NMR studies of protein interactions with nucleic acids : translation termination and transcription

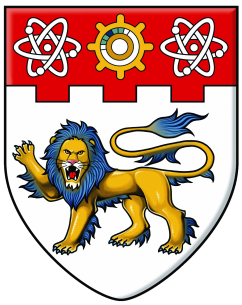
Shubhadra Pillay

2012

Shubhadra Pillay. (2012). NMR studies of protein interactions with nucleic acids : translation termination and transcription. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/54956

https://doi.org/10.32657/10356/54956



NANYANG TECHNOLOGICAL UNIVERSITY

NMR Studies of Protein Interactions with Nucleic Acids: Translation Termination and Transcription

SHUBHADRA PILLAY

SCHOOL OF BIOLOGICAL SCIENCES

2012

NMR Studies of Protein Interactions with Nucleic Acids: Translation Termination and Transcription

SHUBHADRA PILLAY

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University in partial fulfillment of the requirement for the degree of Doctor of Philosophy

2012

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Associate Professor Konstantin Pervushin, my research supervisor, whose encouragement, stimulating suggestions and support thus far has enabled me to develop an understanding of the subject. I would also like to thank Dr. Ralf Jauch and Kamesh Narasimhan, our collaborators from GIS, Biopolis Singapore for their insightful discussions and assistance in keeping my progress on schedule. My grateful thanks are also extended to our collaborators here in SBS, Professor Lars Nordenskiold, Dr. Nikolay Korolev and Dr. Abdollah Allahverdi for their support and help in the course of the NCP project. I am deeply grateful to have worked and had the opportunity of learning from our collaborators. My colleagues have not only supported me, but also helped when in time of difficulty with their patient guidance, enthusiastic encouragement and useful critiques of this research work.

I thank almighty, my parents, brothers, and friends for their constant encouragement without which this thesis would not be possible. The blessing, help and guidance given by them time and again carried me a long way in this journey of which I am about to complete.

Lastly, I offer my deepest regards to all of those who supported me in any respect during the completion of this thesis. I finally thank the School of Biological Sciences, NTU for my research persuasion and NTU research Scholarship.

"...Keep me away from the wisdom which does not cry, the philosophy which does not laugh and the greatness which does not bow before children."

~Khalil Gibran~

CONTENTS

List of Figures

CHAPTER I Introduction

CHAPTER II Translation Termination: NMR studies of human peptide release factor eRF1

CHAPTER III SRY-Related High Mobility Group (SOX-HMG)

List of Tables

Abbreviations

CSP	Chemical shift perturbations
RDC	Residual dipolar couplings
HSQC	Heteronuclear single quantum correlation
TROSY	Transverse relaxation optimized spectroscopy
NOE	Nuclear Overhauser effect
NOESY	Nuclear Overhauser effect spectroscopy
NMR	Nuclear magnetic resonance
PDB	Protein data bank
ppm	parts per million
RMSD	Root mean square deviation
TOCSY	Total correlation spectroscopy
CD	Circular Dichroism
DTT	Dithiothreitol
E.coli	*Escherichia coli*
IPTG	Isopropyl β-D-1-tiogalactopyranoside
Pf1	*Pseudomonas aeruginosa*
SDS	Sodium dodecyl sulfate polyacrylamide
Tris	Tris(hydroxymethyl)aminomethane
KCl	Potassium chloride
kDa	kilo Dalton
mM	millimolar
MES	2-(N-morpholino)ethanesulfonic acid
OD	Optical density
H/D	Hydrogen/deuterium
NTD	N-terminal domain
NCP	Nucleosome core particle
eRF	Eukaryotic release factor
POM	Polyoxometalate
DMSO	Dimethyl sulfoxide
TFA	Trifluoro acetic acid

Abstract

My thesis includes three subprojects related to NMR-based studies of three protein classes involved in DNA/RNA recognition. My focus is to use the advanced solution NMR methods to answer critical biological questions complementing collaborative efforts within each individual subproject.

In the first subproject I studied structural changes in a stop-codon recognizing protein induced by mutations altering stop-codon recognition specificity. Ribosome is one of the most sophisticated supramolecular complexes in the cell. Some of the processes that occur within this machinery include translation initiation, elongation, and termination, all of which involve coding mRNA, a set of aminoacyl-tRNAs, and various protein factors acting in a concerted manner. Translation termination requires class I release factors: RF1 and RF2 in Bacteria, and eRF1 in Eukarya. eRF1 is able to decode all three stop codons. Until now, the mechanism of stop codon recognition by eRF1 leading to hydrolysis of the ester bond on the peptidyl tRNA remains obscure. We working as a team at Prof. Pervushin's lab (Li Yan and Leo Wong) proposed a model of the binding mode of eRF1 to pre-translation termination ribosomal complex based on our solution structures of wild-type and several mutants of eRF1's N-domain as well as its interactions with model RNA and M- and C- domains of the same protein. Due to the global character of structural perturbation induced by mutations, Residual Dipolar Coupling (RDC) data were obtained using ^{15}N labeled sample aligned with *Pf1* phages. The RDC measurements obtained were then applied in the structure calculation and refinement of the eRF1 mutant $Q^{122}FM(Y)F^{126}$ using structure calculation software. The most interesting mutant, $Q^{122}FM(Y)F^{126}$ restricts the decoding capability of eRF1 to UGA codon only. From the 3D structures, we established that the mutations alter conformation and dynamics of the GTS loop distant from the sites of mutations. Based on published biochemical and mutagenesis studies, we propose that the GTS loop forms a switch that allows reading of the multiple codons. NMR analysis revealed that helix $\alpha 1$ of N-domain interacts specifically with double-stranded RNA with a bulge or internal loop resembling a mismatch of H44 in 18S rRNA of Eukaryotic ribosome. From these results, a 3D model of eRF1 interactions with ribosome in pre-termination state is proposed.

In my second project we collaborated with Dr. Ralf Jauch in order to provide structural basis for activity of novel drug candidates interfering with transcription factor/cognate DNA interactions. The DNA binding domains of transcription factors

have so far been considered too impervious to be tackled as drug targets although up-regulated transcription factors are a major cause of cancer and other diseases. Here we identified a Dawson-POM as an unconventional but potent compound to inhibit the DNA binding activity of Sox2. We used NMR to locate binding site of the drug candidates on Sox2. The mode of interaction of the Dawson-POM with the Sox2-HMG domain involves predominantly electrostatic interactions at the pocket just outside of the DNA binding region, but still adequately positioned to compete with the negatively charged DNA backbone. In summary, the inhibitory mechanism demonstrated here could eventually spawn the development of modified classes of POM based drugs to specifically combat aberrant gene expression.

The objective of the third project is to assess the structural flexibility of N-terminal tail domain of Histone 4 when it is in the compact nucleosome array using deuterium/protons exchange experiment in NCP. ^{15}N labeled Histone 4 was prepared and 2D [^{1}H-^{15}N]-TROSY series of experiments were carried out for protein recovered from reconstituted NCPs in H_2O followed by controlled exposure to D_2O. The sequential assignment of the backbone H4 was used to identify the dynamics of the NCP as well as arrays in the compact form. We observed that that although in compact states, the magnitude of protection is not as pronounced as expected. This suggests that the NCP is still flexible, albeit being in a compact state.

The first two projects are either prepared or published. The last project is in active development.

