

https://dspace.org/

# The characterization and crystallization of the TBR1 T-box domain in the presence and absence of the T-box Binding Element

**Mayet, Riyaadh**

**2023**
https://hdl.handle.net/10539/37000

Downloaded from a DSpace repository. https://dspace.org/




# The characterization and crystallization of the TBR1 T-box domain in the presence and absence of the T-box Binding Element

by

**Riyaadh Mayet**

**(1063698)**

**Dissertation**

Submitted in fulfilment of the requirements for the degree

**Master of Science**

in

**Molecular and Cell Biology**

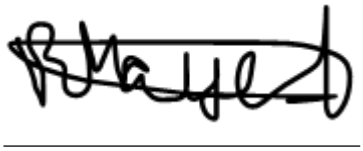
In the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Dr Sylvia Fanucchi

July 2022

# Declaration

I, **Riyaadh Mayet (1063698)**, declare that this dissertation is my own work, unaided work. It is being submitted for the degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.



05/07/2022

## Abstract

TBR1 is a neuron-specific transcription factor involved in numerous developmental events in the brain. It has recently emerged as a master regulator of the genes implicated in autism spectrum disorders. The protein contains an evolutionarily conserved DNA-binding domain, known as the T-box, and binds to a consensus DNA sequence known as the T-box Binding Element. The key to understanding the function of a macromolecule, such as the TBR1 T-box domain, is to determine and understand its structure at all levels. The aim of this study was to determine the DNA-binding mechanism of the protein through structural characterizations, DNA-binding studies, X-ray crystallography and computational methods. The TBR1 T-box domain was successfully overexpressed in *E. coli*. The protein was purified by liquid chromatography, and its purity was confirmed using SDS-PAGE and absorbance spectroscopy. The protein was confirmed to be correctly folded through intrinsic tryptophan fluorescence. The secondary structure and thermal stability were characterized by far UV circular dichroism. The protein was β-sheeted and had a $T_m$ of 63 °C. The secondary and tertiary structures of the protein are conserved upon DNA-binding. Under reducing conditions, the protein is monomeric in solution and binds the DNA as a monomer. Furthermore, the protein binds the DNA with high affinity in the nanomolar range ($K_D$ = 179.6 nM), and the affinity is unaffected by the presence of $Mg^{2+}$. After several rounds of optimization, very thin plate-like protein crystals were obtained. These crystals did not yield any significant diffraction. Protein modelling, disorder predictions and molecular docking were then used to predict the structures of the protein in the presence and absence of DNA. The structure of the protein, both in the presence and absence of DNA, was very similar to other T-box proteins. DNA-binding results in conformational changes in the side chains of residues present in the protein-DNA interface. The TBR1 T-box domain uses the same DNA-binding mechanism utilized by the TBX5 T-box domain. The protein contacts the DNA in the minor groove by insertion of helix $3_{10}$C, which is an inducible recognition element that only becomes structured upon DNA-binding. The results were also used to make structural interpretations of pathogenic point mutations in the TBR1 T-box domain.

## Dedications

**– To my people –**

Mom

Nani and Nana

Mohammed

Humairaa

Abii

Odiloo

*“I am not afraid of storms, for I am learning how to sail my ship”*

-Louisa May Alcott

# Acknowledgements

**To my supervisor, Dr Sylvia Fanucchi.** Thank you for all your unwavering support and guidance throughout my academic journey. I have thoroughly enjoyed the time spent working through all our crazy crystal problems together and I hope we can have more successful endeavours in the future.

**To Professor Yaseen Sayed and Dr Ikechukwu Achilonu.** Thank you for all your support, time and for having confidence in my abilities. The PSFRU is a wonderful learning environment thanks to your dedication and resourcefulness.

**To all my friends at the PSFRU.** Thank you for allowing me to ask so many questions.

**To my mentors Ashleigh Blaine, Heather Donald, Aasiya Lakhi, and Dr Monare Thulo.** Thank you for always trying to answer all my questions and taking the time to help me solve my problems. I have learned a great deal from you all.

**To my family and friends.** There are too many things to say thank you for! Thank you for allowing me the freedom to pursue my dreams.

**To my supervisor, Dr Sylvia Fanucchi, and the National Research Foundation.** Thank you for your financial assistance without which this project would not have been possible.

# Research outputs

**Biophysics in Africa, African Physical Society**:

- Oral presentation
- The characterization and crystallization of the TBR1 T-box domain in the presence and absence of T-box DNA
- Riyaadh Mayet and Dr Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 22 – 26 March 2021

**12th Wits Cross Faculty Postgraduate Symposium**:

- Poster presentation
- The characterization and crystallization of the TBR1 T-box domain in the presence and absence of T-box DNA
- Riyaadh Mayet and Dr Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 26 – 28 July 2021

# Contents

Declaration ..... i
Abstract ..... ii
Dedications ..... iii
Acknowledgements ..... iv
Research outputs ..... v
Contents ..... vi
List of Abbreviations ..... viii
List of Figures ..... xii
List of Tables ..... xv
1. Introduction ..... 1
1.1. Gene expression ..... 1
1.2. The importance of gene regulation in development ..... 2
1.3. Transcription factors ..... 4
1.3.1. The structure of human DNA ..... 4
1.3.2. The structure and function of a transcription factor ..... 5
1.3.3. DNA-binding domains ..... 8
1.3.4. Mechanisms of transcription ..... 11
1.3.5. The regulation of transcription factors ..... 13
1.4. The TBR1 T-box transcription factor ..... 14
1.4.1. The T-box family ..... 15
1.4.2. The T-box DNA-binding domain ..... 19
1.4.3. TBR1 T-box protein-protein interactions ..... 24
1.4.4. Clinical significance ..... 26
2. Rationale ..... 28
3. Aim and objectives ..... 30
4. Materials and methods ..... 31
4.1. Protein preparation ..... 32
4.1.1. Plasmid and protein construct ..... 32
4.1.2. Transformation ..... 34
4.1.3. Heterologous protein expression ..... 35
4.1.4. Purification ..... 36
4.1.5. Assessment of protein purity and concentration ..... 41
4.2. DNA preparation ..... 43
4.3. Characterization of protein structure and stability ..... 44
4.3.1. Circular dichroism spectropolarimetry ..... 44

4.3.2. Intrinsic tryptophan fluorescence spectroscopy ..... 46
4.4. DNA-binding studies ..... 48
4.4.1. Electrophoretic mobility shift assay ..... 48
4.4.2. Fluorescence anisotropy ..... 49
4.5. Protein crystallography ..... 51
4.5.1. Protein crystallization ..... 55
4.6. *In silico* analysis ..... 57
4.6.1. *Ab initio* protein modelling ..... 57
4.6.2. Disorder predictions ..... 59
4.6.3. Molecular docking ..... 59
5. Results ..... 61
5.1. Protein preparation ..... 61
5.1.1. Plasmid sequencing ..... 61
5.1.2. Protein purification and SDS-PAGE ..... 62
5.1.3. Protein purity and concentration determination ..... 66
5.2. Characterization of protein structure and stability ..... 67
5.2.1. Intrinsic tryptophan fluorescence spectroscopy ..... 67
5.2.2. Circular dichroism spectropolarimetry ..... 70
5.2.3. Thermal stability ..... 72
5.3. DNA-binding studies ..... 74
5.3.1. Electrophoretic mobility shift assay ..... 74
5.3.2. Fluorescence anisotropy ..... 75
5.4. Protein crystallography ..... 77
5.5. *In silico* analysis ..... 85
5.5.1. *Ab initio* protein modelling ..... 86
5.5.2. Disorder predictions ..... 93
5.5.3. Molecular docking ..... 94
6. Discussion ..... 98
6.1. The structure and DNA-binding mechanism of the TBR1 T-box domain ..... 99
6.2. The crystallization of the TBR1 T-box domain ..... 105
7. Conclusion ..... 109
8. References ..... 110

# List of Abbreviations

$A_{260}$ – absorbance at 260 nm

$A_{280}$ – absorbance at 280 nm

ADP – adenosine diphosphate

APS – ammonium persulfate

ASD – autism spectrum disorder

ATP – adenosine triphosphate

AU – absorbance units

BLAST – Basic Local Alignment Search Tool

CaBLAM – C-alpha Based Low-resolution Annotation Method

CASP – Critical Assessment of Structure Prediction

CD – circular dichroism

CNS – central nervous system

CRE – *cis*-regulatory element

CRR – *cis*-regulatory region

DBD – DNA-binding domain

DNA – deoxyribonucleic acid

dsDNA – double-stranded deoxyribonucleic acid

DTT - dithiothreitol

EDTA – ethylenediaminetetraacetic acid

EMSA – electrophoretic mobility shift assay

ExPASy – Expert Protein Analysis System

FA – fluorescence anisotropy

FT – flow through

GC-rich – guanine and cytosine rich

GDP – guanosine diphosphate

G-factor – grating factor

GTP – guanosine triphosphate

HADDOCK – High Ambiguity Driven Docking

HAT – histone acetyltransferase

HD – hanging drop

HDAC – histone deacetylase

HEPES - 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HT – high tension

HTH – helix-turn-helix

IDA – iminodiacetic acid

IDR – intrinsically disordered region

Ig – immunoglobulin

IMAC – immobilized metal-ion affinity chromatography

IPTG - isopropyl ß-D-1-thiogalactopyranoside

ITF – intrinsic tryptophan fluorescence

LDDT – Local Distance Difference test

LZ – leucine-zipper

MRE – mean residue ellipticity

mRNA – messenger ribonucleic acid

MWM – molecular weight marker

NMR – nuclear magnetic resonance

PAGE – polyacrylamide gel electrophoresis

PDB – Protein Data Bank

PEG – polyethylene glycol

PIC – pre-initiation complex

PMSF – phenylmethylsulfonyl fluoride

PrDOS – Protein Disorder prediction Server

PSSM – position specific scoring matrix

RE – regulatory element

RMSD – root mean square deviation

RNA – ribonucleic acid

RNA pol II – RNA polymerase II

ROX – 6-Carboxyl-X-Rhodamine

Rpm – revolutions per minute

SD – sitting drop

SDS – sodium dodecyl sulphate

SEC – size exclusion chromatography

SOC – Super Optimal broth with Catabolite repression

SSD – signal sensing domain

SSL – single-site long

SVM – support vector machine

SW – salt wash

TAD – *trans*-activation domain

TBE – T-box Binding Element

TBR1 – T Brain-Related protein 1

TEMED - tetramethylethylenediamine

TF – transcription factor

TRE – *trans*-regulatory element

TSS – transcription start site

UV – ultraviolet

VDRC – vapor diffusion rate control

XRC – X-ray crystallography

YT – yeast tryptone

ZF – zinc-finger

# List of Figures

Figure 1. A schematic representation of a typical gene regulatory network. .....3

Figure 2. Simple representation of a prototypical transcription factor. .....5

Figure 3. Cartoon representation of the structural motifs used to bind DNA. .....9

Figure 4. Multiple sequence alignment of the T-box domains of the T-box family. .....16

Figure 5. Phylogenetic tree for the T-box family of transcription factors. .....17

Figure 6. An annotated crystal structure of TBX5 T-box domain in the DNA-bound form. .....20

Figure 7. Structural alignment of the T-box domains from TBX1, TBX3 and TBX5 in the DNA-bound state. .....21

Figure 8. Crystal structures of the DNA-free and DNA-bound TBX5 T-box domain. .....23

Figure 9. Schematic representation of the pET-11a plasmid with the TBR1 T-box insert. .....33

Figure 10. Putative structure of the metal co-ordination complex used to purify proteins in IMAC. .....38

Figure 11. A depiction of the principles behind size-exclusion chromatography. .....39

Figure 12. A typical solubility phase diagram used to explain the thermodynamics of crystal formation. .....53

Figure 13. The results from Sanger sequencing shown as a protein sequence alignment. .....62

Figure 14. Immobilized metal-ion affinity chromatography purification of the TBR1 T-box domain. .....63

Figure 15. Size exclusion chromatography purification of the TBR1 T-box domain. .....65

Figure 16. Absorbance spectrum of the TBR1 T-box domain. ..... 66

Figure 17. Intrinsic tryptophan fluorescence spectra of the TBR1 T-box domain in the presence and absence of a denaturant. ..... 68

Figure 18. Intrinsic tryptophan fluorescence spectra of the TBR1 T-box domain in the presence and absence of SSL DNA. ..... 69

Figure 19. Circular dichroism spectra of the TBR1 T-box domain in the presence and absence of SSL DNA. ..... 71

Figure 20. Thermal unfolding curves of the TBR1 T-box domain in the presence and absence of SSL DNA. ..... 73

Figure 21. Electrophoretic mobility shift assay showing the TBR1 T-box domain bound to SSL DNA. ..... 75

Figure 22. Fluorescence anisotropy DNA-binding assay of the TBR1 T-box domain and SSL DNA with and without $Mg^{2+}$. ..... 76

Figure 23. A shower of tiny protein-DNA microcrystals resulting from unwanted excessive nucleation. ..... 78

Figure 24. A shower of larger protein-DNA microcrystals obtained by reducing the protein concentration. ..... 79

Figure 25. Large protein-DNA crystals obtained by crystal seeding of tiny microcrystals. ..... 82

Figure 26. A bouquet of plate-like protein-DNA crystals obtained by vapor diffusion rate control. ..... 84

Figure 27. Very thin plate-like protein-DNA crystals before and after staining with Izit Crystal Dye. ..... 85

Figure 28. The predicted structure of the TBR1 T-box domain obtained from RoseTTaFold. 88

Figure 29. The error estimate plot used to validate the local accuracy of the predicted TBR1 T-box domain structure. ..... 89

Figure 30. Structural alignment of the predicted TBR1 T-box domain with the crystal structure of the TBX5 T-box domain. ..... 92

Figure 31. Disorder probability prediction for the TBR1 T-box domain. ..... 94

Figure 32. Structural alignment of the TBX5 crystal structure with TBX5 docked to its DNA using HADDOCK. ..... 975

Figure 33. The predicted structure of the TBR1 T-box domain in the presence of single site long DNA**.** ..... 97

Figure 34. The predicted structure of the TBR1 T-box domain showing the location of the tryptophan residues. ..... 978

# List of Tables

Table 1. All-atom contact analysis of the predicted TBR1T-box domain structure. ..... 91

# 1. Introduction

## 1.1. Gene expression

Gene expression is the phenotypic manifestation of genes, described as the process by which genetic information is used to synthesize functional gene products such as proteins (Crick, 1958). The central dogma of molecular biology explains how these gene products are synthesized from sequence information contained within the DNA (Crick, 1970). The first step in gene expression is transcription. In this RNA polymerase-mediated process, the information contained in DNA is used to synthesize an mRNA (messenger RNA) molecule, through the combined action of the transcriptional machinery (Georges et al., 2010).

Genes can be expressed in a constitutive or facultative manner. Constitutive, or "housekeeping" genes, are continuously transcribed in all cells at a constant rate and are responsible for the maintenance of basic cellular functions such as respiration and metabolism. Facultative genes have variable activity and are only transcribed as and when they are required, such as in response to environmental changes (Geisel, 2011; Latchman, 2005). Since these genes are only transcribed at specific times, in particular cells, and at varying rates, it makes sense for facultative gene transcription to be highly regulated. This is evidenced by the fact that the DNA content of all cells is roughly the same while the mRNA content is not (Alberts et al., 2002; Latchman, 2020). Transcriptional regulation plays a significant role in vital life processes such as cell cycle progression, development, and cellular differentiation (Levine, 2010). The spatiotemporal regulation used in these processes is brought about by a combination of *cis*-regulatory regions and *trans*-regulatory elements in the DNA.

*Trans*-regulatory elements are regions of the DNA which regulate the transcription of distal genes by encoding for DNA-binding proteins known as transcription factors (TFs) (Gilad et al., 2008). TFs bind to various *cis*-regulatory response elements in the *cis*-regulatory regions, where they regulate the transcription of an adjacent gene (Teif, 2010). TFs can activate (by binding to enhancer regions) or repress (by binding to silencer regions) transcription by promoting or inhibiting, respectively, the formation and recruitment of the preinitiation complex to the transcription start site (Luse, 2013; Ranish et al., 1999). The importance of TFs is evidenced by the fact that the more complex vertebrates have more transcription factors per gene than the less complex invertebrates, and that the most complex multicellular

organisms have the most diverse, complex and evolutionarily advanced range of TFs (Levine and Tjian, 2003; Mendoza et al., 2013). Other factors can also influence the formation and recruitment of the preinitiation complex, such as alternative splicing, post translational modifications of TFs and epigenetic effects in the *cis*-regulatory regions (Benayoun and Veitia, 2009; Mortillaro et al., 1996; Pan et al., 2010).

### 1.2. The importance of gene regulation in development

The human brain is perhaps the most important and complex organ in the human body. Its importance stems from the fact that it is responsible for a myriad of functions including the generation and control of the sensory and motor system, regulation of autonomic functions, and control of thought processes such as emotion, cognition, and language (Hall and Guyton, 2011; Hickok, 2009; Lindquist et al., 2012; Nestler et al., 2015). Key developmental events in the brain, such as the differentiation of neural structures, result in the acquisition of cognition and language. Since these events are governed by a complex genetic architecture, cognition and language may have a neuromolecular basis (Graham and Fisher, 2013).

One of the landmarks of brain development is the differentiation of neurons into their respective cell types. Cell differentiation occurs as a result of facultative gene expression, and since this is highly dependent on gene regulation, differentiation depends on the gene regulatory network (Leon and Davidson, 2007).

A typical gene regulatory network is depicted in ***Figure 1***, to show how gene transcription is highly regulated through a variety of mechanisms. In the cascade on the left of ***Figure 1***, a signal molecule (which may be peptide or chemical in nature) binds to a receptor protein which is either located in the cell membrane or within the cytoplasm itself. This activates a cascade of protein kinases which interact with each other in a stepwise fashion, often activating downstream members by phosphorylation. Ultimately, this leads to the activation of transcription factor A (TF A) which can now bind *cis*-regulatory elements, found upstream of the gene to be transcribed, and act as either an activator or repressor of transcription (Pierce, 2012). The cascade on the right of ***Figure 1*** shows that the regulatory network has many nodes, and that a product of gene expression can be used to regulate that gene itself. Firstly, the expressed protein may act as a signal molecule which activates a phosphorylation cascade leading to the activation and DNA-binding of a TF, much in the same way as the cascade on the left of ***Figure 1***. Secondly, the expressed protein can sequester an inhibitor of

TF B thereby allowing TF B to freely bind DNA (Latchman, 1996). Lastly, the expressed protein can take part in a feedback loop in which it binds directly to the *cis*-regulatory elements, thereby blocking, or promoting the association of the TFs, respectively, with the DNA (Davidson and Levin, 2005). These are a few ways in which a gene regulatory network might govern differentiation into the various neuronal cell types.

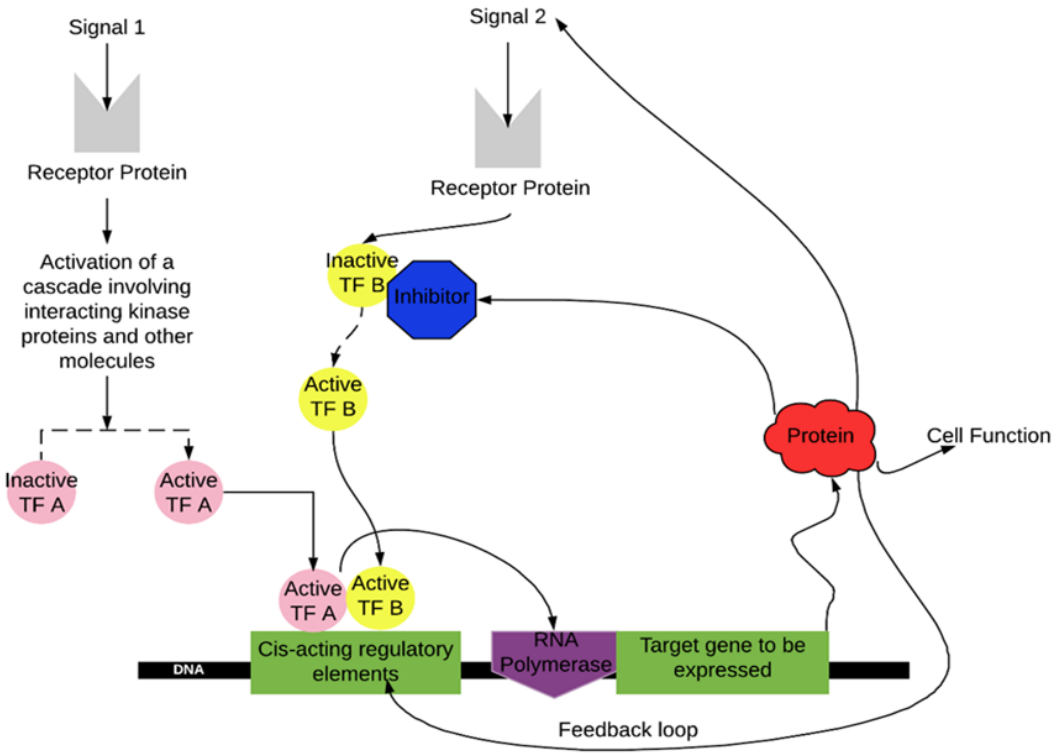



***Figure 1. A schematic representation of a typical gene regulatory network.*** *The levels of mRNA and/or protein expressed is regulated by the interaction of TFs with each other, the DNA, as well as with other molecules. In the cascade on the left, the binding of signal 1 (which could be a ligand or a peptide) to a receptor protein activates a cascade of protein kinases which ultimately activate TF A, allowing it to bind to the cis-acting regulatory elements within the DNA. In the cascade on the right, the protein that is expressed can regulate the very gene which transcribed it through the various mechanisms shown above. TF A, TF B, signal 1 and signal 2 are random in nature and have been used for the purposes of this illustration.*

It can therefore be seen that TFs are key role players in the gene regulatory network, and they may act directly or indirectly to upregulate or downregulate the transcription of genes, via association with other TFs and proteins as well as the DNA itself. Hence, the differentiation of neural structures into the various cell types of the brain is governed by the activity of the neuron-specific TFs.

## 1.3. Transcription factors

TFs are multi-domain proteins that control the rate of gene transcription by directly binding to specific DNA sequences found either up- or downstream of the gene to be transcribed (Karin, 1990; Latchman, 1997). They make up approximately 10% of the 20 000 proteins encoded for in the human genome (Lambert et al., 2018; Ponomarenko et al., 2016). The fact that TFs constitute the single largest family of human proteins alludes to their centralized role in cellular function, making them clinically relevant. At the transcriptional level, gene regulation ensures that constitutive gene expression is prevented when facultative expression is required, thereby preventing wasteful energy expenditure in an organism.

### 1.3.1. The structure of human DNA

To understand how the structure of a TF is suited to its DNA-binding function, it is first necessary to consider the structure of human DNA. Human DNA exists predominantly in the B-DNA form (Kastenholz et al., 2006). It is made up of two antiparallel polynucleotide strands (each nucleotide consists of a deoxyribose sugar, a phosphate, and a nitrogenous base) which wrap around a common central axis with a right-handed twist. The two strands wrap around each other and can only be separated by unwinding the helix (Travers and Muskhelishvili, 2015). The nitrogenous bases occupy the core of the helix, while the sugar-phosphate moieties wrap around the periphery thereby minimizing the electrostatic repulsion between the negatively charged phosphate groups. The planes of the bases are oriented almost perpendicularly to the helix-axis, and each base is hydrogen-bonded to a base on the opposite side (Schweitzer and Kool, 1995). As per Chargaff's rules, adenine preferably binds to thymine while cytosine preferably binds to guanine (Rudner et al., 1968).

B-DNA has two surface-exposed grooves, called the major groove and minor groove, which wind between the sugar-phosphate chains as a result of the helix-axis passing through the centre of each base (Lavery et al., 2009). The grooves are of unequal size because the top edge of each base pair is structurally different from the bottom edge, and the ribose residues are of an asymmetric nature (Wing et al., 1980).

It should be noted that nucleic acids, such as DNA, are conformationally variable molecules (Lodish et al., 2000). Double-helical DNA can form various structural features, besides B-DNA, depending on environmental factors like pH and the presence of cations, as well as the sequence of bases present. B-DNA can undergo a reversible conformational change to a wider

and flatter right-handed helix sometimes seen in the context of TF DNA-binding, known as A-DNA (Kulkarni and Mukherjee, 2017).

### 1.3.2. The structure and function of a transcription factor

TFs are modular in nature – meaning that they contain various domains within which core functions are embedded (Latchman, 1997). For a TF to regulate gene expression at the transcriptional level, it should be able to do 3 things: Firstly, it should be able to bind the DNA in a sequence specific manner. Secondly, it should be able to interact with other factors to enhance or repress transcription. Thirdly, its synthesis and activity should be regulated stringently (Latchman, 1990). TFs are the ideal candidates for this function, since each of the domains that make up the structure can carry out at least one of the functions mentioned above. The modular structure shown in ***Figure 2*** depicts a prototypical TF with a random number and order of domains. The DNA-binding domain (DBD), shown in red in ***Figure 2***, often consists of one or more basic α-helices. These helices can recognize a specific DNA sequence (response element) usually located in the major groove of the DNA. The transactivation domain (TAD), shown in yellow in ***Figure 2***, allows the TF to interact with the pre-initiation complex consisting of RNA polymerase II (RNA pol II) and a host of other protein factors such as co-activators (histone acetyltransferases (HATs) and mediator proteins) and co-repressors (histone deacetylases (HDACs)), thereby activating or repressing transcription. The signal-sensing domain (SSD) or ligand-binding domain, shown in green in ***Figure 2***, allows the TF to bind to various cofactors causing conformational changes in the TF, resulting in activation or repression. The SSD is responsible for modulating the synthesis and/or activity of the TF. Lastly, TFs may also contain a dedicated domain to facilitate dimerization. It should be noted that all these functions are sometimes contained within just one domain (Wärnmark et al., 2003).




***Figure 2. Simple representation of a prototypical transcription factor.*** *The DNA-binding domain (DBD) is red, the transactivation domain (TAD) is yellow, and the signal sensing domain (SSD) is green. The order and number of these domains vary amongst transcription factors, sometimes the activity of the TAD and SSD are contained in one domain. As can be seen by the blue spaces (connecting regions), the domains are separate from each other and represent distinct functional entities. Each of the domains are made up of one or more structural motifs.*

TFs recognize DNA in a sequence specific manner through macromolecular recognition. The TF presents a distinct three-dimensional shape or pattern on its surface. This pattern is complementary, both chemically and structurally, to the surface of the DNA (Garvie and Wolberger, 2001). Once this recognition occurs, the various intermolecular and interatomic interactions that govern DNA-binding can take place. The sequence specificity exhibited by TFs in macromolecular recognition arises from direct interactions between the side chains of the amino acids and the specific bases in the DNA. These interactions are mediated by hydrogen-bonding and the formation of salt bridges resulting from the highly electronegative oxygen atoms present in phosphodiester linkages. TFs can use a variety of structural motifs within DBDs to recognise and interact with DNA. For example, a TF may bind to a DNA sequence by inserting its α-helix into the major groove (Garvie and Wolberger, 2001). This is possible since the diameter of an α-helix is 1.2 nm, while the major groove of B-DNA measures 1.2 nm deep and 0.8 nm wide (Pabo and Sauer, 1984). Thus, one side of the α-helix can fit snugly into the major groove on the surface of the DNA. Whilst in this groove, the side chains adjacent to the groove form hydrogen-bonding partners with the edges of the base pairs that line the floor of the major groove. Furthermore, the pattern presented in the major groove is distinctive of the base pairs present and can thus be used as an identification matrix. This can easily be understood by inspecting the structure of B-DNA as well as the base pairs present. The positions of the hydrogen-bonding acceptors in the major groove vary with both their orientation and identity, whereas in the minor groove the positions of the hydrogen-bonding partners are sequence independent (Rohs et al., 2010). The minor groove is also too small for an α-helix to fit into, however in some cases, the minor groove can be widened through DNA-bending (Coll et al., 2002; Müller and Herrmann, 1997). Base sequence recognition isn't the only way that sequence specific recognition can occur. A TF can recognize DNA via a shape readout in which local conformational changes in B-DNA (resulting from the effect that the sequence has on local DNA structure) can be recognized by the TF, or vice versa. (Rohs et al., 2010).

Even though DNA-binding is a pre-requisite for the regulation of transcription by TFs, it is not in itself sufficient. TFs typically contain one or more TADs, as described in ***Figure 2***, which serve as sites for interaction with other proteins such as transcriptional co-regulators (Glass and Rosenfeld, 2000). Examples of these protein-protein interactions include dimerization

and domain-swapping through hydrogen-bonding, the formation of salt bridges and Van der Waals forces. Through these interactions, gene transcription is activated or repressed by mechanisms such as chromatin remodelling and the recruitment of other cofactors to the pre-initiation complex. These cofactors include mediator proteins, histone acetyltransferases (HATs) and histone deacetylases (HDACs) (Schaefer et al., 2011) . TADs are classified by the presence of certain amino acid sequences which facilitate their activity; or are simply the most abundant residue present in the sequence. By this classification three such TADs exist: The acidic domain, the glutamine-rich domain and the proline-rich domain (Mitchell and Tjian, 1989). The acidic domain is characterized by the presence of glutamate and aspartate residues which confer a negative charge, via deprotonation, on the TAD at physiological pH. This is due to the presence of a carboxyl group in their side chains, with a p$K_a$ of 3.8 for aspartic acid and 4.2 for glutamic acid. These domains also fit the description of an acidic α-helix (Hollenberg and Evans, 1988), which may mediate transcription through non-specific interactions with general TFs (GTFs) which are responsible for basal transcription, as well as DNA polymerase II (Buratowski et al., 1988). Glutamine-rich domains consist of approximately 25% glutamine residues and can facilitate the upregulation of genes by up to 200-fold, through interaction with the specificity protein 1 (SP1) TF which bridges the gap between various TFs and the transcriptional machinery (Xiao and Jeang, 1998). These domains bind to GC-rich response elements in a number of cellular promoters (Courey et al., 1989). The proline-rich TADs consist of approximately 25% proline residues and have been shown to activate numerous heterologous promoters. These domains interact with various factors making up the pre-initiation complex, and are also required for the specific interactions that mediate the initiation of transcription (Mermod et al., 1989). Other less abundant TADs may not fall into these categories. It should be noted that since the amino acid sequence doesn't necessarily dictate the functional pathway, TADs are sometimes also classified by whether they regulate the initiation or elongation processes of transcription (Frietze and Farnham, 2011).

Some TFs can induce changes in gene expression upon the binding of a ligand. These TFs contain an SSD which allows for the macromolecular recognition, and binding, of the ligand. This binding results in the activation of TFs through phosphorylation mediated by conformational changes, and often oligomerization (Burkhard et al., 1999). TFs of this nature

generate signals via various protein kinase pathways (Osborne et al., 2001). Strikingly, the binding of a ligand results in conformational changes within the DBDs and the SSDs, thereby modulating gene expression. For example, when steroids and other hydrophobic molecules bind to nuclear hormone receptors such as the oestrogen receptor located in the cell membrane, the ligand-receptor complex can modify gene expression by binding to various response elements in the DNA. This is achieved through structural rearrangements which do not alter the affinity or specificity for DNA, but instead promote the interaction with co-activators or co-repressors respectively (Berg et al., 2002).

### 1.3.3. DNA-binding domains

The defining feature of the TF family is the presence of the DBD, which allows for the sequence-specific recognition of the DNA as well as DNA-binding. Every DBD contains at least one structural motif used to bind DNA, which is one of several motifs making up the compact topology that characterizes the secondary structure of a DBD. There are various types of DNA-binding structural motifs that include the helix-turn-helix (HTH), zinc finger (ZF), leucine zipper (LZ) and immunoglobulin-like (Ig-like) domains. These have been depicted in ***Figure 3***.

#### *1.3.3.1. Zinc finger*

The zinc finger (ZF) domain, shown in ***Figure 3(A)***, is a structural motif characterized by the co-ordination of one or more divalent metal ions, like $Zn^{2+}$. Cysteine can coordinate these metal ions due to the electrostatic attraction between the highly electronegative thiolate groups and the positively charged metal ions. Histidine can coordinate these metal ions due to the electrostatic attraction between the highly electronegative deprotonated amine group present in the imidazole ring and the positively charged metal ions. The number and order of these residues is used to classify the ZFs as $Cys_2His_2$, $Cys_4$ and $Cys_6$. Each of these types represent remarkably different protein folds and as such can perform a wide variety of functions (Krishna et al., 2003). The region between the cysteine and/or histidine residues contains several conserved phenylalanine and leucine residues, which due to their bulky side chains, form a loop projecting out of the protein surface. This loop, or finger, is anchored at the base by the $Zn^{2+}$ co-ordination. The ZF region also often contains 2 anti-parallel β-strands packed against an α-helix. The α-helix contacts the DNA in the major groove (Luisi et al., 1991), as is also the case with HTH motifs, and interacts directly with several nucleotides.

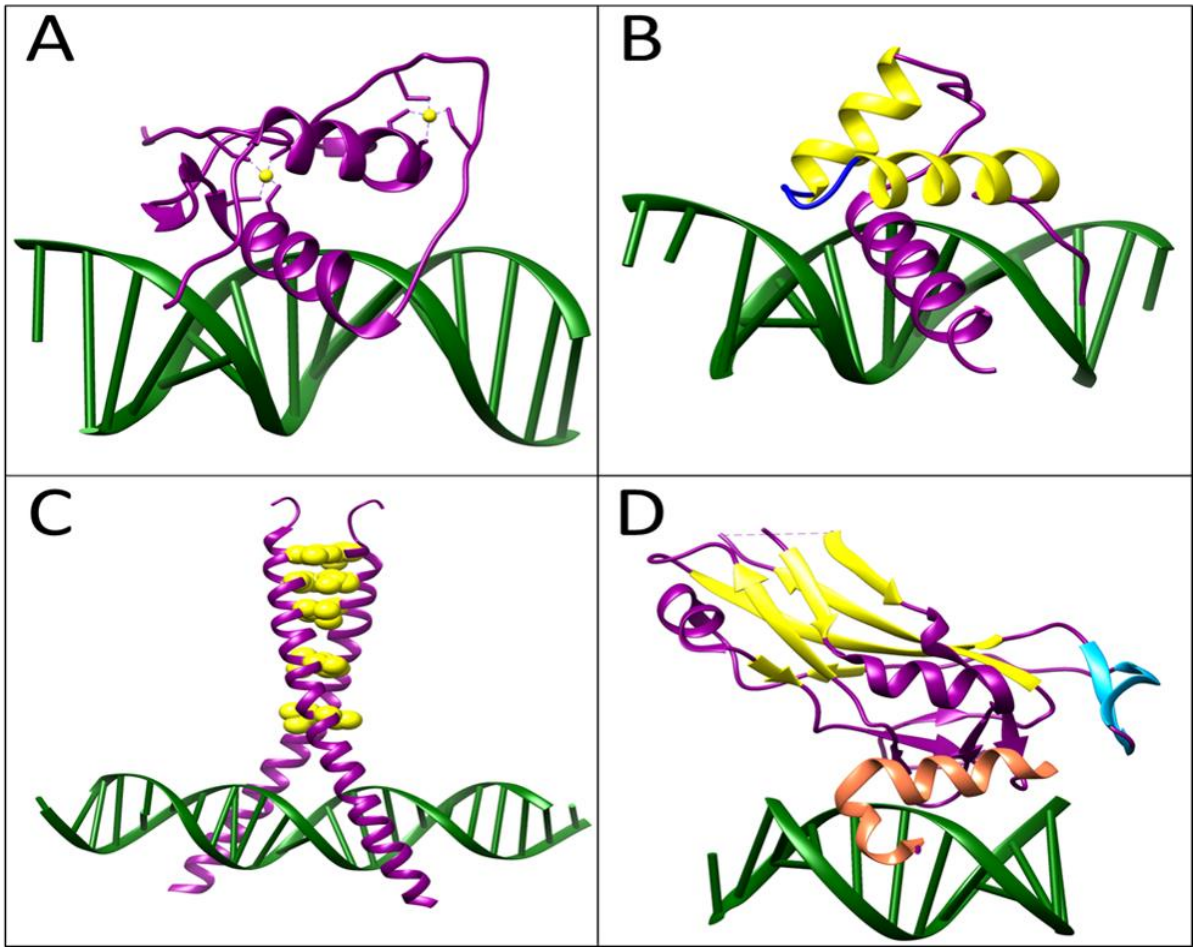



***Figure 3. Cartoon representation of the structural motifs used to bind DNA.*** *These models are some of the most common types used to bind DNA in eukaryotes. The proteins have been shown in purple, the DNA has been shown in green and important protein residues have been shown in yellow, blue, and orange. (A) The zinc finger domain from the glucocorticoid receptor (GR, PDB ID: 1R4R). In this model, the two* $Zn^{2+}$ *ions are coordinated by four cysteine residues, whose side chains have been shown as sticks. The metal chelation stabilizes the fold of the protein such that the purple α-helix at the bottom can be inserted into the major groove of the DNA. (B) The helix-turn-helix motif present in paired-box protein 6 (Pax-6, PDB ID: 6PAX). This structural motif is made up of two α-helices as shown in yellow, separated by a short turn shown in dark blue. The first of the two helices serve as a recognition helix while the second helix stabilizes the fold of the protein allowing the recognition helix to be locked into the major groove. (C) The leucine zipper motif found in the cAMP-responsive element binding protein (CREB, PDB ID: 1DH3). This motif is made up of two extended α-helices, shown in purple, in which approximately every seventh residue is a leucine. The bulky nature of the methyl groups in the side chains, shown by the yellow spheres, allow the helices to interdigitate such that the top of the zipper is in a closed state. The bottom of the zipper, which interacts with DNA, is clamped into the major grooves as a result. (D) The immunoglobulin-like domain from the t-box transcription factor 5 (TBX5, ODB ID: 2X6V). The core of the domain is made up of a seven-stranded beta barrel, as shown in yellow, closed off by two anti-parallel beta strands shown in blue. This structure stabilizes the fold of the protein and allows the mutually perpendicular alpha helices shown in orange to be inserted into the minor groove. The structures were rendered with UCSF Chimera v 1.16 (Pettersen et al., 2004).*

#### 1.3.3.2. Helix-turn-helix

The HTH motif consists of two α-helices, as shown in yellow in ***Figure 3(B)***. The helices are connected by an extended amino acid chain, as shown in blue. The first α-helix is termed the recognition helix as it recognizes specific base pairs in the DNA by the principles of macromolecular recognition mentioned above (Latchman, 2010). The residues found in the recognition helix vary amongst TFs, which may give rise to sequence specificity. The second helix of the motif stabilizes the recognition helix through hydrophobic interactions that keep the angle between the two α-helices constant thereby locking the recognition helix into place (Matthews et al., 1982). TFs possessing the HTH motif almost always dimerize, with the monomers being separated by approximately one turn of DNA (3.4 nm), allowing both recognition helices to fit into the major groove. The recognition helix can bind DNA by forming hydrogen-bonding partners with the distinctive recognition matrix presented by the base pairs presented in the major groove. Additionally, the shape readout mentioned above may be utilized to recognize contours in the DNA that have arisen from the base sequence present (Latchman, 1997; Schnepf et al., 2020).

#### 1.3.3.3. Leucine zipper

The LZ motif consists of two extended α-helices as shown in purple in ***Figure 3(C)***. In each of these α-helices, there is a 35 amino acid long stretch in which approximately every seventh residue is a leucine, resulting in the presence of a leucine every 2 turns on the same side of the α-helix as shown by the yellow spheres (Schumacher et al., 2000). Leucine residues are classified as non-polar due to the presence of methyl groups in their side chains. As a result of this hydrophobicity, the leucine side chains of each α-helix point toward each other and interact, thereby promoting the interdigitation of the helices and burying the hydrophobic residues. This facilitates dimerization through the formation of stable non-covalent linkages. Additionally, the bulky methyl groups in the side chains result in steric hinderance which fixes the top end of the zipper in a closed state. Unlike the HTH motif, the LZ motif doesn't bind the DNA. Instead, LZ formation mediates structural changes in the adjacent region of the protein found to be rich in basic amino acids which can then bind the acidic DNA (Latchman, 2010). This basic DNA-binding region of each α-helix tracks along the major groove of the DNA in opposite directions, resulting in a scissor-like clamp on the DNA.

#### *1.3.3.4. Immunoglobulin-like*

The Ig-like domain is a 125 amino acid long DNA-binding structural motif. It is made up of a β-barrel consisting of 7 - 9 anti-parallel β-strands, as shown in yellow in ***Figure 3(D)***, arranged in a Greek-key topology (Stirnimann et al., 2010a). Ig-like domains are classified according to their topological subtypes. The most common subtypes are the constant (C) and variable (V) domains, which are named according to their occurrences in the constant and variable regions of the protein, respectively. The C-subtype is further divided into the C1 and C2 (or s) classes. The C1 class contains immunoglobulins, T-cell receptors, and major histocompatibility proteins while the C2 (or s) class contains non-immunoglobulin-related molecules (Horstkorte and Fuss, 2012). The TBR1 T-box domain thus contains a C2 (or s) type immunoglobulin fold. The beta barrel is closed off at one end by 2 anti-parallel beta strands, shown in blue. The immunoglobulin domains are unusual when compared to the other DNA-binding domains, in that their core folds and features may vary because of the variability introduced by the flexible loops in the β-barrel. The large core of the protein stabilizes the fold and allows the mutually perpendicular alpha-helices, shown in orange, to interact with the DNA in the minor groove. This is unusual since all the other DBDs bind DNA in the major groove.

### 1.3.4. Mechanisms of transcription

The first step in the eukaryotic transcription of an exon is the recruitment and binding of RNA pol II to the promoter region (located directly upstream of the gene to be transcribed), together with the recruitment of the transcriptional machinery. This critical step is extensively regulated by the activity of the TFs which either stabilize or block the binding of RNA pol II to the promoter and is the first mechanism by which TFs regulate transcription (Gill, 2001). RNA pol II cannot independently initiate transcription due to its low DNA-binding affinity, and instead it requires the presence of many other factors to recruit it to the promoter region and stabilize its binding. This can best be seen in the activity of the general transcription factors (GTFs) responsible for constitutive transcription, which results in the formation of the PIC required for the binding of RNA pol II. Specifically, the TFIIA-TFIID complex first binds to the promoter, followed by the binding of TFIIB. Only once this is complete can the RNA pol II-TFIIF complex bind to the DNA, followed by the binding of TFIIE and TFHII (Orphanides et al., 1996). TFs that upregulate transcription are called activators. These activators increase the binding affinity of RNA pol II and the transcriptional machinery for the promoter. They do so through conformational changes brought about by protein-protein interactions, and by stabilizing a

setup that increases the binding of other GTFs on the DNA (Gill, 2001; Ma, 2011). This same process of constitutive transcription may be analysed to show how TFs can also downregulate transcription by directly or indirectly blocking the binding of RNA pol II. TFs that downregulate expression are called repressors and do so by binding to the promoter region thereby directly preventing the binding of activators, RNA pol II and the transcriptional machinery; or by recruiting other factors with the same function. Repressors can also downregulate expression by binding to regions adjacent to the gene to be regulated, resulting in local changes in DNA conformation, causing the promoter to become inaccessible to the transcriptional machinery (Latchman, 2020, 1997).

