120,121] ssDNA folded quickly (within 10–15 ns) to a compact globular structure (Fig. 5.2). To ascertain that the ssRNA structure did not also fold to this structure on a longer timescale, the simulation was extended to 80 ns and no indication of further folding was observed.

SMD simulations of stretching the ssDNA coiled conformer to its extended state indicated that the globular state was approximately 10–20 kcal/mol more stable than the extended in this case. As this is not the structure that has been previously observed for ssDNA,[120,121] it is possible that the force field parameters used, which were not explicitly designed for single-stranded nucleic acids, may influence this structure, introducing artificial backbone flexibility. Furthermore, as previous studies have been conducted using different sequences and conditions,[120,121] they may not be directly comparable to the results reported here. For ssRNA, the globular state was also found to be approximately 15–30 kcal/mol more stable, in contrast to the observed structure in the unbiased simulation, indicating that the extended conformation of the simulated ssRNA is a metastable conformation. It therefore appears that the conformational fluctuations required for ssRNA to adopt this structure are such that it happens on a very long timescale, longer than 80 ns.

5.3.2 Relative stability of purine DD.D and DD.R triplexes

The relative stability of purine DD.D and DD.R triplexes was compared through FEP calculations of the mutation of the TFO sugars from deoxyribose to ribose and back. Starting the unbound TFO from an extended conformation, it was found that for the antiparallel purine

triplex, the DNA triplex was 2.2 ± 1.9 kcal/mol more stable than that containing the RNA TFO (Table 5.1). Over this approximately 40 ns simulation period, and the preceding 40 ns of equilibration, the TFO was never seen to spontaneously adopt the globular structure observed for the equilibrium simulations for ssDNA.

The annihilation calculations, in which the entire TFO is annihilated, differ from the mutation calculations in that the free energy is calculated as the difference between the annihilation of the TFO, either DNA or RNA, bound to the duplex, and free in solution, giving a dissociation free energy of either the DNA or RNA TFO. The mutation simulations, on the other hand, give the relative free energy difference between the two structures. The difference between the dissociation free energies, calculated by annihilation, of the DNA and RNA TFOs should then give the same relative free energy difference as obtained from the mutation. Annihilation of the entire TFO starting from the final configuration of the unbiased 40 ns simulations discussed in section 5.3.1, where ssDNA adopted a globular structure and ssRNA a mostly extended one, indicated that RNA should be favoured by approximately 30 kcal/mol, where this value is the difference between the dissociation free energies of the RNA and DNA triplexes. It therefore appears that the starting structure of the unbound TFO has a large effect on the relative stabilities calculated using FEP. This indicates at least two different stable or metastable states, globular and extended, of the single–stranded 15-base nucleic acid molecules studied.

In examining the trajectory of the RNA TFO annihilation when free in solution, it can be seen that, despite clearly favouring the extended structure in the unbiased simulation, within the first window of the FEP annihilation (after less than 12 ns of additional simulation), it briefly adopts the globular structure (Fig. 5.3) which appears to be the most stable structure by SMD. As previously mentioned, SMD results indicate that for RNA, the coiled conformation should be approximately 15–30 kcal/mol more stable. Taking this into account, the annihilation of the unbound RNA TFO in the globular state can be expected to be about 15–30 kcal/mol higher than that calculated for the extended one. The relative stability of the RNA TFO, would then be similar to that of the DNA TFO, and within error of the result of the mutation. A comparison of the results obtained by mutation from the extended state, and annihilation, which includes the difference between globular and extended states obtained by SMD, is outlined in table 5.1.

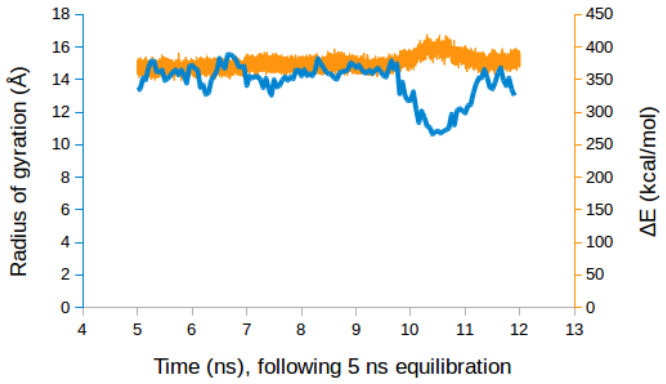


Fig. 5.3: Radius of gyration (blue) and ΔE (orange) over the first window of the FEP annihilation of the free RNA TFO. The decrease in the radius of gyration indicates the structure folding up. This corresponds to an increase in ΔE, indicating that the annihilation of this free globular structure is less favourable.

Table 5.1: Free energy of conversion between a 15-base, purine DNA and RNA TFO with a DNA duplex. A negative value indicates that RNA is favoured. Mutation and annihilation refer to the method by which the free energy difference was calculated. For annihilation, the reported value is the difference in ΔG_{dissoc} for DNA and RNA, calculated through the annihilation of each TFO in the bound and free states. The relative free energy difference for the mutation simulations is calculated directly through mutation of the deoxyribose atoms to ribose.

Thermodynamic cycle	$\Delta\Delta G_{\text{dissoc}}$ (kcal/mol)
Annihilation	-9.53 ± 16.6
Mutation	2.24 ± 1.90

Although contradicting the experimental results, which suggest that purine RNA TFOs should form a less stable triplex, these results can be rationalised physically and therefore may indicate a barrier to formation that prevents the structure's experimental formation. It should also be noted that the experimental sequences differ slightly from the one examined here, having different G and A compositions, which may additionally affect the stability. However, as the base sequences for DNA and RNA TFOs examined here are the same, being a $(GAA)_5$ TFO bound antiparallel to the duplex, this DNA–RNA substitution should not be expected to change the interactions of the pairing bases significantly. Structurally, very little difference is observed between the purine DD.D and DD.R triplexes. As discussed in chapter 4, the backbone of the purine triplex is significantly distorted due to poor formation of the A*A reverse Hoogsteen pair. While in the case of the DD.R triplex, this distortion appears to be not as great, a lack of adenine pairing is still observed, indicating that the relative stability should not be greatly affected. A detailed structural analysis is presented in the section that follows, comparing the structures of purine and pyrimidine, DD.D and DD.R, triplexes.

5.3.3 Structural evidence for enhanced stability of DD.R triplexes

While the results presented here for the purine triplex indicate that, even if marginally less stable, a triple helix should still form, in considering the pyrimidine triplexes it is well known that the DD.R triplex is more stable than the DD.D.[51,52,54,55] Two reasons have been proposed for this additional stability when the TFO is in the parallel pyrimidine class: a lower degree of structural rearrangement required to form the triplex, and potential hydrogen bonding between the ribose OH group and phosphate oxygens on the purine chain of the duplex. Here both are examined for all four triplexes.

Structural rearrangement

Structural changes occurring on TFO binding were compared on two levels. Firstly, the overall structure of the triplex, effectively the width of the helix, was compared for all four structures. On a smaller scale, structural rearrangements, quantified by the change in sugar pucker which should accurately represent changes between A and B form structures, were examined.

Examining the positioning of backbone phosphorus atoms, in all cases, the helix was found to expand, relative to the duplex when the TFO bound (Fig. 5.4). Generally, the increase in diameter was approximately 2 Å. It is not clear what causes this change in structure, but likely the expansion of the major groove required for TFO binding also induces other structural distortions, leading to the pattern observed here. While interesting, as it occurs to

a similar extent for all four triplexes it is unlikely to influence the relative stability of one over the other and does not provide any useful insight into the reasons which may be important for triplex stability.