CHAPTER I

CHAPTER I

Protein – DNA Interaction Studies

1.1 Protein-DNA Interactions

Genes are controlled within an organism through different course of events. It could be either during cell development and growth, or induced by external and/or environmental factors, or perhaps through hundreds of routine activities that cells undergo. In order to understand this mechanism of control, we need to study the DNA-binding proteins that are involved if not manipulate these interactions [1, 2]. These mechanisms are regulated via co-interacting partner proteins, DNA modifications, protein modifications and presence of small molecule binders. In light of this, the structure of proteins or nucleic acids have always been the focal point of studies, as well as the realization of importance on understanding mechanisms of molecular assemblies, and complex structures [3]. Such realizations are the main determinants towards the development of new technologies in this field.

1.2 Techniques to investigate protein-DNA interactions

Various techniques have been developed over the years to study protein –DNA interactions in many different aspects. These techniques function to address the questions that are posed by scientist to further understand protein – DNA interactions; questions pertaining to strength of binding, surface of interaction and factors required for an interaction to occur, and environmental components that would make or break the physical complex. To name a few of the common techniques used are: filter binding [4], gel mobility shift assay [5], foot-printing [6], and spectroscopic techniques (please refer to this review for detailed information [7, 8]).

Filter binding is a rather fast and reproducible method that uses the concept of retention based on complex formed. A nitrocellulose filter is used in different environments to test its effect on complex formation and retention [4]. Almost all studies involving protein – DNA interactions are begun with the gel mobility shift assay or Electrophoresis

Mobility Shift Assay (EMSA) [5]. It is a powerful technique that utilizes the physical properties of the bound complex versus the non-bound components [5], in which the results are based on the outcome of either the positions and/or the amounts of DNA of interest observed in the bands.

In DNA foot printing, the complexes undergo chemical modifications. The cleavage pattern of the complex is compared to that of the unbound DNA (sometimes labeled) and this is then used to decipher the sequence involved in binding or to some extent the residues that are within the vicinity of the binding domain. There are two broad categories for this technique: protection [9] and interference assays [10]. Another class of techniques that is commonly applied in studying protein – DNA interactions is spectroscopy. In most cases, the conformation of the molecule inferred from difference in absorption of fluorescence is used to follow a reaction and its changes. The effect of Circular Dichroism (CD) is used to determine difference in specific as well as non-specific binding in such complexes [7, 11]. This technique is suitable because nucleic acids contribute marginally to the signals in the far UV region, where peptide bonds of proteins are dominant. In addition to this, proteins hardly affect the near UV region, where signals are predominantly from nucleic acids. Therefore, changes incurred by both the protein and nucleic acids can be observed upon complex formation [12]. However, structural detail and analysis at a molecular level is plausible only with spectroscopic techniques such as X-ray crystallography and Nuclear Magnetic Resonance (NMR). With regards to the theme of the methodology employed in this thesis, specific techniques that are used in NMR will be discussed in the following sections.

1.3 NMR-based methods in investigating protein-DNA interactions

One of the major contributions of NMR in studying protein nucleic-acid interactions is its ability to look at the dynamics of such interactions, where changes to structural order could be monitored upon complex formation and vice versa [2]. Information of these specific and non-specific complexes have been successfully attained due to development of techniques in the areas of isotope labeling for both proteins and nucleic acids [2, 13], and more robust synthesis methods. One of the first complexes solved by NMR was the *Drosophila antennapedia* mutant homeo domain in complex with a 14-mer DNA [14].

The sizes of these complexes are rather large and the advent of transverse relaxation optimized spectroscopy (TROSY) [15]-based techniques have made studying these complexes feasible. Labeled proteins allow the use of 2- and 3-dimensional experiments possible to facilitate resonance assignment of complexes, obtaining relaxation data, and extraction of structural data pertaining to interaction surfaces and dynamics of interactions. The techniques involved are chemical shift mapping [16, 17], measurement of hydrogen exchange rates [18], isotope editing [19], and more recently Residual Dipolar Coupling (RDC) [20, 21] and paramagnetic spin labeling (PRE) [2, 3].

1.3.1 Chemical Shift (CS) Mapping

Chemical shifts are sensitive to the local environment of the nuclei of interest; they are very useful in identifying not only the points of interaction but also other possible changes that occur in the protein or DNA. However, this information should be used with caution because as mentioned above, not all residues that are perturbed are involved in interaction [22]. This technique is used to determine the interaction sites or surface between two entities, i.e., proteins and nucleic acids [16]. When the interaction of the protein and DNA is specific, tight binding is usually observed and this leads to an exchange rate that is slow and thus observable in the NMR time scale, which would cause two different peaks corresponding to the two different states, bound and free [13]. Two methods of analysis are available in CS mapping, whereby one could analyze the interaction from the perspective of the protein or the nucleic acid [3]. Mapping changes that occur to the proteins is a preferred choice in earlier stages of the study simply because DNA chemical changes are usually more difficult to observe [13].

Variations in the CS is monitored by adding increasing amounts of DNA to the protein from which the binding constant could be extracted through analyzing the titration curves. Intermediate to slow exchange is by far more challenging to study as the consequence of different bound and unbound states results in line broadening, which makes analysis rather obscure.

1.3.2 Hydrogen exchange rates

The amide protons in a protein exchanges with the hydrogen present in the solvent through a process known as hydrogen exchange. This exchange is advantageous in characterizing proteins as it is applicable in different conditions/environments, can be detected for specific residues of interest and most importantly, it can be used to study conformational changes, dynamics and stability [23]. The exchange that occurs between the amide protons and the hydrogen in the solvent occurs at a characteristic rate that is catalyzed by H^+ and OH^-, is dependent on environmental factors namely, temperature, pH, hydrogen isotope present, steric blocking and inductive effects of adjacent side chains [24]. Studies have shown that in a structured vicinity of a protein, the NH exchange rate is lesser in comparison to an unstructured region, and this lesser exchange phenomena is reported as the protection factor [23]. Apart from looking at the protein by itself, these H/D exchange studies could also be performed on protein complexes followed by separation of the individual components prior to detection and analysis.

Hydrogen exchange rates observed for amide protons when in complex compared to free state can be used to elucidate structural information concerning dynamics of the protein-DNA complex and this information could also be used to infer possible binding sites [18]. The technique evaluates the protection experienced by the amide proton of interest (H_2O is exchanged for D_2O). The exchange experienced by amide protons conveys the position of the residues on the protein – whether on the surface or buried within the protein core [18]. Changes in the hydrogen exchange data are indicative of possible interaction surfaces due to the blocking of the surface (amide proton) by the DNA in complex formation. An observation of the reverse-exposure of a before buried amide proton could be indicative of conformational changes [3, 25]. These experiments could also be used quantitatively in determining structural stability of a complex [18].

1.3.3 Residual Dipolar Coupling (RDC)

Residue specific resonance assignment is a prerequisite for NMR-based structure reconstruction. The backbone of the protein is first assigned by correlating the amide group (N_H, H_N) of a particular spin system with the intra and inter residue cross-peaks of the backbone and also the side chain atoms [26]. Upon completion of the sequence specific

assignment and side chain assignments, structure calculation is done to obtain the 3-dimensional structure of the protein [27]. The necessary structural restraints are extracted from NMR parameters. Among the restraints required for structure calculation are distance restraints from nuclear Overhauser effects (NOEs), hydrogen bond restraints that can be obtained from hydrogen exchange experiments of labile protons, dihedral angle restraints that are derived from scalar couplings, chemical shifts, and global restraints from Residual Dipolar Coupling (RDCs) [28-30]. In determining the structure of larger macromolecules, the key NMR data used are the NOE and scalar J couplings, which are predominantly local information. The NOEs dependence on the r^{-6} limits it to the distance of 5 Å between two atoms, the accuracy of which usually decreases from the real value due to the decreased precision of measurement with increased distance (weaker NOE) [31]. These global restraints aid in understanding the global nature of the protein as well as identify inter domain interaction, that are present within a particular macromolecule [32]. A change in the traditional method of NOE-constraints used in determining structures have been suggested, where backbone information complemented by RDCs is used instead to compute structures for a wide range of proteins [33].