The DNA in the nucleus does not exist in a linear conformation. Instead, it is a supercoiled structure formed by wrapping the DNA tightly around positively charged proteins known as histones. The negatively charged DNA is electrostatically attracted to the positively charged histones, resulting in a highly condensed structure in which the DNA is inaccessible to RNA pol II and the TFs (Narlikar et al., 2002). As a result, the transcriptional machinery simply cannot bind the DNA or recruit RNA pol II, and as such cannot initiate or regulate transcription. However, if genes are to be transcribed, the DNA needs to be made transiently accessible to these various factors by pulling it away from the histones. Changes in chromatin structure therefore allow genes to be turned on and off, and this is the second mechanism by which TFs regulate expression. TFs often possess histone acetyltransferase (HAT) or histone deacetylase (HDAC) activity, which are opposing mechanisms used to relax or condense the DNA-histone complexes respectively (Ng and Littman, 2016). HDACs remove an acetyl group from acetylated lysine residues present in the histone tails resulting in an increased positive charge on the histone. This increases the attraction between the DNA and the histone, making the DNA inaccessible and downregulating transcription (Milazzo et al., 2020). HATs work in the opposite manner by acetylating lysine residues present in histone tails. This acetylation decreases the positive charge on the histone, thereby decreasing the DNA-histone attraction allowing for the DNA to be transiently pulled away from the histone. This makes the DNA accessible such that RNA pol II and the transcriptional machinery can bind. In addition to possessing HDAC or HAT activity, TFs can also recruit other factors possessing HDAC or HAT activity, thereby regulating expression in a similar manner. Sometimes, TFs may also alter chromatin structure in an ATP-dependant manner. This is done by actively sliding the histones

along the DNA in the wound state, exposing certain regions of the genome that are to be transcribed (Bowman, 2010).

The third mechanism by which TFs regulate gene expression is through the recruitment of co-activator or corepressor proteins to the TF-DNA complex. Co-activators and corepressors are regulatory proteins that interact with and activate TFs (activators or repressors) to regulate gene expression (Xu et al., 1999). To upregulate transcription, co-activators bind to activators and cause a conformational change in the DBD of the activator allowing it to bind to the DNA and subsequently recruit RNA pol II. More importantly, co-activators often have histone acetyltransferase activity, and by recruiting them to the response elements in the DNA, TFs upregulate transcription by providing access to the DNA in the manner mentioned above (Chen et al., 1997). Remarkably, activators and co-activators can bind the DNA several hundred base pairs apart. This means that when the activator interacts with the co-activator, two distant regions in the genome can be brought together, facilitating crosstalk between distal genetic elements (Matharu and Ahituv, 2015). Co-activators also function to bridge the gap between various components in the transcription machinery, functioning as molecular adaptors in signalling pathways (Edwards, 2000; Xu et al., 1999). To downregulate transcription, corepressors bind to repressors resulting in conformational changes in the DBD that allow the repressor to bind to the promoter region and prevent the binding of an activator which can recruit RNA pol II. Additionally, corepressors often have HDAC activity, and by recruiting them to response elements, TFs regulate transcription by rendering the genome inaccessible (Jenster, 1998). The delicate balance between activation and repression is maintained by the competitive binding of co-activators and co-repressors to the TF, as well as the fact that some protein regulatory elements can function as either a co-activator or a corepressor depending on the presence of other signals such as the binding of a ligand.

### 1.3.5. The regulation of transcription factors

Since TFs are responsible for the fundamentally important process of transcription and hence regulating the amount of gene products available to the cell, it makes sense for them to be highly regulated (Casamassimi and Ciccodicola, 2019; Latchman, 2011; Sperling, 2007). The first point at which TFs are regulated is at their synthesis. A vast range of facultative TFs are only synthesized in specific cell types, which allows for the cell-specific transcription of specific genes (Sonenberg and Hinnebusch, 2009). The transcription of a TF is extensively

regulated by other TFs. The resulting implication is that TFs can regulate themselves via a negative feedback loop (Hermsen et al., 2010), as seen in the cascade on the right of ***Figure 1***. In a negative feedback loop, the product of a gene (in this case a TF) can bind to that very same gene. By doing this, it blocks the binding of other TFs (activators), cofactors and the transcriptional machinery, turning the gene off and preventing further synthesis of that TF (Pan et al., 2010). The process of translation of a TF is also regulated through the levels of mRNA synthesized. For example, to increase the amount of TF that is synthesized, increased levels of regulatory mRNA block translation initiation codons found upstream of the correct translation start site, thereby forcing the increased use of the correct start codon (Kearse and Wilusz, 2017; Pan et al., 2010).

The second point at which TFs are highly regulated is their activity. Genes need to be regulated in a spatiotemporal manner to ensure that TFs are not permanently switched on. The 3 mechanisms by which the activity of a TF is regulated are ligand binding, phosphorylation and the availability of other TFs and cofactors (Berg et al., 2002; Bohmann, 1990; Glass and Rosenfeld, 2000). TFs are often synthesized in the cytoplasm and the binding of the correct ligand is required to localize the TF to the nucleus since it may not possess any nuclear localization sequence (Whiteside and Goodburn, 1993). For many TFs retained in the nucleus, ligand binding is required to activate the TF by inducing conformational changes in the DBD that favour DNA binding and the recruitment of other factors (Berg et al., 2002). Furthermore, ligand binding can cause TFs to disassociate from co-repressors which TFs are complexed with in their inactive state (Jenster, 1998). Signalling pathways transmit gene regulatory signals through the cytoplasm using protein kinases, which ultimately affect the phosphorylation state of the TF in order to activate or deactivate it (Hunter and Karin, 1992). When TFs interact with other proteins via the transactivation domain shown in ***Figure 2***, their structure is altered in a way that changes the affinity of the TF for DNA and other factors, resulting in the regulation of their activity.

### 1.4. The TBR1 T-box transcription factor

Cognition, perhaps the most salient feature of our species, is defined as the mental processes involved in acquiring knowledge and understanding. These include problem-solving, judgement and memory, which are higher-order brain functions that encompass language, perception, and imagination. The cerebrum, which is the largest part of the brain, is

responsible for cognition (Ackerman, 1992). The outer layer of the cerebral cortex is termed the cerebrum, while the inner layers make up the cerebral hemispheres. The cerebral cortex itself is divided into the neocortex and allocortex, which make up the outermost layer and inner layers of the cerebrum respectively (Arnould-Taylor, 1998). The neocortex is made up of 6 different layers, with each layer having distinct neuronal cell types and functions (Purves et al., 2001). These functions include motor and sensory control, as well as a means of integrating information in a way that gives us a perceptual experience of the world. It is thus clear that regions of the brain responsible for cognition contain a multitude of neuronal cell types. Since the entire central nervous system (CNS) is derived from the primitive embryonic ectoderm, these cells need a way in which to differentiate, which is where TF play a key role. They permit cell differentiation by regulating which genes are expressed, as well as when, where, and how often.

The TBR1 T-box TF is a neuron-specific TF that is involved in the differentiation and migration of neurons, as well as regulating cortical development specifically in layer VI of the developing 6-layered human neocortex (Deriziotis et al., 2014; Hoed et al., 2018). It is encoded for in humans by the *TBR1* gene located at chromosome 2q24.2 (Bulfone et al., 1995). The TBR1 T-box can function both as an activator and repressor of transcription (Han et al., 2011; Hsueh et al., 2000). The TBR1 TF is a member of the T-box family of TFs which share the evolutionarily conserved T-box DNA-binding domain. The protein belongs to the TBR1 subfamily of TFs containing TBR1, TBR2 and TBX21 (Mizuguchi et al., 2012; Wilson and Conlon, 2002a).

### 1.4.1. The T-box family

The T-box domain, shared by the T-box family of TFs, is both necessary and sufficient for sequence-specific DNA-binding (Kispert and Hermann, 1993). Since the T-box domain is the functional domain of the T-box family, only the T-box domain of TBR1 was analysed and investigated by the methods described below. T-box TFs display a remarkably high sequence similarity within the T-box domain, as depicted by residues of a strongly similar nature (shown as the same colour), in the multiple sequence alignment of the T-box TFs shown in ***Figure 4***. The differences in amino acid residues at certain positions have been used to sub-categorize the T-box family into five subfamilies as shown in ***Figure 5***, namely T, TBX1, TBX2, TBX6 and TBR1. The TBR1 T-box TF belongs to the TBR1 subfamily also consisting of TBR2 and TBX21. From amongst the T-box proteins, only the structures of TBX1 (PDB: 4A04), TBX3 (PDB: 1H6F),

TBX21 (PDB: 5T1J) and TBX5 (PDB: 2X6V) have been determined (Coll et al., 2002; El Omari et al., 2012; Liu et al., 2016; Stirnimann et al., 2010). Since there is a high sequence homology in the T-box domain between the TBR1 T-box domain and these known structures, they have been used to predict the structure of the TBR1 T-box as well as provide insight into the exact DNA-binding mechanism utilized.




***Figure 4. Multiple sequence alignment of the T-box domains of the T-box family.*** *The T-box domains of human TBXT, TBX1, TBX3, TBX 5, TBX 21 and TBR1 have been aligned using ClustalOmega and have been coloured according to residues of strongly similar properties. The UNIPROT accession numbers are provided on the figure itself. The alignment shows a high degree of homology in the T-box domain of T-box family proteins, and the differences observed account for the classification into subfamilies. There is a great deal of sequence variation from residues 99-120, indicating a variable region. Of the proteins shown only the structures of TBX1, TBX3, TBX5 and TBX21 have been determined.*

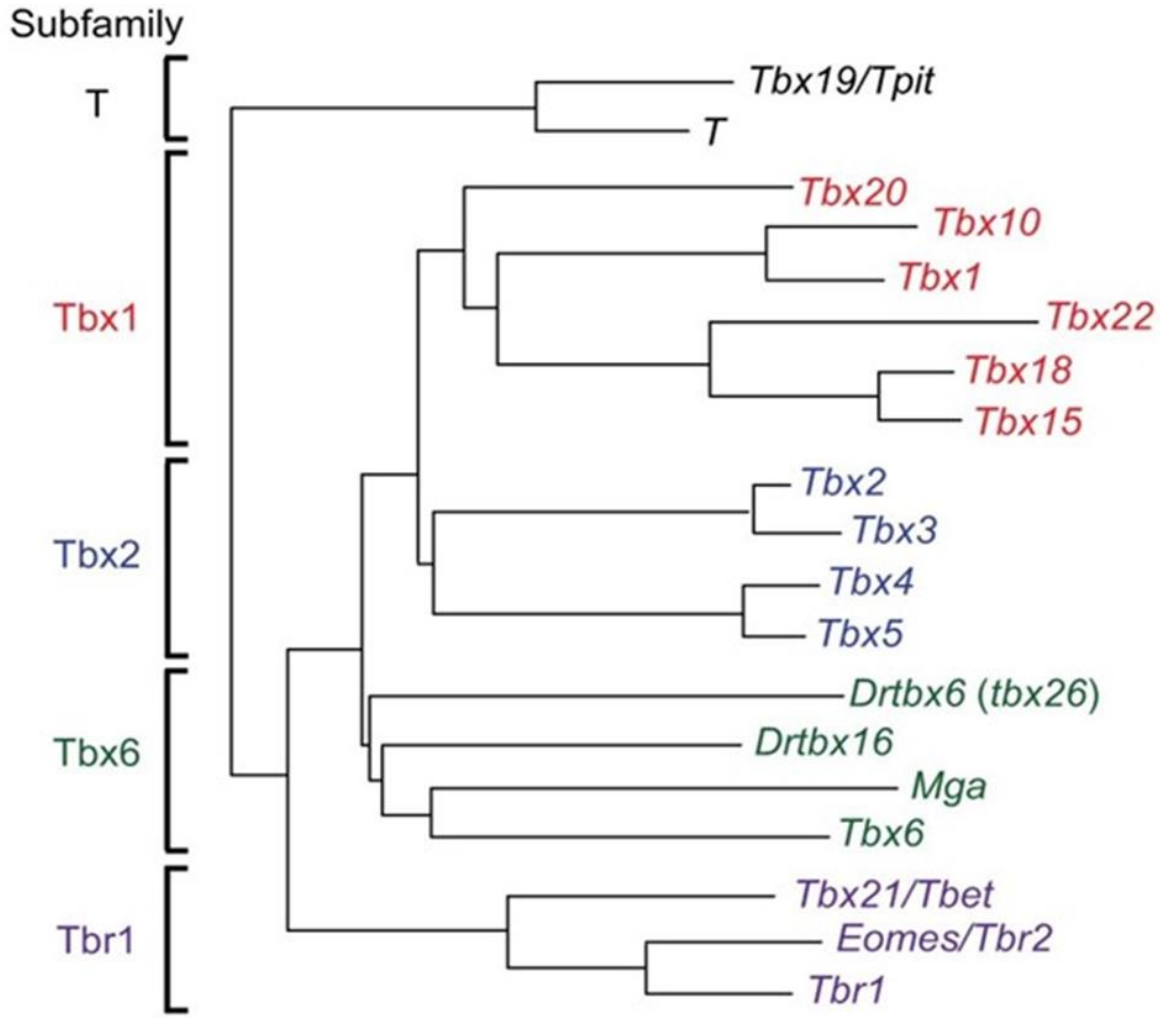



***Figure 5. Phylogenetic tree for the T-box family of transcription factors.*** *The T-box family of transcription factors share an evolutionarily conserved DNA-binding domain and are subclassified into 5 families based upon the variability present in the flexible loop region found in the T-box DNA-binding domain. The 5 T-box subfamilies are: T, TBX1, TBX2, TBX6 and TBR1. The TBR1 T-box transcription factor belongs to the TBR1 subfamily, which also consists of TBX21 and TBR2. The TBR1 subfamily is most closely related to the TBX6 subfamily. This figure was taken from Papaioannou et al., 2014.*

The average molecular mass of T-box TFs is approximately 60 kDa. The T-box DBD makes up a third of this mass making it a relatively large DBD. In addition to the DBD (Liu et al., 2016a), T-box proteins typically also contain a TAD such as shown in ***Figure 2***, used to interact with other TFs and coregulators. The order of the domains varies between members of the T-box family and sometimes the activity of the DBD and the TAD is contained within one domain. Additionally, the C' region following the T-box domain contains a nuclear localization signal since the T-box domain needs to bind DNA in the nucleus to regulate transcription. As shown in ***Figure 4***, there are varying degrees of homology, in the T-box domain, between individual T-box family members. Additionally, there is little to no sequence homology in residues 99 -

120, which could account for the specificity of the T-box domains for their target sites. However, the specificity observed does account for the DNA-binding affinity, suggesting and supporting the notion that the T-box domain could also mediate other functions such as protein-protein interactions (Wilson and Conlon, 2002). The residues of T-box proteins that precede and follow the DBD exhibit variations in primary sequence and length.

Despite the amino acid sequence variations present in the T-box domain among the various T-box family members shown in ***Figure 4***, all T-box proteins bind to the same 8 base pair - long DNA consensus sequence 5' – TCACACCT – 3' known as the T-box binding element (TBE). This has been demonstrated by examining downstream gene targets, as well as through binding-site selection studies (Wilson and Conlon, 2002). In vitro studies have shown that T-box family members preferentially bind DNA sequences containing two or more TBEs. For this reason, previous studies on T-box domains have often utilized palindromic T-box DNA (Coll et al., 2002, p. 21; El Omari et al., 2012; Liu et al., 2016; Müller and Herrmann, 1997). The relevance and significance of these findings should however be questioned since only half-sites containing one TBE have so far been identified *in vivo*. There is some degree of uncertainty surrounding the exact mechanism used by the T-box to bind DNA, such as whether it binds as a monomer or dimer, how many TBEs are required, and the residues involved. T-box proteins have been shown to bind a single TBE as in the case of TBX1, as well as palindromic sites containing two TBEs such as in the case of Xbra (the Xenopus homolog of the mammalian protein brachyury) (El Omari et al., 2012; Müller and Herrmann, 1997). Some T-box proteins, like TBX5, bind DNA as a monomer while others, like TBX21, bind DNA as a dimer (Liu et al., 2016; Stirnimann et al., 2010). These differences may be accounted for by the sequence variations present in the T-box DBD, as well as the variations present in the neighbouring regions. The T-box domain recognizes and binds to the TBE via the formation of an inducible recognition element, helix $3_{10}$C, which only becomes structured upon DNA-binding (Stirnimann et al., 2010a).

The T-box family of TFs has been shown to both activate and repress transcription (Papaioannou, 2014; Paxton et al., 2002). This transcriptional activity lies within the C' region of the protein and within the T-box DBD. In addition to transcriptional regulation, there is emerging evidence that T-box proteins can mediate gene regulation through epigenetic modifications such as the recruitment of methyltransferase activity (Papaioannou, 2014).

Given the important gene regulatory roles played by T-box TFs, it is not surprising that T-box genes are evolutionarily conserved. The T-box family play a host of critical roles, ranging from organogenesis to cellular differentiation and migration. This is supported by the fact that the T-box genes are tightly regulated and mutations in these genes result in dramatic embryonic phenotypes (Coll et al., 2002a). Mutant alleles reflect a phenotype even in the heterozygous state, and this haploinsufficiency suggests that the local concentrations of T-box proteins are critical for their function. Mutations in the T-box family are implicated in a number of developmental disorders such as Ulnar-mammary syndrome, DiGeorge syndrome and Holt-Oram syndrome (Coll et al., 2002; El Omari et al., 2012; Stirnimann et al., 2010). Some subfamilies of the T-box family are neuron specific, and a consequence of these conditions is that mutations in the T-box domain are often associated with neurodevelopmental disorders with cognitive impairments, such as autism spectrum disorders (ASDs), which often include problems in the development of speech and language (Deriziotis et al., 2014b; Hoed et al., 2018; O'Roak et al., 2012a, 2014a).

### 1.4.2. The T-box DNA-binding domain

The structure of the TBR1 T-box has not been determined. Since the T-box domain is evolutionarily conserved amongst T-box family members, any of the crystallographic structures available in the Protein Data Bank (PDB) can be used to study the structure and DNA-binding of the T-box domain to some extent. The crystal structure of TBX5 in the DNA-bound form shown in ***Figure 6*** is a good example. The TBX5 T-box DBD interacts with the promoter regions of various genes including the cardiac-specific *Atrial Natriuretic Factor (ANF)* gene. The gene contains an 8 base-pair long consensus DNA half site 3' - (A/G) GGTGT(G/C/T)(A/G) – 5' (Ghosh, 2001). This sequence is non-palindromic, contains a single TBE, and had been chosen because only half-sites have so far been identified in natural promoters of the T-box target genes (Bruneau et al., 2001).

The structure of the TBX5 T-box domain in ***Figure 6*** shows that the T-box domain is made up of a seven-stranded β-barrel core (A, B, C, C', E, F and G) and is defined as having a C2 (or s) type immunoglobulin fold (as described above) with a Greek key topology (Bork et al., 1994). The β-barrel is composed of two antiparallel β-sheets, the first consisting of strands A, B and E and the second consisting of strands C, C', F and G. The β-barrel is closed off towards the DNA by two smaller β-sheets, the first consisting of strands c, c', f and g and the second of

strands e, e’ and b. There are also α-helices α1 and α4 between the β-stands in the barrel, as well as the C-terminal helices α3 and $3_{10}$C. Helices α3 and $3_{10}$C are mutually perpendicular and $3_{10}$C is inserted into the minor groove of the DNA. DNA contact is also made in the major groove via the loop region between strands c and c’. The variable region, indicated by the dashed line between strands F and G in ***Figure 6***, shows little to no electron density suggesting that it is flexible and dynamic (Stirnimann et al., 2010a). The structures of the T-box domains from TBX1, TBX3 and TBX5, in the DNA-bound form, have been aligned and depicted in ***Figure 7***. They are fairly superimposable with a root mean square deviation (RMSD) of less than 1Å between the structures (El Omari et al., 2012).

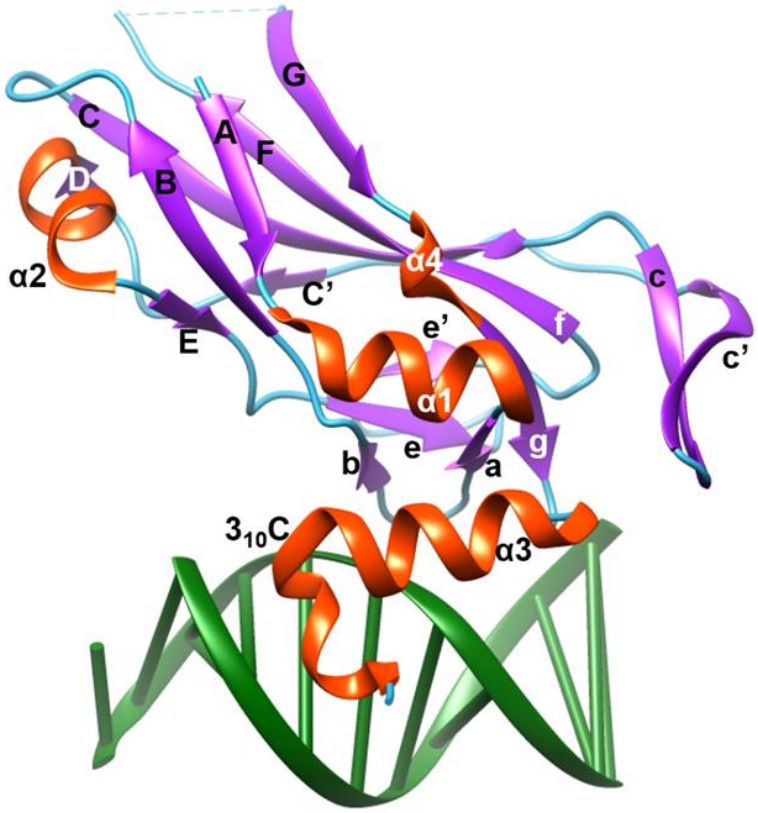



***Figure 6. An annotated crystal structure of TBX5 T-box domain in the DNA-bound form.*** *The core of the T-box domain is a seven-stranded β-barrel consisting of strands A, B, E, C, C’, F and G. The barrel is closed off by two β-sheets consisting of strands c, c’, f, g, e, e’ and b. There are two C’ α-helices α3 and $3_{10}$C which serve as DNA recognition helices, as well as helices α1, α2 and α4 between the strands of the β-barrel. The T-box domain interacts with the DNA in the minor groove via the C’ helices, and in the major grove via the loop between strands c and c’. The variable region of the T-box domain is shown by the dashed line between strand F and G. The TBX5 T-box domain has been coloured according to its secondary structure with helices in orange and strands in purple. The DNA has been shown in green. The PDB accession number for TBX5 in the DNA-bound state is 2X6V (Stirnimann et al., 2010). The structures were rendered with UCSF Chimera v 1.16 (Pettersen et al., 2004).*

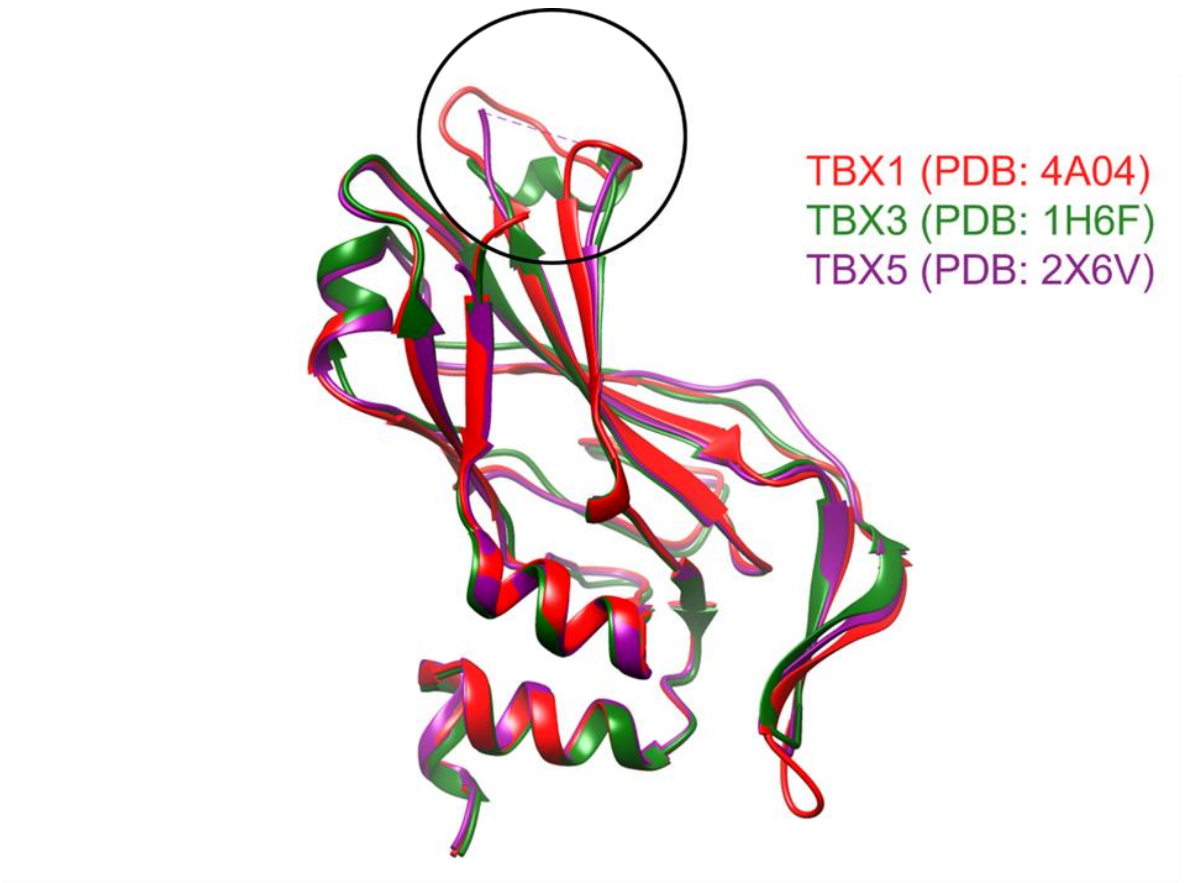



***Figure 7. Structural alignment of the T-box domains from TBX1, TBX3 and TBX5 in the DNA-bound state.*** *The T-box domains of TBX1 (red), TBX3 (green) and TBX5 (purple) have been superimposed with a root mean square deviation of less than 1Å. The major structural components of the T-box DBD are conserved. There are however some structural differences such as the variable loop region, circled in black, and the loop region used to contact the DNA in the major groove. The PDB accession numbers for TBX1, TBX3 and TBX5 are 4A04, 5T1J and 2X6V respectively (El Omari et al., 2012, Coll et al., 2002 and Stirnimann et al., 2010). The structures were aligned and rendered with UCSF Chimera v 1.16 (Pettersen et al., 2004).*

The immunoglobulin fold and Greek key topology is conserved amongst these structures as expected, as are all the other major structural features such as helices α3 and $3_{10}$C that contact the minor groove of the DNA, as well as the loops protruding down from the barrel towards the DNA. However, there are some notable structural differences between these T-box domains, which could account for the varying affinities that T-box TFs have for their cognate binding sites, as well as the array of protein-protein interactions displayed by the T-box family of TFs. The first notable difference is the exclusive presence of an α-helix in the variable loop region (shown in the black circle) of TBX3 (green) in ***Figure 7*** (Coll et al., 2002a). This α-helix is absent in the TBX1 and TBX5 structures which assume a flexible loop conformation instead (El Omari et al., 2012; Stirnimann et al., 2010a). The reason for these differences is the presence of four extra residues in the variable loop region of TBX3. This

could account for the dimeric nature of TBX3 and the monomeric nature of both TBX1 and TBX5 respectively (Coll et al., 2002; El Omari et al., 2012; Liu et al., 2016; Müller and Herrmann, 1997; Stirnimann et al., 2010). The presence of these extra residues increases the dimerization interface between two TBX3 monomers (Coll et al., 2002). The second difference is in the loop region, between strands c and c' that contacts the DNA via the major groove. In TBX1, this loop region is extended when compared to TBX3 and TBX5, as seen in ***Figure 7***. The significance of this conformational difference is unknown. It could be possible that the extended-loop conformation facilitates protein-protein interactions by stabilizing the DNA-bound state of TBX1.

To study the exact mechanism by which DNA binding takes place, the conformational changes that occur upon DNA-binding should be studied. The T-box domain from TBX5 will be discussed here as it is the only T-box structure available in both the apo and DNA-bound states (Stirnimann et al., 2010). The changes induced to facilitate and maintain DNA-binding have been depicted in ***Figure 8***. When the apo and DNA-bound structures are superimposed, there are only minor conformational changes detected with an RMSD of only 0.5Å. The major difference is the exclusive presence of helix $3_{10}$C in the DNA-bound state. When the protein is not bound to DNA, the residues in helix $3_{10}$C do not show any electron density, meaning that this region is either highly unstructured or unwinds in the absence of DNA. The mechanism used to bind DNA is thus the formation of an inducible recognition element $3_{10}$C which only becomes structured upon DNA-binding. The amino acid residues involved in interacting with the DNA can be categorised broadly into 3 groups according to the conformational changes observed upon binding. The first group involves residues which do not undergo any conformational changes, and this makes up most of the residues in the protein. These residues are in the loop region, between strands c and c' in ***Figure 6***, that contacts the DNA. The second group of residues undergoes minor conformational changes within the side chains and these residues are also found within the loop region. The third and most important group of residues includes those that undergo major conformational changes of their side chains, and these residues are unsurprisingly found within the $3_{10}$C helix. The $3_{10}$C helix positions the residues Phe232 and Phe236 so that their bulky benzene rings contact the DNA directly via guanine residues in the minor groove (Stirnimann et al., 2010).

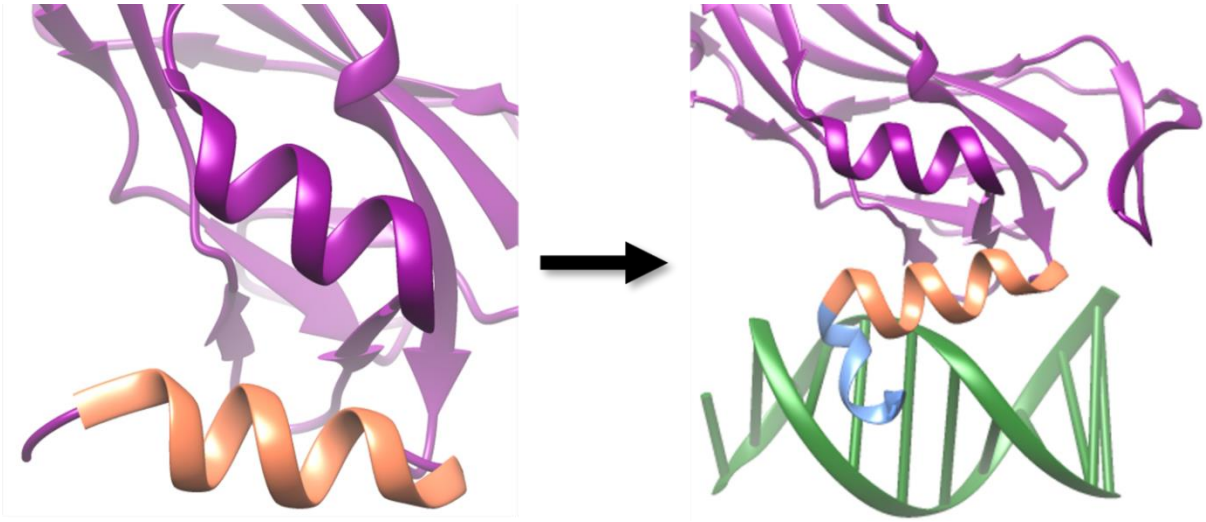


***Figure 8. Crystal structures of the DNA-free and DNA-bound TBX5 T-box domain.*** *The structure of the DNA-free TBX5 T-box domain (purple) on the left is very similar to the structure of the DNA-bound TBX5 T-box on the right, except for the C' α-helix $3_{10}$C (blue) which is only present in the DNA-bound state. Upon DNA-binding, the TBX5 T-box induces the formation of this $3_{10}$C helix which inserts itself into the minor groove of the DNA. As a result, the minor groove widens allowing for amino acid residues to come into direct contact with guanine residues in the DNA. The T-box DNA-binding mechanism is unlike any other, and even though the helices α3 (orange) and $3_{10}$C seem to assume a helix-loop-helix conformation, they bind DNA in the minor groove and are mutually perpendicular which is uncharacteristic of a helix-loop-helix motif. The T-box DNA (green) is natural, non-palindromic and comes from the atrial natriuretic factor promoter. The PDB accession numbers for the TBX5 T-box DNA-unbound and DNA-bound states are 2X6V and 2X6U respectively (Stirnimann et al., 2010). The structures were rendered with UCSF Chimera v 1.16 (Pettersen et al., 2004).*

The conformational changes that occur upon DNA binding allow for the bulky benzyl-side chains to be wedged into the DNA between the sugar moieties on opposite strands. As a result, the DNA undergoes a protein-induced conformational change in the minor groove. Accordingly, comparison of TBX5-bound DNA and β-DNA reveal that the minor groove is indeed widened by TBX5 binding (Stirnimann et al., 2010a). The TBX5-bound DNA is slightly unwound when compared to β-DNA, and this results in the bases being positioned closer to the $3_{10}$C helix. It is thus clear that the formation of the helix is essential for DNA-binding. However, presence of the $3_{10}$C helix is not sufficient for transcriptional regulation. WT TBX5 binds the specific TBE with weaker affinity than it does non-specific DNA, and when there is a mutation in the β-barrel, the binding of specific and non-specific DNA is similar (Stirnimann et al., 2010a). This suggests that while helix $3_{10}$C is essential for DNA-binding, the β-barrel is essential for the specific recognition of the TBE.

### 1.4.3. TBR1 T-box protein-protein interactions

As previously mentioned, for a TF to regulate gene transcription it should be able to interact with other TFs, coregulators and the transcriptional machinery; in addition to sequence-specific DNA-binding. The TBR1 TF is part of a larger gene regulatory network, such as the one shown in ***Figure 1***, which has been implicated in the aetiology of ASDs. It has recently emerged as a master regulator of ASD-related genes such as *RELN*, *GRIN2B* and *AUTS2* amongst others (Hoed et al., 2018). In this gene regulatory network, previous studies have shown that the TBR1 T-box interacts directly with FOXP2, BCL11A and CASK; all of which have been implicated in cortical development (Canovas et al., 2015; Deriziotis et al., 2014; Hsueh et al., 2000). Aberrant intermolecular interactions with these proteins could be the molecular basis for autism and it is therefore important to understand how these interactions occur.

#### *1.4.3.1. FOXP2*

Forkhead box protein 2 (FOXP2) is a TF that is responsible for the development of speech and language in humans (Graham and Fisher, 2013). It has been shown to interact with the TBR1 TF, although not much is known about the interaction besides the fact that there is one. Since FOXP2 has been implicated in speech and language and TBR1 has been implicated in ASDs, an aberrant molecular interaction between FOXP2 and TBR1 could explain why speech and language deficits are a core feature of neurodevelopmental disorders (Hoed et al., 2018). The importance of the FOXP2-TBR1 interaction in the normal development of speech and language may stem from their co-regulatory roles in *AUTS2* gene expression. The *AUTS2* gene is one of six genes implicated in the aetiology of ASDs and is a known TBR1 target gene (Bedogni et al., 2010). Additionally, the induction of FOXP2 expression has resulted in the upregulation of the *AUTS2* gene. The mechanism by which FOXP2 interacts with TBR1 is poorly understood, however it is thought that the T-box DBD is essential in this interaction since mutations within the T-box completely abolish the interaction with FOXP2 (Deriziotis et al., 2014). It has also been hypothesized that the DNA-binding forkhead domain is critical for the interaction, however it is not clear whether mutations in the forkhead domain affect the interaction or co-localization of the proteins since they are both neuron-specific in this context (Deriziotis et al., 2014).

#### *1.4.3.2. CASK*

The calcium/calmodulin dependant serine protein kinase (CASK) is a membrane associated guanylate kinase, which catalyses the reaction between adenosine triphosphate (ATP) and

guanosine diphosphate (GDP) to yield adenosine diphosphate (ADP) and guanosine triphosphate (GTP). Like the TBR1 T-box, CASK is expressed in the developing neocortex where it plays a role in neural development and synaptic function, and is a part of a larger gene regulatory network that utilizes a cascade of phosphorylation events to relay information to TFs (Hsueh et al., 2000). Interestingly, heterozygous mutations of CASK have been reported in ASDs (Sanders et al., 2012) and this finding has been used to rationalize the TBR1-CASK interaction. Interaction between the T-box DBD of TBR1 and the guanylate kinase domain of CASK causes the localization of CASK from the plasma membrane to the nucleus, where it co-operatively regulates the TBR1 target genes *RELN* and *GRIN2B* which have also been implicated in ASDs (Hsueh et al., 2000). The fact that DNA-binding by the TBR1 T-box DBD is required for CASK interaction suggests that DNA-binding causes conformational changes in TBR1 which mediate the protein-protein interaction, as previously suggested. The T-box DBD is responsible for interacting with the guanylate kinase domain of CASK, since truncating mutations in the T-box completely abolish the TBR1-CASK interaction. It has also been shown that several clinically relevant mutations in the T-box domain result in the formation of nuclear aggregates, causing aberrant nuclear localization of the TBR1-CASK complex (Deriziotis et al., 2014).

#### *1.4.3.3. BCL11A*

The B-cell lymphoma/leukaemia 11A (BCL11A) protein is a ZF-like TF that has diverse roles ranging from chromatin remodelling to brain development (Kadoch et al., 2013; Kuo et al., 2010). Since BCL11A is expressed in the deep layers of the developing neocortex and has been shown to interact with CASK, it is no surprise that it interacts with TBR1 as well (Kadoch et al., 2013). As previously mentioned, the developing neocortex contains six layers, and the differentiation of neurons within these layers relies on combinatorial control of a gene regulatory network. In layer V the BCL11A TF represses the *TBR1* gene while in layer VI low levels of BCL11A are required for TBR1 expression, revealing a reciprocal expression pattern of BCL11A and TBR1 in the developing neocortex (Canovas et al., 2015). It was first hypothesized that the N-terminal region of BCL11A mediated the TBR1 interaction, however mutants lacking the N-terminal domain still retained the ability to colocalize and interact with TBR1 in the nucleus suggesting the interaction might be due to some other consequence. Subsequently a C-terminally truncated BCL11A mutant was generated and showed complete loss of interaction with TBR1, suggesting that the C-terminal domain is responsible for TBR1

interaction (Hoed et al., 2018). Together these findings suggest that TBR1 expression in layer V is downregulated by direct association of TBR1 with the C-terminal domain of BCL11A. This sequestration prevents the TBR1 T-box DBD from upregulating its own expression by preventing it from binding to the TBE.

### 1.4.4. Clinical significance

As of 2016 it has been estimated that 1 in every 54 children is affected with ASDs, with cases occurring across all racial, socioeconomic, and ethnic groups (Centre for Disease Control, 2016). Autistic individuals often have communication deficits resulting in impaired social interactions and repetitive tendencies. The spectral nature of this disorder and the lack of information about the molecular mechanisms underlying the disease, has made it difficult to diagnose affected individuals and to characterize their symptoms. Even though inherited mutations account for approximately 40% of the risk of developing ASDs, rare genetic variants play a major role in the aetiology of ASDs as the effect of individual common traits is negligible. Accordingly, de novo loss-of-function mutations in any of the ASD-related genes (*CHD8*, *DYRK1A*, *GRIN2B*, *PTEN*, *TBR1, TBL1XR1, RELN* and *AUTS2*) are enough to cause ASDs (O'Roak et al., 2012). Strikingly three of these genes (*RELN*, *GRIN2B* and *AUTS2*) are regulated by TBR1 which is a putative master regulator of a gene regulatory network important in neocortical development that is recurrently mutated in cases of ASDs (Deriziotis et al., 2014). Insufficiencies in the T-box family of TFs have already been implicated in human disease as is the case of TBX3 (Ulnar-mammary syndrome) and TBX5 (Holt-Oram syndrome) (El Omari et al., 2012; Stirnimann et al., 2010).

The clinical significance of the TBR1 T-box TF is evident from mice studies, which show that homozygous mutants lack cortical lamination while heterozygous mutations result in autistic-like behaviours. These mutations include de novo missense variants, frameshift mutations and even whole gene deletions (Hoed et al., 2018). The detection of these mutants has pointed to the haploinsufficiency of TBR1 as the main pathogenic mechanism for ASDs. Additionally, heterozygous mutants could exert a dominant-negative effect on TBR1 by interfering with normal protein function, although lack of evidence for TBR1 homodimerization opposes this hypothesis. Mutations in TBR1 could result in aberrant molecular interaction with the proteins FOXP2, CASK and BCL11A as well as with the TBE and other regulatory elements. This notion is supported by the fact that neurodevelopmental

disorders often include impairments in speech and language (explained by the FOXP2 and BCL11A interactions), as well as intellectual disabilities (explained by the CASK interaction).

## 2. Rationale

The development of the human brain consists of a series of dynamic and adaptive processes that are governed by a constantly changing, but highly organized, gene regulatory network. Cellular differentiation is highly dependent on gene regulatory networks since differentiation is the direct result of facultative gene expression. This process relies on TFs which regulate genes in a spatiotemporal manner. Neuron-specific TFs hence play a key role in neural development. The TBR1 TF is responsible for the differentiation of subsets of neurons within the developing neocortex. It has diverse developmental roles and can activate or repress transcription. The *TBR1* gene is one of six genes that have been implicated in ASDs, and a heterozygous mutation in any of these genes could be enough to have clinical implications. Moreover, the TBR1 TF has recently emerged as a master regulator of the ASD-related genes, suggesting that mutations in the TBR1 T-box domain could underlie the altered neuromolecular networks observed in ASDs (Hoed et al., 2018). The primary molecular function of the TBR1 T-box domain is to bind the TBE in a sequence specific manner. As with other members of the family, this binding could induce conformational changes via the flexible loop region, which may mediate the protein-DNA and protein-protein interactions mentioned above. These interactive processes are what allows the TBR1 T-box TF to regulate gene transcription and are of fundamental importance to its function (Deriziotis et al., 2014). The key to understanding the function of a macromolecule, such as the TBR1 T-box domain, is to determine and understand its structure at all levels.

The DNA-binding mechanism of the TBX5 T-box domain has been described (Stirnimann et al., 2010a). Although the T-box domain is shared between TBR1 and TBX5, it is still necessary to determine the exact DNA-binding mechanism of the TBR1 T-box for several reasons. Firstly, TBX5 is a cardiac-specific TF while TBR1 is neuron specific (Huang and Hsueh, 2015; Stirnimann et al., 2010a). An implication of this is that TBR1 might have different expression patterns and will interact with a different set of co-factors which could alter its DNA-binding mechanism. Secondly, the variations present between the flexible loop regions of TBR1 and TBX5 could result in different DNA-binding mechanisms. Thirdly, the variations present in the C' residues 166-181 might be of significance since this region is utilised by TBX5 to bind DNA through the formation of helix $3_{10}$C (Stirnimann et al., 2010). Lastly, a detailed understanding of the TBR1 T-box DNA-binding mechanism will shed more light on how TBR1 regulates ASD-related

genes, which will enhance our understanding of the molecular mechanisms that govern neurodevelopmental disorders. In this study, the hypothesis is that the TBR1 T-box domain will bind DNA via an inducible recognition element that only becomes structured upon DNA-binding.