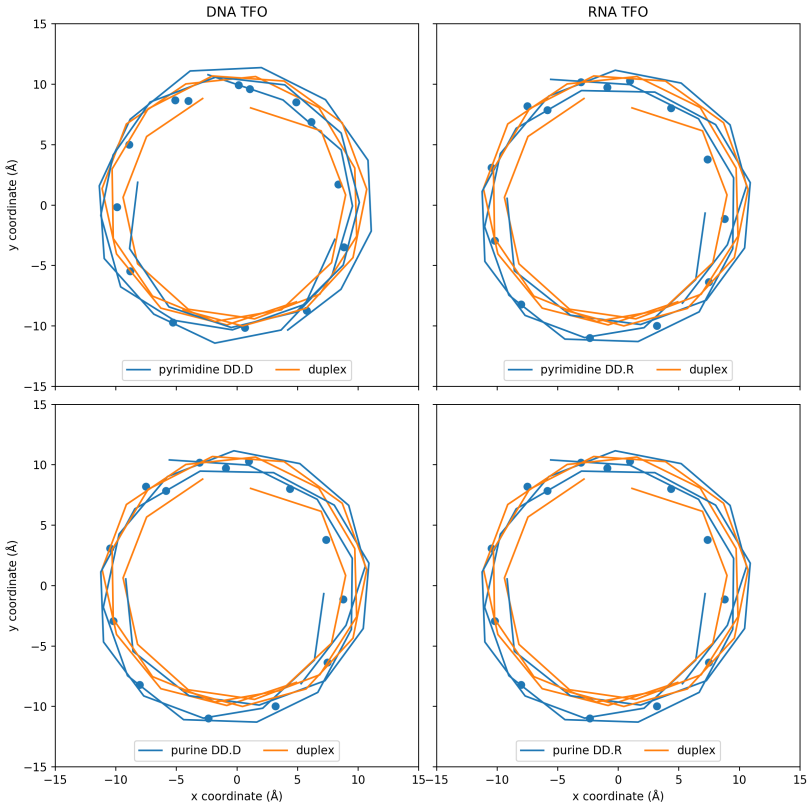
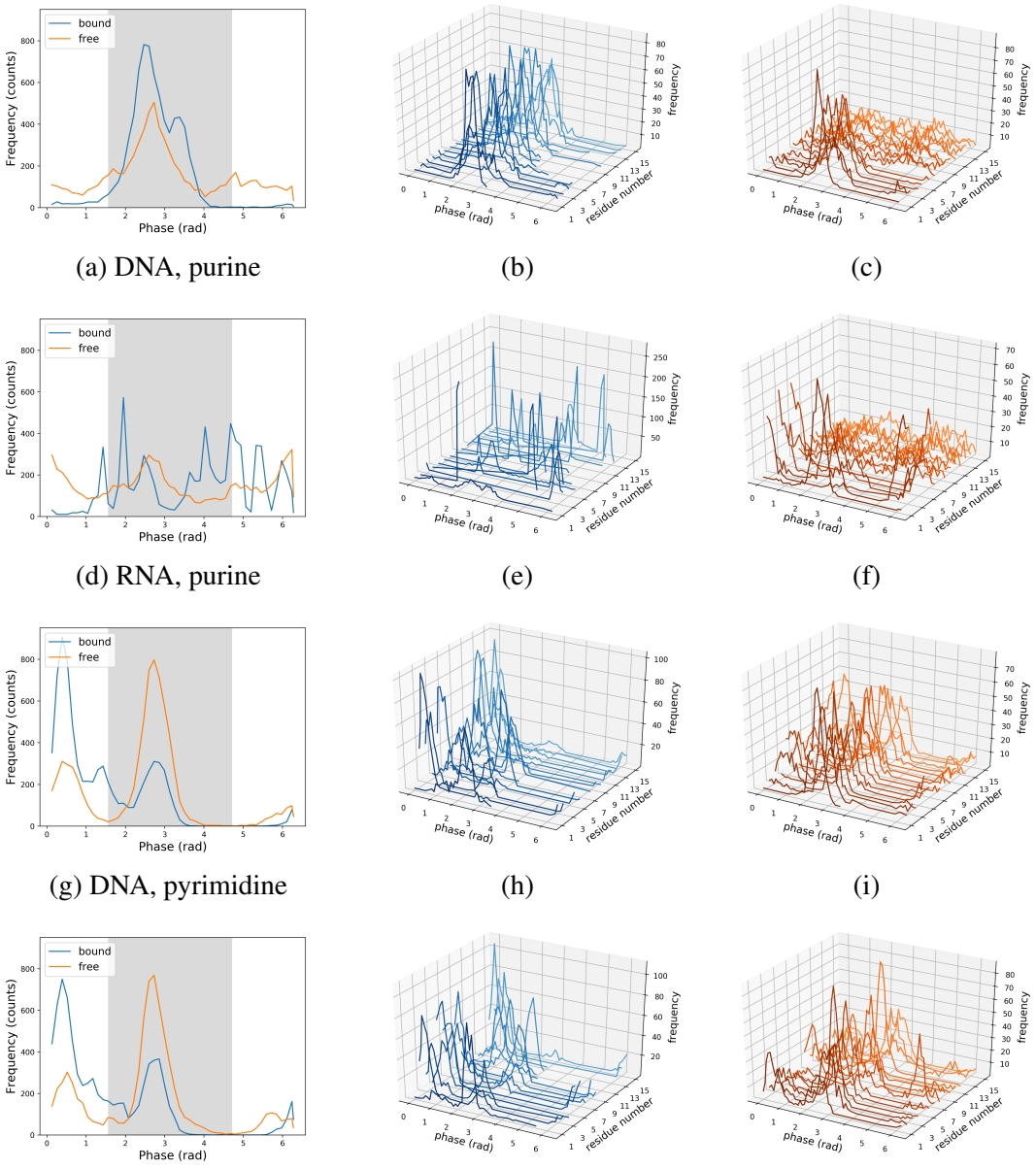


Fig. 5.4: Comparison of the backbone structures, looking down the length of the triplex, between purine and pyrimidine triplexes with DNA and RNA backbones, relative to the DNA duplex. Lines indicate the backbone atoms of the duplex, while the points are the positions of the phosphorus atoms of the TFO.

As the structure of the hybrid DD.R triplex has been shown to include an RNA TFO with mixed sugar puckers while the duplex maintains its B conformation, the changes in sugar puckering for the four triplexes, which is a useful measure for structural conformation, can be used to give insight into this process. The sugar pucker was examined over the final 20 ns of the 40 ns trajectories for each of the four triplexes. Over this time period, the frequency of sugars having a particular conformation was calculated. In the purine case, significant difference is observed between the changes required of DNA and RNA TFOs on binding (Figs. 5.5a, 5.5d). While the DNA TFO does not change pucker significantly, the RNA ribose sugars go from effectively rotating freely to each base favouring a specific, and generally different, phase. If these changes in phase must occur in a concerted fashion, this may introduce a large kinetic barrier which may inhibit the binding of the RNA TFO, potentially providing an explanation for why this triplex has not been reported to form experimentally, despite being quite thermodynamically stable.

On the other hand, although significant rearrangement is required for pyrimidine triplexes going from the free to bound states, this change is consistent between DNA and RNA TFOs. Both change from conformations that are predominantly south (S) to favouring north (N) conformations. It is therefore unlikely that this has a large effect on the relative stabilities of

the DNA and RNA TFOs as previously proposed.



(a) DNA, purine (b) (c)

(d) RNA, purine (e) (f)

(g) DNA, pyrimidine (h) (i)

(j) RNA, pyrimidine (k) bound (l) free

Fig. 5.5: Changes in the frequency of bases having a particular sugar pucker phase going from free in solution (blue) to bound (orange). Grey shaded regions indicate the phases corresponding to south puckers. Left: overall change in sugar pucker. Centre, right: the favoured sugar pucker of each individual base of the 15-base TFO in bound and free states.

