Probing the Stability and Folding Mechanism of FOXP2-FHD

Ilan Kirkel
563242

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg in fulfilment of the requirements for the degree of Master of Science

August 2020

Declaration

I, Ilan Kirkel (student number: 563242), am a student registered for the degree of Master of Science (MSc) in the academic year 2019.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this, or any other University.
- The information used in the dissertation has not been obtained by me while employed by, or working under the aegis of, any person or organisation other than the University.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: 

Date: 11 August 2020

Abstract

The forkhead box (FOX) family of transcription factors is categorised by the presence of a canonical DNA binding winged helix motif, termed the forkhead domain (FHD). Of the FOX transcription factors, the FHD of the FOXP subfamily has the unique capability to form dimers by three-dimensional domain swapped dimerisation. A definitive role for domain swapping has not been established, although it has been speculated to facilitate interchromosomal association and thus enable coregulation of distal genetic elements. The FHD of FOXP2 assembles as a domain swapped dimer upon folding; however, it exhibits an unexpectedly low propensity to do so, despite dimerisation being a proposed prerequisite for its ability to regulate transcriptional activity. This investigation aims to elucidate the propensity of FOXP2-FHD to dimerise through probing its folding mechanism and structural stability. In this pursuit, experiments on wild-type FOXP2-FHD were conducted, and contrasted against a dimer-prone mutant (Y540F) which incorporated a substitution mutation uniquely present in the obligately-dimer forming FHD of FOXP3. Low-resolution DNA binding studies by electrophoretic mobility shift assay showed both proteins to have the ability to bind DNA as both dimers and monomers. Analysis by size exclusion chromatography revealed that while the wild-type and Y540F mutant readily dissociate from a dimeric to monomeric conformation post-purification, such dissociation decreases by two orders of magnitude upon denaturation and refolding. Unexpectedly, the wild-type is less prone to dissociate into a monomer in comparison with the Y540F mutant. Equilibrium unfolding studies reveal the wild-type to fold by a two-state mechanism; in comparison, the Y540F mutant displays a yet undocumented folding pathway, whereby it unfolds by a two-state mechanism and refolds by a three-state mechanism. Dissection of the refolding data for the Y540F mutant reveals the formation of a compact, stable intermediate comprising of hydrophobically buried secondary structural elements that otherwise participate in domain swapping. Overview of the data suggests the low dimerisation propensity for the wild-type and Y540F mutant to be a consequence of the bacterial expression system utilised for overexpression. Furthermore, dimerisation for the Y540F mutant is further impeded by the increased compactness and stability of its hydrophobically buried secondary structural elements that are elsewise required for domain swapping. This work suggests the reduced dimerisation propensity of FOXP2-FHD hence described in the literature to be a product of the expression protocol utilised but, to the author's knowledge, provides the first evidence for a disparate two-/three-state unfolding/refolding mechanism for protein folding.

Acknowledgements

To my supervisor, Dr Sylvia Fanucchi, thank you for your endless support and encouragement throughout my masters studies and time at the PSFRU, amidst all the bumps along the way. Your guidance and passion for research has fostered my curiosity and excitement for biochemistry – feelings I hope to always hold close as I pursue my future academic endeavours.

To the FOXy group, I will not forget our times optimising, experimenting, and bantering, and will remain forever in awe of your uncompromising commitment to French toast.

I am grateful to the University of the Witwatersrand and the National Research Foundation for their generous provision of funding in support of my studies and research.

Table of Contents

List of Abbreviations

α-IPMS – alpha-isopropylmalate synthase

μ_{eff} – net effective dipole moment

$\Delta G_{dimerisation}$ – Gibbs free energy of dimer formation

ΔG_{H2O} – Gibbs free energy of formation in water

ΔH – change in enthalpy

ΔH_m – change in enthalpy at thermal denaturation midpoint

ΔS – change in entropy

ΔS_m – change in entropy at thermal denaturation midpoint

β_T – position of transition state on the reaction coordinate

$\Delta G^{\ddagger}$ - Gibbs free energy of transition

BSA – bovine serum albumin

CD – circular dichroism

CD2 – cluster of differentiation 2

CI2 – chymotrypsin inhibitor 2 protein

CSAW – conditioned self-avoiding walk

DBD – DNA binding domain

DNA – deoxyribonucleic acid

DNAse – deoxyribonuclease

dsDNA – double stranded deoxyribonucleic acid

DVD – developmental verbal dyspraxia

EMSA – electrophoretic mobility shift assay

ETS – E26 transformation specific

Fkh – forkhead gene

FOX – forkhead box

FOXP – forkhead box protein, P subfamily

FPLC – fast protein liquid chromatography

GdnHCl – guanidine hydrochloride

HLH – helix-loop-helix

HTH – helix-turn-helix

IMAC – immobilised metal affinity chromatography

K_d – rate for dimer dissociation into monomer

LB – lysogeny broth

m – cooperativity of folding

PAGE – polyacrylamide gel electrophoresis

PCR – polymerase chain reaction

PDB – protein database

REM – random energy model

RNAse – ribonuclease

SASA – solvent accessible surface area

SCOP – structural classification of protein (database)

SDS – sodium dodecyl sulfate

SEC – size exclusion chromatography

SOC – super-optimal broth with catabolite repression

ssDNA – single stranded deoxyribonucleic acid

STP – standard temperature and pressure

T – temperature

TF – transcription factor

UV – ultra-violet

wHTH – winged helix-turn-helix

wtFOXP2-FHD – wild-type forkhead box P2 protein forkhead domain

List of Figures

List of Tables

List of Equations

1. $c = \frac{A_{280}}{\varepsilon \cdot l}$
2. $\varepsilon = 5550 \cdot n_{Trp} + 1340 \cdot n_{Tyr} + 150 \cdot n_{Cys}$
3. $K_d = \frac{[Monomer]^2}{[Dimer]}$
4. $Absorbance\ fraction = \frac{Area\ under\ individual\ peak}{Total\ area\ under\ both\ peaks}$
5. $\theta_{MRE} = \frac{100 \cdot \theta_{mdeg}}{c \cdot n \cdot l}$
6. $Y_{normalised} = \frac{(Y_{obv} - Y_U)}{(Y_N - Y_U)}$
7. $f_U + f_N = 1$
8. $Y = Y_N f_N + Y_U f_U$
9. $f_U = \frac{(Y_N - Y)}{(Y_N - Y_U)}$
10. $f_N = \frac{(Y_U - Y)}{(Y_U - Y_N)}$
11. $K = \frac{[U]}{[N]} = \frac{f_U}{f_N}$
12. $K = \frac{Y - Y_N}{Y_U - Y}$
13. $\Delta G = -RT\ ln(K)$
14. $\Delta G = \Delta G_{H2O} - m[D]$
15. $f_U = \frac{Y_N + Y_U \cdot e^{-(\Delta G_{H2O} - m[D])/RT}}{1 + e^{-(\Delta G_{H2O} - m[D])/RT}}$
16. $f_U + f_I + f_N = 1$
17. $Y = Y_N f_N + Y_I f_I + Y_U f_U$
18. $K_{N \leftrightarrow I} = \frac{[I]}{[N]} = \frac{f_I}{f_N}$
19. $K_{I \leftrightarrow U} = \frac{[U]}{[I]} = \frac{f_U}{f_I}$
20. $\Delta G_{N \leftrightarrow I} = -RT\ ln(K_{N \leftrightarrow I})$
21. $\Delta G_{I \leftrightarrow U} = -RT\ ln(K_{I \leftrightarrow U})$
22. $\Delta G_{N \leftrightarrow I} = \Delta G_{H2O\ N \leftrightarrow I} - m_{N \leftrightarrow I}[D]$
23. $\Delta G_{I \leftrightarrow U} = \Delta G_{H2O\ I \leftrightarrow U} - m_{I \leftrightarrow U}[D]$
24. $f_U = \frac{Y_N + Y_I \cdot e^{\frac{-(\Delta G_{H2O\ N \leftrightarrow I} - m_{N \leftrightarrow I}[D])}{RT}} + Y_U \cdot e^{\frac{-(\Delta G_{H2O\ N \leftrightarrow I} - m_{N \leftrightarrow I}[D])}{RT}} \cdot e^{\frac{-(\Delta G_{H2O\ I \leftrightarrow U} - m_{I \leftrightarrow U}[D])}{RT}}}{1 + e^{\frac{-(\Delta G_{H2O\ N \leftrightarrow I} - m_{N \leftrightarrow I}[D])}{RT}} + e^{\frac{-(\Delta G_{H2O\ N \leftrightarrow I} - m_{N \leftrightarrow I}[D])}{RT}} \cdot e^{\frac{-(\Delta G_{H2O\ I \leftrightarrow U} - m_{I \leftrightarrow U}[D])}{RT}}}$

Chapter 1: Introduction

1.1 Protein Folding

The process of protein folding refers to the mechanisms by which a polypeptide chain adopts a compact, three-dimensional structure through the formation and organisation of secondary and tertiary structural elements. The success of folding is crucial for proper protein function; incorrect or incomplete protein folding may ultimately manifest as cellular dysfunction and disease (Luheshi *et al.*, 2008). Proteins display a uniquely high degree of intrachain organisation in comparison with other soluble polymers (Dill *et al.,* 1995), allowing proteins to consistently, reliably and expeditiously evolve their functional tertiary structures determined by the sequence of their polypeptide chains (Dill *et al.,* 1995). Effective protein folding underpins the proper physiological functioning of the cell. However, as of yet, no unified folding mechanism has been described which seamlessly predicts, *de novo*, protein folding across the spectrum of all proteins, along with their potential intermediates and molten globule conformations simply from their linear amino acid sequences.

As first described in the research conducted by Anfinsen regarding the refolding of denatured and reduced bovine ribonuclease, it was concluded that all the information required for the correct assumption of three-dimensional structure was encoded in the linear amino acid chain of the protein (Anfinsen, 1973; Lecomte and Matthews, 1993). Upon the renaturation of unfolded bovine ribonuclease in the presence of a reducing agent, the protein was observed to reestablish its proper secondary and tertiary structure – thereby reassuming full functionality and homogeneity in solution. Protein refolding was hence observed to be a spontaneous process: the combination of enthalpy gain, coupled with the hydrophobic effect, outweighed the accompanying loss of entropy of formation upon refolding, allowing for the proper resumption of protein function (Pace, 1986; Alber, 1989). Specifically, the gain in enthalpy arose from the formation of intermolecular bonds between residues in the polypeptide chain, and with surrounding solvent molecules. The hydrophobic effect was enacted by the expulsion of water from the hydrophobic interior by hydrophobic residues of the protein leading to hydrophobic collapse. However, the reaction, occurring *in vitro*, required significantly longer to reach folded equilibrium in comparison to the reaction that would occur *in vivo*. This

prompted the question of how protein folding is conducted with such efficiency in the cell. If the folding behaviour of ribonuclease is to serve as a model for protein folding behaviour in general, a discrepancy arises between folding *in vivo* and *in vitro*. Within the cell, protein folding is observed to be rapid and highly cooperative in nature, which stands in contrary to the observation of ribonuclease folding *in vitro*. This question is further compounded by the Levínthal paradox. Levínthal postulated that as a result of the exponential number of adoptable conformations available to an unfolded polypeptide chain (as a result of its large number of degrees of freedom), a single protein would take longer than the age of universe to sample each conformation to arrive at its correct, native conformation (Levínthal, 1969). This stands in contradiction with observed protein folding rates, which can occur on the micro- to millisecond scale. Certainly, if protein folding occurs over timescales of billions of years, then it would not be observed at all, and life would not exist given the speed at which proteins must be synthesised and folded in the cell in order to execute its many processes. It stands to reason that proteins cannot exhaustively sample all available folding configurations along the path to their final folded state. Since Levínthal's postulate in 1969, advances in spectroscopic and analytical techniques, alongside progress made in computational simulation, have greatly aided in elucidating the discrete mechanisms of protein folding.

1.1.1 Folding Models

The thermodynamic hypothesis proposed by Anfinsen, 1973, and the kinetic hypothesis by Levínthal, 1969, exist as the most broadly accepted, foundational explanations of protein folding. Protein folding can additionally be explained in terms of energetics. As a protein folds from a random-coil unfolded state towards its native structure, its free energy of conformation decreases until it reaches a free energy minimum characteristic of its native fold (Dill and Alonso, 1988; Dill *et al.*, 2008) (Figure 1). However, the folding pathway undertaken by a protein in achieving a free energy minimum is neither smooth nor uniform; as a protein folds it may assume conformations that exhibit local free energy minima which are higher than the free energy minimum of the native fold, although low enough to allow for partially folded conformations to exist transiently (Dobson, 2004). These local minima are presented as energy barriers which impede the progress of folding towards a global energy minimum, thus slowing the rate of folding. Small energy barriers are more easily overcome without hampering the rate of folding, and usually correspond with the formation of secondary structural elements (Morris and Searle, 2012). Larger barriers may manifest in the formation of stable structural

intermediates, or structural aggregates, which considerably limit the rate by which a protein folds to its native state (Morris and Searle, 2012). The concept of an uneven energy landscape along a "folding funnel" most aptly demonstrates this idea (Figure 1).

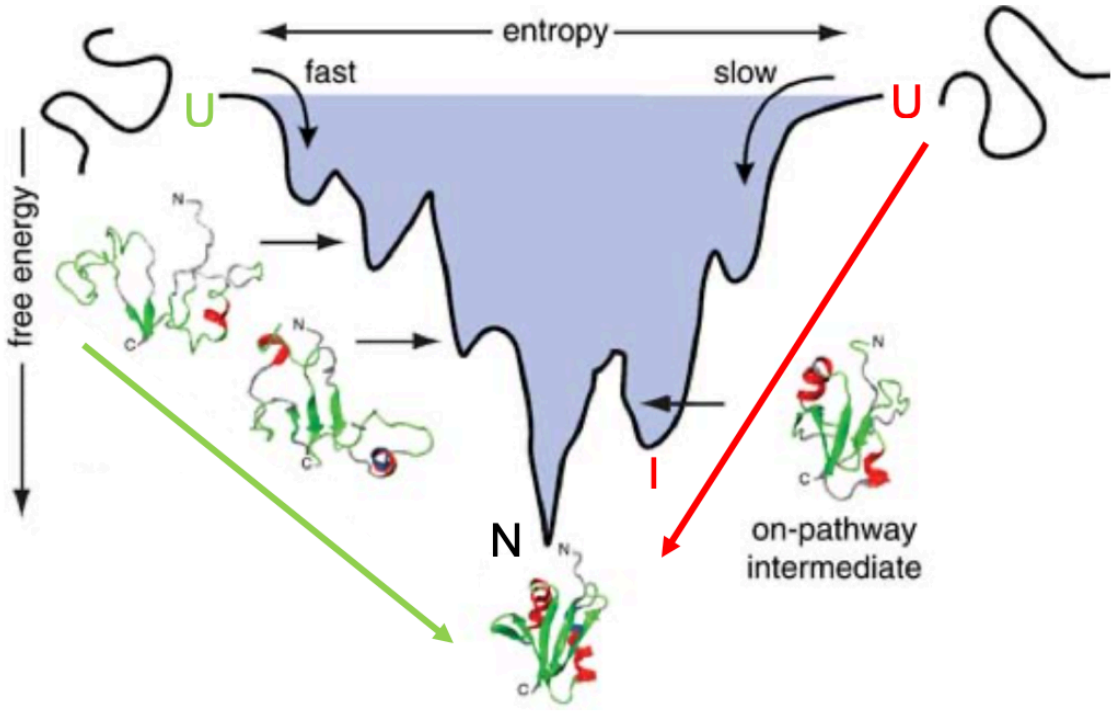


Figure 1: Schematic of the energy landscape of the protein folding funnel

The energy landscape of the folding funnel is non-uniform, and is lined with troughs corresponding to a decrease of free energy of conformation, associated with both the gradual restriction of the conformational state of the protein and the formation of folded structural components. Folding is additionally accompanied by a loss of entropy by virtue of the formation of ordered secondary and tertiary structural components. An unfolded polypeptide chain may encounter multiple energy barriers toward assuming its final folded state, which affect its rate of folding. Random sampling of folded conformations present no major free energy barriers, and folding progresses rapidly, exhibiting a two-state folding mechanism (U → N) (green arrow). In addition to these minor free energy barriers, the folding pathway may be populated by an intermediate state, which is trapped in a low-energy conformation (red arrow). This causes a folding bottleneck to arise which slows folding considerably, and results in a three-state folding mechanism (U → I → N). Adapted from Morris and Searle, 2012.

As embellishments on the concept of a folding funnel, models to describe protein folding such as the diffusion-collision model (Karplus and Weaver, 1979; Karplus and Weaver, 1994), nucleation-condensation model (Fersht, 1997) and hydrophobic collapse model (Sosnick *et al.*, 1996) have been proposed. Models for protein folding are by no means limited to these three propositions: theoretical models supported by statistical mechanics such as the random energy model (REM) and conditioned self-avoiding walk (CSAW) model have been suggested (Derrida, 1980; Pande *et al.*, 1997; Lei and Huang, 2012). However, these models are founded in heavy mathematical formalism within the fields of replica field theory and Itô calculus, and so are beyond the scope of this investigation. Therefore, focus will be placed on a description of the three aforementioned models. While the diffusion-collision, nucleation-condensation, and hydrophobic collapse models have been proposed independently of one another, it is more probable that a description of protein folding on a case-to-case basis requires an overlap of aspects from each model.

1.1.1.1 The Diffusion-Collision Model

The diffusion-collision model views a folded protein as a constituent array of discrete, fluctuating quasiparticles – termed microdomains (Karplus and Weaver, 1994). A microdomain is defined as either an emerging secondary structure (such as an α-helix or β-strand), or transient hydrophobic bundle. Each microdomain moves within the solvent in a diffusive manner, intermittently colliding with other microdomains not necessarily located adjacently in the linear sequence of the polypeptide (Karplus and Weaver, 1994). Upon collision, driven by random external forces, such microdomains may conjoin into multimicrodomain intermediates – assuming correct orientation is achieved. Through a sequence of microdomain-microdomain interactions, folding proceeds either according to a unique folding pathway, or along multiple parallel pathways. The progression of folding along a single or set of pathways is limited by the stability of interacting microdomains, obstructions to their ability to interact, and the viscosity of the solution within which the protein folds (Karplus and Weaver, 1994). According to the model, folding is assumed to be a stepwise process that ultimately leads to final, packed tertiary structure, possibly through the development of an imitative, molten globule state (Ohgishi and Wada, 1983; Ptitsyn, 1987). Such a model allows for reduced concern for individual amino acid behaviour during folding. Rather, it considers groups of amino acids as individual segments, decreasing the range of configurational space available and simplifying the description of the folding pathway for

globular proteins. In turn, this view may help resolve the Levínthal paradox by limiting the conformational search space undertaken by a protein during folding. The model is limited, however, by its assumption of a protein initially existing in a truly unfolded, random coiled state – one that is not observed *in vivo* as a result of both molecular crowding within the cell, as well as from co-translational folding which occurs during protein biosynthesis (Karplus and Weaver, 1994; le Coeur, 2009; Waudby *et al.,* 2019). Within the cell, protein folding occurs almost simultaneously as the nascent polypeptide chain exits from the ribosome, thus preventing the existence of a true unfolded polypeptide state post-translation (Waudby *et al.,* 2019). However, in the presence of GdnHCl and at high temperatures, an unfolded protein does, in fact, closely resemble a random coil conformation (Alber, 1989). Hence, the merit of the model might be restricted to *in vitro* studies and as a basis for computer simulation of protein folding (Baldwin, 2000) (Figure 2).

1.1.1.2 The Nucleation Condensation Model

The nucleation-condensation model, unlike the diffusion-collision and hydrophobic collapse models, does not view protein folding as a stepwise process (Fersht, 1997). Rather it asserts that a few select, neighbouring residues along the length of the unfolded polypeptide adopt a stable secondary structural configuration (nucleation) – most commonly an α-helix – which acts as a template around which higher order structure is assumed cooperatively by the remainder of the protein (condensation) (Fersht, 1997). The formation of the initial helical turn is energetically unfavourable as a result of the large reduction in entropy during nucleation. Once achieved, however, both local and long range residues (in sequence, not space) interact with the nucleus to form a more disperse nucleus surrounding it, subsequently propagating into the final tertiary structure. It must be stated that "nucleation" can be understood ambiguously in this context: in the literature, "nucleation" can refer to both the formation of the stable secondary structural nucleus (structural nucleation) and the assumption of higher order structure of the remainder of the protein around such nucleus (kinetic nucleation) (Wetlaufer, 1990). This model most accurately describes the folding of small globular proteins exhibiting a two-state kinetics, as such proteins do not generally form intermediates, occupying a single transition state between unfolded and native configurations (Fersht, 1997) (Figure 2).

1.1.1.3 The Hydrophobic Collapse Model

The hydrophobic collapse model was first described as a solution to the folding mechanism of cytochrome *c* (Sosnick *et al.*, 1994). It was observed that the rate limiting step of the two step folding mechanism of cytochrome *c* involved a large scale condensation effect whereby half of the molecular surface (predominantly constituting hydrophobic residues) collapsed to form the core of the folding protein (Sosnick *et al.*, 1996). Hydrophobic collapse is spontaneous, and is driven by the hydrophobic effect, whereby water molecules are expelled from the protein interior, thus increasing the entropy of the system (Cui *et al.*, 2014). The collapse involves multiple residues "searching" across multiple unfavourable side chain combinations, until an energetically favourable interaction is attained, which is followed by swift inter-side chain association and chain collapse. Once global collapse has occurred, a portion of the polypeptide chain loops remains solvent exposed, which subsequently folds along a series of intermediate states until a final native state is achieved (Sosnick *et al.*, 1996; Lapidus *et al.*, 2017). As the collapse is rapid there is a possibility of misfolding occurring, especially in the case of larger, more intrinsically disordered proteins. This itself results in an additional misfold-reorganisation barrier that must be surpassed before achievement of a native state, although chaperone proteins, disulphide isomerases and prolyl isomerases assist in overcoming such a barrier were it to arise (Sosnick *et al.*, 1996) (Figure 2).

It must be noted that these proposed models only consider protein folding *in vitro*, and are best employed when used to describe protein folding empirically observed within *in vitro* conditions. More recent models have been suggested (Lei and Huang, 2010; Vitalis and Caflisch, 2012) although these approach protein folding from a theoretical perspective, and model protein folding in an *in silico* environment. In the literature to date, none of the models are especially favoured over one another, although foundational aspects underlying each model, for example, hydrophobic collapse via the hydrophobic effect (Sosnick *et al.*, 1996; Cui *et al.*, 2014; Lapidus *et al.*, 2017) appear to be universally accepted.

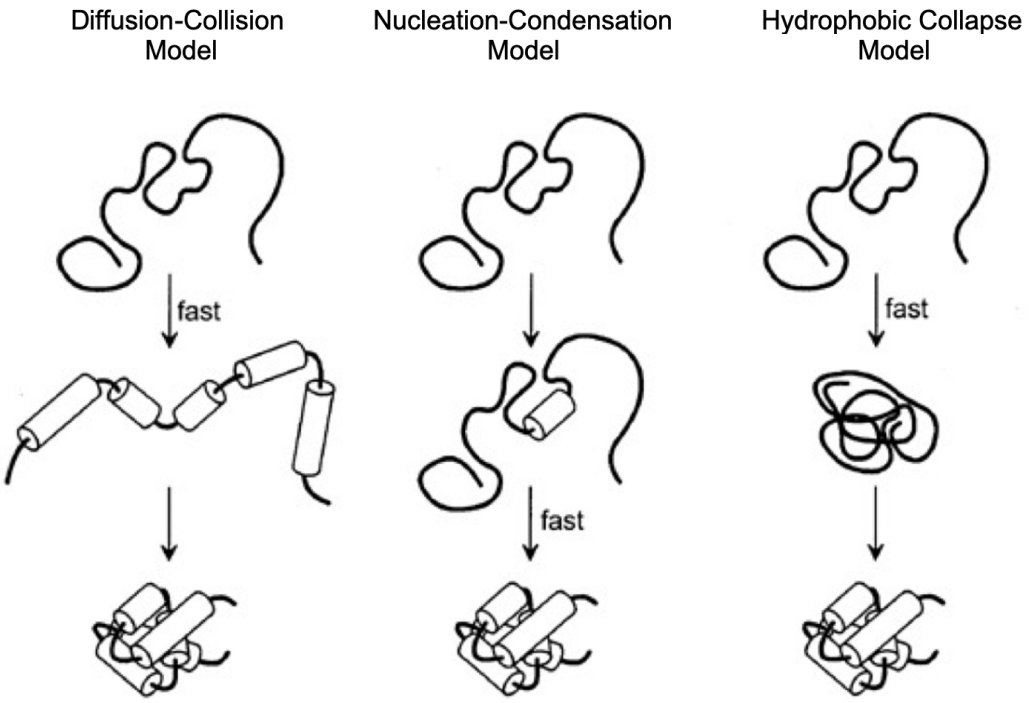


Figure 2: Protein folding models

Three models that describe protein folding initially view the unfolded protein as a random coil. According to the diffusion collision model, proteins initially and rapidly form transient, secondary structural microdomains, and then fold according to the speed of the interactions between stable microdomains. The nucleation-condensation model proposes that a single, stable secondary structural element must fold first to act as a nucleus for subsequent, cooperative folding. The hydrophobic collapse model posits that upon a favourable interaction between multiple hydrophobic side chains, global collapse occurs whereby hydrophobic side chains are subsequently buried in the protein interior. This expels water molecules from the protein interior, increasing the entropy of the system, while simultaneously increasing enthalpy by the formation of intermolecular bonds, governed by van der Waals forces in the interior, and electrostatic interactions with solvent molecules. The remaining solvent exposed main chain then samples through combinations of stable conformations until the native fold is achieved. Obtained from Liu *et al.*, 2012.

1.1.2 Contributing Factors to Protein Stability

The stability of a protein is dictated by the dynamic interplay between internal and external forces both driving and opposing the maintenance of its native state. Proteins exist in a marginally stable state (at STP) with a free energy of folding calculated to be between 20 – 85 $kJ.mol^{-1}$ (approximately 5 – 20 $kcal.mol^{-1}$), for most proteins (Privalov and Khechimashvili, 1974; Savage *et al.*, 1993). The forces governing protein stability include electrostatic interactions between side chains and solvent molecules, hydrogen bonding between surface polar / charged side chains and solvent molecules, local intrinsic interactions mediated electrostatically and by hydrogen bonding and dispersion forces, helix stabilisation, and the hydrophobic effect (Alber, 1989; Dill, 1995; Facchiano *et al.*, 1998; Nicolau *et al*, 2014; Herschlag and Pinney, 2018).

1.1.2.1 The Hydrophobic Effect

The most prominent driving force of protein stability and folding is the hydrophobic effect. The effect is described as the aversion of water to nonpolar residues (Langerstrom-Lang, 1959; Alber, 1989). As first reasoned by Kauzmann, 1959, the formation of one hydrophobic interaction between hydrophobic residues would, in turn, allow for the formation of a full hydrogen bond between water molecules and exposed polar side chains. The strength of this hydrogen bond is an order of magnitude greater than a similar bond formed between water and such nonpolar side chains, thus increasing the free energy of formation of the protein as a result of increased solvation (Kauzmann, 1959). Therefore, from an energetics perspective, the hydrophobic effect can be justified as being thermodynamically favourable through an increase of enthalpy. Nonpolar amino acids collectively contribute to the free energy of stabilisation of a protein, whereby an accumulation of free energy occurs during their bulk transfer from solvent exposure to the hydrophobic interior (Alber, 1989). The magnitude of free energy gained is dependent on the hydrophobicity of each nonpolar side chain. Hydrophobicity is an empirically derived enthalpic parameter, calculated by measuring the free energy of transfer of nonpolar side chains from water to an organic solvent, or to the vapour phase (Tanford, 1962; Pace *et al.,* 2011). It has been estimated that each $Å^2$ of hydrophobically buried surface area contributes between 92 – 105 J of free energy to the total free energy of stabilisation for a protein (Richards, 1997). In addition, the burial of a -CH2- group upon folding has been observed to contribute approximately 4,6 $kJ.mol^{-1}$ to the free energy of stabilisation of a protein

(Pace *et al.*, 2011). Again, from a thermodynamic perspective, the formation of the hydrophobic core plays a considerable role in stabilising the structure of a protein. Further evidence from solved crystal structures of multiple globular proteins confirms side chain hydrophobicity to be a dominant contributor to protein stability (Nicolau *et al*, 2014). These structures have consistently shown sequestration of nonpolar residues in the core of the folded protein, whereby solvent contact is avoided (Nicolau *et al*, 2014). In addition, nonpolar residues involved in the formation of the hydrophobic interior of a protein appear to be more strongly conserved in terms of sequence and structure in comparison to surface exposed residues (Lim and Sauer, 1989; Chowriappa *et al.*, 2009; Gowder *et al.*, 2014). This may result from their presence being so essential in their contribution to free energy of stabilisation, that mutation would make folding unviable (Chowriappa *et al.*, 2009). While the hydrophobic effect contributes primarily to protein stability and folding, additional forces play a supplementary role in the overall stability of a protein.

1.1.2.2 Electrostatic Interactions

Electrostatic interactions between ionisable groups on the surface, or buried within a protein, additionally contribute to protein stability. In early studies, electrostatic interactions were presumed to be the predominant force in influencing protein folding and stability (Jacoesen and Linderstrøm-Lang, 1949). This was justified as acids and bases were used as the earliest protein denaturants to probe structure (Jacoesen and Linderstrøm-Lang, 1949; Dill, 1995). This has long since been accepted not to be the case, as the hydrophobic effect is observed to be more dominant in terms imparting stability to a protein (Dill, 1995). Electrostatic interactions which act to stabilise protein are largely governed by ionic pairing between charged residues, although on average, only a third of the total charged residues constituting a protein are involved in forming ion pairs (Alber, 1989). Stabilising electrostatic forces within a protein and on its surface exist primarily as ion pairs between two charged amino acids in close proximity (salt bridges), where each have an opposite charge (usually as a result of proton exchange) and are thereby attracted to one another (Mirsky and Pauling, 1936; Bosshard *et al.* 2004). An appreciation of their contribution to protein stability is compounded by the observation that ionisable groups are not randomly distributed over protein surfaces, and hence must have purposive effects on stabilising the structure of a protein (Wada and Nakamura, 1981; Crevenna *et al.*, 2012). In addition, protein stability is sensitive to destabilising mutations affecting charged residues (Börjessen and Hünenberger, 2003).

Characteristic of the role of electrostatics on protein stability is that electrostatic effects are not confined by residue proximity with regard to their location in the overall structure of a protein. It is estimated that the electrostatic energy of interaction between charged groups positioned at a distance of 10 Å, and carrying a net unit of charge, exhibit a binding energy of approximately 1 $kJ.mol^{-1}$ (Zhang *et al.*, 2012). Therefore, electrostatic interactions nonlocally act to stabilise protein structure. Nonpolar residues can additionally provide minor contributions to stability by electrostatic interaction (Alber, 1989). Indeed, in terms of aromatic residues, edge-to-face packing of aromatic rings may form electrostatic attractions between their electron-poor aromatic hydrogens and electron-rich π-orbitals of adjacent aromatic rings (Alber, 1989; Zhou, 2015).

1.1.2.3 Hydrogen Bonding

A major contributor to the temperature-independent enthalpy of stabilisation of a protein arises from hydrogen bond formation. The strength of the bond is determined by molecular geometry and bond angle, and is further impacted by enthalpic contributions arising from dispersion, electrostatic and steric repulsion interactions (Vinogradov and Linnell, 1971). Hydrogen bonding occurs when a proton is shared along a linear axis between an electronegative donor and acceptor atom. Rarely in the case of protein structure are hydrogen bond donors and acceptors observed to be unpaired (Schellman, 1958; Alber, 1989; Herschlag and Pinney, 2018). Hydrogen bonds positively contribute to the stability of a protein by virtue of their significant heat of formation, calculated to be -6,3 $kJ.mol^{-1}$ or larger (Susi and Timasheff, 1964). More dramatically, hydrogen bonds between charged groups can contribute up to 25 $kJ.mol^{-1}$ in binding energy (Alber, 1989; Herschlag and Pinney, 2018). It is popularly viewed that as long as there is at least an equal number of hydrogen bonds formed in a protein's unfolded and folded states, the effect of hydrogen bonding will play a role in establishing a protein's specificity of folding (Alber, 1989; Fitzpatrick *et al.*, 2011; Mallamace *et al.,* 2018). In addition, it stands to reason that hydrogen bonding will only increase protein stability, from an energetics perspective, if the number hydrogen bonds formed in the folded state is equal to or greater than the number formed in the unfolded state (Alber, 1989). Hydrogen bonding is not limited to residue-residue interactions: indeed, hydrogen bonding between buried water molecules and buried residue side chains can act to stabilise protein structure – in some cases critically so. (Alber *et al.*, 1987; Alber *et al.*, 1988; Takano *et al.*, 2003; Renthal, 2008). In terms of secondary structure, hydrogen bonding forms the foundation for β-sheet and α-helix

formation. Hydrogen bonding is the primary driver of the transition from the disordered random coil-state of a polypeptide chain to an ordered secondary structural element (Schellman, 1958; Scheiner and Hillenbrand, 1985). In terms of α-helical formation, once the energy requirements of the initial helical turn in the transition are overcome, hydrogen bonding between intrachain residues in the helix become more energetically favorable in comparison to solvent hydrogen bonding, thus stabilising the helix. This is additionally true of the stabilization and formation of beta sheets, although the rate of formation is considerably slower (Anufrieva, *et al.,* 1967; Mattice and Scheraga, 1984). From a hierarchical perspective, the presence of and propensity of formation of α-helices too contributes to the stabilisation of a protein.

1.1.2.4 Helix Stabilisation

The degree of a protein's stability has been shown to be correlated with its propensity for helix formation. Helices, in turn, are inclined towards greater stability as their chain length and nucleation propensities increase (Zimm and Bragg, 1959; Kim and Baldwin, 1984; Facchiano *et al.*, 1998; Yakimov *et al*., 2016), whereby nucleation is defined as initial energy barrier that must be overcome by residues in initiating helix formation (Miller *et al.*, 2002). Helix chain length stability results from the summation of free energy contributed by each individual residue along the helix (Zimm and Bragg, 1959; Scheraga, *et al.*, 2002; Yakimov *et al.,* 2016). Shorter helices have been observed to be unstable in isolation, or fail to nucleate, in aqueous solution (Dyson *et al.*, 1988b). In addition, overall protein stability can be positively affected if helix formation leads to the burial of nonpolar side chains (Richards and Richmond, 1978), and can variably be affected if helix residues interact electrostatically or by nonpolar, aromatic interactions (Shoemaker et al., 1985). However, the effect of helix formation on protein stability is not only confined to the interaction between helix residues. Helix-helix interactions can occur electrostatically on a macroscale by virtue of the helix exhibiting a macrodipole moment (Arridge and Cannon, 1964; Miller *et al.*, 2002; Sengupta *et al.*, 2005). This macrodipole develops from the cumulative sum of microscopic dipole moments arising from individual peptide bonds in the helix (Creighton, 1993; Sengupta *et al.*, 2005), therefore imparting the helix with a partially negative and positive charge at its C- and N-termini, respectively (Hol, 1985; Sengupta *et al.*, 2005). This allows helices to interact electrostatically in an antiparallel fashion, with each helix co-stabilising each other, and by extension, stabilising the overall protein (Yeates and Padilla, 2002; Sengupta *et al.*, 2005). The net

effective dipole moment (μ_{eff}) of a helix depends on the positions of its N- and C-termini with respect to the aqueous phase although, perhaps counterintuitively, μ_{eff} does not depend on the overall solvent exposure of the helix (Sengupta *et al.*, 2005). When both termini, or at least one terminus, are buried (at least 7,5 Å from the protein surface), μ_{eff} is large and the effect of antiparallel helix interaction is pronounced, whereby the helix behaves as it would in a model nonpolar medium ($\varepsilon = 2$, where ε is the dielectric constant). On the contrary, if both termini are solvent exposed, μ_{eff} is small and the helix participates in electrostatic interactions as one would expect of a fully solvent exposed helix in a model polar medium ($\varepsilon = 80$). Therefore, while helical bundles exhibiting buried termini act to stabilise a protein, isolated helices with buried termini act to destabilise it (Sengupta *et al.*, 2005). Indeed, the contrary is also true with respect to solvent exposed helical termini.

Overall protein stability arises from varying contributions from each of the aforementioned forces. Although each force and their effect on protein stability can be satisfactorily described in terms of their chemistry and thermodynamics, unravelling each of their individual contributions to protein stability on a per-protein basis is challenging. Commonly, site-directed mutagenesis and thermal or chemical equilibrium unfolding are relied upon to probe the effect of individual residues on protein stability and folding, and to yield thermodynamic parameters of folding such as conformational stability and cooperativity.

1.1.3 Protein Folding Mechanisms

As a polypeptide chain emerges from the ribosome following translation, protein folding spontaneously occurs whereby a protein adopts its native conformation from its initially close-to-fully unfolded state (Englander and Maine, 2014). The ideal unfolded state is defined as the set of all possible random-coil conformations an unfolded protein can occupy, with each conformation having comparable free energy (Creighton, 1990). The unfolded state is not purely random in the statistical sense; unfolded proteins still portray inter-residue interactions and fluctuating degrees of residual structure (Plaxco and Gross, 2001). Under physiological conditions, proteins exist at an equilibrium of native, and partially to fully unfolded states, which oscillate around either a local or global energy minimum (Walters *et al.*, 2009). The folding pathway of a protein, from an energetics perspective, can hence be described as being populated with a discrete number of folded protein states at varying degrees of stability, all of

which seek to minimise their relative free energies of conformation (Walters, *et al.*, 2009). As each discrete state exists at equilibrium, isolating each equilibrium state over the full set of states along a protein's folding pathway, from its unfolded to native conformation, will yield an overview of a protein's sequential folding pathway (Walters, *et al.*, 2009). However, this view is restricted by the assumption that each point along the folding pathway is significantly populated by intermediates of a similar folded state (Walters, *et al.*, 2009). Indeed, in the case of sequential folding, points along a folding pathway can be populated with intermediates at similar folded states and at similar free energies of conformation. However, on the contrary, the points along the folding pathway may additionally be populated with intermediates of multiple different folded states, although at similar conformation energies. These intermediates form simultaneously and in parallel, and all eventually coalesce to the native state (Wallace and Matthews, 2002; Walters, *et al.* 2009). The rate at which this coalescence occurs is best described by folding kinetics. The kinetic folding pathway, for small proteins, usually occurs by a two-state mechanism where the protein folds over a single transition from an unfolded to folded state whereby no stable intermediates are formed. For larger proteins exhibiting an increased quantity of structural elements, protein folding grows in complexity and may be further complicated by the development of stable folded intermediates. In these cases, the folding pathway will be rate limited by the development of each stable intermediate, and will fold by an (n + 2)-state mechanism, where n is the number of intermediate folded states. A kinetic description of folding becomes even more convoluted in the case of oligomerisation, whereby one must account for oligomer association and dissociation, alongside folding which may occurring distinctly or simultaneously with folding (Schlunegger, *et al.*, 1997; Walters, *et al.* 2009; Englander and Maine, 2014). Regardless of the number of intermediates formed, or complexity of the folding pathway, the folding reaction of any given protein must follow an overall $U \leftrightarrow N$ mechanism as opposed to a $U \leftrightarrow N_1 + N_2 + \cdots + N_i = \sum N_i$ (Rose *et al.*, 2006). In other words, a polypeptide chain should not adopt multiple native state conformations.