The structure of a molecule defines its function and properties. This is true for all matter including macromolecules such as proteins. The primary method used to determine the structure of a molecule at atomic resolution (crystal structure) is X-ray crystallography (XRC). The atomic-resolution data provided by XRC is unrivalled since it provides information about the protein at all levels (primary, secondary, tertiary and quaternary). It yields data regarding several parameters including the lengths and types of chemical bonds formed, the disorder of atoms within a molecule and the residues involved in macromolecular interactions (Rupp, 2013). The models provided can also be used in computational studies such as molecular docking and molecular dynamics simulations. The technique of macromolecular crystallography is hence a fundamental tool in molecular biology as it provides structural information from which functional information can be derived and so provides a more focused approach for future studies (Smyth and Martin, 2000). The information obtained from atomic level structural data can be used to determine the mechanisms underlying DNA-protein and protein-protein interactions. Considering how useful this technique can be in elucidating the DNA-binding mechanism of a TF, we attempted to crystallize the protein in the presence and absence of the TBE with the aim of elucidating the conformational changes that take place upon DNA binding, which is the first step in understanding how pathogenic mutations in the T-box domain can affect DNA binding. In the absence of crystallographic structures, protein modelling and molecular docking may be used to obtain a predicted DNA-binding mechanism.

In the broader context, this will enhance our understanding of DBDs as well how TFs interact with an array of macromolecules to regulate gene transcription. Finally, by understanding how TBR1 carries out its functions, we may begin to understand what goes wrong, at the molecular level, in neurodevelopmental disorders such as ASDs. This is the first step in properly defining ASDs, and subsequently developing diagnostic tools and treatments thereof.

## 3. Aim and objectives

The aim of this study was to determine the structure of the TBR1 T-box domain in the presence and absence of the TBE, with the hope of elucidating the DNA-binding mechanism of the protein.

To this end, the specific objectives of this project were to:

1. Transform T7 Express Competent *E. coli* cells with a pET-11a vector containing the gene (coding sequence) for the human TBR1 T-box domain (residues 203-396).
2. Overexpress and purify the TBR1 T-box domain using immobilized metal-ion affinity chromatography and size exclusion chromatography.
3. Characterize the structure of the TBR1 T-box domain, in the presence and absence of DNA, through far UV circular dichroism and intrinsic tryptophan fluorescence.
4. Study the interaction between the TBR1 T-box domain and DNA using electrophoretic mobility shift assays and fluorescence anisotropy.
5. Determine the DNA-binding mechanism of the TBR1 T-box domain through X-ray crystallography, protein modelling, disorder predictions and molecular docking.

# 4. Materials and methods

Since the T-box domain is evolutionarily conserved amongst T-box family members, one would expect the TBR1 T-box domain to bind DNA by similar mechanisms used by TBX5. This however remains uncertain since there aren't any other T-box crystal structures available, both in the presence and absence of the TBE, to validate this hypothesis. There are several sequence variations present in the variable loop region as well as the C-terminal α-helices of the T-box family of proteins which could imply variation in the binding mechanism, making it necessary to determine the specific DNA binding mechanism of TBR1 T-box domain. Additionally, the TBR1 T-box domain is neuron specific which could have specific implications for DNA-binding and protein-protein interactions undertaken by this protein compared to TBX5 whose T-box domain is cardiac-specific. The structural characterizations, DNA-binding studies, crystal trials and *in silico* studies have been used collectively in this study to try and answer the question of how the TBR1 T-box domain binds DNA.

T7 Express Competent *E. coli* cells were transformed with a pET-11a vector expression system containing the gene for the TBR1 T-box domain. The protein was overexpressed and subsequently purified using liquid chromatography. The structure was thoroughly characterized using far UV circular dichroism and intrinsic tryptophan fluorescence. These studies were done in the presence and absence of the DNA to confirm that the protein had been properly folded and that it was functional, as well as to assess the structural changes that took place upon DNA binding. The interaction between the TBR1 T-box domain and the DNA was studied by electrophoretic mobility shift assays and fluorescence anisotropy. Once the protein had been fully characterized, both structurally and functionally, crystal trials were performed to determine the conditions required for crystallization of the protein both in the presence and absence of the DNA. The conditions were then optimized using several techniques. Since the crystals did not yield any significant diffraction pattern, despite our best efforts to do so, *in silico* studies were used in an attempt to understand the DNA-binding mechanism through alternate means of studying the structure.

The TBR1 T-box domain has two surface-exposed cysteine residues and can therefore dimerize in solution via the formation of disulphide bridges. However, under physiologically relevant conditions such as the reducing environment of the nucleus, the T-box domain is monomeric in solution (Blaine et al., 2021). As such, all experiments were performed in the

presence of 2 mM DL-Dithiothreitol (DTT), to prevent the formation of physiologically irrelevant disulphide-linked dimers.

## 4.1. Protein preparation

This study was done exclusively on the T-box domain of the TBR1 TF since it is the functional DNA-binding domain of the protein, and the aim was to determine the DNA-binding mechanism. A pure and soluble protein sample was needed to perform the experiments in this study.

### 4.1.1. Plasmid and protein construct

The codon-optimized gene sequence for the TBR1 T-box domain (residues 203 – 396 of the full-length protein) was constructed and inserted into a pET-11a plasmid (GenScript, USA). The codons for the T-box domain were optimised for use in *E. coli* in order to increase the expression level of heterologous proteins by ensuring translational efficiency of the target gene through synonymous codon usage (Mauro and Chappell, 2014). The TBR1 T-box gene sequence, shown in ***Figure 9***, contains an N-terminal streptavidin-tag (Strep-tag II) (dark green), a hexahistidine-tag (His-tag) (orange), a thrombin cleavage site (dark blue) and the TBR1 T-box gene. The His-tag was included to facilitate downstream purification with immobilized metal-ion affinity chromatography (IMAC). The Strep-tag II was added to facilitate streptavidin-affinity chromatography in case the protein was not sufficiently pure after IMAC, however, it was not required.

Plasmids are circular, dsDNA molecules that replicate independently from the cell's chromosomal DNA and are naturally found in bacteria and yeasts. They undergo DNA replication prior to each cell division, much like chromosomal DNA, ensuring a continuous propagation through successive host-cell generations resulting in the production of many copies of the plasmid (Rosano and Ceccarelli, 2014). Plasmids are also relatively cheap, stable and easy to work with making them ideal for producing large amounts of protein in a short time (Rosano and Ceccarelli, 2014). Due to their circular nature, plasmids may be limited by insert size and are inappropriate for larger sequences. The pET-11a plasmid used was 5.7 kbp long, while the gene that was inserted was only 684 bp long, which made it size-appropriate for this study. The plasmid, shown in ***Figure 9***, was used to transform T7 Express pLysS Competent *E. coli* cells (New England Biolabs, USA). This vector expression system has routinely provided the highest levels of heterologous protein expression when using the T7lac

promoter in bacteria, while providing the lowest levels of basal expression (Rosano and Ceccarelli, 2014). The T7lac promoter cassette contains a T7 promoter (red) and T7 terminator (red), the lac Operator (purple) and a ribosomal binding site. The T7 promoter drives the high-level expression of the genes downstream of it, provided a T7 RNA polymerase is present (Rong et al., 1998). As such, the gene for the TBR1 T-box domain was inserted between the *Nde*I and *BamH*I restriction sites. The *AmpR* gene has been used for bacterial selection following transformation. The pET-11a plasmid also contains a pBR322 origin of replication which results in a high plasmid copy number. The *AmpR* gene confers ampicillin resistance to the bacteria by coding for the β-lactamase enzyme required to degrade the β-lactam ring in ampicillin, allowing for bacterial selection after transformation (Neu, 1969).

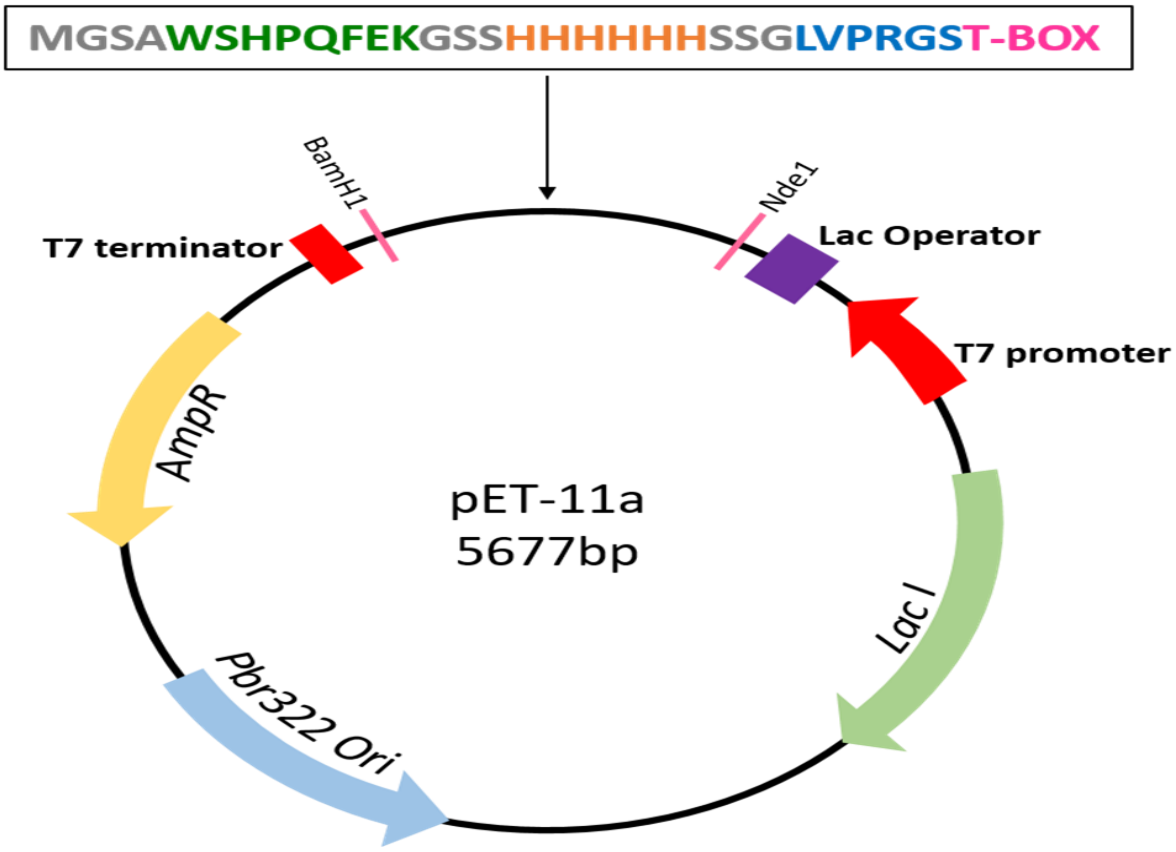



***Figure 9. Schematic representation of the pET-11a plasmid with the TBR1 T-box insert.*** *The plasmid was used to transform T7 Express pLysS Competent E. coli cells. It is 5 677 bp long and contains a T7 promoter (red), a T7 terminator (red), an AmpR gene (yellow), the LacI gene (light green) and the pBR322 origin of replication (light blue). The T7 cassette allows for overexpression to be controlled with isopropyl ß-d-1-thiogalactopyranoside (IPTG). The LacI gene prevents leaky basal expression while the AmpR gene allows for the selection of positive transformants. The gene for the TBR1 T-box domain (residues 203 – 396 of the full-length protein) was inserted between the NdeI and BamHI restriction sites. The N-terminus is fused to a streptavidin affinity tag (The Strep-tag II) (dark green), a hexahistidine tag (His-tag) (orange) and a thrombin cleavage site (dark blue).*

Even though plasmids can replicate independently within a host-cell, they still require the host-cell replication machinery such as the polymerases required to initiate plasmid replication (Rosano and Ceccarelli, 2014). Bacteria are often used as expression vectors for plasmids as they replicate quickly resulting in the production of many plasmids, and hence increased protein production. Furthermore, the passage of a plasmid into the bacterial cell is easily achieved since bacteria do not contain nuclei, and there is therefore no need for any complex subcellular localization (Theriot, 2013). The T7 RNA polymerase required to initiate transcription from the T7 promoter is not present in the plasmid and instead comes from the bacterial host-cell. It is coded for in the host-cell by the *T7 RNA polymerase* gene under the control of the LacUV5 promoter, which is inducible with a non-hydrolysable allolactose analogue isopropyl β-d-1-thiogalactopyranoside (IPTG) (Jeong et al., 2009; Rosano and Ceccarelli, 2014). The TBR1 T-box was thus only expressed in the presence of IPTG, providing a means by which gene expression could be switched on or off. The T7 Express Competent *E. coli* cells have been chosen because they possess the λDE3 lysogen, containing the gene for the T7 RNA polymerase under the control of a LacUV5 promoter. This strain is also protease deficient which made it ideal for the overexpression of the T-box domain.

The presence of the *LacI* gene (light green) is particularly interesting as it encodes for the lac repressor protein in the plasmid, which can bind to and block the expression of the *lacO* gene, preventing leaky basal expression of the TBR1-T-box TF in the absence of IPTG. Furthermore, the lac repressor protein binds to and represses the lacUV5 promoter in the host-cell thereby preventing the transcription of the T7 RNA polymerase in the absence of IPTG. The addition of IPTG blocks the inhibitory action of the lac repressor, which induces the expression of the T7 RNA polymerase in the host-cell, as well as preventing lac repressor inhibition in the gene of interest within the plasmid (Studier et al., 1990).

### 4.1.2. Transformation

The pET-11a plasmid, containing the TBR1 T-box domain insert, was used to transform T7 Express Competent *E. coli* cells using the heat shock method. An aliquot of competent *E. coli* cells was thawed on ice for 15 minutes. To initiate transformation, 2 µL of the plasmid (≈50 ng/µL) was incubated with 50 µL of the cells on ice for 30 minutes. This was done to stabilize the lipid membranes in the cells. The cells were heat-shocked at 42 °C for 45 seconds, after which they were immediately incubated on ice for 5 minutes and subsequently pelleted. The

increased temperature increases the kinetic energy of the molecules in the fluid membrane of the cell, thereby destabilizing it and making it more permeable to the plasmid. By lowering the temperature after the heat-shock, the cells were allowed to recover. The cell pellet was resuspended in 1 mL of freshly prepared sterile Super Optimal broth with Catabolite repression (SOC) media (2% (w/v) tryptone, 0.5% (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$ and 20 mM glucose), pre-warmed to 37 °C. SOC media is used to assist in cell recovery following the stresses induced by heat-shock. The transformed cells were then grown for 1 hour at 37 °C, while shaking at 250 rpm to allow for aeration. Following growth, the cells were spread onto freshly prepared sterile LB-agar (1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl and 1.5% (w/v) agar) plates, supplemented with 0.1 mg/mL of ampicillin (Melford, UK) used to select for successful transformants. Spread plates of empty SOC medium, LB-agar and untransformed *E. coli* cells were performed as controls to ensure that nothing was contaminated. The plated cells were grown for 16 hours at 37 °C. An isolated colony was then selected and used to inoculate 100 mL of fresh sterile 2 x yeast extract-tryptone (YT) media (1.6% tryptone (w/v), 1% (w/v) yeast extract and 0.5% (w/v) NaCl) containing 0.1 mg/mL ampicillin. The ampicillin was added to prevent the growth of any untransformed cells as well as to prevent any contamination. A small, isolated colony was selected to ensure that all the cells were genetically identical. The culture was grown for 16 hours at 37 °C, while shaking at 230 rpm to allow for aeration.

Small volume (1 mL) glycerol stocks were prepared by mixing this culture with 80% glycerol (v/v) in a 1:1 ratio. The glycerol stocks were flash-frozen in liquid nitrogen and stored at -80 °C until required. The culture was also used to isolate plasmids using the GeneJet plasmid mini prep kit (Thermo Fisher Scientific, USA), as per the suggested protocol. These plasmids (≈100 ng/µL) were sent for sequencing (Inqaba Biotechnical Industries, RSA) to ensure the integrity of the TBR1 T-box gene.

### 4.1.3. Heterologous protein expression

The optimal conditions used to overexpress the TBR1 T-box domain had already been determined in our laboratory and as such, there was no need to do any overexpression trials (Blaine et al., 2021).

Following bacterial transformation with the pET-11a plasmid, the TBR1 T-box domain was produced by heterologous protein overexpression in T7 Express pLysS Competent *E. coli* cells.

This was done to have enough protein for downstream applications in this project. The small volume glycerol stocks (1 mL) were used in a 1 in 1000 dilution to inoculate freshly prepared, sterile 2 x YT media containing 0.1 mg/mL ampicillin. The antibiotic was added to prevent the growth of unsuccessful transformants as well as contaminants. The culture was grown overnight at 37 °C, while shaking at 230 rpm to allow for aeration and growth. Larger volumes of freshly prepared sterile 2 x YT media, containing 0.1 mg/mL ampicillin, were inoculated with a 1 in 50 dilution of the overnight culture. The cells were grown until an $OD_{600}$ of 0.6 AU was obtained, since this indicated that the *E. coli* cells were in the exponential phase of growth, which is optimal for the induction of heterologous expression. The culture was subsequently cold-shocked on ice for 30 minutes, after which overexpression was induced with 0.2 mM IPTG (Melford, UK), which effectively turned the lacUV5 promoter on. Cold-shocking the cells prior to IPTG induction causes the rate of protein synthesis to be lower, resulting in soluble and correctly- folded protein. This is because a decrease in temperature slows down the cells' metabolism and growth rate, allowing the protein sufficient time to fold. The cells were then incubated at 20 °C for 24 hours, while shaking at 230 rpm. Reducing the temperature to 20 °C decreases the rate of protein-folding, thereby reducing the possibility of obtaining misfolded and non-functional protein. Following incubation, the cells were pelleted by centrifugation (5000 xg for 30 minutes, at 4 °C) and resuspended in 50 mL of equilibration buffer (500 mM NaCl, 20 mM imidazole and 20 mM tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl) pH 7.5) per litre of culture. The resuspended cell pellet was stored at -20 °C for no more than one month.

### 4.1.4. Purification

The protocol used to purify the TBR1 T-box domain had already been optimized in our laboratory and as such, there was no need for any further optimization (Blaine et al., 2021).

To isolate the protein from the insoluble fraction, complete cell lysis was carried out. The frozen cell lysates were left to thaw completely at 20 °C, after which 1 mM PMSF (Roche, Germany) and 0.1 mg/mL lysozyme (Sigma-Aldrich, USA) were added. PMSF is a serine protease inhibitor which prevents proteolytic degradation of the overexpressed protein by contaminating proteases. Lysozyme cleaves the peptidoglycan component present in gram-positive cell walls, causing it to become ruptured (Kirby, 2001). The mixture was incubated for 30 minutes at 20 °C, while inverting slowly. The cell lysate was sonicated 10 times for 45

seconds, while on ice. The lysate was then incubated with 0.01 mg/mL DNase 1 (Merck, Germany) for 30 minutes at 20 °C. DNase 1 was added to fragment bacterial DNA thereby preventing DNA-contamination in downstream applications, as well as to reduce the viscosity of the solution (Lazarus and Wagener, 2019). The soluble and insoluble fractions of the lysate were then separated by centrifugation (18 000 xg for 30 minutes, at 4 °C). The supernatant was subsequently filtered with a 0.45 µm filter to remove any insoluble matter. The supernatant was used in the subsequent purification steps since it contained soluble protein.

#### *4.1.4.1. Immobilized metal-ion affinity chromatography*

The His-tagged TBR1 T-box domain was purified from the soluble crude cell lysate by immobilized metal-ion affinity chromatography (IMAC), which is based on the known affinity of divalent metal cations such as $Ni^{2+}$, $Cu^{2+}$, $Co^{2+}$ and $Zn^{2+}$ for cysteine or histidine residues. ***Figure 10*** has been used to show how IMAC is used to fractionate and purify proteins. A chelating agent such as iminodiacetic acid (IDA) is covalently bound to a solid support matrix, such as Sepharose, used to entrap metal ions (Gaberc-Porekar and Menart, 2001). The IDA ligand shown in ***Figure 10*** is tridentate and thus forms three dative co-ordinate covalent bonds with the metal ion. The entrapped metal ion then forms two co-ordinate covalent bonds with alternating histidine residues present in the His-tag fused to the TBR1 T-box domain, thereby immobilizing it from the solution (Block et al., 2009). Metal ions are acidic and have an affinity for the exposed amino acid residues on the surface of a protein. More specifically, the imidazole ring in histidine is rich in electron donor groups which form co-ordinate bonds with divalent metal cations. Since the TBR1 T-box was fused to a His-tag, it could be immobilised from solution via the mechanisms mentioned above. The protein was eluted using a relatively high concentration of imidazole, which competes with the histidine residues for the metal ions displacing the protein from the IDA ligand. IMAC is ideal for protein purification as the nucleases and proteases present in the crude cell lysate cannot react with or interfere with the chemically inert Sepharose solid support matrix. Other benefits include ligand stability, the ability to purify large volumes of lysate, mild elution conditions that will not denature the protein, simple protocol regeneration and a low cost(Gaberc-Porekar and Menart, 2001). With IMAC, a 95% pure protein sample can be obtained, with a 90% recovery rate from each elution step (Bornhorst and Falke, 2000).

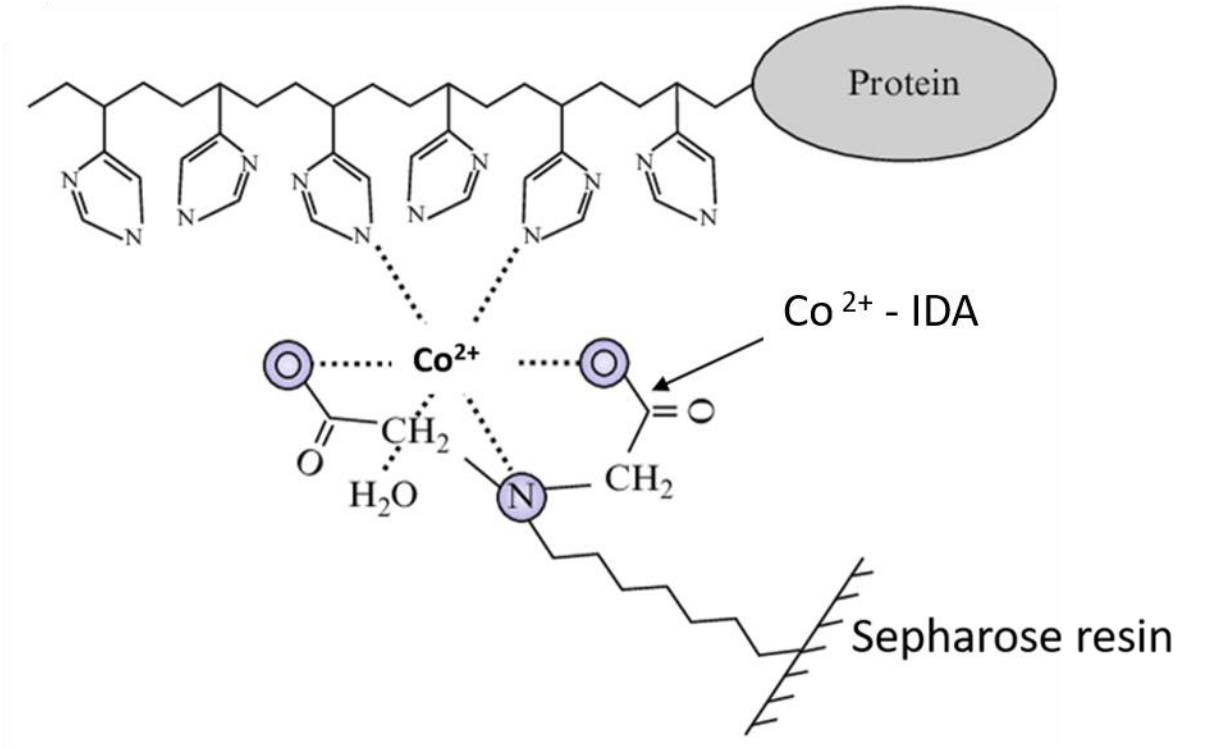



***Figure 10. Putative structure of the metal co-ordination complex used to purify proteins in IMAC.*** *The iminodiacetic acid (IDA) ligand is a chelating agent that has been covalently bound to a solid support matrix consisting of Sepharose. It is used to entrap metal ions such as $Co^{2+}$ through co-ordinate covalent bonds from the carbon and oxygen and nitrogen atoms shown in blue. The immobilized metal-ion is then used to form two co-ordinate bonds with the imidazole ring of two alternating histidine residues in the His-tag, thereby purifying the His-tagged TBR T-box domain. The protein can be eluted from this complex by adding a high concentration of imidazole which will compete with and displace the histidine residues from the metal ions. The co-ordinate bonds have been shown as dashed lines while the covalent bonds are shown as solid lines. This image was adapted from Block et al., 2009.*

The TBR T-box domain was purified on a 5 mL IMAC column (1.6 x 2.5 cm HisTrap column (GE Healthcare, USA) connected to an ÄKTA Prime Liquid Purification System (GE Healthcare, USA). The column was pre-packed with Sepharose, used the IDA ligand and was charged with $Co^{2+}$ ions. All steps were carried out at a flow rate of 5 mL/minute and the pressure limit was set to 0.3 MPa.

The IMAC column was equilibrated with 10 column volumes of equilibration buffer to ensure that the protein molecules interacted with the ligand and were retained by the medium. Subsequently, 50 mL of the filtered supernatant (containing the soluble fraction of the lysate) was loaded onto the column and the flow through was collected. A salt detergent wash was then carried out by passing 5 column volumes of salt wash (1.5 M NaCl, 30 mM imidazole, 0.5% Tween20 and 20 mM Tris-HCl pH 7.5) through the column, and subsequently collected.

The high salt concentration removed DNA contamination while the detergent (Tween 20) removed non-specifically bound proteins from the column. The column was then re-equilibrated with 10 column volumes of equilibration buffer. The His-tagged-TBR1 T-box domain was then eluted from the column by using 5 column volumes of elution buffer (500 mM NaCl, 300 mM imidazole and 20 mM Tris-HCl pH 7.5) and collected. All the samples that were collected (from the flow-through, salt wash and elution) were assessed on a 12.5% SDS-PAGE gel, as described in section 4.1.5.1.

#### *4.1.4.2. Size exclusion chromatography*

Size exclusion chromatography (SEC), also known as gel-filtration chromatography, was used to further purify the TBR1 T-box domain because the protein was not completely pure after IMAC. SEC was used to get rid of contaminants such as proteins with histidine-rich clusters that resemble the His-tag. In column-based SEC, molecules are separated according to their size, by a molecular sieve mechanism. ***Figure 11*** has been used to show the principle behind SEC. The stationary phase consists of a spongy gel consisting of porous, hydrated beads (blue) made from dextran (Superdex). The pores of the beads span a narrow size range of molecular masses.

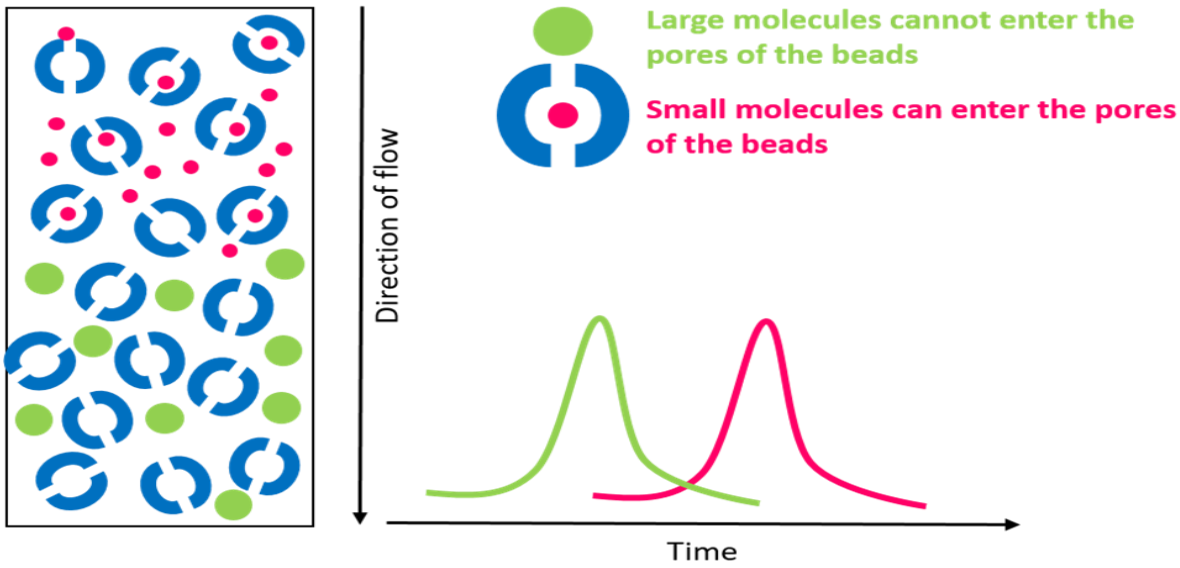



***Figure 11. A depiction of the principles behind size-exclusion chromatography.*** *SEC is used to separate molecules based on their size and shape, or more accurately their dynamic volume. The stationary phase is made of porous dextran beads (blue). Molecules that are too large to fit through the pores (green) are defined as the size exclusion limit and will be eluted first. Molecules with molecular masses lower than the exclusion limit (pink) will then be eluted from the column in the order of decreasing size. Smaller molecules elute later because they have a larger distance to travel, as a result of entering the porous beads.*

The porosity of the beads, and hence the resolution of the column, is determined by the amount of glycerol ether units that crosslink the hydroxyl groups of poly-glucose chains (Chavda and Patel, 2011). The molecular mass of the smallest molecule that is unable to penetrate the pores in the beads is appropriately defined as the size exclusion limit and represents molecules too large to enter the beads. These have been presented as the large green circles in ***Figure 11***. They move through the column without being impeded and are hence eluted first, as seen in the chromatogram. Molecules that have molecular masses below the exclusion limit will then elute in order of decreasing molecular mass. This is because smaller molecules, depicted by the small pink circles in ***Figure 11,*** enter the pores in the beads and thus have a larger volume to traverse.

There is a linear relationship between the elution time (or volume) and the log of the molecular mass, over a considerable range of molecular mass. As such, the molecular weight of an unknown sample can be interpolated from a standard curve of elution time (or volume) versus the log of the molecular weight, assuming that the macromolecules all have the identical shape (Hong et al., 2012). The exclusion limit is to some extent also a function of the hydrodynamic volume. This is because linear elongated molecules are less likely to enter the porous beads when compared to spherical globular molecules. At low flow rates however, the hydrodynamic volume has negligible effects on the molecular separation as the molecules have a larger time to access the porous beads in the column (Grubisic et al., 1967). The advantages of SEC include a high-resolution separation of large and small molecules with minimal eluent volume, as well as minimal sample loss since there are no chemical reactions between the protein and the inert porous beads of the stationary phase.

The TBR T-box domain was further purified by size exclusion chromatography (SEC) on a HiLoad® 16/600 Superdex® 75 preparatory grade column (16 cm x 60 cm (GE Healthcare, USA)) in conjunction with an ÄKTA Prime Liquid Purification System (GE Healthcare, USA). The protein was further purified by SEC as there were some contaminants carried over from IMAC. The column was pre-packed with agarose-dextran having an average particle size of 34 µm. The flow rate was set to 1 mL/minute, the pressure limit was set to 0.3 MPa. The SEC column was equilibrated with 2 column volumes of equilibration buffer, after which 2 mL of the sample was loaded. To prevent unwanted dimer formation through disulphide bonds, the purification was carried out in the presence of 2 mM DTT (reducing agent). The standards

were determined using the Low range gel filtration calibration kit (GE Healthcare, USA), as per the prescribed protocol. All the samples that were collected were assessed on a 12.5% SDS-PAGE gel, as described in section 4.1.5.1.

### 4.1.5. Assessment of protein purity and concentration

In order to obtain relevant information and to ensure that the conclusions being made were about the protein of interest, the purity and homogeneity of the TBR1 T-box domain was thoroughly assessed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) and absorbance spectroscopy.

#### *4.1.5.1. SDS – PAGE*

SDS-PAGE is used to separate proteins based on their molecular weight. In this technique, proteins migrate through a porous polyacrylamide gel in response to an applied electric filed. The rate of migration is inversely proportional to the molecular mass, provided the ratio of charge to mass remains constant (Laemmli, 1970). Proteins do not inherently have a constant charge to mass ratio, and therefore, they need be fully denatured (linearized) and coated with a uniform charge. The samples are denatured by heat and reduced with β-mercaptoethanol. Sodium dodecyl sulphate, an anionic surfactant, is used to coat the proteins with a uniform negative charge. SDS-PAGE is advantageous because of the small sample volume required as well as the high sensitivity offered by this technique. The resolution of SDS-PAGE is greatly improved by using a discontinuous buffer system, consisting of a stacking gel on top of a separating gel. As a result of the differences in the pH, ionic strength and pore size between the gels, the proteins are concentrated into a narrow zone in the stacking gel. This leads to a better separation at the beginning of electrophoresis (Smith, 1984). In the stacking gel (pH 6.8), the large, negative chloride ions migrate the fastest due to their strong ionic charge, followed by the negatively charged proteins and finally the glycinate ions due to their relatively weak ionic charge. In the separating gel (pH 8.8), the glycinate ions assume a large negative charge and hence overtake the protein, creating a uniform electrical field for proteins to migrate in solely based on size. At the boundary between the two gels, the glycinate ions sweep past the protein as a result of encountering a relatively large increase in pH (Smith, 1984).

SDS-PAGE was used to determine the size of the TBR1 T-box domain as well as to detect any contaminants. The size of the TBR1 T-box was predicted to be ≈25 kDa based on the sequence

by the Expert Protein Analysis system (ExPASy) ProtParam tool. The samples were prepared by diluting the protein twice in reducing sample buffer (125 mM Tris-HCl pH 6.8, 4% (w/v) SDS, 20% (v/v) glycerol, 10% (v/v) β-mercaptoethanol, 3.5 µg/mL bromophenol blue) and then boiling it for 5 minutes at 95 °C. This was done to ensure that the protein was fully linearized and coated in a uniform negative charge. The samples were then thoroughly mixed to ensure proper mixing.

The Bio-Rad Mini Protean™ Tetra Cell electrophoresis set (Bio-Rad Laboratories, USA) was used to conduct SDS-PAGE. After the gel-casting apparatus had been setup, 12.5% acrylamide separating gel (12.5% (w/v) acrylamide, 1.08% (w/v) bis-acrylamide, 250 mM Tris-HCl pH 8.8, 0.1% (w/v) SDS, 10% (w/v) ammonium per sulphate and 0.2% (v/v) TEMED) was prepared and poured. Once it had polymerised, a 4% acrylamide stacking gel (4% (w/v) acrylamide, 0.36% (w/v) bis-acrylamide, 50 mM tris-HCl pH 6.8, 0.1% (w/v) SDS, 0.005% ammonium per sulphate and 0.2% (v/v) TEMED) was prepared and poured. A 10-well comb was inserted into the stacking gel after it was poured. Once polymerization had occurred, the gels were placed into the electrophoresis tank containing the SDS-PAGE electrophoresis buffer (250 mM tris-HCl pH 8.3, 192 mM glycine and 0.15% (w/v) SDS).

Sample volumes of 10 µL were loaded into each well. Electrophoresis was subsequently carried out in SDS-PAGE electrophoresis buffer, at 165 V until the dye front reached the bottom of the gel. The gels were stained with Coomassie stain solution (0.1% (w/v) Coomassie Brilliant Blue R-250 in 1:5:4 (v/v/v) acetic acid-methanol-water) for 3 hours. The gels were subsequently destained with destain solution (1:5:4 (v/v/v) acetic acid-methanol-water) for 16 hours, or as required. The molecular weight of the samples was determined relative to the migration of the proteins in the BLUeye Prestained protein Ladder (Merck, Germany). A Molecular Imager® Gel Doc™ XR system (Bio-Rad Laboratories, USA) was used to document the gels, as well as quantitively assess protein purity through densitometric analysis. Protein samples that were at least 95% pure were used further in this study.

#### *4.1.5.2. Absorbance spectroscopy*

Proteins absorb light at 280 nm ($A_{280}$) primarily due to the aromatic amino acid residues tyrosine, tryptophan and phenylalanine (Schmid, 2001). This is because of the resonance brought about by the delocalisation of π-electrons in the aromatic rings (Anthis and Clore, 2013). The $A_{280}$ is therefore used to quantitatively determine protein concentration using the

Beer-Lambert law. Nucleic acids absorb light at 260 nm due to the resonance structure in purine and pyrimidine bases (Olson and Morrow, 2012). The absorbance at 260 nm ($A_{260}$) is therefore used to detect nucleic acid contamination. The $A_{280}/A_{260}$ is thus used as an indicator of the amount of nucleic acid contamination in a purified protein sample and should be as high as possible (≥ 1.8).

Absorbance spectroscopy was used to determine the concentration of the protein, as well as to quantify the contamination. The protein was first centrifuged to remove any aggregation (12 000 xg for 15 minutes, at 4 °C). The absorbance spectrum was obtained from 240 to 350 nm, in triplicate, on a Jasco V-630 spectrophotometer (Jasco, USA) in scanning mode. All experiments were performed in the presence of 2 mM DTT to prevent the formation of unwanted dimers. The buffer contributions were accounted for and the absorbance at 340 nm ($A_{340}$) was subtracted from the $A_{280}$ to correct for protein aggregation. The resultant $A_{280}$ was used to determine the molar concentration of the protein according to the Beer-Lambert law as follows,

$$c = \frac{A_{280}}{\varepsilon \cdot \ell}$$

where $c$ is the protein concentration (M), $A_{280}$ is the corrected absorbance at 280 nm (dimensionless), $\varepsilon$ is the molar extinction coefficient at 280 nm and $\ell$ is the path length (cm). The molar extinction coefficient is the sum of the aromatic contributions of all the aromatic amino acids present in the protein, and was predicted to be 36 440 $M^{-1}.cm^{-1}$ for TBR1 T-box by the ExPASy ProtParam tool (Gasteiger et al., 2005).

## 4.2. DNA preparation

Since the DNA-binding T-box domain is evolutionarily conserved amongst T-box family members, they all bind to an 8 base-pair long consensus sequence 5’ – TCACACCT – 3’, known as the TBE (Wilson and Conlon, 2002). The T-box crystal structures have been determined in the presence of various DNA sequences including a palindromic sequence containing 2 TBEs and a single site short sequence containing 1 TBE (Coll et al., 2002; Müller and Herrmann, 1997; Stirnimann et al., 2010). These sequences are not present in the genome and were probably used to promote the crystallization process or to increasing the DNA-binding affinity, and thus do not have any real physiological relevance. The DNA chosen for this experiment was a natural single site long DNA sequence (SSL) 5’ – CCCAATTTCACACCTTCCTCA – 3’ that

has been found in the Auts2 promoter in the brain and thus has physiological relevance (Bedogni et al., 2010).The SSL DNA has been used in all downstream applications.

The DNA oligonucleotides were synthesized by Integrated DNA technologies (USA). The lyophilised DNA was solubilized to a concentration of 1 mM in Milli-Q® water and stored at -20 °C until required. The SSL DNA used for fluorescence anisotropy was also synthesized by Integrated DNA Technologies (USA) and labelled at the 5' end with 5-carboxy-X-rhodamine (ROX). The lyophilised DNA was solubilized to a concentration of 100 µM in Milli-Q® water and stored at -20 °C until required.

## 4.3. Characterization of protein structure and stability

### 4.3.1. Circular dichroism spectropolarimetry

Briefly, circular dichroism (CD) is defined as the unequal absorption of left- and right-handed circularly polarized light (Woody, 1995). Electromagnetic radiation in the form of light waves is a transverse wave, which consists of a magnetic field and an electric field that oscillate perpendicular to each other, as well as perpendicular to the direction of propagation of the wave. In linearly polarized light, the electric field vector oscillates only in one plane. By passing this light through the appropriate prisms, the electric field vector will oscillate in a plane about its direction of propagation while having a constant magnitude, thus circularly polarizing the incident ray (Greenfield, 2006a). When viewing this sinusoidal wave from the front it appears as the resultant of two vectors of the same length which trace out circles, one which rotates clockwise ($E_R$) and one which rotates anti-clockwise ($E_L$). When asymmetric molecules, such as chiral carbon centres present in polypeptide backbones, interact with light they absorb the left- and right-handed circularly polarized light to different extents and with different refractive indices. This is because the peptide bonds present have two distinct electronic transitions in the far UV wavelength range (180 – 250 nm); the *n* to $\pi^*$ transition (215 – 230 nm) and the $\pi$ to $\pi^*$ (185 – 200 nm). CD is thus reported as the difference in absorbance between $E_L$ and $E_R$ ($\Delta E$) by an asymmetric molecule and can be expressed as the degrees of ellipticity (Greenfield, 2006). For proteins, the $\Delta E$ is indicative of the secondary structural composition when measured in the far UV range. This results in distinct CD-spectra for different secondary structural features present in the protein. The typical far-UV CD spectrum of a β-sheeted protein will have trough at around 218 nm and a peak at around 195 nm. The typical far-UV CD spectrum of a α-helical protein will have troughs at around 208 nm and 222

nm, and a peak at around 190 nm (Greenfield, 2006; Woody, 1995). CD measurements are experimentally limited by the use of aqueous buffers that absorb light in the same wavelength where the secondary structures differentially absorb circularly polarized light.

Far-UV CD spectropolarimetry was used to characterize the secondary structure of the TBR1 T-box domain in the presence and absence of SSL DNA. This was done to confirm that the protein was predominantly β-sheeted as expected, which would mean that it is correctly folded, and to determine if DNA-binding caused any changes to the secondary structure of the protein. The TBR1 T-box domain was dialyzed against DNA-binding buffer (10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.5 and 100 mM NaCl). The far-UV CD spectra were obtained from 200 – 250 nm, in triplicate, on a Jasco J-1500 spectropolarimeter (Jasco, USA). Each replicate was averaged from 10 accumulations. The spectra were obtained at 20 °C, using a bandwidth of 5 nm, a scanning speed of 200 nm/minute and a pathlength of 2 mm. For spectra in the absence of SSL DNA, the sample consisted of 10 µM protein and 2 mM DTT. For spectra in the presence of SSL DNA, the sample consisted of 10 µM protein, 2mM DTT and 5 µM SSL DNA. The samples were diluted five times with Milli-Q® water such that the final NaCl concentration was only 20 mM. This was done to minimize the noise caused by the presence of chloride ions, which absorb light in the far-UV wavelength range. The far-UV signal that was measured was recorded in millidegrees (mdeg) and subsequently normalized to mean residue ellipticity as follows,

$$\theta_{MRE} = \frac{\theta\ (mdeg) \times\ 10^6}{\ell \cdot C \cdot n}$$

where $\theta_{MRE}$ is the mean residue ellipticity (deg.cm$^2$.dmol$^{-1}$), $\theta\ (mdeg)$ is the recorded signal (mdeg), $\ell$ is the pathlength (mm), $C$ is the protein concentration (µM) and $n$ is the number of peptide bonds in the protein backbone. For samples with no DNA, the average spectrum was corrected by subtracting the buffer spectrum. For samples with DNA, the average spectrum was corrected by subtracting the spectrum of 5 µM DNA in buffer. Noisy data, with a high tension (voltage) over 700 V, was discarded.