From calculations of the pseudorotational free energy landscapes, the barrier separating C2'-endo and C3'-endo, conformations typical of S and N sugars, respectively, was found to be approximately 0.7 kcal/mol for purine ssDNA, and 1.8 kcal/mol for ssRNA (Fig. 5.6). These values are on the order of 1–2.5 $k_{\mathrm{B}}T$ at 310 K, indicating that although both C2'-endo and C3'-endo states are quite stable, transitions between the two may be quite slow, particularly for RNA. As changes in all RNA bases are required for the binding of this triplex, this likely contributes a large energetic penalty to binding. Furthermore, the barrier between conformers is larger for RNA, indicating that this required change may not be as facile as it would be in

the case of DNA.

HYDROGEN BONDING BETWEEN OH AND OP

Hydrogen bonding between the new ribose 2'-OH group and a phosphate oxygen on the adjoining strand potentially may influence the stability of the DD.R triplex, as it is an interaction which is not possible when the TFO is deoxyribose based. As it requires the two groups to be in close proximity, it is only possible when the TFO binds parallel to the purine strand of the duplex, indicating a pyrimidine TFO. The average interatomic distance was calculated between the ribose H2' and phosphate OP1 atoms over the final 20 ns of the 40 ns trajectory for pyrimidine TFOs with both DNA and RNA backbones.

Although the starting conformation of the RNA TFO was such that the groups of interest were within 1–3 Å of each other, capable of hydrogen bonding, withing 50 ps of simulation they had moved to a distance of 5 Å, which then increased to an average equilibrium distance of approximate 7 Å, much too far for significant hydrogen bonding (Fig. 5.7).

Overall, the average H–O distance was shorter for the pyrimidine (parallel) triplexes than their purine (antiparallel) equivalents, as expected due to the arrangement of the TFO in the major groove, but there was little difference between DNA– and RNA–based TFOs and the distance was too great to indicate significant hydrogen bond formation (Table 5.2). The similarity of the pyrimidine DNA and RNA triplexes is further indication that the potential for the ribose 2'-OH to form hydrogen bonds (which contrasts with the deoxyribose 2'-H inability to do so), is not responsible for the formation and stabilisation of this triplex.

Table 5.2: Average H–O distance between ribose 2'-OH (or deoxyribose 2'-H) and phosphate O for the four triplexes of interest.

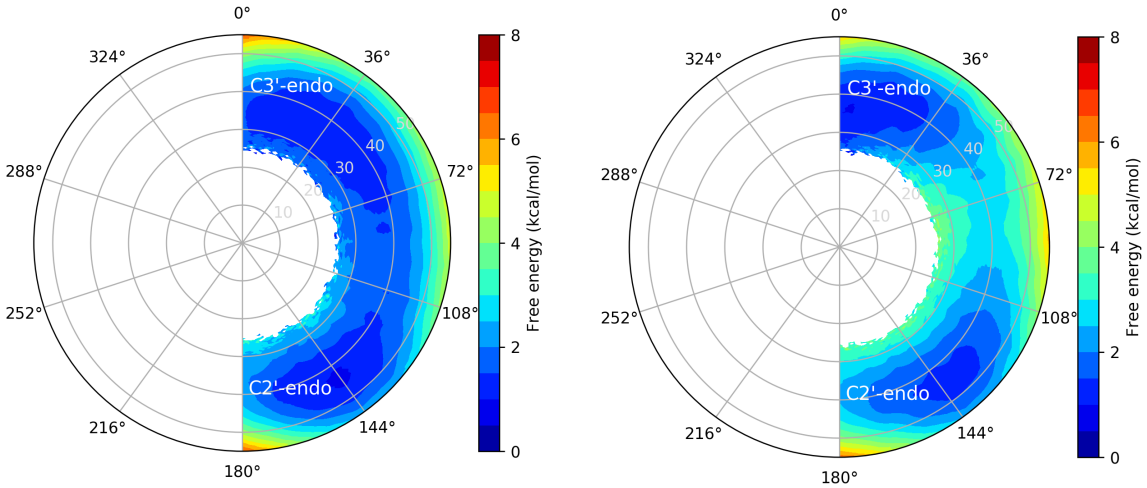
Triplex	OH-OP distance (Å)
Purine DD.D	9.8 ± 1.6
Purine DD.R	10.6 ± 3.2
Pyrimidine DD.D	6.7 ± 1.5
Pyrimidine DD.R	7.0 ± 1.8

SUMMARY

Overall, for the pyrimidine triplexes, very little difference was found between DNA- and RNA-based TFOs to indicate that one would be more stable than the other. Additionally, the average O–H distance is too far to be implicated in stabilising these parallel triplexes. The results are summarised in table 5.3. For the purine triplexes, although the DD.R triplex was found to be stable enough that it should form, the large rearrangement required of the RNA backbone may explain its absence in the literature. As expected in this case, the average O–H distance is also too far for hydrogen bonding.

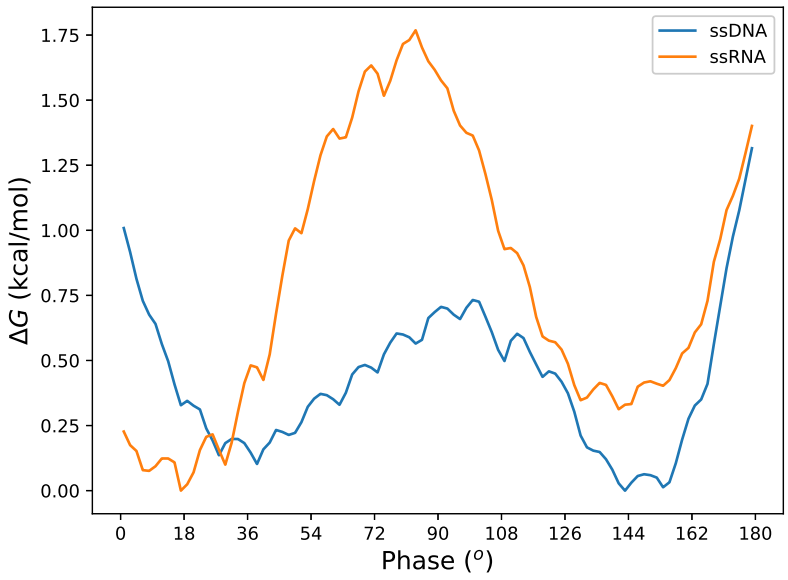
5.4 CONCLUSIONS

Although many previous studies have examined the relative stability of DD.D and DD.R triplexes with a pyrimidine third strand, few have been able to form the DD.R purine equivalent to study. It has therefore been proposed that for a purine motif triplex, the DD.R triplex is significantly less stable than the analogous DD.D triplex. Here, through MD simulation of



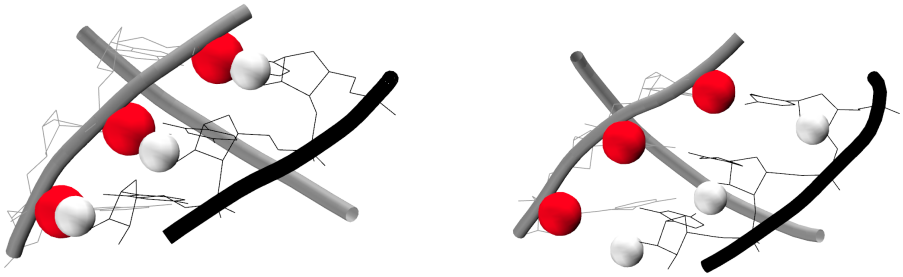
(a) ssDNA

(b) ssRNA



(c) One dimensional

Fig. 5.6: Eastern hemisphere of the pseudorotational free energy landscape of a single base in purine single stranded DNA and RNA. The one dimensional free energy is calculated at amplitude of 41°, found to be the most stable amplitude, and shifted so the minimum of each plot is at zero.