Experimentally, the determination of the folding mechanism of a protein is commonly achieved through equilibrium unfolding studies, which relies on isolating discrete folded states by the sequential denaturation of a protein. Denaturation may be employed chemically (though the addition of urea, guanidine hydrochloride (GdnHCl) or altering solution pH) or physically (by the incremental increase of temperature or pressure) (Privalov, 1979). Equilibrium unfolding studies can yield details pertaining to the thermodynamic parameters of protein folding such as

its conformational stability (in terms of free energy of conformation - ΔG_{H2O}), its cooperativity of folding (*m*) and, by extension, its solvent accessible surface area (SASA) of its various folded states. In deducing ΔG_{H2O} and *m,* additional thermodynamic parameters such as the enthalpic (ΔH) and entropic ($T\Delta S$) contributions to protein stability may be calculated by virtue of the Gibbs-Helmholtz equation for free energy ($\Delta G = \Delta H - T\Delta S$). Equilibrium unfolding studies have been shown to be very effective in elucidating and comparing the folding mechanisms and stabilities of small, monomeric proteins such as individual protein domains (Pace, 1986), as well as larger monomeric or oligomeric proteins (Walters, *et al.* 2009; Dìaz-Villanueva, *et al.,* 2015).

1.2 Protein Domains and Domain Folding

1.2.1 The Concept of a Protein Domain

Domains are defined as regions of a polypeptide chain that are spatially distinct, fold in isolation, and are biochemically functional (Ponting and Russell, 2002). At the advent of the term, a domain referred exclusively to a functional segment of a protein, without much regard to its structure. Indeed, in the first solved structures of lysozyme and ribonuclease, distinct, spatially separate, active regions were observed and termed “domains” as a result of their intrinsic biochemical function (Blake and Peacocke, 1965; Kartha, *et al.*, 1967). In contrast, domains are currently classified according to their structure. Modern methods employ both *in vitro* and *in silico* approaches to classify domains by their three-dimensional architecture in terms of geometry and degree of compactness (Ponting and Russell, 2002). This characterisation, based on a three-dimensional structure, has allowed for domains to be used as identifiers and predictors of function – in some cases – across varying species and contexts, whereby sequence is not a comparable factor, although structures show a degree of similarity (Russell, 1998; Russell, 2001).

However, the proposition that a shared similarity between domain structures can always be used to infer a similarity between each domain’s function does not hold true in all instances. In fact, certain proteins that show similar folds often have no related function at all. This is justified as domains with common structure do not always share common ancestry (Russell, 2001). Rather, a more reliable determinant of function arising from structure lies in the presence of similar active surfaces shared by structurally similar domains. A prominent

example is in the case of ligand binding, whereby it was observed (Russell, 1998) that nine superfolds including β/α barrels, doubly wound βαβ folds and β-propellers, among others, have preferential ligand binding surfaces despite arising from unrelated roots of ancestry (Russell, 1998).

As evidenced by homology modelling across a variety of protein databases, up to two-thirds of proteins in prokaryotes consist of two or more domains comprising their full length, with this fraction expected to be higher among eukaryotes (Teichman *et al.*, 1998; Gernstein, 1998). However, in terms of domain variability within proteins, it has been observed that a significant proportion (up to 25 %) of eukaryotic proteins contain domain repeats (Marcotte *et al.*, 1999; Schüler and Bornberg-Bauer, 2016) with 98 % of domains in human proteins being duplicates (Muller *et al.*, 2002). Such a prevalence of domain repeats across multiple protein families has been justified by the proposition that gene duplication events followed by eventual sequence divergence act as the major drivers of novel protein emergence (Muller *et al.,* 2002; Lynch, 2005). Through duplication and sequence divergence, domains can gradually acquire an increased propensity for interaction via their interfaces, or develop the ability to form higher-ordered architectures (Vogel *et al.*, 2004). Moreover, it is within the subtle variations between such evolved domain interfaces that lead to the rise of the multiple biochemical functions observed across domain families (Bhaskara and Srinivasan, 2011).

1.2.2 The Variability of Single-Domain Two-State Protein Folding

At the advent of protein folding studies, proteins were assumed to all fold through intermediate transition states that were stable enough to be characterised (Kim and Baldwin, 1990). It was shown in 1991, however, that the formation of intermediates was not a prerequisite for fast protein folding, whereby chymotrypsin inhibitor 2 (CI2) was observed to rapidly fold without displaying such intermediate states (Jackson and Fersht, 1991). Since the pioneering studies on CI2, there has been substantial thermodynamic and kinetic evidence for many small, single domain proteins exhibiting folding without the formation of intermediates (Jackson, 1998), thus according to a two-state folding mechanism.

The two-state mechanism is the simplest model of protein folding. As aforementioned, in two-state folding, only the native and unfolded state of a protein populate the folding pathway,

whereby both unfolding and refolding are monophasic processes (Jackson, 1998). Although many small, single domain proteins exhibit simple two-state folding kinetics, the rates at which these proteins unfold and refold (k_U and k_F, respectively) as well as the position of their transition states on the reaction coordinate (β_T) vary broadly (Jackson, 1998; Englander and Maine, 2014). Indeed, a comparison of k_U and k_F of small single domain proteins that display two-state folding kinetics has shown that these quantities span a range of up to six orders of magnitude across various protein species (Jackson, 1998; Corrales *et al.*, 2015). Given this large range, applications of machine learning have been employed to predict the theoretical folding rates for uninvestigated proteins by virtue of amino acid sequence (Corrales *et al.*, 2015). Unfortunately, the quantity of experimental data in informing the machine learning approach is too lacking for it to be reliable (Corrales *et al.*, 2015). The position of the transition state on the reaction coordinate, β_T, is a measure of likeness between the transition state during folding and the native state with respect to solvent accessible surface area (SASA) (Jackson, 1998). Hence, β_T describes the compactness of a protein in its transition state in comparison with the compactness of the native state. β_T can have values between 0 – 1, whereby 1 indicates a transition state which is very alike the native state with respect to SASA, and 0 indicates a transition state SASA closer to the unfolded state (Jackson, 1998). β_T is useful in informing similarities in folding between homologous proteins, whereby comparable β_T values indicate similar transition states between such proteins. In terms of engineered proteins with incorporated mutations to probe structure, comparing β_T values between both engineered and wild-type variants can be equally informative. Indeed, as with folding rates, single domain proteins show high variability in β_T. Single domain proteins have been observed to exhibit a β_T between 0,39 – 0,94; in other words, single domain proteins may exhibit a β_T over almost two-thirds of the total range of permitted β_T values (Jackson, 1998; Fersht, 2004). From these two considerations alone, it is evident that single domain proteins can exhibit drastically difference mechanisms of folding even in light of the apparent simplicity of two-state folding kinetics.

More pertinent to this investigation is the folding of single domain DNA binding proteins. As can be expected, DNA binding proteins too show high variability in folding even if observed to fold by two-state kinetics. For example, in the case of human papillomavirus E2 protein, a nonnative-like monomeric intermediate exists early during its folding pathway which is only detectable by kinetic analysis by stopped-flow, and unobserved in equilibrium studies (Mok *et*

al., 1996). The folding behaviour of single domain DNA binding proteins may further be impacted by the presence of DNA: in the case of Arc-repressor protein, the presence of ssDNA increased the folding rate of the protein in comparison with dsDNA (Marcovitz and Levy, 2013) although a high concentration of ssDNA were observed to impede folding rates of the protein (Rentzeperis *et al.*, 1999).

1.3 DNA Binding Domains and Dimerisation

1.3.1 DNA Binding Domains

Transcription factors (TFs) are multidomain proteins that serve to up- or downregulate gene expression through specific DNA sequence recognition and association (Kaufmann and Knöchel, 1996). TFs bind directly to local or distal control elements, commonly upstream of their target genes, whereby sequence-specific recognition is mediated by a constituent DNA binding domain (DBD) (Kaufmann and Knöchel, 1996). The DBD interacts noncovalently with both the sugar-phosphate backbone and nucleotide bases of the DNA in order to facilitate transcriptional initiation and the recruitment of additional TFs. The DBD accomplishes sequence-specific DNA binding through the presence of a secondary structural recognition element, which forms hydrogen bonds between the DBD and DNA nucleotide bases (Kaufmann and Knöchel, 1996). The various structures of DBDs promote DNA binding by two sets of interactions: firstly, the recognition element of the DBD forms a tight shape complementarity between the concave DNA major groove, and its complementarily convex DBD binding element (Noy *et al.*, 2016). Shape complementarity is governed by an accumulation of van der Waals forces between the recognition element and the DNA, as well as electrostatic attractions between the negatively charged DNA backbone and positively charged residue side chains (Privalov *et al.*, 2011). Secondly, specific hydrogen bonding occurs between the residue side chains of the DBD and the acceptor or donor groups of the nitrogenous bases of the DNA. Both interactions are necessary to facilitate the formation of a stable protein-DNA complex (Privalov *et al.*, 2011).

In eukaryotes, a wide range of TFs exhibit common DBDs, hence permitting their classification according to their class of DBD (Funnell and Crossley, 2012). Within the array of eukaryotic transcription factors, the most prominent DBDs include the zinc finger, zipper-type domains (including the leucine zipper and helix-loop-helix (HLH)), E26 transformation-specific (ETS)

domain, and helix-turn-helix (HTH) (Funnell and Crossley, 2012). While TFs featuring a zinc finger or ETS domain are able to associate with DNA as monomers to facilitate transcription, homo- or heterodimerisation is a requirement between TFs containing zipper-type or HTH motifs in order to achieve proper DNA recognition and binding (Funnell and Crossley, 2012) (Figure 3).

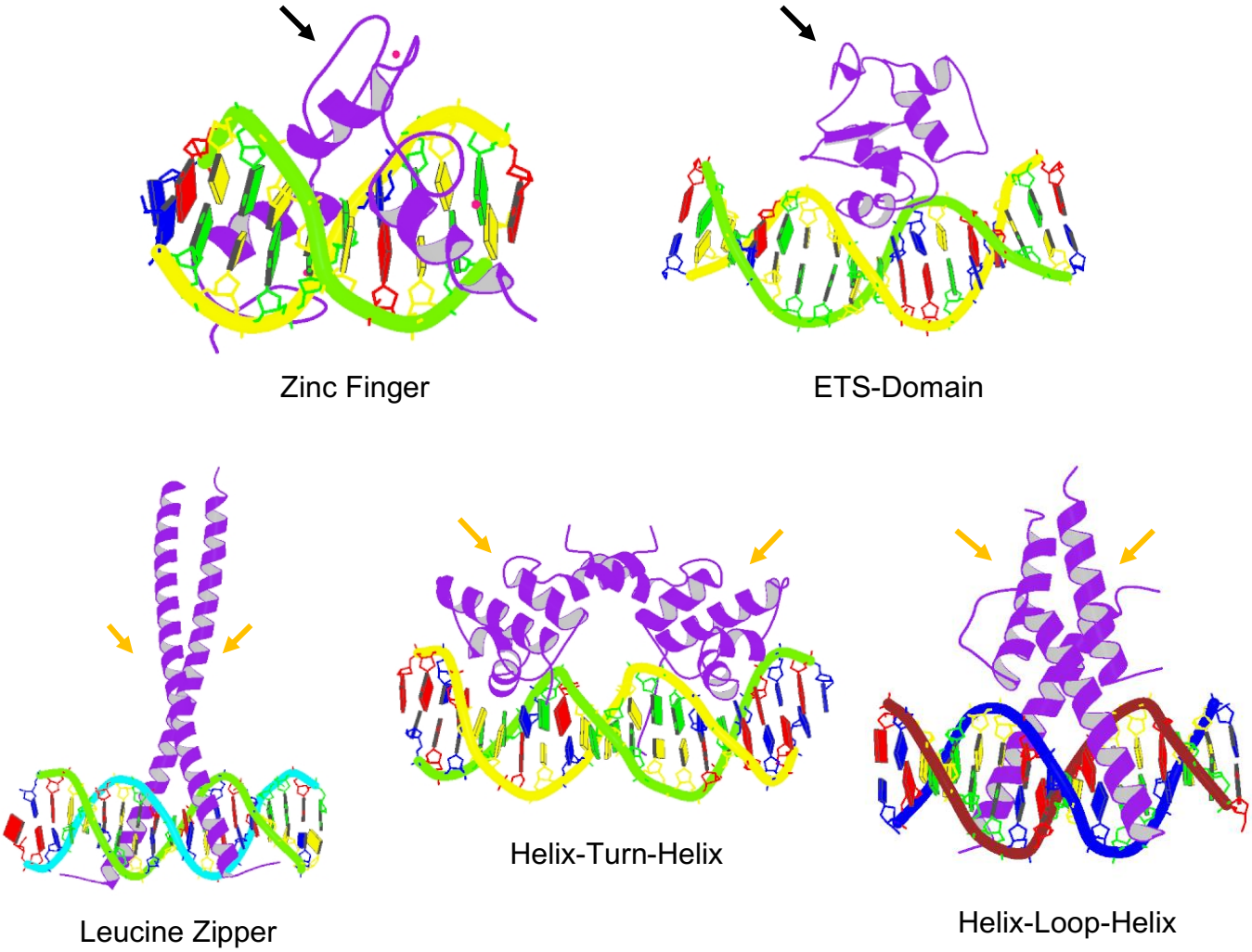


Figure 3: Association of DNA binding domains with DNA.

Crystal structures of the most prominently observed DBDs in eukaryotes are displayed. The zinc finger and ETS-domain motifs are able to bind DNA as monomers (black arrows), while the leucine zipper, helix-turn-helix (HTH) and helix-loop-helix (HLH) bind DNA as dimers (orange arrows). Featured DBDs include: (1) Zinc Finger: Zif268 protein, Elrod-Erickson *et al.* (1996), 1AAY; (2) ETS-Domain: ETS-1 protein, Werner *et al.* (1997), 2STT; (3) Leucine Zipper: human c-Jun, Hong *et al.* (2017), 5T01; (4) Helix-Turn-Helix: λ-repressor operator – *Escherichia* λ-phage, Beamer and Pabo (1992), 1LMB; (5) Helix-Loop-Helix: E47/NeuroD1 protein, Longo *et al.*, (2008), 2QL2.

1.3.1.1 Zipper-type Domains: The Leucine Zipper and Helix-Loop-Helix

The zipper-type domain family consist of two subfamilies: the leucine zipper and the HLH. The leucine zipper consists of a pair of extended α-helices exhibiting a conserved seven-amino acid (heptad) motif repeated along the length of the helix. TFs in humans containing leucine zippers commonly display five or six heptads, although leucine zippers containing between three and nine heptads have been described (Vinson *et al.*, 2006). Within each heptad, a leucine residue typically exists at the fourth residue position – leading to the regular interspersion of leucine along the helix – while a hydrophobic residue generally occupies the first residue position. Dimerisation of the leucine zipper domain is achieved through "zippering" of residues one through four, resulting in the formation of a hydrophobic core within a coiled-coil architecture. Upon the completion of dimerisation, the adjacent, positively charged N-terminus of the coiled-coil is brought into close enough proximity to allow for sequence recognition and subsequent DNA binding (O'Shea *et al.*, 1991; Pu and Struhl, 1993). Through this mechanism, homo- and heterotypic dimer combinations can form, allowing for more complex and intricate transcriptional regulation.

The HLH resembles the leucine zipper by, too, exhibiting an extended α-helical structure. However, it differs by the presence of a loop within the motif resulting in a discontinuous, helical architecture. The presence of the loop allows for greater helical flexibility with regard to positioning along the DNA strand. The C-terminus of the HLH motif consists mainly of positively charged residues and is inserted into the major groove of the DNA, while the loop stabilises the protein-DNA complex through the creation of additional contacts with the DNA backbone (Massari and Murre, 2000). Similarly, as in the case of the leucine zipper, the HLH is able to both homo- and heterodimerise, further increasing combinatorial control over transcriptional activity (Funnell and Crossley, 2012).

1.3.1.2 The Helix-Turn-Helix Domain

The HTH motif is a recognition element found in TFs of both prokaryotes and eukaryotes (Harrison, 1990). The motif exclusively binds DNA through insertion in the major groove of DNA. Sequence recognition is conducted by the "recognition (or probe) helix" which forms direct contact between the amino acids side chains of the domain and the nucleotide bases in the DNA sequence (Luscombe *et al.*, 2000). Additionally, non-specific binding may occur

between the amino acid side chains and the DNA backbone or minor groove, facilitated electrostatically or by means of bridging water molecules (Luscombe *et al.*, 2000). The HTH motif generally exists as a cluster of three to six α-helices which overlap to provide a stabilising hydrophobic core (Luscombe *et al.*, 2000). While the HTH group of TFs comprises of sixteen homologous protein families with significant structural similarity, only slight homology is observed between the HTH motifs of each family of TFs (Luscombe *et al.*, 2000). This variation allows for a greater number of distinct DNA sequences to be recognised. In addition, an array of HTH motifs, similar in structure but diverse in terms of sequence recognition, can be derived through the addition or truncation of secondary structural elements on the simple tri-α-helical HTH motif (Aravind, *et al.*, 2005) (Figure 4). As a result, variations such as the winged HTH (wHTH) and multihelical bundle can be realised (Aravind, *et al.*, 2005).

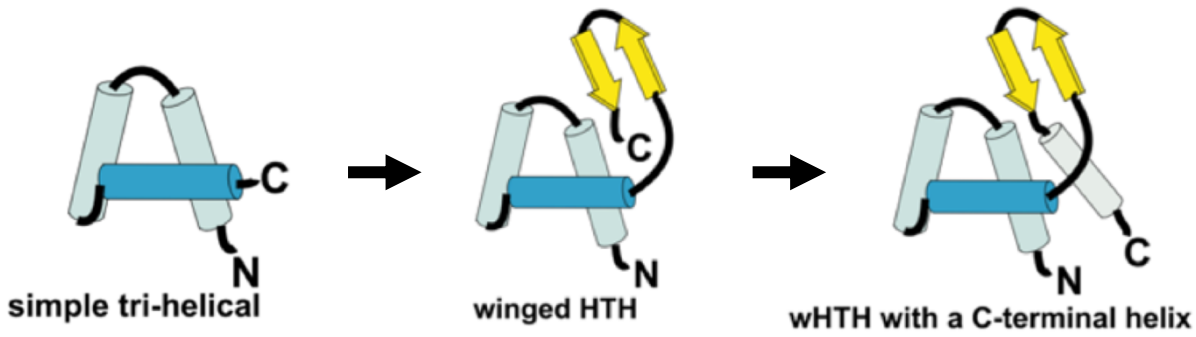


Figure 4: Secondary structural evolution from simple tri-α-helical HTH to wHTH

The tri-α-helical HTH is the foundational motif upon which all variations of the HTH are derived. The tri-α-helical HTH is designated as a winged HTH (wHTH) by the addition of a C-terminal β-sheet, forming the "wing" of the domain. The further addition of a C-terminal α-helix is a further variation on the wHTH motif – one which most closely resembles the architecture of the FOXP2-FHD. Blue cylinders indicate α-helices and yellow arrows indicate β-strands. Obtained from Aravind, *et al.*, 2005.

1.3.1.3 The Winged HTH and Forkhead Domain

The winged HTH (wHTH) is an extension of the simple tri-α-helical HTH, distinguished by the presence of two β-strand hairpins which C-terminally append the third α-helix, altogether forming a "wing" (Aravind, *et al.*, 2005) (Figure 4). In some wHTH configurations, there can

further exist a third, linker β-strand between the first and second helix of the motif. The wings of the wHTH participate in DNA backbone bonding, and can additionally associate with the minor groove of the DNA in order to increase stability and selectivity to the protein-DNA complex (Aravind, *et al*., 2005). One of the more prominent wHTH motifs is the forkhead (FOX) domain – a defining motif of the FOX superfamily of TFs.

The forkhead domain was first discovered in *Drosophila melanogaster* as a part of a protein encoded by the homeotic *fork head* (*fkh)* gene, which is implicated in the terminal, rather than segmental, development of the fly embryo (Weigel and Jackie, 1990). The Fkh protein was observed to be localised to the nucleus, leading to the proposition that it may serve a role in transcriptional regulation (Weigel and Jackie, 1990). Through *in situ* localisation studies, the same motif was subsequently discovered to be the DBD in rodent hepatocyte nuclear factor 3 (HNF-3) upon observing a high degree of sequence similarity between the DBDs of both HNF-3 and Fkh (Weigel and Jackie, 1990). Since this observation, the forkhead domain has been found to exist extensively in DNA binding proteins across the animal kingdom, in species ranging from fungi to vertebrates (Lai, *et al.*, 2003).

Structurally, the forkhead domain displays a compact α/β topology with two wings (W1 and W2), three α-helices (H1, H2 and H3) and a twisted antiparallel β-sheet consisting of three strands (S1, S2 and S3) (Gajiwala and Burley, 2000) (Figure 5). The wings are comprised of secondary structural elements that extend laterally and perpendicularly from the H3 axis. Both wings are composed of large loops which flank H3, one located between S2 and S3 (W1), and the other on the C-terminus of the DBD (W2) (Gajiwala and Burley, 2000). The presence of these wings has inspired an alternative term for proteins exhibiting the motif, namely “winged helix proteins”, as the motif imitates the likeness of a butterfly. H3, known as the recognition helix, is directly involved in DNA binding by association with the major groove of DNA. Binding of the motif is accomplished by specific DNA sequence recognition by H3, which forms hydrogen bonds with the respective nucleotides in the DNA binding sequence located in the major grove of DNA, alongside W2 associating with the minor groove by van der Waals forces (Gajiwala and Burley, 2000; Stroud *et al.*, 2006). Most of the binding interactions are electrostatically mediated, either by polar residues or through bridging water molecules, or in the case of W2, by hydrophobic interactions.

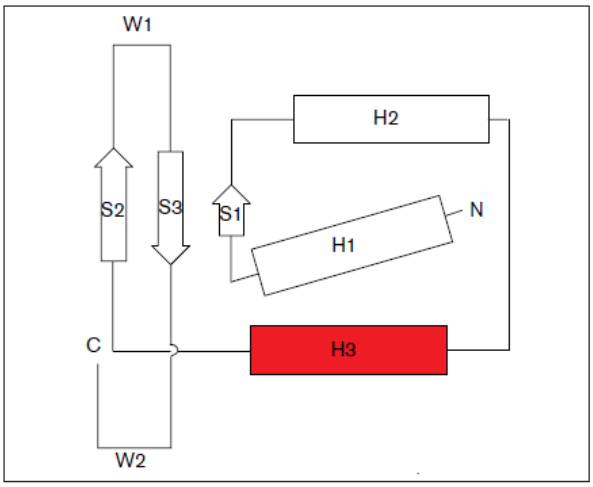


Figure 5: Topology of a typical forkhead box motif

Two wings (W1 and W2), threeα-helices (H1, H2 and H3) and a twisted antiparallel β-sheet consisting of three strands (S1, S2 and S3) can been observed to be arranged in the order N-H1-S1-H2-H3-S2-(W1)-S3-(W2)-C. H3 (red) in involved in core binding site recognition in the major groove. Obtained from Gajiwala and Burley, 2000.

1.3.2 Dimerisation of DNA Binding Domains

The homo- or heterodimerisation between zipper-type, HLH, and HTH DBDs is accompanied by a degree of conformational change which affects each participating DBD monomer. As aforementioned, the leucine zipper dimerises by the formation of noncovalent, hydrophobic interactions between interspersed leucine residues along pairing helices (Vinson *et al.*, 2006). This results in the formation of a coiled-coil architecture with either helix coiling around each other in parallel (O'Shea *et al.*, 1991; Pu and Struhl, 1993) (Figure 6). In comparison, the conformational change accompanying dimerisation of the HLH domain is more pronounced. The HLH domain is comprised of both a shorter and longer α-helix, linked by a flexible hinge-loop region. As a result of the flexibility of the hinge-loop region, dimerisation of the HLH domain is facilitated by the folding and packing of each opposing short and long helices of each monomer against one another to form a stable dimer (Massari and Murre, 2000) (Figure 6). Lastly, the HTH domain exists as a weakly-coupled dimer which is stabilised primarily by electrostatic or hydrophobic interactions, and supplemented by stacking forces, between the participant helices of interacting monomers (Funnell and Crossley, 2012).

HTH dimerisation occurs at a single interface between participating monomers, whereby residues (commonly phenylalanaine, tyrosine and arginine) intertwine to form a stable electrostatic or hydrophobic interface, hence promoting dimer stability by virtue of additional bond formation (Killykelly *et al.,* 2015; Galilee *et al.,* 2016).

1.3.3 Three Dimensional Domain Swapped Dimerisation

Three-dimensional domain swapping is defined as the process by which two or more identical monomers exchange part of their structure to form intertwined oligomers (Bennett *et al.*, 1994; Rousseau *et al.*, 2003) (Figure 6). The swapped subdomain can vary broadly in terms of simplicity and size, ranging from a simple secondary structural motif, such as a singular α-helix or β-strand, to an entire globular protein domain (Rousseau *et al.*, 2003). Each swapped domain is associated with a nonidentical hinge-loop of variable residue identity, length and flexibility that affects the overall favourability of domain swapping (Rousseau *et al.*, 2003).

Domain swapping was first observed in crystal structures of the diphtheria toxin (Bennett *et al.*, 1994; Bennett, *et al.*, 1995), whereby it was later suggested that domain swapping may not result in exclusively structural consequences, but could additionally have implications concerning misfolding, aggregation, the regulation of protein function, and the structural diversification of oligomers during evolution (Bennett *et al.*, 1994; Bennett *et al.*, 1995; Schlunegger *et al.*, 1997). Such implications may arise from the development of a secondary surface formed after the domain swapped oligomerisation event. This resulting secondary surface is absent in the free, yet undimerised monomer (which exhibits a native primary surface), but becomes apparent after domain swapping occurs (Figure 6).

Moreover, domain-swapping is not only confined to the exchange of domains between only two monomers: theoretically, any number of monomers which exhibit domain swapping can form a domain swapped oligomer such as in the case of cyclic domain swapping, or more dramatically in the case of aggregation and amyloid formation through linear (open-ended) domain swapped oligomerisation (Figure 6) (Rousseau *et al.*, 2003). Domain swapped oligomers have been observed to acquire additional functions previously absent from the protein as a monomer through the formation of secondary active surfaces, although the

emergent functions post-domain swapping differ depending on the protein in question (Vitagliano *et al.*, 1999; Gouldson *et al.*, 2000) (Figure 6).

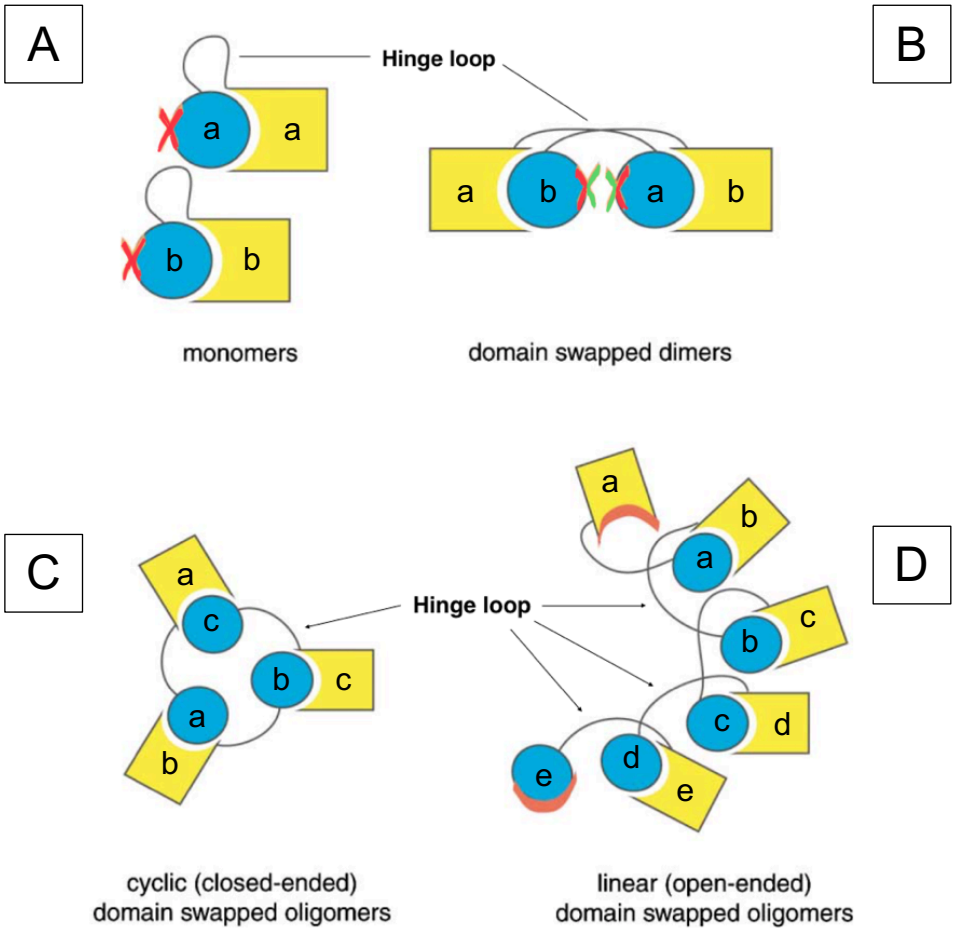


Figure 6: Schematic of domain swapped dimerisation

(A) As monomers, only primary surfaces are present (red crosses) associated with the domain capable of domain swapping (blue circle) connected to the protein (yellow square) via a hinge-loop of variable flexibility. **(B)** Once domain swapped, a secondary surface (green part of crosses) associated with the acquisition of additional function is revealed without a change in conformation in the tertiary structures of the protein (yellow squares) or the exchanged domains (blue circles). **(C)** Through domain swapped oligomerisation, higher order oligomers can arrange themselves in a cyclical fashion or **(D)** extend as a linear, unterminated chain seen in the case of amyloid formation and protein aggregation. Adapted from Rousseau, *et al.*, (2003).

In terms of energetics, three-dimensional domain swapping can be viewed as a thermodynamically unfavourable process as the rotational and translational entropy associated with domain swapping favours the formation of the monomer (Bennett *et al.*, 1994; Schlunegger *et al.*, 1997; Rousseau *et al.*, 2001; Rousseau *et al.*, 2003). This was first observed in the domain swapped dimerisation of diphtheria toxin (Bennett *et al.*, 1994) whereby the Gibbs free energy of formation of the dimer ($\Delta G_{dimerisation}$) was positive in relation to the free energy of formation of the monomer. The change in free energy was attributed to the decrease in entropy associated with domain swapped dimer formation (Bennett *et al.*, 1994).

Alongside increased entropic requirements, the formation of domain swapped oligomers is additionally affected by free energy changes generated by the rise of new intermolecular interfaces and, most importantly, by the structural conformation of the hinge-loop involved in domain swapping (Rousseau *et al.*, 2001). Indeed, the major determinant of domain swapping is the hinge-loop – the sole region involved in domain swapping which occupies an alternate form in monomeric and domain swapped oligomeric protein conformations. The flexibility (or rather the strain) of the hinge-loop can have a dramatic effect on the propensity for domain swapped oligomers to form (Rousseau *et al.*, 2003). Shortening the hinge-loop increases its strain, causing it to act like a compressed molecular spring that relieves its intrinsic strain by adopting an alternate conformation without affecting the overall structure of the protein (Rousseau *et al.*, 2003). Such shortening has been shown to increase the formation of domain swapped dimers through the destabilisation of the monomer, forcing it to adopt a more favourable conformation, as seen in staphylococcal nuclease, suc1, and CD2, among others (Rousseau *et al.*, 2001; Green *et al.*, 1995; Parker *et al.*, 1998). Furthermore, strain can be controlled through the introduction or substitution of less flexible residues, such as proline, to the hinge-loop region, inducing greater inflexibility and thus driving the process of domain swapping (Rousseau *et al.*, 2003). The lengthening of the hinge-loop has also shown the propensity to induce domain swapped oligomerisation, as seen in CI2, which forms domain swapped dimers, as well as higher order oligomers, upon the lengthening of the hinge-loop by 5 – 10 residues (Chen *et al.*, 1999). Overall, manipulating the strain on the hinge-loop either by destabilising the monomer through an induction of conformational strain, or increasing the flexibility of the hinge-loop by inserting additional residues can have an effect on the favourability of three-dimensional domain swapping.

1.3.4 The Mechanism of Three Dimensional Domain Swapped Dimerisation

Three-dimensional domain swapping has been observed to occur in a variety of proteins that perform a diverse array of functions (Rousseau *et al.*, 2012). First observed in diphtheria toxin (Bennett *et al.*, 1994), further instances of three-dimensional domain swapping have been observed across multiple organisms (Rousseau *et al.*, 2012) Examples of such include: the coat protein from Simian virus 40 (Stehle and Harrison, 1996), proteins that function in DNA recombination in *E. coli* and the regulation of the cell cycle (Khazanovich *et al.*, 1996, Bourne *et al.*, 2000), α-IPMS in *M. tuberculosis* which is involved in leucine biosynthesis (Koon *et al.*, 2004), and prion protein (Knaus *et al.*, 2001). Three dimensional domain swapping has additionally been simulated *in silico* for ribonuclease A (Esposito and Daggett, 2005).

In general, proteins that exhibit three-dimensional domain swapping have comparable energies per conformation of both the monomeric and oligomeric forms of the protein (Schlunegger *et al.*, 1997) (Figure 7). This is the case as the formation of a domain swapped oligomer is accompanied by a conformational change of the monomer, however, the energy of interactions, atom-for-atom on the whole, for either the monomer or oligomer is unchanged: only the atoms involved in domain swapping have to form new intermolecular interactions (Schlunegger *et al.*, 1997). The monomeric and domain swapped oligomeric forms of a protein are separated by an energy barrier that must be overcome for the successful formation of a stable domain swapped oligomer (Schlunegger *et al.*, 1997) (Figure 7). The magnitude of the energy barrier is dependent on three factors: firstly, the difference in translational and rotational entropy between the monomeric and oligomeric configuration of the protein; secondly, the energy change of the hinge loop and its further interactions with the oligomeric protein; thirdly, the energy difference in the formation of the secondary surface of the oligomer, that arises only from domain swapping, which is absent from the monomer (Schlunegger *et al.*, 1997). Hence, a mechanism that would promote three-dimensional domain swapping should: decrease the difference in entropy between monomer and oligomer, favour the energy change associated with the conformational change of the hinge loop, and form a secondary surface that is stabilised by attractive interactions within the surface. In addition, in order for three-dimensional domain swapping to occur, a mechanism that allows first for the partial unfolding of the monomer followed then by domain swapping must be present (Schlunegger *et al.*, 1997; Medina *et al.*, 2016). Indeed to date (Bennett *et al.*, 1994; Schlunegger *et al.*, 1997; Ha *et al.*, 2016; Medina *et al.*, 2016), all domain swapping proteins follow this model, without

thermodynamic intermediates forming, thereby following a folding mechanism denoted as $N_2 \leftrightarrow 2U$ (in terms of domain swapped dimerisation), whereby N_2 refers to the formation of a dimer and U the unfolded monomer.

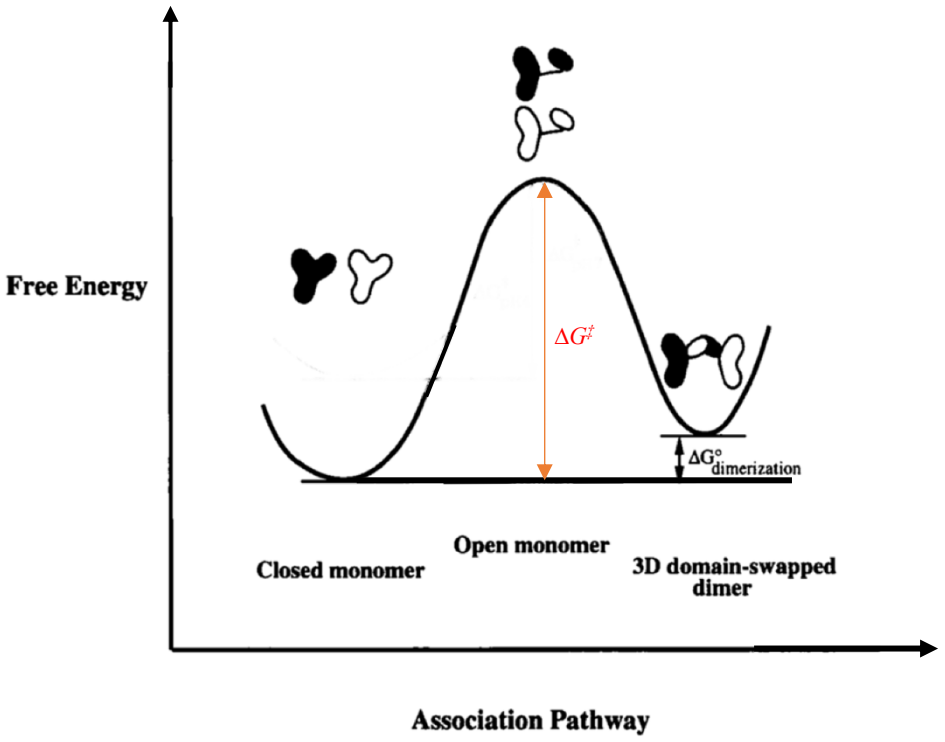


Figure 7: The free energy relationship between monomer and dimer for domain swapped dimerisation

As a monomer, a protein able to domain swap dimerise exists at a more stable, lower energy state in comparison with the three-dimensional domain swapped dimer form, although their respective minimum energies are comparable (black double sided arrow). The formation of monomer and dimer is separated by an energy barrier ($\Delta G^{\ddagger}$) (red arrow). $\Delta G^{\ddagger}$ can be affected by differences in entropy between the monomer and dimer conformations, hinge loop tension or flexibility, or enthalpic contributions arising from attractive interactions within the secondary surface formed by domain swapped dimerisation. Adapted from Bennett *et al.*, 1994.