Circular dichroism spectropolarimetry is also often used to study thermal unfolding. Proteins have a distinct CD spectrum in the far-UV wavelength region, because of the secondary structural content. As the temperature of the protein increases and it unfolds, the secondary structure of the protein becomes less defined, and the characteristic far-UV CD spectrum of

the protein becomes deconvoluted (Greenfield, 2006; Ireland et al., 2018). The thermal melt is monitored at a wavelength that gives the strongest far-UV CD signal, to maximize the signal difference between the native and unfolded states of the protein. The loss or gain in the far-UV signal can be used to construct an unfolding curve (Greenfield, 2006). These are used to determine the melting temperature of the protein, as well as to detect any thermodynamically stable intermediates in the unfolding pathway (Greenfield, 1999). Equilibrium unfolding can also be studied through this technique, only if the unfolding reaction is highly reversible.

The melting temperature of the TBR1 T-box domain, in the presence and absence of SSL DNA, was monitored by far-UV CD spectropolarimetry. This was done to determine if DNA-binding had any significant effect on the thermal stability of the protein. An increase in thermal stability upon DNA-binding may suggest that helix $3_{10}$C becomes structured upon DNA-binding. The TBR1 T-box domain was dialyzed against DNA-binding buffer. For samples in the absence of SSL DNA, the sample consisted of 10 µM protein and 2 mM DTT. For samples in the presence of SSL DNA, the sample consisted of 10 µM protein, 2mM DTT and 5 µM SSL DNA. The thermal melts of the protein were obtained on a Jasco J-1500 spectropolarimeter (Jasco, USA) by monitoring the far-UV CD signal at 213 nm. The far-UV CD spectrum of the protein showed a characteristic trough at 213 nm, which is why it was chosen to monitor thermal unfolding. The thermal melts were conducted over a temperature range of 20 – 80 °C, in duplicate, with a bandwidth of 5 nm and a heating rate of 1 °C/minute. Noisy data, with a high tension (voltage) over 700 V, was discarded. Equilibrium unfolding studies could not be carried out as the unfolding reaction was only ≈ 60% reversible.

### 4.3.2. Intrinsic tryptophan fluorescence spectroscopy

Fluorescence is the phenomenon by which a molecule, known as a fluorophore, emits light after being excited by electromagnetic radiation (Lakowicz, 2006). Electromagnetic radiation is absorbed by the fluorophore, causing its electrons to become excited to higher energy levels. The fluorophore first loses some of this energy in order to return to a lower energy level, through non-radiative energy decay caused by heat dissipation and vibrational rotations. The fluorophore then returns to the ground state by emitting light (Lakowicz, 2006). Due to the non-radiative energy decay, the light that is emitted is of a longer wavelength than that used to excite the fluorophore. The intensity of the emission is dependent on the nature

and concentration of the fluorophore, as well as its environment. The fluorophore may also return to the ground state by losing energy through quenching, which can be static or dynamic (Lakowicz, 2006).

Proteins may exhibit intrinsic fluorescence if they possess amino acids that act as fluorophores, such as tryptophan, tyrosine, phenylalanine, and cysteine. Phenylalanine and cysteine do not significantly contribute to fluorescence due to their low quantum yields (Lakowicz, 2006). Tryptophan is the best candidate for fluorescence as it has a much higher quantum yield than tyrosine and can be selectively excited at 295 nm. Intrinsic tryptophan fluorescence (ITF) can be used as a probe to assess the tertiary structure of a protein because the indole group of tryptophan is solvatochromic, meaning its emission wavelength is dependent on the polarity of its environment (Lakowicz, 2006). When tryptophan residues become buried within the protein core, like when the protein is folded, a blue shift in the fluorescence emission occurs (hypsochromic) because of decreased tryptophan exposure to the polar solvent. When tryptophan residues become exposed to the polar solvent, as in the case of protein-unfolding, a red-shift occurs in the fluorescence emission (bathochromic) (Eftink, 2000).

ITF spectroscopy was used to determine if the TBR1 T-box domain had been properly folded. The spectra were obtained in the presence or absence of a denaturant (8 M urea), to observe spectral differences between the native and unfolded proteins. The protein was first dialyzed against DNA-binding buffer. The protein was then incubated with 8 M urea for 1 hour, at 20 °C. Using an excitation wavelength of 295 nm, the emission spectra were obtained from 280 – 450 nm, in triplicate, on a Jasco FP-6300 spectrofluorometer (Jasco, USA). Each replicate was averaged from 3 accumulations. The spectra were obtained at 20 °C using excitation and emission bandwidths of 5 and 2.5 nm respectively, a scanning speed of 200 nm/minute and a path length of 1 cm. For spectra in the absence of urea, the sample consisted of 2 µM protein and 2 mM DTT. For spectra in the presence of urea, the sample consisted of 2 µM protein, 2mM DTT and 8 M urea. For samples with no urea, the average spectrum was corrected by subtracting the buffer spectrum. For samples with urea, the average spectrum was corrected by subtracting the spectrum of 8 M urea in buffer. The data was normalized by dividing each corrected fluorescence intensity value by the maximum fluorescence intensity value of the

protein in the absence of the denaturant. The data was normalized to make the spectra easier to compare.

ITF spectroscopy was also used to determine if there are any changes in tertiary structure that occur upon DNA-binding. ITF spectra were obtained in the presence of absence of SSL DNA. The protein was first dialyzed against DNA-binding buffer. For samples in the presence of DNA, the sample was incubated with SSL DNA for 1 hour, at 20 °C. Using an excitation wavelength of 295 nm, the emission spectra were obtained from 280 – 450 nm, in triplicate, on a Jasco FP-6300 spectrofluorometer (Jasco, USA). Each replicate was averaged from 3 accumulations. The spectra were obtained at 20 °C using excitation and emission bandwidths of 5 and 2.5 nm respectively, a scanning speed of 200 nm/minute and a path length of 1 cm. For spectra in the absence of DNA, the sample consisted of 2 µM protein and 2 mM DTT. For spectra in the presence of DNA, the sample consisted of 2 µM protein, 2mM DTT and 1 µM SSL DNA. For samples with no DNA, the average spectrum was corrected by subtracting the buffer spectrum. For samples with DNA, the average spectrum was corrected by subtracting the spectrum of 1 µM SSL DNA in buffer. The data was normalized by dividing each corrected fluorescence intensity value by the maximum fluorescence intensity value of the protein in the absence of DNA. The data was normalized to make the spectra easier to compare.

## 4.4. DNA-binding studies

### 4.4.1. Electrophoretic mobility shift assay

An electrophoretic mobility shift assay (EMSA) is a technique that is routinely used to assess DNA-binding. The assay is based on the electrophoretic separation of protein-DNA complexes from free DNA in solution, and is conducted on a porous polyacrylamide gel (Hellman and Fried, 2007). In an EMSA, DNA is used as a probe to detect protein-DNA interactions. Since the rate of migration through a porous gel is dependent on size, protein-DNA complexes will migrate slower than free DNA or free protein. The bands from protein-DNA complexes are thus shifted upwards with respect to the bands of the free DNA in the reaction mixture. Since EMSAs are a low-resolution technique, and the binding reaction is not equilibrated, quantitative analyses are avoided.

An EMSA was carried out to confirm the interaction between the TBR1 T-box domain and SSL DNA and thus confirm that the protein was properly folded and functional. The EMSA was also used to get an idea of the protein concentration required for saturated DNA-binding. This

information was required for downstream applications such as fluorescence anisotropy and crystal trials. The TBR1 T-box domain was dialyzed against DNA-binding buffer. The protein-DNA samples were prepared in EMSA binding-buffer (10 mM HEPES pH7.5, 100 mM KCl, 1 mM $MgCl_2$, 10% glycerol and 0.1 mg/mL bovine serum albumin (BSA)). The concentration of the SSL DNA was kept constant at 0.5 µM and the protein concentrations were increased from 0 to 8 µM. The molar ratios of DNA:Protein used were 3:1, 1:1, 1:5, 1:10 and 1:20. An 8% polyacrylamide gel was prepared (8% (w/v) acrylamide, 0.69% (w/v) bis-acrylamide, 5xTBE buffer (450 mM Tris-HCl, 450 mM boric acid, 13 mM $Na_2EDTA.H_2O$, pH 8.3). The Bio-Rad Mini Protean™ Tetra Cell electrophoresis set (Bio-Rad Laboratories, USA) was used to conduct electrophoresis. The samples, gels, and electrophoresis buffer (1xTBE) were all preequilibrated to 4 °C for 1 hour. A sample volume of 50 µL was used. Electrophoresis was carried out at 165 V for 1.5 hours, at 4 °C. The gels were then stained with a 1 in 10 000 dilution of GelRed® Nucleic Acid Gel Stain (Anatech, RSA) in Milli-Q® water for 5 minutes. The gels were immediately viewed, under UV-light, with a Molecular Imager® Gel Doc™ XR system (Bio-Rad Laboratories, USA).

The resolution of the EMSAs was greatly improved by conducting pre-electrophoresis of the gel at 75 V for 45 minutes, reducing the sample volume to 4 µL, and using elongated gel-loading tips to load the samples.

### 4.4.2. Fluorescence anisotropy

According to Brownian motion, molecules in solution are constantly moving due to inelastic collisions with other molecules. One such movement is rotational diffusion, or the tendency of a molecule to 'tumble' in solution such that the thermal equilibrium of the system is maintained (Miller, 1981). The fluorescent emission of a molecule is dependent on both the environment in which a fluorophore resides, as well as the polarization of the excitation light used (Favicchio et al., 2009). When a fluorophore is excited with polarized light it will emit light that is only partially polarized (Gijsbers et al., 2016). This means that when fluorophores in solution are excited with plane polarized light, they will emit light that is only partially polarized in that plane. The extent to which the emitted light is partially polarized is defined as the fluorescence anisotropy of a molecule and is described by the equation,

$$r = \frac{I_v - I_h}{I_v - 2I_h}$$

where $r$ is the anisotropy, $I_v$ is the intensity of vertically polarized light and $I_h$ is the intensity of horizontally polarized light (Favicchio et al., 2009; Weber, 1953). The degree of depolarization exhibited during the fluorescent lifetime of the excited state is inversely proportional to the rate at which a fluorophore rotates (Moerke, 2009). When the rotational rate of a fluorophore is much faster than the lifetime of the excited state, the emitted light will be completely depolarized, resulting in an anisotropy of 0. However, when the rotational rate is slower than the fluorescent lifetime, the emitted light will have less depolarization and the anisotropy will be 1. If a fluorophore is attached to a protein, its rotational rate is slowed such that the emission remains partially polarized. When a fluorescently labelled DNA molecule binds to a protein, the size of the complex is substantially increased, resulting in the complex tumbling more slowly in solution relative to the free DNA, thereby increasing the anisotropy (Gijsbers et al., 2016). The anisotropy of the fluorescently labelled DNA is monitored, and as the concentration of the protein is increased, the anisotropy increases until all the DNA molecules have been saturated.

Fluorescence anisotropy is an appropriate technique for studying DNA-protein interactions since there is a large size difference between the DNA-protein complex and the free DNA in solution. This is necessary as the anisotropy of a free protein molecule is already quite high to begin with (Lakowicz, 2006). Extrinsic fluorophores are typically used in fluorescence anisotropy as it maximises the signal difference between the free DNA and the protein-DNA complexes in the solution. There are a number of commercially available fluorescent labels used to label DNA such as ATTO™, Alexa Flour® and Rhodamine dyes.

FA was carried out to determine the dissociation constant ($K_D$) for the interaction between the TBR1 T-box domain and SSL DNA, in the presence and absence of $MgCl_2$. This was done because various T-box proteins have been successfully crystallised in the presence of $MgCl_2$. Metal ions have also been shown to regulate DNA-binding of other transcription factors through changes in DNA-binding affinity and specificity (Coll et al., Moll et al., 2002; Stirnimann et al., 2010). The protein was first dialyzed against buffer containing 10 mM HEPES pH 7.5, 100 mM NaCl, 2 mM DTT and 1 mM EDTA. Ethylenediaminetetraacetic acid (EDTA) was used to chelate any $Mg^{2+}$ ions that were present. EDTA was subsequently removed by dialyzing the protein against 10 mM HEPES pH 7.5, 100 mM NaCl and 2 mM DTT, pre-treated with Chelex® Resin (Bio-Rad Laboratories, USA). The SSL DNA was labelled at its 5’ end with

5-carboxy-X-rhodamine (ROX). The experiment was carried out at 20 °C on a Perkin Elmer LS-50B luminescence spectrometer. The excitation and emission wavelengths were 580 and 605 nm respectively, the bandwidth was 5 nm, and the integration time was 3 seconds. The grating factor (G Factor), automatically determined by the instrument, accounts for the instrument's differential transmission of the horizontal and vertical vectors. The G Factor was determined from ROX-labelled SSL DNA and used to correct the anisotropy. For anisotropy in the absence of $Mg^{2+}$, samples were prepared by titrating 200 nM ROX-labelled SSL DNA with a relatively high concentration of the TBR1 T-box domain (≈11 µM). For anisotropy in the presence of $Mg^{2+}$, samples were prepared by titrating 200 nM ROX-labelled SSL DNA containing 100 µM $MgCl_2$ with a relatively high concentration of the TBR1 T-box domain (≈11 µM). The data was obtained in triplicate, and each replicate was averaged from 10 accumulations. The anisotropy of free ROX-labelled DNA was subtracted from each data point in the replicate to obtain the relative anisotropy. The average relative anisotropy values were plotted against the protein concentration (nM) to construct a binding isotherm. The binding isotherm was then fit to a single-site saturation binding model in GraphPad Prism v 9.2.0. The equation for the fit is given by,

$$Y = \frac{(B_{max})(x)}{(K_{\mathrm{D}} + x)}$$

where $Y$ is the relative anisotropy (unitless), $B_{max}$ is the relative anisotropy when binding is saturated (unitless), $x$ is the protein concentration (nM) and $K_{\mathrm{D}}$ is the dissociation constant (unitless).

### 4.5. Protein crystallography

The structure of a molecule defines its function and properties. The primary method used to determine the structure of a molecule, at an atomic resolution, is XRC. A crystal is a solid material whose constituents are arranged in an orderly and repeating pattern in all directions. Under a highly specific set of conditions, protein molecules have the remarkable ability to self-assemble into a crystal (McPherson and Gavira, 2013). Since the molecules making up the crystal have an ordered and repeating pattern, they are used to diffract an incident beam of X-rays into various directions. The angles at which these beams are diffracted, and the relative intensities of these 'reflections' produce a diffraction pattern in the reciprocal space. A Fourier transformation is then used to transform the pattern back into real space atomic co-

ordinates yielding a three-dimensional map of electron density. Finally, an atomic-resolution model of the protein is built into the map of electron density, to yield a three dimensional crystal structure of the protein (Rupp, 2013).

The molecules in a protein crystal are held together by transient network of sparse, weak intermolecular forces known as crystal contacts (McPherson and Gavira, 2013; Rupp, 2013). For crystallization to take place, crystal contacts must take place between specific residues in particular locations along specific directions. The crystallizability of a protein is thus highly dependent on its sequence. Single protein crystals are typically 50 µm – 0.5 mm in size and display well defined faces and sharp edges in three dimensions. Considering the fact that protein molecules have regions of inherent flexibility and mobility, and because crystal contacts are relatively weak, protein crystals are extremely fragile and thus sensitive to slight changes in the environment (Rupp, 2013). Due to the sparse nature of crystal contacts, protein crystals are composed of approximately 50% solvent (Chruszcz et al., 2008; Matthews, 1968).

The crystallization of diffraction-worthy crystals is a formidable task and is hence the rate limiting step in successful crystallography endeavours (DeLucas and Bugg, 1987). The crystallization of a macromolecule requires that the sample solution reaches a supersaturated state under conditions that do not affect the structure or function of the molecule (McPherson and Gavira, 2014). Supersaturation is a non-equilibrium condition in which some quantity of the macromolecule, in excess of the solubility limit, is present in solution. The process of crystal formation occurs via the spontaneous formation of homogenous nuclei (nucleation) and subsequent crystal growth. Nucleation begins by the formation of prenucleation molecular aggregates. When these aggregates reach a critical size, stable nuclei form with some degree of molecular order. Once nuclei exist, crystal growth can occur through the subsequent addition of protein molecules. The phase diagram depicted in ***Figure 12*** can be used to explain the thermodynamic principles that govern crystal formation (Rupp, 2013). Coordinates on the solubility curve correspond to equilibrium concentrations between the sample solution and the precipitant. The phase diagram may be separated into 3 regions: The unsaturated region (blue), the metastable region (green) and the unstable region (red). The unsaturated region is stable and contains a single phase. Nucleation and crystal growth do not occur in this region of the phase diagram. For crystal formation to occur, the sample

needs to be supersaturated by increasing the protein and/or precipitant concentration. Once the solubility line has been surpassed, the metastable region is reached. Since the equilibrium concentrations are altered in this region, the concentration of the protein in solution is decreased by the formation and development of a crystalline state. The metastable region is divided into the lower limits, in which crystal growth occurs (growth zone), and the upper limits, in which crystal nucleation occurs (labile zone). The sample should thus be supersaturated into the nucleation zone in which crystal formation can be initiated. However, if the sample is too deep into the labile zone, many small nuclei will form resulting in the depletion of protein from the solution, which will off course prevent any further crystal growth. The sample should thus ideally be in the lower limits of the labile zone as this will enable nucleation to occur without depleting the protein from the solution too rapidly, allowing for subsequent crystal growth. If the sample is supersaturated beyond the decomposition line, spontaneous decomposition will occur resulting in protein precipitation.

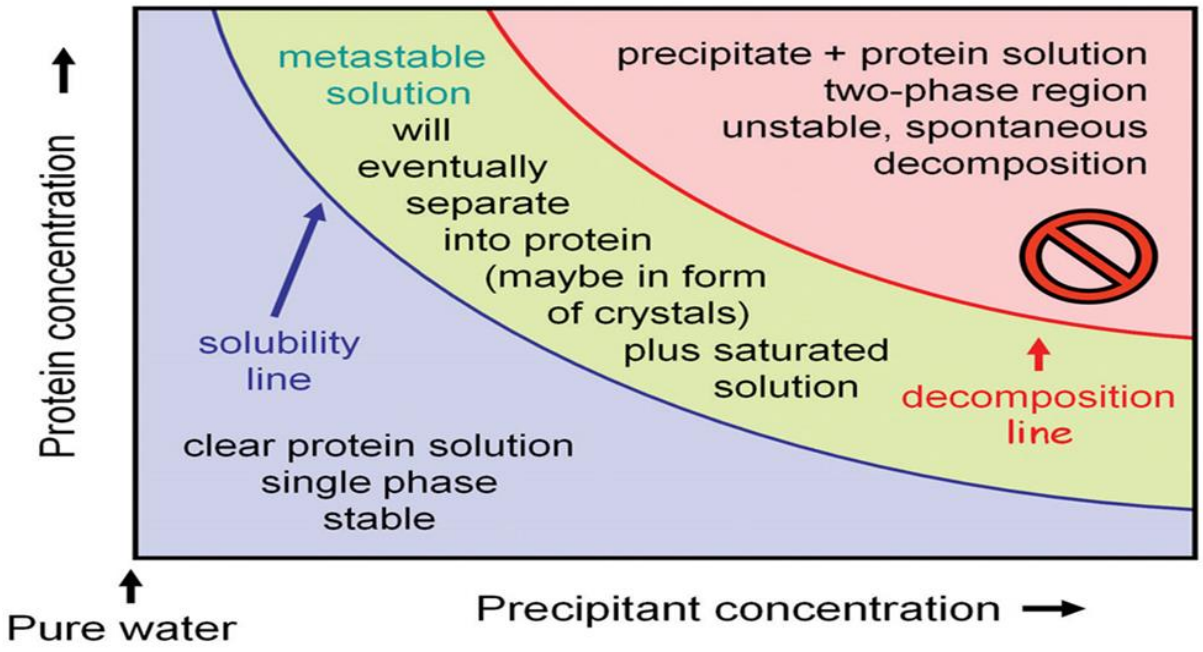



***Figure 12. A typical solubility phase diagram used to explain the thermodynamics of crystal formation.*** *The blue region represents the unsaturated solution in which nucleation and growth do not occur. The sample is then supersaturated by increasing the protein and/or precipitant concentration resulting in a metastable solution. The upper limits of the metastable region (labile zone) favours nucleation while the lower limits (growth zone) favour the growth of crystals from homogenous nuclei. This figure was reproduced with permission from Biomolecular Crystallography by Bernhard Rupp, © 2009-2014 Garland Science/Taylor & Francis LLC.*

Protein crystallization is achieved through a variety of physical means, which trace different paths along the phase diagram in ***Figure 12***, such as batch and vapor-diffusion (Watts, 1993). In batch experiments, the protein is directly mixed with precipitant solution and immersed under a layer of paraffin oil. Precipitant solution, more appropriately referred to as crystallization solution or mother liquor, is actually a combination of precipitant, buffer, salt, and additives used to promote the crystallization of a protein. The volume of the drop remains constant while the concentration of the protein in solution is decreased through the formation of crystals which re-establish the equilibrium via directly attaining supersaturation (Chayen et al., 1992). In vapor-diffusion experiments, a drop consisting of protein and crystallization solution is placed above or beside a reservoir containing an obviously higher precipitant solution concentration. The concentration gradient created by the precipitant solution is the driving force for crystal formation since water will evaporate from the drop until its precipitant solution concentration is equal to that of the reservoir (Rhodes, 2006). This increases the concentrations of both the protein and the precipitant solution, thereby promoting supersaturation. Vapor-diffusion experiments are set up as either a hanging-drop or a sitting-drop. In the hanging-drop method, a drop of protein and crystallization solution is placed on the underside of a coverslip and used to seal a reservoir containing precipitant solution. In the sitting-drop method, a drop of protein and crystallization solution is placed on a raised platform which is sealed within the reservoir (McRee and David, 1999; Rhodes, 2006). The use of various methods and setups increases the number of thermodynamic and kinetic routes that can be taken to achieve supersaturation, thereby increasing the chances of success.

Even though the crystallizability of a protein is highly dependent on its amino acid sequence, there are many variables which affect whether a protein will crystallize, and to what extent (McPherson and Gavira, 2013; McRee and David, 1999; Rhodes, 2006; Rupp, 2013). The sample should be as pure and homogenous as possible. The buffer that the protein is solubilized in will affect its solubility and stability, and hence its crystallization. The salt and buffer concentrations should be kept as low as possible while still maintaining the solubility and stability of the protein in solution. The chaotropic nature of salts commonly used in protein buffers, such as NaCl, disrupts crystallization at high concentrations by interfering with hydrogen bonded networks (crystal contacts). The buffer concentrations are kept low to

ensure that the crystallization solution alters the pH of the drop instead of it being controlled by the protein buffer. The protein concentration is one of the most important factors to consider since it influences whether supersaturation will be achieved, as well as the rates of crystal nucleation and subsequent growth. Proteins crystallize in concentrations ranging from 5 to 25 mg/mL, on average. If the protein concentration is too low, supersaturation will not be achieved. If the protein concentration is too high, precipitation will occur. The ratio of protein to crystallization solution will also affect many aspects of crystallization. When the ratio is high, there won't be enough precipitant to cause supersaturation and when it is too low, precipitation will occur. There are additional variables to consider when crystallizing a protein in the presence of a ligand such as DNA. The crystallization of a protein-DNA complex is highly dependent on the DNA sequence used, its length, whether it has sticky or blunt ends and the presence of palindromic sequences (Jordan et al., 1985). The DNA is often kept in 1.2 molar excess of the protein because DNA annealing is not 100% perfect, the estimations of DNA and protein concentrations are approximate, and DNA longer than 12 base pairs rarely crystallizes on its own (Hollis, 2007).

### 4.5.1. Protein crystallization

The TBR1 T-box domain was dialyzed against Buffer A (20 mM Tris-HCl pH 7.5, 100 mM NaCl and 2mM DTT), Buffer B (10 mM HEPES pH 7.5, 100 mM NaCl and 2mM DTT) or Buffer C (20 mM HEPES pH 7.5, 200 mM NaCl, 2mM DTT, 5mM $MgCl_2$). These buffers were chosen based on previous T-box crystal structures as well as past experiences in our laboratory. The protein was then centrifuged (12 000 xg for 15 minutes, at 4 °C) to remove aggregation and concentrated to 1.4 mM using an Amicon Ultra-0.5 mL Centrifugal Filter (Merck, Germany) with a 10 kDa cut-off. The protein was then flash frozen in liquid nitrogen and stored at -80 °C until required.

All crystallization experiments were carried out at 20 °C, using the Oryx8™ protein crystallization robot (Douglas Instruments, UK) fitted with a three-bore tip system. The hanging-drop experiments were carried out in a Qiagen EasyXtal™ 15-well hanging-drop plate system (Molecular Dimensions, UK). In these experiments, 1 - 5 µL drops were sealed over 0.5 - 1 mL of mother liquor. The sitting drop experiments were carried out in Swissci 48-Well MRC Maxi Optimization Plates (Molecular Dimensions, UK). In these experiments, 1 - 5 µL drops were sealed beside 200 - 350 µL of mother liquor. The micro-batch under oil experiments

were carried out in Swissci 96-Well MRC Under Oil Plates (Hampton research, USA). In these experiments, a 3 µL drop was immersed under 15 µL of Al's oil (50% (v/v) paraffin oil and 50% (v/v) silicon oil). The crystallization conditions used were as follows: Natrix, Natrix 2 and Index crystallization screens (Hampton, USA) as well as crystal solution 1 (5% (w/v) polyethylene glycol (PEG) 6000 and 100 mM citrate pH 5), crystal solution 2 (40 mM magnesium acetate, 50 mM sodium cacodylate pH 6.0, 30% (v/v) 2-methyl-2,4-pentanediol), crystal solution 3 (100 mM 4-morpholineethanesulfonic acid pH 6.5, 200 mM $MgCl_2$ and 25% (w/v) PEG 4 000), crystal solution 4 (100 mM 4-morpholineethanesulfonic acid pH 5.6, 200 mM $MgCl_2$, 200 mM NaCl and 5% (w/v) PEG 6000) and crystal solution 5 (85 mM sodium citrate pH 5.1, 1.7 M sodium formate and 15% (v/v) ethylene glycol). The crystal solutions were obtained from previous T-box crystal structures (Coll et al., 2002; El Omari et al., 2012; Stirnimann et al., 2010). The following drop volume ratios of protein to mother liquor were used in all experiments: 2:1, 3:2, 1:1, 2:3 and 1:2.

Crystal trails in the absence of DNA were performed after dialyzing the protein against Buffer A, Buffer B or Buffer C respectively. The protein was diluted to concentrations ranging from 0.001 – 1 mM. Crystal trials were set up using all of the methods listed above, along with all the kits and crystal solutions listed.

Crystal trails in the presence of DNA were performed after dialyzing the protein against Buffer A, Buffer B or Buffer C respectively. The protein was diluted to concentrations ranging from 0.25 – 0.75 mM. The protein was mixed with SSL DNA in the molar ratios 1.2:1, 1:1 and 1:1.2. I was unable to separate the unbound DNA by SEC because the sample loss was too high, and the mixture would re-equilibrate (to yield unbound DNA) from the time SEC was performed to the time the trials were carried out. The mixture was incubated at 20 °C for 2 hours to allow for the DNA-binding reaction to reach equilibrium. Crystal trials were set up using all of the methods listed above, along with all the kits and crystal solutions listed.

Crystals were harvested with a CrystalCap SPINE HT goniometer base fitted with either a 0.05 – 0.1 mm, 0.1 – 0.2 mm or 0.2 – 0.3 mm Mounted CryoLoop™ (Hampton Research, USA). The crystals needed to be kept at ≈100 K to prevent radiation damage when diffraction was carried out. At this temperature however, the crystals were susceptible to the formation of ice crystals which would hinder the data collection process. The crystals were thus cryoprotected by 3 dips in either 100 (v/v) Parabar (Hampton Research, USA), 30% (v/v)

glycerol (made up with mother liquor) or 30% PEG (from the mother liquor). The presence of the crystal in the loop was confirmed using a darkfield protein crystallography microscope. The crystal was then immediately submerged in liquid nitrogen and taken for data collection.

The data collection was attempted with the in-house light source, the Bruker D8 Venture Dual Wavelength Hybrid Diamond Anode X-ray Diffractometer (Bruker, USA). The instrument is equipped with a 1 µS 3.0 and 1 µS DIAMOND microfocus X-ray sources, and a PHOTON II detector. Data collection was attempted at 100 K.

## 4.6. *In silico* analysis

Since the crystal structures of the TBR1 T-box domain were unobtainable through XRC, the DNA-binding mechanism was predicted through *in silico* studies. This was also done to help interpret the *in vitro* experiments. Since the only T-box crystal structure available in both the DNA-bound and unbound form is that of TBX5, the hypothesis was that the TBR1 T-box domain similarly binds DNA via the formation of an inducible recognition element that becomes structured upon DNA-binding (Stirnimann et al., 2010). The intrinsically disordered regions of the protein were predicted, as was the three-dimensional structure of the protein in the absence of DNA. Finally, molecular docking was used to obtain a binding pose of the TBR1 T-box domain bound to SSL DNA.

### 4.6.1. *Ab initio* protein modelling

Due to the recent advent of significant computational abilities, the prediction of three-dimensional protein structures from a sequence of amino acids has become increasingly common. The use of computational approaches is necessitated by the high cost and extensive time associated with experimental structure determination (Pearce and Zhang, 2021). The approaches used to predict protein structure can be categorized as either template-based (homology) or template-free (*ab initio*) modelling (Pearce and Zhang, 2021). In homology modelling, the structure of a protein is predicted using knowledge of the primary sequence and structural similarities with closely related proteins. The steps in homology modelling include the identification and selection of templates used for structural alignments, backbone modelling, loop modelling, the addition of side chains and model refinement (Hameduh et al., 2020). Recent advances have allowed for homology modelling to be improved by iterative approaches which allow the secondary structures to be rethreaded through the final structure in an attempt to refine it (Yang and Zhang, 2015). *Ab initio* modelling is used to predict protein

structure without the use of global template information obtained from experimentally obtained protein structures (Pearce and Zhang, 2021). This method relies on the generation and assembly of local structural fragments. These fragments are generated from unrelated proteins based on sequence similarity, secondary structure, solvent accessibility, and torsion angles (Pearce and Zhang, 2021). Fragment assembly is advantageous because it reduces the entropy of the conformational search space while ensuring that the local structural fragments are properly defined.

Recent developments in artificial intelligence have drastically improved *ab initio* modelling (Hameduh et al., 2020). Machine learning, a subset of artificial intelligence, is when a computer program can make predictions based on a labelled data set on which it has been trained. Deep learning, a subset of machine learning, is when a computer program can make predictions based on an unlabelled data set on which it has been trained. The field of protein structure prediction has been completely revolutionized by the advent of deep learning techniques, as evidenced by the performance of the AlphaFold 2 software in the recent CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiments (Jumper et al., 2021; Skolnick et al., 2021). AlphaFold 2 is currently the most accurate deep learning-based structure prediction tool (Jumper et al., 2021). Its use is however limited by the need for prior programming knowledge as well as the inability to predict structures from user defined sequences. The RoseTTaFold software is a deep learning-based structure prediction tool that allows for structures to be predicted from user-defined sequences (Baek et al., 2021). The accuracies of the predicted structures are almost as good as that of AlphaFold 2, based on comparisons of the predicted structures with those available in the Protein Data Bank. It makes use of a three-track neural network to make accurate predictions of protein structures and interactions. The information obtained at the one-dimensional sequence level, the two-dimensional distance map level and three-dimensional coordinate level are successively transformed and integrated. This allows the network to collectively reason about connections within and between sequences, distances, and coordinates (Baek et al., 2021).

The sequence of the TBR1 T-box domain, outlined in ***Figure 9***, was submitted to the Robetta webserver (https://robetta.bakerlab.org/) in FASTA format. Using default parameters, the RoseTTaFold algorithm was used to obtain a crystal structure of the TBR1 T-box domain using a deep learning-based modelling method. The structure was then validated using MolProbity

which considers rotamers, Ramachandran outliers, covalent geometries, and backbone torsion angles. (http://molprobity.biochem.duke.edu/index.php) (Williams et al., 2018).

### 4.6.2. Disorder predictions

Intrinsically disordered regions (IDRs) are parts of a protein that do not possess any secondary structure (Babu, 2016). These regions are thus more flexible and dynamic and are thought to contribute to conformational changes that take place when a TF binds DNA (Ishida and Kinoshita, 2007). The presence of many IDRs in a protein has implications for structural characterizations as well as XRC. The PrDOS (Protein Disorder Prediction System) webserver (https://prdos.hgc.jp/) is used to predict the IDRs in a protein from its amino acid sequence (Ishida and Kinoshita, 2007). A sequence profile is created through multiple sequence alignments with homology models and scored by a position-specific score matrix (PSSM). Disorder predictions are made based on both local amino acid sequence information, and template homology modelling. The predictions based on local amino acid sequence information are made by a set of machine algorithms known as support vector machines (SVMs). The PSSM is used in a conventional homology search (PSI-BLAST) to create homology models. The combined prediction score is the weighted average of the results from the two predictions. The results from local amino acid sequences are weighted ten times more than that obtained from homology modelling (Ishida and Kinoshita, 2007).

The IDRs were identified to try and explain why the protein was so difficult to crystallize as well as to help explain the putative role of helix $3_{10}$C in the DNA-binding mechanism since, in the case of TBX5, it only becomes structured in the presence of DNA (Stirnimann et al., 2010). The sequence of the TBR1 T-box domain, outlined in ***Figure 9***, was submitted to the PrDOS webserver in FASTA format. The default parameters were used to predict the intrinsically disordered regions of the protein.

### 4.6.3. Molecular docking

Molecular docking is a computational tool used to predict the preferred orientation of a molecule, such as a protein, when bound to another molecule, such as a ligand (DNA) (Meng et al., 2011). Briefly, docking is achieved by sampling various conformations of the ligand within the active site of its binding partner and subsequently ranking these conformations by a scoring function. The HADDOCK (High Ambiguity Driven protein−protein Docking) webserver is a freely available platform for modelling biomolecular interactions (Dominguez

et al., 2003). Unlike conventional methods, HADDOCK relies on experimentally obtained biophysical and biochemical data from chemical shift perturbations obtained with nuclear magnetic resonance (NMR) or mutagenesis studies. This data is introduced to the docking process through ambiguous interaction restraints (AIRs), which are ambiguous distances between all the residues involved in the interaction. The scoring function ranks models based on their intermolecular energies, which is the sum of the energy contributions from electrostatic interactions, Van der Waals interactions and the AIRs (Dominguez et al., 2003). In HADDOCK, the user defines the residues that are directly involved in binding (active), thus defining the active site, leading to more accurate predictions.

Molecular docking was used to obtain a binding pose for the interaction between the TBR1 T-box domain and SSL DNA, in the hopes that it would provide some information about the DNA-binding mechanism, as well as help to interpret the *in vitro* data obtained in the presence of SSL DNA. A model of SSL DNA, in the B-form, was created in the Accelrys Discovery studio v 4.1. The predicted TBR1 T-box domain crystal structure obtained from RoseTTaFold was docked to SSL DNA using the HADDOCK v 2.4 webserver (https://wenmr.science.uu.nl/haddock2.4/). The active residues in the protein were defined as Ile226, Arg232, Thr370, Ala371, Phe387, Gly390 and Phe391 of the original sequence. The residues were chosen based on homologous T-box crystal structures (Coll et al., 2002; El Omari et al., 2012; Stirnimann et al., 2010). The active residues in the SSL DNA were defined as the TBE (5' – TTCACACCT – 3') since all T-box proteins bind to these residues. The docking protocol was verified by docking the TBX5 crystals structure to its DNA using the same parameters (positive control). If the protocol wasn't flawed, the docked control would align with the crystal structure of TBX5 in the presence of DNA.

# 5. Results

## 5.1. Protein preparation

The aim of this study was to determine the DNA-binding mechanism of the TBR1 T-box domain through structural characterizations, DNA-binding studies and XRC. To that effect, it was necessary to have sufficient quantities of pure protein. T7 express pLysS Competent *E. coli* cells were transformed with a pET-11a plasmid containing the TBR1 T-box domain gene sequence. After confirming the sequence, the protein was successfully overexpressed at 20 °C for 24 hours following induction with 0.2 mM IPTG. The protein was then successfully purified by IMAC and SEC. Thereafter, the purity was quantitatively and qualitatively assessed by SDS-PAGE and absorbance spectroscopy. The protein preparation was successful, and it could therefore be used for downstream applications.

### 5.1.1. Plasmid sequencing

The protein sequence was verified to ensure that the TBR1 T-box domain was successfully inserted into the *E. coli* cells without any mutations. This was done to confirm that the protein being studied was indeed the TBR1 T-box domain. The pET-11a plasmids containing the TBR1 T-box domain gene were isolated from T7 Express pLysS Competent *E. coli* cells with the GeneJet plasmid mini prep kit (Thermo Fisher Scientific, USA). The plasmids were then sent to Inqaba Biotechnical Industries (RSA) for Sanger sequencing. The resultant gene sequence was translated into an amino acid sequence using the translate tool in ExPASy. The resultant amino acid sequence (Query) was aligned to the codon optimized protein sequence (Reference) with the Emboss Needle tool which uses the Needleman-Wunsch algorithm (and visualized with Jalview v 2 (***Figure 13***). The sequences have 100% alignment indicating that the TBR1 T-box domain did not undergo any mutations.




***Figure 13. The results from Sanger sequencing shown as a protein sequence alignment.*** *The gene sequence from sequencing was translated using the translate tool in ExPASy. The sequence obtained from Sanger sequencing (Query) was aligned to the codon optimized sequence (Reference). The asterisk indicates identical residues. There is a 100% alignment between the sequences indicating that the gene that was inserted was the same as the gene that was isolated. The sequences were aligned and visualized using Clustal Omega.*

### 5.1.2. Protein purification and SDS-PAGE

#### *5.1.2.1. Immobilized metal-ion affinity chromatography*

The TBR1 T-box domain was overexpressed for 24 hours at 20°C, after induction with 0.2 mM IPTG. The protein was overexpressed with an N-terminal His-tag, allowing it to be purified from the soluble lysate by IMAC. The resultant IMAC elution profile and SDS-PAGE gel is shown in ***Figure 14***.

The soluble lysate was loaded onto an IMAC column resulting in the flow through of any unbound contaminant proteins. The flow through peak was high due to the presence of a large number of proteins naturally expressed by the *E. coli* cells, as shown by the bands in the flow through (FT) lane of the gel. The column was then washed with detergent and a high salt concentration resulting in the removal of any non-specifically bound contaminant proteins and/or DNA. The salt wash peak was relatively small, indicating a low concentration of such contaminants. These did not show up on the gel (SW) indicating that they were of a relatively low concentration, or that the contaminants were DNA. The low levels of non-specifically bound proteins could be accounted for by the high specificity with which His-tags bind $Co^{2+}$. The column-bound proteins were then eluted with a high concentration of imidazole,

resulting in a relatively large elution peak. The tailing in the elution peak was the result of uneven levels of retention in the column, at different points during the elution. The eluted fractions were pooled together so the tailing was insignificant in any case. The TBR1 T-box domain was not completely pure after IMAC as the eluted fraction contained a few contaminants, as seen by the bands in the elution lane (E) of the gel. The molecular weight of the protein was ≈26 kDa, in accordance with its predicted molecular weight. The protein was thus further purified by SEC.

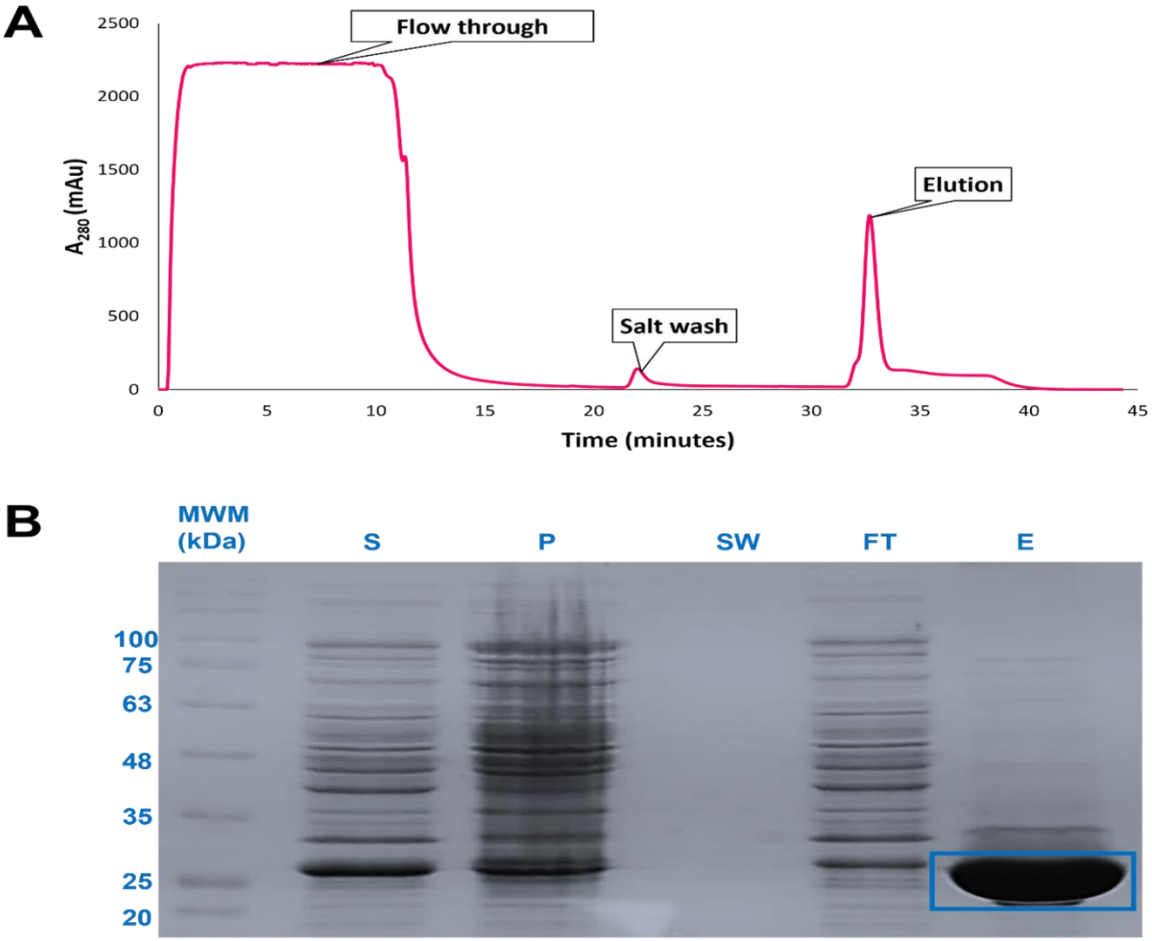



***Figure 14. Immobilized metal-ion affinity chromatography purification of the TBR1 T-box domain.*** *(A) Elution profile from immobilized metal-ion affinity chromatography. (B) SDS-PAGE gel with the samples collected from immobilized metal-ion affinity chromatography. S: supernatant. P: pellet. SW: salt wash. FT: flow through. E: elution. The protein of interest is shown in the blue box. The soluble lysate was loaded onto the column. The flow through contained many unbound contaminant proteins, as can be seen by the large number of bands in the flow through lane of the gel. Non-specifically bound contaminants were then removed by a salt wash containing detergent and a high salt concentration. These contaminants could not be detected by SDS-PAGE as indicated by the absence of any bands in the salt wash lane of the gel. Column-bound proteins were then eluted with a high concentration of imidazole. The elution was successful as shown by the bands in the elution lane of the gel. The presence of many bands in the elution lane indicate that the protein wasn't completely pure after IMAC and required subsequent purification via SEC.*

#### 5.1.2.2. Size exclusion chromatography

The TBR1 T-box domain wasn't completely pure after IMAC, and was therefore further purified by SEC. In order to separate the monomeric species from the dimeric species, SEC was carried out under reducing conditions (2 mM DTT). The resultant elution profile is shown in ***Figure 15 (A)***. The elution profile shows two distinct peaks. The molecular weight corresponding to each peak was determined by linear interpolation of a standard curve constructed from a set of standards of known molecular weight (Low-range gel filtration calibration kit (GE Healthcare, USA)). The standard curve is shown in ***Figure 15 (B)***. The first peak in the elution profile corresponded to a size of 50.01 kDa and the second peak corresponded to a size of 25.75 kDa. The molecular weight of the TBR1 T-box domain was predicted to be 25.6 kDa by the ExPASy ProtParam tool (Gasteiger et al., 2005). The first peak thus represented a dimeric species while the second peak represented a monomeric species.