(a) Starting structure (b) Equilibrated structure

Fig. 5.7: Starting vs equilibrium structure of parallel DD.R triplex. At equilibrium OH' (H in white) and OP (O in red) groups are too distant for hydrogen bonding. TFO is shown in black.

Table 5.3: Summary of the differences between triplexes formed with the 4 different TFOs under study. [a] Favoured pucker refers to the US simulations of the single–stranded purine structures and is the thermodynamically most stable puckering phase for the single base studied. Height of the barrier between states is given in brackets.

	Pu DD.D	Pu DD.R	Py DD.D	Py DD.R
Helix width	↑ ~2 Å	↑ ~2 Å	↑~2 Å	↑ ~2 Å
Pucker: free to bound	S + random → S	random → banded	S → N	S → N
Favoured pucker[a]	C2'-endo (0.7 kcal/mol)	C3'-endo (1.6 kcal/mol)	-	-
O-H distance (Å)	9.8 ± 1.6	10.6 ± 3.2	6.7 ± 1.5	7.0 ± 1.8

the DD.R purine triplex, we found that the DD.R triplex is slightly less stable than the DD.D equivalent, but should still be stable enough to form under standard conditions. Additionally, the conformation of the free TFO backbone was found to have a large effect on the value of ΔG for the annihilation of the unbound triplex. However, if the free energy associated with the transition from globular to extended ssRNA is also considered, comparing the stability difference between globular ssDNA and globular ssRNA, and extended ssDNA and extended ssRNA, they do not differ significantly.

Structurally, very little difference was found between triplexes formed with DNA or RNA TFOs to explain their reported relative stabilities. In all cases, the helix width increased by approximately 2 Å upon binding, an effect which may be due to the widening of the major groove on triplex binding and its related structural distortions. Further study of this phenomenon is required to better understand the mechanisms of this slight widening although it is unlikely to contribute to the relative stabilities.

While there was little difference in the changes required of sugar pucker for the pyrimidine DD.D and DD.R triplexes, examination of the purine case indicated that significant rearrangement of the RNA TFO is required for binding, suggesting a large kinetic barrier. When considering the energetic barrier to the N–S transition for the purine TFO, this is much greater for ssRNA than ssDNA, on the order of 2.5 $k_{\mathrm{B}}T$. It is therefore likely that the rearrangement required for the RNA TFO to bind plays a large role in preventing the experimental formation of the DD.R purine triplexes, even if they are more stable. Although it has been proposed that a larger rearrangement of the DNA pyrimidine TFO is responsible for the

lower reported stability of this triplex relative to the DD.R pyrimidine triplex, we do not find any difference in the degree of rearrangement required. This therefore does not explain the relative stability differences. It is likely that the sugar pucker does not completely describe all the structural changes required on TFO binding, and examining other structural parameters may therefore provide insight into the stabilities of the pyrimidine triplexes.

The potential for pyrimidine ribose 2'-OH groups to form hydrogen bonds with the parallel purine strand of the duplex provides another logical explanation for the observed stability of the DD.R pyrimidine triplex. However, upon examination of the equilibrium structures of the DD.D and DD.R pyrimidine triplexes, little difference was observed between the O–H distances, which were too far for hydrogen bonding (at approximate 7 Å) in both cases. As the DD.D triplex is not capable of forming these hydrogen bonds, the similar distance for both TFOs indicates further that these hydrogen bonds are not in fact forming. It is therefore unlikely that this significantly contributes to the stability of the RNA triplex over the DNA.

In general, from the simulations carried out, structural properties previously hypothesised to contribute to the relative stabilities of the DD.D and DD.R pyrimidine triplexes do not appear to play a role. It appears that neither structural rearrangement, in terms of sugar pucker, nor hydrogen bonding between TFO and duplex backbones, provide a satisfactory explanation for the additional stability of pyrimidine RNA triplexes over their DNA counterparts. Further study of these triplexes is therefore required before the effect of backbone composition on stability can be understood and generalised for the design of more stable TFOs with potential in antigene therapy and other applications.

CHAPTER

SIX

CONCLUSIONS AND FUTURE DIRECTIONS

6.1 CONCLUSIONS

In this work, a comprehensive study of the structure and stability of nucleic acid triple helices has been presented. The stability of these triplexes, focusing on the $(GAA)_n$ sequence known to be associated with the neurodegenerative disease Friedreich's ataxia, was examined from two directions, with aims of both destabilising the structures, through the binding of ligands, and stabilising them, through the choice of triplex–forming oligomer (TFO) sequence and backbone composition. The use of molecular dynamics simulation has provided insight into the processes that occur upon binding of minor-groove binder netropsin, as well as possible explanations for reported relative stabilities of TFOs with differing compositions, both of backbone and bases.

Chapter 3 presented an investigation into the effect of netropsin, which is known to destabilise triplexes, on the triplex structure and thermodynamics. It was shown that the use of free energy methods such as alchemical free energy perturbation (FEP) can calculate, with reasonable accuracy, the expected thermodynamics of large flexible biomolecules. However, it was also found that these systems relax very slowly and that calculated free energy is therefore hugely dependent on simulation time. Using these methods, it was found that netropsin destabilises this triple helix by approximately 15 kcal/mol when bound opposite to the TFO, an effect which appears to be highly localised. This thermodynamic effect was supported by structural analyses which show a localised narrowing of the minor groove for the 4–5 bases where netropsin is bound. This agrees with previous computational results, but is the opposite effect as described by solid-state experimental studies. It therefore appears that the triple helix structure with netropsin may be significantly different in solid and solution phases. Furthermore, the TFO appears to be pushed slightly (2–3 Å) out of the major groove in the region opposite netropsin's binding site. However, this change occurs without a significant loss of hydrogen bonding sites within 3 Å of each other, indicating that this may occur in regions where hydrogen bonding contacts are already poor. Given the observed structural changes, it is likely that the interplay of structural distortions occurring in the triplex structure on netropsin binding play a significant role in its ability to destabilise these structures. However, as this is appears to be a highly localised phenomenon, a high concentration of netropsin would likely be required to achieve significant, therapeutic, destabilisation. As

netropsin can be considered a typical minor-groove binding ligand, these results are likely generalisable to other minor-groove binders.

While netrospin is known to bind in the minor groove of triple helices, binding in either of the other possible grooves formed between TFO and duplex has not previously been reported and it is not clear why it has such a significant preference for the minor groove. This has been investigated, also in chapter 3, with studies of netropsin binding in both the Crick–Hoogsteen (C–H) and Watson–Hoosteen (W–H) grooves formed by the division of the major groove into two asymmetric grooves by TFO binding. Through short free energy calculations, it was found that the strength of the van der Waals interactions between ligand and triplex plays a large role in stabilising the ligand in the minor groove. Relative stabilities of the three possible binding sites were determined as minor > W–H > C–H, although the stability in the minor and W–H grooves was shown to be potentially quite close. The width of the groove was shown to play a significant role, the W–H being too wide for interactions with both walls of the groove and the C–H potentially too narrow, forcing the ligand out of the groove onto the face of the triplex. It therefore appears that a fortunate groove width results in the preference of netropsin for minor groove binding, which is encouraged by loss of solvent molecules from the spine of hydration at the binding site.