It is known that a magnetic nucleus is capable of creating a local magnetic field around its neighboring nuclei. When these nuclei are subjected to an external magnetic field, the local field becomes modified such that the original Larmor frequency is changed. This effect that is termed dipole-dipole interaction is also commonly known as dipolar coupling. The magnitude of the interaction experienced by the nuclei is dependent on the distance between the interacting nuclei and the angle between the internuclear vector and magnetic field [3, 34].

RDCs provide both short range and long range structural information, which can be incorporated in the molecular dynamics calculations [35]. From an RDC, it is possible to determine the distance as well as the angles formed by a vector of two atoms in a tensor axis system. Due to this dependence on distance, orientation and dynamics, it is therefore an essential restraint in characterizing structure. In order to obtain RDCs that can be used in molecular dynamics calculations, the resulting alignment observed should not cause severe coupling interactions or distort the spectra (this is usually caused by interaction of the

alignment media with the macromolecule). Therefore it is essential to introduce an alignment media that only weakly interacts with the protein of interest [20].

i. Measurement of Residual Dipolar Coupling

The measured dipolar coupling is dependent on the gyromagnetic ratio of the coupling spins of interest, in addition to the distance between them. For a particular aligned state (depending on the media used), the coupling observed between H^{α} and C^{α} is double that of NH; C' and N is 1/10 that of NH [20, 28]. Based on these observations, the couplings measured in this study would mainly focus on the N-H bonds, which have a relatively large gyromagnetic ratio. For reviews on measurements of residual dipolar coupling, please refer to [36-38].

In an isotropic (alignment media independent) solution, the scalar J_{NH} is measured for each amide group. The experiment is then repeated in the presence of an alignment media, whereby the sum of the scalar and dipolar coupling: $J_{NH} + D_{NH}$ is measured [21]. Based on the spectra obtained (results and discussion), the coupling is measured from the peak separation between the peaks in the TROSY and anti-TROSY spectra. **Figure 1.1** illustrates the peak separation observed resulting from the J_{NH} scalar coupling of a magnitude of ~92 Hz. Increasing concentration of alignment media (*Pf1*), causes the peak separation, $J_{NH} + D_{NH}$, to change.

When the degree of alignment increases (higher concentration of alignment media), the ^{1}H-^{1}H dipolar coupling would also increase, thus becoming comparable to the $3J_{HH}$ scalar couplings. This would in turn affect the line-width of the ^{1}H dimension and result in reduced sensitivity and increased overlap in the spectrum. Due to the above mentioned phenomena, it is preferable to measure RDCs from the splitting in the ^{15}N dimension compared to the ^{1}H dimension [26].

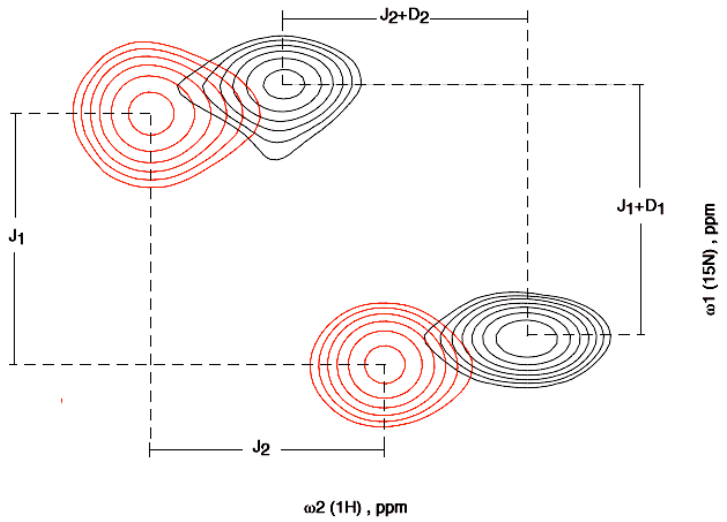


Figure 1.1. TROSY and anti-TROSY spectra of wildtype-eRF1 acquired at 700MHz in *Pf1* alignment media. Two well resolved peaks are observed for each experiment. Residual dipolar coupling is then measured by taking the difference between the TROSY and anti-TROSY peaks of both isotropic conditions (J1 and J2, red peaks) and anisotropic conditions (J1+D1 and J2+D2, black peaks) with J corresponding to the scalar coupling and D corresponding to dipolar coupling.

ii. Alignment media

Proteins are generally ordered by the introduction of larger oriented macromolecules into the sample, as they are not sufficiently anisotropic, thus unable to produce molecular alignment. Only the proteins near the surface of the orientating molecule would become ordered and to average this ordering effect over the entire ensemble of proteins, it is necessary that the rate of translational diffusion is sufficiently rapid in order to distribute the anisotropy over all the molecules [34, 39].

There are diverse types of alignment media available that can provide this small ordering required for NMR but at the same time maintain the macromolecule of interest, each of which with its own unique characteristic [28, 40]. Alignment is adjusted by varying the concentration of the media in order to obtain couplings between closely spaced nuclei. The choice of alignment media to be used critically depends on its inert effects to the protein of interest, apart from being stable in a range of pH and temperature. In addition to stability, the media should not interact in a destructive manner so as to reduce the resolution of the spectra or increase line widths [41]. With these criteria in mind, several alignment media were tried and tested based on the property of the protein of interest, namely filamentous bacteriophages *Pf1* [42], Poly(ethylene glycol)/alcohol mixtures (commonly known as Otting media) [43], and strained polyacrylamide gels [44]. It is good practice to

use more than one alignment media when obtaining RDCs. This is done to prevent the averaging of the induced alignment over the ensemble, should the protein under study interact with the alignment media.

a. Filamentous Bacteriophages, Pf1

Pf1 are filamentous bacteriophages, consisting of single stranded circular DNA, 60 Å in diameter and ca. 20,000 Å long [45]. The coat protein has α-helical structure that is parallel to the axis of the phage, which is believed to be the main source of the observed anisotropic susceptibility. The versatility of this bacteriophages is what makes it so widely used; stable over a range of concentrations, pH, temperatures and salt concentrations, in addition to being fully aligned at all magnetic field strengths [26, 40, 42].