The samples collected from SEC were quantitatively and qualitatively assessed using SDS-PAGE. The resultant gel is depicted in ***Figure 15 (C)***. The contamination lane (C) was empty, indicative of a relatively low concentration of contaminants in the elution profile. The dimer lane (D) contained one intense band, corresponding to the predicted size of the protein. This suggested that the first peak in the elution profile corresponded to a dimer since it was ≈50 kDa. This was unexpected since the experiment was carried out in an excess of DTT which was supposed to reduce the disulphide linked dimers. The 'dimer peak' in the elution profile can therefore be explained by the presence of monomer-bound contamination corresponding to the size of the dimer. Another possible explanation for the presence of the dimer peak is that the disulphide bonds were incompletely reduced. It is also possible that the dimers are not all disulphide linked and are instead formed by concentration-driven associations governed by non-covalent interactions. The monomer lane (M) contained one single intense band corresponding to the predicted size of the protein, indicating that the sample was completely pure and thus ready for downstream applications. The purity of the protein was determined to be >98% by densitometric analysis (results not shown). The molecular weight corresponding to the band of interest was determined by linear interpolation of a standard curve constructed from a set of standards of known molecular weight (BLUeye Prestained protein Ladder (Merck, Germany). The standard curve is shown in ***Figure 15 (D)***. The band of interest corresponds to a molecular weight of 26 kDa. The molecular weights determined

from SEC and SDS-PAGE were approximately the same and correlated well with the size predicted by the ExPASy ProtParam tool.

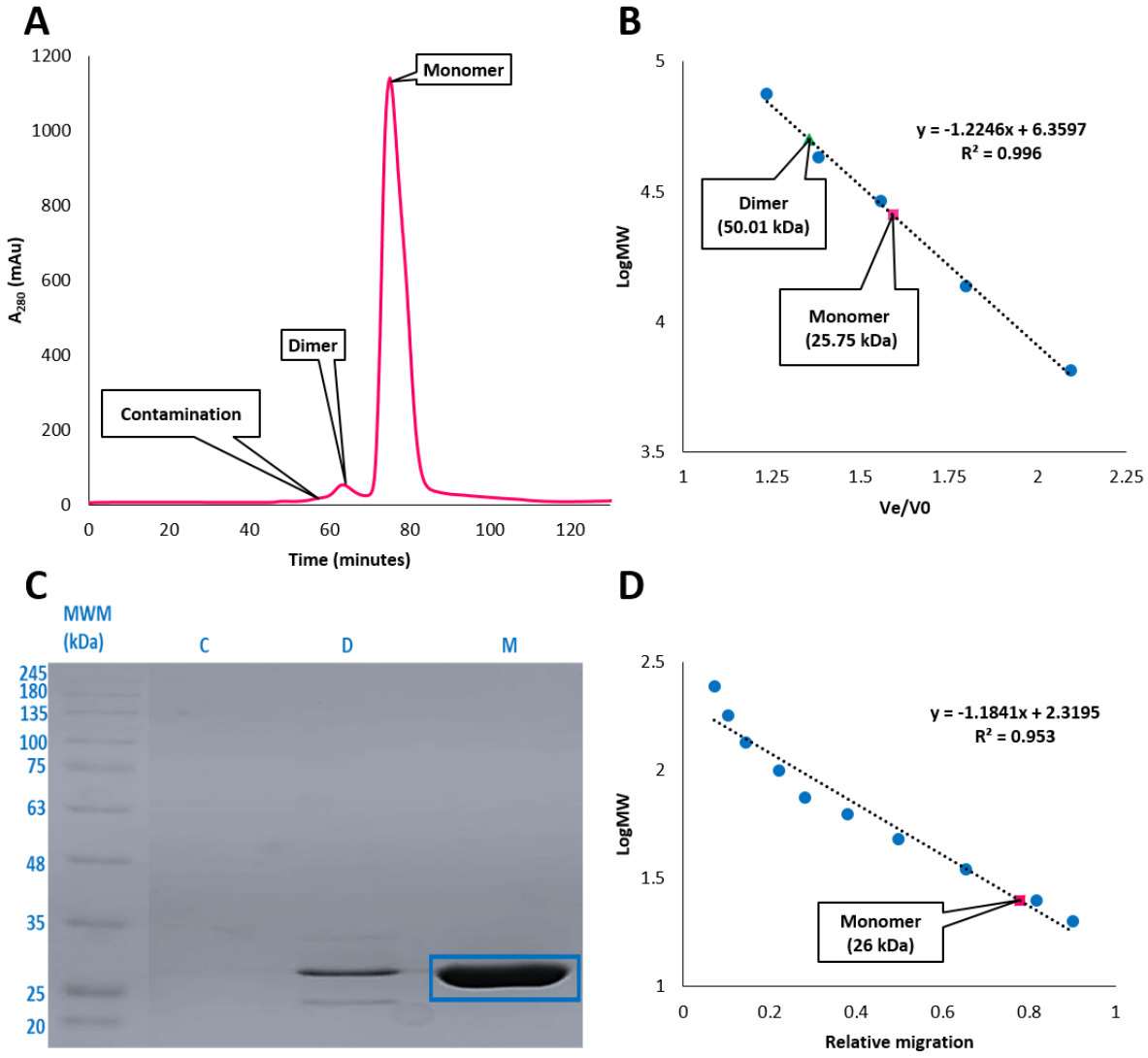



***Figure 15. Size exclusion chromatography purification of the TBR1 T-box domain.*** *The size exclusion chromatography elution profile (A) of the TBR1 T-box domain shows 2 distinct peaks. The elution time for each peak was used to determine the MW of the protein by linear interpolation of a standard curve (B) constructed from standards of known molecular weight (Low-range gel filtration calibration kit (GE Healthcare, USA)), where Ve is the elution volume and V0 is the void volume of the column. The elution time of the first peak indicates a molecular weight of 50.01 kDa, while the elution time of the second peak indicates a molecular weight of 25.75 kDa. This corresponds to the predicted size of the dimer and monomer, respectively. Samples from size exclusion chromatography were quantitatively assessed using SDS-PAGE. C: contamination, D: dimer and M: monomer. The size of the unknown band in the blue box was determined by linear interpolation of a standard curve (D) constructed from standards of known molecular weights (BLUeye Prestained protein Ladder (Merck, Germany)). The dimer lane has one intense band and two lighter bands. The intense band corresponds to the predicted molecular weight of the protein. The monomer lane has one single intense band, corresponding to the predicted size of the protein. This*

*indicates that the protein was successfully and completely purified. The molecular weight determined from size exclusion chromatography correlates well with the molecular weight determined from SDS-PAGE.*

### 5.1.3. Protein purity and concentration determination

Absorbance spectroscopy was used to determine the concentration of the protein, and to determine if there was any nucleic acid contamination. The resultant absorbance spectrum shown below in ***Figure 16*** is characteristic of a protein absorbance spectrum. There is a peak at 280 nm due to the presence of the aromatic amino acids like tyrosine, tryptophan and phenylalanine which absorb at 280 nm. The absorbance at 280 nm was used to calculate the protein concentration of 61.71 µM using an extinction coefficient of 36 565 $M^{-1}.cm^{-1}$. Even though the absorbance value was greater than 1, a dilution series was used to ensure that the linearity was maintained even at a relatively high protein concentration (data not shown). On average, ~28.8 mg of protein was obtained per litre of culture. There is a shoulder at 295 nm due to the presence of five tryptophan residues which exclusively absorb UV light at 295 nm. The absence of absorbance at 340 nm indicates that the protein is free from aggregation. Nucleic acids absorb UV light at 260 nm. The $A_{280}/A_{260}$ ratio is 1.79 indicating that the protein is sufficiently free from nucleic acid contamination.

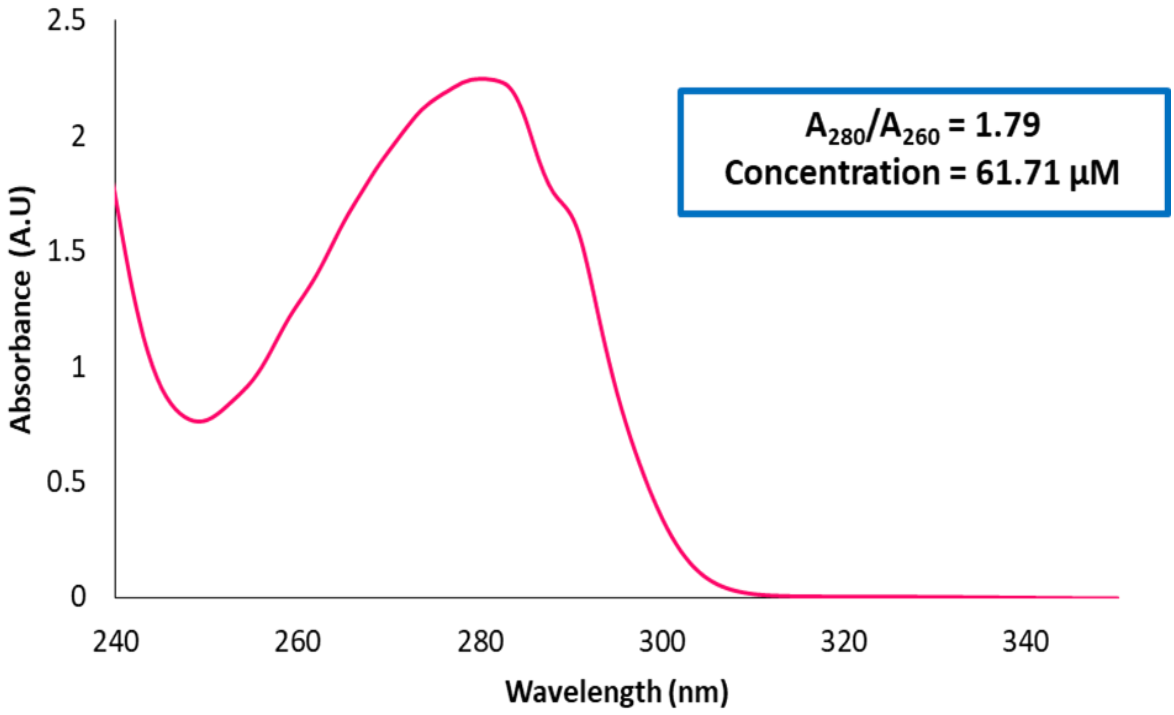



***Figure 16. Absorbance spectrum of the TBR1 T-box domain.*** *The spectrum has a peak at 280 nm due to the presence of aromatic amino acids, and a shoulder at 295 nm due to the presence of five tryptophan residues. A lack of absorbance at 340 nm indicates that the protein is free from any detectable aggregation. The parameters obtained from the absorbance*

*spectrum are shown in the blue box. The absorbance at 280 nm was used to calculate the protein concentration using an extinction coefficient of 36565 $M^{-1}.cm^{-1}$. The protein concentration was 61.71 µM. The A280/A260 ratio is 1.79 indicating that the protein is free from nucleic acid contamination.*

## 5.2. Characterization of protein structure and stability

After obtaining sufficient quantities of pure homogenous protein, the structure and thermal stability of the protein were characterized in the presence and absence of SSL DNA. ITF spectroscopy was successfully used to show that the protein had been properly folded and that there are no differences in the tertiary structure of the protein upon DNA-binding. Far UV CD spectropolarimetry was successfully used to show that the protein had a predominantly β-sheeted secondary structure, and that there are no differences in the secondary structure of the protein upon DNA-binding. It was also used to show that thermal stability of the protein remains unchanged upon DNA-binding.

### 5.2.1. Intrinsic tryptophan fluorescence spectroscopy

The tertiary structure of the TBR1 T-box domain was studied by intrinsic tryptophan fluorescence spectroscopy. Tryptophan residues contain a solvatochromic indole fluorophore which can be exclusively excited at 295 nm. When tryptophan residues are in a completely polar environment, such as when they are exposed to solvent, they will emit light at 350 nm (Lakowicz, 2006). However, when they are in a completely hydrophobic environment, such as the inside of the protein's core, they will emit light at 320 nm (Lakowicz, 2006). The indole fluorophore is bulky and non-polar resulting in it being buried in the hydrophobic core of a folded protein. These qualities make tryptophan the ideal probe to easily study tertiary structure by spectroscopy. Additionally, the TBR1 T-box domain has five tryptophan residues which are well spread throughout the protein, making ITF spectroscopy a good probe for global structure. To confirm that the protein had been properly folded, the ITF spectra (***Figure 17***) were determined in the presence and absence of a denaturant (8 M urea), after excitation at 295 nm. The ITF emission spectrum of the native protein (in the absence of 8 M urea) shows a peak at 330 nm and a shoulder at 340 nm, presumably due to the tryptophan residues being in two distinct environments, with different degrees of solvent exposure. The ITF emission spectrum of the unfolded protein (in the presence of 8 M urea) shows a single peak at 350 nm suggesting that the tryptophan residues are all in the same environment. In the presence of urea, the emission maximum is thus shifted by 10 nm with respect to the native protein. This bathochromic or red shift is evidence that the tryptophan residues in the native protein

were in a hydrophobic environment, presumably the interior of the protein, suggesting that the protein was probably correctly folded. The T-box crystal structures show that the tryptophan residues are well buried in the core of the protein (Coll et al., 2002; El Omari et al., 2012; Liu et al., 2016; Müller and Herrmann, 1997; Stirnimann et al., 2010). If the protein was unfolded, the tryptophan residues in the native protein would have been in a polar environment and the bathochromic shift observed in the presence of urea would not occur. There is a significant decrease in intensity (quenching) from the native to the unfolded proteins. When the protein is unfolded, its tryptophan residues are solvent exposed. As a result, some of the light emitted by the solvent exposed indole fluorophores is absorbed by surrounding water molecules. The bathochromic shift and quenching that occurred upon denaturation was evidence that the TBR1 T-box domain was correctly folded during overexpression. This meant that the protein could be used for downstream applications.

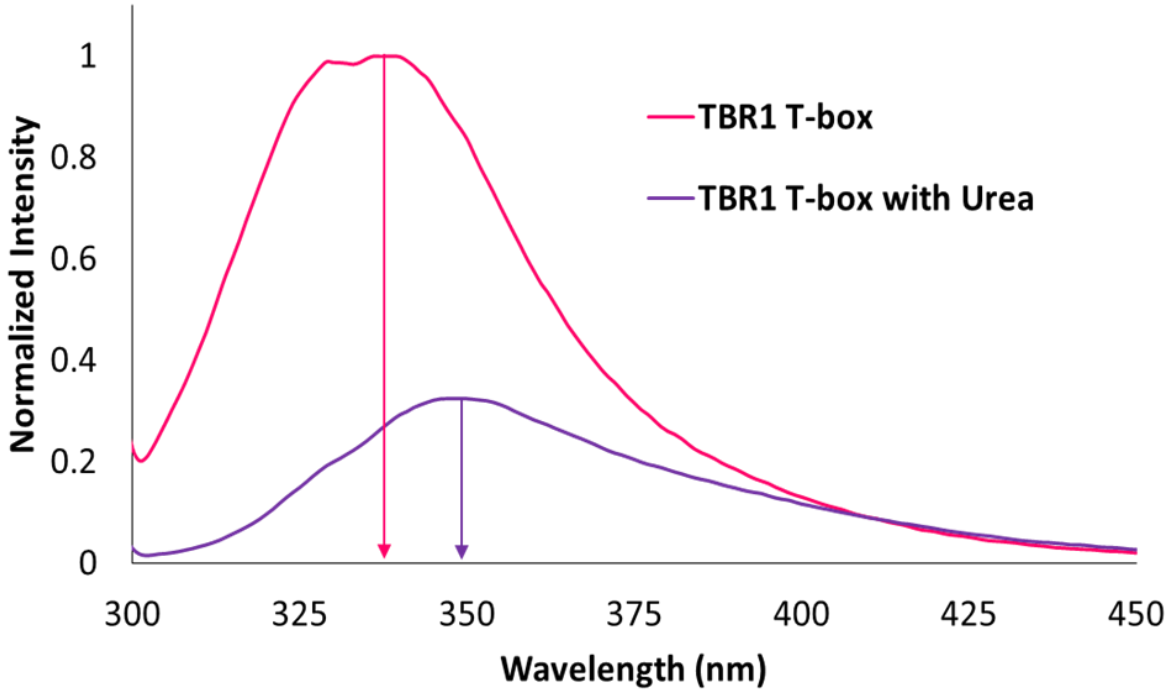



***Figure 17. Intrinsic tryptophan fluorescence spectra of the TBR1 T-box domain in the presence and absence of a denaturant.*** *The spectra were obtained after exciting the tryptophan residues at 295 nm. The protein (2 µM) was denatured with 8M urea. The spectrum of the denatured protein was normalized against the spectrum of the wild type. The native spectrum has a peak at 330 nm and a shoulder at 340 nm suggesting that the tryptophan residues are in two different environments. The denatured spectrum has a single peak at 350 nm. The bathochromic shift observed from the native to the unfolded proteins is evidence that the native protein was properly folded. The quenching observed upon denaturation is further evidence that the native protein was properly folded.*

To determine whether DNA-binding had any effect on the tertiary structure of the protein, the ITF spectra were determined in the presence and absence of SSL DNA, as shown in ***Figure 18***. There is a significant amount of quenching observed in the presence of SSL DNA, based on the results of a two-sample t-test. DNA can quench intrinsic tryptophan fluorescence by a mechanism known as photoinduced electron transfer (Lakowicz, 2006c). In this distance-dependant process, the guanine bases in DNA donate electrons to tryptophan residues (Lakowicz, 2006c). This suggested that the TBR1 T-box domain was successfully bound to SSL DNA. In the absence of SSL DNA, the ITF emission spectrum of the protein shows a peak at 330 nm and a shoulder at 340 nm, as mentioned above. Once again, this is presumably due to the tryptophan residues being in two distinct environments, with varying degrees of solvent exposure. In the presence of SSL DNA, the ITF emission spectrum of the protein shows a peak at 332 nm and a shoulder at 342 nm. This suggests that, even in the presence of SSL DNA, the tryptophan residues are still in two distinct environments. The bathochromic shift of 2 nm observed upon DNA binding is insignificant, based on the results of a two-sample t-test. The data thus suggests that the tertiary structure of the protein is conserved upon DNA-binding. The data also suggests that tryptophan residues are in close proximity to the DNA even though their tertiary structure is conserved. The data is consistent with the TBX5 crystal structures which also suggest that the tertiary structure is conserved upon DNA-binding (Stirnimann et al., 2010).

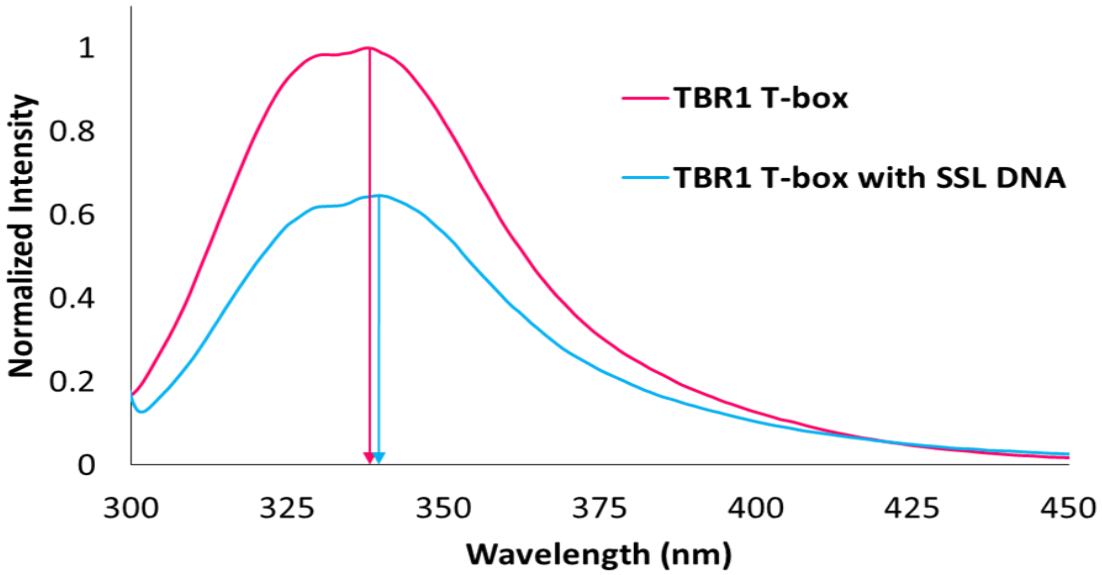



***Figure 18. Intrinsic tryptophan fluorescence spectra of the TBR1 T-box domain in the presence and absence of SSL DNA.*** *The spectra were obtained after exciting the tryptophan residues at 295 nm. There is significant quenching in the presence of*

*DNA, suggesting that the protein (2 µM) was successfully bound to the DNA (1 µM). The spectrum in the absence of SSL DNA has a peak at 330 nm and a shoulder at 340 nm because the tryptophan residues are in two distinct environments. The spectrum in the presence of SSL DNA has a peak at 332 nm and a shoulder at 342 nm. The 2 nm shift observed upon DNA-binding is insignificant. DNA-binding thus has no effect on the tertiary structure of the protein.*

### 5.2.2. Circular dichroism spectropolarimetry

T-box proteins, like the TBR1 T-box domain, are predominantly β-sheeted because the core of the protein contains a seven-stranded β-barrel. Circular dichroism spectropolarimetry was used to study the secondary structure of the TBR1 T-box domain. The far UV CD spectrum of a typical β-sheeted protein contains a trough at around 218 nm and a peak at around 195 nm (Greenfield, 2006; Woody and Kozlowski, 2002). To determine whether DNA-binding had any effect on the secondary structure of the protein, the the far UV CD spectra were determined in the presence and absence of SSL DNA. The spectra, shown in ***Figure 19 (A)***, contain a trough at 213 nm. This indicates that the protein is predominantly β-sheeted, both in the absence and presence of DNA. Previous findings suggest that T-box proteins are β-sheeted and that the secondary structure of the protein remains unchanged upon DNA binding (Stirnimann et al., 2010). The data thus correlates well with previous findings. The data could not be reliably dissected into the proportions of secondary structural elements because the data from 180 nm to 200 nm was unusable due to high levels of noise.

Chloride ions absorb light in the far UV range of wavelengths used to study the secondary structure of proteins by far UV circular dichroism. The inclusion of chloride ions therefore results in a very noisy spectrum from which relevant data cannot be confidently deduced. However, chloride ions had to be included in the buffer to maintain the solubility of the protein. The protein was thus diluted to its final concentration using Milli-Q™ water such that the final NaCl concentration was only 20 mM. This slowed for an acceptable CD signal while keeping noise levels down to an acceptable minimum. The high tension (HT) voltage is used as a quantitative measure of noise and should be kept below 600 V for the CD data to be meaningful. This measure has been included, shown in ***Figure 19 (B)***, to prove that noise levels were kept well below 600 V even though chloride ions were present in the buffer. The data that obtained is thus acceptable. The data obtained from 180 nm to 200 nm was removed as the HT voltage exceeded 600 V. The characteristic peak of β-sheeted proteins, at 195 nm, was thus not visible. However, the trajectory of the data suggests that this would have been the case if chloride ions were omitted from the buffer.

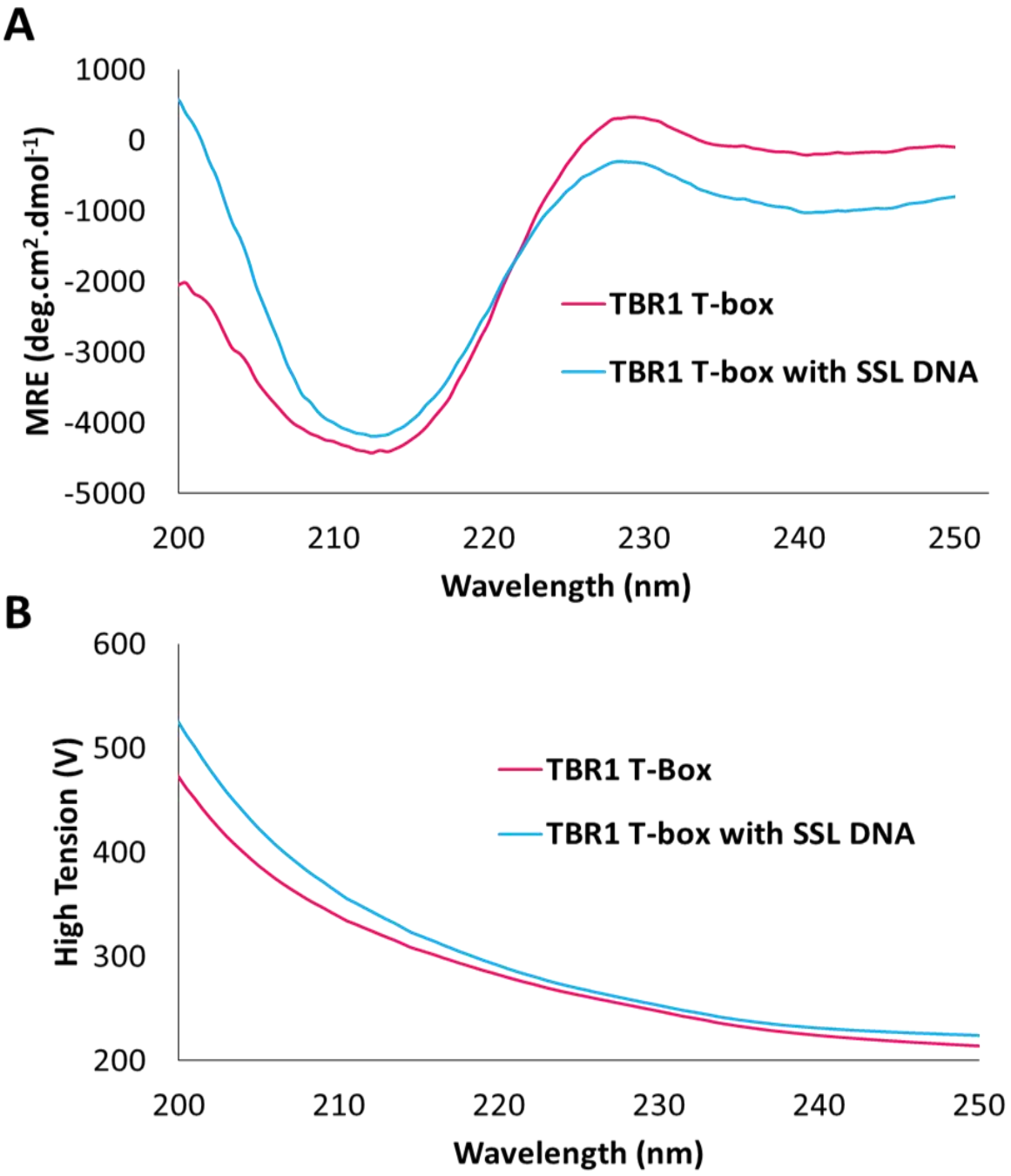



***Figure 19. Circular dichroism spectra of the TBR1 T-box domain in the presence and absence of SSL DNA.*** *The far UV circular dichroism spectra (A) were obtained by normalizing the circular dichroism signal (millidegrees) to mean residue ellipticity (MRE), which accounts for the protein concentration and the number of amino acids present. The spectra, both in the absence and presence of DNA, contain a trough at 213 nm. The spectra indicate that the protein is predominantly β-sheeted as expected. The secondary structure thus does not change upon DNA-binding. The differences in the MRE are due to slight differences in protein concentration which were amplified when the sample was diluted prior to analysis. The high tension (HT) voltage spectra (B) have been included to show that noise levels were kept to an acceptable minimum even though chloride ions were present in the buffer. The voltage did not exceed 550 V in the wavelength range indicated both in the absence and presence of DNA.*

### 5.2.3. Thermal stability

The thermal stability of a protein is defined as its ability to resist changes in structure as the temperature is increased. Salt bridges, hydrogen bonding, Van de Waals interactions and disulphide bonds all contribute to the thermal stability of a protein (Pace et al., 2014). By studying the thermal stability of a TF in the presence and absence of DNA, the effect that DNA has on the interactions that stabilize the structure of the protein can be assessed. An increase in thermal stability upon DNA-binding may suggest that helix $3_{10}$C becomes structured upon DNA-binding. The thermal stability of the TBR1 T-box domain was studied, in the presence and absence of SSL DNA, by circular dichroism spectropolarimetry. The loss in the far UV signal at 213 nm was monitored as the temperature was increased from 20 to 90 °C. This was done to determine if DNA-binding had any significant effect on the thermal stability of the protein and also to get an idea of how it could affect crystallization. The resultant raw data is shown below in ***Figure 20 (A)***. The raw data was converted to represent the fraction of unfolded protein at each temperature, and then fit to a sigmoidal dose-response curve using the nonlinear regression tool in GraphPad Prism v 9. The fitted data is shown below in ***Figure 20 (B)***. The thermodynamic parameters of equilibrium unfolding were unobtainable due to the irreversibility of the thermal unfolding.

The TBR1 T-box domain shows a two-state transition from the native (folded) to the denatured (unfolded) state, both in the absence and presence of SSL DNA. The inflection points were used to determine the denaturation midpoints ($T_m$) of the protein in the absence and presence of SSL DNA. In the absence of SSL DNA, the TBR1 T-box domain begins to unfold at ≈55 °C and has a $T_m$ of 63.5 °C. The $T_m$ is consistent with previous findings (Stirnimann et al., 2010). The protein is fully unfolded at ≈75 °C. In the presence of SSL DNA, the TBR1 T-box domain begins to unfold at ≈58 °C and has a $T_m$ of 63.3 °C. The protein is fully unfolded at ≈70 °C. The $T_m$ is thus not significantly affected by DNA-binding. The differences in signal that are observed in the unfolding curves are a result of small differences in protein concentrations which were amplified when the sample was diluted with Milli-Q™ water. The thermal stability of the protein is not affected by DNA-binding. The data is consistent with the findings from ITF spectroscopy and far UV CD which suggested that the global protein structure is conserved upon binding to SSL DNA.

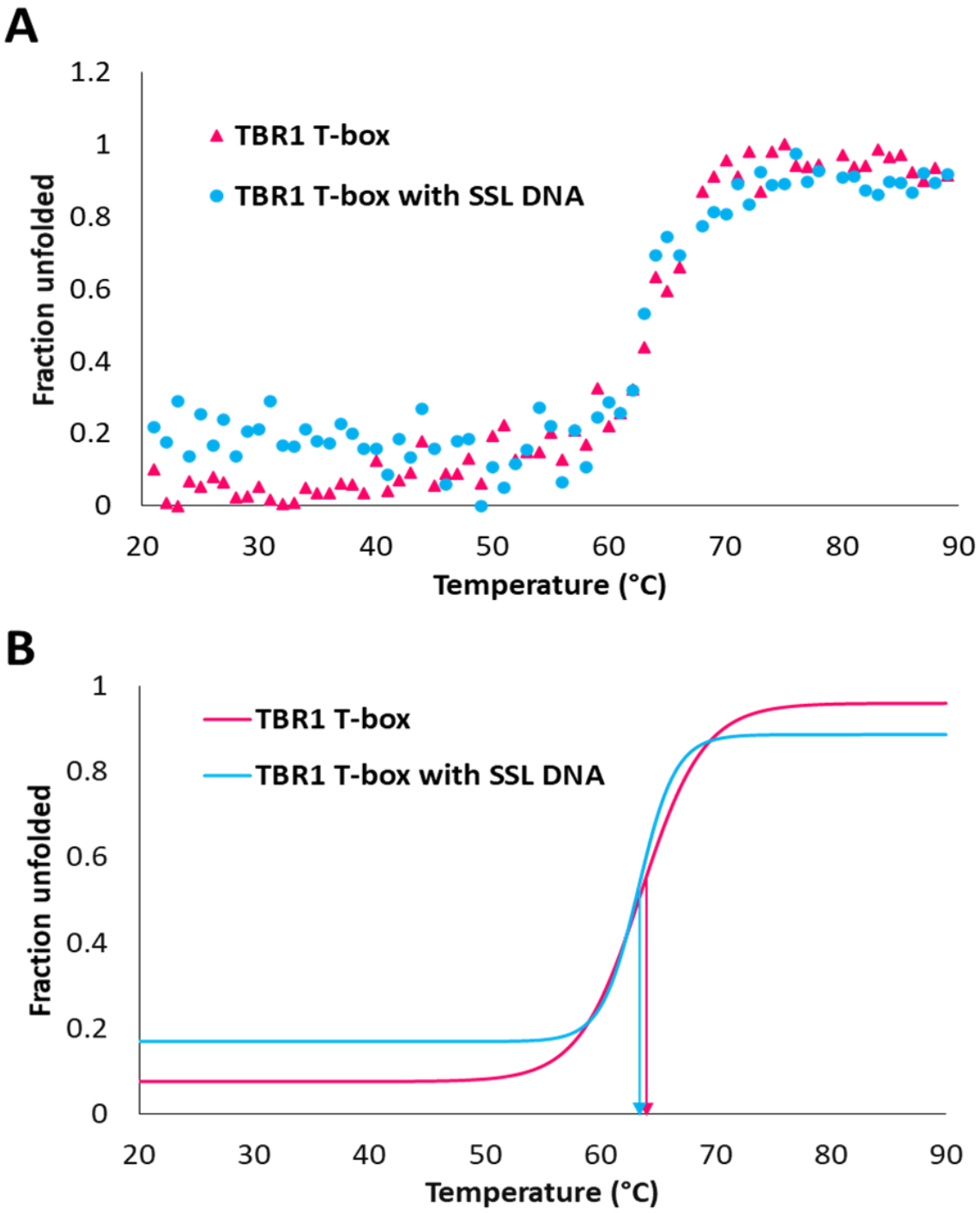



***Figure 20. Thermal unfolding curves of the TBR1 T-box domain in the presence and absence of SSL DNA.*** *The thermal unfolding was probed by monitoring the loss of the far UV circular dichroism signal at 213 nm as the protein was heated from 20 to 90 °C. The raw data was converted to represent the fraction of unfolded protein at each temperature, and then fit to a sigmoidal dose-response curve using the nonlinear regression tool in GraphPad Prism v 9. In the absence of SSL DNA, the TBR1 T-box domain begins to unfold at ≈55 °C and has a $T_m$ of 63.5 °C (pink arrow). The protein is fully unfolded at ≈75 °C. In the presence of SSL DNA, the TBR1 T-box domain begins to unfold at ≈58 °C and has a $T_m$ of 63.3 °C (blue arrow). The protein is full unfolded at ≈70 °C. The thermal stability of the protein is thus not affected by DNA-binding.*

## 5.3. DNA-binding studies

In order for a transcription factor to regulate gene expression at the transcriptional level, it should be able to bind DNA in a sequence specific manner. The structural studies showed that there were no significant structural changes in the TBR1 T-box domain that occurred upon binding to SSL DNA. The DNA-binding function was confirmed by an EMSA which also showed that saturated DNA-binding is achieved when the protein is 10x in molar excess of the DNA. The binding affinity was then quantified in the presence and absence of $MgCl_2$ by fluorescence anisotropy. The dissociation constants ($K_D$s) were detected in the nanomolar range and there was no statistically significant difference between the $K_D$s in the presence and absence of $MgCl_2$.

### 5.3.1. Electrophoretic mobility shift assay

An EMSA was used to confirm the interaction between the TBR1 T-box domain and SSL DNA, even though this had already been confirmed by ITF spectroscopy. It would also give an idea of the protein concentration required for saturated DNA-binding. This information was required for downstream applications such as fluorescence anisotropy and crystal trials. The resultant EMSA is shown below in ***Figure 21***. The band shift with respect to the free DNA, observed in lanes 2 to 6, is proof that the TBR1 T-box domain binds SSL DNA. A negative control should have been included to show that the protein binds the TBE specifically, such as a randomised DNA-sequence of the same bases with the same GC content. The degree of binding increases with protein concentration as expected. Saturated DNA-binding has been achieved by lane 5, where protein was ten times the molar concentration of the DNA, since it does not contain a band corresponding to free DNA. The absence of another band above the protein-DNA complex bands suggests that the protein binds the DNA as a monomer.

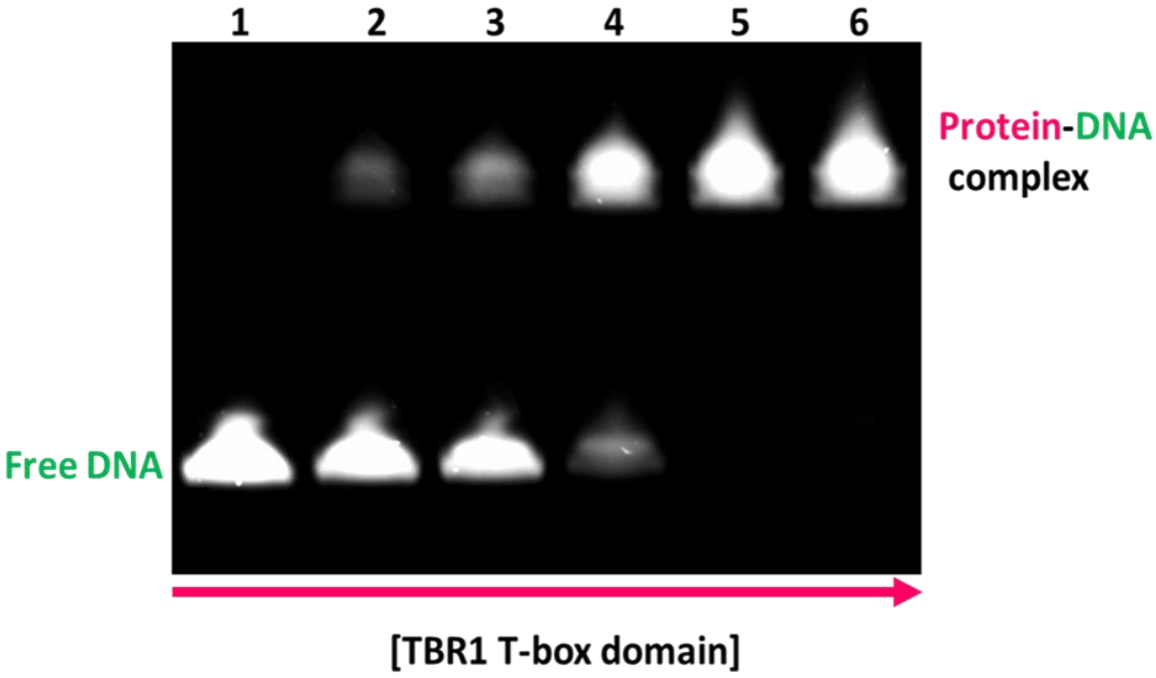


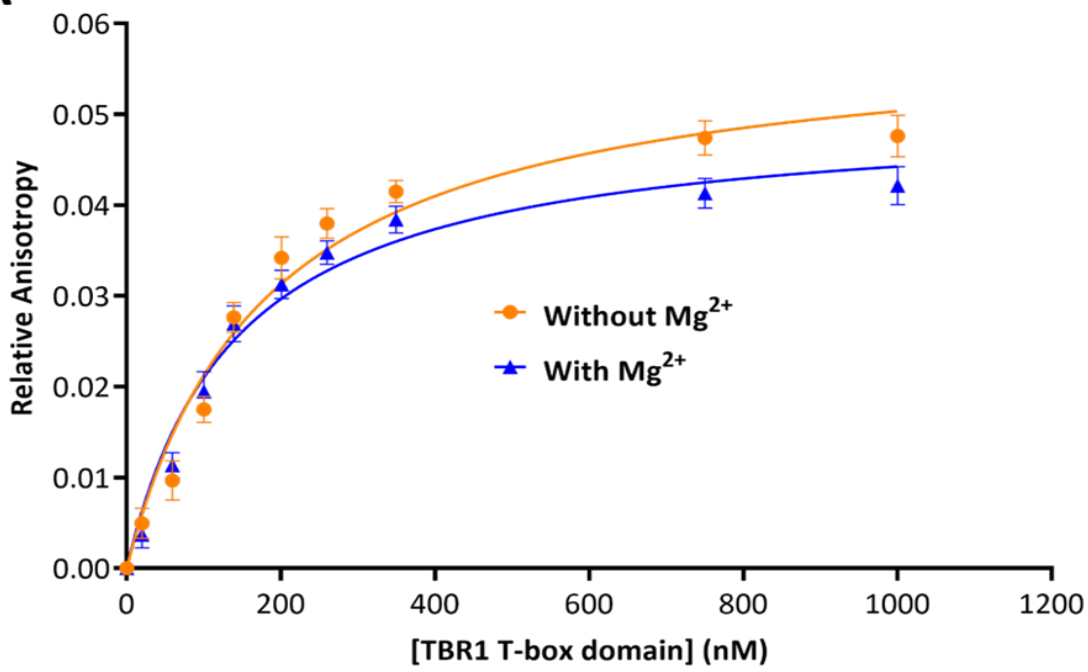

***Figure 21. Electrophoretic mobility shift assay showing the TBR1 T-box domain bound to SSL DNA.*** *Samples were prepared with 0.5 µM SSL DNA and increasing concentrations of the TBR1 T-box domain (0 – 10 µM). Lane 1 contains free SSL DNA. The molar ratios of DNA:Protein used were as follows: Lane 2 – 3:1, Lane 3 – 1:1, Lane 4 – 1:5, Lane 5 - 1:10 and Lane 6 – 1:20. The samples were run on an 8% continuous native polyacrylamide gel, at 150 V for 1.5 hours. Lower bands represent free DNA and upper bands represent the protein-DNA complex. The shift observed, with respect to the free DNA, is proof that the TBR1 T-box domain binds SSL DNA. Saturated DNA-binding is achieved by lane 5 where the protein is ten times more concentrated than the DNA.*

### 5.3.2. Fluorescence anisotropy

Fluorescence anisotropy assays were carried out to determine the dissociation constant ($K_D$) of the TBR1 T-box domain binding to SSL DNA, in the presence and absence of $MgCl_2$. This was done to determine the affinity that the TBR1 T-box domain has for SSL DNA, as well as to determine if the presence of $MgCl_2$ could affect this affinity. ROX-labelled SSL DNA (200 nM) was titrated with a relatively high concentration of the TBR1 T-box domain (≈11 µM), either in the presence or absence of 100 µM $MgCl_2$. The data points were normalized relative to the anisotropy of free ROX-labelled SSL DNA (either in the presence or absence of 100 µM $MgCl_2$). The data was then fitted to a single-site saturation binding model in GraphPad Prism v 8. The isotherms are shown in ***Figure 22 (A).*** The correlation coefficients and parameters obtained from the fit are displayed in a table in ***Figure 22 (B)***. The correlation coefficients suggest that the model is a good fit for the data.