While chapter 3 focused on the destabilisation of triple helices, chapters 4 and 5 presented an investigation into select means by which they can be stabilised, with applications in antigene therapies. In chapter 4, the effect of changing the TFO from a purine to pyrimidine sequence was investigated. The A*A reverse Hoogsteen pair prominent in the purine sequence was found to be very unstable, with hydrogen bonds rarely forming between these bases. This was in contrast to the regular structures and hydrogen bonding patterns of the G*G, A*T, and G*C$^+$ equivalents. The overall result of this instability was a shifting of the TFO into the center of the major groove in A–rich regions, while G regions remained close to the duplex purine strand as expected, and may explain the reported requirement of high G content for purine triplex stability. In addition, large structural rearrangements, in terms of sugar pucker, were required for pyrimidine TFO binding, while changes required when the purine sequence bound were small. This may be indicative of a kinetic barrier that may inhibit the formation of the pyrimidine triplex, even if it is more stable. Overall, these results indicate that, at least for the sequence studied here, pyrimidine TFOs should give more stable triplexes, especially if the requirement for cytosine protonation can be overcome, while the purine sequences may form more readily, despite their lower stability. The G content of the duplex is therefore an important parameter when deciding on choice of triplex.

Finally, the effect of backbone composition was investigated in chapter 5, through comparison of triplexes formed with either a DNA or RNA based TFO. It has previously been shown that pyrimidine triplexes with RNA backbones are more stable than their DNA counterparts, while for the purine case DNA is expected to form the more stable triplex. Free energy calculations conducted on the purine triplexes indicated that the DD.R purine triplex should be stable enough to form. The lack of formation reported in the literature can potentially be explained by the observed significant rearrangement of the ribose sugars required for the binding of the purine RNA TFO to the duplex. This, couple with the energy barrier of approximately 2.5 k_BT between C2'-endo and C3'-endo conformations, likely gives a large barrier which is required to be overcome before the triplex can form. The relative stability of the pyrimidine DD.R triplex has been proposed to be due to the potential for hydrogen bonding between ribose's 2' OH group and the phosphate oxygens when in a parallel orientation, and the requirement of less backbone rearrangement to bind. Simulations performed here

indicate that neither of these explanations are satisfactory. The equilibrium OH–OP distance was found to be too great for hydrogen bonding to occur, and the same trends are observed when examining the change in sugar pucker required of either the duplex or the TFO on binding the strands of differing composition. To properly understand the experimentally observed DD.D and DD.R triplexes, further study of the thermodynamics and kinetics of their formation is required.

Combined, this work gives useful insights into the stability of nucleic acid triple helices, and investigates the way certain features can be manipulated to enhance or reduce this stability as desired. The binding of minor groove binder netropsin was shown to locally narrow the minor groove, decreasing the stability of triplex structures when bound in close proximity to the TFO, an effect which is likely generalisable to other minor groove binders. In terms of enhancing stability, it was found that a pyrimidine sequence should give a more stable triplex, particularly if the requirement for protonated cytosines can be overcome for biological applications. However, the purine should form more readily although A*A Hoogsteen pairs appear to significantly disfavour formation of purine triplexes, as they distort the backbone structure of the TFO. Lastly, the DD.R triplex containing purine third strand was unexpectedly found to have similar stability to the DD.D equivalent, despite DD.R with purine TFOs not being observed in experiment. It was found that significant backbone rearrangement of the DD.R purine triplex likely prevents its formation, despite its stability. Furthermore no structure reason could be found to explain the greater relative stability of the pyrimidine DD.R triplex compared with the corresponding DD.D triplex, despite mechanisms for this stability having been previously proposed.

6.2 Future directions

Although this work provides some useful insights into the stability of triplexes, and different methods to modulate, either enhancing, or reduce, this, further simulations may aid to better answer some of these questions, as well as those raised by the results presented here.

6.2.1 Effect of minor groove binders on triplex stability

The work presented in chapter 3 examined the effect of a single minor-groove binder, netropsin, on triplex stability. As minor-groove binders have, in general, been shown to have triplex destabilising properties,[29] this effect can likely be generalisable to other ligands, such as distamycin or Hoechst-33258, which share structural features with netropsin. Simulation of these, or other similar molecules, will provide further insight into the mechanisms of triplex destabilisation by minor-groove binders and potentially lead to the development of some generalisable rules for design of novel drugs for applications in triplex destabilisation. Furthermore, it was shown that netropsin acts locally in destabilising triplexes, likely requiring large concentrations of the ligand for effective therapeutic destabilisation of long, biologically relevant, triplexes. Determining the approximate concentrations required for effective destabilisation is an important next step in this process. The binding of additional netropsin molecules to the triplex should be considered in order to determine an approximate relationship between the number of bases and netropsin molecules required for destabilisation.

Additionally, the binding of netropsin in the other grooves was considered, but it has been shown in chapter 4 that the C–H and W–H grooves of the purine triplex have neither regular width, nor the asymmetry expected. For better understanding of the ligand's preference for the minor groove, it should also be studied in the pyrimidine grooves, which give a structure

closer to expected. More extensive free energy calculations could also provide insight into this preference.

6.2.2 NATURE OF BASES: PURINE OR PYRIMIDINE

The structural analysis outlined in chapter 4 indicated that for the sequence studied, the pyrimidine triplex is likely to be significantly more stable due to large structural rearrangements and distortions required for the binding of its purine equivalent. As this sequence has a relatively low G content, at approximately 33 %, other sequences with differing G content should also be studied to better quantify the effect of sequence on this property. Furthermore, free energy calculations, similar to those described in chapter 3 should be carried out to quantitatively examine the effect of these structural changes on stability. Structural changes required of the purine TFO to bind to the duplex were not large, however the changes required for the pyrimidine TFO were extensive. The free energy landscape of this transition should be considered, in order to determine whether this process is particularly unfavourable. Umbrella sampling simulations, such as those described in chapter 5, however, only consider a single base in the center of the TFO. Other bases, differing in both identity and location, should also be considered, and compared to the purine equivalent for which little sugar pucker rearrangement was required. Other structural properties, such as the helical roll, rise, and twist, may also have greater significance than the sugar puckering, and further study of these more global properties may be of use.

6.2.3 EFFECT OF BACKBONE COMPOSITION

It has been proposed that triplexes containing a pyrimidine third strand with an RNA backbone are more stable than their counterparts with DNA backbones while for the purine equivalents, the DNA triplex is expected to be more stable than the RNA hybrid.[46,53,55] Here we find that for the purine case, even if slightly less stable, the DD.R triplex should form. A significant structural change is required in the sugar pucker of the RNA TFO on binding which may explain why it is not observed experimentally, suggesting that further study of the kinetics of the binding process in this triplex is required to explain why it might not form. The large effect of unbound TFO structure on calculated free energy warrants further study, and whether the globular single stranded nucleic acid structure is an artifact of the force-field should also be examined.

Of the two potential reasons for the enhanced stability of the pyrimidine DD.R triplex over the DD.D, being hydrogen bonding between backbones of RNA TFO and duplex and a larger structural rearrangement required for formation of the DD.D triplex, neither have provided a satisfactory answer to the question of stability. Although the sugar pucker proved to have very little significance in regards to structural rearrangement of the pyrimidine triplex, other structural parameters, as described in section 6.2.2, should be considered, as the slight change in sugar puckering may be indicative of other, more significant changes. The structure of the duplex moiety in particular should be considered in more detail, as any changes required in this upon TFO binding are likely to disfavour triplex formation.[55]

One further difference between the pyrimidine DNA and RNA triplexes that has not been examined here is the presence of uracil in place of thymine which is likely to also influence the stability.[54,122] Free energy calculations, comparing both the DD.D and DD.R structures with either thymine or uracil should be conducted. As methods such as alchemical FEP allow for the sequential mutation of structures, the free energy changes for the cycle DD.D (T) $\rightarrow$

DD.R (T) $\rightarrow$ DD.R (U) $\rightarrow$ DD.D (U) $\rightarrow$ DD.D (T) could be used to examine the relative contributions of the change from thymine to uracil and deoxyribose to ribose on the stability of the triplexes. Although these changes would be small compared to those outlined in the study conducted in chapter 3, in order to obtain free energies accurate enough to confidently assess this process, significant simulation time would likely still be required.