The magnetic alignment of *Pf1* is determined by the splitting of the D_2O signal, which is due to the large deuterium quadrupole moment that is not isotropically averaged for the water that is bound to the aligned phage particles [42].

b. Strained Polyacrylamide Gels (SAG)

Compressed gel is another method used in achieving molecular alignment. The anisotropic effect is achieved from either compression or stretching (radial compression) of polyacrylamide gels, while maintaining a stable and inert environment for the macromolecules of interest. Often enough, conditions of partial or completely denatured proteins can also be studied. The method adapted in this study is based on stretching (radial compression), omitting the dehydration step. This allowed for the use of lower gel concentrations with lesser effects on the rotational diffusion of the solute [46]. The advantage of this method is in its ability to function over a wide range of temperature, pH and ionic strength, apart from being compatible with most biopolymers. Alignment can be adjusted based on the strain induced, the gel density used, and gel composition. The most favorable feature would be the sample recoverability. The molecular alignment measured is determined by the direction of the induced strain and not the magnetic field [44].

c. Ether/alcohol liquid crystalline phases (Otting Media)

Another medium that is very common due to its versatile nature is the liquid crystalline phase formed from n-Alkyl-poly(ethyleneglycol)/glucopone and n-alkyl alcohol mixtures [43]. This media is considered versatile, as it is mostly insensitive to pH and ionic strength. It induces alignment in the magnetic field rather fast, with narrow linewidth even at high degree of orientation. The alignment is achieved by dissolving the ether to a 5 % w/w in aqueous buffer (10 % D_2O), and adding alcohol in microlitre steps followed by vigorous vortexing. The addition of alcohol is ceased when the biphasic solution becomes transparent and opalescent upon formation of the liquid crystalline phase. Its crucial to avoid using excessive amounts of alcohol as this would decrease the temperature stability of the liquid crystalline phase [3].

CHAPTER II

CHAPTER II
Translation Termination: NMR studies of human peptide release factor eRF1

2.1 Introduction

2.1.1 Protein synthesis

DNA is made up of nucleic acids, which represents an organism's genome. The genetic makeup of each organism is then ensured to be consistent within its species via replication of the genetic information in the DNA, also known as genes that codes for one protein. Gene expression occurs when its genetic information is transferred to the messenger RNA (mRNA) via opening of the DNA double helix via a zipper-like motion to enable the mRNA to assemble on a single strand [47, 48]. The mRNA then receives the genetic information through template complementation; process termed as transcription, which is followed by the diffusion of the mRNA into the cell where it then attaches itself to the ribosome to begin the translation process. The genetic information in the mRNA is translated into specific sequences of amino acids leading to the formation of a polypeptide chain, the success of which is determined by the correct selection of tRNA's [49, 50].

2.1.2 Overview of translation

Eukaryotic protein synthesis is characterized by four stages; the initiation, followed by elongation, termination that releases the nascent peptide, and finally the recycling of the ribosomal subunits to be used in the following round of initiation [51]. The initiation step fundamentally consists of formation of the ternary complex – methionyl initiator transfer RNA (tRNA), eukaryotic initiation factor 2 (eIF2) and GTP- which then binds to the smaller subunit [52]. Binding to the mRNA occurs via scanning in the 5" to 3" direction to locate the initiation codon. A single amino acid specific to the tRNA would be covalently bound to it, and the 3-nucleotide sequence (anticodon) tRNA should match the 3-nucleotide sequence (codon) of the mRNA. Elongation is when the aminoacyl tRNA enters the ribosomal A site for decoding to ensue, thus allowing the catalysis of a peptide bond by the ribosome [49]. The whole process is repeated when the tRNAs and mRNA are translocated to allow the next codon to move into the ribosomal A site. Termination would then occur when the stop codon is present in the ribosomal A site. The nascent peptide is released

following hydrolysis of the ester bond at the ribosomal P site. Re-initiation is induced when the ribosomal subunits release the mRNA and deacylated tRNA [53].

2.1.3 Translation termination

The genetic code comprises of sense codons as well as stop codons. Eukaryotes have three termination codons (commonly known as stop codons) UAA, UAG and UGA [54, 55]. Translation termination occurs when a single stop codon present at the A-site is recognized, resulting in a chain of events whereby the peptidyl-tRNA in the P-site undergoes hydrolysis causing the release of the nascent polypeptide. Release factor proteins mediate the recognition of stop codon, deemed as the highly pertinent step in termination. In the case of eukaryotes, they are known as eukaryotic release factor 1 and 3 (eRF1, eRF3). Unlike prokaryotes, eukaryotes have an omnipotent release factor protein, which enables it to recognize not singular or specific stop codons but rather all 3 stop codons [56, 57]. The efficiency of this process is dependent on the formation of a quaternary complex that involves the ribosome, the two former release factors and GTP [58, 59]. Although extensive work has been done to establish the principal mechanism of stop codon recognition in eukaryotes, it still eludes us. Chavatte *et al.* have shown via crosslinking experiments that eRF1 dictates specificity for stop codon recognition [52] and is supported by various other experimental results, particularly mutagenesis analysis, in establishing the region of recognition on eRF1 [60, 61]. The second release factor, eRF3 consist of a GTP-binding domain at its C-terminal, domain 2 and 3 (beta-barrels) and the less conserved N-terminal domain.

The elucidation of the crystal structure of human eRF1 illustrated the presence of 3 domains [62] of which have been extensively studied over the past 12 years (**Fig. 2.1**). The 3 domains identified were ascertained to having precise roles in the translation termination process. Domain 3 (C terminal end of eRF1) constitutes the binding region that corresponds to eRF3 [56, 63-67]. Studies have shown that domain 3 is not involved in ribosome binding whereby removal of this domain enhances instead of diminishes the termination activity *in vitro* [52]. Apart from this observation, sequence analysis indicates that in comparison to the other 2 domains (N- and M-domain), C domain has evolved the most and would therefore be most unlikely to be involved in binding with a highly conserved ribosome [68].

The NMR structure of this domain is shown to be represented by 3 anti-parallel β-strands and one α-helix [69]. Domain 2 (constitutes the middle of eRF1, M-domain) that mimics the tRNA acceptor stem with its highly conserved GGQ motif is essential for release activity and ribosome binding via peptidyltransferase [62, 70, 71]. It has been proposed that the C domain increases the binding affinity between class I and class II release factors (eRF1 and eRF3) in the presence of GTP [63, 72], and this additional affinity could be caused by the M domain interacting with the switch regions of eRF3's GTPase [67].

Domain 1 (N terminal end of eRF1) is believed to be the decoding site of the release factor based on mutagenesis [52, 60] and sequence analysis from various organisms (**Appendix A1**) [68]. This domain has the highly conserved NIKS motif (positions 61–64), YxCxxxF motif (positions 125–131) [73] and the GTS loop (positions 31-33), which have all been implicated in stop-codon recognition [61, 67, 74, 75]. As previously mentioned, both eRF1 and eRF3 interact via their respective C-terminals in the presence of the eRF1 M-domain, which in fact enhances the GTP binding to eRF3. This stable ternary complex comprising eRF1-eRF3-GTP shows that both the GTPase activity of eRF3 and stop-codon recognition of eRF1 are inter-dependent [76].

Therefore, the current model of translation termination is as follows [49, 77]: binding of the eRF1-eRF3-GTP ternary complex to the pre-termination complex (pre-TC) causes conformational changes in ribosome. These changes lead to a 2-nt forward ribosomal shift. The hydrolysis of GTP in the ternary complex causes peptide release and subsequent conformational changes in eRF1. The M-domain of eRF1 would then be able to enter PTC in the 60S ribosome via its GGQ-loop resulting in peptidyl-tRNA hydrolysis [77].