In the absence of $MgCl_2$, the $K_D$ was 179.6 nM and the $B_{max}$ was 0.0594. In the presence of 100 µM $MgCl_2$, the $K_D$ was 140.3 nM and the $B_{max}$ was 0.0504. There was no statistically significant difference between the $K_D$s, based on a two-sample t-test. In both cases, the $K_D$ is in the nanomolar range, suggesting that the TBR1 T-box domain binds SSL DNA tightly and with high affinity. The difference in the $B_{max}$ values is probably due to slight differences in the concentration of the TBR1 T-box domain used.

**A**




**B**

| $[Mg^{2+}]$ (µM) | $R^2$ | $K_D$ (nM) | $B_{max}$ |
|---|---|---|---|
| 0 | 0.9674 | 179.6 | 0.0594 |
| 100 | 0.9723 | 140.3 | 0.0504 |

***Figure 22. Fluorescence anisotropy DNA-binding assay of the TBR1 T-box domain and SSL DNA with and without $Mg^{2+}$.*** *(A) DNA-binding isotherms of TBR1 T-box domain binding to SSL DNA in the presence and absence of $Mg^{2+}$. In both cases, the data was fitted to a single-site saturation binding model using the non-linear regression tool in GraphPad prism v 8. The experiments were performed in triplicate and the error bars represent the standard error of the mean. (B) Table showing the correlation coefficient for the data fitting, as well as the parameters obtained from the fit. In both cases, the correlation coefficient was >0.95 indicating a good fit. In the absence of $Mg^{2+}$, the $K_D$ was 179.6 and the $B_{max}$ was 0.0594. In the presence of 100 µM $Mg^{2+}$, the $K_D$ was 140.3 and the $B_{max}$ was 0.0504. There was no statistically significant difference between the $K_D$s. In both cases, the $K_D$ is in the nanomolar range, suggesting that the TBR1 T-box domain binds SSL DNA tightly with high affinity.*

## 5.4. Protein crystallography

XRC was carried out to determine the crystal structures of the TBR1 T-box domain in the presence and absence of SSL DNA. The atomic resolution models would have allowed for the determination of the DNA-binding mechanism utilized. Despite innumerable attempts, the protein could not be crystallized in the absence of the DNA.

Crystal trials were then attempted in the presence of SSL DNA with the hope of attaining a protein crystal containing DNA-bound and unbound structures in the same asymmetric unit. After dialyzing the protein against Buffer C, a shower of microcrystals was obtained (***Figure 23 (A)***). The SSL DNA was mixed with the protein at a final concentration of 0.54 mM DNA duplex and 0.45 mM protein (the DNA was in 1.2 molar excess of the protein). The sample was incubated at 20 °C for 1 hour to allow for the DNA-binding reaction to reach equilibrium. Very tiny microcrystals appeared after 1 week at 20 °C in sitting drops containing 1.5 µL of DNA/protein solution and 1.5 µL of crystal solution 3 (100 mM 4-morpholineethanesulfonic acid pH 6.5, 200 mM $MgCl_2$ and 25% (w/v) PEG 4 000), equilibrated against 250 µL of crystal solution 3. The microcrystals were too tiny to manipulate (< 0.01 mm) and would definitely not have diffracted significantly. The crystallization conditions were optimized with the hope of getting fewer (and inevitably larger) crystals. The phase diagram in ***Figure 23 (B)*** was used to understand why microcrystals formed and how to go about preventing excessive nucleation (Rupp, 2013). The protein and/or precipitant concentration was too high, resulting in the sample being too supersaturated and ending up too deep into the labile zone. There were too many excessive nucleation events which depleted all the protein in the solution thus preventing crystal growth. In the next trial, the attention was thus focused on reducing the level of supersaturation such that the sample was not as deep into the labile zone. Another possibility was that the rate of crystal formation was too high, and by slowing down the process, the events governing nucleation would take longer, thus allowing more time for crystal growth to occur before depletion of the protein from the solution.

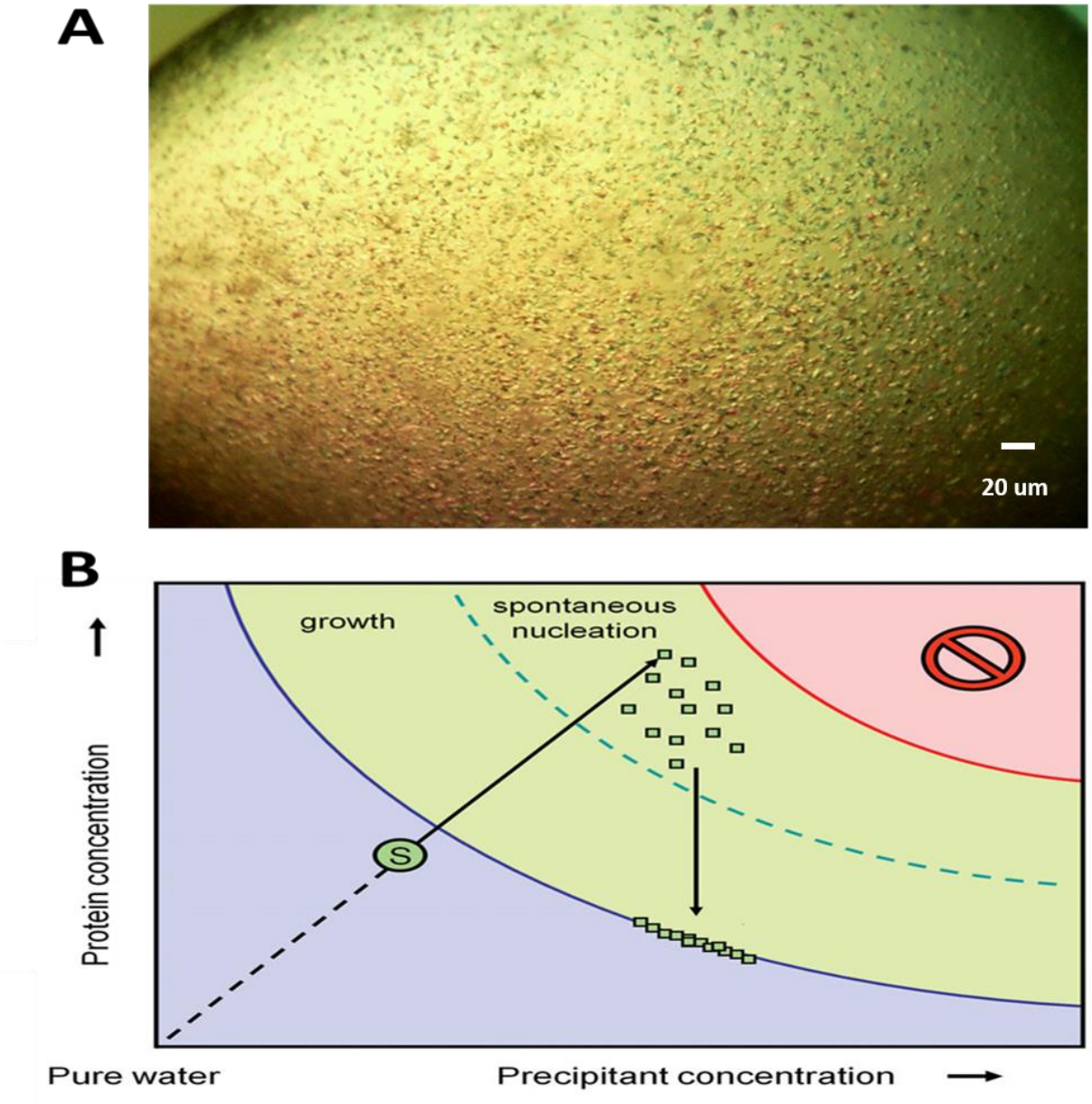



***Figure 23. A shower of tiny protein-DNA microcrystals resulting from unwanted excessive nucleation.*** *Crystal trials of the TBR1 T-box domain in the presence of SSL DNA was performed at 20 °C by using 0.54 mM SSL DNA duplex and 0.45 mM TBR1 T-box domain. A shower of microcrystals (A) appeared after 1 week in sitting drops containing 1.5 µL of DNA/protein solution and 1.5 µL of crystal solution 3 (100 mM 4-morpholineethanesulfonic acid pH 6.5, 200 mM $MgCl_2$ and 25% (w/v) PEG 4 000), equilibrated against 250 µL of crystal solution 3. Since the crystals were too tiny to manipulate or diffract (< 0.01 mm), the crystallization conditions were optimized with the hope of getting fewer (and inevitably larger) crystals. A phase diagram (B) has been used to understand why microcrystals formed and how to go about preventing excessive nucleation. It is clear from the phase diagram that the sample was too supersaturated and thus too deep into the labile zone which promoted excessive nucleation thereby depleting all the protein from the solution and inhibiting crystal growth. In our next trial, the focus was on reducing the supersaturation level such that the sample was closer to the lower limits of the labile zone, indicated by the dashed line in the phase diagram.*

There are a number of ways by which the level of supersaturation can be reduced. These include reducing the concentration of the protein and/or precipitant, reducing the temperature, adding glycerol to inhibit excessive nucleation, and increasing the ratio of protein to precipitant (McPherson and Cudney, 2014; McPherson and Gavira, 2014). The conditions were optimized because this was found to be to be the most influential parameter. After decreasing the protein concentration from 0.45 mM to 0.42 mM, fewer and bigger crystals were obtained, with some notable aggregation. The crystals have been shown in ***Figure 24*** below. Reducing the protein concentration allowed the sample to be closer to equilibrium, which allowed for fewer nucleation events and subsequent growth before equilibrium was attained. Even though these crystals were better, they were still too small (< 0.05 mm) to be used for diffraction. The conditions were further optimized, by using the same approaches mentioned above. The protein concentration was thus reduced to obtain even fewer and hopefully bigger crystals. This did not result in any crystallization, probably because supersaturation had not been attained.

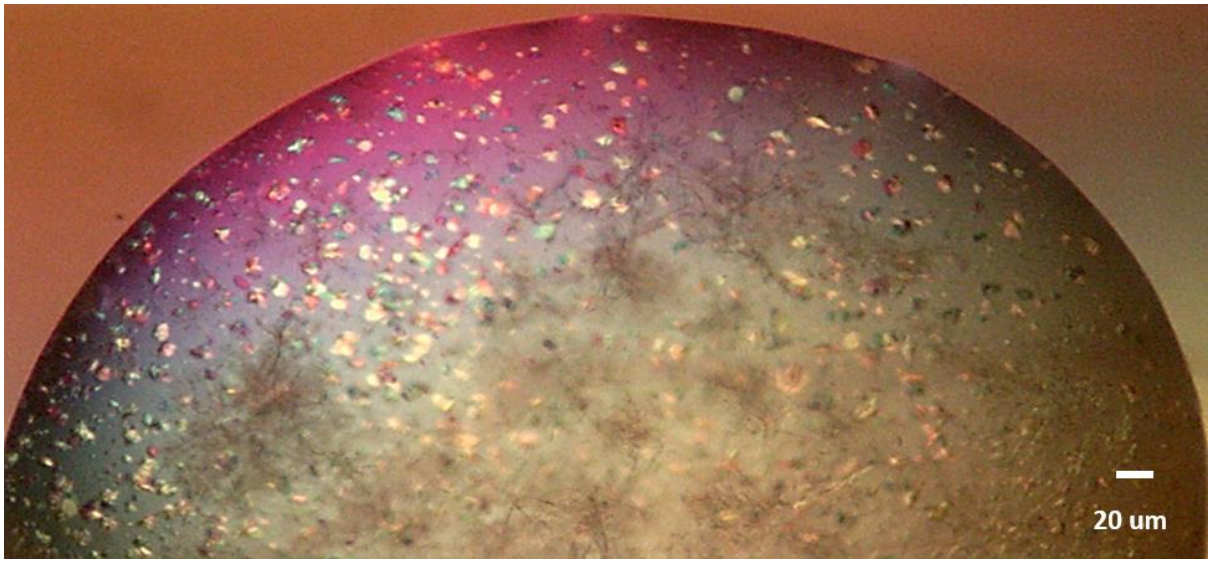



***Figure 24. A shower of larger protein-DNA microcrystals obtained by reducing the protein concentration.*** *Crystal trials of the TBR1 T-box domain in the presence of SSL DNA were performed at 20 °C by using 0.5 mM SSL DNA duplex and 0.42 mM TBR1 T-box domain. A shower of relatively larger microcrystals (A) appeared after 1 week in sitting drops containing 1.5 µL of DNA/protein solution and 1.5 µL of crystal solution 3 (100 mM 4-morpholineethanesulfonic acid pH 6.5, 200 mM $MgCl_2$ and 25% (w/v) PEG 4 000), equilibrated against 250 µL of crystal solution 3. Since the crystals were too tiny to manipulate or diffract (0.03 – 0.05 mm), the crystallization conditions were optimized with the hope of getting fewer and larger crystals, using the same methods that led to these crystals in the first place. The protein concentration was thus reduced. This did not result in any crystallization, probably because supersaturation had not been attained.*

It was evident from the first two crystal trials that a protein concentration of at least 0.42 mM was required to achieve supersaturation. At concentrations below this level, crystal formation did not occur and at concentrations above this level, there were too many excessive nucleation events occurring too quickly. The protein concentration could not be optimized any further. Optimization of the other important variables (precipitant concentration, the ratio of protein to DNA, and the drop ratio) did not yield any significant improvement in crystallization. After the first two trials, there were two important problems hampering crystallization. Firstly, excessive nucleation events were unavoidable. Secondly, the events governing crystal formation were occurring too quickly resulting in aggregation which depleted all the protein from the solution thereby inhibiting crystal growth. The first problem was to be solved through crystal seeding and the second problem through vapor diffusion rate control.

In crystal seeding, an ordered solid phase is used as a surface for the growth of crystals (Rupp, 2013). More simply put, it is the process of adding preformed nuclei to a crystallization solution in order to grow fewer (and inevitably larger) crystals. Seeding is said to be homogenous when the nuclei are made of the same molecules that are to be crystallized, such as in the case of macromolecular seeding. Seeding is said to be heterogeneous when the nuclei are not made of the same molecules that are to be crystallized, such as in the case of epitaxial growth and cross-seeding. The spontaneous formation of prenucleation aggregates is a kinetically demanding step when compared to crystal growth, and as such, molecules prefer to accumulate on a ready-made template (McPherson and Cudney, 2014; McPherson and Gavira, 2014). The crystallization conditions that support nucleation are not necessarily optimal for crystal growth. The processes of nucleation and growth are thus decoupled by transferring nuclei from the labile zone into the growth zone (Till et al., 2013). Seeding has been shown to improve the number, size, and quality of existing crystals. The three approaches to crystal seeding are macro-seeding, micro-seeding, and streak-seeding (Till et al., 2013). In macro-seeding, a single large protein crystal is placed into a drop containing protein and crystallization solution. In micro-seeding, crushed crystal fragments in are placed into a drop containing protein and crystallization solution. In streak-seeding, a fine tool is used to harvest the crystals and transfer them into a drop containing protein and crystallization solution. Regardless of which approach is used, the drops into which the crystals are seeded

should be equilibrated such that the supersaturation levels are high enough to promote crystal growth, but low enough to prevent spontaneous nucleation. The results from the first two crystal trials revealed unwanted excessive nucleation, resulting in the formation of a shower of tiny microcrystals from which no meaningful diffraction data could be attained. Micro-seeding and streak-seeding were used in an attempt to grow fewer and larger crystals. Macro-seeding could not be attempted as I did not have a single large protein crystal to seed.

In order to do micro-seeding, a seed stock was created. To do this, the crystals from the second trial (***Figure 24***) were isolated from the drop and crushed with glass probe. The fragments were then resuspended in 50 µL of crystallization solution. Sitting drop experiments were setup, in which 1 µL of sample was mixed with 0.5 µL of crystallization solution and equilibrated against 250 µL of crystal solution 3. The experiments were sealed and left to equilibrate for 5 – 24 hours. Once the experiments were pre-equilibrated, 1.5 µL of the seed stock was placed into each drop and allowed to crystallize at 20 °C. After 3 days, several large bipyramidal crystals (≈ 0.1 – 0.2 mm) appeared (***Figure 25 (A)***). The crystals were then harvested and subsequently cryoprotected by 3 dips in either 100 (v/v) Parabar (Hampton Research, USA), 30% (v/v) glycerol (made up with mother liquor) or 30% PEG (made up with mother liquor). The presence of the crystal in the loop was confirmed using a darkfield protein crystallography microscope. The crystal was then immediately submerged in liquid nitrogen and taken for data collection. Data collection was unsuccessful due to the incorporation of aggregation which essentially poisoned the crystal lattice. The crystal was thus amorphous and therefore did not yield any significant diffraction data. Aggregates were probably incorporated during harvesting even though the seeds were thoroughly rinsed with mother liquor prior to seeding.

For streak-seeding, sitting drop experiments were set up, in which 1 µL of sample was mixed with 0.5 µL of crystallization solution and equilibrated against 250 µL of crystal solution 3. The experiments were sealed and left to equilibrate for 5 – 24 hours. Once the experiments were pre-equilibrated, crystals were harvested from the second trial (***Figure 24***) with a CrystalCap SPINE HT goniometer base fitted with a 0.05 mm Mounted CryoLoop™ (Hampton Research, USA), and placed directly into the pre-equilibrated drops. After 2 days, a few large bipyramidal crystals (≈ 0.2 – 0.3 mm) appeared (***Figure 25 (B)***). The crystals were then harvested and subsequently cryoprotected by 3 dips in either 100 (v/v) Parabar (Hampton Research, USA),

30% (v/v) glycerol (made up with mother liquor) or 30% PEG (made up with mother liquor). The presence of the crystal in the loop was confirmed using a darkfield protein crystallography microscope. The crystal was then immediately submerged in liquid nitrogen and taken for data collection. Data collection was unsuccessful due to the incorporation of aggregation which essentially poisoned the crystal lattice, in the same way as for micro-seeding. The crystal was thus amorphous and therefore did not yield any significant diffraction data.

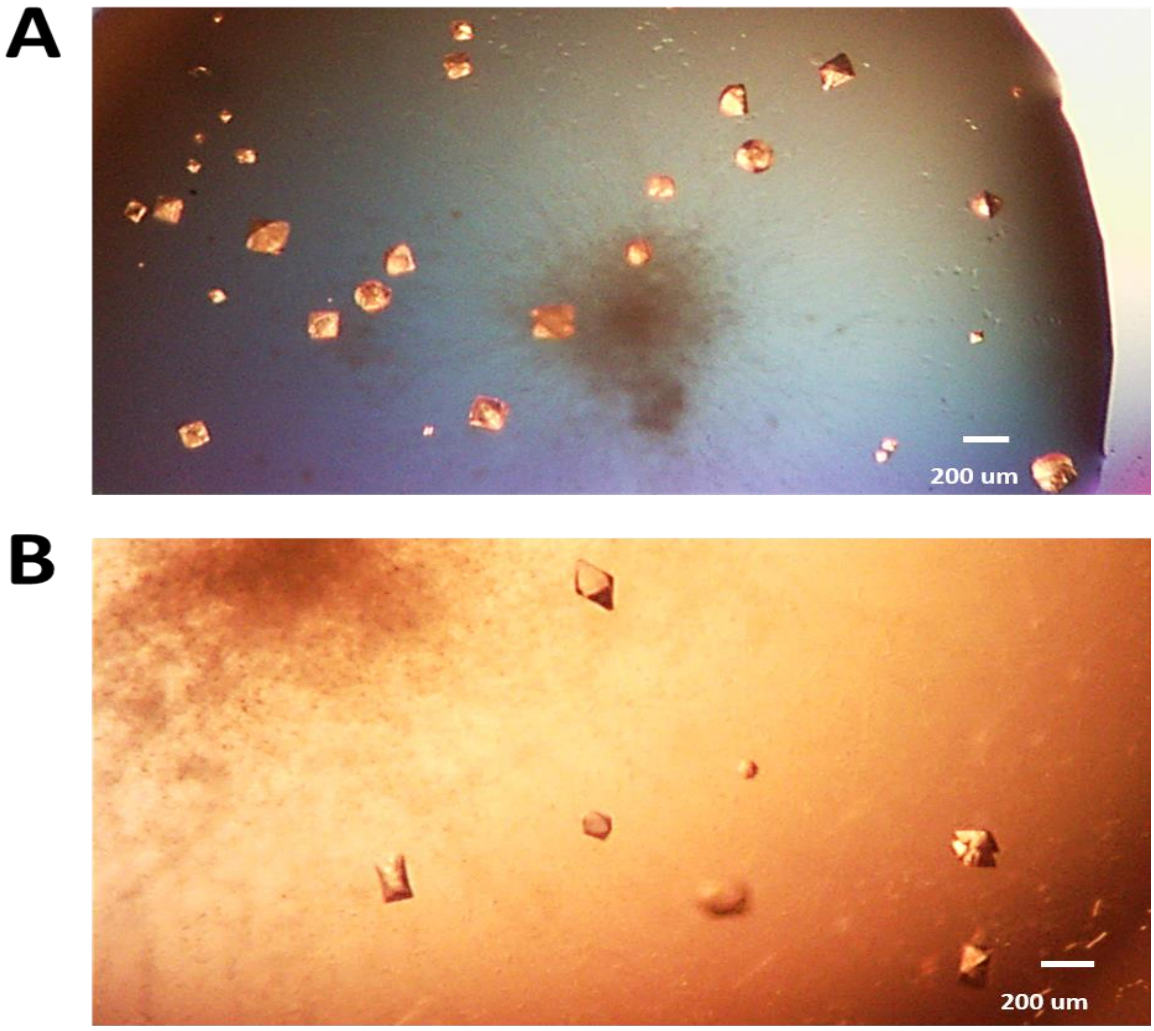



***Figure 25. Large protein-DNA crystals obtained by crystal seeding of tiny microcrystals.*** *(A) Bipyramidal protein-DNA crystals obtained by micro-seeding. The crystals were approximately 0.1 – 0.2 mm in length. (B) Protein-DNA crystals obtained by streak-seeding. The crystals were approximately 0.2 – 0.3 mm in length. In both cases, the crystals were harvested with a CrystalCap SPINE HT goniometer base fitted with either a 0.05 – 0.1 mm, 0.1 – 0.2 mm or 0.2 – 0.3 mm Mounted CryoLoop™ (Hampton Research, USA). The crystals were then cryoprotected by 3 dips in either 100 (v/v) Parabar (Hampton Research, USA), 30% (v/v) glycerol (made up with mother liquor) or 30% PEG (made up with mother liquor) and frozen in liquid nitrogen. The data collection was unsuccessful due to the incorporation of aggregation which essentially poisoned the crystal lattice. The crystal was thus amorphous and therefore did not yield any significant diffraction data.*

Since viable crystals could not be obtained by reducing the number of nuclei through seeding, efforts were focused on slowing down the kinetic processes that govern crystal formation. It was hoped that the protein in the solution would not be depleted before it had the time to contribute to crystal growth. Theoretically, this could be achieved by lowering the temperature, or by lowering the rate of vapor diffusion. The rate of vapor diffusion was optimized simply because there was no appropriate setup to conduct crystal trails at cooler temperatures.

As seen in our trials, protein crystallization often occurs too quickly, producing showers of microcrystals instead of single large crystals that can be used for diffraction. In a method known as vapor diffusion rate control (VDRC), a classical vapor-diffusion experiment is set up with a layer of mineral oil over the reservoir (Chayen, 1997). Since the oil is immiscible, it will float on the top of the reservoir solution and act as a barrier between the reservoir and the drop. Vapor diffusion will take longer to reach equilibrium since the water molecules will move slower through the layer of oil. The extent to which the vapor diffusion is slowed down is a function of the type of oil used, as well as the thickness of the oil layer above the reservoir. A combination of silicon oil and paraffin oil is traditionally used to overlay the reservoirs, and by varying the ratio of paraffin oil to silicon oil, the rate of vapor diffusion can be controlled. A higher percentage of silicon oil leads to a lower rate of vapor diffusion.

In order to try out vapor diffusion rate control, a sitting drop experiment was first set up, identical to the second crystal trial, except that the reservoir was covered with 50 µL of Al's oil (50% (v/v) paraffin oil and 50% (v/v) silicon oil). The SSL DNA was thus mixed with the protein at a final concentration of 0.5 mM SSL DNA duplex and 0.42 mM protein (the DNA was in 1.2 molar excess of the protein). The sample was incubated at 20 °C for 1 hour to allow for the DNA-binding reaction to reach equilibrium. Very thin plate-like crystals (***Figure 26***) appeared after 1 week at 20 °C in sitting drops containing 1.5 µL of DNA/protein solution and 1.5 µL of crystal solution 3, equilibrated against 250 µL of crystal solution 3, which was overlaid with 50 µL of Al's oil. The crystals appear to grow as a cluster of spiralling plates from a common nucleus at the surface of the crystallization support (plate). A single nucleus formed at the surface of the plate (the bottom of the drop) through heterogeneous epitaxy and gave rise to an active crystal. The side of the crystal that was attached to the plate was deprived of growth while the other sides of the crystal were growing, resulting in considerable

stresses on the crystal lattice. To relieve the mechanical stress imposed on it, the crystal splintered causing crystals to grow in all directions, resulting in a crystal bouquet (McPherson and Cudney, 2014).

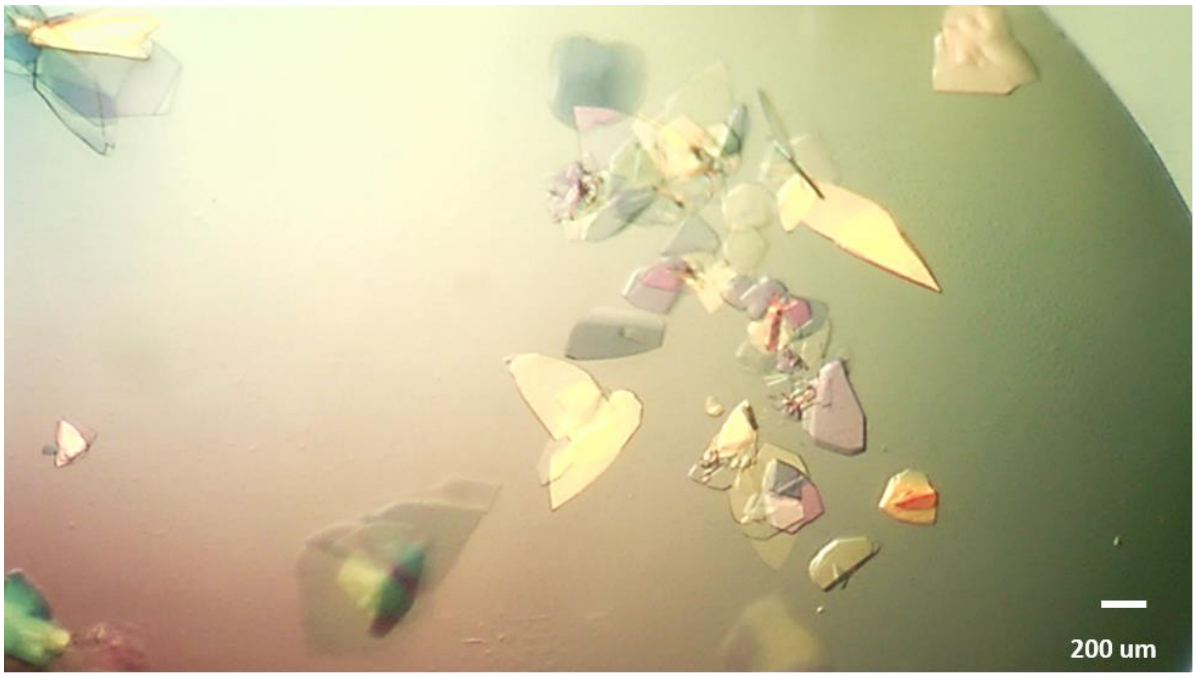



***Figure 26. A bouquet of plate-like protein-DNA crystals obtained by vapor diffusion rate control.*** *Crystal trials of the TBR1 T-box domain in the presence of SSL DNA were performed at 20 °C by using 0.5 mM SSL DNA duplex and 0.42 mM TBR1 T-box domain. A bouquet of very thin plate-like crystals appeared after 1 week in sitting drops containing 1.5 µL of DNA/protein solution and 1.5 µL of crystal solution 3 (100 mM 4-morpholineethanesulfonic acid pH 6.5, 200 mM $MgCl_2$ and 25% (w/v) PEG 4 000), equilibrated against 250 µL of crystal solution 3, overlayed with 50 µL of Al's oil (50% (v/v) paraffin oil and 50% (v/v) silicon oil). The crystals were harvested with a CrystalCap SPINE HT goniometer base fitted with either a 0.2 – 0.3 mm, 0.3 – 0.4 mm or 0.4 – 0.5 mm Mounted CryoLoop™ (Hampton Research, USA). The crystals were then cryoprotected by 3 dips in either 100 (v/v) Parabar (Hampton Research, USA), 30% (v/v) glycerol (made up with mother liquor) or 30% PEG (made up with mother liquor) and frozen in liquid nitrogen. Data collection was unsuccessful as the crystals were just too thin to cause significant diffraction.*

The crystals obtained were indeed protein crystals. This was confirmed by setting up an identical experiment and staining the resultant drop with 0.5 µL of Izit Crystal Dye (Hampton research, USA) as per the manufacturer's protocol. The blue staining against the lighter background, shown in ***Figure 27***, is evidence that the crystals were protein in nature. Protein crystals can take up the dye because of their relatively high solvent content, whereas salt crystals cannot.

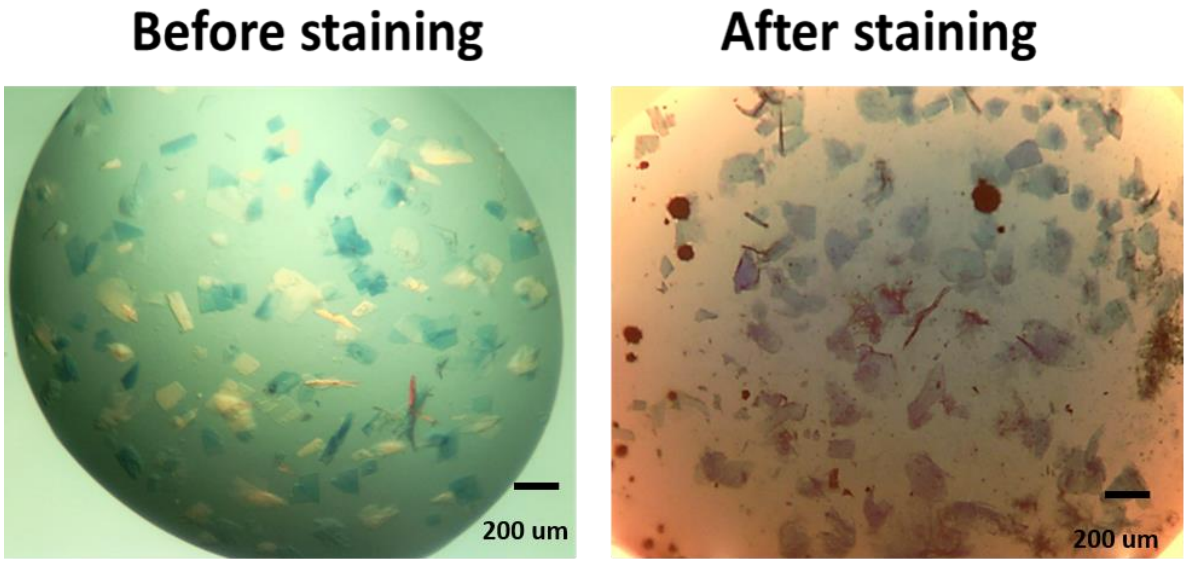



***Figure 27. Very thin plate-like protein-DNA crystals before and after staining with Izit Crystal Dye.*** *Crystal trials of the TBR1 T-box domain in the presence of SSL DNA were performed at 20 °C by using 0.5 mM SSL DNA duplex and 0.42 mM TBR1 T-box domain. Very thin plate-like crystals appeared after 1 week in sitting drops containing 1.5 µL of DNA/protein solution and 1.5 µL of crystal solution 3 (100 mM 4-morpholineethanesulfonic acid pH 6.5, 200 mM $MgCl_2$ and 25% (w/v) PEG 4 000), equilibrated against 250 µL of crystal solution 3, overlayed with 50 µL of Al's oil (50% (v/v) paraffin oil and 50% (v/v) silicon oil). Staining was achieved through the addition of 0.5 µL of Izit Crystal Dye. The picture on the right (after staining) shows that the plate-like crystals are a darker blue when compared to the background, confirming that the crystals were indeed protein crystals.*

The crystals obtained from the bouquet were very fragile and difficult to manipulate, even though they were relatively large (0.3 – 0.4 mm). A few large crystal fragments were harvested by slicing it off from the rest of the bouquet. The crystals did not yield any significant diffraction because the plate-like crystals had a very small unit cell volume and there simply weren't enough protein molecules in the lattice. This was very unfortunate since the TBX3 crystals that diffracted to 1.7 Å were also very thin and plate-like (Coll et al., 2002). The crystallization conditions could not be optimized any further, and therefore *in silico* methods were used to understand the DNA-binding of the TBR1 T-box domain.

## 5.5. *In silico* analysis

The DNA-binding mechanism of the TBR1 T-box domain could not be determined through XRC. The protein did not crystallize in the absence of DNA, while the crystals obtained in the presence of DNA did not diffract significantly due to their poor quality. *In silico* techniques were thus used to find out some information about DNA-binding. This information was also used to help with the interpretation of the *in vitro* results. The three-dimensional structure of

the TBR1 T-box domain in the absence of SSL DNA was predicted using *ab initio* techniques, followed by model validation with well-documented algorithms. The disorder of the protein was predicted in order to understand the role of the $3_{10}$C helix in DNA-binding, and to determine if the crystallization could be improved by removing intrinsically unstructured and disordered regions of the protein. Finally, the predicted structure was docked to SSL DNA to determine a possible mode of DNA-binding for the TBR1 T-box domain.

### 5.5.1. *Ab initio* protein modelling

The three-dimensional structure of the TBR1 T-box domain was predicted by the RoseTTaFold algorithm (*ab initio* modelling). The predicted structure can yield vital information about T-box proteins in the absence of DNA, as there is only one such crystal structure available (Stirnimann et al., 2010). Additionally, it can be used to compare the few T-box crystal structures that are available, which can provide valuable insights into DNA-binding and putative protein-protein interactions. The model was also used to help with the interpretation of the *in vitro* results. Since there is no crystal structure available, the predicted structure was required to obtain a binding pose of the TBR1 T-box domain to SSL DNA through molecular docking. The RoseTTaFold algorithm makes use of a three-track neural network to make accurate predictions of protein structures and interactions (Baek et al., 2021). The information obtained at the one-dimensional sequence level, the two-dimensional distance map level and three-dimensional coordinate level are successively transformed and integrated, allowing the network to collectively reason about connections within and between sequences, distances, and coordinates (Baek et al., 2021). The global accuracy of the predicted structure is indicated by the confidence-score (C – score), which is based on the Local Distance Difference Test (LDDT). A C – score of 0 corresponds to the poorest model while a C – score of 1 corresponds to a perfect model. The LDDT evaluates the differences in the local distance of all the atoms in a model without using superposition and considers the stereochemical plausibility of the model (Mariani et al., 2013). The RoseTTaFold webserver generates multiple conformations of the protein which are grouped into clusters based on their structural similarity. The server then provides the best model in each cluster based on the confidence score. The local accuracy of each amino acid residue is indicated as a distance error in angstroms (error estimate). Since some regions in the protein are void of any structure, a high angstroms error estimate does not necessarily mean that the local structure

could not be accurately predicted. Rather, it could mean that those regions are intrinsically disordered and thus flexible and dynamic.

The predicted structure of the TBR1 T-box domain is shown in ***Figure 28***. The model has a C-score of 0.75 indicating that the global structure is reliable and accurate. The local accuracy of the model is depicted in the error estimate plot in ***Figure 29***. The plot shows that the core of the protein has well defined local structure with the exception of the N and C termini. This was expected since the N-terminus contained the purifications tags and the thrombin cleavage site, which are often quite unstructured. It is interesting that the C-terminus has a high error estimate in the absence of DNA since it was hypothesized that the $3_{10}$C helix becomes structured upon DNA-binding. The regions represented by residues 100 -120, 140 – 160 and 175 - 190 have been shown to interact with the DNA in other crystal structures and as such, it makes sense for them to have less defined local structure.

The predicted structure in ***Figure 28*** shows that the TBR1 T-box domain is predominantly β-sheeted, as is the case for all T-box proteins. The protein contains a seven-stranded β-barrel core (blue), which is closed off by 2 anti-parallel β-strands (orange). The arrangement of the β-strands in the β-barrel is characteristic of the C2 (or s) subtype immunoglobulin fold as expected. The preservation of the core of the protein suggests that it will be able to recognize the TBE as expected of all T-box proteins. The purification tags and thrombin cleavage site (yellow) are quite mobile as expected. At the C-terminus, there are 2 mutually perpendicular α-helices, namely helix α3 (purple) and helix $3_{10}$C (light green). The helical nature of helix $3_{10}$C is quite unexpected since it was hypothesized that it would only become structured upon DNA-binding. However, the error estimate plot in ***Figure 29*** hints at the possibility that this region is inherently unstructured since the local accuracy of the residues in this region approaches 5 Å. The presence of the C-terminal α-helices suggest that the protein would be able to interact with and bind to DNA containing a TBE.



***Figure 28. The predicted structure of the TBR1 T-box domain obtained from RoseTTaFold.*** *The model of the protein has a C - score of 0.75 indicating that the global structure is fairly accurate and reliable. The core of the protein is made up of a seven-stranded β-barrel (blue) suggesting an S – type immunoglobulin fold. The barrel is closed off at one end by two anti-parallel β-strands (orange). The core will allow the protein to recognize the TBE in the DNA. The purification tags and thrombin cleavage site (yellow) are quite unstructured as expected. At the C-terminus, there are 2 mutually perpendicular α-helices, namely helix α3 (purple) and helix $3_{10}$C (light green). These helices will allow the protein to interact with and bind to DNA containing the TBE. The helical nature of $3_{10}$C is quite unexpected since it was hypothesized that it would only become structured in the presence of the DNA.*

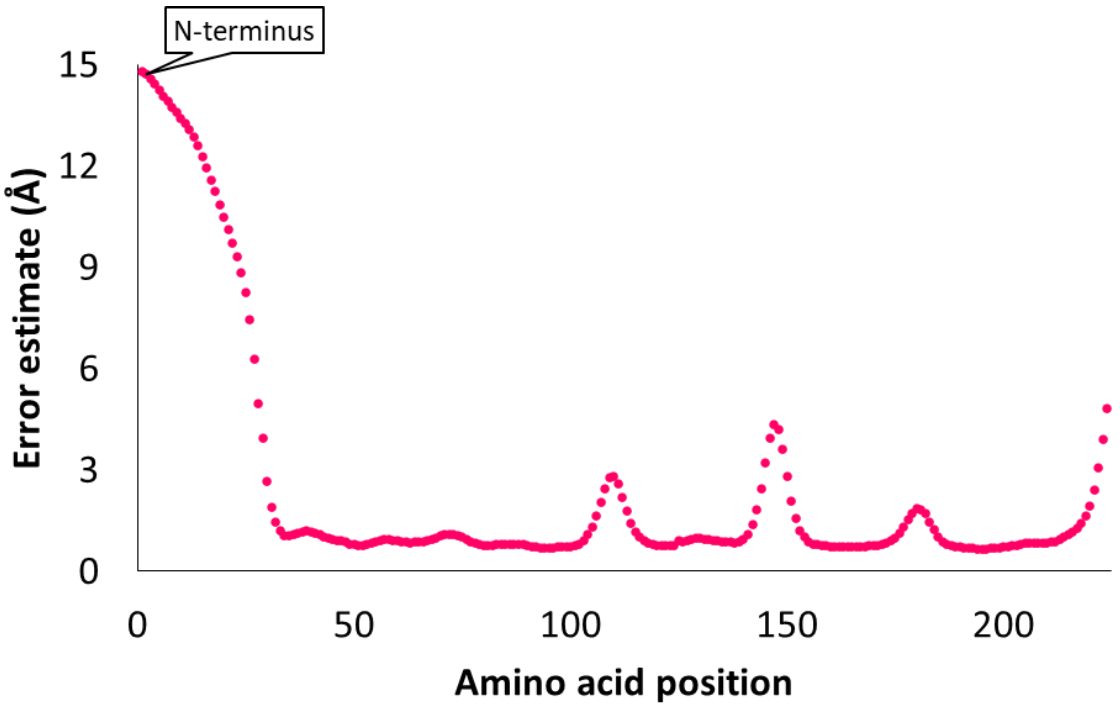



***Figure 29. The error estimate plot used to validate the local accuracy of the predicted TBR1 T-box domain structure.*** *Regions with a high error estimate indicate either a poor local accuracy, or a lack of intrinsic structure. The error estimates are below 5 Å for most regions indicating an accurate local structure. The regions represented by residues 100 -120 and 140 – 160 have been shown to interact with the DNA in other T-box crystal structures, so it makes sense for them to have poorly defined local structures. The purification tags and thrombin cleavage site, located at the N-terminus, is highly dynamic and flexible which explains why the error estimate was above 5 Å. The error estimate is relatively high for the C terminus which contains helix $3_{10}$C. This suggests that helix $3_{10}$C could be unstructured in the absence of DNA.*

The RoseTTaFold - predicted structure of the TBR1 T-box domain was validated using MolProbity. MolProbity is an all-atom structure-validation web-service that evaluates model quality based on the global and local levels of protein structure. It is important to validate the structure of the predicted model because local errors can affect biological interpretation since nearly all structures contain some stereochemical outliers. The all-atom contact analysis provided by MolProbity operates by measuring the amount of overlap in the Van de Waals surfaces between pairs of non-bonded atoms. When the overlap between two non-bonded atoms is less than 0.4 Å, it is denoted as a clash. These clashes cannot occur in the actual molecule suggesting that one of the two atoms in the non-bonded pair is in the incorrect orientation.