1.4 Forkhead Box Transcription Factors

As aforementioned, transcription factors (TFs) are nuclearly localised DNA binding proteins that act to facilitate the regulation of gene expression, through both DNA association and assembly into higher order multimers. The forkhead box (FOX) family of TFs constitutes 50 members across 19 subfamilies in humans (FOXA – FOXS) (Jackson *et al.,* 2010), which are classified according to a conserved forkhead domain (FHD) showing a high degree of sequence homology (Figure 8). The FHD is a subgroup of the wHTH domain, and is the defining DBD found throughout the family (Clark *et al.,* 2003; Lehmann *et al.*, 2003) (SCOP classification number: 46832). FOX proteins are expressed ubiquitously in the body and play critical and complex roles in organogenesis, homeostasis, regulation of senescence, and cell proliferation (Vernes *et al.,* 2007; Jackson, *et al.,* 2010). Furthermore, FOX proteins exhibit pioneering transcriptional activity, having the ability to bind heterochromatic segments of chromatin during the process of cellular differentiation (Zaret and Carroll, 2011).

FOX proteins exhibit a largely versatile degree of transcriptional regulation. In mice, FOXA1 acts as a transcriptional activator in gut development, while in contrast FOXA2 displays transcriptional repression, though both are members of the same subfamily and both are involved in the same developmental processes (Kaestner *et al.,* 1999). Mutations in the forkhead domain of these proteins often display defective phenotypes; a mutation in the aforementioned proteins can lead to deficient insulin secretion in humans, leading to the development of diabetes milletus type II (Vaisse *et al.,* 1997). It is through such observed phenotypes from mutations affecting the forkhead domain that has lead to the discovery of many winged helix proteins involved in development, including multiple members of the FOX family of proteins.


Relative position →
10
20
30
40
50
60
70
80
90
100
110
120
FOXA1|169-260
FOXA2|159-252
FOXA3|116-207
FOXB1|12-103
FOXB2|13-107
FOXC1|77-168
FOXC2|71-162
FOXD1|124-215
FOXD2|126-217
FOXD3|141-235
FOXD4|104-198
FOXE1|53-147
FOXE3|71-165
FOXF1|47-138
FOXF2|99-190
FOXG1|181-275
FOXH1|32-128
FOXI1|123-217
FOXI2|102-196
FOXI3|145-239
FOXJ1|120-210
FOXJ2|66-143
FOXJ3|77-168
FOXK1|305-400
FOXK2|258-353
FOXL1|48-139
FOXL2|54-148
FOXM1|235-327
FOXN1|271-367
FOXN2|112-208
FOXN3|113-204
FOXN4|193-289
FOXO1|159-235
FOXO3|157-251
FOXO4|100-188
FOXO6|88-182
FOXP1|465-555
FOXP2|504-594
FOXP3|337-423
FOXP4|467-559
FOXQ1|119-214
FOXR1|173-272
FOXR2|192-294
FOXS1|17-108
Clustal Consensus

Figure 8

Figure 8: Multiple sequence alignment of the forkhead domain from all FOX family members (*Homo sapiens)* discovered to date

Names of members are recorded in the far-left column followed by actual range of sequence positions in each respective full-length TF; amino acid classes are coloured as follows: nonpolar - green, polar uncharged - brown, polar positively charged - red, polar negatively charged - blue, aromatic - purple, sulphur containing - yellow. Proline is coloured in grey. Relative amino acid position with respect to the positions of the multiple sequence alignment is indicated above the alignment. In the "Clustal Consensus" row: (*) indicates all residues in alignment are identical; (:) indicates conserved substitutions in alignment exhibit strongly similar properties; (.) indicates conserved substitutions in alignment exhibit weakly similar properties. Sequences were aligned with Clustal Omega (EMBL-EBI Online) (Madeira *et al.*, 2019) and formatted with BioEdit v7.2.5 (Hall, 2011). The FOX-FHD has been considerably conserved in terms of sequence over the array of FOX family members. More than 55% of residues across all members display at least 75% similarity in terms of amino acid class (nonpolar, polar uncharged or polar charged) and occupy the same relative sequence position (indicated by blue arrow); 17% of residues across at least 75% of members show identical amino acid class at the same relative position (indicated by (:) in Clustal Consensus); and four residues (I28, W51, L59, F65) occur in all members at the same relative position (indicated by (*) in Clustal Consensus). Notably, the residues in the segment associated with the recognition helix H3 (relative position 51 – 59) shows almost 100% class similarity, most notably R56 and H57 which are implicated in DNA binding through direct nucleotide base association (Stroud *et al.*, 2006; Bandukwala *et al.*, 2011). Specifically, in terms of the FOXP subfamily (black box), a P37A substitution is present, which has been suggested to allow domain swapped dimerisation to exist among in subfamily members (Li *et al.*, 2004).

1.4.1 FOXP Subfamily

The FOXP subfamily is divided into four members which are involved (in humans) in maintaining stem cell pluripotency (FOXP1) (Katoh and Katoh, 2005), the acquisition and development or language and vocalisation (FOXP2) (Shu *et al.,* 2007), maintenance of the immune system (FOXP3) (Wu *et al.,* 2006), and cellular differentiation and oncogenesis (FOXP4) (Teufel *et al.,* 2003). While FOXP3 is exclusively involved in immune system regulation, FOXP1, FOXP2 and FOXP4 additionally participate in the development of brain,

lung and gastrointestinal tract tissue (Katoh and Katoh, 2005; Shu *et al.,* 2007; Sakaguchi *et al.*, 2011). In terms of topology, all four members contain a zinc finger, leucine zipper and conserved forkhead domain. FOXP1 and FOXP2 both additionally exhibit an N-terminal glutamine-rich domain (polyQ domain) and C-terminal acidic-residue rich tail.

The FOXP subfamily is unique among the family of FOX TFs as its members have shown the ability to dimerise by virtue of their constituent FHDs (Li *et al.*, 2004). While members were first assumed to form homo- and heterotypic dimers mediated solely by interactions between leucine zippers (Li *et al.*, 2004), low resolution chromatography experiments have revealed the ability of the FOXP1-, FOXP2- and FOXP3-FHD to form homodimers *in vitro* in the presence (FOXP1 and FOXP2) and absence (FOXP3) of DNA (Stroud *et al.*, 2006; Bandukwala *et al.,* 2011; Medina *et al.*, 2016).

1.4.2 FOXP2

FOXP2 was first identified in two separate cases (a British family – referred to as the "KE family", and a child by the alias "CS") where individuals displayed severe monogenic speech and language disorders (Lai *et al.,* 2003) These individuals, carrying a heterozygous *FOXP2* mutation which was later found to code for a loss-of-function mutant, displayed a spectrum of disorders affecting vocalisation and speech coordination. This syndrome, termed developmental verbal dyspraxia (DVD), impacts the coordinated movements needed for enunciation, and results in inconsistencies in complex speech (Lai *et al.,* 2003). Studies based on mouse models which expressed R552H FOXP2 (an equivalent to the loss-of-function mutant in humans) have further uncovered the physiological and developmental roles of FOXP2 (Groszer *et al.*, 2008; Enard *et al.*, 2009).

FOXP2 acts as a transcriptional repressor and has been shown to contribute to brain and alveolar development (Shu *et al.,* 2007; Lai *et al.,* 2001). As aforementioned, it has been found to be essential in the acquisition of language and the proper development of speech in humans (Lai *et al.,* 2001). FOXP2 is expressed in many regions of the brain such as the inferior frontal cortex and basal ganglia where it plays further developmental roles involved neural maturation (Spiteri *et al.,* 2007). In addition, FOXP2 has been implicated in the regulation of metastasis of breast cancers, various osteosarcomas, and B-cell lymphoma (Cuiffo *et al.*, 2014;

Wong *et al.*, 2016). It was observed by Cuiffo *et al.*, 2014, that repression of FOXP2 expression by mesenchymal stem cell deregulated miRNA promotes breast cancer stem cell morphology and, in turn, leads to less effective prognoses of individuals at higher risk of malignant breast cancer (Cuiffo *et al.*, 2014). As FOXP2 is implicated in both the development of neural networks that are concerned with speech, as well as tumourigenesis and metastasis in the case of malignant breast cancer, it appears a potential candidate for novel gene therapy.

1.4.3 FOXP2 Structure

FOXP2 is comprised of five distinct domains, of which three participate in DNA binding (Lai *et al.*, 2011). From the N- to C-terminus these domains include: a glutamine rich (poly-Q) region, zinc-finger domain, leucine zipper domain, FHD, and acid rich tail (Figure 9). The leucine zipper has been observed to modulate homo- and hetero-typic dimerisation between FOXP1, FOXP2 and FOXP4 (Li *et al.*, 2004), while the FHD has been observed to dimerise by domain swapping (Stroud *et al.*, 2006) (Figure 9).

The FOXP2-FHD predominantly follows the canonical form of the winged-helix motif in terms of structure, displaying a compact α/β topology with two wings (W1 and W2) which form large loop structures, three stacking α-helices (H1, H2 and H3) and a twisted antiparallel β-sheet consisting of three strands (S1, S2 and S3) (Galiwala and Burley, 2000; Stroud *et al.*, 2006) (Figure 9). As with other winged helix motifs, FOXP2-FHD binds to DNA via H3 which identifies the cognate DNA binding sequence and participates in binding by association with the major groove, while W2 associates with the minor groove via electrostatic interactions and dispersion forces (Galiwala and Burley, 2000; Stroud *et al.*, 2006). In comparison to the canonical form of the FHD found across the FOX protein family, FOXP2-FHD shows several secondary structural differences: W1 is truncated to a simple type I turn, W2 forms a fourth α-helix (H5) and a fifth helix (namely a 3_{10}-helix) (H4) occurs between H2 and H3 in W2 (Stroud *et al.*, 2006) (Figure 9). These variations allow the FOXP2-FHD to participate in van der Waals interactions with the DNA backbone and adjacent minor groove, with H5 forming hydrogen bonds to the phosphate backbone upstream of the core recognition site, and W2 forming hydrogen bonds with the backbone of the recognition site (Stroud *et al.*, 2006). The FOXP2-FHD is able to bind DNA as both a monomer and homodimer (Stroud *et al.*, 2006), although the effect of the presence DNA on its propensity to form dimers is, as of yet, unknown.

In a physiological context, FOXP2-FHD may additionally function as a linker between two opposing DNA strands, participating in DNA-looping and mediating interchromosomal association as has been observed in the case of FOXP3-FHD (Stroud *et al.*, 2006; Rousseau, *et al.*, 2012; Chen *et al.*, 2015).

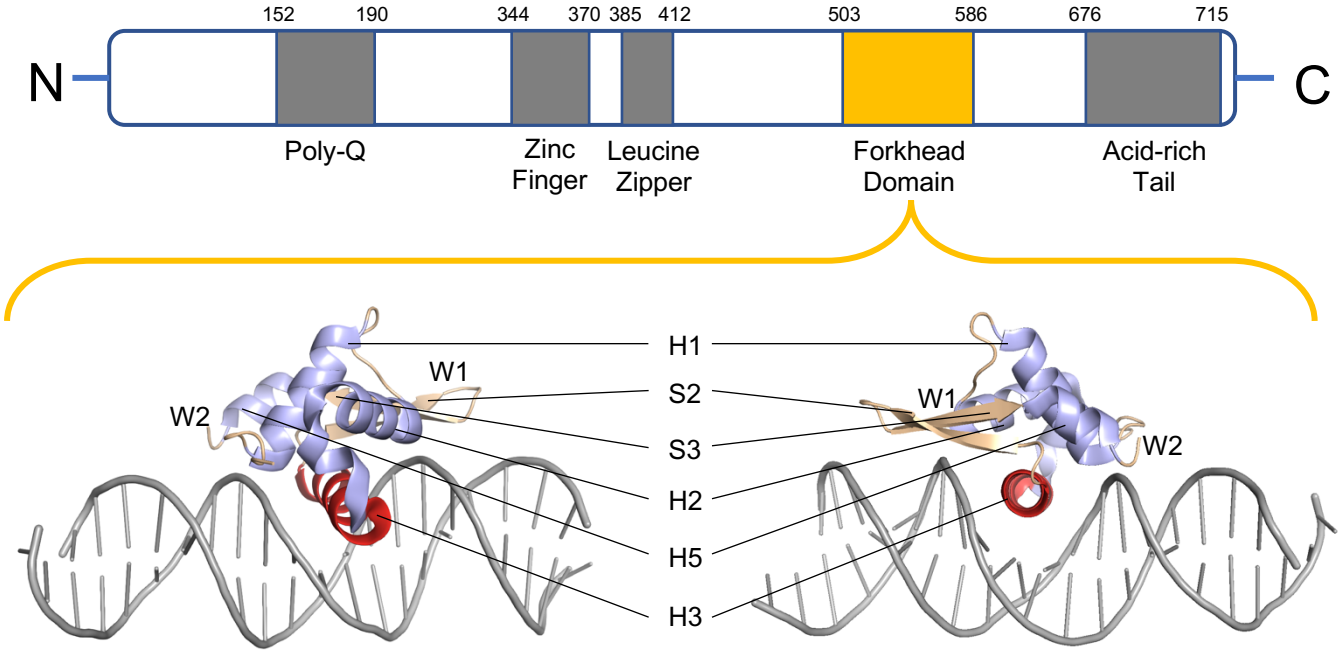


Figure 9: FOXP2 schematic, and cartoon representation, of the FOXP2-FHD bound to DNA with FOX cognate sequence

Full length FOXP2 consists of (from N → C terminus) a glutamine rich region, zinc finger domain, leucine zipper domain, FHD (orange), and acid rich tail. A crystal structure of monomeric FOXP2 (PDB: 2A07) (Stroud *et al.*, 2006) is bound to DNA displaying the aspects of a typical forkhead box motif (rendered in PyMOL, v2.3.4 (DeLano Scientific, 2019)). α-helices (H1, H2 and H5) are coloured in blue; the DNA recognition helix (H3) is coloured in red. The two wings (W1 and W2) can be seen alongside the DNA strand (grey), which act to stabilise the binding interaction between the motif and the DNA backbone. β-strands (S2 and S3) are coloured in beige; S2 and S3 form W1 of the motif, located between H3 and the C-terminus. H5 exists at the start of W2.

1.4.4 Three Dimensional Domain Swapping in the FOXP Subfamily and FOXP2

Domain swapped dimerisation has been documented for the FHDs of FOXP1 – 3 (Stroud, *et al.*, 2006; Bandukwala *et al.*, 2011; Medina *et al.*, 2016). FOXP2-FHD and FOXP3-FHD have been observed to form domain swapped oligomers in solved crystal structures (Stroud, *et al.*, 2006; Bandukwala *et al.*, 2011). In addition, FOXP1-FHD has been shown to form dimers at micromolar concentrations in solution through an energetically favoured folding mechanism (Medina *et al.*, 2016). A three state mechanism ($N_2 \leftrightarrow 2I \leftrightarrow 2U$) was described for FOXP1-FHD folding, whereby during unfolding the dimer was observed to dissociate into partially stable monomeric intermediates before fully unfolding (Medina *et al.*, 2016). This mechanism is not observed for other domain-swapping proteins, in that their unfolding pathways are not populated by stable intermediate states, and unfolding follows a two-state mechanism (Bennett *et al.*, 1994; Schlunegger *et al.*, 1997; Medina *et al.*, 2016). This observation suggested that FOXP1 had a greater propensity for dimer formation, however not as a result of hinge-loop flexibility as required by other domain swapping proteins, but rather because of increased monomer stability (Medina *et al.*, 2016). The FOXP1 monomer was observed to participate in domain swapped dimerisation by virtue of first forming a native-state-like intermediate, proceeded by rapid domain swapping of unfolded secondary structural elements, unlike other domain swapping proteins, which require a close-to-fully unfolded state in order to successfully participate in domain swapping (Rousseau *et al.*, 2012; Medina *et al.*, 2016).

FOXP3-FHD has been proposed to be the most stable domain swapped dimer of the FOXP subfamily (Bandukwala *et al.*, 2011). FOXP3-FHD dimerises by an exchange of α-helices H1 and H2, and β-strands S2 and S3. These motifs are folded upon helix H3 and strands S2 and S3 (Bandukwala *et al.*, 2011), whereby H2 is extended towards the opposing monomer and shifts H3, later to be stabilised by interactions with H1 and H2 of the opposing monomer (Bandukwala *et al.*, 2011) (Figure 10). The dimer is further stabilised through two interfaces generated after domain swapping: the first is a serial network of aromatic residues (Tyr364, Phe367, Phe373, Phe374, and Trp381) between both H2 helices of each monomer (which are known to be highly conserved between members of the FOX TF family), and the second existing as a network of hydrophobic interactions within the hydrophobic interior of the domain swapped dimer between additional residues (Phe340, Leu345, Arg347, Trp348, Trp366, and

Met370) (Bandukwala *et al.*, 2011). The serial aromatic interface is analogous to a similar interface formed in FOXP2-FHD domain swapped dimers (constituted by residues Trp533, Phe538, Ala539, Tyr540 and Phe541) (Figure 10) (Stroud *et al.*, 2006).

Phe373 and Arg347 in FOXP3-FHD are non-homologous with respect to the ordered sequences of other FOXP-FHD members (Figure 8). As a result, these residues have been assumed to offer a critical residue-specific contribution towards the stabilisation of the domain swapped dimer of FOXP3-FHD. This is justified as both residues do not play a role in DNA binding, and the mutation of either residue to a small, nonpolar residue such as alanine results in dimerisation disruption (Bandukwala, *et al.*, 2011). This is considerably true of Phe373, whereby an F373A loss-of-function mutation has been observed to result in the impedance of domain swapped dimerisation altogether (Bandukwala, *et al.*, 2011). In addition, domain swapped dimerisation is seen to be promoted by a phenylalanine residue at the equivalent H2 position in FOXP2-FHD. The Y540F substitution in FOXP2-FHD results in an approximately ten-fold increase in dimer formation in comparison to wild-type FOXP2-FHD in solution, as a result of increased hydrophobicity contributed by the phenylalanine residue within the dimer serial aromatic interface (Perumal *et al.*, 2015).

In FOXP2, the FHD has been shown to exchange α-helices H1 and H2, and β-strands S2 and S3 to form a domain swapped dimer (Stroud *et al.*, 2006). In the domain swapped dimer, H2 is extended through the turn connecting H2 and H3 originally present in the monomer, resulting in an elongated 15 residue helix replacing the shorter helix observed in the monomer. In this region exist residues 538 – 541 (FAYF) which are highly conserved across all FOX members (Figure 10), barring residue Ala539 which is instead occupied by proline in almost all other FOX members (Figure 10). As proline residues confer rigidity to the peptide backbone and discourage helix formation, it has been suggested that substitution to alanine in the FOXP subfamily has been instrumental in its member's increased domain swapped dimerisation propensity (Stroud *et al.*, 2006; Rousseau *et al.*, 2013).

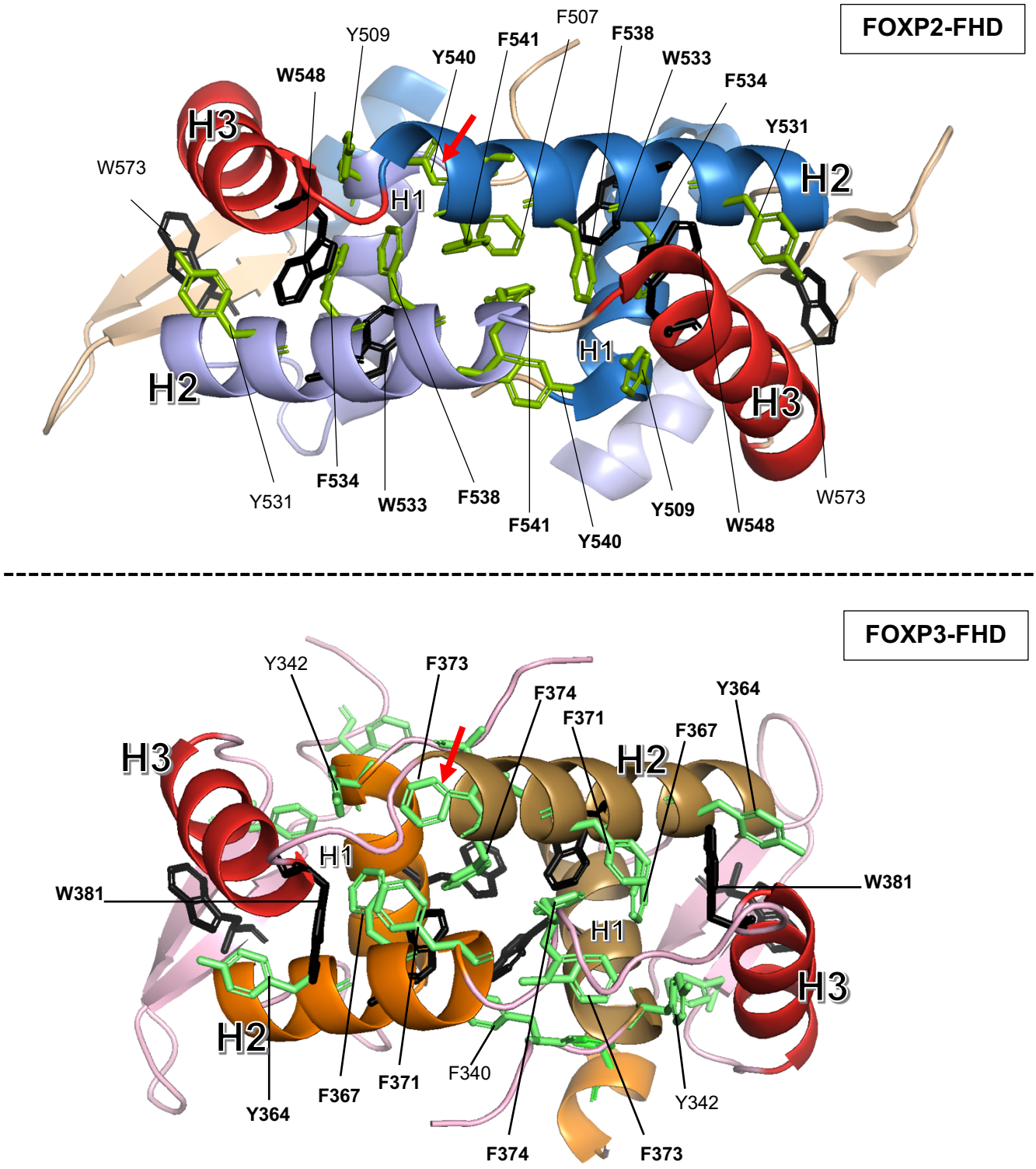


Figure 10: The serial aromatic interfaces of the FOXP2-FHD and FOXP3-FHD

An interface stabilising the FOXP2-FHD dimer involves the interaction of serially located aromatic residues found in H2 of both monomers. The residues that form hydrophobic interactions along the extent of the helix include: Try509, Trp533, Phe534, Phe538, Try540, Phe541 and Trp548 (bold). Phe538, Tyr540 and Phe541 are constituents of the FAYF region

of H2, a sequence which is are highly conserved across the members of the FOXP protein family. For the FOXP3-FHD, a similar stabilising interface exists contributed to by residues: Tyr364, Phe367, Phe371, Phe373, Phe374, and Trp381 (bold). In the case of FOXP3-FHD, phenylalanine exists in place of tyrosine which is present in FOXP2-FHD at the same relative position on H2 (red arrows).

1.5 Research Problem Identification and Aim

In the literature to date, FOXP2-FHD has been observed to exist as a domain swapped dimer in the presence of DNA in a solved crystal structure (PDB Accession Code: 2A07) (Stroud *et al.*, 2006). However, investigations into wild-type FOXP2-FHD dimerisation in solution have shown the protein to not readily dimerise, and if so, only at considerably high concentrations (Perumal *et al*. 2015; Blane and Fanucchi, 2015; Morris and Fanucchi, 2016). Quaternary structural analyses by size exclusion chromatography have shown wild-type FOXP2-FHD to have a K_d for dimer dissociation into monomer in the millimolar range, hence the protein exists predominantly as a monomer in solution at workable experimental concentrations (Perumal *et al.*, 2015). However, an increase in dimer formation for FOXP2-FHD has been engineered through the introduction of substitution mutations affecting the domain swapped dimer interface (Perumal *et al.*, 2015; Morris and Fanucchi, 2016). A variant of FOXP2-FHD – F541C – has been observed to exist as a dimer in solution exacted by the formation of a disulphide bond at the dimer interface, locking the protein in a dimeric conformation (Morris and Fanucchi, 2016). In addition, a Y540F substitution variant was observed to promote dimer formation of the FOXP2-FHD. This substitution resembles the H2-located phenylalanine residue uniquely present in FOXP3-FHD among the FOXP-FHDs, which acts to stabilise the domain swapped dimer as a result of increasing the hydrophobicity of the serial aromatic dimer interface (Figure 10) (Bandukwala *et al.*, 2011; Perumal *et al.,* 2015). In the Y540F variant, the substitution resulted in a decrease of the K_d for FOXP2-FHD dimer dissociation into monomer ~10-fold, allowing for the existence of FOXP2-FHD dimer at micromolar concentrations in solution (Perumal *et al.*, 2015). The dimerisation propensity of FOXP2-FHD additionally appears to be promoted by the presence of DNA. A truncated form of FOXP2, which encompasses the amino acid sequence from the FHD up until the final C-terminal amino acid of the protein, weakly exhibits dimerisation in the presence of DNA but does not form dimers in its absence (Häußermann *et al.*, 2019). Furthermore, the FOXP2-FHD

F541C variant has been observed to exist as a dimer in the presence of DNA (Morris and Fanucchi, 2016). FOXP2-FHD is not unique among the FHDs of the FOXP family in its ability to form domain swapped dimers (Bandukwala *et al.*, 2011; Medina *et al.*, 2016). FOXP1-FHD has been observed to exist as a dimer at micromolar concentrations in solution and is presumed, although has not been confirmed, to form domain swapped dimers in solution (Medina *et al.*, 2016). Moreover, FOXP3-FHD has been observed to exist as a domain swapped dimer in a solved crystal structure (PDB Accession Code: 3QRF) (Bandukwala *et al.*, 2011). Additionally, FOXP3-FHD assumes a dimeric conformation in the absence of DNA – although the dimer was formed in the presence of the protein crosslinker disuccinimidyl subterate (DSS) (Bandukwala *et al.*, 2011). Indeed, in the cases of the FOXP1 – P3 FHDs, the formation of a domain swapped dimer is speculated to have physiological importance through its participation in interchromosomal association through DNA binding by both dimer subunits (Stroud *et al.*, 2006; Bandukwala *et al.*, 2011; Chen *et al.*, 2015; Morris and Fanucchi, 2016; Häußermann *et al.*, 2019).

While it is evident that the FHDs of FOXP1 – P3 exhibit the capability to domain swap dimerise, the process is nonetheless nonspontaneous (Bennett, *et al.*, 1994). In order for domain swapped dimerisation to progress, a protein must occupy a partially unfolded state preceding dimerisation so that secondary structural motifs may exchange and become intertwined (Rousseau, *et al.*, 2013). As folding occurs, there is competition between the probability of the unfolded chain forming a monomer by intrachain hydrophobic collapse, and the probability of interchain interaction between monomers resulting in the formation of a domain swapped dimer (Rousseau, *et al.*, 2013). In terms of the FOXP2-FHD, as the protein folds, H1 and H2 may be exchanged via domain swapping to form a dimer which is stabilised through interactions between the aromatic residues located along H1 and H2 of either monomer (Figure 10). Alternatively, FOXP2-FHD may adopt a monomeric conformation through rapid coalescence of H1 and H2 driven by hydrophobic collapse, whereby both helices fold upon each other to form a stable hydrophobic core incapable of participating in domain swapping. In this case, the aromatic residues of H1 and H2 act to stabilise the FOXP2-FHD monomer through interactions within the hydrophobic core (Figure 9 and 18). The probability of the formation of either dimeric or monomeric FOXP2-FHD is hence determined by the folding pathway the protein undertakes from its unfolded to native state. It is necessary to explore the folding pathway and structural stability of FOXP2-FHD in order to elucidate its propensity to participate in domain swapped dimerisation.

Therefore, this investigation aims to probe the folding mechanism and stability of FOXP2-FHD. As aforementioned, FOXP2-FHD has not been observed to readily dimerise in solution although the FHDs of related FOXP protein members, FOXP1-FHD and FOXP3-FHD, do display such behaviour. By determining the folding mechanism of FOXP2-FHD, this investigation aims to elucidate its propensity to form dimers, as its folding pathway determines whether the protein will adopt a monomeric or dimeric conformation. The folding mechanism of wild-type FOXP2-FHD will be contrasted against the folding mechanism of a Y540F variant. This variant has been observed to form dimers at micromolar concentrations in solution (Perumal, *et al.,* 2015). A comparison of the variant's folding pathway with that of wtFOXP2-FHD will assist in revealing the hence observed decreased propensity of wtFOXP2-FHD to form domain swapped dimers in solution. Overall, a description of the folding mechanism of FOXP2-FHD will allow a better understanding of how its structure contributes to its ability to form domain swapped dimers, which in turn has further implications in understanding its role of transcriptional regulation within the cell.

1.5.1 Objectives

The specific objectives of this investigation were to:

1. Transform a suitable bacterial expression system with recombinant DNA plasmid coding for expression of wtFOXP2-FHD protein
2. Generate a recombinant DNA plasmid coding for the Y540F variation of wtFOXP2-FHD by site-directed mutagenesis, followed by transformation of a suitable bacterial expression system for the expression of the mutant
3. Overexpress and purify soluble wtFOXP2-FHD and Y540F mutant protein
4. Determine the purity, homogeneity and molecular weights of purified wtFOXP2-FHD and Y540F mutant protein by discontinuous tricine SDS-PAGE
5. Ascertain whether DNA binding occurs for wtFOXP2-FHD and the Y540F mutant, and investigate whether the proteins form dimers in the presence of DNA, by electrophoretic mobility shift assay (EMSA)
6. Assess the quaternary structure of wtFOXP2-FHD and the Y540F mutant in solution, and hence calculate their respective dissociation constants, by analytical size exclusion chromatography (SEC)

7. Characterise the secondary structure of wtFOXP2-FHD and the Y540F mutant by far-UV circular dichroism (CD) spectropolarimetry
8. Characterise the tertiary structure of wtFOXP2-FHD and the Y540F mutant by monitoring of intrinsic tryptophan fluorescence emission
9. Determine the folding mechanism of wtFOXP2-FHD and the Y540F mutant by denaturant driven equilibrium unfolding studies, monitored by far-UV CD spectropolarimetry and intrinsic tryptophan fluorescence emission
10. Assess the reversibility of unfolding and percentage protein recovery for wtFOXP2-FHD and the Y540F mutant, and hence calculate the thermodynamic parameters of unfolding for each protein
11. Assess the thermal stability of wtFOXP2-FHD and the Y540F mutant by thermal denaturation monitored by far-UV CD spectropolarimetry

Chapter 2: Materials and Methods

2.1 Materials

T7 Competent *E. coli* and XL-10 Gold Ultracompetent *E. coli* cells were obtained from New England Biolab, USA. GeneJET™ Plasmid Miniprep Kit and PageRuler™ Unstained Protein Ladder was obtained from Thermo Fisher Scientific, USA. The Quikchange™ Site-Directed Mutagenesis Kit was obtained from Stratagene, USA. Ampicillin salt and isopropyl β-D-1-thiogalactopyranoside (IPTG) was obtained from Melford, UK. SYBR Gold Nucleic Acid Gel Stain was obtained from Invitrogen, USA. Lysozyme was obtained from Sigma Aldrich, USA. Deoxyribonuclease I was obtained from Merck, Germany. Duplex DNA oligonucleotide was obtained from WhiteSci Scientific, South Africa. HisTrap™ High Efficiency column and HiLoad® 16/600 Superdex® 75 pg Prep Grade column were obtained from GE Healthcare Bio-Sciences, Sweden. Gel Filtration Low Molecular Weight Kit was obtained from GE Healthcare Life Sciences, USA. All other reagents used were of analytical grade.

2.2 Methods

2.2.1 Plasmid Preparation and Transformation

2.2.1.1 Plasmid Amplification

The coding sequence of wild-type human FOXP2-FHD, corresponding to residues 504 – 594, was codon optimised (GenScript, USA) and inserted into the multiple cloning site of a pET-11a plasmid (Novagen, Germany), under control of an inducible T7 promoter. An N-terminal hexahistidine tag was incorporated for application in downstream purification. Synthesised wtFOXP2-FHD plasmid was amplified in XL-10 Gold Ultracompetent *E. coli* cells for further downstream transformation and site-directed mutagenesis. Cells were transformed with the wtFOXP2-FHD plasmid by heat shock at 42 °C for 75 seconds, followed by immediate setting on ice for 2 minutes and pelleting. Pelleted cells were resuspended in super-optimal broth with catabolite repression (SOC) medium, prewarmed to 37 °C, and incubated with shaking (230 rpm) for one hour at 37 °C. Following incubation, cells were plated on LB agar containing 0,1 mg/ml ampicillin, and further incubated at 37 °C for 20 hours.

LB media containing 0,1 mg/ml ampicillin was inoculated with a single colony of transformed cells and cultured overnight at 37 °C with vigorous shaking. wtFOXP2-FHD plasmid was extracted from cultured cells, and purified, using the GeneJET™ Plasmid Miniprep Kit (Thermo Fisher Scientific, USA) as per protocol provided. Extracted plasmid was sequenced (Inqaba Biotechnical Industries, South Africa) to verify the insert and to confirm the coding sequence was preserved. Extracted wtFOXP2-FHD plasmid was stored in sterile deionised water at -20 °C and transformed XL-10 Gold cells were stored as a suspension in sterile aqueous glycerol solution (80 %v/v) at -80 °C.

2.2.1.2 Site-directed Mutagenesis

Site-directed mutagenesis is a technique that introduces an intentional change to the DNA sequence of a gene in order to elicit an amino acid substitution in the final recombinant protein. This allows for probing the importance of the substituted residue in terms of protein structure/function. The method requires the synthesis of DNA primers which contain the desired mutation and are complementary to the template DNA around the mutation site for successful hybridisation with the gene of interest (Papworth *et al.*, 1996). Following PCR, the amplified gene will contain the mutated site, whereby the gene can be subsequently introduced in a vector and cloned.

The wtFOXP2-FHD plasmid was used as a template for generating the Y540F substitution mutant. LB media with 0,1 mg/ml ampicillin was inoculated with transformed XL-10 Gold cell culture stock and incubated with shaking at 37 °C for 16 hours. wtFOXP2-FHD plasmid was extracted using the GeneJET™ Plasmid Miniprep Kit for use in high-fidelity PCR site-directed mutagenesis.

Primers necessary for the introduction of the desired Y540F mutation in the FOXP2-FHD insert were designed using PrimerX (online) (http://www.bioinformatics.org/primerx) and checked for possible hairpin and loop structure formation prior to synthesis (Gene Runner v3.0.1, Hastings Software Inc., NY, USA). Primers were designed to contain a centred mutation site (bold, underlined) flanked on both sides by bases complementary to the template sequence:

Forward: 5`-GTTTACCCGTACGTTTGCAT**T**TTTCCGTCGCAACGCGGCC-3`
Reverse: 5`-GGCCGCGTTGCGACGGAAA**A**ATGCAAACGTACGGGTAAAC-3`

Primers were synthesised (Inqaba Biotechnical Industries, South Africa) and site-directed mutagenesis was conducted using the Quikchange™ Site-Directed Mutagenesis Kit (Stratagene, USA) according to the protocol provided. The PCR product was sent for sequencing (Inqaba Biotechnical Industries, South Africa) followed by pairwise sequence alignment using the EMBOSS Water online server (EMBL-EBI) (Madeira *et al.*, 2019) against the wtFOXP2-FHD coding sequence to confirm mutagenesis was successful. Success was indicated by the presence of the targeted nucleotide substitution, and the absence of additional, undesired mutations which may have arisen during the PCR reaction.

2.2.1.3 Transformation

T7 Competent *E. coli* cells were used as an expression system for recombinant wtFOXP2-FHD and Y540F mutant protein. Transformed T7 cells were prepared using the amplified wtFOXP2-FHD and mutant plasmids according to the above protocol (section 2.2.1.1), and stored as a suspension in sterile aqueous glycerol solution (80 %v/v) at -80 °C.

2.2.2 Recombinant Protein Overexpression and Purification

In order to procure sufficient protein for downstream assays, manipulation of gene expression in an organism to overexpress the desired protein is necessary. By introducing a vector containing a recombinant gene in proximity to a promoter which can be intentionally and variably activated, overexpression of a protein of interest can be elicited. Such expression can be conducted in a variety of cells, ranging from prokaryotic to insect and mammalian cells (Khan, 2013). The degree of protein overexpression, conservation of native structure, and solubility of the overexpressed protein will vary depending on the expression system used. Commonly, an *E. coli* expression system is utilised for overexpression by virtue of its robustness, dependability, high expression yield, and cost effectiveness. However, if a recombinantly expressed protein requires additional levels of modification during expression, such as post-translational modifications, or if solubility is unattainable in a prokaryotic

expression system, an insect or mammalian cell expression system would be preferred (Khan, 2013).

Both wtFOXP2-FHD and Y540F mutant protein were overexpressed under identical conditions and according to the same protocol: 2x YT media (1,6 %w/v tryptone, 1 %w/v yeast extract, 0,5 %w/v sodium chloride) containing 0,1 mg/ml ampicillin was inoculated with transformed T7 cell stock and incubated with shaking (190 rpm) at 37 °C until OD_{600} measured 0,8 AU. Upon reaching OD_{600} = 0,8 AU, cultures were set on ice for 10 minutes followed by induction with 0,6 mM IPTG. After induction, cultures were incubated with shaking at 18 °C for 20 hours. The cultured cells were harvested by centrifugation at 5000 x *g* for 15 minutes, resuspended in binding buffer (20 mM Tris-HCl buffer at pH 7,5 containing 500 mM sodium chloride and 30 mM imidazole), and stored at -20 °C.