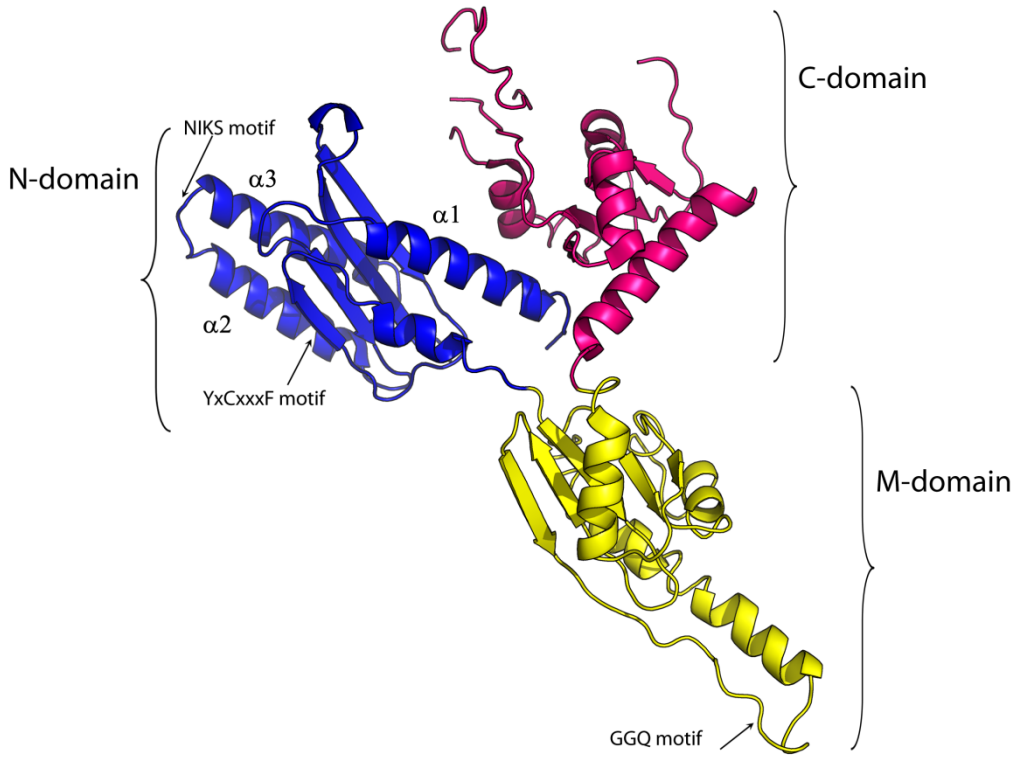


Figure 2.1. Cartoon representation of eRF1 derived from PDB-1DT9 [62]. NIKS, YxCxxxF and GGQ motifs are labeled.

2.1.4 Stop codon recognition

Several models have been proposed to explain the mechanism of stop codon recognition in the N-domain of eRF1. One of the foremost hypotheses concerning the method of binding is the anti-codon mimicry model, which was first postulated for prokaryotic systems [54, 56]. Apart from the mimicry model, a favorite among ardent fans of the translation termination mystery is the cavity binding model, whereby they postulate that nucleotides of the stop codon are accommodated into three defined cavities on the surface of N-domain [60, 68], thus causing conformational changes induced onto the protein through binding with the stop codon at these various sites [78]. The non-linear model is derived from the TASNIKS and YxCxxxF motifs in N-domain of eRF1; the NIKS motif observed to be positioned in proximity to the first uridine, and the YxCxxxF motif in proximity to the

purines in the second and third stop codon positions [79, 80]. Both models are illustrated in **Figure 2.2.**

Although both the non-linear and cavity models share some residues implicated in stop codon recognition, they are not entirely compatible, thus requiring further experiments to resolve this ambiguity. From these efforts, different mutants of human eRF1 with substitutions in N-domain were studied; specifically eRF1 mutant of Stylonychia (ciliate) with UGA-only specificity, $Q^{122}FM(Y)F^{126}$ (T122Q, S123F, L124M and L126F) was shown to exhibit strong UGA unipotency [81] and invariant amino acid E55 and Y125 affected the release factor ability specifically the UAG response and thus may be essential for maintenance of spatial protein structure [74]. Rationale for selection of these specific residues for mutation is discussed in Section 2.1.5. Although the solution structures of the mutant N-domain in comparison to the wild-type (wt) eRF1 showed similar global structure, conformational differences were observed in the mutants specifically in the GTS loop (positions 31-33), which is remote from the mutation sites. We postulate that switching between omnipotency and unipotency of eRF1 may be modulated by distinct conformations of the GTS loop.

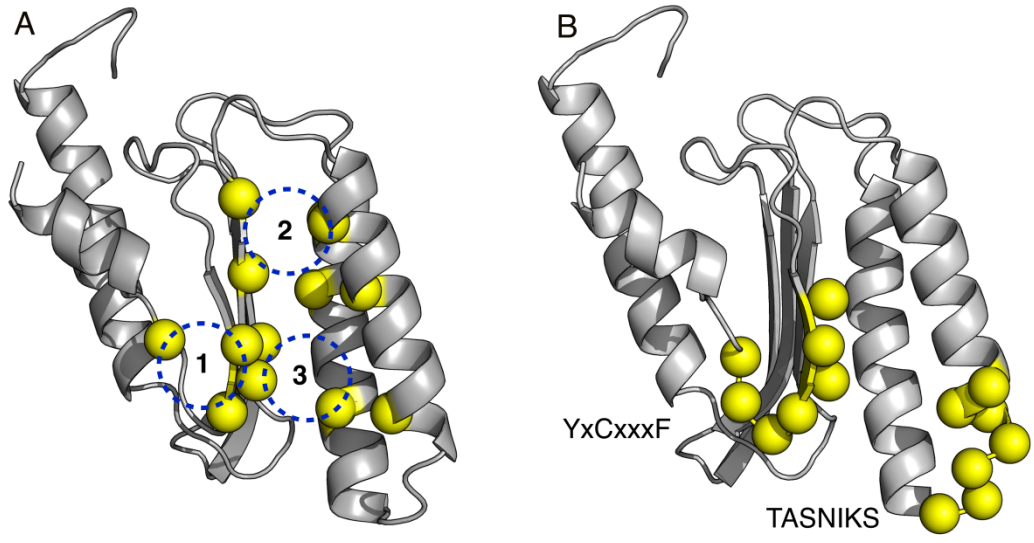


Figure 2.2. Proposed models of stop codon recognition in N-domain of eRF1. N-domain is derived from PDB: 1DT9 [62] A) Cavity binding model, potential stop codon binding pockets are labeled 1, 2, and 3 B) non-linear binding model.

2.1.5 Selection of amino acid positions for mutagenesis

Table 2.1. *In vitro* release activity of N-domain eRF1 mutants (in percent) to the wildtype eRF1

Mutant/codon	UAA	UAG	UGA
F131A[i]	10	7	62
E55Q[ii]	75	35	80
Y125F[ii]	100	34	100
T112Q+S123F+L124M+L126F[i]	0	0	80

i The release activity of mutant eRF1s was measured according to the *in vitro* Caskey's assay. Data referenced from [74].
ii Activity of mutants of N-domain in percent to the human wild type eRF1 (Kisselev, private communication)

In prokaryotes, the recognition of stop codons is unipotent and based on linear motifs, namely PAT in RF1 and SPF in RF2 [82, 83]. In eukaryotes, the case is different altogether as recognition is omnipotent and occurs specifically on eRF1 at the N-domain. The perplexing issue faced thus far is the nature and positions of amino acids that are involved in this recognition especially since eRF1 is a highly conserved family.