The all-atom contact analysis results obtained from MolProbity are shown in ***Table 1***. The clash score reports the number of serious steric overlaps, greater than 0.4 Å, per 1000 atoms. The percentile rank for the clash score is also indicated for the relevant resolution range. The

clash score should be as low as possible, with a high percentile rank across the relevant resolution range. The clash score of 0.28 for our model was in the 99$^{th}$ percentile at all resolutions indicating that there were no serious steric overlaps between non-bonding atoms. Rotamers are the various positions that the sidechain of an amino acid can take since various conformational isomers can be assumed. Favoured rotamers have the lowest possible energy conformation making them the most likely. All the rotamers in our model were favoured indicating that the side chains were in the lowest possible energy conformation. The Ramachandran distribution Z – score describes how normal the model is relative to a reference set of high resolution structures by characterizing the shape of the (ϕ, ψ) angle distribution in the Ramachandran plot. The Z -score of 0.25 ± 0.54 in our model indicates that the angles of the protein backbone are in the preferred conformations. The MolProbity score is the single number that best represents all the protein quality statistics obtained. The MolProbity score of 0.84 lies in the 100$^{th}$ percentile indicating that the model is of an excellent quality. Cβ deviations refer to steric overlaps between the side chains and the protein backbone. There were no Cβ deviations in the protein structure. The number of bad angles is only slightly disallowed for our model and the overall quality could not be improved by adjusting the geometry of the 3 outliers. The outliers contain ring structures in their side chains which suggests that these geometries are not as rigid as expected. CaBLAM (Cα Based Low-resolution Annotation Method) outliers refer to residues that do not have the predicted secondary structure based on the Cα coordinates. The number of CaBLAM outliers was in the allowed range for our model. The CA geometry refers to the geometry of the protein backbone with respect to the alpha carbons. There were no CA geometry outliers in our model. The predicted model of the TBR1 T-box domain thus met nearly all the criteria required for it to be deemed a perfect model. The model could thus be used for downstream applications including structural comparisons with other T-box proteins, and molecular docking.

***Table 1. All-atom contact analysis of the predicted TBR1 T-box domain structure.*** *The structure of the protein was predicted by RoseTTaFold and validated with MolProbity. The values have been highlighted as per the MolProbity guidelines. Values shown in green are good, values shown in yellow are acceptable if justified and values shown in red are not allowed. The results of the analyses suggest that this model is highly plausible, and possibly more stereo-chemically accurate when compared to the T-box crystal structures available in the protein data Bank.*

| | |
|---|---|
| Clashscore | 0.28 (99$^{th}$ percentile)<br>(N = 1784, all resolutions) |
| Poor rotamers (%) | 0 |
| Favored rotamers (%) | 100 |
| Ramachandran outliers (%) | 0 |
| Ramachandran favored (%) | 96.4 |
| Ramachandran distribution<br>Z - score | 0.24 ± 0.54 |
| MolProbity score | 0.84 (100$^{th}$ percentile)<br>(N = 26 765, all resolutions) |
| Cβ deviations > 0.25 Å (%) | 0 |
| Bad bonds (%) | 0 |
| Bad angles (%) | 0.12 |
| CaBLAM outliers (%) | 2.3 |
| CA geometry outliers (%) | 0 |

Once the predicted structure of the TBR1 T-box domain was fully validated using MolProbity, it was compared with the crystal structure of the TBX5 T-box domain (the only T-box crystal structure available in the absence of DNA). The structural alignment is shown in ***Figure 30***. The structures align well with a Cα RMSD of 0.98 Å across 151 pruned atom pairs, and 1.962 Å across all 176 pairs, as determined in UCSF Chimera v 1.16 (Pettersen et al., 2004). There are two notable structural differences between the proteins. The variable loop region in the predicted structure is in a more extended conformation which could allow it to interact with other proteins. The most important structural difference is the exclusive presence of helix $3_{10}$C in the predicted structure. In the TBX5 crystal structure, this helix is unstructured in the absence of DNA and only becomes structured upon binding. Interestingly, the predicted

structure in the absence of DNA aligns even better with the crystal structures of other T-box proteins in the presence of DNA.

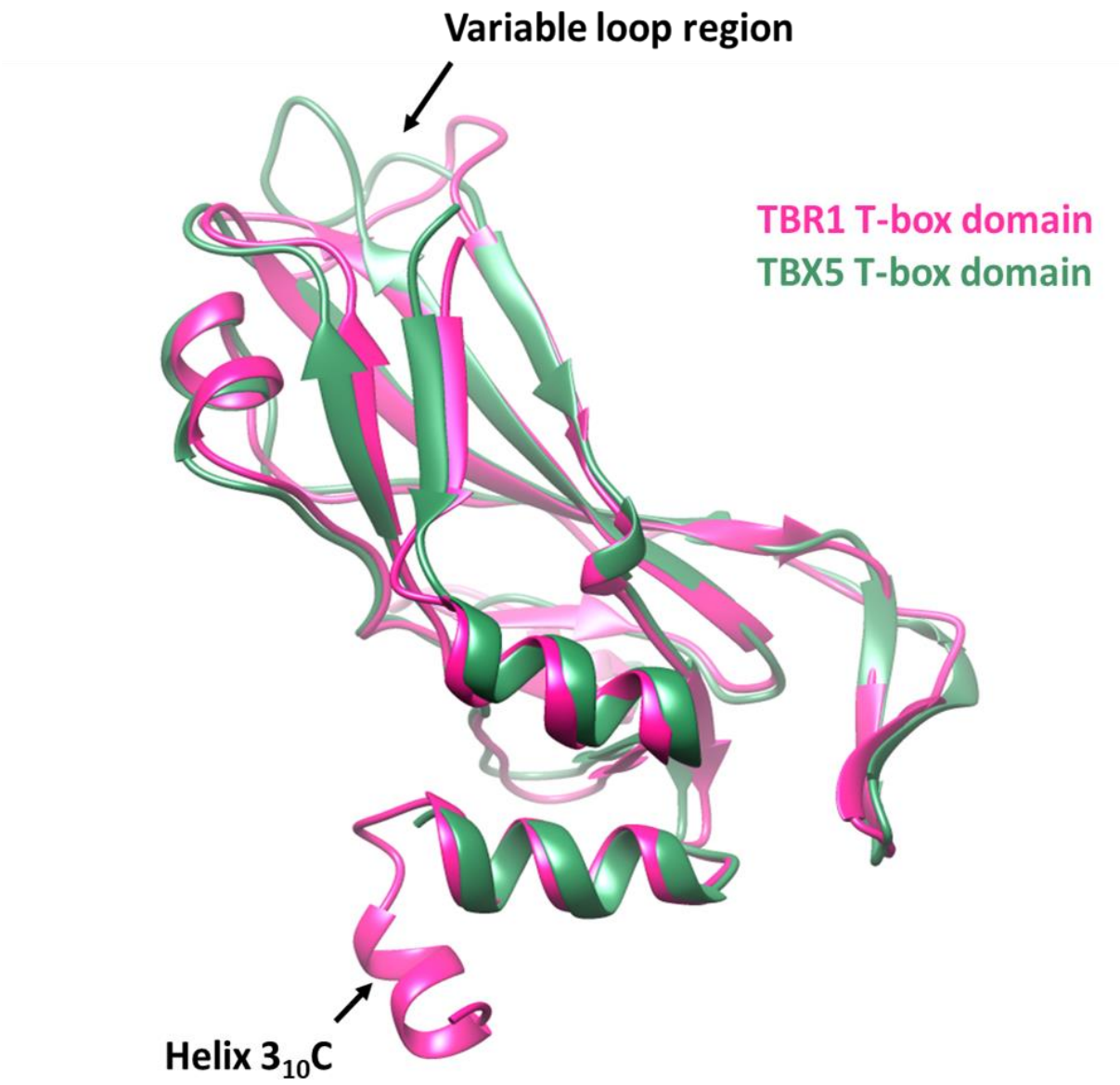



***Figure 30. Structural alignment of the predicted TBR1 T-box domain with the crystal structure of the TBX5 T-box domain.*** *The structures align well with a Cα RMSD of 0.98 Å across 151 pruned atom pairs, and 1.962 Å across all 176 pairs, as determined in UCSF Chimera v 1.16 (Pettersen et al., 2004). There are two notable structural differences between the proteins. The variable loop region in the predicted structure is in a more extended conformation. The most important difference is the exclusive presence of helix $3_{10}$C in the predicted structure. In the TBX5 crystal structure, this helix is unstructured in the absence of DNA and only becomes structured upon binding. The PDB accession number for the TBX5 T-box domain in the absence of DNA is 2X6U (Stirnimann et al., 2010). The structures were aligned and visualized in UCSF Chimera v 1.16 (Pettersen et al., 2004).*

### 5.5.2. Disorder predictions

There are a number of proteins, particularly TFs, which have regions with no secondary structure. These IDRs are flexible and dynamic; and play a role in the molecular recognition of proteins and DNA through conformational changes. The IDRs are characterized by a high frequency of charged and hydrophilic residues as well as low sequence complexity (Uversky, 2013). Disorder-to-order transitions are often the mechanism by which a TF interacts with DNA. This can clearly be seen in the crystal structures of the TBX5 T-box domain wherein helix $3_{10}$C only becomes ordered upon DNA-binding. The flexibility of the IDRs has negative implications for XRC since it hinders the formation of crystal contacts.

The IDRs of the protein were predicted from the amino acid sequence using the PrDOS webserver. The disorder is predicted based on local amino acid sequence information as well as template proteins. The server returns a disorder probability for each residue. At a false positive rate of 5%, the threshold for disorder probability was 0.5. This means that residues with a disorder probability greater than 0.5 are said to be disordered. The predicted disorder probability of the residues in the TBR1 T-box domain are shown below in ***Figure 31***. The N – terminal region of the protein has a very high disorder probability and can most likely assume multiple conformations as expected. This region could significantly have hindered crystallization. The regions represented by residues 100 -120, 140 – 160 and 175 – 190 are predicted to have some disorder. These IDRs can be seen in various T-box crystal structures where they have poor or no electron density. They are thus not thought to have hindered crystallization. The C – terminal region of the protein has a high disorder probability. This makes sense since it contains helix $3_{10}$C, which probably only becomes structured upon DNA-binding. The error estimate plot produced by RoseTTaFold overlays well with the disorder probability plot produced by PrDOS. This correlation confirms the notion that regions with a high error estimate are not necessarily incorrectly modelled using *ab initio* methods, but rather, that these regions just do not have any inherent structure to predict.

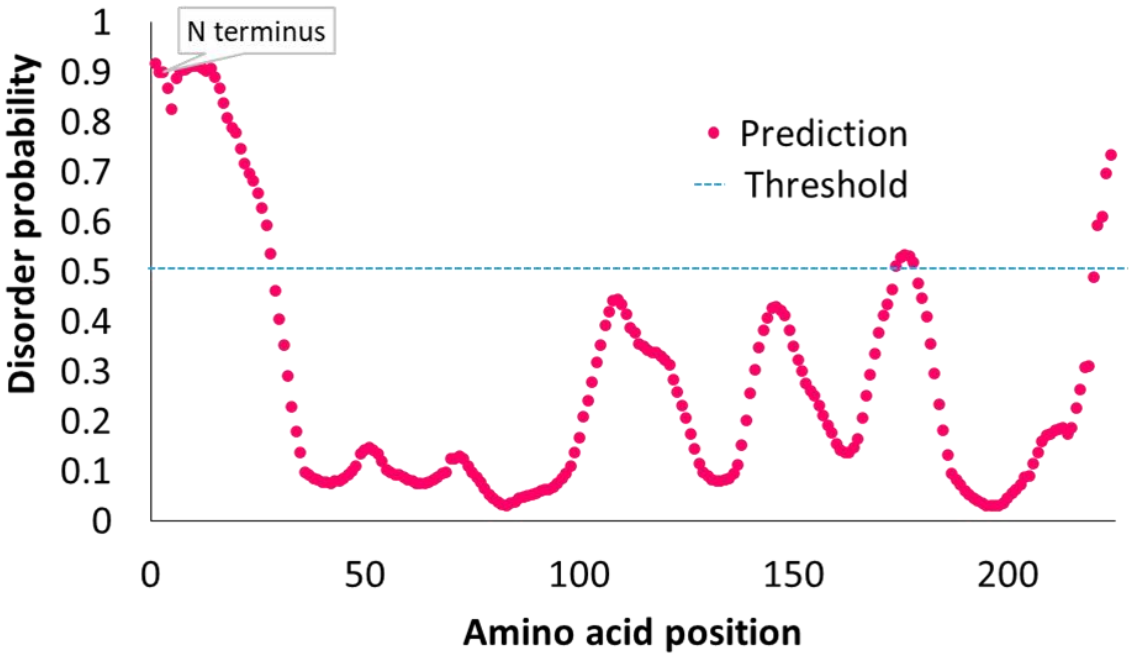



***Figure 31. Disorder probability prediction for the TBR1 T-box domain.*** *At a false positive rate of 5%, the threshold for disorder probability was 0.5. This means that residues with a disorder probability greater than 0.5 are said to be disordered. The N – terminus, residues 175 – 190 and the C – terminus are thus disordered. The disorder in the N – terminus is expected as it contains the purification tags and thrombin cleavage site which are quite flexible. It could thus significantly have hindered crystallization. Residues 175 -190 are slightly disordered. The corresponding region in the TBX5 T-box crystal structure lacks electron density suggesting that it is also disordered, and that this region probably did not hinder crystallization. The C – terminal domain is relatively disordered due to the presence of helix $3_{10}$C, which I predict only becomes ordered upon DNA-binding. The disorder probability was calculated using the PrDOS webserver (Ishida and Kinoshita, 2007).*

### 5.5.3. Molecular docking

Molecular docking is the computational method that predicts how two molecules fit together. As such, it is used to predict the structure of a protein-DNA complex when a crystal structure is unavailable. It provides insight into the number and types of bonds that could form, as well as the residues that are potentially important for DNA-binding affinity and specificity. The predicted structures are often compared to crystal structures or homology models to study the differences and similarities in the DNA-binding of TFs from the same family, to deduce a DNA-binding mechanism in the absence of a crystal structure. The TBR1 T-box domain was docked to SSL DNA with the HADDOCK webserver using the default parameters. The structure of the protein was predicted using RoseTTaFold and the structure of SSL DNA was modelled in Accelrys Discovery studio v 4.1. The active residues in the protein were chosen based on other crystal structures while the active residues in the DNA were defined as those

constituting the TBE. The docking protocol was validated with a positive control in which the TBX5 crystal structure was docked to its DNA sequence using the same parameters in the HADDOCK webserver. This positive control was carried out to ensure the correct docking/binding pose had been obtained. The docked TBX5 structure closely aligns to the crystal structure in the PDB, with a root mean square deviation of 0.6 Å, showing that there were no flaws in the docking protocol.

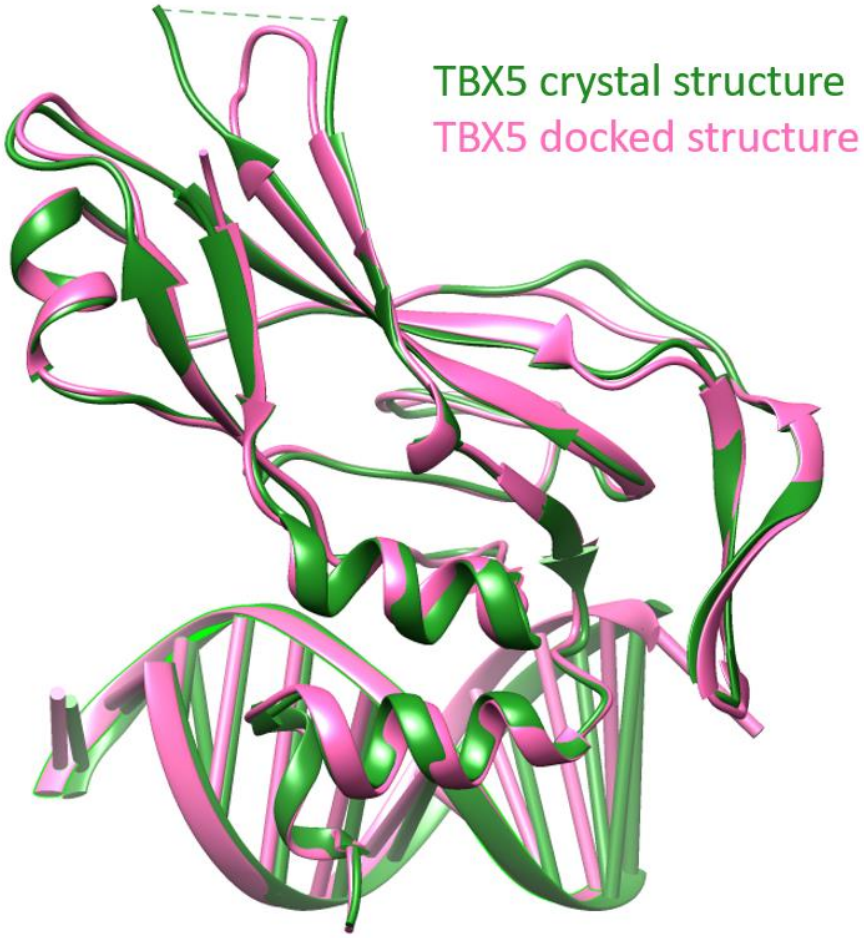



***Figure 32. Structural alignment of the TBX5 crystal structure with TBX5 docked to its DNA using HADDOCK.*** *To verify the docking protocol, the TBX5 crystal structure with no DNA has been docked to its DNA using the same parameters in the HADDOCK webserver. This positive control was carried out to ensure the correct docking/binding pose had been obtained. The docked structure (pink) closely aligns to the crystal structure (green) with a root mean square deviation of 0.6 Å, showing that there were no flaws in the docking protocol. The PDB accession number for TBX5 in the DNA-bound state is 2X6V (Stirnimann et al., 2010). The structures were rendered and aligned with UCSF Chimera v 1.16 (Pettersen et al., 2004).*

The predicted interactions between the TBR1 T-box domain and SSL DNA are displayed in ***Figure 33 (A)***. The predicted structure of the protein-DNA complex is shown in ***Figure 33 (B)***, with a focus on DNA-binding. The protein makes direct contact with bases G11, T12, C32 and

A33 in the minor groove of the DNA via hydrophobic interactions formed by Phe215. This is consistent with the T-box structures that are currently available in which the bulky side chain of Phe215 is used to insert helix $3_{10}$C into the minor groove. This makes sense since these bases constitute the TBE. Helices α3 and $3_{10}$C are mutually perpendicular, and helix α3 positions helix $3_{10}$C such that it can slide along the minor groove into the DNA. The insertion of Phe215 into the minor groove causes the DNA to become slightly enlarged in the minor groove, which is also consistent with the T-box crystal structures. There are a number of hydrophobic interactions formed between the residues in helix $3_{10}$C, and the DNA backbone in the minor groove. The bulky side chain of Phe219 interacts with the sugar moieties of bases A14 and A15, which keeps helix α3 in the minor groove of the DNA. The corresponding residue is involved in hydrophobic contacts in the crystal structures because helix $3_{10}$C is inserted deeper into the minor groove. The residues in helix $3_{10}$C in our model thus have a greater degree of solvent accessibility. There are a number of hydrogen bonds formed between residues protruding from loops extending down from the β-barrel, and the backbone of the DNA in the minor groove. These loops are the most conserved DNA-recognition elements in the superfamily of IL domains and account for the specific recognition of the TBE by T-box proteins. Even though most of the interactions that govern binding take place in the minor groove, DNA recognition also occurred in the major groove. The protein also makes direct contact with bases A26 and T27 in the major groove of the DNA via hydrogen bonds formed with residue Lys56 located between strand a and strand b. This region is outside the TBE, and this interaction could thus explain the varying affinities with which T-box proteins bind DNA.

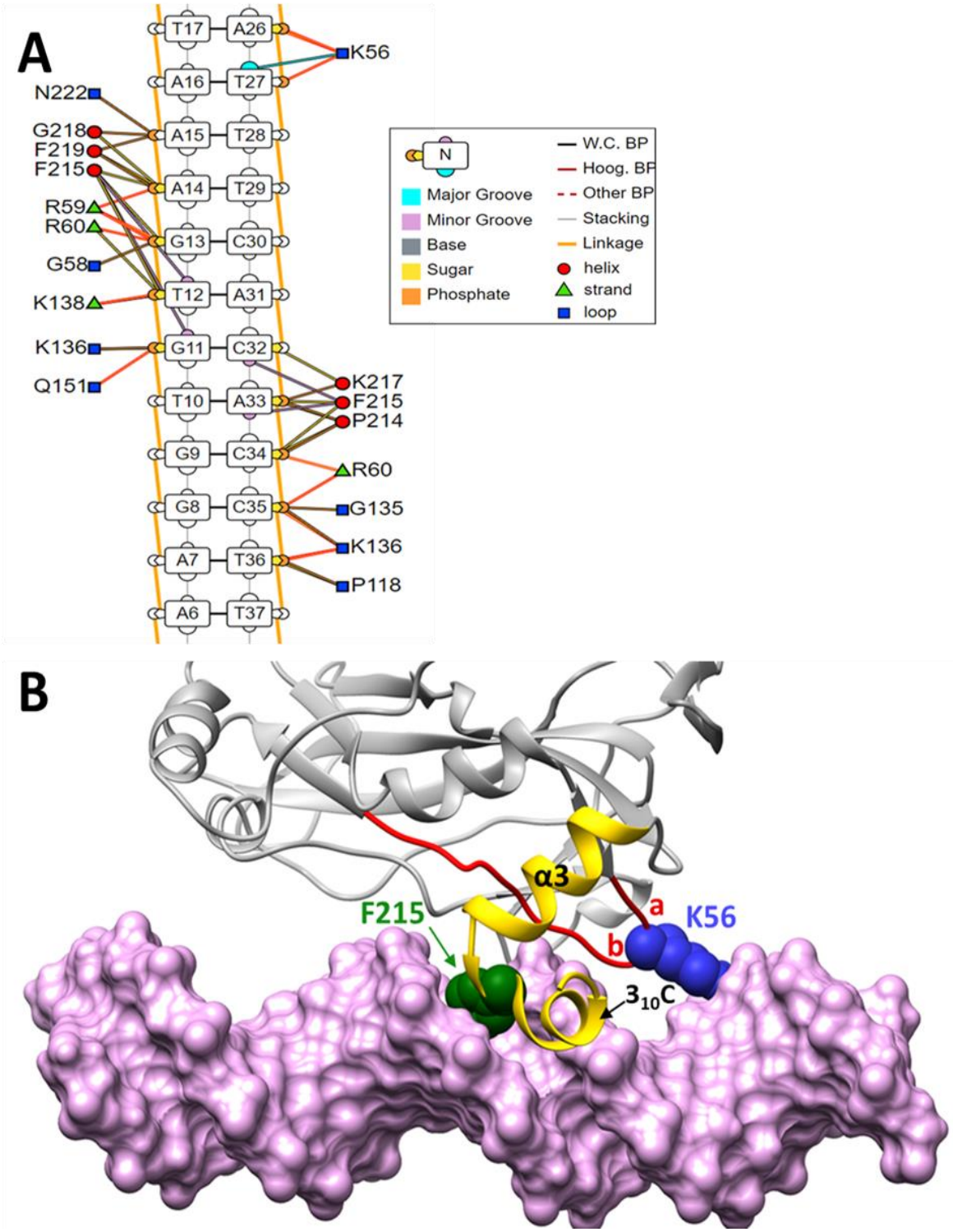



***Figure 33. The predicted structure of the TBR1 T-box domain in the presence of single site long DNA.*** *(A) Schematic diagram of the interactions between the predicted TBR1 T-box domain and single site long DNA. Hydrogen bonds are indicated by red lines. The protein – DNA interface was created using the DNAproDB webserver (https://dnaprodb.usc.edu/index.html) (Sagendorf et al., 2017). (B) Important structural features of the DNA-binding interaction. The protein makes direct contact with bases G11, T12, C32 and A33 in the minor groove of the DNA via hydrophobic interactions formed by Phe215. The protein also makes direct contact with bases A26 and T27 in the major groove of the DNA via hydrogen bonds formed with residue Lys56 located between strand a and strand b. The structure was rendered with UCSF Chimera v 1.16 (Pettersen et al., 2004).*

# 6. Discussion

The TBR1 T-box domain is a neuron-specific TF involved in numerous developmental events in the brain, such as the migration and differentiation of neurons in the neocortex. It has recently emerged as a master regulator of the genes implicated in ASDs suggesting that mutations in the TBR1 T-box domain could underlie the altered neuromolecular networks observed in neurodevelopmental disorders. The primary molecular function of the TBR1 T-box domain is to bind the TBE in a sequence specific manner. The key to understanding the function of a macromolecule, such as the TBR1 T-box domain, is to determine and understand its structure at all levels.

The aim of this study was to determine the DNA-binding mechanism of the TBR1 T-box domain through biophysical characterizations of protein structure, DNA-binding studies, high resolution XRC and computational methods. The hypothesis was that the DNA-binding mechanism utilized by the TBR1 T-box domain is the same as that of the TBX5 T-box domain due to the conservation of the IL domain and the high sequence similarity between the proteins. The TBX5 T-box binds the TBE via the formation of an inducible recognition element, helix $3_{10}$C, which only becomes structured upon DNA-binding (Stirnimann et al., 2010). Additionally, there are various hydrogen bonds formed between loops extending from the β-barrel, and the backbone of the DNA in the minor groove. By understanding how TBR1 carries out its functions of DNA-binding and subsequent transcriptional repression, we may begin to understand what goes wrong, at the molecular level, in neurodevelopmental disorders such as ASDs. This is the first step in properly defining ASDs, and subsequently developing diagnostic tools and treatments thereof.

The DNA-binding mechanism of the TBR1 T-box domain could not be obtained by XRC since the protein could not be crystallized either in the presence or absence of DNA. Structural characterizations, DNA-binding studies and *in silico* methods were thus used in an attempt to explain the DNA-binding mechanism. The TBR1 T-box domain has significant clinical relevance, and as such, the pursuit of these crystal structures should be continued. To this effect, the crystallization of the protein has also been discussed, in an attempt to figure out how best to obtain diffraction-worthy crystals in the future.

## 6.1. The structure and DNA-binding mechanism of the TBR1 T-box domain

The *in vitro* and *in silico* results suggest that the overall structure and fold of the TBR1 T-box domain is highly similar to other T-box proteins. The far UV CD spectrum (***Figure19***) confirms that the protein is β-sheeted. This was expected since the core of the T-box domain is made up of a β-barrel consisting of seven anti-parallel β-strands closed off by lid-like structure consisting of a β-pleated sheet, as seen in the predicted model (***Figure 28***). The β-barrel is composed of a C2 (or s) immunoglobulin fold, consistent with other T-box proteins, and stabilizes the structure of the protein through hydrophobic interactions between highly conserved residues (Coll et al., 2002; El Omari et al., 2012; Liu et al., 2016; Stirnimann et al., 2010).

The ITF spectrum of the protein (***Figure 17***), as well as the predicted model (***Figure 34***), suggest that the conserved tryptophan residues are buried in the hydrophobic interior of the protein and thus contribute to the stabilization of the structure.

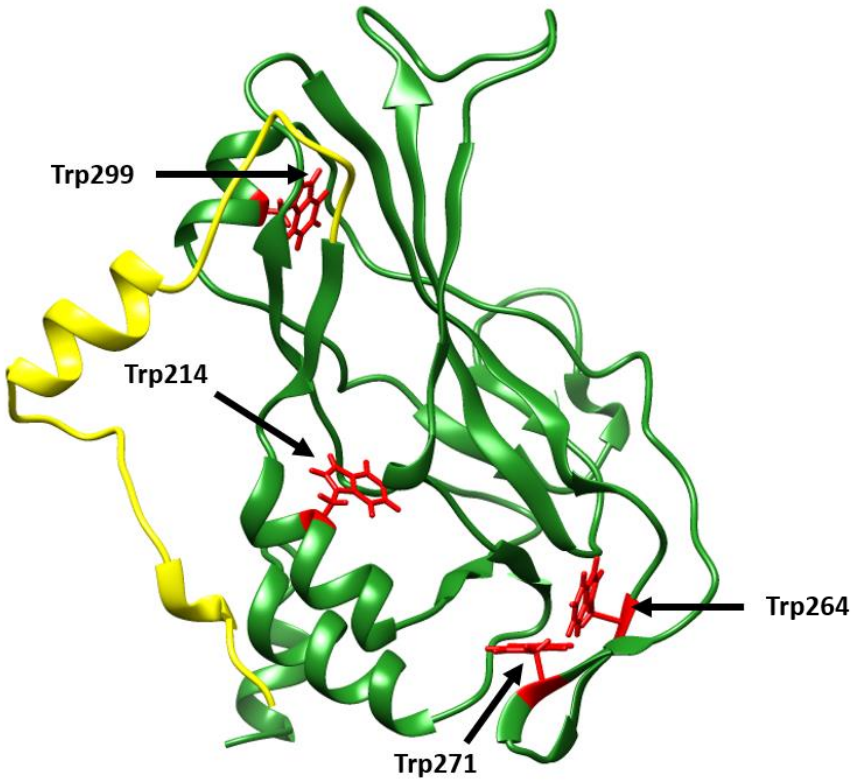



***Figure 34. The predicted structure of the TBR1 T-box domain showing the location of the tryptophan residues.*** *The protein (green) contains four tryptophan residues (red). The side chains of the tryptophan residues point towards the interior cavity of the protein, where they appear to stabilize the fold of the β-barrel. The purification tags have been shown in yellow. The structure was rendered with UCSF Chimera v 1.16 (Pettersen et al., 2004).*

Residue Trp121 in TBX5, corresponding to residue Trp99 in TBR1 (residue Trp271 in the original sequence), is conserved amongst T-box proteins. This residue is located in the lid-like structure and points towards the bottom of the β-barrel where it stabilizes the fold through hydrophobic interactions (Stirnimann et al., 2010). A W121G mutation in TBX5 resulted in a drastic decrease in the conformational stability of the T-box domain (Stirnimann et al., 2010). The removal of the bulky side chain of tryptophan introduced increased flexibility and mobility at the bottom of the β-barrel, leading to a destabilization of the entire structure. The corresponding W271R and W271C mutations in TBR1 have been shown to be pathogenic even though they have no effect on the transcriptional repression (De Rubeis et al., 2014; Hoed et al., 2018; O'Roak et al., 2012). Interestingly, the W271R mutation did not appear to affect the protein function while the W271C mutation had more pronounced effects. In the W271R mutation, the relatively long side chain of arginine is a good replacement for the bulky side chain of tryptophan and thus does not contribute to pathogenicity through the disruption of protein structure. Instead, the expression levels of this mutant have been shown to be much lower than that of the wildtype suggesting that reduced protein levels could contribute to the associated neurodevelopmental phenotype. The predicted model (***Figure 34***) suggests that the W271C mutation results in a large cavity in the core of the protein leading to severe changes in structure. Indeed, the pathogenicity of this mutant has been shown to arise from the presence of abnormal aggregates within the nucleus. Strikingly, both the W271R and W271C mutants have been shown to be more stable than that of the wildtype which can lead to an amplification of the functional effects (Hoed et al., 2018). In the case of W121C, the abnormal aggregation is therefore probably a result of altered protein interactions rather than a decrease in protein stability. The W271C mutant severely decreases the interaction with the FOXP2 TF while the W271R mutant has no effect on the interaction (Deriziotis et al., 2014; Hoed et al., 2018). It is hence possible that the decreased interaction with the FOXP2 TF resulted in the abnormal aggregation of the TBR1 T-box in the nucleus. This is perhaps one of the mechanisms resulting in ASDs with speech and language impediments. It is interesting that the Trp271 mutations in TBR1 resulted in an increase in stability while the corresponding Trp121 mutation in TBX5 did not. The TBR1 T-box domain, like other members of the TBR1 subfamily, contains a second tryptophan residue (Trp264) in the lid -like structure, which is absent in other T-box subfamilies. Trp264 stabilizes the β-barrel in the same way as Trp271,

which would explain why the W271R and W271C mutations did not decrease the stability of the protein in TBR1.

According to the results from molecular docking (***Figure 33***), the global structure of the TBR1 T-box – DNA complex is very similar to the structures of all the other T-box proteins in complex with their respective DNA sequences (Coll et al., 2002; El Omari et al., 2012; Liu et al., 2016; Müller and Herrmann, 1997; Stirnimann et al., 2010). The structural characterizations (***Figure 17*** – ***Figure 20***) show that the secondary structure, tertiary structure, and thermal stability of the protein is conserved upon DNA-binding. This is also consistent with previous findings (Stirnimann et al., 2010). The crystal structures of TBX5 suggest that T-box proteins recognize the TBE via an inducible recognition element at the C terminus of the T-box DBD, helix $3_{10}$C, which only becomes structured upon DNA-binding. In this work, there are no *in vitro* results to prove that helix $3_{10}$C is unstructured in the absence of DNA, and that it only becomes ordered upon DNA binding. There are, however, significant results from the *in silico* analysis to support this notion. The disorder predictions (***Figure 31***) suggest that the last 4 residues, constituting helix $3_{10}$C, are indeed unstructured in the absence of DNA. Even though helix $3_{10}$C is conserved amongst T-box proteins, the modelling results show that the error estimate for this region is high (***Figure 29***). This could thus mean that the structure of helix $3_{10}$C is difficult to predict because it does not have any structure and is thus disordered in the absence of DNA. It is therefore highly likely that the TBR1 T-box domain binds the TBE via the formation of an inducible recognition element, helix $3_{10}$C, which only becomes ordered upon DNA binding. The TBR1 T-box thus recognizes the TBE via indirect or shape readout. The formation and unwinding of helix $3_{10}$C will probably affect the orientation of the C-terminal domain of the protein. This is significant since the C-terminal domain has been shown to mediate protein-protein interactions and transcriptional repression (Deriziotis et al., 2014; Hoed et al., 2018). The transition of helix $3_{10}$C from a disordered to an ordered state should therefore be explored further through high resolution nuclear magnetic resonance, Fourier transform infrared spectroscopy and hydrogen deuterium exchange mass spectrometry. These techniques are ideal for studying the structural dynamics of proteins in solution.

There may be an alternative explanation of the DNA-binding mechanism, if one considers the complete TBR1 structure, as predicted by the AlphaFold neural network, instead of just the T-box domain. It is possible that helix $3_{10}$C is the beginning of a longer α-helix, if one adds

residues traditionally not classified as part of the DNA-binding T-box domain. This would mean that the T-box DNA-binding domain is larger than that reported in the literature and that the C-terminal domain contains a longer α-helix. This may define the DNA-binding mechanism of T-box TFs into the typical helix-turn-helix mechanisms. If this is true, which would of course require experimental evidence, it may render the current TBX5 mechanism as an artifact of the chosen domain sequence for the crystallization experiments. This would also resolve the current contrast in DNA-binding mechanisms which currently suggest that T-box TFs bind DNA in the minor grove while all other TFs bind DNA in the major groove.

The TBR1 T-box domain makes direct contact with bases in the minor groove of the DNA via hydrophobic interactions formed by the bulky side chain of Phe215 (residue Phe387 in the original sequence) (***Figure 33***). This interaction is stabilized by other hydrophobic interactions between residues in helix $3_{10}$C and the sugar-phosphate backbone in the minor groove of the DNA. The data from modelling and docking, in conjunction with other T-box structures, suggest that Arg220 (residue Arg392 in the original sequence) is particularly important for the formation and positioning of helix $3_{10}$C (Coll et al., 2002; El Omari et al., 2012). This residue interacts, via hydrogen bond formation, with the main chain carbonyls of Ile 176 and Asn180 located in helix α3. Additionally, it acts as a cap for the C-terminus by neutralizing its negative charge and stabilizing the interactions between helices α3 and $3_{10}$C. The corresponding R237Q and R237W mutations in TBX5 have been shown to be pathogenic by affecting transcriptional regulation through a strongly reduced DNA-binding affinity (Stirnimann et al., 2010). It is thus likely that mutating this residue in TBR1 would alter the recognition and binding in the minor groove of the DNA. The protein makes direct contact with a base in the major groove via an electrostatic interaction between the positively charged lysine residue Lys56 (residue Lys228 in the original sequence) and the negatively charged DNA backbone (***Figure 33***). A pathogenic K228E mutation has been found in TBR1 and will likely abolish the electrostatic interaction formed with the negatively charged DNA backbone (Deriziotis et al., 2014; O'Roak et al., 2012). Strikingly, this mutation did not have any effect on DNA-binding or transcriptional repression in TBR1, but completely abolished the interaction with the FOXP2 TF (Deriziotis et al., 2014; Hoed et al., 2018). Lastly, several hydrogen bonds are formed between residues protruding down from the β-barrel and the sugar-phosphate backbone of the DNA constituting the TBE. These loops are the most conserved DNA-

recognition elements in the superfamily of IL domains and account for the specific recognition of the TBE by T-box proteins. The TBR1 T-box thus recognizes the TBE by direct or base readout. The results from docking show that there are several electrostatic interactions formed between residues protruding down from the β-barrel, and bases that flank the TBE. These interactions explain the high affinity observed in the DNA binding studies (***Figure 21*** and ***Figure 22***) and this is consistent with previous findings. The ability of T-box proteins to form bonds with bases flanking the TBE could explain the preferential binding of T-box proteins to their individual target sites. The interactions formed between TBR1 and the TBE are very similar to other T-box proteins since all the residues mentioned above are conserved in the T-box family.

There is considerable speculation about the ability of T-box proteins to dimerize upon DNA-binding. The first crystal structure of a T-box domain was that of *Xenopus laevis*. In this structure, a T-box dimer with a small dimerization interface (440 $Å^2$) is bound to a palindromic DNA sequence containing two TBEs (Müller and Herrmann, 1997). The palindromic sequence has not yet been identified in natural target sites of T-box proteins in humans. The monomers, each bound to a single TBE, interact through a combination of hydrogen bonds and hydrophobic interactions formed by residues 125 – 130, which are conserved in the TBX19 subfamily. The relatively small dimerization interface suggests that the protein is monomeric in solution and dimerizes upon DNA-binding. Functionally, this sequential binding allows for rapid complex formation that is not limited by dimerization and would also prevent the T-box from becoming kinetically trapped on low affinity sites in the DNA. In the structure of the TBX3-DNA complex, two T-box monomers are bound to the same palindromic sequence used for Xbra (Coll et al., 2002). The two monomers interact via a tiny interface (180 $Å^2$) composed of residues 238 – 241 (the variable loop region), which is absent in the T-box domain of Xbra. The tiny dimerization interface and the fact that these residues are highly mobile suggest that the 'dimerization' is a crystallographic artifact and thus does not have any significance. This region is highly mobile in all T-box proteins. Even though Xbra and TBX3 have different quaternary structures, the protein-DNA interactions and the DNA conformation is essentially the same. This means that the interactions of one monomer adjacently bound on the palindrome does not affect the binding of the second monomer to the second TBE. The structure of a TBX3 monomer bound to a single TBE in the palindrome is thus an accurate

model of a T-box domain in the presence of a natural TBE. This is well corroborated by the crystal structures of TBX5. The crystal structures of other human T-box proteins (besides TBX21) suggest that binding is monomeric, and any observable dimerization is likely just an artifact of crystallization (Stirnimann et al., 2010). The only T-box protein shown to have a functional dimerization interface is TBX21. In the TBX21 crystal structure, two T-box monomers are bound to each other, and each of them is bound to a single TBE from two separate palindromic sequences (Liu et al., 2016). As a result, the homodimerization of the TBX21 T-box domain allows the DNA to be looped such that two distant TBEs are brought together. The dimerization interface is large (2 478 $Å^2$) and is composed of two loops which are different in sequence and length to analogous loops in other T-box proteins. Loop 1 is composed of residues 248–256 and loop 2 is composed of residues 276–293. These residues are not well conserved in T-box proteins which would explain the lack of this type of dimerization in other T-box proteins. Tbx21 belongs to the TBR1 subfamily, and one would thus expect that the TBR1 T-box domain also dimerizes, in a similar fashion, upon binding to DNA. The results from SEC (***Figure 15***) show that the TBR1 T-box domain is monomeric in solution under reducing conditions. The DNA-binding studies (***Figure 21*** and ***Figure 22***), in conjunction with previous work, show that the TBR1 T-box domain binds the TBE as a monomer under reducing conditions. The predicted model (***Figure 28***) can be used to explain why the TBR1 T-box does not dimerize like TBX21. The residues in loop 1 of TBX21 are identical to that of TBR1 and are exclusive to the TBR1 subfamily. Loop 2 in TBX21 contains several positively charged residues which interact with similarly positively charged residues on its own monomer and on the two-fold related dimer, resulting in salt bridge formation between Arg269 and Glu279, and between Arg184 and Asp188. In the TBR1 T-box domain, Glu279 is replaced by a glycine and Arg184 is replaced by an asparagine residue. This prevents the formation of the salt bridges required to form a stable and relevant dimerization interface and explains why the TBR1 T-box domain is monomeric upon DNA-binding. The monomeric binding of the TBR1 T-box domain to the TBE suggests that pathogenic TBR1 T-box mutations function through a dominant-negative mechanism rather than through haploinsufficiency. It is therefore possible that mutant T-box proteins form dimers with wild type T-box proteins, causing them to form abnormal aggregates in the nucleus.

## 6.2. The crystallization of the TBR1 T-box domain

The TBR1 T-box has significant roles in brain development and is a master regulator of the genes implicated in ASDs. It is therefore important to understand how it functions, and the best way to do this is by studying its structure. The results provided excellent insight into the DNA-binding mechanism of the protein. The data from modelling and molecular docking were interpreted with the results from *in vitro* experiments and suggested that the TBR1 T-box domain binds the TBE in a similar manner to the other T-box proteins. There are still several questions about the intimate details of the structure and DNA-binding of the TBR1 T-box domain which can only be answered by XRC. These include the role of conserved and non-conserved residues in DNA-binding and structural stability, whether or not helix $3_{10}$C is an inducible recognition element and the structural interpretation of pathogenic mutations. Ultimately, there is no substitution for reliable and accurate structural data at an atomic resolution and it is absolutely necessary to continue trying to obtain crystals of the TBR1 T-box domain in the presence and absence of the TBE.

The results from the crystallization suggest that the protein is unlikely to crystallize in the absence of DNA. In the presence of DNA, it is clear from the crystal trials that excessive nucleation and a high rate of crystal formation hamper successful crystallization. After several rounds of optimization, very thin plate-like crystals were obtained. These crystals were too thin and too small to diffract. It is thus unlikely that further optimization of the conditions will lead to diffraction-worthy crystals. The aim in future should therefore either be to use the existent plate-like crystals to get bigger crystals by increasing the third dimension, or to obtain an entirely different crystal lattice with a different space group and morphology, since there is no guarantee that the plate-like crystals will ever grow in the third dimension.

With regards to the TBR1 T-box domain, the most promising result was the formation of very thin plate-like co crystals of the protein-DNA complex. Even though crystals of a similar morphology were diffracted to a resolution of 1.7 Å in the case of TBX3, the crystals did not yield any significant diffraction. Previously, the epitaxial growth of two-dimensional (2D) plate-like crystals on lipid layers has been shown to give rise to three-dimensional (3D) crystals suitable for XRC (Darst et al., 1991; Edward et al., 1994). The growth of the 2D crystals is based on interactions with ligands attached to lipids, or through electrostatic interactions. The TBR1 T-box domain can form 2D crystals in solution, so it can probably form 2D crystals

when immobilized onto a lipid layer, where the growth will be more ordered. The TBR1 T-box domain can be linked to streptavidin via the Strep-tag II, and the streptavidin can then be linked to a biotinylated lipid layer attached to a coverslip of a hanging-drop setup. Instead of using ligands, the protein can be immobilized via an electrostatic interaction. The TBR1 T-box domain is positively charged at a pH of 7 and will thus be attracted to a negatively charged phospholipid layer. The reason why a lipid-immobilized 2D crystal is better than a 2D crystal in solution is that such a crystal will be forced to grow in the third dimension since the growth in the first two dimensions is inhibited by the dimensions of the lipid layer beneath it. This will allow for the formation of a 3D crystal which will hopefully yield significant diffraction.