Prior to purification, both the wtFOXP2-FHD and Y540F mutant cell resuspensions were thawed at room temperature, and lysozyme was added to a final concentration of 0,1 mg/ml. Resuspensions were incubated at room temperature with inversion for 15 minutes to form a lysate. Lysates were sonicated at 55 % amplitude for 20 seconds, followed by resting on ice for 1 minute, repeated eight times. Following sonication, deoxyribonuclease I (DNAse I) was added to lysates, to a final concentration of 0,01 mg/ml, and lysates were incubated at room temperature for a further 30 minutes. DNAse I was added to facilitate digestion of free dsDNA in order to reduce free DNA interaction with overexpressed protein during subsequent centrifugation and purification steps, hence increasing purifiable protein yield. Lysates were centrifuged at 24 000 x *g* for 30 minutes at 4 °C to pellet cell debris and separate the soluble protein product from its insoluble fraction. Pelleted cell debris was discarded, and the supernatant used for subsequent purification.

wtFOXP2-FHD and Y540F mutant protein were purified using Ni^{2+}-charged immobilised metal affinity chromatography coupled with fast protein liquid chromatography (IMAC-FPLC). A 5 mL HisTrap™ High Efficiency column (GE Healthcare Bio-Sciences, Sweden) was initially equilibrated with five column volumes of binding buffer (20 mM Tris-HCl buffer at pH 7,5 containing 500 mM sodium chloride and 30 mM imidazole), followed by loading of supernatant onto the column at 5 mL/min. The column was re-equilibrated with ten column volumes of binding buffer in order to remove residual, non-specifically bound protein

contaminant and cellular debris, and subsequently washed with ten column volumes of salt wash buffer (20 mM Tris-HCl, 1,5 M sodium chloride, 30 mM imidazole at pH 7,5) to remove residual DNA contamination bound to the column. The column was, again, re-equilibrated with binding buffer to remove excess salt introduced from the previous salt wash step. Protein was eluted from the column with elution buffer (20 mM Tris-HCl, 500 mM sodium chloride, 500 mM imidazole at pH 7,5) and eluent fractions were collected for downstream analysis. wtFOXP2-FHD and Y540F protein fractions were dialysed against and stored in storage buffer (10 mM HEPES, 100 mM sodium chloride at pH 7,5) at 4 °C.

2.2.3 Protein Molecular Weight and Purity Determination

In order to confirm overexpression of wtFOXP2-FHD and the Y540F mutant, as well as to identify the molecular weight, purity, and homogeneity of purified protein, discontinuous sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was employed. Discontinuous SDS-PAGE is an analytical technique that allows for the separation of proteins according to their sizes and molecular weights (Laemmli, 1970). Initially, protein samples are denatured through the breaking of disulfide bonds by a reducing agent (such as β-mercaptoethanol or dithiothreitol) at a high temperature followed by saturation in sodium dodecyl sulfate (SDS). SDS is an amphipathic anion that uniformly adheres to and fully denatures the protein, and consequently applies a uniform negative charge along the stretch of the protein (Pitt-Rivers and Impiombato, 1968). Each denatured protein sample is then resolved through a discontinuous polyacrylamide gel (containing both a stacking and separation gel) of discrete pore sizes, within a buffer containing various cations and anions over a voltage gradient. The separation of the protein in the sample occurs in two stages. Firstly, as proteins in the sample migrate through the stacking gel, they are compressed into a distinct, concentrated protein layer as a result of high gel porosity and a steep voltage gradient developed between leading and trailing buffer ions (a process termed isotachophoresis). As proteins traverse the stacking gel, the compressed protein layer further sharpens until it encounters the frontier of the separating gel. At this point, and as a consequence of isotachophoresis, the sample proteins exist in a sufficiently compacted state to allow for their equally-timed permeation of the separating gel, regardless of their respective molecular weights. This protein layer then finely resolves into constituent protein bands within the separating gel (by virtue of higher percentage acrylamide composition), according to retention

effects determined by the molecular weight of each protein. After separation, each protein within the sample may be identified by their respective molecular weight by comparing their extent of retention within the gel to a ladder of known molecular weights.

The molecular weights, purity, and homogeneity of purified wtFOXP2-FHD and Y540F mutant was assessed by discontinuous tricine SDS-PAGE (Schägger, 2006). Reducing sample buffer (150 mM Tris-HCl at pH 7,5, with 12 %w/v SDS, 6 %w/v β-mercaptoethanol, 30 %v/v glycerol, and 0,005 %w/v Coomassie Blue G250) was added to protein samples, and vortexed on high intensity for 10 seconds to ensure adequate mixing of protein and reducing sample buffer. Samples were incubated for 5 minutes at 95 °C and centrifuged at 13 000 x *g* for 10 seconds prior to gel loading. Samples were resolved by electrophoresis using the Bio-Rad miniPROTEAN Electrophoresis System (Bio-Rad, USA) with 10 % separating / 4 % stacking tricine SDS-PAGE gel (Schägger, 2006). Electrophoresis proceeded initially at 40 V until reaching the frontier of the separating gel, followed by electrophoresis at 160 V for 1,5 hours, or until the dye front reached the terminus of the gel. Gels were stained overnight in Coomassie Blue R250 staining solution, and destained in 40 %v/v methanol with 10 %v/v glacial acetic acid until sufficient gel clarity was achieved. Molecular weights of migrated proteins were determined by comparison against protein standards (PageRuler^{TM} Unstained Protein Ladder, Thermo Fisher Scientific, USA). Gels were visualised using a ChemiDoc^{TM} MP (Bio-Rad, USA) and analysed using Image Lab (v5.2.1, Bio-Rad, USA).

2.2.4 Determination of Protein Concentration

The concentration of purified protein was determined by measuring absorbance at 280 nm by UV-Vis spectrophotometry, and calculated according to the Beer-Lambert Law, rearranged as:

$$c = \frac{A_{280}}{\varepsilon \cdot l} \tag{1}$$

where c is the concentration of the sample protein measured in M, A_{280} is the absorbance of the sample at 280 nm (dimensionless), ε is the molar extinction coefficient at 280 nm measured in $M^{-1}cm^{-1}$, and l is the path length of the sample measured in cm. Absorbance was measured at 280 nm, as tryptophan and tyrosine residues absorb UV light most strongly at this

wavelength, and when taken into account with their extinction coefficients, protein concentration can be accurately determined. Cysteine additionally shows slight absorbance at 280 nm. The molar extinction coefficient was calculated according the equation by Perkins, 1986:

$$\varepsilon = 5550 \cdot n_{Trp} + 1340 \cdot n_{Tyr} + 150 \cdot n_{Cys} \quad (2)$$

where n_{Trp}, n_{Tyr}, and n_{Cys} are the number of tryptophan, tyrosine and cysteine residues present in a single, monomeric protein sequence, respectively. The molar extinction coefficients for wtFOXP2-FHD and the Y540F mutant for use in concentration determination were calculated to be ε_{WT} = 22 460 $M^{-1}cm^{-1}$ and ε_{Y540F} = 20 970 $M^{-1}cm^{-1}$, respectively.

2.2.5 DNA Binding Studies

The electrophoretic mobility shift assay (EMSA) is a simultaneously qualitative and quantitative technique that allows for the study of the DNA-binding ability of proteins. EMSA follows principles similar to traditional native polyacrylamide gel electrophoresis (native PAGE). Proteins are electrophoresed through a native polyacrylamide gel of discrete pore size by the flow of buffer ions over a voltage gradient. As they traverse the gel, they exhibit varying degrees of retention depending on their intrinsic structures, sizes and charges (Garner and Revzin, 1986; Lane *et al.*, 1992). The degree of retention is higher for proteins of greater molecular weight and size, while proteins of lower molecular weight and size exhibit smaller levels of retention within the gel. However, in the EMSA, a DNA probe is introduced to each protein sample, and it is the probe which is monitored following electrophoresis, as opposed to the gel front in traditional PAGE. Sample protein in the presence of DNA, regardless of DNA-binding ability, will experience retention effects during electrophoresis. However, by detecting the location of the probe in the gel, it can be inferred whether DNA-binding by a protein had occurred since the probe will appear more greatly retained in the gel (as a result of protein binding the DNA probe and subsequent gel retention) (Garner and Revzin, 1986; Lane *et al.*, 1992). Free DNA experiences the least of gel retention effects (hence migrating fastest through the gel) while proteins binding DNA experience retention effects determined by the factors mentioned above.

EMSAs were conducted for both wtFOXP2-FHD and the Y540F mutant to ensure DNA recognition and binding function were retained post-purification. An 18-base oligonucleotide was synthesised (WhiteSci Scientific, South Africa) for use as the DNA ligand in EMSAs for both wtFOXP2-FHD and Y540F mutant protein. The sequence included the high-affinity cognate DNA binding site (bold) for wtFOXP2-FHD as described by Nelson *et al.*, 2013, (5`-AGG**TGTTTA**CTTTCATAG-3`) to ensure a stable protein-DNA complex was maintained during electrophoresis. Lyophilised oligonucleotide was dissolved in TE buffer. DNA concentration was determined by microvolume spectrophotometry (NanoDrop One UV-Vis Spectrophotometer, Thermo Fisher Scientific).

Protein-DNA samples were prepared in EMSA binding buffer (10 mM HEPES at pH 7,5 with 100 mM KCl, 1 mM EDTA, 1 mM $MgCl_2$, 0,1 mg/ml BSA, and 10 %v/v glycerol), at 4 °C and incubated for 1 hour to allow for protein-DNA complex formation to reach equilibrium at 4 °C before electrophoresis. Samples were prepared with a concentration ratio of protein : DNA incrementally increasing from 1 : 3 to 10 :1, maintaining a constant DNA concentration of 0,5 uM per sample. Gels were cast according to the protocol by Morris and Fanucchi, 2016, and electrophoresis conducted at 4 °C for 1,5 hours using the Bio-Rad miniPROTEAN Electrophoresis System (Bio-Rad, USA). Following electrophoresis, gels were stained with SYBR Gold Nucleic Acid Gel Stain, diluted 1 : 10 000 in TBE buffer, and incubated at room temperature for 2 minutes. Gels were visualised under UV light using a ChemiDoc MP™ (Bio-Rad, USA) and analysed using Image Lab (v5.2.1, Bio-Rad, USA).

2.2.6 Structural Characterisation

2.2.6.1 Quaternary Structural Characterisation

Size exclusion chromatography (SEC) is a chromatographic technique whereby proteins are separated on a basis of their hydrodynamic volume, for the purpose of purification and analysis (Hong, *et al.*, 2012). SEC is conducted by eluting a sample through a column containing resin beads of discrete pore size. Proteins sized smaller than the bead pore size diffuse through the beads, become impeded and, subsequently, elute at a slower rate. Proteins sized larger than the pore size are excluded from the beads and are impeded to a lesser degree, thereby eluting at a faster rate. As SEC separates proteins almost exclusively on the basis of hydrodynamic volume,

it can be used to not only isolate proteins of different size, but also to identify proteins of higher oligomer configurations. In addition, SEC can be used to calculate the dissociation constant (K_d) of oligomers into monomers at equilibrium to assess the stability of the oligomer under a variety of conditions such as temperature and denaturant concentration.

Analytical size exclusion chromatography (SEC) was used to determine the molecular weights for both wtFOXP2-FHD and the Y540F mutant, as well as investigate the proportion of native monomeric and dimeric species at equilibrium for both proteins. SEC was used concurrently to qualitatively assess possible differences in the hydrodynamic volume of the mutant, in comparison to wtFOXP2-FHD, resulting from Y540F substitution. The dimerisation propensity for each protein at equilibrium, dependent on their respective concentrations in solution, was exhibited as a dissociation constant (K_d), where:

$$2\,Monomer \leftrightarrow Dimer$$
$$\therefore K_d = \frac{[Monomer]^2}{[Dimer]} \tag{3}$$

Analytical SEC was conducted using a HiLoad® 16/600 Superdex® 75 pg Prep Grade column (GE Healthcare Bio-Sciences, Sweden) coupled with FPLC . The column was equilibrated with one column volume of equilibration buffer (20 mM Tris-HCl buffer at pH 7,5 containing 500 mM sodium chloride and 30 mM imidazole) overnight. The buffer contained imidazole in order to decrease undesired protein interaction with the column resin. Protein was loaded at concentrations of ~20 uM, ~75 uM and ~150 uM, and elution volumes recorded for monomeric and dimeric species. Prior to loading, exact concentrations of samples was determined by UV-Vis spectrophotometry as described above (section 2.2.4). The molecular weight for each protein species was calculated by interpolation against a standard curve, populated with molecular weights of known protein standards (Gel Filtration Low Molecular Weight Kit, GE Healthcare Life Sciences, USA). The standards consisted of conalbumin (75 kDa), ovalbumin (44 kDa), carbonic anhydrase (29 kDa), ribonuclease A (13,7 kDa) and aprotinin (6,5 kDa).

In order to determine K_d for both wtFOXP2-FHD and the Y540F mutant, the area under the curve measuring absorbance at 280 nm for each peak was integrated with respect to elution time (Äkta PrimeView Software, v. 4.0.2, GE Healthcare Bio-Sciences). The calculated area under each peak related to the total absorbance of the monomeric or dimeric species contributed

during their respective full time eluting. As the flow rate during elution was kept constant (1 mL/min), and as the sample was at equilibrium in terms of monomer / dimer presence, the proportion of absorbance contributed by each species could be represented as a ratio:

$$Absorbance\ fraction = \frac{Area\ under\ individual\ peak}{Total\ area\ under\ both\ peaks} \quad (4)$$

The absorbance fraction for each peak was multiplied by sample concentration, determined prior by UV-Vis spectrophotometry, to calculate the respective concentrations of monomeric and dimeric species eluted. The concentration of either fraction was seen as representative of the concentration of the species present at dynamic equilibrium before loading. The K_d values calculated for different samples concentrations for both wtFOXP2-FHD and the Y540F mutant was determined according to equation (3), and specific K_d for each protein calculated by linear regression of a [monomer]^2 vs. [dimer] plot.

2.2.6.2 Secondary Structural Characterisation

Far-UV circular dichroism (CD) spectropolarimetry was used to determine the secondary structural characteristics of both wtFOXP2-FHD and the Y540F mutant, as well as to determine the degree of isolated or coiled helicity for each protein.

Protein in storage buffer (10 mM HEPES, 100 mM sodium chloride at pH 7,5) was initially dialysed against sodium sulfate buffer (10 mM HEPES, 50 mM sodium sulfate at pH 7,5) to exchange out chloride ions in order to improve signal-to-noise ratio. All experiments were conducted using a Jasco J-810 Spectropolarimeter (Analytical Solutions, South Africa) at a maintained internal cell temperature of 20 °C regulated with a Peltier Thermostatted Cell Holder, using a Hellma® 2 mm quartz absorption cuvette (Merck, USA). Protein concentration used was 10 uM. Spectra were recorded over a range of 250 – 190 nm, or over wavelengths where high tension voltage measured lower than 700 V, using continuous scanning at 100 nm/min with 1 nm scanning increment. Additional scanning parameters for all measurements of both proteins were: data pitch = 0,2 nm, bandwidth = 3 nm, and response time = 1 second. Experiments were conducted in triplicate with eight technical replicates per sample and recorded in millidegree ellipticity (θ_{mdeg}). Buffer contributions were deducted from final

spectra. Triplicate measurements were averaged and data converted to mean residue ellipticity (θ_{MRE}) measured in deg.cm^2.dmol^{-1}, calculated as:

$$\theta_{MRE} = \frac{100 \cdot \theta_{mdeg}}{c \cdot n \cdot l} \quad (5)$$

where θ_{mdeg} is millidegree ellipticity measured in mdeg, c is concentration measured in M, n is number of residues in a monomer chain, and l is path length measured in cm.

2.2.6.3 Tertiary Structural Characterisation

Intrinsic tryptophan fluorescence emission monitoring was used to obtain emission spectra of both wtFOXP2-FHD and the Y540F mutant in order to characterise their tertiary structures. This was achieved by monitoring the fluorescent signal emitted by tryptophan residues at their local positions within the protein. The fluorescent signal of a tryptophan residue is dependent on the local solvent environment encompassing its indole side chain: more hydrophobically buried tryptophan residues display a hypsochromic shift in fluorescence signal, while solvent exposed tryptophan residues exhibit a bathochromic shift. By monitoring such fluorescent shifts, an emission spectrum correlating to protein tertiary structure may be obtained. Proteins were analysed in storage buffer using a Jasco FP-6300 fluorescence spectrophotometer (Analytical Solutions, South Africa). Measurements were conducted at a maintained internal cell temperature of 20 °C regulated with a Peltier Thermostatted Cell Holder, using a Hellma® 10 mm quartz fluorescence cuvette (Merck, USA). A concentration of 1,5 uM was used for each protein sample. Excitation and emission bandwidths were fixed at 5 nm. Emission spectra were recorded over a range of 280 – 500 nm for excitation wavelengths of 280 nm and 295 nm, measured separately, using continuous scanning at 200 nm/min with 1 nm scanning increment. Experiments were conducted in triplicate with three technical replicates per sample and recorded in relative fluorescence intensity (arbitrary units). Buffer contributions were deducted from final spectra and triplicate measurements were averaged.

2.2.7 Guanidine Hydrochloride Equilibrium Unfolding

The folding mechanism of a protein describes the pathway the protein undertakes in order to adopt its native state from its initial unfolded polypeptide chain. The pathway is populated with a discrete number of folded protein states at equilibrium (Walters, *et al.*, 2009). Equilibrium unfolding studies aim to denature a protein at discrete points along its folding pathway, elicited by chemical or physical denaturation (Pace, 1986). The fraction of unfolded protein is determined by spectroscopic probe, such as far-UV CD spectropolarimetry or intrinsic tryptophan fluorescence emission. Through sequentially plotting the fraction of unfolded protein over the range of denaturant conditions, a folding mechanism for a particular protein may be described (Pace, 1986).

Equilibrium unfolding studies were performed to characterise the folding mechanism, cooperativity (m), conformational stability (ΔG_{H20}) and denaturation midpoint ($[GdnHCl]_{1/2}$) of wtFOXP2-FHD and Y540F mutant protein. Unfolding studies were conducted using guanidine hydrochloride (GdnHCl) as a chemical denaturant in separate experiments monitored by far-UV CD spectropolarimetry and intrinsic tryptophan fluorescence emission. GdnHCl was used as it has a considerably stronger denaturation effect in comparison with urea (Morjana *et al.*, 1993). High GdnHCl concentrations more reliably ensure complete denaturation of protein compared to urea, thereby allowing for proper computation of thermodynamic parameters which depends on the establishment of an identifiable pre- and post-transition region.

GdnHCl stock solution (8 M) was prepared fresh prior to each experiment. The final concentration of GdnHCl was determined by refractometry as described by Pace, 1986. GdnHCl was prepared in sodium sulfate buffer (10 mM HEPES at pH 7,5 with 50 mM sodium sulfate) for unfolding monitored by far-UV CD spectropolarimetry. GdnHCl was prepared in storage buffer (10 mM HEPES at pH 7,5 with 100 mM sodium chloride) for monitoring by intrinsic tryptophan fluorescence emission. The pH of denaturant stock solutions was adjusted to 7,5 prior to addition to protein samples. Protein samples were denatured in GdnHCl concentrations ranging from 0 – 6 M, ensuring protein concentration for each sample remained constant after denaturation with GdnHCl. Samples were prepared at 0,25 M increments of GdnHCl over the range mentioned above.

Protein concentrations of 1,5 uM and 10 uM was used for intrinsic tryptophan fluorescence and far-UV CD spectropolarimetry, respectively. Protein was first highly concentrated before denaturation. Samples were prepared by the dilution of this highly concentrated protein-in-buffer with calculated volumes of 8 M GdnHCl stock solution, depending on the final GdnHCl concentration required for the sample. In order to ensure the concentration of protein was constant in each sample, additional volumes of buffer were added. The amount of additional buffer added was dependent on the final desired GdnHCl concentration increment for each sample. Samples were prepared in triplicate at each denaturant concentration increment, and were incubated at room temperature for 30 minutes to reach equilibrium before analysis.

2.2.8 Guanidine Hydrochloride Equilibrium Refolding

GdnHCl stock solution (8 M) was prepared as described above (section 2.2.7). However, in contrast to the preparation of equilibrium unfolding samples, equilibrium refolding samples were prepared by dilution of a concentrated protein-in-GdnHCl solution, as opposed to a protein-in-buffer solution. Protein was first highly concentrated and subsequently denatured to a concentration of 6 M GdnHCl to make up a protein-in-GdnHCl stock solution. The protein-in-GdnHCl stock solution was left to incubate at room temperature for 30 minutes for protein to fully denature and reach equilibrium. Samples were prepared by dilution of this solution with calculated volumes of GdnHCl and buffer in order to arrive at the required sample concentrations of protein and GdnHCl mentioned above (section 2.2.7). Samples were prepared in triplicate and incubated at room temperature for a further 30 minutes to re-establish equilibrium before analysis by far-UV CD and intrinsic tryptophan fluorescence.

2.2.9 Spectroscopic Probes for Folding Studies

Far-UV CD spectropolarimetry and intrinsic tryptophan fluorescence emission were used to probe secondary and local tertiary structural changes, respectively, during unfolding and refolding studies for both wtFOXP2-FHD and Y540F mutant protein.

Secondary structural changes were assessed using a Jasco J-810 Spectropolarimeter (Analytical Solutions, South Africa) (see section 2.2.6.2). In particular, loss / gain of isolated helical content at incremental changes in denaturant concentration was measured by fixed

wavelength measurement for ellipticity values at 222 nm. Measurements were not made at 208 nm as GdnHCl strongly absorbs UV light at wavelengths below 210 nm (Kelly *et al.*, 2005) resulting in poor signal-to-noise ratio. Local tertiary structural changes were monitored using a Jasco FP-6300 fluorescence spectrophotometer (Analytical Solutions, South Africa) (see section 2.2.6.3). Emission spectra were recorded over a range of 280 – 500 nm at an excitation wavelength of 280 nm.

2.2.10 Thermal Unfolding

Unfolding studies included thermal denaturation of wtFOXP2-FHD and the Y540F mutant in order to characterise the enthalpy of unfolding (ΔH_m), transition temperature (T_m) and entropy of unfolding (ΔS_m) for each protein, in the instance that thermal unfolding was observed to be reversible. Unfolding was monitored by CD spectropolarimetry at a wavelength of 222 nm using a Jasco J-810 Spectropolarimeter (Analytical Solutions, South Africa) coupled to a Peltier Thermostatted Cell Holder, using a Hellma® 2 mm quartz absorbance cuvette (Merck, USA). Thermal unfolding was conducted over a temperature range of 20 – 70 °C with data recorded at 1°C increments for both wtFOXP2-FHD and the Y540F mutant.

2.2.11 Data Fitting

Curve fitting of equilibrium unfolding data obtained from far-UV CD and intrinsic tryptophan fluorescence emission measurements was achieved by regression modelling. Regression was conducted in an iterative process of nonlinear curve fitting estimations, constrained by user-defined parameters, until fit convergence was reached (SigmaPlot v14.0). Thermodynamic parameters of folding were calculated computationally from the fitted curve (SigmaPlot v14.0). Both a two-state monomer (N ↔ U) and three state monomer (N ↔ I ↔ U) model was used in data fitting. Prior to fitting, all data was normalised in order to visualise and compare different spectroscopic signals on a single scale according to:

$$Y_{normalised} = \frac{(Y_{obv} - Y_U)}{(Y_N - Y_U)} \quad (6)$$

where Y_{obv} is the signal obtained from either far-UV CD or intrinsic tryptophan fluorescence emission at a specific concentration of GdnHCl, Y_U is the signal of the unfolded protein determined from the linear fit of data in the pre-transition region and extrapolated to zero denaturant, and Y_N is the signal of the native protein determined from the linear fit of data in the post-transition region and extrapolated to zero denaturant.

2.2.11.1 Two state monomeric model (N ↔ U)

The two-state monomeric model assumes that the folding pathway is only populated by protein either in a native (N) or unfolded (U) conformation. Therefore, the sum of the native and unfolded fractions of protein are equal to unity:

$$f_U + f_N = 1 \tag{7}$$

where f_U and f_N are the fractions of unfolded and native protein, respectively. The contribution to signal arises from concentrations of protein in both an unfolded and native conformation, whereby:

$$Y = Y_N f_N + Y_U f_U \tag{8}$$

where Y is the value of the signal obtained, dependent on the respective spectroscopic probe, Y_N and Y_U are the signals of the native and unfolded protein determined from the linear fit of data from the pre- and post-transition regions extrapolated to zero denaturant, respectively, and f_N and f_U are the fractions of native and unfolded protein, respectively. Substituting equations (7) and (8) yields an expression for f_U and f_N as:

$$f_U = \frac{(Y_N - Y)}{(Y_N - Y_U)} \quad \text{and} \quad f_N = \frac{(Y_U - Y)}{(Y_U - Y_N)} \tag{9, 10}$$

The equilibrium constant for the unfolding reaction can be described as:

$$K = \frac{[U]}{[N]} = \frac{f_U}{f_N} \tag{11}$$

and therefore by substitution of equations (9), (10), and (11), and rearrangement:

$$K = \frac{Y - Y_N}{Y_U - Y} \tag{12}$$

Gibbs-Helmholtz free energy change (ΔG) may be expressed as:

$$\Delta G = -RT\ ln(K) \tag{13}$$

where R is the gas constant, equivalent to 8,3145 J.mol^{-1}.K^{-1}, T is the temperature measured in Kelvin, and K is the equilibrium constant calculated in equation (12). If it is assumed that ΔG for each step of the reaction is linearly dependent on the denaturant concentration (Pace, 1986), the free energy change in the absence of denaturant may be calculated as:

$$\Delta G = \Delta G_{H2O} - m[D] \tag{14}$$

where ΔG_{H2O} is the free energy change in the absence of denaturant, m is the cooperativity of the reaction which relates to solvent accessible surface area of the protein during unfolding, and [D] refers to the concentration of denaturant. Therefore, substitution of equation (13) into equation (14) gives:

$$f_U = \frac{Y_N + Y_U \cdot e^{-(\Delta G_{H2O} - m[D])/RT}}{1 + e^{-(\Delta G_{H2O} - m[D])/RT}} \tag{15}$$

f_U was plotted against [GdnHCl]. Data for unfolding fitted to this equation was used to calculate ΔG_{H2O} and m for unfolding, and lastly, $[\text{GdnHCl}]_{1/2}$.

2.2.11.2 Three-state monomeric model (N ↔ I ↔ U)

The three-state monomeric model assumes that the folding pathway is populated by protein in a native (N), unfolded (U), or stable intermediate (I) conformation. Therefore, the sum of the native, accumulated intermediate, and unfolded fractions of protein are equal to unity:

$$f_U + f_I + f_N = 1 \tag{16}$$

where f_U, f_I, and f_N are the fractions of unfolded, intermediate and native protein, respectively. The contribution to signal arises from concentrations of protein in both an unfolded, intermediate, and native conformation:

$$Y = Y_N f_N + Y_I f_I + Y_U f_U \tag{17}$$

where Y is the value of the signal obtained, dependent on the respective spectroscopic probe, Y_N, Y_I, and Y_U are the signals of the native, intermediate, and unfolded protein determined from the linear fit of data from the pre-transition, stable intermediate, and post-transition regions extrapolated to zero denaturant, respectively. Two equilibrium constants describe the full reaction, the first describing N ↔ I transition and the second the I ↔ U transition:

$$K_{N\leftrightarrow I} = \frac{[I]}{[N]} = \frac{f_I}{f_N} \quad \text{and} \quad K_{I\leftrightarrow U} = \frac{[U]}{[I]} = \frac{f_U}{f_I} \tag{18, 19}$$

where $K_{N\leftrightarrow I}$ and $K_{I\leftrightarrow U}$ are the equilibrium rate constants for the native to intermediate transition, and intermediate to unfolded transition, respectively.

The Gibbs-Helmholtz free energy change (ΔG) for each equilibrium constant, according to equation (13), may be expressed as:

$$\Delta G_{N\leftrightarrow I} = -RT\ ln(K_{N\leftrightarrow I}) \quad \text{and} \quad \Delta G_{I\leftrightarrow U} = -RT\ ln(K_{I\leftrightarrow U}) \tag{20, 21}$$

If it is assumed that ΔG for each step of the reaction is linearly dependent on the denaturant concentration (Pace, 1986), the free energy change in the absence of denaturant for each transition may be calculated as:

$$\Delta G_{N\leftrightarrow I} = \Delta G_{H2O\ N\leftrightarrow I} - m_{N\leftrightarrow I}[D] \quad \text{and} \quad \Delta G_{I\leftrightarrow U} = \Delta G_{H2O\ I\leftrightarrow U} - m_{I\leftrightarrow U}[D] \quad (22, 23)$$

where $\Delta G_{N\leftrightarrow I}$ and $\Delta G_{I\leftrightarrow U}$ is the free energy change in the absence of denaturant for the N ↔ I and I ↔ U transitions, respectively. $m_{N\leftrightarrow I}$ and $m_{I\leftrightarrow U}$ is the cooperativity of the reaction for the N ↔ I and I ↔ U transitions, respectively, which relates to solvent accessible surface area of the protein during unfolding. [D] refers to the concentration of denaturant.

Substitution and rearrangement of equations (16) through (23) yields an expression for f_U:

$$f_U = \frac{Y_N + Y_I \cdot e^{\frac{-(\Delta G_{H2O\ N\leftrightarrow I} - m_{N\leftrightarrow I}[D])}{RT}} + Y_U \cdot e^{\frac{-(\Delta G_{H2O\ N\leftrightarrow I} - m_{N\leftrightarrow I}[D])}{RT}} \cdot e^{\frac{-(\Delta G_{H2O\ I\leftrightarrow U} - m_{I\leftrightarrow U}[D])}{RT}}}{1 + e^{\frac{-(\Delta G_{H2O\ N\leftrightarrow I} - m_{N\leftrightarrow I}[D])}{RT}} + e^{\frac{-(\Delta G_{H2O\ N\leftrightarrow I} - m_{N\leftrightarrow I}[D])}{RT}} \cdot e^{\frac{-(\Delta G_{H2O\ I\leftrightarrow U} - m_{I\leftrightarrow U}[D])}{RT}}} \quad (24)$$

f_U was plotted against [GdnHCl]. Data for unfolding fitted to this equation was used to calculate $\Delta G_{H2O\ N\leftrightarrow I}$, $\Delta G_{H2O\ I\leftrightarrow U}$, $m_{N\leftrightarrow I}$, and $m_{I\leftrightarrow U}$ for unfolding, followed by $[\text{GdnHCl}]_{1/2}$ for the $N \leftrightarrow I$ and $I \leftrightarrow U$ transitions.

2.2.11.3 Single Data Set Curve Fitting and Global Curve Fitting

Both single data set curve fitting and global curve fitting were conducted for the data sets obtained from equilibrium unfolding studies monitored by both far-UV CD and intrinsic tryptophan fluorescence. A parameterised two-state (N ↔ U) and three-state (N ↔ I ↔ U) folding model was applied to each data set in order to obtain a curve of best fit. Parameters for the two-state model included Y_U and Y_N where each value refers to the signal of the unfolded protein determined from the linear fit of data in the pre- and post-transition region, and extrapolated to zero, respectively. The three-state model had an additional parameter – Y_I – which referred to the linear fit of the data in the region of the stable intermediate, extrapolated to zero. The model fit was accepted if the R^2-value for the fit was >0,95. This indicates that the variance of the regression model accounts for at least 95 % of the variance of the observed values, meaning that the fitted curve significantly modelled a two- or three-state folding mechanism (depending on the fit).

Single data set curve fitting was conducted to assess if the equilibrium unfolding data could be accurately predicted by a two- or three-state folding mechanism. Global curve fitting was conducted to calculate the thermodynamic parameters for each equilibrium unfolding study (incorporating both unfolding and refolding data for fitting) monitored by either far-UV CD or intrinsic tryptophan fluorescence. Global curve fitting allows for the simultaneous analysis of multiple data sets (for example, data for all unfolding and refolding studies obtained from both spectroscopic techniques) in order to derive global model parameters that are consistent with all data sets under the same experimental conditions (such as ΔG_{H2O}, m and $[\text{GdnHCl}]_{1/2}$ over the same range of GdnHCl concentrations) (Herman and Lee, 2012).

Chapter 3: Results

FOXP2-FHD has been observed to not readily form domain swapped dimers in solution, despite this being a prerequisite for its involvement in transcriptional regulation. Successful domain swapped dimerisation depends on the folding pathway undertaken by participating unfolded monomers and the potential interchain interactions that occur between them along such a pathway. Therefore, in order to elucidate the reduced dimerisation propensity of FOXP2-FHD in solution, the folding mechanism and stability of wtFOXP2-FHD, alongside a dimer-prone mutant (Y540F), was investigated. DNA binding studies were conducted for both wtFOXP2-FHD and the Y540F mutant to ensure DNA recognition and binding function were retained post-purification, and to investigate the effect of DNA presence on the dimerisation propensity for both proteins. Quaternary structural analysis (by SEC) was conducted to examine the K_d of dissociation of dimer into monomer for both proteins in solution, before and after denaturation. The secondary and tertiary structures of wtFOXP2-FHD and the Y540F mutant were characterised (by far-UV CD and intrinsic tryptophan fluorescence emission, respectively) to obtain spectral signatures for each protein, and to observe if there existed any significant structural differences between the wild-type and mutant protein as a result of the Y540F mutation. Lastly, equilibrium unfolding studies in the presence of GdnHCl, and thermal unfolding studies, were conducted. This was to establish the folding pathway and stability for each protein, and to calculate their corresponding thermodynamic parameters of folding, which describe their respective free energies of conformation, folding cooperativity, enthalpies of unfolding, and denaturant / temperature unfolding midpoints. Overview and comparison of DNA binding studies, structural characterisation, and equilibrium unfolding studies for both wtFOXP2-FHD and the dimer prone Y540F mutant, will shed light on the folding mechanism and stability of FOXP2-FHD, and hence, allow for a better understanding of how its structure contributes to its ability to domain swap dimerise.

3.1 Plasmid Sequence Verification

Sequencing of the recombinant pET-11a plasmid coding for wtFOXP2-FHD showed that the codon optimised hexahistidine-tagged wtFOXP2-FHD gene insert was present (Figure 11). In addition, sequencing results obtained for plasmid DNA coding for the Y540F mutant confirmed the incorporation of a single mutation at the desired location with no additional

mutations incorporated during the thermocycling reaction (Figure 11A). This resulted in a substitution of phenylalanine for tyrosine at position 540 in the FOXP2-FHD (Figures 11A and 11B). Both the sequenced wtFOXP2-FHD and Y540F mutant plasmids were extracted from their respective transformed T7 Competent *E. coli* cell lines, indicating that transformation was successful for both expression systems.

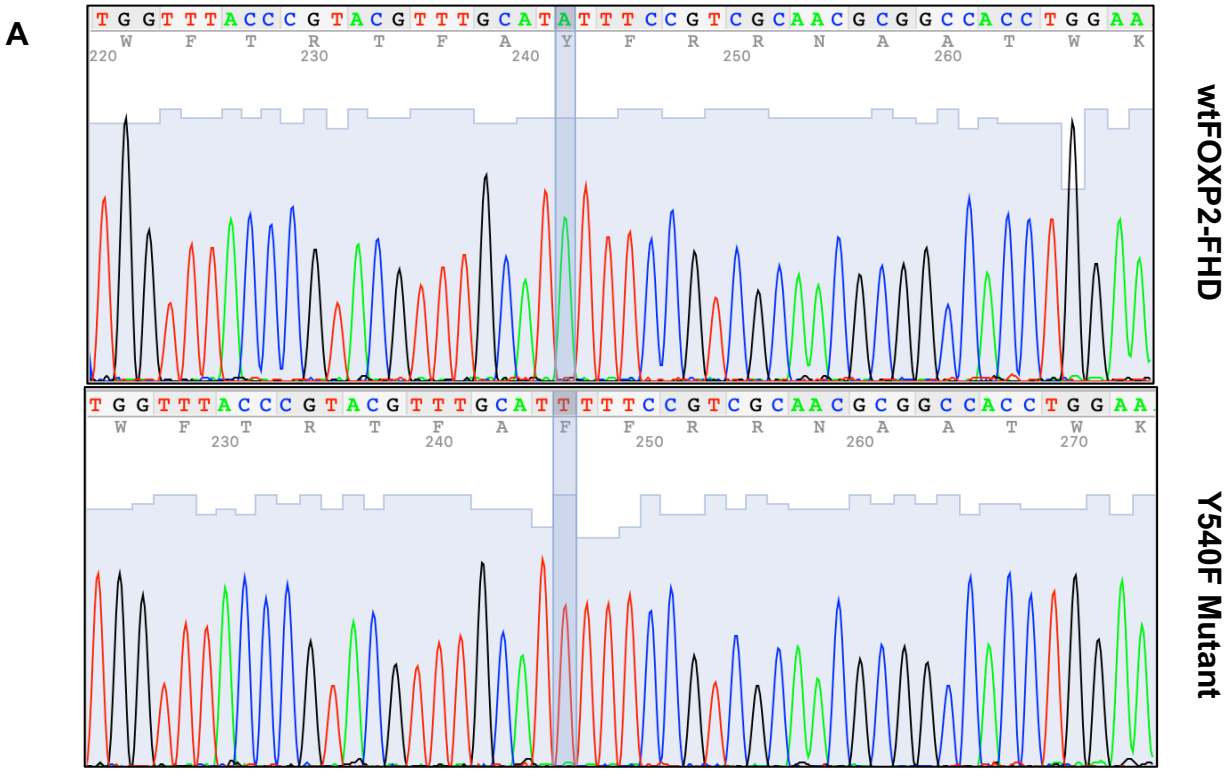


Figure 11: Chromatograms of sequence identity for wtFOXP2-FHD and Y540F mutant, and amino acid sequence pairwise alignment

(A) Thymine was successfully substituted for adenine (highlighted bar) following site-directed mutagenesis, resulting in a Tyr → Phe amino acid substitution. **(B)** Pairwise alignment of the amino acids sequences of wtFOXP2-FHD and the Y540F mutant, indicating the location of the Tyr → Phe amino acid substitution with respect to the full sequence of each protein.

3.2 Protein Purification and Molecular Weight Determination

Following overexpression and cell lysis (see section 2.2.2), wtFOXP2-FHD and Y540F mutant protein was purified from soluble cell lysate using Ni^{2+}-charged IMAC-FPLC (see section 2.2.2). Lysate was loaded onto the column and fractions collected during elution with elution buffer (20 mM Tris-HCl, 500 mM sodium chloride, 500 mM imidazole at pH 7,5) when absorbance was measured to exceed the baseline value (Figure 12A and 12B). Both wtFOXP2-FHD and Y540F mutant protein eluted from the IMAC column in two peaks: one corresponding to column-bound protein contamination, and the other to overexpressed protein (Figure 12A and 12B).

Collected eluent was analysed by discontinuous tricine SDS-PAGE to determine protein purity and molecular weight for both wtFOXP2-FHD and Y540F mutant. The samples resolved by electrophoresis included: supernatant from centrifuged lysate, resuspended insoluble pellet (shown for wild-type), elution flow through, and fractions eluted with elution buffer (Figures 12C and 12D). Purified wtFOXP2-FHD was resolved alongside the Y540F elution fraction to qualitatively compare the sizes and gel migration distances of both proteins (Figure 12D). A protein molecular weight ladder was resolved alongside samples to determine molecular weights of protein in samples. Both wtFOXP2-FHD and Y540F mutant proteins were observed to fully bind the Ni^{2+}-charged IMAC column evidenced by absence of overexpressed protein in their relative flow through eluents (Figures 12C and 12D). Purification of wtFOXP2-FHD showed a low volume of column-bound contaminant, and eluted with a purity of 97,9 % (Figure 12C). Y540F mutant showed the presence of considerable protein contaminants in the elution fraction, however eluted without significant contamination of the protein product, with a purity of 95,5 % (Figure 12D). The bands relating to wtFOXP2-FHD and Y540F mutant were similar in size and gel migration distance (Figure 12D).