Suggested models for the decoding site of eRF1 involve several residues that are conserved in both eRF1 and aRF1, namely E55 and Y125. The conserved nature of these amino acids throughout eRF1 and aRF1protein family suggests an essential structural and functional role. In addition, the location of these residues on the surface of the molecule presents possible sites of ligand interactions, e.g. mRNA [74]. Position E55 and Y125 have been shown to participate in the recognition of stop codon based on indirect evidence [60, 68] and direct evidence [74]. Kolosov *et. al.* introduced point mutations at these invariant positions and examined *in vitro* their functional as well as physical properties. Y125 was shown as essential in maintaining the structure of the decoding site via its interactions with E55, with both residues implicated in recognition of G of UAG codon. This selectivity however was not due to the binding of the mutant proteins to the ribosome, as GTPase activity of eRF3 in the binding assay was not affected. Apparently the hydrogen bond that is formed between these two residues is essential for the UAG-dependent RF activity. **Table 2.1** summarizes the release factor activity when these mutations were introduced into the protein and **Figure 2.3** illustrates the positions of these mutations on the 3-dimensional structure of N-domain eRF1.

Chimeric molecules of the N-terminal *Stylonychia* eRF1 with the M and C domains of human eRF1 were constructed to prove that stop codon recognition lies within the N-domain, and it was observed to retain UGA specificity. Seit-Nebi *et al.* did a study of chimeras with swapped N-domain sequences between ciliate and human protein. Functional analysis of these chimeras highlighted importance of the QFM tripeptide in this UGA specificity, whereas chimeras of *Paramecium* eRF1 showed UGA specificity to be confined to the NIKS and YxCxxxF motifs instead [81]. They focused on the YxCxxxF motif in N-domain as it comprises of semi-conserved and invariant amino acid residues located near the famed NIKS motif [61], and also because certain residues namely Y125, C127 and F131 are the slowest evolving in N-domain [68]. All these residues when mutated were shown to affect the RF activity of eRF1 *in vitro*. It is suggested that both the YxCxxxF and NIKS motif are involved in stop codon recognition with the invariant amino acids playing a crucial role in decoding, however not in peptidyl-tRNA hydrolysis [73].

A few groups have addressed the relevance of the GTS loop in stop codon recognition. The first to report these residues to be involved in recognition were Liang *et al*., where sequence and computational analysis of conserved regions on the protein were studied based on 3 different factors i) conservation, ii) accessibility to other macromolecules, and iii) structural environment. They concluded with the hypothesis that five conserved amino acid sites (31, 32, 62, 63, and 127) and three class-specific sites (57, 70, and 126) trigger stop codon recognition. Cheng *et al.* proceeded by a systematic functional mutational analysis on the residues of eRF1 involved in stop codon recognition. The mutations on eRF1 that showed a tendency of UGA unipotency were T32 (of the GTS loop), I35 (positioned above the loop) and C127 (YxCxxxF motif region), and those that exhibited UAA and UGA dual potency was E55 [67]. Codon specificity of mutations at Thr32, Ile35, Glu55, Val71, and Cys127 implies that they are crucial for stop codon recognition. The crystal structure of the complex indicates that these residues define the recognition pocket based on their polarity [67]. Apart from these findings, E55 and Y125 showed only UAG recognition *in vivo*, similar to previously reported studies done *in vitro* [74].

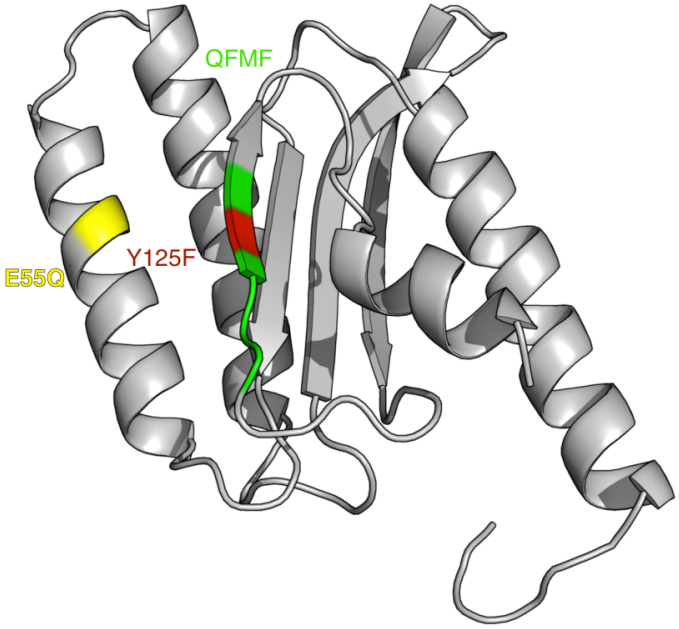


Figure 2.3. Activity associated mutations introduced in N-domain of eRF1. The E55Q mutation is colored yellow and believed to abolish hydrogen bonds with Y125F colored in red. The unipotent-associated mutation, QFM_F is colored in green. The N-domain is derived from crystallographic data [62].

2.1.6 Medical implications of premature termination

The efficiency of translation termination depends on the recognition of the premature termination codons (PTC) by either the eRF1/eRF3 complex or the near-cognate tRNA. Several factors such as nonsense mutations, frame-shift deletions or abnormal splicing of the mRNA causes this PTC. The eRF1/eRF3 complex would result in premature termination whereas the tRNA would interact with the PTC, resulting in 'read-through'. This latter choice of interaction allows ribosome to continue translation until the stop codon is reached and a full polypeptide is synthesized [84]. Nonsense mutations cause approximately 20-40% of the individual cases of more than 240 different inherited diseases, including cystic fibrosis, hemophilia, Duchenne muscular dystrophy, and Marfan syndrome [85]. Correcting nonsense mutations in the regulatory genes to permit synthesis of the respective proteins should cause death of the cancer cells [50, 86]. The presence of PTC in this gene inhibits cell proliferation or promotes apoptosis due to the failure in producing proteins responsible for these functions [87].

2.1.7 Overview of Study

By solving the solution structures of N-domain of both the wild-type (wt) eRF1 and the UGA-unipotent mutant (denoted as $Q^{122}FM(Y)F^{126}$ henceforth), we found that those point mutations alter conformation of the strictly conserved GTS loop (positions 31–33) remote from the mutation sites. This indicates that switching between omnipotence and unipotency of eRF1 may be modulated by distinct conformations of the GTS loop. We thus looked at the other functional mutations to elucidate the structural implications they might have on N-domain of the protein. Based on NMR Residual Dipolar Coupling (RDCs) and chemical shift perturbation (CSP) data, we are able to elucidate that these RF activity-related mutations cause local changes, specifically at the GTS loop and the TASNIKS motif, which are required or involved in stop codon recognition. This conformationally altered state relative to the wildtype tells us the regions that mediate stop codon recognition in the translation termination process. Based on these results as well as recently published study [88], we propose a structural model (**Fig. 2.4**) that encompasses currently known interactions between N-domain of eRF1 and the A site of eukaryotic ribosome. The model shows that it is possible for the GTS loop to contact the stop codon while helix a1 is positioned next to the decoding region of Helix 44. Hence, a major domain rearrangement between N-domain and C-domain is likely to occur in translation termination during which N-domain accommodates itself into the A site.