APPENDIX A: NETROSPIN PARAMETERS

Table 6.1: Netropsin charges used in this work, calculated using antechamber.

Name	Type	Charge (e)	Name	Type	Charge (e)	Name	Type	Charge (e)
C1	CZ	0.7985	N5	NA	0.049200	C13	CD	-0.1733
N1	NH	-0.8237	C8	C3	-0.1076	H13	H4	0.2010
HN1	HN	0.4311	H81	H1	0.09190	C15	C	0.6992
HN2	HN	0.4311	H82	H1	0.0919	O3	O	-0.5915
N2	NH	-0.9697	H83	H1	0.0919	N8	N	-0.5820
HHN2	HN	0.4830	C7	CD	-0.1501	HN8	HN	0.3500
H3	HN	0.4830	H7	H4	0.1913	C16	C3	0.0340
N3	NH	-0.3735	C9	C	0.6637	HH16	H1	0.0979
HN3	HN	0.2977	O2	O	-0.5229	H9	H1	0.0979
C2	C3	-0.0556	N6	N	-0.4842	C17	C3	-0.1794
H21	H1	0.0862	HN6	HN	0.3286	HH17	HC	0.1269
H22	H1	0.0862	C10	CC	0.1648	H10	HC	0.1269
C3	C	0.5374	C11	CC	-0.3924	C18	C2	0.5798
O1	O	-0.4952	H11	HA	0.2136	N9	NH	-0.6872
N4	N	-0.3630	C12	CD	-0.1209	HN9	HN	0.4139
HN4	HN	0.3241	N7	NA	0.0713	HHN1	HN	0.4139
C4	CC	0.0495	C14	C3	-0.1081	N10	NH	-0.8269
C5	CC	-0.3673	HH14	H1	0.0883	HHN0	HN	0.4475
H5	HA	0.2023	H4	H1	0.0883	H12	HN	0.4475
C6	CD	-0.0951	H6	H1	0.0883			

Table 6.2: Bond parameters for netropsin. Bond energy is calculated as $U_{\text{bond}} = K_b(b - b_0)^2$ for spring constant K_b and equilibrium bond length b_0

Atom 1	Atom 2	K_b (kcal/mol/Å^2)	b_0 (Å)
CZ	NH	487.80	1.339
HN	NH	401.20	1.014
C3	NH	332.70	1.458
C3	H1	335.90	1.093
C	C3	328.30	1.508
HN	N	410.20	1.009
CC	N	426.00	1.380
CC	CC	418.30	1.429
CC	CD	504.00	1.371
CC	HA	347.20	1.085
CD	NA	438.80	1.371
C	CD	377.40	1.462
C3	NA	334.70	1.456
Cd	H4	350.10	1.083
C3	N	330.60	1.460
C3	C3	303.10	1.535
C3	HC	337.30	1.092
C2	C3	328.30	1.508
C2	NH	462.60	1.355

Table 6.3: Angle parameters for netropsin, calculated as $U_{\text{angle}} = K_\theta(\theta - \theta_0)^2$ for θ_0, the angle made by atoms 1–3, and spring constant K_θ.

Atom 1	Atom 2	Atom 3	K_θ (kcal/mol/rad^2)	θ (°)
CZ	NH	HN	48.790	121.240
CZ	NH	C3	63.010	125.510
NH	CZ	NH	72.970	120.170
HN	NH	HN	40.050	114.850
NH	C3	H1	49.730	109.960
NH	C3	C	67.360	109.430
HN	NH	C3	46.460	114.950
C3	C	O	68.030	123.110
C3	C	N	67.860	115.150
H1	C3	H1	39.180	109.550
H1	C3	C	47.630	107.660
C	N	HN	49.210	118.460
C	N	CC	65.240	124.190
O	C	N	75.830	122.030
N	CC	CC	67.950	119.890
N	CC	CD	70.720	115.520
HN	N	CC	48.080	118.710
CC	CC	HA	47.460	119.260
CC	CC	CD	68.160	114.190
CC	CD	NA	72.910	109.420
CC	CD	H4	47.190	129.110
CC	CD	C	65.280	121.320
HA	CC	CD	48.350	122.890
CD	NA	C3	62.560	125.090
CD	NA	CD	68.940	109.900
CD	C	O	68.910	125.710
CD	C	N	70.190	111.860
NA	CD	C	69.960	114.105
NA	C3	H1	49.900	109.450
NA	CD	H4	50.220	119.660
C	N	C3	63.920	121.350
N	C3	H1	49.820	109.320
N	C3	C3	65.850	112.130
HN	N	C3	46.040	116.780
C3	C3	Hc	46.370	110.050
C3	C3	C2	63.530	111.440
H1	C3	C3	46.360	110.070
C3	C2	NH	66.690	118.590
HC	C3	HC	39.430	108.350
HC	C3	C2	47.030	110.490
C2	NH	HN	49.620	114.890
NH	C2	NH	74.460	112.720

Table 6.4: Dihedral parameters for netropsin, where the dihedral energy is calculated as $U_{\text{dihedral}} = K_\chi(1+\cos(n(\chi)-\delta)$ for χ, the angle between the planes of atoms 1–3 and 2–4, and spring constant K_χ. X represents a wildcard.

Atom 1	Atom 2	Atom 3	Atom 4	K_χ (kcal/mol)	multiplicity (n)	δ (°)
NH	CZ	NH	HN	0.675	2	180.0
NH	CZ	NH	C3	0.675	2	180.0
X	C3	NH	X	0.000	2	0.0
X	C	C3	X	0.000	2	180.0
H1	C3	C	O	0.800	1	0.0
H1	C3	C	O	0.080	3	180.0
X	C	N	X	2.500	2	180.0
HN	N	C	O	2.500	2	180.0
HN	N	C	O	2.000	1	0.0
X	N	CC	X	1.650	2	180.0
X	CC	CC	X	4.000	2	180.0
X	CC	CD	X	4.000	2	180.0
X	CD	NA	X	1.700	2	180.0
X	C	CD	X	2.875	2	180.0
X	C3	NA	X	0.000	2	0.0
X	C3	N	X	0.000	2	0.0
C3	C3	N	C	0.500	4	180.0
C3	C3	N	C	0.150	3	180.0
C3	C3	N	C	0.530	1	0.0
X	C3	C3	X	0.156	3	0.0
X	C2	C3	X	0.000	2	0.0
X	C2	NH	X	0.675	2	180.0

Table 6.5: Improper parameters for netropsin, calculated as $U_{\text{improper}} = K_\psi(\psi-\psi_0)^2$ for ψ_0, the angle between the planes of atoms 1–3 and 2–4, and spring constant K_ψ. X represents a wildcard.

Atom 1	Atom 2	Atom 3	Atom 4	K_ψ (kcal/mol/rad^2)	multiplicity (n)	ψ_0 (°)
NH	NH	CZ	NH	1.1	2	180.0
CZ	HN	NH	HN	1.1	2	180.0
C3	CZ	NH	HN	1.1	2	180.0
X	X	N	HN	1.1	2	180.0
CC	CD	CC	N	1.1	2	180.0
CC	CD	CC	HA	1.1	2	180.0
C	CC	CD	NA	1.1	2	180.0
C3	CD	NA	CD	1.1	2	180.0
CC	H4	CD	NA	1.1	2	180.0
C3	NH	C2	NH	1.1	2	180.0
C2	HN	NH	HN	1.1	2	180.0

BIBLIOGRAPHY

[1] Watson, J. D.; Crick, F. H. C. *Nature* **1953**, *171*, 737–738.

[2] Hoogsteen, K. *Acta Crystallogr.* **1963**, *16*, 907–916.