The sequence of the protein is the most important variable to consider when crystallizing a protein, since it determines the crystal contacts that will be formed. A different crystal lattice can be obtained by altering the sequence of the protein with the aim of promoting the formation of new and improved crystal contacts. Care should be taken so as not to interfere with the DNA-binding function of the TF. The propensity of protein molecules to associate and form crystal contacts is highly dependent on the physical chemistry and topology of the molecular surface. The presence of long and flexible side chains on the surface of the molecule presents an entropic impediment to crystallization by hindering the formation of crystal contacts (Derewenda, 2010; Price and Nagai, 1995). Furthermore, TFs typically have highly disordered N and C termini which could essentially generate structural heterogeneity and thus impede the formation of crystal contacts. The surface entropy can be reduced effectively through limited proteolysis in which flexible and unstructured regions of the protein are selectively removed to promote the formation of crystal contacts (Derewenda, 2010; McPherson and Cudney, 2014; Price and Nagai, 1995; Rupp, 2013). The predicted model and disorder predictions (***Figure 28***, ***Figure 29,*** and ***Figure 31***) show that the TBR1 T-box domain used in this work contains three unstructured and flexible regions. The first region is the N terminus which contains a Strep-tag II, His-Tag, and thrombin cleavage site (residues 1 – 30). The second region is one of the loops protruding out from the conserved β-barrel (residues 174 – 178 (residues 347 – 350 in the original sequence)). The third region is the C terminus which contains the end of helix $3_{10}$C (residues 222 – 225 (residues 395 – 398 in the original sequence). The first region is a good candidate for limited proteolysis since these residues were only added to facilitate the purification of the protein and can be removed

without fearing that the DNA-binding function of the protein will be affected. According to the predicted model (***Figure 28*** and ***Figure 29***), the His-tag forms an alpha helix while the Strep-Tag and thrombin cleavage site form long and flexible surface-exposed moieties. It thus makes sense to remove the thrombin cleavage site and Strep-tag II so that it does not interfere with the formation of stable crystal contacts. The second and third regions should not be removed by proteolysis because they have functional significance. The protein was rendered highly insoluble after the tags were cleaved off. The buffers should thus be optimized to improve the solubility of the protein after thrombin cleavage. Alternatively, the purification protocol should be altered such that it does not require the use of extrinsic affinity tags.

One of the most significant variables in the crystallization of a protein-DNA complex is the DNA (Hollis, 2007). The sequence of the nucleotide, its length, and the composition of its ends all influence the formation of crystal contacts. A close inspection of the T-box crystal structures shows that, in most cases, the DNA is involved in the formation of crystal contacts (Coll et al., 2002; El Omari et al., 2012; Liu et al., 2016; Müller and Herrmann, 1997). With regards to the sequence, a 21 base-pair long SSL DNA sequence was used, containing one TBE flanked by six non-specific bases. This sequence was located in the Auts2 promoter and thus has physiological relevance. Traditionally, T-box proteins have been crystallized in complex with a 22 base-pair long palindromic DNA sequence containing two TBEs flanked by two non-specific bases (Coll et al., 2002; El Omari et al., 2012; Liu et al., 2016; Müller and Herrmann, 1997). The relevance of using the palindromic sequence was questioned since no such palindromic sequences have so far been identified *in vivo*. Our results, in conjunction with the T-box crystal structures, suggest that the TBR1 T-box domain binds SSL DNA much in the same way as it does the palindrome, which means that palindromic DNA could be used to obtain a physiologically relevant crystal structure. The TBX5 T-box domain was crystallized in complex with DNA containing one TBE with flanking bases. The sequence used can thus be modified by reducing the number of non-specific bases outside the TBE. This will reduce the entropy of the target molecule and hopefully allow for crystallization to occur. The composition of the 3' and 5' ends are also particularly important because dsDNA oligonucleotides often pack end-to-end within a crystal lattice. Blunt DNA and DNA with non-complementary base-pair overhangs will prevent this sort of packing, while DNA with complementary base-pair

overhangs will promote it through Watson-Crick or Hoogsteen base pairing. In the case of T-box, many of the crystal structures show end-to-end crystal packing and it is therefore a good idea to use DNA with complementary base-pair overhangs for the crystallization of the TBR1 T-box domain. The end-to-end crystal packing of the DNA can be stabilized by covalent crosslinking, via intra-crystal ligation of the DNA (Ward et al., 2022) . In this technique, the terminal 3’ or 5’ phosphates were crosslinked with 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC), resulting in a pseudo-continuous dsDNA helix which increased the probability of forming crystal contacts. The mechanism of EDC is similar to that of DNA ligase (Ward et al., 2022). This technique could be used for the TBR T-box because EDC is water soluble and the molecular docking results show that the terminal bases of SSL DNA are not involved in DNA-binding.

# 7. Conclusion

The TBR1 T-box domain, like other T-box TFs, is a predominantly β-sheeted protein due to the presence of an evolutionarily conserved DBD stabilized by tryptophan residues. The protein binds specifically to SSL DNA containing one TBE with an affinity in the nanomolar range (($K_D$ = 179.6 nM). The secondary and tertiary structures of the protein, as well as its thermal stability, are conserved upon DNA-binding to SSL DNA. The TBR1 T-box domain uses the same DNA-binding mechanism utilized by the TBX5 T-box domain. The protein contacts the DNA in the minor groove by insertion of helix $3_{10}$C, which only becomes structured upon DNA-binding. Residue F215 forms multiple hydrophobic interactions directly with the base pairs in the TBE, causing the minor groove to become splayed open. The specificity with which the protein binds the DNA is a result of the enthalpically favourable hydrogen bonds formed upon DNA binding. These bonds are formed between residues protruding from the loops extending down from the β-barrel, and the backbone of the DNA in the minor groove. The high affinity with which the protein binds the DNA is a result of the entropically favourable electrostatic interactions formed upon DNA-binding. Under reducing conditions, the TBR1 T-box domain is monomeric in solution and binds SSL DNA as a monomer. The results were consistent with previous findings. The DNA-binding mechanism used by T-box proteins is likely conserved due to the presence of conserved structural elements in the protein-DNA interface. The protein could not be crystallized in the presence or absence of SSL DNA due to the formation of an unfavourable plate-like crystal form. In future, the sequence of the protein and/or the DNA should be engineered with the aim of obtaining new and improved crystal contacts.

# 8. References


Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2002. An Overview of Gene Control. Mol. Biol. Cell 4th Ed.

Anthis, N.J., Clore, G.M., 2013. Sequence-specific determination of protein and peptide concentrations by absorbance at 205 nm. Protein Sci. Publ. Protein Soc. 22, 851–858. https://doi.org/10.1002/pro.2253

Babu, M.M., 2016. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. Biochem. Soc. Trans. 44, 1185–1200. https://doi.org/10.1042/BST20160172

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., Millán, C., Park, H., Adams, C., Glassman, C.R., DeGiovanni, A., Pereira, J.H., Rodrigues, A.V., van Dijk, A.A., Ebrecht, A.C., Opperman, D.J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M.K., Dalwadi, U., Yip, C.K., Burke, J.E., Garcia, K.C., Grishin, N.V., Adams, P.D., Read, R.J., Baker, D., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876. https://doi.org/10.1126/science.abj8754

Benayoun, B.A., Veitia, R.A., 2009. A post-translational modification code for transcription factors: sorting through a sea of signals. Trends Cell Biol. 19, 189–197. https://doi.org/10.1016/j.tcb.2009.02.003

Berg, J.M., Tymoczko, J.L., Stryer, L., 2002. Transcriptional Activation and Repression Are Mediated by Protein-Protein Interactions. Biochem. 5th Ed.

Blane, A.A., 2021. The FOXP2-TBR1 interaction and its role in the regulation of DNA binding. PhD Thesis, University of the Witwatersrand, South Africa.

Block, H., Maertens, B., Spriestersbach, A., Brinker, N., Kubicek, J., Fabis, R., Labahn, J., Schäfer, F., 2009. Chapter 27 Immobilized-Metal Affinity Chromatography (IMAC): A Review, in: Burgess, R.R., Deutscher, M.P. (Eds.), Methods in Enzymology, Guide to Protein Purification, 2nd Edition. Academic Press, pp. 439–473. https://doi.org/10.1016/S0076-6879(09)63027-5

Bornhorst, J.A., Falke, J.J., 2000. [16] Purification of proteins using polyhistidine affinity tags, in: Methods in Enzymology. Elsevier, pp. 245–254. https://doi.org/10.1016/S0076-6879(00)26058-8

Bulfone, A., Smiga, S.M., Shimamura, K., Peterson, A., Puelles, L., Rubenstein, J.L., 1995. T-brain-1: a homolog of Brachyury whose expression defines molecularly distinct domains within the cerebral cortex. Neuron 15, 63–78. https://doi.org/10.1016/0896-6273(95)90065-9

Buratowski, S., Hahn, S., Sharp, P.A., Guarente, L., 1988. Function of a yeast TATA element-binding protein in a mammalian transcription system. Nature 334, 37–42. https://doi.org/10.1038/334037a0

Burkhard, P., Tai, C.-H., Ristroph, C.M., Cook, P.F., Jansonius, J.N., 1999. Ligand binding induces a large conformational change in O-acetylserine sulfhydrylase from Salmonella

typhimurium11Edited by R. Huber. J. Mol. Biol. 291, 941–953. https://doi.org/10.1006/jmbi.1999.3002

Casamassimi, A., Ciccodicola, A., 2019. Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation. Int. J. Mol. Sci. 20, 1281. https://doi.org/10.3390/ijms20061281

Chavda, H., Patel, C., 2011. Effect of crosslinker concentration on characteristics of superporous hydrogel. Int. J. Pharm. Investig. 1, 17–21. https://doi.org/10.4103/2230-973X.76724

Chayen, N.E., 1997. A novel technique to control the rate of vapour diffusion, giving larger protein crystals. J. Appl. Crystallogr. 30, 198–202. https://doi.org/10.1107/S0021889896013532

Chayen, N.E., Shaw Stewart, P.D., Blow, D.M., 1992. Microbatch crystallization under oil — a new technique allowing many small-volume crystallization trials. J. Cryst. Growth 122, 176–180. https://doi.org/10.1016/0022-0248(92)90241-A

Chruszcz, M., Potrzebowski, W., Zimmerman, M.D., Grabowski, M., Zheng, H., Lasota, P., Minor, W., 2008. Analysis of solvent content and oligomeric states in protein crystals—does symmetry matter? Protein Sci. Publ. Protein Soc. 17, 623–632. https://doi.org/10.1110/ps.073360508

Coll, M., Seidman, J.G., Müller, C.W., 2002. Structure of the DNA-Bound T-Box Domain of Human TBX3, a Transcription Factor Responsible for Ulnar-Mammary Syndrome. Structure 10, 343–356. https://doi.org/10.1016/S0969-2126(02)00722-0

Courey, A.J., Holtzman, D.A., Jackson, S.P., Tjian, R., 1989. Synergistic activation by the glutamine-rich domains of human transcription factor Sp1. Cell 59, 827–836. https://doi.org/10.1016/0092-8674(89)90606-5

Crick, F., 1970. Central Dogma of Molecular Biology. Nature 227, 561–563. https://doi.org/10.1038/227561a0

Crick, F.H., 1958. On protein synthesis. Symp. Soc. Exp. Biol. 12, 138–163.

Darst, S.A., Kubalek, E.W., Edwards, A.M., Kornberg, R.D., 1991. Two-dimensional and epitaxial crystallization of a mutant form of yeast RNA polymerase II. J. Mol. Biol. 221, 347–357. https://doi.org/10.1016/0022-2836(91)80223-h

Davidson, E., Levin, M., 2005. Gene regulatory networks. Proc. Natl. Acad. Sci. U. S. A. 102, 4935. https://doi.org/10.1073/pnas.0502024102

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., Singh, T., Klei, L., Kosmicki, J., Fu, S.-C., Aleksic, B., Biscaldi, M., Bolton, P.F., Brownfeld, J.M., Cai, J., Campbell, N.G., Carracedo, A., Chahrour, M.H., Chiocchetti, A.G., Coon, H., Crawford, E.L., Crooks, L., Curran, S.R., Dawson, G., Duketis, E., Fernandez, B.A., Gallagher, L., Geller, E., Guter, S.J., Sean Hill, R., Ionita-Laza, I., Jimenez Gonzalez, P., Kilpinen, H., Klauck, S.M., Kolevzon, A., Lee, I., Lei, J., Lehtimäki, T., Lin, C.-F., Ma'ayan, A., Marshall, C.R., McInnes, A.L., Neale, B., Owen, M.J., Ozaki, N., Parellada, M., Parr, J.R., Purcell, S., Puura, K., Rajagopalan, D., Rehnström, K., Reichenberg,

A., Sabo, A., Sachse, M., Sanders, S.J., Schafer, C., Schulte-Rüther, M., Skuse, D., Stevens, C., Szatmari, P., Tammimies, K., Valladares, O., Voran, A., Wang, L.-S., Weiss, L.A., Jeremy Willsey, A., Yu, T.W., Yuen, R.K.C., Cook, E.H., Freitag, C.M., Gill, M., Hultman, C.M., Lehner, T., Palotie, A., Schellenberg, G.D., Sklar, P., State, M.W., Sutcliffe, J.S., Walsh, C.A., Scherer, S.W., Zwick, M.E., Barrett, J.C., Cutler, D.J., Roeder, K., Devlin, B., Daly, M.J., Buxbaum, J.D., 2014. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209–215. https://doi.org/10.1038/nature13772

DeLucas, L.J., Bugg, C.E., 1987. New directions in protein crystal growth. Trends Biotechnol. 5, 188–193. https://doi.org/10.1016/S0167-7799(87)80006-9

Derewenda, Z.S., 2010. Application of protein engineering to enhance crystallizability and improve crystal properties. Acta Crystallogr. D Biol. Crystallogr. 66, 604–615. https://doi.org/10.1107/S090744491000644X

Deriziotis, P., O'Roak, B.J., Graham, S.A., Estruch, S.B., Dimitropoulou, D., Bernier, R.A., Gerdts, J., Shendure, J., Eichler, E.E., Fisher, S.E., 2014a. De novo TBR1 mutations in sporadic autism disrupt protein functions. Nat. Commun. 5, 4954. https://doi.org/10.1038/ncomms5954

Edward, A.M., Darst, S.A., Hemming, S.A., Li, Y., Kornberg, R.D., 1994. Epitaxial growth of protein crystals on lipid layers. Nat. Struct. Biol. 1, 195–197. https://doi.org/10.1038/nsb0394-195

Eftink, M.R., 2000. Intrinsic Fluorescence of Proteins, in: Lakowicz, J.R. (Ed.), Topics in Fluorescence Spectroscopy: Volume 6: Protein Fluorescence, Topics in Fluorescence Spectroscopy. Springer US, Boston, MA, pp. 1–15. https://doi.org/10.1007/0-306-47102-7_1

El Omari, K., De Mesmaeker, J., Karia, D., Ginn, H., Bhattacharya, S., Mancini, E.J., 2012. Structure of the DNA-bound T-box domain of human TBX1, a transcription factor associated with the DiGeorge syndrome. Proteins Struct. Funct. Bioinforma. 80, 655–660. https://doi.org/10.1002/prot.23208

Favicchio, R., Dragan, A.I., Kneale, G.G., Read, C.M., 2009a. Fluorescence spectroscopy and anisotropy in the analysis of DNA-protein interactions. Methods Mol. Biol. Clifton NJ 543, 589–611. https://doi.org/10.1007/978-1-60327-015-1_35

Favicchio, R., Dragan, A.I., Kneale, G.G., Read, C.M., 2009b. Fluorescence spectroscopy and anisotropy in the analysis of DNA-protein interactions. Methods Mol. Biol. Clifton NJ 543, 589–611. https://doi.org/10.1007/978-1-60327-015-1_35

Frietze, S., Farnham, P.J., 2011. Transcription Factor Effector Domains, in: Hughes, T.R. (Ed.), A Handbook of Transcription Factors. Springer Netherlands, Dordrecht, pp. 261–277. https://doi.org/10.1007/978-90-481-9069-0_12

Gaberc-Porekar, V., Menart, V., 2001. Perspectives of immobilized-metal affinity chromatography. J. Biochem. Biophys. Methods 49, 335–360. https://doi.org/10.1016/S0165-022X(01)00207-X

Garvie, C.W., Wolberger, C., 2001. Recognition of Specific DNA Sequences. Mol. Cell 8, 937–946. https://doi.org/10.1016/S1097-2765(01)00392-6

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A., 2005. Protein Identification and Analysis Tools on the ExPASy Server, in: Walker, J.M. (Ed.), The Proteomics Protocols Handbook, Springer Protocols Handbooks. Humana Press, Totowa, NJ, pp. 571–607. https://doi.org/10.1385/1-59259-890-0:571

Geisel, N., 2011. Constitutive versus Responsive Gene Expression Strategies for Growth in Changing Environments. PLoS ONE 6, e27033. https://doi.org/10.1371/journal.pone.0027033

Georges, A.B., Benayoun, B.A., Caburet, S., Veitia, R.A., 2010. Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from? FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. 24, 346–356. https://doi.org/10.1096/fj.09-142117

Gijsbers, A., Nishigaki, T., Sánchez-Puig, N., 2016. Fluorescence Anisotropy as a Tool to Study Protein-protein Interactions. J. Vis. Exp. JoVE. https://doi.org/10.3791/54640

Gilad, Y., Rifkin, S.A., Pritchard, J.K., 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. TIG 24, 408–415. https://doi.org/10.1016/j.tig.2008.06.001

Glass, C.K., Rosenfeld, M.G., 2000. The coregulator exchange in transcriptional functions of nuclear receptors. Genes Dev. 14, 121–141. https://doi.org/10.1101/gad.14.2.121

Graham, S.A., Fisher, S.E., 2013. Decoding the genetics of speech and language. Curr. Opin. Neurobiol., Neurogenetics 23, 43–51. https://doi.org/10.1016/j.conb.2012.11.006

Greenfield, N.J., 2006a. Using circular dichroism spectra to estimate protein secondary structure. Nat. Protoc. 1, 2876–2890. https://doi.org/10.1038/nprot.2006.202

Greenfield, N.J., 2006b. Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. Nat. Protoc. 1, 2527–2535. https://doi.org/10.1038/nprot.2006.204

Greenfield, N.J., 1999. Applications of circular dichroism in protein and peptide analysis. TrAC Trends Anal. Chem. 18, 236–244. https://doi.org/10.1016/S0165-9936(98)00112-5

Grubisic, Z., Rempp, P., Benoit, H., 1967. A universal calibration for gel permeation chromatography. J. Polym. Sci. [B] 5, 753–759. https://doi.org/10.1002/pol.1967.110050903

Hall, J.E., Guyton, A.C., 2011. Guyton and Hall textbook of medical physiology, 12th ed. ed. Saunders/Elsevier, Philadelphia, Pa.

Hameduh, T., Haddad, Y., Adam, V., Heger, Z., 2020. Homology modeling in the time of collective and artificial intelligence. Comput. Struct. Biotechnol. J. 18, 3494–3506. https://doi.org/10.1016/j.csbj.2020.11.007

Han, W., Kwan, K.Y., Shim, S., Lam, M.M.S., Shin, Y., Xu, X., Zhu, Y., Li, M., Šestan, N., 2011. TBR1 directly represses Fezf2 to control the laminar origin and development of the corticospinal tract. Proc. Natl. Acad. Sci. 108, 3041–3046. https://doi.org/10.1073/pnas.1016723108

Hellman, L.M., Fried, M.G., 2007. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. Nat. Protoc. 2, 1849–1861. https://doi.org/10.1038/nprot.2007.249

Hermsen, R., Ursem, B., Wolde, P.R., 2010. Combinatorial Gene Regulation Using Auto-Regulation. PLoS Comput. Biol. 6, e1000813. https://doi.org/10.1371/journal.pcbi.1000813

Hickok, G., 2009. The functional neuroanatomy of language. Phys. Life Rev. 6, 121–143. https://doi.org/10.1016/j.plrev.2009.06.001

Hoed, J., Sollis, E., Venselaar, H., Estruch, S.B., Deriziotis, P., Fisher, S.E., 2018. Functional characterization of TBR1 variants in neurodevelopmental disorder. Sci. Rep. 8, 14279. https://doi.org/10.1038/s41598-018-32053-6

Hollenberg, S.M., Evans, R.M., 1988. Multiple and cooperative trans-activation domains of the human glucocorticoid receptor. Cell 55, 899–906. https://doi.org/10.1016/0092-8674(88)90145-6

Hollis, T., 2007. Crystallization of protein-DNA complexes. Methods Mol. Biol. Clifton NJ 363, 225–237. https://doi.org/10.1007/978-1-59745-209-0_11

Hong, P., Koza, S., Bouvier, E.S.P., 2012. A review of size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates. J. Liq. Chromatogr. Relat. Technol. 35, 2923–2950. https://doi.org/10.1080/10826076.2012.743724

Horstkorte, R., Fuss, B., 2012. Chapter 9 - Cell Adhesion Molecules, in: Brady, S.T., Siegel, G.J., Albers, R.W., Price, D.L. (Eds.), Basic Neurochemistry (Eighth Edition). Academic Press, New York, pp. 165–179. https://doi.org/10.1016/B978-0-12-374947-5.00009-2

Hsueh, Y.-P., Wang, T.-F., Yang, F.-C., Sheng, M., 2000a. Nuclear translocation and transcription regulation by the membrane-associated guanylate kinase CASK/LIN-2. Nature 404, 298–302. https://doi.org/10.1038/35005118

Huang, T.-N., Hsueh, Y.-P., 2015. Brain-specific transcriptional regulator T-brain-1 controls brain wiring and neuronal activity in autism spectrum disorders. Front. Neurosci. 9. https://doi.org/10.3389/fnins.2015.00406

Ireland, S.M., Sula, A., Wallace, B. a., 2018. Thermal melt circular dichroism spectroscopic studies for identifying stabilising amphipathic molecules for the voltage-gated sodium channel NavMs. Biopolymers 109, e23067. https://doi.org/10.1002/bip.23067

Ishida, T., Kinoshita, K., 2007. PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res. 35, W460-464. https://doi.org/10.1093/nar/gkm363

Jeong, H., Barbe, V., Lee, C.H., Vallenet, D., Yu, D.S., Choi, S.-H., Couloux, A., Lee, S.-W., Yoon, S.H., Cattolico, L., Hur, C.-G., Park, H.-S., Ségurens, B., Kim, S.C., Oh, T.K., Lenski, R.E., Studier, F.W., Daegelen, P., Kim, J.F., 2009. Genome Sequences of Escherichia coli B strains REL606 and BL21(DE3). J. Mol. Biol. 394, 644–652. https://doi.org/10.1016/j.jmb.2009.09.052

Jordan, S.R., Whitcombe, T.V., Berg, J.M., Pabo, C.O., 1985. Systematic variation in DNA length yields highly ordered repressor-operator cocrystals. Science 230, 1383–1385. https://doi.org/10.1126/science.3906896

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J.,

Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Karin, M., 1990. Too many transcription factors: positive and negative interactions. New Biol. 2, 126–131.

Kastenholz, M.A., Schwartz, T.U., Hünenberger, P.H., 2006. The Transition between the B and Z Conformations of DNA Investigated by Targeted Molecular Dynamics Simulations with Explicit Solvation. Biophys. J. 91, 2976–2990. https://doi.org/10.1529/biophysj.106.083667

Kirby, A.J., 2001. The lysozyme mechanism sorted — after 50 years. Nat. Struct. Biol. 8, 737–739. https://doi.org/10.1038/nsb0901-737

Krishna, S.S., Majumdar, I., Grishin, N.V., 2003. SURVEY AND SUMMARY: Structural classification of zinc fingers. Nucleic Acids Res. 31, 532–550.

Kulkarni, M., Mukherjee, A., 2017. Understanding B-DNA to A-DNA transition in the right-handed DNA helix: Perspective from a local to global transition. Prog. Biophys. Mol. Biol., Exploring mechanisms in biology: simulations and experiments come together 128, 63–73. https://doi.org/10.1016/j.pbiomolbio.2017.05.009

Laemmli, U.K., 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680–685. https://doi.org/10.1038/227680a0

Lakowicz, J.R. (Ed.), 2006a. Introduction to Fluorescence, in: Principles of Fluorescence Spectroscopy. Springer US, Boston, MA, pp. 1–26. https://doi.org/10.1007/978-0-387-46312-4_1

Lakowicz, J.R. (Ed.), 2006b. Fluorophores, in: Principles of Fluorescence Spectroscopy. Springer US, Boston, MA, pp. 63–95. https://doi.org/10.1007/978-0-387-46312-4_3

Lakowicz, J.R. (Ed.), 2006c. Quenching of Fluorescence, in: Principles of Fluorescence Spectroscopy. Springer US, Boston, MA, pp. 277–330. https://doi.org/10.1007/978-0-387-46312-4_8

Lakowicz, J.R. (Ed.), 2006d. Protein Fluorescence, in: Principles of Fluorescence Spectroscopy. Springer US, Boston, MA, pp. 529–575. https://doi.org/10.1007/978-0-387-46312-4_16

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., Weirauch, M.T., 2018. The Human Transcription Factors. Cell 172, 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Latchman, D., 2005. Gene Regulation, 5th ed. Taylor & Francis, London. https://doi.org/10.4324/9780203016336

Latchman, D.S., 2020. Gene Control, 2nd ed. Garland Science, New York. https://doi.org/10.1201/9781317407751

Latchman, D.S., 2011. Transcriptional Gene Regulation in Eukaryotes, in: ELS. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0002322.pub2

Latchman, D.S., 2010. Gene Control, 2nd ed. Garland Science.

Latchman, D.S., 1997. Transcription factors: An overview. Int. J. Biochem. Cell Biol. 29, 1305–1312. https://doi.org/10.1016/S1357-2725(97)00085-X

Latchman, D.S., 1996. Inhibitory transcription factors. Int. J. Biochem. Cell Biol. 28, 965–974. https://doi.org/10.1016/1357-2725(96)00039-8

Latchman, D.S., 1990. Eukaryotic transcription factors. Biochem. J. 270, 281–289.

Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D., Zakrzewska, K., 2009. Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Res. 37, 5917–5929. https://doi.org/10.1093/nar/gkp608

Lazarus, R.A., Wagener†, J.S., 2019. Recombinant Human Deoxyribonuclease I. Pharm. Biotechnol. 471–488. https://doi.org/10.1007/978-3-030-00710-2_22

Leon, S.B.-T., Davidson, E.H., 2007. Gene Regulation: Gene Control Network in Development. Annu. Rev. Biophys. Biomol. Struct. 36, 191–212. https://doi.org/10.1146/annurev.biophys.35.040405.102002

Levine, M., 2010. Transcriptional enhancers in animal development and evolution. Curr. Biol. CB 20, R754-763. https://doi.org/10.1016/j.cub.2010.06.070

Levine, M., Tjian, R., 2003. Transcription regulation and animal diversity. Nature 424, 147–151. https://doi.org/10.1038/nature01763

Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., Barrett, L.F., 2012. The brain basis of emotion: A meta-analytic review. Behav. Brain Sci. 35, 121–143. https://doi.org/10.1017/S0140525X11000446

Liu, C.F., Brandt, G.S., Hoang, Q.Q., Naumova, N., Lazarevic, V., Hwang, E.S., Dekker, J., Glimcher, L.H., Ringe, D., Petsko, G.A., 2016a. Crystal structure of the DNA binding domain of the transcription factor T-bet suggests simultaneous recognition of distant genome sites. Proc. Natl. Acad. Sci. 113, E6572–E6581. https://doi.org/10.1073/pnas.1613914113

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J., 2000. Structure of Nucleic Acids. Mol. Cell Biol. 4th Ed.

Luisi, B.F., Xu, W.X., Otwinowski, Z., Freedman, L.P., Yamamoto, K.R., Sigler, P.B., 1991. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. Nature 352, 497–505. https://doi.org/10.1038/352497a0

Luse, D.S., 2013. The RNA polymerase II preinitiation complex. Transcription 5, e27050. https://doi.org/10.4161/trns.27050

Mariani, V., Biasini, M., Barbato, A., Schwede, T., 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29, 2722–2728. https://doi.org/10.1093/bioinformatics/btt473

Matthews, B.W., 1968. Solvent content of protein crystals. J. Mol. Biol. 33, 491–497. https://doi.org/10.1016/0022-2836(68)90205-2

Matthews, B.W., Ohlendorf, D.H., Anderson, W.F., Takeda, Y., 1982. Structure of the DNA-binding region of lac repressor inferred from its homology with cro repressor. Proc. Natl. Acad. Sci. U. S. A. 79, 1428–1432. https://doi.org/10.1073/pnas.79.5.1428

Mauro, V.P., Chappell, S.A., 2014. A critical analysis of codon optimization in human therapeutics. Trends Mol. Med. 20, 604–613. https://doi.org/10.1016/j.molmed.2014.09.003

McPherson, A., Cudney, B., 2014. Optimization of crystallization conditions for biological macromolecules. Acta Crystallogr. Sect. F Struct. Biol. Commun. 70, 1445–1467. https://doi.org/10.1107/S2053230X14019670

McPherson, A., Gavira, J.A., 2014. Introduction to protein crystallization. Acta Crystallogr. Sect. F Struct. Biol. Commun. 70, 2–20. https://doi.org/10.1107/S2053230X13033141

McRee, D.E., David, P.R., 1999. Practical Protein Crystallography. Elsevier Science.

Mendoza, A. de, Sebé-Pedrós, A., Šestak, M.S., Matejčić, M., Torruella, G., Domazet-Lošo, T., Ruiz-Trillo, I., 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc. Natl. Acad. Sci. 110, E4858–E4866. https://doi.org/10.1073/pnas.1311818110

Mermod, N., O'Neill, E.A., Kelly, T.J., Tjian, R., 1989. The proline-rich transcriptional activator of CTF/NF-I is distinct from the replication and DNA binding domain. Cell 58, 741–753. https://doi.org/10.1016/0092-8674(89)90108-6

Miller, J.N., 1981. General considerations on fluorescence spectrometry, in: Miller, J.N. (Ed.), Standards in Flourescence Spectrometry: Ultraviolet Spectrometry Group, Techniques in Visible and Ultraviolet Spectometry. Springer Netherlands, Dordrecht, pp. 1–7. https://doi.org/10.1007/978-94-009-5902-6_1

Mitchell, P.J., Tjian, R., 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science 245, 371–378. https://doi.org/10.1126/science.2667136

Mizuguchi, R., Naritsuka, H., Mori, K., Yoshihara, Y., 2012. Tbr2 Deficiency in Mitral and Tufted Cells Disrupts Excitatory-Inhibitory Balance of Neural Circuitry in the Mouse Olfactory Bulb. J. Neurosci. 32, 8831–8844. https://doi.org/10.1523/JNEUROSCI.5746-11.2012

Moerke, N.J., 2009. Fluorescence Polarization (FP) Assays for Monitoring Peptide-Protein or Nucleic Acid-Protein Binding. Curr. Protoc. Chem. Biol. 1, 1–15. https://doi.org/10.1002/9780470559277.ch090102

Moll, J.R., Acharya, A., Gal, J., Mir, A.A., Vinson, C., 2002. Magnesium is required for specific DNA binding of the CREB B-ZIP domain. Nucleic Acids Res. 30, 1240–1246.

Mortillaro, M.J., Blencowe, B.J., Wei, X., Nakayasu, H., Du, L., Warren, S.L., Sharp, P.A., Berezney, R., 1996. A hyperphosphorylated form of the large subunit of RNA polymerase II is associated with splicing complexes and the nuclear matrix. Proc. Natl. Acad. Sci. U. S. A. 93, 8253–8257.

Müller, C.W., Herrmann, B.G., 1997. Crystallographic structure of the T domain–DNA complex of the Brachyury transcription factor. Nature 389, 884–888. https://doi.org/10.1038/39929

Nestler, E.J., Hyman, S.E., Holtzman, D.M., Malenka, R.C., 2015. Higher Cognitive Function and Behavioral Control, in: Molecular Neuropharmacology: A Foundation for Clinical Neuroscience. McGraw-Hill Education, New York, NY.

Neu, H.C., 1969. Effect of β-Lactamase Location in Escherichia coli on Penicillin Synergy. Appl. Microbiol. 17, 783–786.

Olson, N.D., Morrow, J.B., 2012. DNA extract characterization process for microbial detection methods development and validation. BMC Res. Notes 5, 668. https://doi.org/10.1186/1756-0500-5-668

O'Roak, B.J., Stessman, H.A., Boyle, E.A., Witherspoon, K.T., Martin, B., Lee, C., Vives, L., Baker, C., Hiatt, J.B., Nickerson, D.A., Bernier, R., Shendure, J., Eichler, E.E., 2014a. Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat. Commun. 5, 5595. https://doi.org/10.1038/ncomms6595

O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J.B., Turner, E.H., Levy, R., O'Day, D.R., Krumm, N., Coe, B.P., Martin, B.K., Borenstein, E., Nickerson, D.A., Mefford, H.C., Doherty, D., Akey, J.M., Bernier, R., Eichler, E.E., Shendure, J., 2012a. Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. Science 338, 1619–1622. https://doi.org/10.1126/science.1227764

Osborne, C.K., Schiff, R., Fuqua, S.A., Shou, J., 2001. Estrogen receptor: current understanding of its activation and modulation. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 7, 4338s–4342s; discussion 4411s-4412s.

Pabo, C.O., Sauer, R.T., 1984. Protein-DNA recognition. Annu. Rev. Biochem. 53, 293–321. https://doi.org/10.1146/annurev.bi.53.070184.001453

Pace, C.N., Scholtz, J.M., Grimsley, G.R., 2014. Forces Stabilizing Proteins. FEBS Lett. 588, 2177–2184. https://doi.org/10.1016/j.febslet.2014.05.006

Pan, Y., Tsai, C.-J., Ma, B., Nussinov, R., 2010. Mechanisms of transcription factor selectivity. Trends Genet. TIG 26, 75–83. https://doi.org/10.1016/j.tig.2009.12.003

Papaioannou, V.E., 2014. The T-box gene family: emerging roles in development, stem cells and cancer. Development 141, 3819–3833. https://doi.org/10.1242/dev.104471

Paxton, C., Zhao, H., Chin, Y., Langner, K., Reecy, J., 2002. Murine Tbx2 contains domains that activate and repress gene transcription. Gene 283, 117–124. https://doi.org/10.1016/s0378-1119(01)00878-2

Pearce, R., Zhang, Y., 2021. Deep learning techniques have significantly impacted protein structure prediction and protein design. Curr. Opin. Struct. Biol., Protein-Carbohydrate Complexes and Glycosylation ● Sequences and Topology 68, 194–207. https://doi.org/10.1016/j.sbi.2021.01.007

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera: A visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612. https://doi.org/10.1002/jcc.20084

Pierce, B.A., 2012. Genetics: a conceptual approach, 4th ed. ed. W.H. Freeman, New York.

Ponomarenko, E.A., Poverennaya, E.V., Ilgisonis, E.V., Pyatnitskiy, M.A., Kopylov, A.T., Zgoda, V.G., Lisitsa, A.V., Archakov, A.I., 2016. The Size of the Human Proteome: The Width and Depth. Int. J. Anal. Chem. 2016, 1–6. https://doi.org/10.1155/2016/7436849

Price, S.R., Nagai, K., 1995. Protein engineering as a tool for crystallography. Curr. Opin. Biotechnol. 6, 425–430. https://doi.org/10.1016/0958-1669(95)80072-7

Ranish, J.A., Yudkovsky, N., Hahn, S., 1999. Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB. Genes Dev. 13, 49–63.

Rhodes, G., 2006. Crystallography Made Crystal Clear. Elsevier. https://doi.org/10.1016/B978-0-12-587073-3.X5000-4

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., Mann, R.S., 2010. Origins of Specificity in Protein-DNA Recognition. Annu. Rev. Biochem. 79, 233–269. https://doi.org/10.1146/annurev-biochem-060408-091030

Rong, M., He, B., McAllister, W.T., Durbin, R.K., 1998. Promoter specificity determinants of T7 RNA polymerase. Proc. Natl. Acad. Sci. 95, 515–519. https://doi.org/10.1073/pnas.95.2.515

Rosano, G.L., Ceccarelli, E.A., 2014. Recombinant protein expression in Escherichia coli: advances and challenges. Front. Microbiol. 5, 172. https://doi.org/10.3389/fmicb.2014.00172

Rudner, R., Karkas, J.D., Chargaff, E., 1968. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. Proc. Natl. Acad. Sci. 60, 921–922. https://doi.org/10.1073/pnas.60.3.921

Rupp, B., 2013. Macromolecular Crystallography: Overview, in: Roberts, G.C.K. (Ed.), Encyclopedia of Biophysics. Springer, Berlin, Heidelberg, pp. 1346–1353. https://doi.org/10.1007/978-3-642-16712-6_655

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., Walker, M.F., Ober, G.T., Teran, N.A., Song, Y., El-Fishawy, P., Murtha, R.C., Choi, M., Overton, J.D., Bjornson, R.D., Carriero, N.J., Meyer, K.A., Bilguvar, K., Mane, S.M., Šestan, N., Lifton, R.P., Günel, M., Roeder, K., Geschwind, D.H., Devlin, B., State, M.W., 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485, 237–241. https://doi.org/10.1038/nature10945

Schaefer, U., Schmeier, S., Bajic, V.B., 2011. TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. Nucleic Acids Res. 39, D106–D110. https://doi.org/10.1093/nar/gkq945

Schmid, F.-X., 2001. Biological Macromolecules: UV-visible Spectrophotometry, in: ELS. John Wiley & Sons, Ltd. https://doi.org/10.1038/npg.els.0003142

Schnepf, M., von Reutern, M., Ludwig, C., Jung, C., Gaul, U., 2020. Transcription Factor Binding Affinities and DNA Shape Readout. iScience 23, 101694. https://doi.org/10.1016/j.isci.2020.101694

Schumacher, M.A., Goodman, R.H., Brennan, R.G., 2000. The Structure of a CREB bZIP·Somatostatin CRE Complex Reveals the Basis for Selective Dimerization and Divalent Cation-enhanced DNA Binding. J. Biol. Chem. 275, 35242–35247. https://doi.org/10.1074/jbc.M007293200

Schweitzer, B.A., Kool, E.T., 1995. Hydrophobic, Non-Hydrogen-Bonding Bases and Base Pairs in DNA. J. Am. Chem. Soc. 117, 1863–1872. https://doi.org/10.1021/ja00112a001

Skolnick, J., Gao, M., Zhou, H., Singh, S., 2021. AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. J. Chem. Inf. Model. 61, 4827–4831. https://doi.org/10.1021/acs.jcim.1c01114

Smith, B.J., 1984. SDS Polyacrylamide Gel Electrophoresis of Proteins, in: Proteins. Humana Press, New Jersey, pp. 41–56. https://doi.org/10.1385/0-89603-062-8:41

Smyth, M.S., Martin, J.H., 2000. x ray crystallography. Mol. Pathol. MP 53, 8–14. https://doi.org/10.1136/mp.53.1.8

Sperling, S., 2007. Transcriptional regulation at a glance. BMC Bioinformatics 8, S2. https://doi.org/10.1186/1471-2105-8-S6-S2

Stirnimann, C.U., Ptchelkine, D., Grimm, C., Müller, C.W., 2010. Structural Basis of TBX5–DNA Recognition: The T-Box Domain in Its DNA-Bound and -Unbound Form. J. Mol. Biol. 400, 71–81. https://doi.org/10.1016/j.jmb.2010.04.052

Teif, V.B., 2010. Predicting Gene-Regulation Functions: Lessons from Temperate Bacteriophages. Biophys. J. 98, 1247–1256. https://doi.org/10.1016/j.bpj.2009.11.046

Theriot, J.A., 2013. Why are bacteria different from eukaryotes? BMC Biol. 11, 119. https://doi.org/10.1186/1741-7007-11-119

Till, M., Robson, A., Byrne, M.J., Nair, A.V., Kolek, S.A., Shaw Stewart, P.D., Race, P.R., 2013. Improving the Success Rate of Protein Crystallization by Random Microseed Matrix Screening. J. Vis. Exp. JoVE 50548. https://doi.org/10.3791/50548

Travers, A., Muskhelishvili, G., 2015. DNA structure and function. FEBS J. 282, 2279–2295. https://doi.org/10.1111/febs.13307

Uversky, V.N., 2013. The alphabet of intrinsic disorder. Intrinsically Disord. Proteins 1, e24684. https://doi.org/10.4161/idp.24684

Ward, A.R., Dmytriw, S., Vajapayajula, A., Snow, C.D., 2022. Stabilizing DNA–Protein Co-Crystals via Intra-Crystal Chemical Ligation of the DNA. Crystals 12, 49. https://doi.org/10.3390/cryst12010049

Wärnmark, A., Treuter, E., Wright, A.P.H., Gustafsson, J.-Å., 2003. Activation Functions 1 and 2 of Nuclear Receptors: Molecular Strategies for Transcriptional Activation. Mol. Endocrinol. 17, 1901–1909. https://doi.org/10.1210/me.2002-0384

Watts, A., 1993. Crystallization of nucleic acids and proteins. A practical approach: edited by A. Ducruix and R. Giegé, Oxford University Press; Oxford, 1992; xxiv + 331 pages. £25.00. ISBN 019-963246-4. FEBS Lett. 319, 283–284. https://doi.org/10.1016/0014-5793(93)80565-C

Weber, G., 1953. Rotational Brownian motion and polarization of the fluorescence of solutions. Adv. Protein Chem. 8, 415–459. https://doi.org/10.1016/s0065-3233(08)60096-0

William Studier, F., Rosenberg, A.H., Dunn, J.J., Dubendorff, J.W., 1990. [6] Use of T7 RNA polymerase to direct expression of cloned genes, in: Methods in Enzymology. Elsevier, pp. 60–89. https://doi.org/10.1016/0076-6879(90)85008-C

Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., Jain, S., Lewis, S.M., Arendall, W.B., Snoeyink, J., Adams, P.D., Lovell, S.C., Richardson, J.S., Richardson, D.C., 2018. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. Publ. Protein Soc. 27, 293–315. https://doi.org/10.1002/pro.3330

Wilson, V., Conlon, F.L., 2002. The T-box family. Genome Biol. 3, reviews3008.1-reviews3008.7.

Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K., Dickerson, R.E., 1980. Crystal structure analysis of a complete turn of B-DNA. Nature 287, 755–758. https://doi.org/10.1038/287755a0

Woody, R.W., 1995. [4] Circular dichroism, in: Methods in Enzymology. Elsevier, pp. 34–71. https://doi.org/10.1016/0076-6879(95)46006-3

Woody, R.W., Koslowski, A., 2002. Recent developments in the electronic spectroscopy of amides and alpha-helical polypeptides. Biophys. Chem. 101–102, 535–551. https://doi.org/10.1016/s0301-4622(02)00187-4

Xiao, H., Jeang, K.-T., 1998. Glutamine-rich Domains Activate Transcription in Yeast *Saccharomyces cerevisiae*. J. Biol. Chem. 273, 22873–22876. https://doi.org/10.1074/jbc.273.36.22873

Yang, J., Zhang, Y., 2015. Protein Structure and Function Prediction Using I-TASSER. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al 52, 5.8.1-5.815. https://doi.org/10.1002/0471250953.bi0508s52