Molecular weights of wtFOXP2-FHD and the Y540F mutant were determined by extrapolation by standard curve (Figures 12E and 12F). wtFOXP2-FHD was calculated to have a molecular weight of 15,2 kDa (Figure 12E) and the Y540F mutant protein was calculated to have a molecular weight of 13,6 kDa. The sample of wtFOXP2-FHD protein run alongside the Y540F mutant was calculated to have a molecular weight of 14,5 kDa (Figure 12F). The theoretical molecular weights of wtFOXP2-FHD and Y540F mutant are 12,2 kDa and 12,1 kDa, respectively. Although there is a discrepancy between the calculated molecular weights and

theoretical molecular weights of both proteins, it is not uncommon for this to be observed for molecular weight determination by SDS-PAGE, as it is a low-resolution molecular weight determination technique (Matsumoto *et al.*, 2019). Regardless of the discrepancy in determined molecular weight, it can be confidently assumed that each resolved protein corresponds to the desired, overexpressed recombinant protein because: (1) each expressed protein contained a hexahistidine Ni^{2+}-affinity tag, (2) was purified by Ni^{2+}-affinity chromatography, (3) eluted with >95 % purity, and (4) is observed to exist predominantly after column washing with elution buffer, and not during previous washing steps (with equilibration / salt buffer) (Figures 12C and 12D).

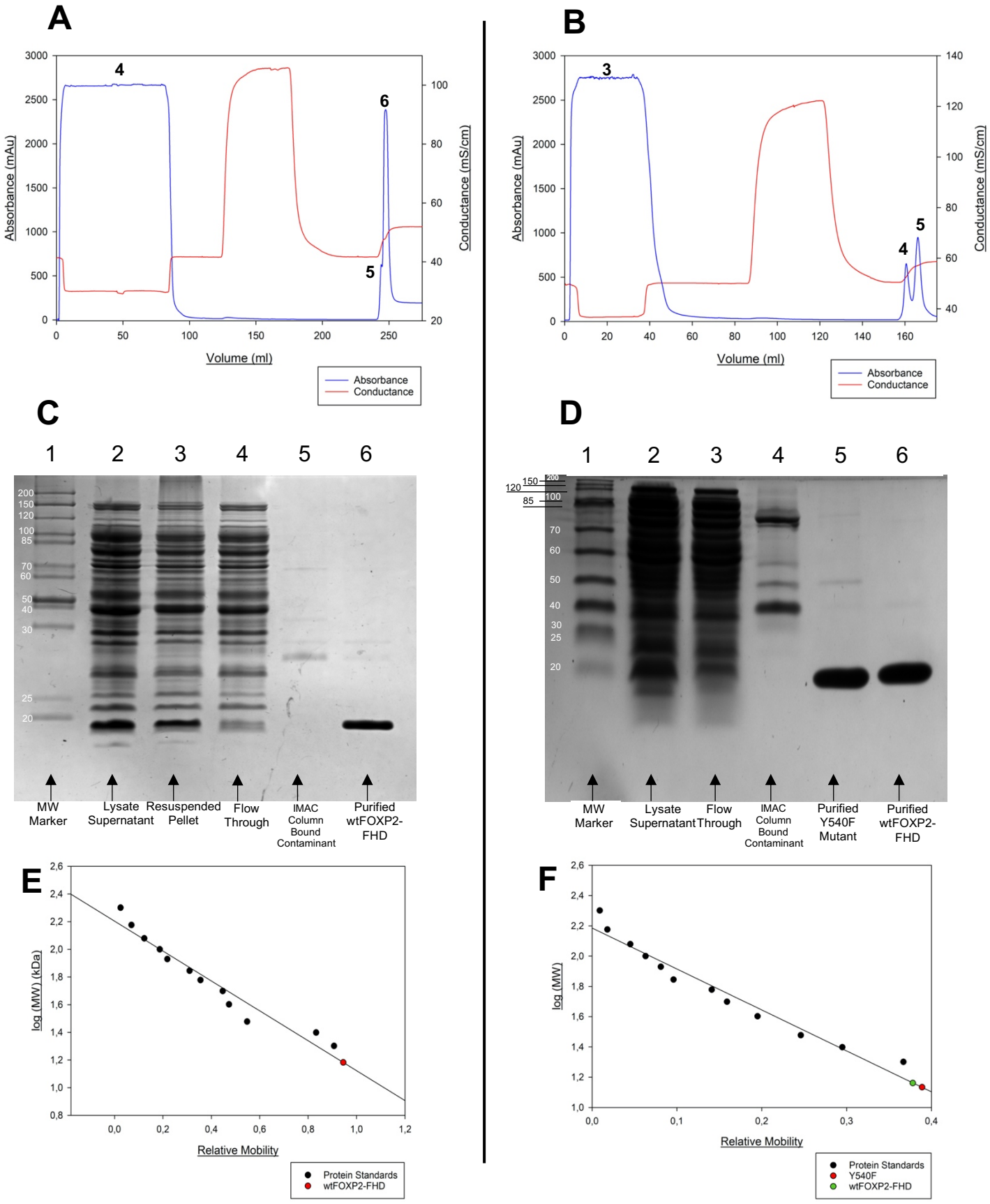


Figure 12: Purification of wtFOXP2-FHD and the Y540F mutant

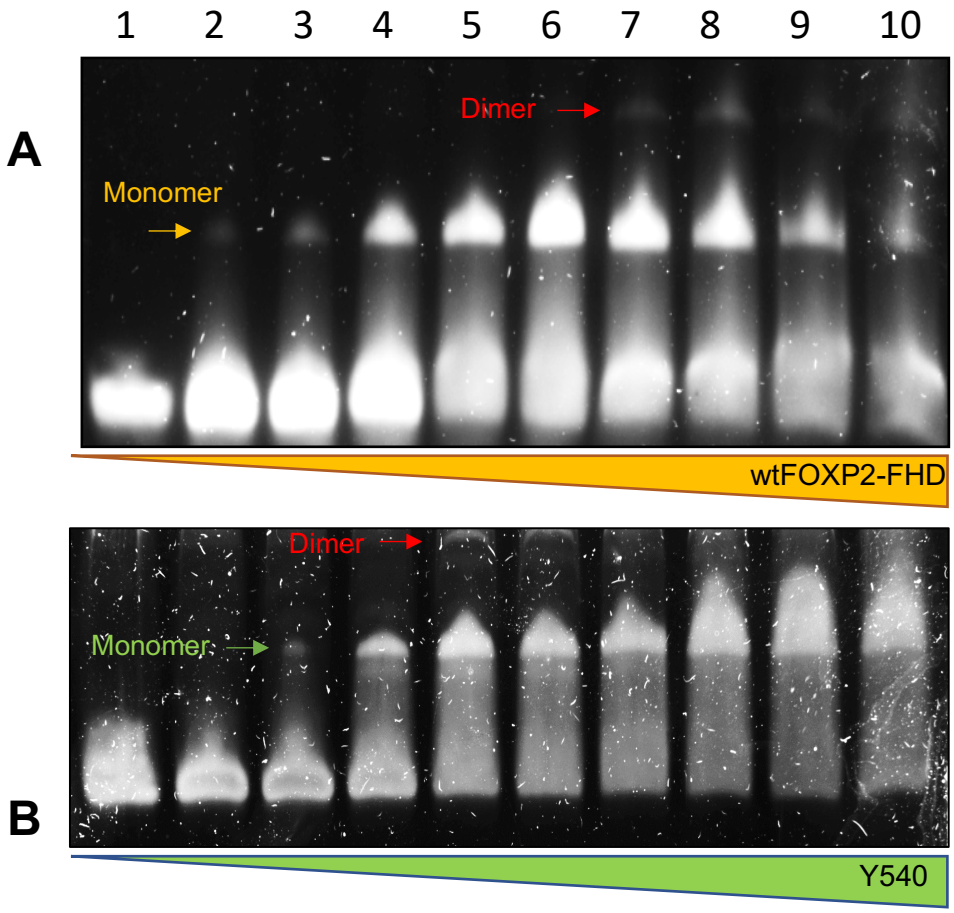
(A) Elution profile of wtFOXP2-FHD purification. Supernatant was injected onto the IMAC column and flow through eluted over 80 mL at 5 mL/min (4), followed by salt wash, and

elution with elution buffer (5, 6). Two fractions were collected during elution with elution buffer: one corresponding to column bound contaminant (5) and the other to purified wtFOXP2-FHD protein (6). The number indicated above eluted peak (4, 5, 6) relate to gel lanes in (C). **(B)** Elution profile of the Y540F mutant. Supernatant was injected onto the column and eluted over 40 mL at 5 mL/min (3), followed by salt wash, and elution with elution buffer (4, 5). Two fractions were collected during elution with elution buffer: one corresponding to column bound contaminant (4) and the other to Y540F mutant protein (5). The number indicated above eluted peak (3, 4, 5) relates to gel lanes in (D). **(C)** Tricine SDS-PAGE gel of wtFOXP2-FHD purification. Lanes 1 – 6 correspond to: molecular weight marker (1), lysate supernatant (2), resuspended cell pellet (3), eluted flow through (4), column bound contaminant (5), and purified wtFOXP2-FHD protein (6). **(D)** Tricine SDS-PAGE gel of Y540F mutant purification. Lanes 1 – 6 correspond to: molecular weight marker (1), lysate supernatant (2), eluted flow through (3), column bound contaminant (4), purified Y540F mutant protein, and previously purified wtFOXP2-FHD protein for comparison (6). **(E)** Standard curve of wtFOXP2-FHD purification. A line of best fit was fitted by least means squares regression to the logarithm of protein standard molecular weights against relative mobility of the standards ($R^2 = 0,96$). wtFOXP2-FHD was extrapolated to have a molecular weight of 15,2 kDa (red). **(F)** Standard curve of Y540F mutant purification. A line of best fit was fitted by least means squares regression to the logarithm of protein standard molecular weights against relative mobility of the standards ($R^2 = 0,95$). Y540F mutant was extrapolated to have a molecular weight of 13,6 kDa (red) in comparison with wtFOXP2-FHD which had a differing molecular weight of 14,5 kDa (green).

3.3 DNA Binding Studies

EMSAs were conducted for both wtFOXP2-FHD and the Y540F mutant to ensure DNA recognition and binding function were retained post-purification. As aforementioned in section 1.5, the presence of DNA has been observed to promote dimerisation of FOXP2-FHD and a F541C variant (Morris and Fanucchi, 2016; Häußermann *et al.*, 2019). Therefore, EMSAs were additionally conducted to investigate the effect of DNA presence on the dimerisation propensity for wtFOXP2-FHD and the Y540F mutant. wtFOXP2-FHD was observed to form a protein-DNA complex at a ratio of DNA : protein (D : P) = 3 :1 (Figure 13A). A possible dimeric form of wtFOXP2-FHD was observed to bind DNA at a D : P ratio of 1 : 5, although it was hardly detectable (Figure 13A). Y540F mutant protein was observed to form a

protein-DNA complex at a D : P ratio of 2 : 1, also with a possible dimeric state observed at a ratio of 1 : 2 (Figure 13B).



C

Lane	1	2	3	4	5	6	7	8	9	10
Protein (uM)	0	0,17	0,25	0,5	1,0	1,5	2,5	4,0	5,0	10,0
DNA (uM)	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5
Ratio DNA : Protein (D : P)	1 : 0	3 : 1	2 : 1	1 : 1	1: 2	1 : 3	1 : 5	1 : 8	1 : 10	1 : 20

Figure 13: EMSA of wtFOXP2-FHD and Y540F protein with Nelson DNA oligonucleotide

(A) wtFOXP2-FHD formed a protein-DNA complex at a DNA : protein ratio of 3 : 1 (orange arrow) and assumed a possible dimeric conformation at a ratio of 1 : 5 (red arrow). **(B)** Y540F mutant formed a protein-DNA complex at a DNA : protein ratio of 2 : 1 (green arrow) and assumed a possible dimeric conformation at a ratio of 1 : 2 (red arrow). **(C)** Table relating lane number to protein and DNA concentrations, and their ratios.

3.4 Structural Characterisation

3.4.1 Quaternary Structural Characterisation

Analytical SEC was used to determine the molecular weights for both wtFOXP2-FHD and the Y540F mutant, as well as investigate the proportion of monomeric and dimeric species at equilibrium for both proteins. Analytical SEC was initially performed on native, freshly purified protein, whereby protein samples were loaded at three increasing concentrations to investigate the effect of concentration on dimer formation propensity. wtFOXP2-FHD eluted as two peaks which correspond to the dimeric and monomeric species of the protein (Figure 14A). For each sample concentration for wtFOXP2-FHD (138,2 uM, 73,5 uM and 18,2 uM), the monomeric protein species eluted at the 84,3 mL, and the dimeric species eluted at 73,3 mL (Figure 14A). Furthermore, the Y540F mutant eluted as two peaks also corresponding to the dimeric and monomeric species of the protein. At each sample concentration (117,8 uM, 65,1 uM and 17,7 uM), the monomeric species eluted at the 83,1 mL, and the dimeric species eluted at 71,8 mL (Figure 14A). The molecular weights of elution fractions were determined by interpolation by standard curve (Figure 14B). wtFOXP2-FHD monomeric and dimeric species were calculated to have a molecular weight of 12,3 kDa and 24,9 kDa, respectively (Figure 14B). The molecular weights of the monomeric and dimeric species of the Y540F mutant were calculated to be 13,4 kDa and 27,1 kDa, respectively (Figure 14B). These molecular weights aligned more consistently with the theoretically calculated molecular weights for wtFOXP2-FHD and the Y540F mutant in comparison to those determined by SDS-PAGE (section 3.2).

The concentrations of wtFOXP2-FHD and Y540F mutant protein affects each protein's dimerisation propensity. For wtFOXP2-FHD, as concentration increased from 18,2 – 138,2 uM, the percentage of dimer eluted decreased from 8,3 % to 6,6 % (Table 1). For the Y540F mutant, percentage of dimer presence was initially observed to slightly decrease from 2,5 – 2,3 % as the concentration increased from 17,7 – 65,1 uM, and then increased to 4,5 %, as concentration increased from 65,1 – 117,8 uM (Table 1). The K_d value for dimer dissociation into monomer was calculated as the gradient of the graph [monomer]^2 against [dimer] (Figure 15). The K_d for wtFOXP2-FHD and the Y540F mutant was calculated to be 1,6 mM and 2,5 mM, respectively, according to the equation (3).

Table 1: Concentrations and Percentages of wtFOXP2-FHD and Y540F Mutant Species Eluted from SEC Column at Increasing Total Concentration

Protein	**Total Concentration (uM)**	**Dimer Eluted (%)**	**Monomer Eluted (%)**	**Dimer (uM)**	**Monomer (uM)**
wtFOXP2-FHD	18,2	8,3	91,7	1,2	17,0
	73,5	7,7	92,3	5,6	67,8
	138,2	6,6	93,4	11,5	126,7
Y540F	17,7	2,5	97,5	0,5	17,2
	65,1	2,3	97,7	1,5	63,6
	117,8	4,5	95,5	5,3	112,5

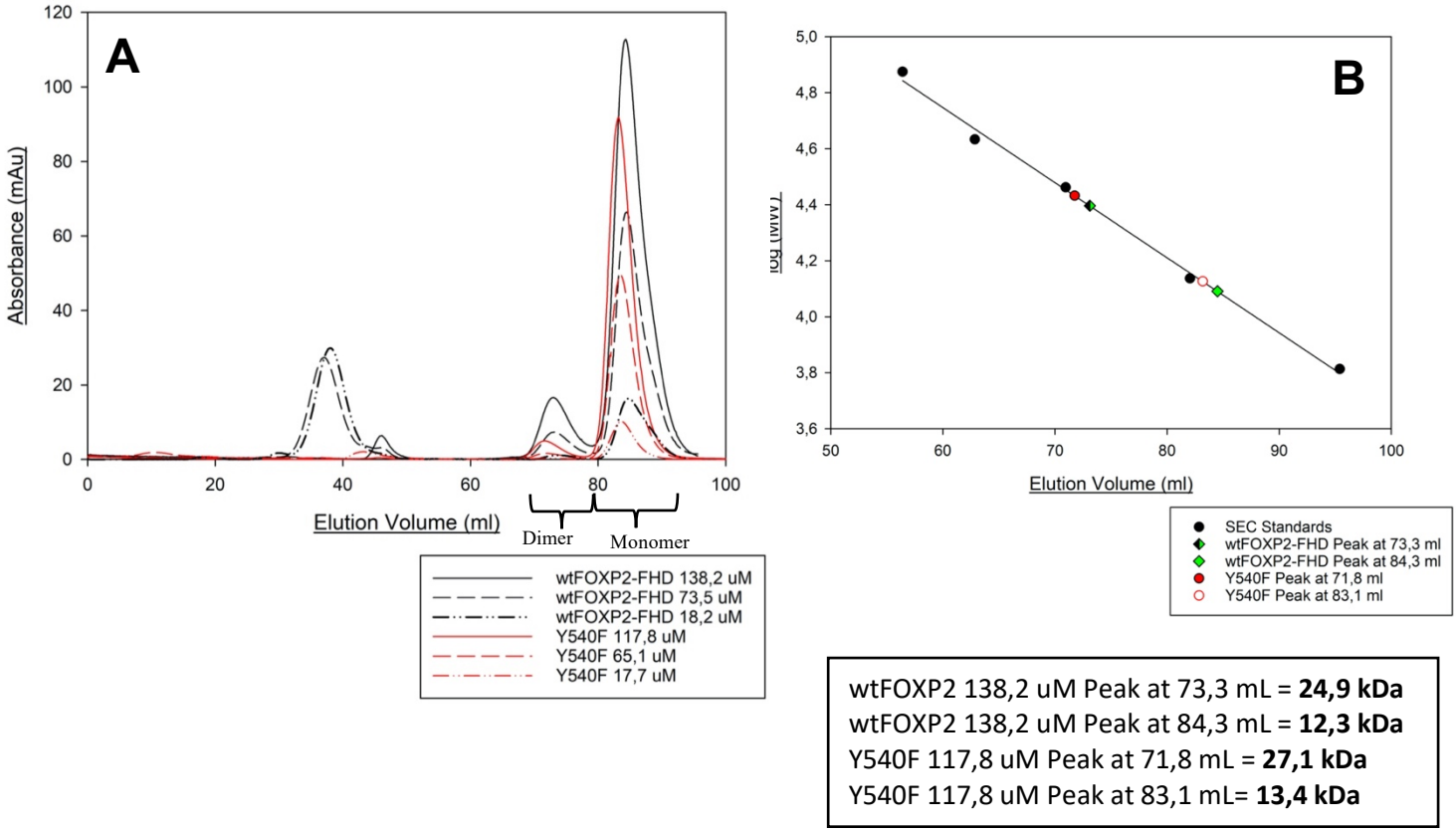


Figure 14: Analytical SEC for increasing concentrations of wtFOXP2-FHD and Y540F mutant protein

(A) wtFOXP2-FHD protein was loaded at concentrations 138,2 uM (solid black line), 73,5 uM (dashed black line) and 18,2 uM (dash-dotted black line), and the Y540F mutant at concentrations 117,8 uM (solid red line), 65,1 uM (dashed red line) and 17,7 uM (dash-dotted red line). Fractions eluted for wtFOXP2-FHD at 73,3 mL and 84,3 mL, and for Y540F mutant at 71,8 mL and 83,1 mL, corresponding respectively to the elution of dimeric and monomeric species of each protein. **(B)** Standard curve of SEC protein standards against wtFOXP2-FHD and Y540F elution fractions. A line of best fit was fitted by least mean squares regression to the logarithm of protein standard molecular weights against relative mobility of the standards (R^2 = 0,996). The wtFOXP2-FHD fractions eluted at 73,3 mL and 84,3 mL were interpolated to have molecular weights of 24,9 kDa and 12,3 kDa, respectively, and the Y540F mutant fractions eluted at 71,8 mL and 83,1 mL were interpolated to have molecular weights of 27,1 kDa and 13,4 kDa, respectively

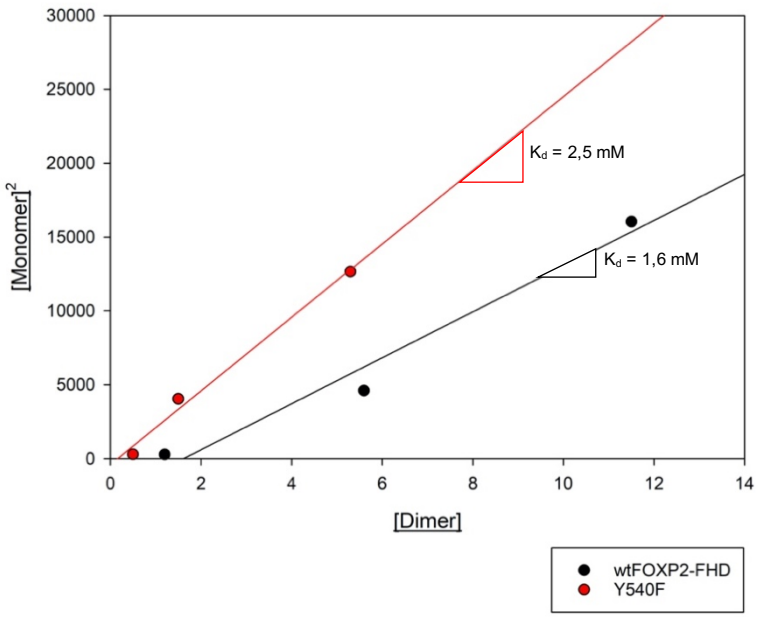


Figure 15: K_d Determination of dimer dissociation into monomer for wtFOXP2-FHD and the Y540F mutant

The gradient of the square of eluted monomer concentration against dimer concentration indicated K_d for both wtFOXP2-FHD and Y540F mutant protein. A line of best fit was fitted by least mean squares regression for each protein ($R^2_{wtFOXP2\text{-}FHD} = 0,97$; $R^2_{Y540F} = 0,99$). The K_d for wtFOXP2-FHD and Y540F mutant was calculated as 1,6 mM and 2,5 mM, respectively.

Analytical SEC was subsequently conducted to investigate changes in dimer formation propensity after refolding in GdnHCl. Both proteins were denatured in 6 M GdnHCl, allowed to equilibrate, and refold to a GdnHCl concentration of 1,38 M, and eluted in GdnHCl buffer of the same concentration. wtFOXP2-FHD and Y540F mutant were loaded at concentrations of 61,5 uM and 65,3 uM, respectively, determined by UV-Vis spectroscopy. Both proteins eluted at lower elution volumes in comparison to the elution profile of native, freshly purified protein before denaturation. wtFOXP2-FHD eluted in two fraction at 64,49 mL and 76,12 mL, and Y540F mutant eluted at 65,25 mL and 76,5 mL (Figure 16). Each fraction was assumed to correspond with dimeric and monomeric species, eluting at lower volumes as a result of the protein still exhibiting partial denaturation, and hence exhibiting a larger hydrodynamic volume. The difference in elution volume correlating to the difference between dimer and monomer peaks of refolded wtFOXP2-FHD and Y540F mutant in GdnHCl was 11,6 mL and 11,3 mL, respectively. This difference very closely compares with the difference in elution

volume between dimer and monomer peaks of native protein before denaturation (wtFOXP2-FHD elution volume difference = 11,0 mL and Y540F elution volume difference = 11,3 mL).

Under these conditions, wtFOXP2-FHD eluted predominately as a dimeric species (55,1 % of total eluted protein) and the Y540F mutant eluted mainly as a monomeric species (64,8 % of total eluted protein) (Figure 16). These percentages differ considerably from the percentage of eluted dimer and monomer for freshly purified wtFOXP2-FHD and Y540F mutant. After refolding, there was an increase of 47,4 and 32,9 percentage points for eluted dimer for wtFOXP2-FHD and the Y540F mutant, respectively. The K_d for each protein could not be determined by linear regression as only one data point was available and so it is rather expressed as a simple ratio according to equation (3). The K_d for refolded wtFOXP2-FHD and Y540F mutant is 22,5 uM and 77,9 uM, respectively. This relates to an ~69-fold and ~31-fold increase in dimer presence of wtFOXP2-FHD and Y540F mutant protein, respectively, following denaturation in GdnHCl and equilibrium refolding.

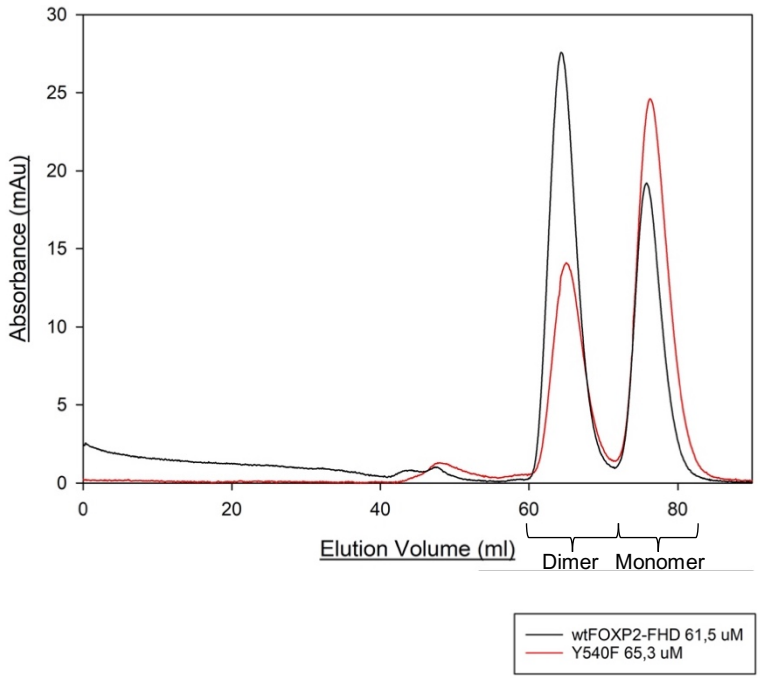


Figure 16: Elution profile of wtFOXP2-FHD and Y540F mutant after equilibrium refolding in GdnHCl

wtFOXP2-FHD and Y540F protein were loaded at concentrations of 61,5 uM and 65,3 uM, respectively. wtFOXP2-FHD eluted first at a lower elution volume in comparison to the Y540F mutant. wtFOXP2-FHD was observed to exist predominantly as a dimeric species (55,1 % total protein) as opposed to a monomeric species (44,9 % total protein) (black solid line). In the case of the Y540F mutant, both a dimeric species (35,2 % total protein) alongside a mainly monomeric species (64,8 % total protein) was observed (red solid line).

3.4.2 Secondary Structural Characterisation

Far-UV CD spectropolarimetry was used to determine the secondary structural profile of both wtFOXP2-FHD and the Y540F mutant. Data was recorded between wavelengths 250 – 206 nm (Figure 17A), while high tension voltage remained below 700 V (Figure 17B). wtFOXP2-FHD and Y540F mutant showed minima at 222 nm and 208 nm, indicating predominantly α-helical structure (Figure 17A).

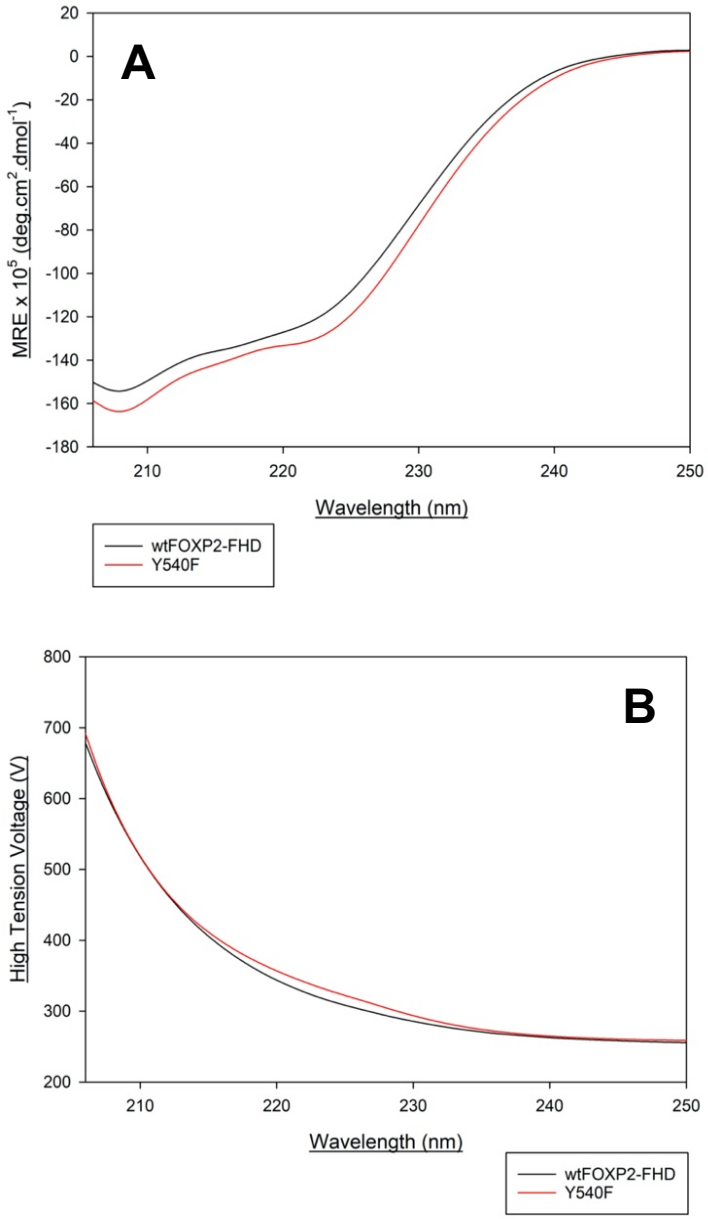


Figure 17: Far-UV CD spectra for wtFOXP2-FHD and Y540F mutant.

(A) The CD spectra for wtFOXP2-FHD (solid black line) and the Y540F mutant (solid red line) showed minima at 222 nm and 208 nm, indicating predominantly α-helical structure. **(B)** High tension voltage of measurements made in (A). Data was collected for wtFOXP2-FHD (solid black line) and Y540F (solid red line) while high tension voltage remained below 700 V.

3.4.3 Tertiary Structural Characterisation

Intrinsic tryptophan fluorescence emission was used to obtain emission spectra for both wtFOXP2-FHD and the Y540F mutant in order to characterise their tertiary structures. This was accomplished through monitoring the shift in fluorescence signal emitted by tryptophan residues in each protein, as a result of their exposure to the local solvent environment, and hence, positional relationship to tertiary structure. Both wtFOXP2-FHD and Y540F mutant have three tryptophan residues: one buried in the hydrophobic core (Trp548) and two solvent exposed (Trp533 and Trp573) (Figure 18).

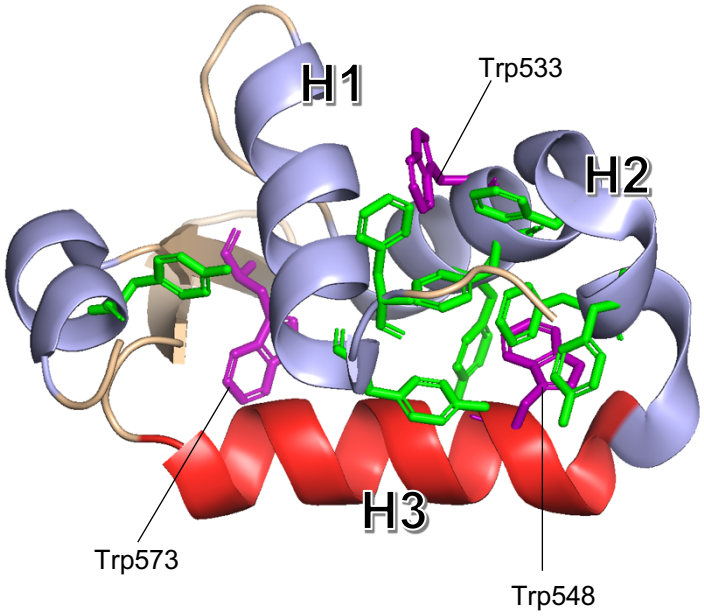


Figure 18: Location of tryptophan residues in wtFOXP2-FHD.

wtFOXP2-FHD has three tryptophan residues (purple): one buried in the hydrophobic core (Trp548) and two solvent exposed (Trp533 and Trp573). The hydrophobic core is comprised of Trp548 located on H3, and aromatic residues located on H1 and H2 of the protein (green).

The fluorescence spectra were obtained at excitation wavelengths of 280 nm and 295 nm. At 280 nm, tryptophan and tyrosine residues contribute to fluorescence signal, while at 295 nm, tryptophan alone contributes to the signal. An excitation wavelength of 280 nm was used to account for differences in fluorescence signal between wtFOXP2-FHD and the Y540F mutant which may arise from the Tyr → Phe substitution. An excitation wavelength of 295 nm was used to solely monitor tryptophan residues, and hence observe if the tryptophan residues of the

Y540F mutant occupy different degrees of solvent exposure resulting from the Tyr → Phe substitution in comparison with wtFOXP2-FHD. Emission maxima for both proteins were observed at 330 nm and 340 nm. These maxima correspond to fluorescence emission from the hydrophobically buried Trp548 residue, and solvent exposed Trp533 and Trp573 residues, respectively (Figure 19). This was observed for both excitation wavelengths (Figure 19). The fluorescence intensity of the Y540F mutant spectrum excited at 280 nm is greater than that of the wtFOXP2-FHD spectrum as a result of technical error, whereby a higher concentration (~0,1 uM) of protein was excited for Y540F (Figure 19). The effect of tyrosine to phenylalanine substitution in the Y540F mutant was undetectable in the emission spectra monitored at an excitation wavelength of 295 nm. This indicates that the tryptophan residues for both wtFOXP2-FHD and the Y540F are at the same degree of solvent exposure for each protein despite the Tyr → Phe substitution in the mutant.

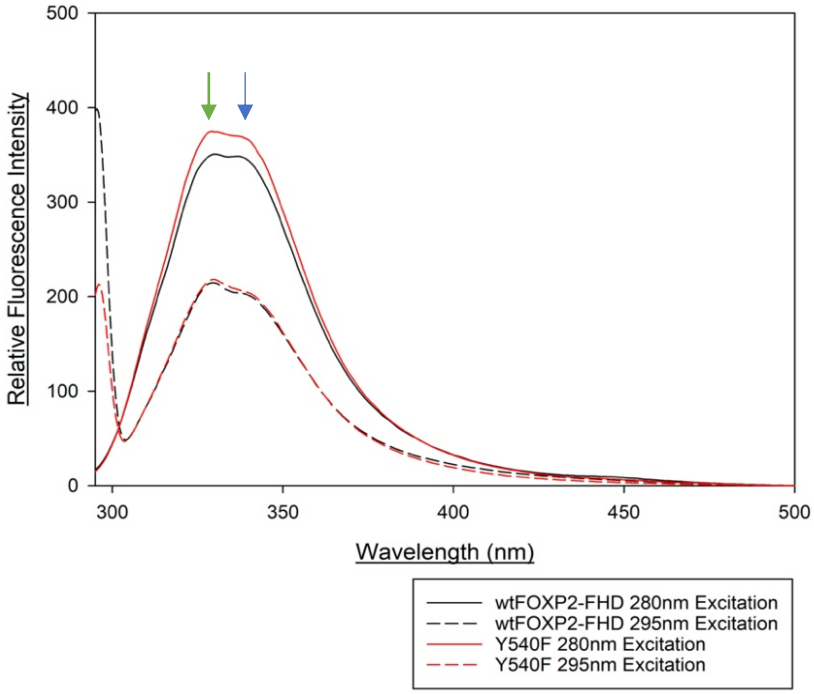


Figure 19: Fluorescence emission spectra for wtFOXP2-FHD and theY540F mutant at excitation wavelengths 280 nm and 295 nm.

Emission maxima at 330 nm (green arrow) and 340 nm (blue arrow) are present in spectra for both wtFOXP2-FHD and Y540F mutant at excitation wavelengths of 280 nm and 295 nm. The Y540F mutant excited at 280 nm (solid red line) emitted at a higher intensity in comparison with wtFOXP2-FHD excited at the same wavelength (solid black line). Both proteins showed overlapping emission spectra when excited at 295 nm (dashed black line – wtFOXP2-FHD, dashed red line – Y540F mutant). This indicates that the tryptophan residues for both wtFOXP2-FHD and the Y540F are at the same degree of solvent exposure for each protein.

3.5 Equilibrium (Un)Folding Studies

Equilibrium unfolding studies were performed to characterise the folding mechanism, cooperativity (m), conformational stability (ΔG_{H2O}) and denaturation midpoint ($[GdnHCl]_{1/2}$) of wtFOXP2-FHD and Y540F mutant protein. The folding mechanism of a protein describes the pathway the protein undertakes in order to adopt its native state from its initial unfolded polypeptide chain. The pathway is populated with a discrete number of folded protein states at

equilibrium (Walters *et al.*, 2009). Equilibrium unfolding studies aim to denature a protein at discrete points along its folding pathway. The fraction of unfolded protein is determined by spectroscopic probe, such as far-UV CD spectropolarimetry or intrinsic tryptophan fluorescence emission. Through sequentially plotting the fraction of unfolded protein over the range of denaturant conditions, a folding mechanism for a particular protein may be described (Pace, 1986). Studies were conducted using GdnHCl as a chemical denaturant, and were monitored spectroscopically by far-UV CD spectropolarimetry for ellipticity values at 222 nm, and by intrinsic tryptophan fluorescence emission. For intrinsic tryptophan fluorescence, an excitation wavelength of 280 nm was opted for in order to monitor possible changes in fluorescence signal resulting from the Tyr → Phe substitution in the Y540F mutant as tyrosine and phenylalanine contribute differently to fluorescence emission signal. Monitoring at 280 nm hence allowed for the comparison of intrinsic tryptophan fluorescence emission data obtained for wtFOXP2-FHD and the Y540F mutant. Tertiary structural changes were monitored at emission wavelengths 330 nm and 340 nm with an excitation wavelength of 280 nm. Emission wavelengths of 330 nm and 340 nm correspond, respectively, to intrinsic tryptophan fluorescence emission of hydrophobically buried Trp548, and solvent exposed Trp533 and Trp573, respectively. The monitoring of the change in tertiary structure with respect to denaturant concentration at both emission wavelengths was useful in order to simultaneously account for tertiary structural changes affected at the protein surface and hydrophobic core.

Curve fitting and global curve fitting were conducted for the data sets obtained from equilibrium unfolding studies monitored by both far-UV CD and intrinsic tryptophan fluorescence. A parameterised two-state U ↔ N unfolding and three-state U ↔ I ↔ N model was applied to each data set in order to obtain a curve of best fit. The model fit was accepted if the R^2-value for the fit was greater than 0,95.