[3] Parvathy, V. R.; Bhaumik, S. R.; Chary, K. V. R.; Govil, G.; Liu, K.; Howard, F. B.; Miles, H. T. *Nucleic Acids Res.* **2002**, *30*, 1500–1511.

[4] Cubero, E.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **2001**, *123*, 12018–12025.

[5] Dickerson, R. *Nucleic Acids Res.* **1989**, *17*, 1797–1803.

[6] Dickerson, R.; Drew, H.; Conner, B.; Wing, R.; Fratini, A.; Kopka, M. *Science* **1982**, *216*, 475–485.

[7] Frank-Kamenetskii, M. D.; Mirkin, S. M. *Annu. Rev. Biochem.* **1995**, *64*, 65–95.

[8] Doluca, O.; Withers, J. M.; Filichev, V. V. *Chem. Rev.* **2013**, *113*, 3044–3083.

[9] Saini, N.; Zhang, Y.; Usdin, K.; Lobachev, K. S. *Biochimie* **2013**, *95*, 117–123.

[10] Paugh, S. W. et al. *PLoS Comput. Biol.* **2016**, *12*, 1–20.

[11] Belotserkovskii, B. P.; Liu, R.; Tornaletti, S.; Krasilnikova, M. M.; Mirkin, S. M.; Hanawalt, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 12816–12821.

[12] Bacolla, A.; Wells, R. D. *J. Biol. Chem.* **2004**, *279*, 47411–47414.

[13] Tsuda, N.; Matsumoto, A.; Ito, A.; Uneda, T.; Tanabe, A.; Obika, S.; Imanishi, T. *Nucleic Acids Symp. Ser.* **2005**, *49*, 335–336.

[14] Hélène, C.; Thuong, N. T.; Harel, A. *Ann. N. Y. Acad. Sci.* **1992**, *660*, 27–36.

[15] Wells, R. D. *Trends Biochem. Sci.* **2007**, *32*, 271–278.

[16] Grabczyk, E.; Usdin, K. *Nucleic Acids Res.* **2000**, *28*, 2815–2822.

[17] Campuzano, V.; Montermini, L.; Molto, M. D.; Pianese, L.; Cossee, M.; Cavalcanti, F.; Monros, E.; Rodius, F.; Duclos, F.; et al., *Science* **1996**, *271*, 1423–1427.

[18] Patel, P. I.; Isaya, G. *Am. J. Hum. Gene.* **2001**, *69*, 15–24.

[19] Felsenfeld, G.; Davies, D. R.; Rich, A. *J. Am. Chem. Soc.* **1957**, *79*, 2023–2024.

[20] Laughton, C.; Neidle, S. *Nucleic Acids Res.* **1992**, *20*, 6535–6541.

[21] Arcella, A.; Portella, G.; Ruiz, M. L.; Eritja, R.; Vilaseca, M.; Gabelica, V.; Orozco, M. *J. Am. Chem. Soc.* **2012**, *134*, 6596–6606.

[22] Mariappan, S. S.; Cheng, X.; van Breemen, R. B.; Silks, L. A.; Gupta, G. *Anal. Biochem.* **2004**, *334*, 216–226.

[23] Plum, E. G.; Breslauer, K. J. *J. Mol. Biol.* **1995**, *248*, 679–695.

[24] Roberts, R. W.; Crothers, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 4320–4325.

[25] Park, Y.-W.; Breslauer, K. J. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 6653–6657.

[26] Durand, M.; Thuong, N. T.; Maurizot, J. C. *J. Biol. Chem.* **1992**, *267*, 24394–24399.

[27] Escudé, C.; Nguyen, C. H.; Kukreti, S.; Janin, Y.; Sun, J.-S.; Bisagni, E.; Garestier, T.; Hélène, C. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 3591–3596.

[28] Keppler, M. D.; James, P. L.; Neidle, S.; Brown, T.; Fox, K. R. *Eur. J. Biochem.* **2003**, *270*, 4982–4992.

[29] Jain, A. K.; Bhattacharya, S. *Bioconjug. Chem.* **2010**, *21*, 1389–1403.

[30] Stonehouse, T. J.; Fox, K. R. *BBA - Gene Structure and Expression* **1994**, *1218*, 322–330.

[31] Durand, M.; Thuong, N.; Maurizot, J. *Biochimie* **1994**, *76*, 181–186.

[32] Wartell, R. M.; Larson, J. E.; Wells, R. D. *J. Biol. Chem.* **1974**, *249*, 6719–6731.

[33] Van Hecke, K.; Nam, P. C.; Nguyen, M. T.; Van Meervelt, L. *FEBS J.* **2005**, *272*, 3531–3541.

[34] Nunn, C. M.; Garman, E.; Neidle, S. *Biochemistry* **1997**, *36*, 4792–4799.

[35] Wellenzohn, B.; Flader, W.; Winger, R. H.; Hallbrucker, A.; Mayer, E.; Liedl, K. R. *Biopolymers* **2002**, *61*, 276–286.

[36] Dolenc, J.; Oostenbrink, C.; Koller, J.; van Gunsteren, W. F. *Nucleic Acids Res.* **2005**, *33*, 725–733.

[37] Kopka, M. L.; Yoon, C.; Goodsell, D.; Pjura, P.; Dickerson, R. E. *Proc. Natl. Acad. Sci. U. S. A.* **1985**, *82*, 1376–1380.

[38] Soliva, R.; Laughton, C. A.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **1998**, *120*, 11226–11233.

[39] Avendaño, C.; Menéndez, J. C. In *Medicinal Chemistry of Anticancer Drugs*; Carmen, A., , Menéndez, J. C., Eds.; Elsevier: Amsterdam, 2008; pp 177–198.

[40] Nielsen, P. E. *Bioconjug. Chem.* **1991**, *2*, 1–12.

[41] Chalikian, T. V.; Plum, G. E.; Sarvazyan, A. P.; Breslaver, K. J. *Biochemistry* **1994**, *33*, 8629–8640.

[42] Buske, F. A.; Bauer, D. C.; Mattick, J. S.; Bailey, T. L. *Genome Res.* **2012**, *22*, 1372–1381.

[43] Evans, K.; Bhamra, I.; Wheelhouse, R. T.; Arnold, J. R. P.; Cosstick, R.; Fisher, J. *Chemistry – A European Journal* **2015**, *21*, 7278–7284.

[44] Faruqi, A. F.; Krawczyk, S. H.; Matteucci, M. D.; Glazer, P. M. *Nucleic Acids Res.* **1997**, *25*, 633–640.

[45] Vasquez, K. M.; Wensel, T. G.; Hogan, M. E.; Wilson, J. H. *Biochemistry* **1995**, *34*, 7243–7251.

[46] Mills, M.; Arimondo, P. B.; Lacroix, L.; Garestier, T.; Klump, H.; Mergny, J.-L. *Biochemistry* **2002**, *41*, 357–366.

[47] Vekhoff, P.; Ceccaldi, A.; Polverari, D.; Pylouster, J.; Pisano, C.; Arimondo, P. B. *Biochemistry* **2008**, *47*, 12277–12289.

[48] Sun, J.-S.; Hélène, C. *Curr. Opin. Struct. Biol.* **1993**, *3*, 345–356.

[49] Seela, F.; Shaikh, K. I. *Org. Biomol. Chem.* **2006**, *4*, 3993–4004.

[50] Debin, A.; Laboulais, C.; Ouali, M.; Malvy, C.; Le Bret, M.; Svinarchuk, F. *Nucleic Acids Res.* **1999**, *27*, 2699–2707.

[51] Wang, S.; Kool, E. T. *Nucleic Acids Res.* **1995**, *23*, 1157.

[52] Roberts, R. W.; Crothers, D. M. *Science* **1992**, *258*, 1463–1466.

[53] Gotfredsen, C. H.; Schultze, P.; Feigon, J. *J. Am. Chem. Soc.* **1998**, *120*, 4281–4289.