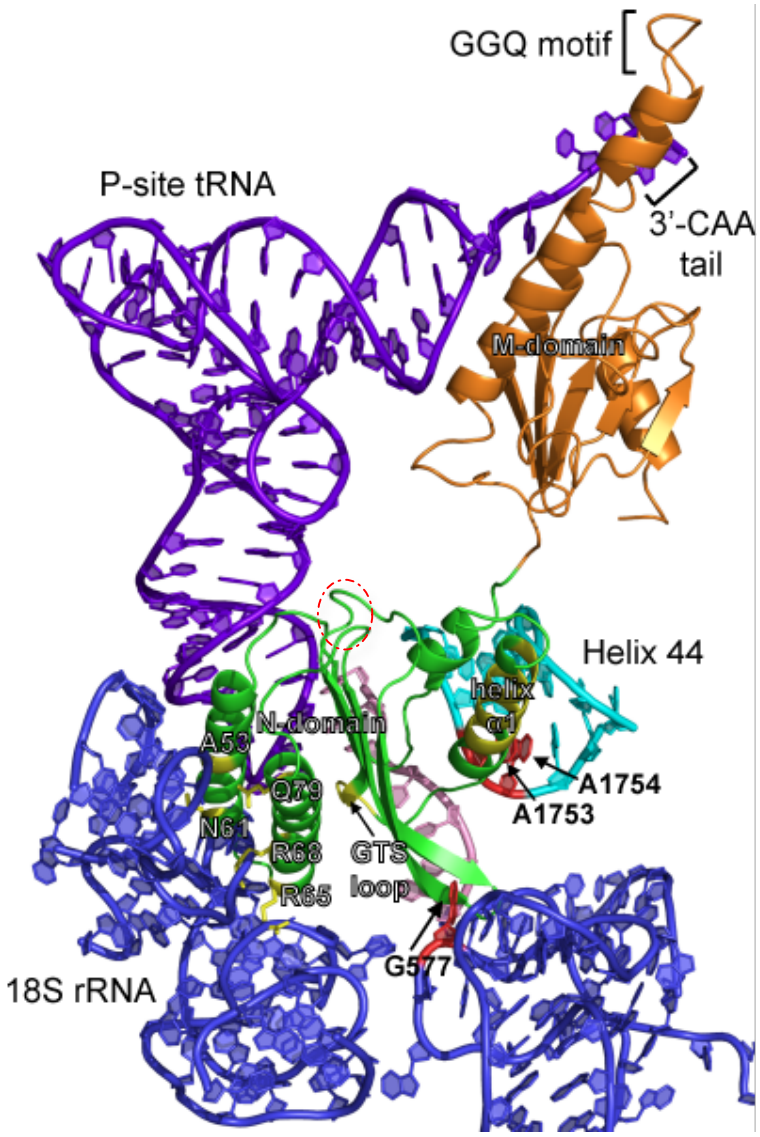


Figure 2.4. Our model of eRF1 bound to the A-site of Eukaryotic Ribosome (PDB ID: 3IZ7) [88]. The truncated NM-domain of eRF1 (green and orange) (PDB ID: 1DT9) was docked onto the A-site of plant *Triticum aestivum* 18S rRNA (blue) with P-site tRNA (purple) and mRNA (pink), based on our results on the interaction between helix α1 of N-domain and the decoding region of helix 44 (cyan), as well as on the role of the GTS loop in stop codon recognition.

2.2 Materials

2.2.1 *E.coli* strains and plasmids

The bacterial strains used in this study for protein expression were *E.coli* BL21 (Rosetta) host cells obtained from Novagen.

2.2.2 Growth media and antibiotics

All buffers were prepared in 1 liter stock, unless otherwise stated and filtered using 2 µm nylon filter. The antibiotics were prepared at 1M stock solutions and kept at -20 °C. The growth media used was prepared according to the compositions stated below, pH adjusted to 7 and autoclaved prior to use. LB agar-plates were prepared by adding 2% (w/v) of agar to LB medium and autoclaved, ampicillin and chloramphenicol was added to the media when temperature was ca. 40 °C.

Table 2.2. Growth media and antibiotics

Name	Material	Amount	Procedure
LB medium	Tryptone	10g	pH adjusted to
	Yeast extract	5g	7.0 using NaOH
	NaCl	5g	
M9 salts	$Na_2HPO_4.7H_2O$	64 g	dissolve the salts in
5XM9 (1L)	KH_2PO_4	15 g	ddH_2O
	NaCl	2.5 g	
	NH_4Cl	5g	
M9 minimal medium	5XM9 salts	200 ml	filter all chemicals into medium
	Glucose	2.0 g	top up to final volume
	$MgSO_4.7H_2O$	0.494 g	of 1L with ddH_2O
	$CaCl_2.2H_2O$	0.0152 g	
	Thiamine	0.010 g	
	$FeSO_4.7H_2O$	0.010 g	
IPTG stock	IPTG	1M	filtered 0.2 µm, storage -20°C
Ampicilin	Ampicilin	100 mg/ml	filtered 0.2 µm, storage -20°C
Chloramphenicol		35 mg/ml	filtered 0.2 µm, storage -20°C

Table 2.3. Buffers used in purification

Name	Material	Amount	Procedure
Binding buffer	K-PO_4 buffer	20 mM	pH 6.8
	KCl	100 mM	
	Imidazole	25 mM	
	DTT	2 mM	
	PMSF	1 mM	
	Protease inhibitor cocktail		
Elution buffer	K-PO_4 buffer	20 mM	pH 6.8
	KCl	100 mM	
	DTT	2 mM	
	PMSF	1 mM	
NMR buffer (100 ml)	MES	20 mM	pH 6.0
	KCl	100 mM	
	DSS	0.1 mM	
	DTT	2 mM	
	H_2O	96%	
	D_2O	4%	

2.3 Methods

2.3.1 eRF1 Mutagenesis, Expression and Purification

Competent Rosetta cells were used in the transformation process by electroporation. An overnight starter culture was used to inoculate a 10 ml Luria Broth (LB) medium culture, which was incubated at 37 °C/180 rpm. Expression was induced with 0.2 mM IPTG at 25 °C, overnight. The cells were then harvested by centrifugation at 8000 rpm (JLA 9.1, Beckmann) at 4°C for 20 mins and lysed in 20 mM HEPES at pH 7.5, 300 mM NaCl, 25 mM imidazole, with 2 mM DTT.

The cell pellet was then homogenized in Lysis buffer by sonication to obtain the cell extract. The cell lysate was consequently centrifuged at 18 000 rpm (JA 25.5, Beckmann) at 4°C for 50 minutes to remove unwanted cell debris. The lysate was purified using FPLC system (AKTA) by loading onto a Nickel based-resin which was pre-equilibrated with buffer A (20 mM HEPES, pH 7.5, 300 mM NaCl). The protein was eluted with 250 mM imidazole in the buffer. The purity and identity of the fractions was determined using SDS-PAGE and mass spectrometry, respectively. Following this step, a buffer exchange was

performed to remove the high amount of Imidazole present in the protein sample. The exchange buffer used were 20 mM MES at pH 6.0, 100 mM KCl, 0.05%NaN_3. The protein samples were then concentrated using Sartorius filter tubes with 10 kD cut-off (Sartorius Stedim) as the molecular weight of the protein is 17 kD. The NMR samples consisted of 90 % H_2O/10 % D_2O, with 5 % DSS as internal standard.

2.3.2 Concentration determination

Protein concentration was determined using NanoDrop (Thermo Fisher Scientific). The absorbance of the purified protein was measured at 280 nm. The concentration was then calculated using the Beer-Lambert law , $A_{280} = \varepsilon.c.d.$ Where A_{280} is the absorbance at wavelength of 280 nm, ε is the extinction coefficient of the protein of interest obtained from its sequence using the Scripps Research Institute Protein Calculator v3.3, *c* is the concentration of the protein observed and *d* is path length.