3.5.1 Recoverability and Reversibility of wtFOXP2-FHD and Y540F Mutant

In order for thermodynamic parameters to be derived from equilibrium unfolding studies, reversibility of the folding pathway of a protein must be established (Pace, 1986), and recoverability of native protein following denaturation must be confirmed. wtFOXP2-FHD and Y540F mutant protein were denatured in GdnHCl and left to reach equilibrium. Changes in secondary and tertiary structure recoverability was assessed by far-UV CD spectropolarimetry

at a wavelength of 222 nm, and intrinsic tryptophan fluorescence at excitation wavelength of 280 nm. Spectra were collected for native protein, denatured protein in 6 M GdnHCl, and refolded protein at a GdnHCl concentration of 0,5 M (Figure 20). wtFOXP2-FHD exhibited a percentage recovery of native signal of ~89 % when calculated using far-UV CD (Figure 20A), and ~100 % using intrinsic tryptophan fluorescence signals (Figure 20B), respectively. The percentage recovery of native signals from the Y540F mutant were similar, exhibiting ~82 % recovery when calculated using far-UV CD (Figure 20C), and ~96 % recovery when calculated using intrinsic tryptophan fluorescence emission (Figure 20D). The reversibility of denaturation is evidenced by the overlap of unfolding and refolding curves (Figures 20B and 20D).

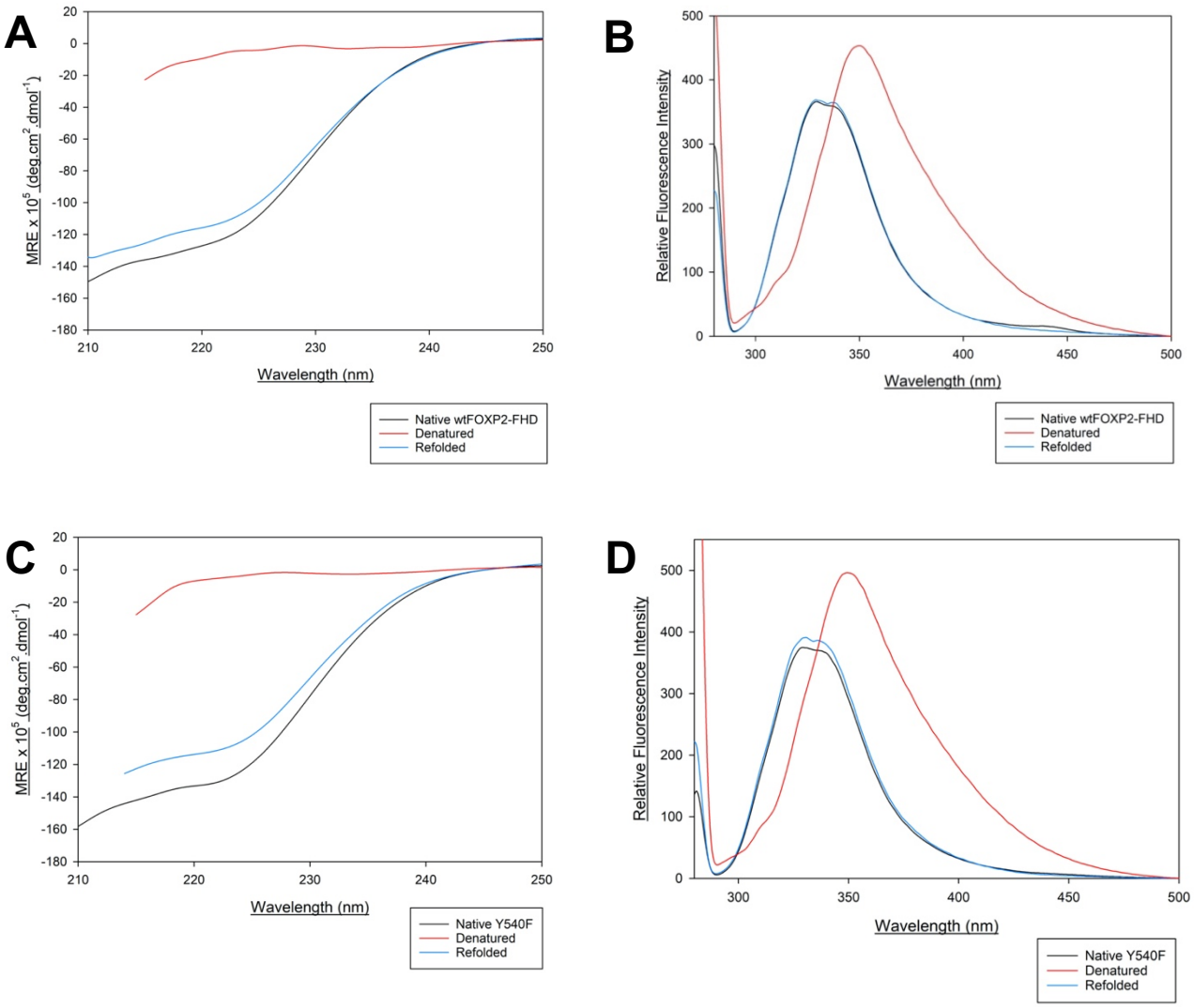


Figure 20: Recoverability of wtFOXP2-FHD and Y540F mutant protein after denaturation in GdnHCl.

Percentage recovery was calculated by monitoring secondary and tertiary structures of wtFOXP2-FHD and the Y540F mutant. Secondary structure was monitored by far-UV CD for values of ellipticity at 222 nm below a high-tension voltage of 700 V. Tertiary structure was monitored by intrinsic tryptophan fluorescence emission at an excitation wavelength of 280 nm. **(A and B)** wtFOXP2-FHD exhibited a percentage recovery of native signal of ~89 % and ~100 % when calculated from far-UV CD spectropolarimetry (A) and intrinsic tryptophan fluorescence (B) signals. **(C and D)** The Y540F mutant exhibited a percentage recovery of native signal of ~82 % and ~96 % when calculated from far-UV CD spectropolarimetry (C) and intrinsic tryptophan fluorescence emission (D) signals.

3.5.2 Secondary Structural Monitoring: GdnHCl Equilibrium (Un)Folding Studies

3.5.2.1 wtFOXP2-FHD

Both the unfolding and refolding of wtFOXP2-FHD exhibited a single transition from N ↔ U when monitored using far-UV CD (Figure 21A). Data for unfolding and refolding was superimposable for the post-transition and pre-transition regions of the data sets but became more sparsely distributed with respect to each data set in the transition region between GdnHCl concentrations of 2 – 3,5 M (Figure 21A). Despite this dispersion, a two-state folding curve could be fit for both unfolding data (R^2 = 0,98) and refolding data (R^2 = 0,99) monitored by far-UV CD (Figure 21A). In addition, the fitted curves did not overlay. A global fit of wtFOXP2-FHD equilibrium unfolding and refolding studies monitored by far-UV CD satisfied a two-state folding mechanism (Figure 21B). From this fit, thermodynamic parameters for unfolding were calculated, whereby: ΔG_{H2O} = 25,8 ± 2,7 kJ.mol^{-1}, m = 6,7 ± 0,7 kJ.mol^{-1}.M^{-1}, and $[GdnHCl]_{1/2}$ = 3,8 M.

3.5.2.2 Y540F Mutant

Secondary structural unfolding of the Y540F mutant underwent a single transition from N ↔ U over a GdnHCl concentration range of 3,25 M – 4,5 M (Figure 21C). In contrast, the refolding of the secondary structure displayed two transitions: the first between 3,5 M – 4,5 M and the second between 1,5 M and 2,25 M GdnHCl (Figure 21C). A stable secondary structural intermediate was observed over GdnHCl concentrations of 2,25 M – 3,5 M, exhibiting a unfolded fraction of 39 – 42 % secondary structure (Figure 21C). The data for unfolding was fitted by regression modelling to a two-state unfolding curve (R^2 = 0,99), and the data for refolding to a three-state curve (R^2 = 0,996). The folding mechanism of the Y540F mutant was disparate, whereby the secondary structure of the mutant unfolds by a two state mechanism N → U and refolds according to a three-state mechanism U → I → N (Figure 21D). As folding for the Y540F mutant is not reversible by virtue of the unfolding and refolding curves not being superimposable, thermodynamic parameters for the Y540F mutant could not be determined.

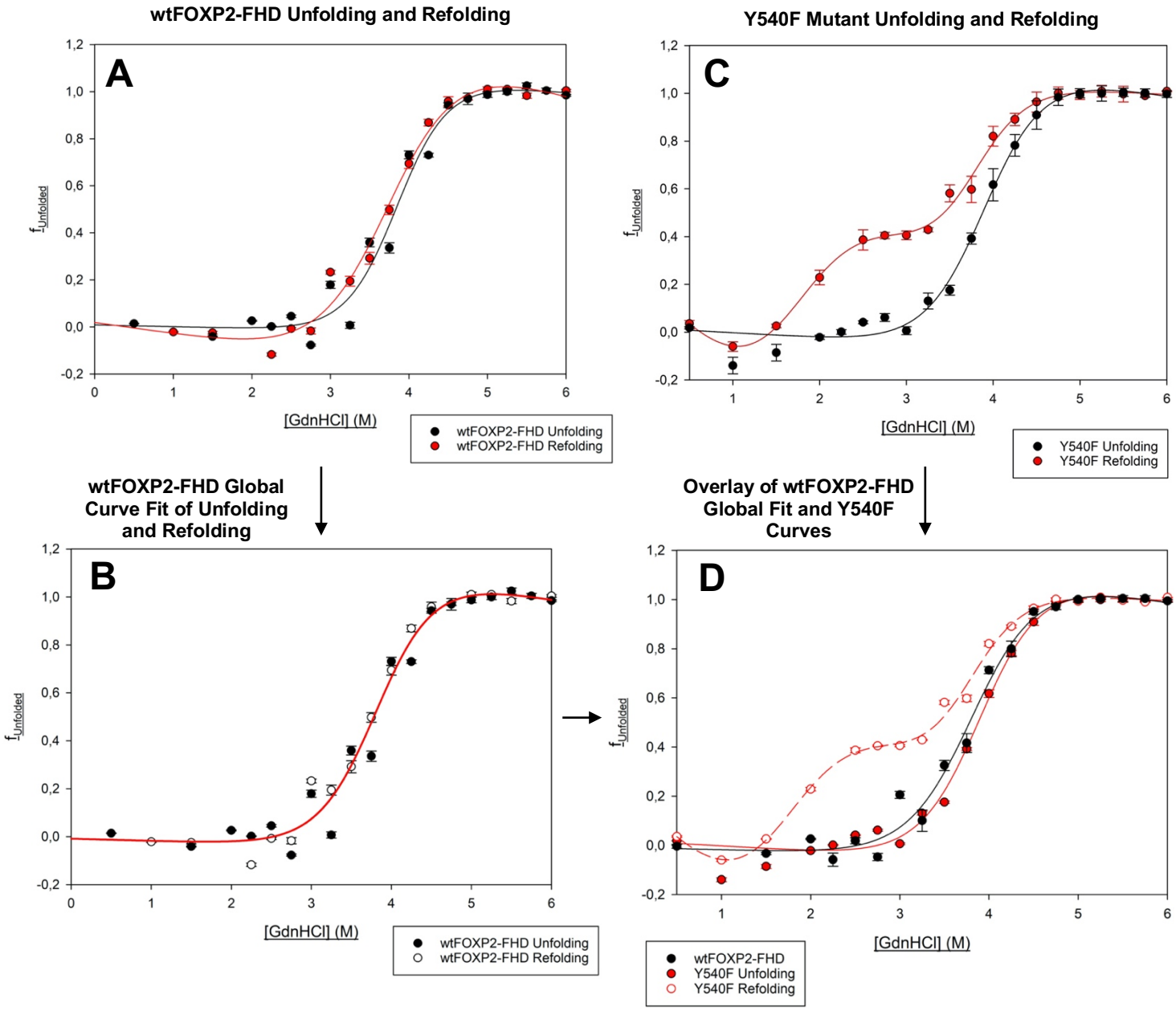


Figure 21: Far-UV CD Monitoring of ellipticity at 222 nm for wtFOXP2-FHD and the Y540F mutant

(A) Data for the fraction of unfolded wtFOXP2-FHD protein plotted against GdnHCl concentration during unfolding (black circle) and refolding (red circle). Error bars indicate a standard error of the mean. The unfolding and refolding of wtFOXP2-FHD is fitted to a two-state folding mechanism, N ↔ U. The two-state fit for unfolding and refolding had an R^2 = 0,98 and R^2 = 0,99, respectively. **(B)** A two-state folding mechanism was fit to data points from unfolding and refolding of wtFOXP2-FHD monitored by far-UV CD (R^2 = 0,98). Error bars indicate standard error of the mean. From this fit, thermodynamic parameters for unfolding were calculated, where: ΔG_{H2O} = 25,8 ± 2,67 kJ.mol^{-1}, m = 6,7 ± 0,7 kJ.mol^{-1}.M^{-1}, and $[GdnHCl]_{1/2}$ = 3,8 M. **(C)** Data for the fraction of unfolded Y540F mutant protein plotted

against GdnHCl concentration during unfolding (black circle) and refolding (red circle). Error bars indicate a standard error of the mean. The unfolding of Y540F mutant is fitted to a N → U two-state folding mechanism (R^2 = 0,99) and refolding is fit to a U → I → N three-state folding mechanism (R^2 = 0,996). **(D)** wtFOXP2-FHD structure monitored by far-UV CD folds according to a N ↔ U two-state mechanism (black solid line), while the Y540F mutant unfolds by a N → U two-state mechanism (red solid line) and refolds by a three-state U → I → N mechanism (red dashed line). The denaturation midpoint of wtFOXP2-FHD unfolding occurs at a GdnHCl concentration in comparison with the unfolding curve of the Y540F mutant.

3.5.3 Tertiary Structural Monitoring: GdnHCl Equilibrium (Un)Folding Studies

3.5.3.1 wtFOXP2-FHD

Intrinsic tryptophan fluorescence emission monitored at wavelengths 330 nm and 340 nm revealed that both tertiary structural unfolding and refolding occurred over a single transition (Figure 22A). Differences in $[GdnHCl]_{1/2}$ for unfolding and refolding were observed depending on the monitored wavelength. For unfolding and refolding, intrinsic tryptophan fluorescence monitored at 340 nm – relating to solvent exposed tryptophan residues Trp533 and Trp573 – revealed a lower $[GdnHCl]_{1/2}$ in comparison to the $[GdnHCl]_{1/2}$ from fluorescence monitored at 330 nm, relating to the hydrophobically buried tryptophan Trp548 (Figure 22A). This is not evidence for hysteresis; rather, an equilibrium of unfolded / refolded species which exhibit fully solvent exposed Trp533 and Trp573 residues occurs at a lower GdnHCl concentration in comparison to an equilibrium of unfolded / refolded species exhibiting fully solvent exposed Trp548. This implies that the tertiary structural elements of the protein which are more solvent exposed unfold initially, and then are gradually followed by the unfolding of interior tertiary structure. In terms of refolding, this implies that interior tertiary structure refolds first, followed by solvent exposed tertiary structure. Two-state unfolding curves were fitted to the data ($R^2_{330nm\ Un.}$ = 0,999; $R^2_{340nm\ Un.}$ = 0,999; $R^2_{330nm\ Re.}$ = 0,998; $R^2_{340nm\ Re.}$ = 0,998) (Figure 22A). While the fitted curves for unfolding and refolding monitored at emission wavelengths of either 330 nm or 340 nm are superimposable, each pair of unfolding / refolding of curves for emission at 330 nm or 340 nm is not superimposable with one another (Figure 11C). Therefore, the

folding of interior and solvent exposed tertiary structure for wtFOXP2-FHD is reversible, but does not occur over the same GdnHCl concentration.

A two-state unfolding curve was globally fit to the data for wtFOXP2-FHD intrinsic tryptophan fluorescence ($R^2 = 0,997$) in order to obtain the thermodynamic parameters of tertiary structural folding, calculated as: ΔG_{H2O} = 31,9 ± 0,8 kJ.mol^{-1}, m = 8,2 ± 0,2 kJ.mol^{-1}.M^{-1}, and $[GdnHCl]_{1/2}$ = 3,9 M (Figure 22B).

3.5.3.2 Y540F Mutant

Intrinsic tryptophan fluorescence emission monitored at wavelengths of 330 nm and 340 nm revealed that both tertiary structural unfolding and refolding occurred over a single transition (Figures 22C). Differences in $[GdnHCl]_{1/2}$ for unfolding and refolding were observed depending on the monitored wavelength. For unfolding, intrinsic tryptophan fluorescence monitored at 340 nm – relating to solvent exposed tryptophan residues Trp533 and Trp573 – revealed a lower $[GdnHCl]_{1/2}$ in comparison to $[GdnHCl]_{1/2}$ from emisison monitored at 330 nm, relating to the hydrophobically buried tryptophan Trp548 (Figure 22C). This would imply that an equilibrium of unfolded / refolded species which exhibit fully solvent exposed Trp533 and Trp573 residues occurs at a lower GdnHCl concentration in comparison to an equilibrium of unfolded / refolded species exhibiting fully solvent exposed Trp548. Therefore, the tertiary structural elements of the protein which are more solvent exposed unfold initially, and then are gradually followed by the unfolding of interior tertiary structure.

Tertiary structural refolding monitored at 330 nm and 340 nm also revealed a lower $[GdnHCl]_{1/2\ 340nm}$ than $[GdnHCl]_{1/2\ 330nm}$ (Figure 22C). In comparison with unfolding, the data for tertiary structural refolding exhibited a more significant difference between denaturation midpoints when monitored at either 330 nm and 340 nm (Figure 22C), whereby $[GdnHCl]_{1/2\ 330nm}$ = 3,7 M and $[GdnHCl]_{1/2\ 340nm}$ = 3,6 M. Therefore, half of the folded fraction of Y540F mutant protein exhibits tryptophan residue Trp548 reburied within the protein interior at a higher concentration of GdnHCl (3,7 M) in comparison to Trp533 and Trp573 which only return to their respective degree of solvent exposure at a lower [GdnHCl] (3,6 M). This implies that interior tertiary structural components first refold, followed by solvent exposed tertiary structure – a clear reflection of the aforementioned unfolding mechanism.

The fitted curves generated from equilibrium unfolding data monitored at 330 nm and 340 nm significantly overlay (Figure 22C). This implies that the interior and surface exposed tertiary structure of the Y540F mutant denatures to similar degrees of solvent exposure over the sequential increase of GdnHCl concentration. In addition, the fitted curve for refolding monitored at 330 nm overlays with the corresponding unfolding curve monitored at 330 nm (corresponding to Trp548) (Figure 22C). This suggests that the interior tertiary structure of the protein unfolds and refolds according to the same mechanism. However, while both equilibrium unfolding and refolding monitored at 340 nm each exhibit a two-state folding mechanism, there is evidence for hysteresis as both curves differ significantly by their respective $[GdnHCl]_{1/2}$, and do not overlay with one another (Figure 22C). This indicates that the surface exposed tertiary structure of the Y540F mutant unfolds and refolds according to differing folding pathways, although each mechanism is two-state. Therefore, while both the interior and surface exposed tertiary structure of the Y540F mutant unfolds by the same mechanism, the refolding pathway of tertiary structure differs between interior and surface exposed structure.

A global fit of data ($R^2 = 0,99$) obtained for Y540F mutant equilibrium unfolding and refolding studies monitored by intrinsic tryptophan fluorescence at 330 nm and 340 nm fitted a two-state mechanism of unfolding, and yielded overall thermodynamic parameters for unfolding of: ΔG_{H2O} = 28,9 ± 1,8 kJ.mol^{-1}, m = 8,0 ± 0,5 kJ.mol^{-1}.M^{-1}, and $[GdnHCl]_{1/2}$ = 3,6 M (Figure 22D). Overall, while wtFOXP2-FHD and the Y540F mutant both fold according to a two-state mechanism, each protein exhibits a significantly different folding pathway, evidenced by the significant difference in $[GdnHCl]_{1/2}$ for each protein (Figure 22E).

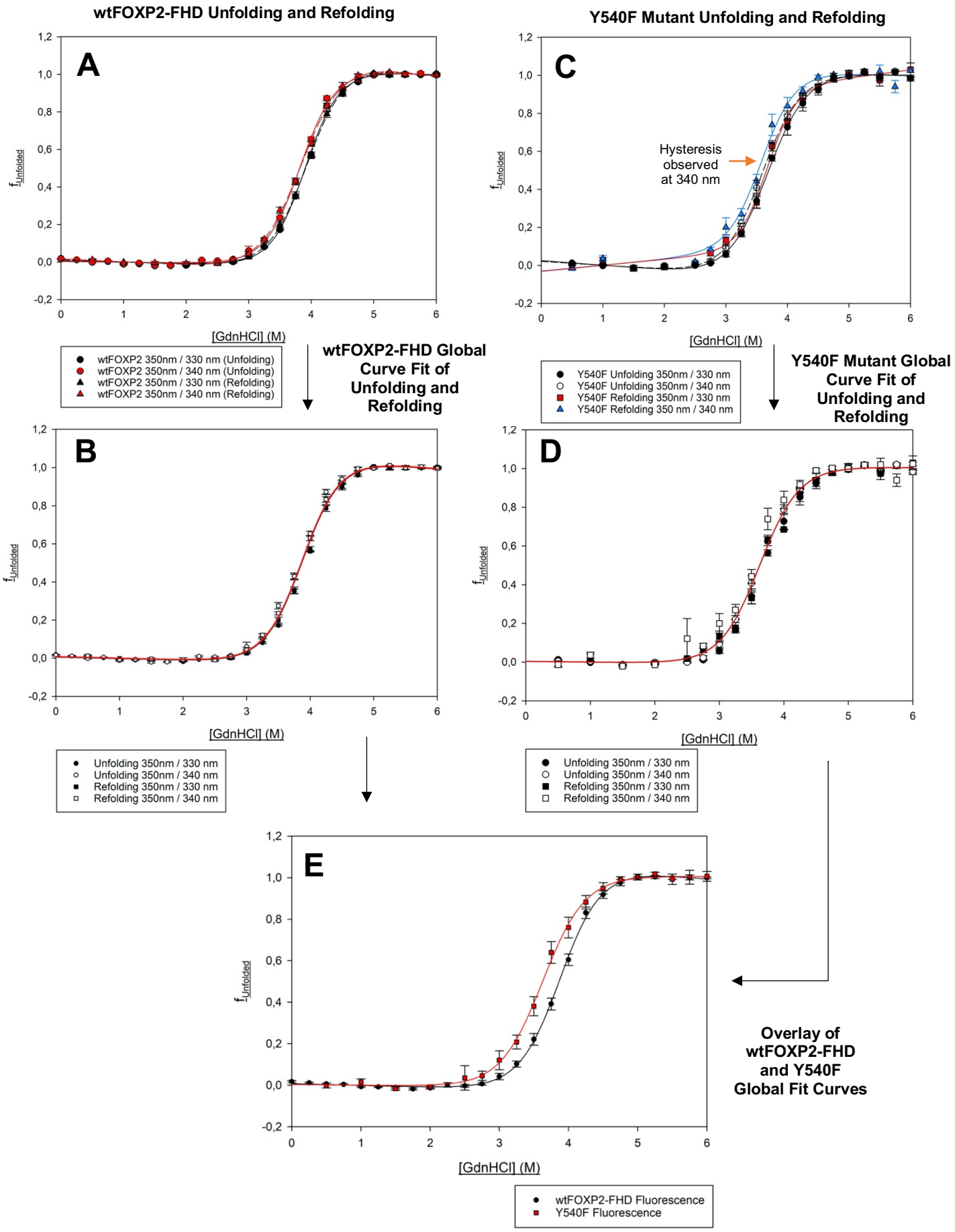

wtFOXP2-FHD Unfolding and Refolding
A
f Unfolded
[GdnHCl] (M)
wtFOXP2 350nm / 330 nm (Unfolding)
wtFOXP2 350nm / 340 nm (Unfolding)
wtFOXP2 350nm / 330 nm (Refolding)
wtFOXP2 350nm / 340 nm (Refolding)
wtFOXP2-FHD Global Curve Fit of Unfolding and Refolding
B
Unfolding 350nm / 330 nm
Unfolding 350nm / 340 nm
Refolding 350nm / 330 nm
Refolding 350nm / 340 nm
Y540F Mutant Unfolding and Refolding
C
Hysteresis observed at 340 nm
Y540F Unfolding 350nm / 330 nm
Y540F Unfolding 350nm / 340 nm
Y540F Refolding 350nm / 330 nm
Y540F Refolding 350 nm / 340 nm
Y540F Mutant Global Curve Fit of Unfolding and Refolding
D
E
Overlay of wtFOXP2-FHD and Y540F Global Fit Curves
wtFOXP2-FHD Fluorescence
Y540F Fluorescence

Figure 22

Figure 22: Intrinsic tryptophan fluorescence monitoring of wtFOXP2-FHD folding at emission wavelengths 330 nm and 340 nm

The fraction of unfolded wtFOXP2-FHD was plotted against GdnHCl concentration. All errors bars indicate standard error of the mean. The fitted curves separately overlay for unfolding and refolding monitored at 330 nm and 340 nm. **(A)** Unfolding of wtFOXP2-FHD monitored at emission wavelengths of 330 nm (black circle) and 340 nm (red circle). Data for both wavelengths was fit to a N ↔ U two-state unfolding curve (R^2_{330nm} = 0,999 and R^2_{340nm} = 0,999) (black solid line – 330 nm, red solid line – 340 nm). Refolding of wtFOXP2-FHD monitored at emission wavelengths of 330 nm (black square) and 340 nm (red square). Data for both wavelengths was fitted to a two-state unfolding curve (R^2_{330nm} = 0,998 and R^2_{340nm} = 0,998) (black solid line – 330 nm, red solid line – 340 nm). **(B)** A two-state folding mechanism was fitted to data points from unfolding and refolding of wtFOXP2-FHD monitored at emission wavelengths for 330 nm and 340 nm (R^2 = 0,997). Error bars indicate standard error of the mean. From this fit, thermodynamic parameters for tertiary structural unfolding were calculated, whereby: ΔG_{H2O} = 31,9 ± 0,8 kJ.mol^{-1}, m = 8,2 ± 0,2 kJ.mol^{-1}.M^{-1}, and $[GdnHCl]_{1/2}$ = 3,9 M. **(C)** Overlays of fluorescence monitored unfolding at 330 nm and 340 nm. The fitted curves for unfolding monitored at 330 nm (black solid line) and 340 nm (black dashed line), as well as refolding monitored at 330 nm (red solid line) significantly overlay. However, the fitted curve for refolding monitored at 340 nm (blue solid line) does not overlay with any other curves; therefore, hysteresis (orange arrow) is observed when monitoring fluorescence emission at 340 nm – the emission wavelength corresponding to surface exposed tryptophans Trp533 and Trp573. However, no hysteresis is observed when monitoring unfolding / refolding at 330 nm – the emission wavelength corresponding to buried residue Trp548. **(D)** A two-state folding mechanism was fitted to data points from unfolding and refolding of the Y540F mutant monitored at emission wavelengths for 330 nm and 340 nm (R^2 = 0,99). Error bars indicate standard error of the mean. From this fit, thermodynamic parameters for unfolding were calculated, whereby: ΔG_{H2O} = 28,9 ± 1,8 kJ.mol^{-1}, m = 8,0 ± 0,5 kJ.mol^{-1}.M^{-1}, and $[GdnHCl]_{1/2}$ = 3,6 M. **(E)** Folding monitored by intrinsic tryptophan fluorescence shows wtFOXP2-FHD and the Y540F mutant have significantly different folding pathways, although both fold according to a two-state mechanism. The denaturation midpoint for the Y540F mutant (red solid line) is considerably lower in comparison with that of wtFOXP2-FHD (black solid line).

3.5.4 Contrast of Secondary and Tertiary Structural Folding Pathways

3.5.4.1 wtFOXP2-FHD

wtFOXP2-FHD was observed to unfold and refold by a two state mechanism, when monitoring secondary or tertiary structure. A global fit of data obtained from wtFOXP2-FHD unfolding and refolding monitored by both far-UV CD and intrinsic tryptophan fluorescence fitted a two-state mechanism of unfolding ($R^2 = 0{,}997$), and yielded overall thermodynamic parameters for unfolding of: ΔG_{H20} = 29,0 ± 1,0 kJ.mol^{-1}, m = 7,5 ± 0,3 kJ.mol^{-1}.M^{-1}, and $[GdnHCl]_{1/2}$ = 3,9 M (Figure 23).

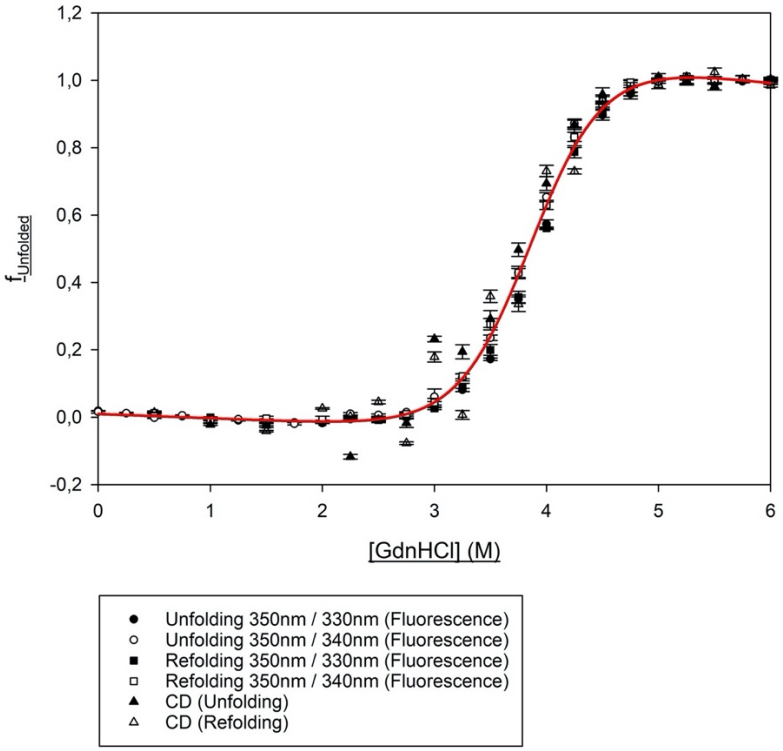


Figure 23: Global fit of two-state folding mechanism for wtFOXP2-FHD.

Data collected for unfolding and refolding of wtFOXP2-FHD monitored by far-UV CD spectropolarimetry and intrinsic tryptophan fluorescence was fitted to a two-state folding mechanism N ↔ U ($R^2 = 0{,}997$). Error bars indicate standard error of the mean. The global thermodynamic parameters calculated for wtFOXP2-FHD unfolding were: ΔG_{H2O} = 29,0 ± 1,0 kJ.mol^{-1}, m = 7,5 ± 0,3 kJ.mol^{-1}.M^{-1}, and $[GdnHCl]_{1/2}$ = 3,9 M.

3.5.4.2 Y540F Mutant

The comparison of fitted curves for the Y540F mutant monitored by both far-UV CD and intrinsic tryptophan fluorescence exhibited significant differences with regard to structural unfolding and refolding (Figure 24A). During unfolding (N → U), the denaturation midpoint of unfolded secondary structure occurred at $[GdnHCl]_{1/2}$ = 3,9 M while denaturation midpoint of unfolded tertiary structure occurred at $[GdnHCl]_{1/2}$ = 3,6 M (Figure 24A). In terms of refolding, secondary structure followed a three state mechanism (U → I → N), reaching a denaturation midpoint for the U → I transition at $[GdnHCl]_{1/2\ U\rightarrow I}$ = 3,7 M and a denaturation midpoint for the I → U transition at $[GdnHCl]_{1/2\ I\rightarrow N}$ = 1,7 M (Figure 24B). Tertiary structural refolding (U → N) monitored at 330 nm reached its denaturation midpoint at $[GdnHCl]_{1/2}$ = 3,7 M, while tertiary structure monitored at 340 nm reached its denaturation midpoint at $[GdnHCl]_{1/2}$ = 3,6 M (Figure 24B). Comparison of fitted curves show that Y540F unfolds and refolds according to two significantly different pathways (Figure 24B). Both the secondary and tertiary structure of the Y540F mutant unfolds by a two-state mechanism, although the secondary structure has its denaturation midpoint at a higher $[GdnHCl]_{1/2}$ in comparison with the tertiary structure. Therefore, the tertiary structure of the protein unfolds at a lower concentration of GdnHCl than the secondary structure, and hence is more sensitive to unfolding in GdnHCl (Figure 24B). When the protein refolds, the secondary structure follows a three-state mechanism, forming a stable intermediate, while the tertiary structure refolds by a two state mechanism (Figure 24B). In addition, the interior tertiary structure of the protein (monitored by fluorescence of the Trp548 residue at 330 nm) refolded along the same pathway as it did during unfolding, indicated by the superimposition of the unfolding and refolding curves at 330 nm (Figure 24B). The surface exposed tertiary structure monitored by fluorescence of the Trp533 and Trp573 residues at 340 nm, refolded over a lower concentration of GdnHCl within the 0 – 6 M range (Figure 24B), and therefore folded at a later stage in comparison with the interior tertiary structure. Furthermore, the curve for refolding of secondary structure overlays with the curve for tertiary structural refolding monitored at 330 nm during its U → I transition, deviating from the buried tertiary structure refolding curve at $[GdnHCl]_{1/2\ U\rightarrow I}$ (Figure 24C). At the highest GdnHCl concentration where the secondary structural intermediate is stable, the fraction of unfolded buried tertiary structure is ~0,20 and unfolded solvent exposed tertiary structure is ~0,35. The intermediate remains stable between GdnHCl concentrations of 2,25 – 3,25 M (Figure 24B). At the start of the I → N transition of

the secondary structure, tertiary structure monitored at both 330 nm and 340 nm have almost completely refolded (Figure 24B). At the end of the I → N transition both secondary and tertiary have completely refolded.

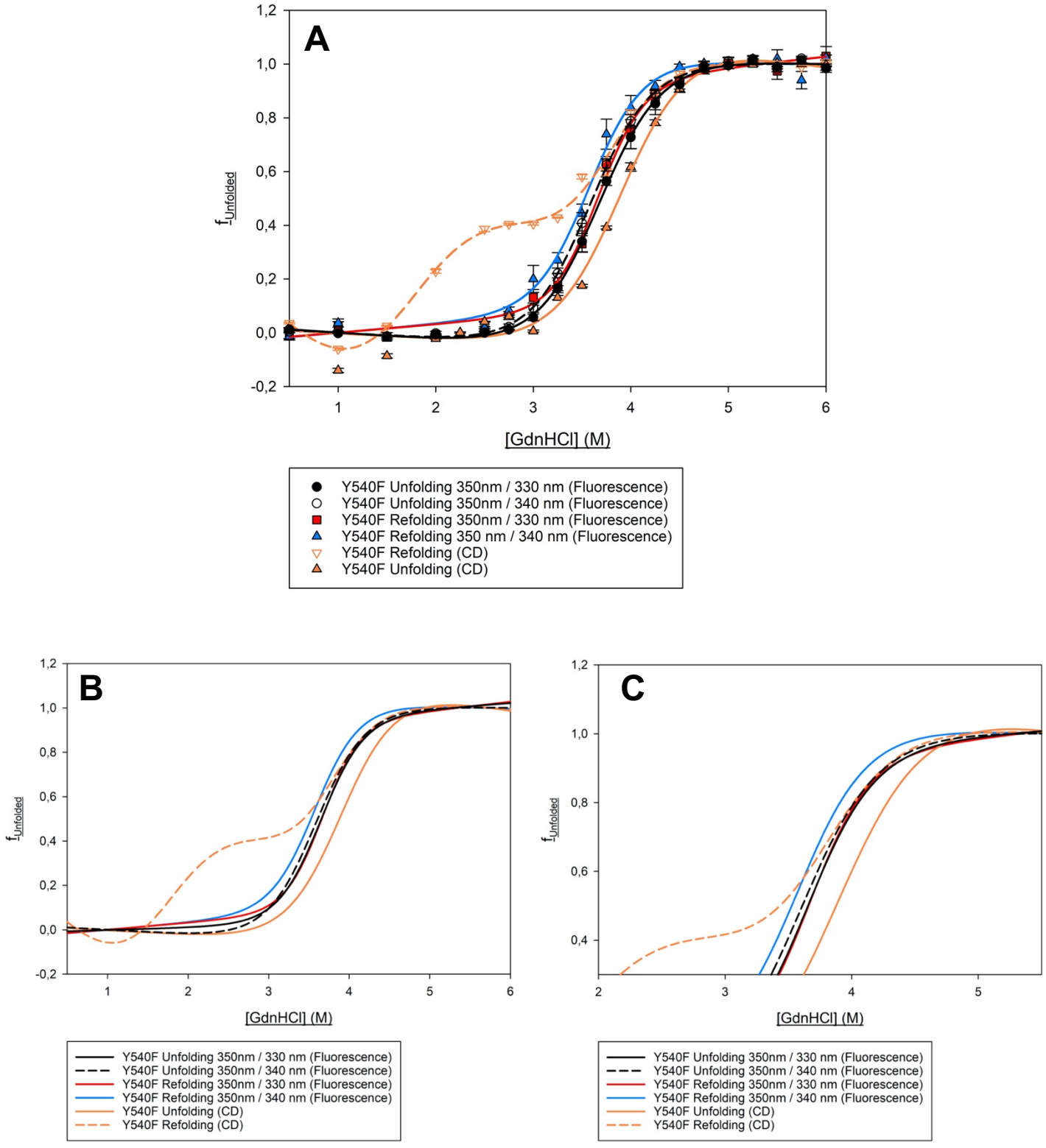


Figure 24

Figure 24: Folding curves for the Y540F mutant for unfolding and refolding monitored by far-UV CD and intrinsic tryptophan fluorescence.

(A) Overlay of data points and fitted curves for unfolding and refolding monitored by both far-UV CD and intrinsic tryptophan fluorescence. **(B)** Fitted curves for unfolding and refolding monitored by both far-UV CD and intrinsic tryptophan fluorescence. The curves for Y540F unfolding and refolding monitored at 330 nm (black and red solid lines, respectively) are superimposable, and unfolding monitored at 340 nm (dashed black line) is partially superimposable with unfolding and refolding monitored at 330 nm. Refolding monitored at 340 nm is not completely superimposable within any other curves, but is partially superimposable with the U → I transition state for refolding monitored by far-UV CD (dashed orange line). The unfolding transition monitored by far-UV CD (solid orange line) exists at the highest $[GdnHCl]_{1/2}$ in comparison with other curve transition regions. **(C)** Focus on aspects of curves between GdnHCl concentrations of 2 – 5,5 M and f_U of 0,35 – 1,2. It is more easily observed that the secondary structure refolding curve is superimposed on the curve for tertiary structural refolding monitored at 330 nm during its U → I transition. The secondary structure curve deviates from superimposition over the buried tertiary structure refolding curve at $[GdnHCl]_{1/2\ U\rightarrow I}$.