[54] Han, H.; Dervan, P. B. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 3806–3810.

[55] Asensio, J. L.; Carr, R.; Brown, T.; Lane, A. N. *J. Am. Chem. Soc.* **1999**, *121*, 11063–11070.

[56] Escudé, C.; Francçois, J.-C.; Sun, J.-S.; Ott, G.; Sprinzl, M.; Garestier, T.; Hélène, J.-C. *Nucleic Acids Res.* **1993**, *21*, 5547.

[57] Semerad, C. L.; Maher, L. J. *Nucleic Acids Res.* **1994**, *22*, 5321–5325.

[58] Suresh, G.; Priyakumar, U. D. *Phys. Chem. Chem. Phys.* **2014**, *16*, 18148–18155.

[59] Shimizu, M.; Konishi, A.; Shimada, Y.; Inoue, H.; Ohtsuka, E. *FEBS Lett.* **1992**, *302*, 155–158.

[60] Cheatham, T. E.; Kollman, P. A. *J. Am. Chem. Soc.* **1997**, *119*, 4805–4825.

[61] Jain, A.; Rajeswari, M. R.; Ahmed, F. *J. Biomol. Struct. Dyn.* **2002**, *19*, 691–699.

[62] Medhi, C.; Parajuli, R. *Journal of Molecular Structure: THEOCHEM* **2005**, *717*, 59 – 66.

[63] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

[64] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press, 2001.

[65] Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

[66] Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

[67] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. *J. Comput. Phys.* **1977**, *23*, 327–341.

[68] Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D. *J. Chem. Theory Comput.* **2012**, *8*, 348–362.

[69] Cheatham, T. E.; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862.

[70] Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829.

[71] Ivani, I. et al. *Nat. Methods* **2016**, *13*, 55–58.

[72] Zgarbová, M.; Šponer, J.; Otyepka, M.; Cheatham, T. E.; Galindo-Murillo, R.; Jurečka, P. *J. Chem. Theory Comput.* **2015**, *11*, 5723–5736.

[73] Galindo-Murillo, R.; Robertson, J. C.; Zgarbová, M.; Šponer, J.; Otyepka, M.; Jurečka, P.; Thomas E. Cheatham, I. *J. Chem. Theory Comput.* **2016**, *12*, 4114–4127.

[74] Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[75] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

[76] Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D. *Chem. Phys. Lett.* **2006**, *418*, 245–249.

[77] Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

[78] Brown, J. A.; Bulkley, D.; Wang, J.; Valenstein, M. L.; Yario, T. A.; Steitz, T. A.; Steitz, J. A. *Nat Struct Mol Biol* **2014**, *21*, 633–640.

[79] Olsen, G. L.; Louie, E. A.; Drobny, G. P.; Sigurdsson, S. T. *Nucleic Acids Res.* **2003**, *31*, 5084–5089.

[80] Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33–38.

[81] Harvey, S. C.; Prabhakaran, M. *J. Am. Chem. Soc.* **1986**, *108*, 6128–6136.

[82] Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

[83] Altona, C.; Sundaralingam, M. *J. Am. Chem. Soc.* **1972**, *94*, 8205–8212.

[84] Huang, M.; Giese, T. J.; Lee, T.-S.; York, D. M. *J. Chem. Theory Comput.* **2014**, *10*, 1538–1545.

[85] Genheden, S.; Ryde, U. *Expert Opin. Drug Discov.* **2015**, *10*, 449–461.

[86] Dolenc, J.; Baron, R.; Oostenbrink, C.; Koller, J.; van Gunsteren, W. F. *Biophys. J.* **2006**, *91*, 1460–1470.

[87] Schlitter, J. *Chem. Phys. Lett.* **1993**, *215*, 617–621.

[88] Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

[89] Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. *J. Chem. Phys.* **2003**, *119*, 3559–3566.

[90] Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

[91] Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *100*, 9025–9031.

[92] Pohorille, A.; Jarzynski, C.; Chipot, C. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.

[93] Pearlman, D. A. *J. Phys. Chem.* **1994**, *98*, 1487–1493.

[94] Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[95] Sambriski, E.; Schwartz, D.; de Pablo, J. *Biophys. J.* **2009**, *96*, 1675–1690.

[96] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

[97] Grossfield, A. *http://membrane.urmc.rochester.edu/content/wham/*

[98] Grossfield, A.; Zuckerman, D. M. *Annual reports in computational chemistry* **2009**, *5*, 23–48.

[99] Case, D. et al. AMBER 2017, University of California, San Francisco. 2017.

[100] Wang, J.; Wang, W.; A., K. P.; Case, D. A. *J. Mol. Graph. Model.* **2006**, *25*.

[101] Jakalian, A.; Bush, B.; Jack, B.; Bayly, C. *J. Comp. Chem.* **2000**,

[102] Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798–2814.

[103] Vergara-Jaque, A.; Comer, J.; Monsalve, L.; González-Nilo, F. D.; Sandoval, C. *J. Phys. Chem. B* **2013**, *117*, 6801–6813.

[104] Häse, F.; Zacharias, M. *Nucleic Acids Res.* **2016**, *44*, 7100.

[105] Lomzov, A. A.; Vorobjev, Y. N.; Pyshnyi, D. V. *J. Phys. Chem. B* **2015**, *119*, 15221–15234.

[106] Harris, L.-A.; Williams, L. D.; Koudelka, G. B. *Nucleic Acids Res.* **2014**, *42*, 14053–14059.

[107] Hamelberg, D.; McFail-Isom, L.; Williams, L. D.; Wilson, W. D. *J. Am. Chem. Soc.* **2000**, *122*, 10513–10520.

[108] Shields, G. C.; Laughton, C. A.; Orozco, M. *J. Am. Chem. Soc.* **1997**, *119*, 7463–7469.

[109] Buske, F. A.; Mattick, J. S.; Bailey, T. L. *RNA Biol.* **2011**, *8*, 427–439.

[110] Tung, C.-S.; Carter, E. S., II *Bioinformatics* **1994**, *10*, 427.

[111] Lu, X. J.; Olson, W. K. *Nucleic Acids Res.* **2003**, *31*, 5108–5121.

[112] Lu, X. J.; Olson, W. K. *Nature Protocols* **2008**, *3*, 1213–1227.

[113] Goh, G. B. Development and Application of Constant pH Molecular Dynamics (CPHMD$^{MS\lambda D}$) for investigating pH-mediated Transient conformational state and their effects on nucleic acid and protein activity. Ph.D. thesis, 2015.

[114] Peters, M.; Rozas, I.; Alkorta, I.; Elguero, J. *J. Phys. Chem. B* **2003**, *107*, 323–330.

[115] Lee, J. S.; Woodsworth, M. L.; Latimer, L. J.; Morgan, A. R. *Nucleic Acids Res.* **1984**, *12*, 6603–6614.

[116] Ono, A.; Ts'o, P. O. P.; Kan, L. S. *J. Am. Chem. Soc.* **1991**, *113*, 4032–4033.

[117] Devi, G.; Yuan, Z.; Lu, Y.; Zhao, Y.; Chen, G. *Nucleic Acids Res.* **2014**, *42*, 4008–4018.

[118] Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem.* **1996**, *100*, 1206–1215.

[119] Fiorin, G.; Klein, M. L.; Hénin, J. *Mol. Phys.* **2013**, *111*, 3345–3362.

[120] Plumridge, A.; Meisburger, S. P.; Pollack, L. *Nucleic Acids Res.* **2016**, *45*.

[121] Sim, A. Y. L.; Lipfert, J.; Herschlag, D.; Doniach, S. *Phys. Rev. E* **2012**, *86*, 021901.

[122] Povsic, T. J.; Dervan, P. B. *J. Am. Chem. Soc.* **1989**, *111*, 3059–3061.