3.5.5 Comparison of Folding Pathways for wtFOXP2-FHD and the Y540F Mutant

The folding pathways of wtFOXP2-FHD and the Y540F mutant differ significantly when monitored by both far-UV CD spectropolarimetry and intrinsic tryptophan fluorescence emission. Global fits of the data show that, as previously mentioned, the secondary structure of wtFOXP2-FHD folds according to a N ↔ U two-state mechanism, while the Y540F mutant unfolds by a N → U two-state mechanism and refolds by a three-state U → I → N mechanism (Figure 25). In terms of tertiary structure folding, both wtFOXP2-FHD and the Y540F mutant fold according to a N ↔ U two-state mechanism, although the Y540F mutant enters its transition region at a lower GdnHCl concentration in comparison with wtFOXP2-FHD (Figure 25).

A comparison of all globally fitted unfolding curves shows superimposition of wtFOXP2-FHD tertiary structural folding and Y540F mutant secondary structural unfolding (Figure 25). The

secondary structure for wtFOXP2-FHD unfolds by a two-state mechanism but shows lower cooperativity in comparison with secondary structure folding for the Y540F mutant (Figure 25). Tertiary structural folding of the Y540F mutant overlays partially with the U $\rightarrow$ I transition region of the mutant, and exhibits no superimposition with any other curves (Figure 25).

Thermodynamic parameters calculated from global fits indicate that the difference in conformational stability, cooperativity and denaturation midpoint between wtFOXP2-FHD and the Y540F mutant monitored by far-UV CD is negligible, with slightly higher values observed for the Y540F mutant ($\Delta(\Delta G_{H2O})$ = 2,2 kJ.mol^{-1}, Δm = 0,4 kJ.mol^{-1}.M^{-1}, and $\Delta[\text{GdnHCl}]_{1/2}$ = 0,1 M) (Table 2). This indicates that the secondary structure of the Y540F mutant is very slightly more conformationally stable and compact than the secondary structure of wtFOXP2-FHD. The thermodynamic parameters calculated for folding monitored by intrinsic tryptophan fluorescence emission also show a negligible difference for ΔG_{H2O} and m of tertiary structural folding between wtFOXP2-FHD and the Y540F mutant. The tertiary structure of the Y540F mutant is slightly less conformationally stable and compact in comparison with wtFOXP2-FHD ($\Delta(\Delta G_{H20})$ = -3,1 kJ.mol^{-1}, Δm = -0,3 kJ.mol^{-1}.M^{-1}) (Table 2). In addition, the denaturation midpoint of Y540F tertiary structure is significantly lower than that for wtFOXP2-FHD ($\Delta[\text{GdnHCl}]_{1/2}$ = -0,3 M) (Table 2).

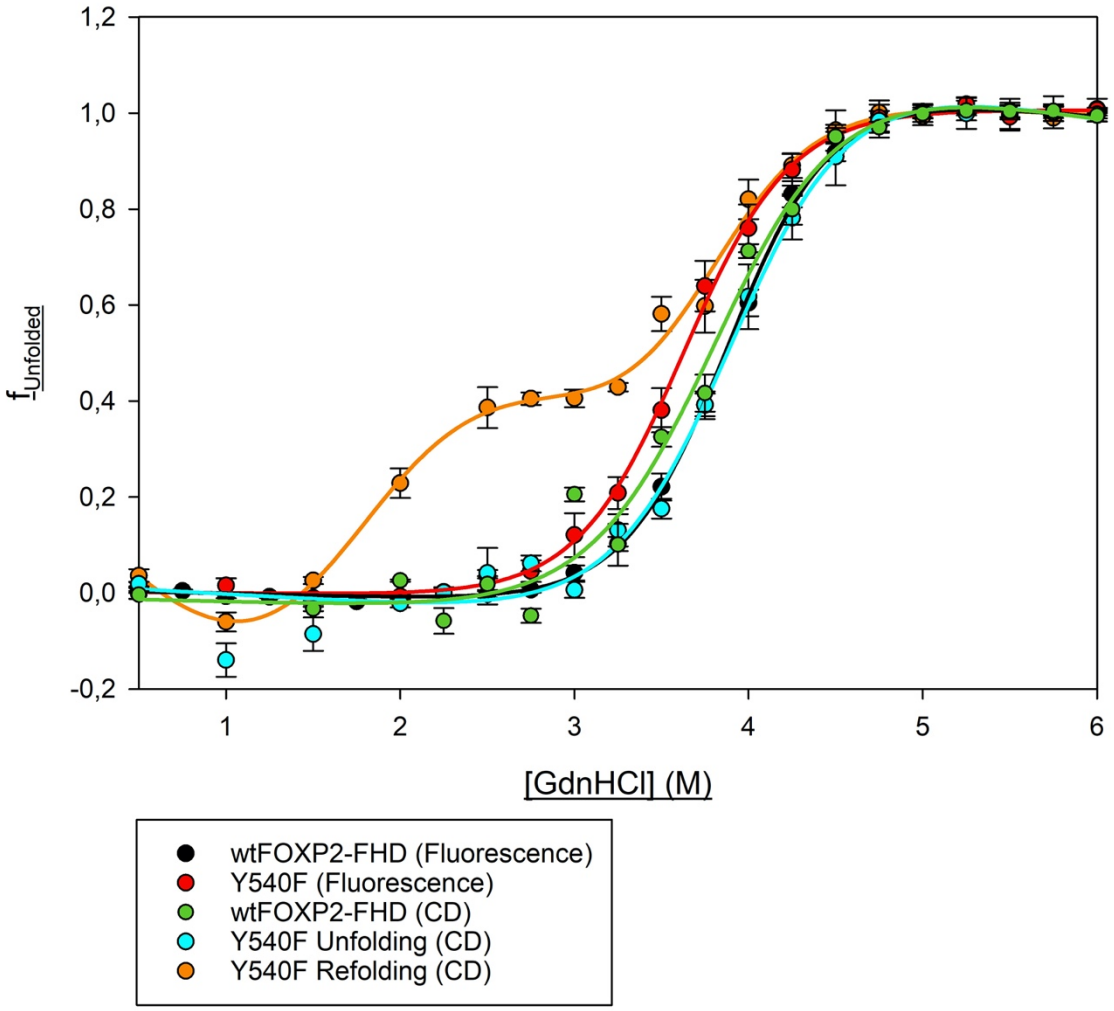


Figure 25: Overlay of global folding curves for wtFOXP2-FHD and Y540F mutant folding monitored by far-UV CD and intrinsic tryptophan fluorescence emission.

Superimposition of wtFOXP2-FHD tertiary structural folding (black line) and Y540F mutant secondary structural unfolding (cyan line) are superimposable. In terms of increasing $[GdnHCl]_{1/2}$, the curves are ordered: wtFOXP2-FHD tertiary structure folding (black) and Y540F mutant secondary structure unfolding (cyan) > wtFOXP2-FHD secondary structure folding (green) > Y540F mutant tertiary structure refolding (red). The secondary structure of the Y540F mutant exhibits refolding according to a three-state mechanism (orange).

Table 2: Thermodynamic Parameters of Global Folding Fits for wtFOXP2-FHD and the Y540F Mutant.

Values for $\Delta(\Delta G_{H2O})$, Δm and $\Delta[\text{GdnHCl}]_{1/2}$ are calculated as Value_{Y540F} - $\text{Value}_{wtFOXP2\text{-}FHD}$. (*) indicates Y540F mutant refolding subtracted by wtFOXP2-FHD. (**) indicates Y540F mutant refolding subtracted by Y540F unfolding.

Spectroscopic Probe	Protein	ΔG_{H20} (kJ.mol^{-1})	*m*-value (kJ.mol^{-1}.M^{-1})	$[\text{GdnHCl}]_{1/2}$ (M)	$\Delta(\Delta G_{H20})$ (kJ.mol^{-1})	Δm-value (kJ.mol^{-1}.M^{-1})	$\Delta[\text{GdnHCl}]_{1/2}$ (M)
Circular Dichroism	wtFOXP2-FHD	25,8 ± 2,7	6,7 ± 0,7	3,8			
	Y540F Unfolding	28,0 ± 2,8	7,2 ± 0,8	3,9	2,2	0,4	0,1
	Y540F $\text{Refolding}_{U\rightarrow I}$	30,7 ± 5,7	8,2 ± 1,3	3,7	5,0* 2,8**	1,5* 1,1**	-
	Y540F $\text{Refolding}_{I\rightarrow N}$	10,3 ± 3,7	6,1 ± 2,2	1,7	-15,5* -17,6**	-0,6* -1,1**	-
Intrinsic Tryptophan Fluorescence	wtFOXP2-FHD	31,9 ± 0,8	8,21 ± 0,21	3,9			
	Y540F	28,9 ± 1,8	8,0 ± 0,5	3,6	-3,1	-0,3	-0,3

3.5.6 Thermal Unfolding

wtFOXP2-FHD and the Y540F mutant were thermally denatured over a temperature range of 20 – 70 °C monitored by far-UV CD (Figure 26). Data was collected at 1 °C increments. Both proteins unfolded from their native to unfolded states over a single N $\rightarrow$ U transition. (Figure 26). The transition region for wtFOXP2-FHD was between 49 – 61 °C, or 322,15 – 334,15 K, and between 50 – 61 °C, or 323,15 – 334,15 K, for the Y540F mutant. However, as the thermal unfolding reaction for both wtFOXP2-FHD and the Y540F mutant was irreversible (data not shown), thermodynamic parameters such as enthalpy of unfolding (ΔH_m), transition temperature (T_m) and entropy of unfolding (ΔS_m) could not be accurately determined.

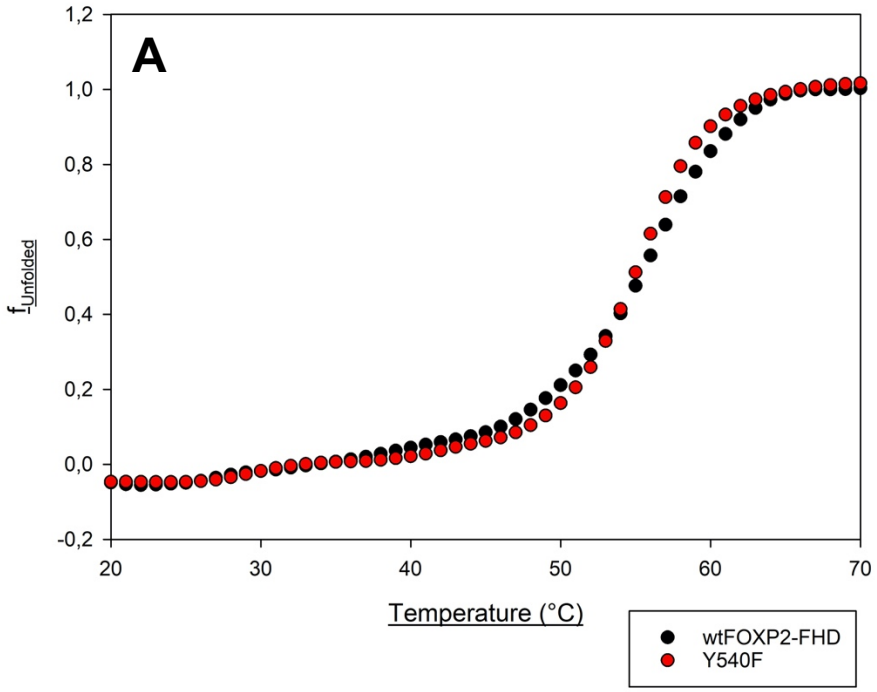


Figure 26: Thermal unfolding and van't Hoff plots of wtFOXP2-FHD and the Y540F mutant

(A) wtFOXP2-FHD (black circles) and the Y540F mutant (red circles) were observed to unfold over a single transition step from N → U. The unfolding reaction was observed to be irreversible (data not shown).

Chapter 4: Discussion

Among the FOX family of TFs, the FHD of the FOXP subfamily shows the unique capability of forming domain swapped dimers (Stroud *et al.,* 2006; Bandukwala *et al.*, 2011; Medina *et al.*, 2016). The formation of such dimers has been speculated to have importance in facilitating interchromosomal association, thus enabling the transcriptional regulation of distal genetic elements (Chen *et al.*, 2015). Domain swapping is inherently energetically unfavourable and its occurrence depends on sufficient interchain interaction between folding monomers as they tend towards a native conformation. While the FOXP2-FHD assembles as a domain swapped dimer upon folding, it exhibits an unexpectedly low propensity to do so, despite dimerisation being a proposed requirement for its ability to regulate transcriptional activity. This investigation aimed to elucidate the propensity of FOXP2-FHD to form dimers through probing its folding mechanism and structural stability.

4.1 wtFOXP2-FHD and the Y540F Mutant Exist as Dimers at Micromolar Concentrations in Solution

In the literature to date, wtFOXP2-FHD has been observed to form homodimers in a single solved crystal structure in the presence of DNA (Stroud *et al.*, 2006). Furthermore, a small fraction of dimeric wtFOXP2-FHD has been observed to exist in solution alongside its predominant monomeric form (Perumal *et al.,* 2015; Blane and Fanucchi, 2015; Morris and Fanucchi, 2016). Moreover, a FOXP2-FHD F541C variant has been engineered to form dimers in the presence of DNA while in solution (Morris and Fanucchi, 2016), as has a truncated form of FOXP2 encompassing the FHD and remaining C-terminal sequence of the protein (Häußermann *et al.*, 2019). However, in the absence of DNA, wtFOXP2-FHD does not readily form dimers and has been calculated to exist only at millimolar concentrations in solution (Perumal *et al.*, 2015). In this study, quaternary structural characterisation of wtFOXP2-FHD by analytical SEC has shown that wtFOXP2-FHD dimer can exist at micromolar concentrations in solution (Figure 16). Notably, this is not observed of native wtFOXP2-FHD post-purification, but rather only after denaturation and refolding in GdnHCl. After purification, the K_d for dimer dissociation into monomer for wtFOXP2-FHD was calculated as 1,6 mM (Figure 15) but after refolding to a concentration of 1,38 M GdnHCl the K_d decreased

~69-fold to 22,5 uM. This stark contrast in K_d may be a result of GdnHCl facilitating dimer formation, as GdnHCl has been observed to stabilise protein structure by virtue of its salting-in effect (Bhuyan, 2002). However, the lack of dimer presence in the native protein may also be attributed to the bacterial expression system used to overexpress wtFOXP2-FHD. In two similar studies where K_d was calculated for FOXP1-FHD (Chu *et al,* 2011; Medina et al, 2016), protein was expressed in a BL21 (DE3) bacterial expression system and the K_d was observed to be in the micromolar range when analysed by SEC. Indeed, the K_d of FOXP1-FHD dimer dissociation into monomer calculated by Chu *et al.,* 2011, is closely comparable to that observed in this investigation for wtFOXP2-FHD. As FOXP1-FHD and FOXP2-FHD are both related, exhibit high sequence homology, have been observed to form both homo- and hetero-dimers, and are implicated in similar DNA binding processes, it is likely that FOXP2-FHD should form dimers at concentrations similar to that of FOXP1-FHD in solution. Therefore, while it cannot be assumed that the K_d measured for wtFOXP2-FHD after refolding reflects an accurate K_d for dimer dissociation into monomer in the absence of denaturant, it is evident that wtFOXP2-FHD dimers can exist at micromolar concentrations in solution.

The Y540F mutant also shows a substantial increase in dimer formation after refolding in GdnHCl (Figure 6). After expression in T7 Competent *E. coli*, the K_d of dimer dissociation into monomer for the mutant was calculated as 2,6 mM (Figure 15) but after refolding to a concentration of 1,38 M GdnHCl the K_d decreased ~31-fold to 77,9 uM (Figure 16). This difference in K_d may also be a consequence of expression in the chosen bacterial expression system. However, one cannot be truly confident that this explanation fully accounts for the difference in K_d as, in the literature, the Y540F mutant was expressed in a bacterial system identical to this study and was observed to have a K_d at a micromolar concentration, albeit higher than observed here (Perumal *et al.*, 2015). Quaternary structural analysis by SEC of the Y540F mutant expressed by Perumal *et al.*, 2015, revealed a K_d of 246 uM after protein purification, and without refolding in denaturant, although expression was under control of a different plasmid. In light of the discrepancy between K_d values for dimer dissociation into monomer calculated in this study and by Perumal *et al*., 2015, the Y540F mutant can be only qualitatively described as having a K_d at micromolar concentrations in solution.

wtFOXP2-FHD displays a greater propensity for dimerisation in comparison with the Y540F mutant. The Tyr → Phe substitution mutation was chosen in order to probe its effect on dimer

formation as it is uniquely observed in the amino acid sequence of FOXP3-FHD – a member that obligately forms dimers in solution (Bandukwala *et al.*, 2011). It was observed that the K_d of the Y540F mutant, after purification, was 1,6 times greater than that of wtFOXP2-FHD, and 3,4 times greater after refolding in GdnHCl. Therefore, the mutation discourages the dimerisation of the FOXP2-FHD in solution. However, this is only the case in the absence of DNA. The Y540F substitution mutation does not impede the DNA binding function of FOXP2-FHD. Dimeric species of both proteins were observed in the presence of DNA at a ratio of protein : DNA as low as 2,5 uM : 0,5 uM for wtFOXP2-FHD, and 1,0 uM : 0,5 uM for the Y540F mutant when probed by EMSA (Figure 3). Therefore, while the substitution mutation discourages dimer formation in the absence of DNA, it may act to promote dimerisation in the presence of DNA by up to 2,5-fold. This may be as an extended consequence of the increased stability of the hydrophobic core of the Y540F mutant, as will be discussed further (sections 4.3 and 4.4). The Tyr → Phe substitution mutation results in the increased compactness and stability of the hydrophobic interior, contributed to predominantly by buried aromatic residues on H1 and H2 of the protein. One such aromatic residue that contributes to the hydrophobic core is Trp548, located on H3 of the protein, the recognition helix which mediates DNA sequence recognition and sequence specific binding through embedment in the major groove of DNA. In order for major groove insertion to occur and instigate development of the protein-DNA complex, tight shape complementarity between the recognition helix and the DNA groove must exist (Noy *et al.*, 2016). It is possible that the increased stability of the core, contributed to by Trp548, extends to stabilise H3, creating a more definite shape complementarity between helix and groove, thus promoting the formation of the protein-DNA complex.

4.2 wtFOXP2-FHD Folds by a Two-State Mechanism

Small, globular proteins are commonly observed to fold by a two-state mechanism (U ↔ N) whereby only the unfolded and folded states populate the folding pathway (Jackson, 1998). The folding pathway can be monitored by a variety of spectroscopic techniques such as fluorescence spectroscopy, CD spectropolarimetry and nuclear magnetic resonance spectroscopy (Beatrice *et al.*, 2001). Denaturation is typically induced over a gradient of increasing pressure, temperature or chemical denaturant. wtFOXP2-FHD was denatured in GdnHCl and its unfolding monitored by both far-UV CD spectropolarimetry and intrinsic

tryptophan fluorescence emission. Far-UV CD was used to monitor secondary structural changes while intrinsic tryptophan fluorescence was monitored at wavelengths of 330 nm and 340 nm in order to probe tertiary structural changes affected at the protein surface and within the hydrophobic interior. Monitoring at both these locations within the protein was possible by virtue of the FOXP2-FHD containing a buried tryptophan residue (Trp548) and two solvent exposed tryptophan residues (Trp533 and Trp573). There was sufficient resolution between the fluorescence signals emitted by the tryptophan residues at each location in FOXP2-FHD in order for monitoring at both 330 nm and 340 nm to be possible, and hence tertiary structural changes at the protein surface and interior to be discernible (Figure 19).

Unfolding of wtFOXP2-FHD was fully reversible and ~100 % recoverable when monitored by fluorescence emission (Figure 20). It was observed to be ~89 % recoverable when monitored by far-UV CD. The difference in percentage recoverability when monitored by both spectroscopic probes is possibly justified as technical error arising from imprecision of protein concentrations for measured native and refolded protein. The method to investigate recoverability depends on two dilution steps: first with denaturant and then with buffer, and hence an accumulation of pipetting error is possible. However, the difference in percentage may indicate that the secondary structure of the protein does not fully refold to its native conformation in the presence of GdnHCl. Far-UV CD monitoring at low concentrations of GdnHCl, in the range of 2 – 3,5 M, revealed data points for f_U against [GdnHCl] to be disperse around the two-state fit (Figure 21), and therefore the native secondary structure may not be equally present upon refolding for all wtFOXP2-FHD protein macromolecules in solution.

Despite such dispersion of data points, wtFOXP2-FHD folding was able to be fitted to a two-state folding mechanism monitored by far-UV CD (R^2 = 0,98) (Figure 21). Folding was also fitted to a two-state mechanism monitored by intrinsic tryptophan fluorescence (R^2 = 0,997) (Figure 22). Notably, wtFOXP2-FHD unfolding and refolding monitored at wavelengths of 330 nm and 340 nm showed a slight difference in $[GdnHCl]_{1/2}$ ($\Delta[GdnHCl]_{1/2\ 340\text{-}330nm}$= 0,1 M). Therefore, solvent exposed tryptophan residues Trp533 and Trp573 become fully solvent exposed at a lower GdnHCl concentration in comparison with buried Trp548. This implies that the solvent exposed tertiary structure of wtFOXP2-FHD unfolds before the interior tertiary structure, and conversely, the interior tertiary structure refolds before that which is solvent exposed. A N ↔ U two-state folding mechanism was globally fitted for wtFOXP2-FHD

(R^2 = 0,997) taking into account data for both far-UV CD and intrinsic tryptophan fluorescence emission at 330 nm and 340 nm, revealing thermodynamic parameters for folding of: ΔG_{H20} = 29,0 ± 1,0 kJ.mol^{-1}, m = 7,5 ± 0,3 kJ.mol^{-1}.M^{-1}, and $[GdnHCl]_{1/2}$ = 3,9 M (Figure 23).

4.3 The Y540F Mutant Forms a Stable Intermediate Governed by Secondary Structural Elements Upon Refolding in GdnHCl

As for wtFOXP2-FHD, the folding of the Y540F mutant was monitored spectroscopically by far-UV CD spectropolarimetry and by intrinsic tryptophan fluorescence emission at wavelengths of 330 nm and 340 nm. The Y540F mutant was found to be ~96 % recoverable when monitored by intrinsic tryptophan fluorescence and ~82 % recoverable when monitored by far-UV CD (Figure 20). This too may indicate that either the total native secondary structure of the mutant is not assumed upon refolding in GdnHCl, or that a small minority of Y540F protein macromolecules in solution do not reassume their native secondary structure upon refolding.

A description of the folding mechanism of the Y540F mutant is further complicated by its secondary structural unfolding and refolding pathways. In unfolding, the mutant was observed to follow a two-state unfolding pathway characterised by one transition of N → U over GdnHCl concentrations of 3,25 M – 4,5 M. In contrast, upon refolding, the mutant underwent two transitions from U → I → N, with U → I between 3,25 M – 4,5 M and I → N between 1,5 M and 2,25 M GdnHCl. A stable secondary structural intermediate was observed to form between GdnHCl concentrations of 2,25 M – 3,25 M (Figure 21). This intermediate was only observed when monitoring equilibrium refolding by far-UV CD, and was not observed when monitoring unfolding or refolding by intrinsic tryptophan fluorescence emission (Figure 22). In addition, the $[GdnHCl]_{1/2}$ for secondary structural unfolding was observed to be 0,2 M higher than the $[GdnHCl]_{1/2}$ during refolding along the U → I transition (Table 2). This raises the questions of why the secondary structure of the mutant is less sensitive to denaturant during unfolding, and more pertinently, what accounts for a stable secondary structural intermediate to form during refolding, but not during unfolding. The presence of a secondary structural intermediate is certainly statistically significant as the data falls within a confidence interval of 95 % for a three-state fit (Figure 25). At first, it may seem plausible that GdnHCl plays a role in stabilising the secondary structure of the mutant as has been documented in the literature for other proteins

(Morjana *et al.*, 1993; Monera *et al.*, 1994; Bhuyan, 2002; Povarova, *et al.*, 2010). However, as this intermediate is not additionally observed during unfolding, it seems less to be the reason.

In terms of tertiary structure, the mutant exhibited a similar denaturation midpoint for unfolding monitored at 330 nm and refolding monitored at 340 nm (Figure 22). The denaturation midpoint of refolding monitored at 340 nm was lowest in comparison. In addition, the $[GdnHCl]_{1/2}$ of the tertiary structure of the mutant occurs at the start of the formation of the stable intermediate (Figure 27). Furthermore, the tertiary structure is almost completely refolded by the beginning of the I $\rightarrow$ N secondary structural transition state (Figure 25). This can be interpreted as that the intermediate state of the mutant forms when half of the tertiary structure has refolded, and subsequently, the secondary structure only completely refolds once the complete tertiary structure has refolded (Figure 27). However, as tertiary structure is inferred by solvent exposure of tryptophan residues by intrinsic tryptophan fluorescence, it can be alternatively stated that all monitored tryptophan residues exist at the same degree of solvent exposure as the native protein. The stable secondary structural intermediate forms at a fraction unfolded (f_U) secondary structure of 38 – 41 %, equivalent to a folded fraction of 59 – 62 % (Figure 27). The hydrophobic interior of the protein is formed by aromatic amino acids located at the start of H1, and then along H2 until Trp548 located on H3 (Figure 21). With regard to secondary structure, 64 % of the total helical content is bounded by aromatic residues that contribute to stabilising the hydrophobic interior, with Trp548 buried in the core, and Trp533 in close proximity to the core although more solvent exposed (Figure 21 and 27). With these considerations, the reduction in secondary structural sensitivity to GdnHCl may be explained, as well as a folding mechanism for the Y540F mutant described as follows.

Firstly, the Y540F mutation is located on H2 in close proximity to the hydrophobic core of the protein. Mutation from tyrosine to phenylalanine causes the amino acid side chain to become more hydrophobic and thus interact more strongly within the hydrophobic core. As Phe540 is located on H2, this brings H2 closer to the interior through the formation of van der Waals interactions, and hence causes the protein to be more compact. It is this compactness that allows for the secondary structure to more readily resist denaturing by GdnHCl, as it reduces the solvent accessible surface area available to the denaturant (Figure 27).

Secondly, the increase in core hydrophobicity also explains the three-state mechanism observed during refolding. In this regard, the U $\rightarrow$ I transition can be correlated with the formation of the secondary structure associated with the hydrophobic interior of the protein, and the I $\rightarrow$ N transition with the refolding of remaining secondary structure not found in the interior (Figure 27). In terms of helical content, 64 % of the helices are bounded by aromatic amino acids that contribute to the stability of the hydrophobic interior of the protein (Figure 27). The intermediate state of the Y540F mutant is observed when the folded secondary structure fraction is 59 – 62 % of the total folded secondary structure (Figure 27). It is possible that the mutation allows for hydrophobic collapse to be more imminent as a result of increased hydrophobicity, whereby first the hydrophobic interior of the protein forms, followed by the remainder of secondary structure (H3, and H4), which themselves are not stabilised by hydrophobic interactions. This explanation is further supported by the conformational stability calculated for unfolding and refolding monitored by far-UV CD (Table 2). The ΔG_{H2O} for secondary structural unfolding (ΔG_{H2O} = 28,0 ± 2,8 kJ.mol^{-1}) and U $\rightarrow$ I refolding (ΔG_{H2O} = 30,7 ± 5,7 kJ.mol^{-1}) are within error of one another, indicating that the native and refolded intermediate states exhibit similar conformational stability, which may be a result from the hydrophobicity of the core. In addition, there may be a correlation between the ΔG_{H2O} for I $\rightarrow$ N refolding (ΔG_{H2O} = 10,3 ± 3,7 kJ.mol^{-1}) and the remainder of the secondary structure that refolds to a native state, as approximately one third of the secondary structure refolds from I $\rightarrow$ N and the $\Delta G_{H2O\ I\rightarrow N}$ is approximately one third of $\Delta G_{H2O\ U\rightarrow I}$. Lastly, it must be acknowledged that although FOXP2-FHD has been observed to form domain swapped dimers at micromolar concentrations (Figure 16), it is unlikely that the equilibrium unfolding data represents domain swapping. Far-UV CD and intrinsic tryptophan fluorescence were conducted at protein concentrations of 1,5 uM and 10 uM, respectively, far below the calculated K_d for the Y540F mutant.

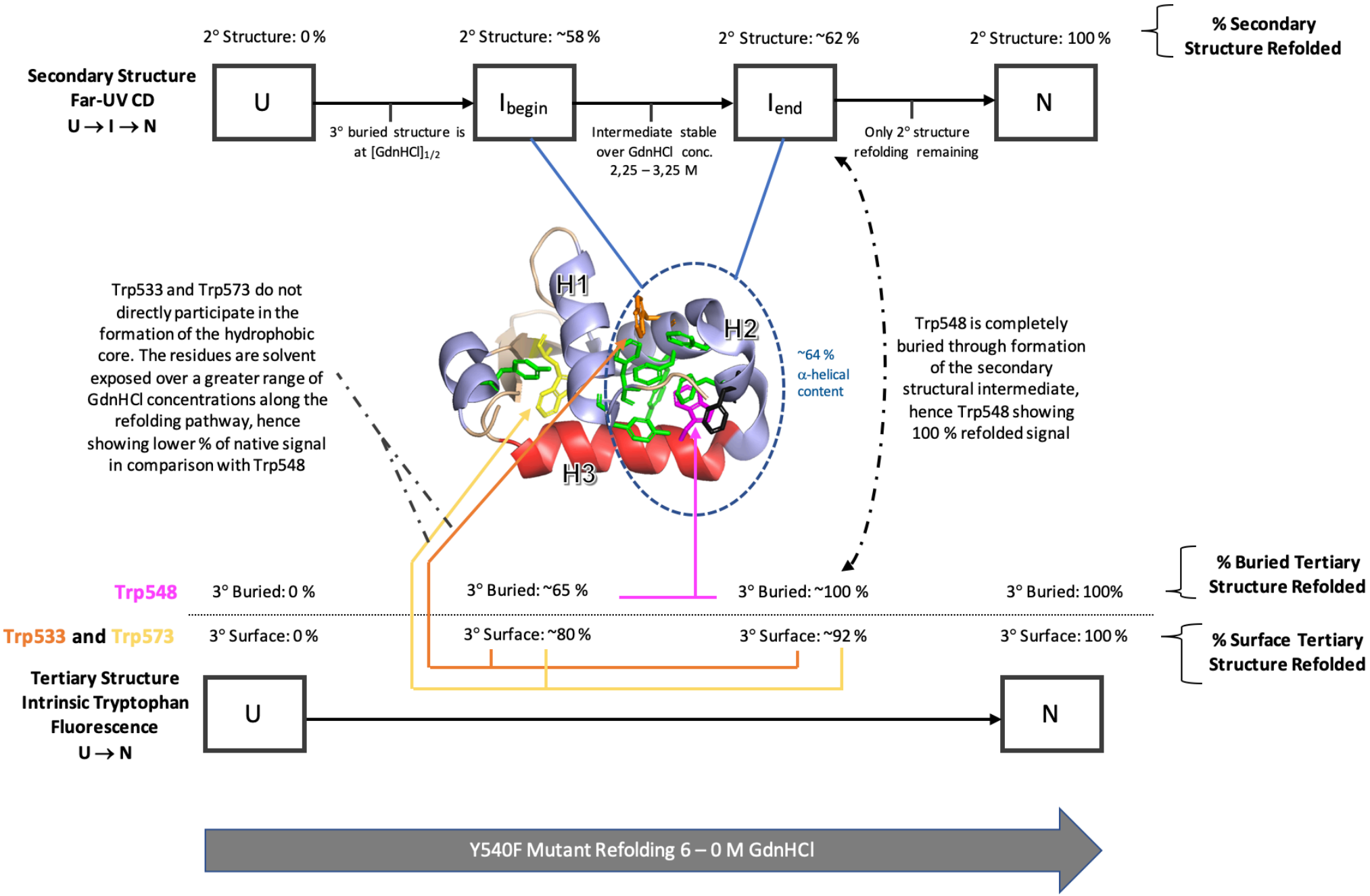

2° Structure: 0 %
2° Structure: ~58 %
2° Structure: ~62 %
2° Structure: 100 %
% Secondary Structure Refolded
Secondary Structure Far-UV CD U → I → N
U
I begin
I end
N
3° buried structure is at [GdnHCl]1/2
Intermediate stable over GdnHCl conc. 2,25 – 3,25 M
Only 2° structure refolding remaining
Trp533 and Trp573 do not directly participate in the formation of the hydrophobic core. The residues are solvent exposed over a greater range of GdnHCl concentrations along the refolding pathway, hence showing lower % of native signal in comparison with Trp548
H1
H2
H3
~64 % α-helical content
Trp548 is completely buried through formation of the secondary structural intermediate, hence Trp548 showing 100 % refolded signal
Trp548
3° Buried: 0 %
3° Buried: ~65 %
3° Buried: ~100 %
3° Buried: 100%
% Buried Tertiary Structure Refolded
Trp533 and Trp573
3° Surface: 0 %
3° Surface: ~80 %
3° Surface: ~92 %
3° Surface: 100 %
% Surface Tertiary Structure Refolded
Tertiary Structure Intrinsic Tryptophan Fluorescence U → N
U
N
Y540F Mutant Refolding 6 – 0 M GdnHCl

Figure 27

Figure 27: Sequential formation of the stable secondary structural intermediate of Y540F mutant

The Y540F mutant exhibits a three state U → I → N refolding mechanism when monitored by far-UV CD and a two state U → N refolding mechanism when monitored by intrinsic tryptophan fluorescence. At 6 M GdnHCl, both the secondary and tertiary structure of the Y540F mutant are completely unfolded ("U" blocks). At the U → I denaturation midpoint for refolded secondary structure, the U → N denaturation midpoint for refolded tertiary structural is reached ("U" → "I_{begin}"). A stable secondary structural intermediate comprising of ~58 – 62 % refolded secondary structure exists over a range of GdnHCl concentrations between 2,25 M and 3,25 M ("I_{begin}" and "I_{end}" blocks). This percentage of secondary structure closely compares with the percentage of α-helical content (64 %) comprising the hydrophobic interior of the native protein. Over the concentration range of GdnHCl corresponding to formation of the stable secondary structural intermediate ("I_{begin}" → "I_{end}"), Trp548 (magenta) is buried in the hydrophobic interior of the protein, hence exhibiting 100 % native signal despite the protein not yet being fully refolded. Unlike Trp548, residues Trp533 (orange) and Trp573 (yellow) do not play a direct role in the formation of the hydrophobic interior and, therefore, exhibit a signal corresponding to solvent exposure over a greater range of concentrations of GdnHCl. At the start of the I → N transition region for secondary structural refolding ("I_{end}" → "N") , close to 100 % of tertiary structure has refolded (or rather, Trp548, Trp533 and Trp573 exhibit a fluorescence signal equivalent to that observed of the native state). The remainder of secondary structure refolds over GdnHCl concentrations of 0 – 2,25 M, until the protein reaches its native conformation ("N" blocks).

4.4 The Secondary Structural Stability of the Y540F Mutant May Account for its Reduced Dimerisation Propensity

The differences of ΔG_{H2O}, m and $[GdnHCl]_{1/2}$, alongside the rationale behind the Y540F mutant exhibiting a three-state refolding pathway when monitored by far-UV CD may further explain the differences in K_d calculated for wtFOXP2-FHD and the Y540F mutant (section 3.4.1). Upon refolding, the K_d for wtFOXP2-FHD was ~3,5-fold lower than the Y540F mutant meaning the wild type is over three times more likely to dimerise. FOXP2-FHD forms dimers by three-dimensional domain swapping whereby H1 and H2 between monomers are exchanged and intertwined, as observed in the crystal structure by Stroud *et al*., 2006. Three-dimensional

domain swapping is an overall energetically unfavourable process, whereby the energetic requirements for domain swapping are offset by the formation of bonds between swapped dimer interfaces, and by the degree of hinge-loop flexibility (Bennett *et al.*, 1994; Green *et al.*, 1995; Parker *et al.*, 1998; Rousseau *et al.*, 2001). In terms of thermodynamics, the entropic loss upon forming a higher order conformation should be compensated for by an enthalpic gain arising from the formation of new intermolecular surfaces following domain swapping (Rousseau *et al.*, 2001). The Y540F mutant exhibits a greater secondary structural conformational stability in its hydrophobic interior, of which H1 and H2 are the main contributors. It is possible that increased conformational stability of the interior which involves H1 and H2 impedes the ability of the Y540F mutant to participate in domain swapped dimerisation as a result of increased compactness. Therefore, while the Y540F substitution mutation increases the stability and compactness of the hydrophobic core of FOXP2-FHD, it decreases its propensity to engage in domain swapped dimerisation.

4.5 Conclusions

In this study, wtFOXP2-FHD was observed to have a K_d of dimer dissociation into monomer at a micromolar concentration in the absence of DNA, where up till present it has been undocumented in the literature. The Y540F mutant too exhibited a K_d in the micromolar range. wtFOXP2-FHD has a K_d for dimer dissociation into monomer ~3,5 times lower in comparison to the Y540F mutant, suggesting wtFOXP2-FHD is able to dimerise more readily in solution. In terms of folding, wtFOXP2-FHD folds by a two-state mechanism, evident by the absence of any intermediate along its folding pathway. In contrast, the Y540F mutant exhibits a characteristically disparate and unexpected folding mechanism, whereby its secondary structure unfolds by a two-state mechanism and refolds according to a three-state mechanism. Its tertiary structure, however, appears to unfold and refold according to a two-state mechanism. The three-state refolding pathway of the secondary structure of the Y540F mutant is attributed to the increased hydrophobicity contributed by the Tyr $\rightarrow$ Phe substitution which, in turn, results in the formation of a stable, compact folding intermediate. The Y540F mutant's greater conformational stability may give justification for its lower propensity to dimerise. As H1 and H2 are exchanged during domain swapped dimerisation and are the main elements of the hydrophobic interior, the greater hydrophobicity and compactness of the protein interior may impede the partial unfolding of H1 and H2 required for domain swapped dimerisation.

Hence, a higher K_d of dimer dissociation into monomer is observed with regard to the Y540F mutant. Although the Y540F mutation acts to stabilise and increase the compactness of FOXP2-FHD, this increase in conformational stability appears to be at the cost of a reduction in domain swap dimerisation propensity. wtFOXP2-FHD may, therefore, exist more prominently as a domain swapped dimer as a result of reduced conformational stability from the presence of Tyr540. This confers a greater flexibility to wtFOXP2-FHD to participate in the exchange of H1 and H2 between monomers which, following domain swapping, are subsequently stabilised by Tyr540 in the serial aromatic interface of the domain swapped dimer (Figure 10).

In this investigation, a description of the folding mechanism and stability of wtFOXP2-FHD and the Y540F mutant have been outlined from a thermodynamic perspective. However, in order to more comprehensively describe the folding mechanism of FOXP2-FHD, additional kinetic equilibrium unfolding studies would be of benefit. Furthermore, unfolding studies at a higher concentration of protein, whereby FOXP2-FHD exists predominantly as a dimer, would allow for a more complete understanding of its mechanism of domain swapped dimerisation. Such a description of this mechanism would greatly increase our understanding of how the structure of FOXP2-FHD contributes to its role of transcriptional regulation within the cell.

References

Alber, T. (1989). Mutational effects on protein stability. *Annual Review of Biochemistry*, **58**(1), 765 – 792.

Alber, T., Bell, J. A., Sun, D. P., Nicholson, H., Wozniak, J. A., Cook, S., and Matthews, B. W. (1988). Replacements of Pro86 in phage T4 lysozyme extend an alpha-helix but do not alter protein stability. *Science*, **239**(4840), 631 – 635.

Alber, T., Sun, D. P., Nye, J. A., Muchmore, D. C., and Matthews, B. W. (1987). Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry*, **26**(13), 3754 – 3758.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**(4096), 223 – 230.

Anufrieva, E. V., Birshtein, T. M., Nekrasova, T. N., Ptitsyn, O. B., and Sheveleva, T. V. (1967). The models of the denaturation of globular proteins. II. Hydrophobic interactions and conformational transition in polymethacrylic acid. In *Journal of Polymer Science Part C: Polymer Symposia,* **16**(6) 3519 – 3531

Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M., and Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiology Reviews*, **29**(2), 231 – 262.

Arridge, R. G. C., and Cannon, C. G. (1964). Calculation of the CONH dipole contribution to lattice energies of amides, polyamides, and polypeptides. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, **278**(1372), 91 – 109.

Baldwin, R. L., and Zimm, B. H. (2000). Are denatured proteins ever random coils?. *Proceedings of the National Academy of Sciences*, **97**(23), 12391 – 12392.

Bandukwala, H.S., Wu, Y., Feuerer, M., Chen, Y., Barboza, B., Ghosh, S., Stroud, J.C., Benoist, C., Mathis, D., Rao, A. and Chen, L. (2011). Structure of a domain swapped FOXP3 dimer on DNA and its function in regulatory T cells. *Immunity*, **34**(4), 479 – 491.

Beamer, L. J., and Pabo, C. O. (1992). Refined 1.8 Å crystal structure of the λ repressor-operator complex. *Journal of Molecular Biology*, **227**(1), 177 – 196.

Becktel, W. J., and Schellman, J. A. (1987). Protein stability curves. *Biopolymers: Original Research on Biomolecules*, **26**(11), 1859 – 1877.

Bennett, M.J., Choe, S. and Eisenberg, D. (1994). Domain swapping: entangling alliances between proteins. *Proceedings of the National Academy of Sciences*, **91**(8), 3127 – 3131.

Bennett, M.J., Schlunegger, M.P. and Eisenberg, D. (1995). 3D domain swapping: a mechanism for oligomer assembly. *Protein Science*, **4**(12), 2455 – 2468.

Berkovits – Cymet, H. J., Amann, B. T., and Berg, J. M. (2004). Solution structure of a CCHHC domain of neural zinc finger factor-1 and its implications for DNA binding. *Biochemistry*, **43**(4), 898 – 903.

Bhaskara, R. M., and Srinivasan, N. (2011). Stability of domain structures in multi – domain proteins. *Scientific Reports*, **1**, 40.

Bhuyan, A. K. (2002). Protein stabilization by urea and guanidine hydrochloride. *Biochemistry*, **41**(45), 13386 – 13394.

BinSheng, Y. (2011). The Stability of a Three-State Unfolding Protein. *Thermodynamics: Physical Chemistry of Aqueous Systems*, **365**, 1 – 12.

Blake, A., and Peacocke, A. R. (1965). Optical rotatory dispersion and secondary structure of ribonucleic acid in mammalian ribosomes. *Nature*, **208**(5017), 1319 – 1320.

Blane, A., and Fanucchi, S. (2015). Effect of pH on the Structure and DNA Binding of the FOXP2 Forkhead Domain. *Biochemistry*, **54**(25), 4001 – 4007.

Börjesson, U., and Hünenberger, P. H. (2003). Effect of mutations involving charged residues on the stability of staphylococcal nuclease: a continuum electrostatics study. *Protein Engineering*, **16**(11), 831 – 840.

Bosshard, H. R., Marti, D. N., and Jelesarov, I. (2004). Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *Journal of Molecular Recognition*, **17**(1), 1 – 16.

Chen, Y., Chen, C., Zhang, Z., Liu, C.C., Johnson, M.E., Espinoza, C.A., Edsall, L.E., Ren, B., Zhou, X.J., Grant, S.F. and Wells, A.D. (2015). DNA binding by FOXP3 domain swapped dimer suggests mechanisms of long-range chromosomal interactions. *Nucleic Acids Research*, **43**(2), 1268 – 1282.

Chen, Y.W., Stott, K. and Perutz, M.F. (1999). Crystal structure of a dimeric chymotrypsin inhibitor 2 mutant containing an inserted glutamine repeat. *Proceedings of the National Academy of Sciences*, **96**(4), 1257 – 1261.

Chowriappa, P., Dua, S., Kanno, J., and Thompson, H. W. (2008). Protein structure classification based on conserved hydrophobic residues. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **6**(4), 639 – 651.

Chu, Y. P., Chang, C. H., Shiu, J. H., Chang, Y. T., Chen, C. Y., & Chuang, W. J. (2011). Solution structure and backbone dynamics of the DNA-binding domain of FOXP1: Insight into its domain swapping and DNA binding. *Protein Science*, **20**(5), 908 – 924.

Corrales, M., Cuscó, P., Usmanova, D. R., Chen, H. C., Bogatyreva, N. S., Filion, G. J., and Ivankov, D. N. (2015). Machine learning: how much does it tell about protein folding rates?. *PloS one*, **10**(11), 1 – 12.

Creighton, T. E. (1990). Protein folding. *The Biochemical Journal*, **270**(1), 1–16.

Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*. Macmillan. 134 – 147

Crevenna, A. H., Naredi-Rainer, N., Lamb, D. C., Wedlich-Söldner, R., and Dzubiella, J. (2012). Effects of Hofmeister ions on the α-helical structure of proteins. *Biophysical Journal*, **102**(4), 907 – 915.

Cui, D., Ou, S., and Patel, S. (2014). Protein-spanning water networks and implications for prediction of protein–protein interactions mediated through hydrophobic effects. *Proteins: Structure, Function, and Bioinformatics*, **82**(12), 3312 – 3326.

Cuiffo, B. G., Campagne, A., Bell, G. W., Lembo, A., Orso, F., Lien, E. C., Bhasin, M. K., Raimo, M., Hanson, S. E., Marusyk, A. (2014) MSC-regulated microRNAs converge on the transcription factor FOXP2 and promote breast cancer metastasis. *Cell Stem Cell,* **15**, 762–774.

Derrida, B. (1980). Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, **45**(2), 79.

Díaz-Villanueva J. F., Díaz-Molina R., García-González V. (2015). Protein Folding and Mechanisms of Proteostasis. *Int J Mol Sci*., **16**(8), 17193–17230.

Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry*, **24**(6), 1501–1509.

Dill, K. A., and Alonso, D. O. V. (1988). Conformational entropy and protein stability. *Protein Structure and Protein Engineering,* **39**, 51 – 58.

Dill, K. A., Bromberg, S., Yue, K., Chan, H. S., Ftebig, K. M., Yee, D. P., and Thomas, P. D. (1995). Principles of protein folding—a perspective from simple exact models. *Protein Science*, **4**(4), 561 – 602.

Dill, K. A., Ozkan, S. B., Shell, M. S., and Weikl, T. R. (2008). The protein folding problem. *Annu. Rev. Biophys.*, **37**, 289 – 316.

Dobson, C. M. (2004). Principles of protein folding, misfolding and aggregation. *Seminars in Cell and Developmental Biology* **15**(1), 3 – 16.

Dyson, H. J., Lerner, R. A., and Wright, P. E. (1988). The physical basis for induction of protein-reactive antipeptide antibodies. *Annual Review of Biophysics and Biophysical Chemistry*, **17**(1), 305 – 324.

Dyson, H. J., Rance, M., Houghten, R. A., Wright, P. E., and Lerner, R. A. (1988). Folding of immunogenic peptide fragments of proteins in water solution: II. The nascent helix. *Journal of Molecular Biology*, **201**(1), 201 – 217.

Elrod-Erickson, M., Rould, M. A., Nekludova, L., and Pabo, C. O. (1996). Zif268 protein–DNA complex refined at 1.6 Å: a model system for understanding zinc finger–DNA interactions. *Structure*, **4**(10), 1171 – 1180.

Enard, W., Gehre, S., Hammerschmidt, K., Hölter, S. M., Blass, T., Somel, M., Brückner, M. K., Schreiweis, C., Winter, C., Sohr, R. (2009) A Humanized Version of Foxp2 Affects Cortico-Basal Ganglia Circuits in Mice. *Cell,* **137**, 961–971.

Englander, S. W., and Mayne, L. (2014). The nature of protein folding pathways. *Proceedings of the National Academy of Sciences*, **111**(45), 15873 – 15880.

Esposito, L., and Daggett, V. (2005). Insight into ribonuclease A domain swapping by molecular dynamics unfolding simulations. *Biochemistry*, **44**(9), 3358 – 3368.

Facchiano, A. M., Colonna, G., and Ragone, R. (1998). Helix stabilizing factors and stabilization of thermophilic proteins: an X – ray based study. *Protein Engineering*, **11**(9), 753 – 760.

Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Current Opinion in Structural Biology*, **7**, 3–9.

Fersht, A. R. (2004). Relationship of Leffler (Brønsted) α values and protein folding Φ values to position of transition-state structures on reaction coordinates. *Proceedings of the National Academy of Sciences*, **101**(40), 14338 – 14342.

Fitzpatrick, A. W., Knowles, T. P., Waudby, C. A., Vendruscolo, M., and Dobson, C. M. (2011). Inversion of the balance between hydrophobic and hydrogen bonding interactions in protein folding and aggregation. *PLoS Computational Biology*, **7**(10).

Funnell, A.P.W. and Crossley, M. (2012) Homo- and Heterodimerization in Transcriptional Regulation, Protein Dimerization and Oligomerization in Biology. *Advances in Experimental Medicine and Biology*, **747**. Springer, New York. 45 – 61.

Gajiwala, K. S., and Burley, S. K. (2000). Winged helix proteins. *Current Opinion in Structural Biology*. **10***,* 110–116.

Galilee, M., Britan-Rosich, E., Griner, S.L., Uysal, S., Baumgärtel, V., Lamb, D.C., Kossiakoff, A.A., Kotler, M., Stroud, R.M., Marx, A. and Alian, A. (2016). The preserved HTH-docking cleft of HIV-1 integrase is functionally critical. *Structure*, **24**(11), 1936 – 1946.

Garner, M. M. and Revzin, A. (1986). The use of gel electrophoresis to detect and study nucleic acid-protein interactions. *Trends Biochem Sci,* **11**, 395 – 396.

Gouldson, P.R., Higgs, C., Smith, R.E., Dean, M.K., Gkoutos, G.V. and Reynolds, C.A. (2000). Dimerisation and domain swapping in G-protein-coupled receptors: a computational study. *Neuropsychopharmacology*, **23**, S60 – S77.

Gowder, S. M., Chatterjee, J., Chaudhuri, T., and Paul, K. (2014). Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins. *The Scientific World Journal*, **2014**, 1 – 7

Green, S.M., Gittis, A.G., Meeker, A.K. and Lattman, E.E. (1995). One-step evolution of a dimer from a monomeric protein. *Nature Structural and Molecular Biology*, **2**(9), 746 – 751.

Greenfield, N. J., and Fasman, G. D. (1969). Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, **8**(10), 4108 – 4116.

Groszer, M., Keays, D. A., Deacon, R. M. J., de Bono, J. P., Prasad – Mulcare, S., Gaub, S., Baum, M. G., French, C. A., Nicod, J., Coventry, J. A. (2008) Impaired Synaptic Plasticity and

Motor Learning in Mice with a Point Mutation Implicated in Human Speech Deficits. *Curr. Biol.* **18**, 354–362.

Ha, J. H., Karchin, J. M., Walker-Kopp, N., Castañeda, C. A., & Loh, S. N. (2015). Engineered domain swapping as an on/off switch for protein function. *Chemistry & Biology*, **22**(10), 1384 – 1393.

Hall, T., Biosciences, I., and Carlsbad, C. (2011). BioEdit: an important software for molecular biology. *GERF Bull Biosci*, **2**(1), 60 – 61.

Harrison, S. C., and Aggarwal, A. K. (1990). DNA recognition by proteins with the helix – turn – helix motif. *Annual Review of Biochemistry*, **59**(1), 933 – 969.

Häußermann, K., Young, G., Kukura, P., and Dietz, H. (2019). Dissecting FOXP2 oligomerization and DNA binding. *Angewandte Chemie*, **131**(23), 7744 – 7749.

Herman, P., and Lee, J. C. (2012). The advantage of global fitting of data involving complex linked reactions. *Allostery*. Springer, New York. 399 – 421.

Herschlag, D., and Pinney, M. M. (2018). Hydrogen bonds: Simple after all?. *Biochemistry*, **57**(24), 3338 – 3352.

Hol, W. G. (1985). The role of the α-helix dipole in protein function and structure. *Progress in Biophysics and Molecular Biology*, **45**(3), 149 – 195.

Hong, P., Koza, S. and Bouvier, E. S. (2012). A review size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates. *Journal of Liquid Chromatography and Related Technologies*, **35**(20), 2923 – 2950.

Hong, S., Wang, D., Horton, J. R., Zhang, X., Speck, S. H., Blumenthal, R. M., and Cheng, X. (2017). Methyl-dependent and spatial-specific DNA recognition by the orthologous transcription factors human AP-1 and Epstein-Barr virus Zta. *Nucleic Acids Research*, **45**(5), 2503 – 2515.

Jackson, B.C., Carpenter, C., Nebert, D.W. and Vasiliou V. (2010). Update of human and mouse forkhead box (FOX) gene families. *Human Genomics,* **4**(5), 345 – 352.

Jackson, S. E. (1998). How do small single – domain proteins fold? *Folding and Design*, **3**(4), R81–R91.

Jackson, S. E., and Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2. 2. Influence of proline isomerization on the folding kinetics and thermodynamic characterization of the transition state of folding. *Biochemistry*, **30**(43), 10436–10443.

Jacoesen, C. F., and Linderstrøm-Lang, K. (1949). Salt linkages in proteins. *Nature*, **164**(4166), 411 – 412.

Kaestner K. H., Katz J., Liu Y., Drucker D. J., Schutz G. (1999). Inactivation of the winged helix transcription factor HNF3alpha affects glucose homeostasis and islet glucagon gene expression in vivo. *Genes Dev,* **13***,* 495 – 504

Karplus, M., and Weaver, D. L. (1979). Diffusion–collision model for protein folding. *Biopolymers: Original Research on Biomolecules*, **18**(6), 1421 – 1437.

Karplus, M., and Weaver, D. L. (1994). Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Science*, **3**(4), 650 – 668.

Kartha, G., Bello, J., and Harker, D. (1967). Tertiary structure of ribonuclease. *Nature*, **213**(5079), 862 – 865.

Katoh M. and Katoh M. (2005). Human FOX gene family. *International Journal of Oncology,* **25**(5), 1495–1500.

Kaufmann E. and Knöchel W. (1996). Five years on the wings of fork head. *Mech. Dev.,* **57**(1), 3 – 20.

Kauzmann, W. (1959). Some Factors in the Interpretation of Protein Denaturation. *Advances in Protein Chemistry*, **14**(C), 1 – 63.

Kelly, S. M., Jess, T. J., and Price, N. C. (2005). How to study proteins by circular dichroism. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, **1751**(2), 119 – 139.

Khan, K.H. (2013). Gene expression in mammalian cells and its applications. *Adv Pharm Bull*, **3**(2), 257 – 263.

Khazanovich, N., Bateman, K.S., Chernaia, M., Michalak, M. and James, M.N.G. (1996). Crystal structure of the yeast cell-cycle control protein, p13 suc1, in a strand-exchanged dimer. *Structure*, **4**(3), 299 – 309.

Killikelly, A., Benson, M.A., Ohneck, E.A., Sampson, J.M., Jakoncic, J., Spurrier, B., Torres, V.J. and Kong, X.P. (2015). Structure-based functional characterization of repressor of toxin (Rot), a central regulator of Staphylococcus aureus virulence. *Journal of Bacteriology*, **197**(1), 188 – 200.

Kim, J. G., and Hudson, L. D. (1992). Novel member of the zinc finger superfamily: A C2-HC finger that recognizes a glia-specific gene. *Molecular and Cellular Biology*, **12**(12), 5632 – 5639.

Kim, P. S., and Baldwin, R. L. (1984). A helix stop signal in the isolated S-peptide of ribonuclease A. *Nature*, **307**(5949), 329 – 334.

Kim, P. S., and Baldwin, R. L. (1990). Intermediates In The Folding Reactions Of Small Proteins. *Annual Review of Biochemistry*, **59**(1), 631–660.

Knaus, K. J., Morillas, M., Swietnicki, W., Malone, M., Surewicz, W. K., and Yee, V. C. (2001). Crystal structure of the human prion protein reveals a mechanism for oligomerization. *Nature Structural Biology*, **8**(9), 770 – 774.

Koon, N., Squire, C. J., and Baker, E. N. (2004). Crystal structure of LeuA from Mycobacterium tuberculosis, a key enzyme in leucine biosynthesis. *Proceedings of the National Academy of Sciences*, **101**(22), 8295 – 8300.

Laemmli, V.K., (1970). Determination of protein molecular weight in polyacrylamide gels. *Nature*, **227**, 680 – 685.

Lai C.S., Gerrelli D., Monaco A.P., Fisher S.E., Copp A.J. (2003). FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder. *Brain*. **126**(11), 2455 – 2462.

Lai, C. S. L., Fisher, S. E., Hurst, J. A., Vargha – Khadem, F. and Monaco, A. P. (2001) A forkhead – domain gene is mutated in a severe speech and language disorder. *Nature,* **413**, 519–523.

Lane, D., Prentki, P. and Chandler, M. (1992). Use of gel retardation to analyze protein – nucleic acid interactions. *Microbiological Reviews*, **56**(4), 509 – 528.

Lapidus, L. J. (2017). Protein unfolding mechanisms and their effects on folding experiments. *F1000 Research*, **6**, 1723 – 1742.

Le Coeur, C., Demé, B., and Longeville, S. (2009). Compression of random coils due to macromolecular crowding. *Physical Review E*, **79**(3), 1 – 4

Lecomte, J. T., and Matthews, C. R. (1993). Unraveling the mechanism of protein folding: new tricks for an old problem. *Protein Engineering, Design and Selection*, **6**(1), 1 – 10.

Lehmann, O.J., Sowden, J.C., Carlsson, P., Jordan, T., and Bhattacharya, S.S. (2003). Fox's in development and disease. *Trends in Genet*ics, **19***,* 339 – 344.

Lei, J., and Huang, K. (2012). Protein Folding: a Perspective from Statistical Physics. *Proteomics Research Journal*, **3**(1 / 2), 155 – 158.

Levínthal, C. (1969). Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, **65**, 44 – 45

Li, S., Weidenfeld, J. and Morrisey, E. E. (2004) Transcriptional and DNA binding activity of the Foxp1/2/4 family is modulated by heterotypic and homotypic protein interactions. *Mol. Cell. Biol.*, **24**, 809 – 22.

Lim, W. A., and Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of λrepresser. *Nature*, **339**(6219), 31 – 36.

Liu, S. Q., Ji, X. L., Tao, Y., Tan, D. Y., Zhang, K. Q., and Fu, Y. X. (2012). Protein folding, binding and energy landscape: A synthesis. *Protein engineering*. IntechOpen.

Longo, A., Guanga, G. P., and Rose, R. B. (2008). Crystal Structure of E47− NeuroD1/Beta2 bHLH Domain-DNA Complex: Heterodimer Selectivity and DNA Recognition. *Biochemistry*, **47**(1), 218 – 229.

Luheshi, L. M., Crowther, D. C., and Dobson, C. M. (2008). Protein misfolding and disease: from the test tube to the organism. *Current Opinion in Chemical Biology*, **12**(1), 25 – 31.

Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biology*, **1**(1), 1 – 24.

Lynch, M. (2005). Simple evolutionary pathways to complex proteins. *Protein science*, **14**(9), 2217 – 2225.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., Potter, S.C., Finn, R.D. and Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, **47**(W1), W636 – W641.

Mallamace, D., Fazio, E., Mallamace, F., and Corsaro, C. (2018). The role of hydrogen bonding in the folding/unfolding process of hydrated lysozyme: A review of recent NMR and FTIR results. *International Journal of Molecular Sciences*, **19**(12), 3825 – 3831.

Malleshappa Gowder, S., Chatterjee, J., Chaudhuri, T., and Paul, K. (2014). Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins. *The Scientific World Journal*, **2014,** 32 – 39.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**(5428), 751 – 753.

Marcovitz, A., and Levy, Y. (2013). Obstacles may facilitate and direct DNA search by proteins. *Biophysical Journal*, **104**(9), 2042 – 2050.

Massari, M. E., and Murre, C. (2000). Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Molecular and Cellular Biology*, **20**(2), 429 – 440.

Matsumoto, H., Haniu, H., and Komori, N. (2019). Determination of protein molecular weights on SDS-PAGE. *Electrophoretic Separation of Proteins,* **1**, 101 – 105.

Mattice, W. L., and Scheraga, H. A. (1984). Matrix formulation of the transition from a statistical coil to an intramolecular antiparallel β sheet. *Biopolymers: Original Research on Biomolecules,* **23**(9),1701 – 1724.

Medina, E., Córdova, C., Villalobos, P., Reyes, J., Komives, E.A., Ramírez-Sarmiento, C.A. and Babul, J., 2016. Three-dimensional domain swapping changes the folding mechanism of the forkhead domain of FoxP1. *Biophysical Journal*, **110**(11), 2349 – 2360.

Miller, J. S., Kennedy, R. J., and Kemp, D. S. (2002). Solubilized, spaced polyalanines: a context-free system for determining amino acid α-helix propensities. *Journal of the American Chemical Society*, **124**(6), 945 – 962.

Mirsky, A. E., and Pauling, L. (1936). On the structure of native, denatured, and coagulated proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **22**(7), 439 – 449.

Mok, Y. K., Bycroft, M., and de Prat-Gay, G. (1996). The dimeric DNA binding domain of the human papillomavirus E2 protein folds through a monomeric intermediate which cannot be native – like. *Nature Structural Biology*, **3**(8), 711 – 717.

Monera, O. D., Kay, C. M., and Hodges, R. S. (1994). Protein denaturation with guanidine hydrochloride or urea provides a different estimate of stability depending on the contributions of electrostatic interactions. *Protein Science*, **3**(11), 1984 – 1991.

Morjana, N. A., McKeone, B. J., and Gilbert, H. F. (1993). Guanidine hydrochloride stabilization of a partially unfolded intermediate during the reversible denaturation of protein disulfide isomerase. *Proceedings of the National Academy of Sciences*, **90**(6), 2107 – 2111.

Morris, E. R., and Searle, M. S. (2012). Overview of protein folding mechanisms: experimental and theoretical approaches to probing energy landscapes. *Current Protocols in Protein Science*, **68**(1), 28 – 32.

Morris, G., and Fanucchi, S. (2016). A key evolutionary mutation enhances DNA binding of the FOXP2 forkhead domain. *Biochemistry*, **55**(13), 1959 – 1967.

Murphy, K. P. and Freire, E. (1992) Thermodynamics of structural stability and cooperative folding behaviours in proteins. *Adv. Protein Chem.,* **43**, 313–361.

Nelson, C. S., Fuller, C. K., Fordyce, P. M., Greninger, A. L., Li, H. and De Risi, J. L. (2013). Microfluidic affinity and ChIP-seq analyses converge on a conserved FOXP2-binding motif in chimp and human, which enables the detection of evolutionarily novel targets. *Nucleic Acids Research*, **41**(12), 5991 – 6004.

Nicolau Jr, D. V., Paszek, E., Fulga, F., and Nicolau, D. V. (2014). Mapping hydrophobicity on the protein molecular surface at atom – level resolution. *PloS one*, **9**(12), e114042

Noy, A., Sutthibutpong, T., and Harris, S. A. (2016). Protein/DNA interactions in complex DNA topologies: expect the unexpected. *Biophysical Reviews*, **8**(1), 145 – 155.

O'Shea, E. K., Klemm, J. D., Kim, P. S., and Alber, T. (1991). X – ray structure of the GCN4 leucine zipper, a two – stranded, parallel coiled coil. *Science*, **254**(5031), 539 – 544.

Ohgushi, M., and Wada, A. (1983). 'Molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Letters*, **164**(1), 21 – 24.

Pace, C. N. (1986). Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods in Enzymology*, **131**, 266 – 280.

Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Shirley, B. A., Hendricks, M. M., Iimura, S., Gajiwala, K., Scholtz, J. M. and Grimsley, G. R. (2011). Contribution of hydrophobic interactions to protein stability. *Journal of molecular biology*, **408**(3), 514 – 528.

Pande, V. S., Grosberg, A. Y., and Tanaka, T. (1997). Statistical mechanics of simple models of protein folding and design. *Biophysical journal*, **73**(6), 3192 – 3210.

Papworth, C., Bauer, J.C., Braman, J. and Wright, D.A., (1996). Site-directed mutagenesis in one day with> 80% efficiency. *Strategies*, **9**(4). 38 – 40.

Parker, M. J., Dempsey, C. E., Hosszu, L. L., Waltho, J. P. and Clarke, A. R. (1998). Topology, sequence evolution and folding dynamics of an immunoglobulin domain. *Nature Structural and Molecular Biology*, **5**(3), 194 – 198.

Perumal, K., Dirr, H. W. and Fanucchi, S. (2015) A Single Amino Acid in the Hinge Loop Region of the FOXP Forkhead Domain is Significant for Dimerisation. *Protein J.,* **34**, 111–121.

Pitt-Rivers, R. and Impiombato, F. A. (1968). The binding of sodium dodecyl sulphate to various proteins. *Biochemical Journal*, **109**(5), 825 – 830.

Plaxco, K. W., and Gross, M. (2001). Unfolded, yes, but random? Never! *Nature Structural Biology*, **8**(8), 659–660.

Plotkin, S. S., and Onuchic, J. N. (2002). Understanding protein folding with energy landscape theory Part II: Quantitative aspects. *Quarterly Reviews of Biophysics*, **35**(3), 205 – 286.

Ponting, C. P., and Russell, R. R. (2002). The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure*, **31**(1), 45 – 71.

Privalov, P. L., Dragan, A. I. and Crane-Robinson, C. (2011). Interpreting protein/DNA interactions: Distinguishing specific from non-specific and electrostatic from non- electrostatic components. *Nucleic Acids Research*, **39**(7), 2483 – 2491.

Privalov, P. L., and Khechinashvili, N. N. (1974). A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *Journal of Molecular Biology*, **86**(3), 665 – 684.

Ptitsyn, O. B. (1987). Protein folding: hypotheses and experiments. *Journal of Protein Chemistry*, **6**(4), 273 – 293.

Pu, W. T., & Struhl, K. (1993). Dimerization of leucine zippers analyzed by random selection. *Nucleic Acids Research*, **21**(18), 4348 – 4355.

Renthal, R. (2008). Buried water molecules in helical transmembrane proteins. *Protein Science*, **17**(2), 293 – 298.

Rentzeperis, D., Jonsson, T., and Sauer, R. T. (1999). Acceleration of the refolding of Arc repressor by nucleic acids and other polyanions. *Nature Structural Biology*, **6**(6), 569 – 573.

Richards, F. M. (1997). Protein stability: still an unsolved problem. *Cellular and Molecular Life Sciences CMLS*, **53**(10), 790 – 798.

Richards, F. M., and Richmond, T. (1978). Solvents, interfaces and protein structure. *Ciba Found. Symp.*, **60**, 23 – 45.

Rose, G. D., Fleming, P. J., Banavar, J. R., and Maritan, A. (2006). A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences*, **103**(45), 16623 – 16633.

Rousseau, F., Schymkowitz, J. and Itzhaki, L.S. (2012). Implications of 3D domain swapping for protein folding, misfolding and function. *Protein Dimerisation and Oligomerization in Biology.* Springer, New York., 137 – 152

Rousseau, F., Schymkowitz, J.W. and Itzhaki, L.S. (2003). The unfolding story of three-dimensional domain swapping. *Structure*, **11**(3), 243 – 251.

Rousseau, F., Schymkowitz, J.W.H., Wilkinson, H.R. and Itzhaki, L.S. (2001). Three-dimensional domain swapping in p13suc1 occurs in the unfolded state and is controlled by conserved proline residues. *Proceedings of the National Academy of Sciences*, **98**(10), 5596 – 5601.

Russell, A. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?. *Journal of Structural Biology*, **134**, 191 – 203.

Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *Journal of Molecular Biology*, **279(**5), 1211 – 1227.

Sakaguchi, S., Miyara, M., Costantino, C. M. and Hafler, D. A. (2010) FOXP3 + regulatory T cells in the human immune system. *Nat. Rev. Immunol*., **10,** 490 – 500.

Savage, H. J., Elliott, C. J., Freeman, C. M., and Finney, J. L. (1993). Lost hydrogen bonds and buried surface area: rationalising stability in globular proteins. *Journal of the Chemical Society, Faraday Transactions*, **89**(15), 2609 – 2617.

Schägger, H. (2006). Tricine–SDS-PAGE. *Nature Protocols Electronic Edition*, **1**(1), 16 – 22.

Scheiner, S., and Hillenbrand, E. A. (1985). Modification of pK values caused by change in H-bond geometry. *Proceedings of the National Academy of Sciences*, **82**(9), 2741 – 2745.

Schellman, J. A. (1958). The factors affecting the stability of hydrogen – bonded polypeptide structures in solution. *The Journal of Physical Chemistry*, **62**(12), 1485 – 1494.

Scheraga, H. A., Vila, J. A., and Ripoll, D. R. (2002). Helix–coil transitions re – visited. *Biophysical Chemistry*, **101**, 255 – 265.

Schlunegger, M. P., Bennett, M. J. and Eisenberg, D. (1997). Oligomer formation by 3D domain swapping: a model for protein assembly and misassembly. *Advances in Protein Chemistry*, **50**, 61 – 122.

Schüler, A., and Bornberg-Bauer, E. (2016). Evolution of protein domain repeats in Metazoa. *Molecular Biology and Evolution*, **33**(12), 3170 – 3182.

Sengupta, D., Behera, R. N., Smith, J. C., and Ullmann, G. M. (2005). The α helix dipole: screened out?. *Structure*, **13**(6), 849 – 855.

Sharrocks, A. D. (2001). The ETS-domain transcription factor family. *Nature Reviews Molecular Cell Biology*, **2**(11), 827 – 837.

Shoemaker, K. R., Kim, P. S., Brems, D. N., Marqusee, S., York, E. J., Chaiken, I. M., Stewart, J. M., and Baldwin, R. L. (1985). Nature of the charged-group effect on the stability of the C-peptide helix. *PNAS*, **82**(8), 2349 – 2353

Shu, W., Lu, M. M., Zhang, Y., Tucker, P. W., Zhou, D. and Morrisey E. E. (2007). Foxp2 and Foxp1 cooperatively regulate lung and esophagus development. *Development.* **134**(10), 1991–2000

Sosnick, T. R., Mayne, L., and Englander, S. W. (1996). Molecular collapse: The rate-limiting step in two-state cytochrome c folding. *Proteins: Structure, Function, and Bioinformatics*, **24**(4), 413 – 426.

Sosnick, T. R., Mayne, L., Hiller, R., and Englander, S. W. (1994). The barriers in protein folding. *Nature Structural Biology*, **1**(3), 149 – 156.

Spiteri, E., Konopka, G., Coppola, G., Bomar, J., Oldham, M., Ou, J., Vernes, S.C., Fisher, S.E., Ren, B. and Geschwind, D.H. (2007). Identification of the transcriptional targets of FOXP2, a gene linked to speech and language, in developing human brain. *The American Journal of Human Genetics*, **81**(6), 1144 – 1157.

Stehle, T. and Harrison, S. C. (1996). Crystal structures of murine polyomavirus in complex with straight-chain and branched-chain sialyloligosaccharide receptor fragments. *Structure*, **4**(2), 183 – 194.

Stroud, J. C., Wu, Y., Bates, D. L., Han, A., Nowick, K., Paabo, S., Tong, H., and Chen, L., (2006). Structure of the forkhead domain of FOXP2 bound to DNA. *Structure*, **14***,* 159 − 166.

Susi, H., Timasheff, S. N., and Ard, J. S. (1964). Near infrared investigation of interamide hydrogen bonding in aqueous solution. *J. Biol. Chem*, **239**, 3051 – 3054.

Takano, K., Yamagata, Y., and Yutani, K. (2003). Buried water molecules contribute to the conformational stability of a protein. *Protein Engineering*, **16**(1), 5 – 9.

Tanford, C. (1962). Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *Journal of the American Chemical Society*, **84**(22), 4240 – 4247.

Temel, D. B., Landsman, P., and Brader, M. L. (2016). Orthogonal methods for characterizing the unfolding of therapeutic monoclonal antibodies: differential scanning calorimetry, isothermal chemical denaturation, and intrinsic fluorescence with concomitant static light scattering. *Methods in Enzymology*, **567**, 359 – 389.

Teufel, A., Wong, E.A. and Mukhopadhyay, M. (2003). FoxP4, a novel forkhead transcription factor. *Biochimica et Biophysica Acta.,* **1627**(2–3), 147 – 152.

Vaisse C., Kim J., Espinosa R. III, Le Beau M. M., Stoffel M. (1997). Pancreatic islet expression studies and polymorphic DNA markers in the genes encoding hepatocyte nuclear factor-3alpha, -3beta, -3gamma, -4gamma, and -6. *Diabetes,* **46***,* 1364 – 1367.

Vernes, S.C., Spiteri, E., Nicod, J., Groszer, M., Taylor, J.M., Davies, K.E., Geschwind, D.H. and Fisher, S.E. (2007). Highthroughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders. *The American Journal of Human Genetics,* **81***,* 1232 – 1250.

Vinogradov, S. N., and Linnell, R. H. (1971). *Hydrogen bonding*. Van Nostrand Reinhold Company, London, 32 – 43.

Vinson, C., Acharya, A., and Taparowsky, E. J. (2006). Deciphering B – ZIP transcription factor interactions in vitro and in vivo. *Biochimica et Biophysica Acta (BBA) – Gene Structure and Expression*, **1759**(1 – 2), 4 – 12.

Vitagliano, L., Adinolfi, S., Sica, F., Merlino, A., Zagari, A. and Mazzarella, L. (1999). A potential allosteric subsite generated by domain swapping in bovine seminal ribonuclease. *Journal of Molecular Biology*, **293**(3), 569 – 577.

Vitalis, A., and Caflisch, A. (2012). 50 Years of Lifson–Roig models: application to molecular simulation data. *Journal of Chemical Theory and Computation*, **8**(1), 363 – 373.

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, **14**(2), 208 – 216.

Wada, A., and Nakamura, H. (1981). Nature of the charge distribution in proteins. *Nature*, **293**(5835), 757 – 758.

Wallace, L. A., and Matthews, C. R. (2002). Sequential vs. parallel protein-folding mechanisms: experimental tests for complex folding reactions. *Biophysical Chemistry*, **101**, 113 – 131.

Walters, J., Milam, S. L., and Clark, A. C. (2009). Practical approaches to protein folding and assembly: spectroscopic strategies in thermodynamics and kinetics. *Methods in Enzymology*, **455**, 1 – 39.

Waudby, C. A., Dobson, C. M., and Christodoulou, J. (2019). Nature and regulation of protein folding on the ribosome. *Trends in Biochemical Sciences*, **44(**11), 914 – 926.

Weigel, D. and Jackie, H. (1990). The fork head domain: a novel DNA binding motif of eukaryotic transcription factors. *Cell.* **63**, 455 – 456.

Werner, M. H., Clore, G. M., Fisher, C. L., Fisher, R. J., Trinh, L., Shiloach, J., and Gronenborn, A. M. (1997). Correction of the NMR structure of the ETS1/DNA complex. *Journal of Biomolecular NMR*, **10**(4), 317 – 328.

Wetlaufer, D. B. (1990). Nucleation in protein folding - confusion of structure and process. *Trends in Biochemical Sciences*, **15**(11), 414 – 415.

Wong, K. K., Gascoyne, D. M., Soilleux, E. J., Lyne, L., Spearman, H., Roncador, G., Pedersen, L. M., Møller, M. B., Green, T. M. and Banham, A. H. (2016) FOXP2 – positive diffuse large B-cell lymphomas exhibit a poor response to R-CHOP therapy and distinct biological signatures. *Oncotarget*, **7**, 52940 – 52956.

Wu Y., Borde M., Heissmeyer V., Feuerer M., Lapan A. D., Stroud J. C., Bates D. L., Guo L., Han A., Ziegler S. F., Mathis D., Benoist C., Chen L., Rao A. (2006). FOXP3 controls regulatory T cell function through cooperation with NFAT. *Cell,* **126**(2), 375 – 387

Yakimov, A. P., Afanaseva, A. S., Khodorkovskiy, M. A., and Petukhov, M. G. (2016). Design of stable α – helical peptides and thermostable proteins in biotechnology and biomedicine. *Acta Naturae*, **8,**(4 (31)), 70 – 81.

Yeates, T. O., and Padilla, J. E. (2002). Designing supramolecular protein assemblies. *Current Opinion in Structural Biology*, **12**(4), 464 – 470.

Yeates, T. O., Komiya, H., Rees, D. C., Allen, J. P., and Feher, G. (1987). Structure of the reaction center from Rhodobacter sphaeroides R-26: membrane-protein interactions. *Proceedings of the National Academy of Sciences*, **84**(18), 6438 – 6442.

Zaret, K. S., and Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes and Development*, **25**(21), 2227 – 2241.

Zhang, F., Roth, R., Wolf, M., Roosen – Runge, F., Skoda, M.W., Jacobs, R.M., Stzucki, M. and Schreiber, F. (2012). Charge-controlled metastable liquid–liquid phase separation in protein solutions as a universal pathway towards crystallization. *Soft Matter*, **8**(5), 1313 – 1316.

Zhang, Z., Witham, S., and Alexov, E. (2011). On the role of electrostatics in protein–protein interactions. *Physical Biology*, **8**(3), 035001.

Zhou, R. (2015). Modeling of Nanotoxicity. *Springer*, 169 – 184.

Zimm, B. H., and Bragg, J. K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *The Journal of Chemical Physics*, **31**(2), 526 – 535.