2.3.3 Circular Dichroism Spectroscopy

All CD experiments were recorded at the pH of the corresponding NMR samples using a Chirascan CD spectrometer (Applied Photophysics, UK). Near and far-UV measurements (200-320 nm) were both performed using a quartz cell with a path length of 1.0 cm and 0.01 cm respectively and the temperature was maintained at 25 °C. CD spectra were acquired using sample concentrations of 100 μM in NMR buffer (20 mM MES at pH 6.0, 100 mM KCl). Three replicates were acquired and averaged. All spectra were corrected against the buffer baseline and processing of spectra was done using the available software Chirascan.

2.3.4 NMR sample preparation

Purified protein samples were concentrated using Sartorius filter tubes with 10 kD cut-off (Sartorius Stedim) as the molecular weight of the protein is 17 kD. The NMR buffer used were 20 mM MES at pH 6.0, 100 mM KCl. The NMR samples consisted of 90 % H_2O/10 % D_2O, 5 % DSS as internal standard, and transferred to 5 mm or 3 mm NMR tubes (Norell).

2.3.5 NMR data acquisition and processing

NMR experiments were performed on a Bruker 600 and 700 MHz spectrometer, equipped with Bruker ^{1}H, ^{13}C, ^{15}N triple-resonance probe-heads with three-axis, Ultrashield gradient coils. The spectra were collected at a regulated temperature of 298K. The residual HDO resonance signal was suppressed with presaturation. A combination of experiments was used to derive the assignments of the backbone for eRF1 mutants (E55Q). $^{1}H^{N}$ and ^{15}N resonances observed from the TROSY experiments were correlated with their corresponding inter- and intra-residue spin systems from 3D experiments namely ^{15}N-resolved NOESY and ^{15}N-resolved TOCSY. The ^{15}N-resolved NOESY spectra were collected in the phase-sensitive manner, with mixing times of 200 msec. The experiment was recorded using 320 × 40 × 1k complex points with spectral widths of 7 kHz (F1), 2 kHz (F2) and 9 kHz (F3), and a relaxation delay of 1.1 s. For ^{15}N-resolved TOCSY, a mixing time of 200 msec was used. The experiment was recorded using 240 × 80 × 1k complex points with spectral widths of 7 kHz (F1), 2 kHz (F2) and 9 kHz (F3), and a relaxation delay of 1 s. The raw data was processed using TopSpin 2.1 software, and the chemical shifts were referenced directly (^{1}H) to the frequency of DSS. Peak picking and spectral analysis was done using CARA. The resonances were verified by reference to 2D [^{15}N, ^{1H}]-TROSY spectra of wildtype at pH 6.0.

2.3.6 NMR structure of mutants of N-domain of eRF1

Residue specific resonance assignment is a prerequisite for NMR-based structure reconstruction. The backbone of the protein is first assigned by correlating the amide group (N_H, H_N) of a particular spin system with the intra- and interresidue chemical shifts of the backbone and also the side chain atoms [26]. Upon completion of the sequence specific assignment and side chain assignments, structure calculation is done to obtain the 3-dimensional structure of the protein. The necessary structural restraints are extracted from NMR parameters. Among the restraints required for structure calculation are distance restraints from nuclear Overhauser effects (NOEs), hydrogen bond restraints that can be obtained from hydrogen exchange experiments of labile protons, dihedral angle restraints that are derived from scalar couplings, chemical shifts, and global restraints from Residual Dipolar Coupling (RDCs) [28]. These global restraints aid in understanding the global

nature of the protein as well as identify inter domain interaction, if any, that are present within a particular macromolecule. A change in the traditional approach of NOE-constraint based approach in determining structures have been suggested, where backbone information complemented by RDCs is used instead to compute structures for a wide range of proteins [33].

2.3.7 Residual Dipolar Coupling (RDC) Alignment

i. Filamentous Bacteriophage, *Pf1*

Pf1 was purchased from Asla Biotech. The sample preparation below is for a *Pf1* phage in 10 mM potassium phosphate buffer, pH 6.0 [40]. *Pf1* was added to the eRF1 mutant QFM_F (0.65 mM) at different concentrations to determine the optimum concentration to induce partial alignment of the protein. Concentrations tried were 20, 15 and 10 mg/ml *Pf1*. Approximately 10 mg/ml bacteriophage *Pf1* was found to be sufficient for protein alignment. The sample appeared to be clear and free of precipitates, as opposed to the other two higher concentrations tried. 2D [^{15}N,^{1}H]-TROSY and anti-TROSY experiments were measured for the sample in both isotropic and anisotropic conditions. The same concentration (10 mg/ml) was consequently used for the eRF1-wt and mutant E55Q and Y125F.

ii. Strained Polyacrylamide Gels (SAG)

It takes approximately 2-3 days for complete sample preparation. 250 µl of gel solution would make sample volume height of approximately 20 mm. Q^{122}FM(Y)F^{126} sample concentration prior to diffusion into the gel is 0.6 mM. After diffusion, the concentration will drop to half its original amount, ~ 0.3 mM. 2D [^{15}N,^{1}H]-TROSY and anti-TROSY experiments were measured for the sample in both isotropic and anisotropic conditions. 3 different percentages of gel were tried, 5, 7 and 10 % (**Table 2.4**) and for all three conditions, the protein was first introduced into the sample via co-polymerization with the gel. This method proved to be unsuccessful; therefore we proceeded with the diffusion method, also using 3 different percentages of gel as previously mentioned. The 7% gel for the diffusion method showed the best results for SAG.

Table 2.4. Gel composition for alignment

	7% (ml)	5% (ml)
40% Acryl/Bisacryl	43.75	31.25
APS (0.1%)	2.5	2.5
TEMED (0.1%)	1.75	1.75
H2O	202	214.5
Total Volume	250	250

iii. Ether/alcohol liquid crystalline phases

Another alignment media that was tried was the ether/alcohol liquid crystalline phases (Otting media). This media is suitable for proteins, DNA and protein/DNA complex. A stock solution of C8E5 in NMR buffer was prepared to yield a final concentration of 30 %. This was then diluted to 5 % for the working protein sample. The correct weight of octanol was added, drop wise, while vigorously shaking (vortex) to achieve a final molar ratio of 0.85. The biphasic solution becomes transparent and opalescent upon formation of the Lα phases [43]. Higher amounts of alcohol decrease the temperature stability of the liquid crystalline phase. It is advisable to store the samples at room temperature and these samples are stable at least over several months. The NMR sample consisted of 5 µl DSS, 10 µl D_2O, 105 µl protein (0.6 mM), 30 µl of C8E5: octanol solution.

2.3.8 Data analysis

Combinations of programs were used throughout this study in analyzing the spectral data. **Table 2.5** list all the programs used in this study.

Table 2.5. Software used throughout this study

Software	Application	References
TopSpin 2.1	Processing of spectra	Bruker Biospin
CARA	Assignment and spectral analysis	www.cara.nmr
PALES	RDC analysis	[89]
TALOS+	Secondary structure prediction	[90]
CYANA	Structure calculation	Peter Guntert
CNS	Structure calculation and refinement	[91, 92]
MacPyMol	Molecular visualization and graphics	www.pymol.org

2.4 Results and Discussion

2.4.1 eRF1 expression and purification

In order to proceed with NMR structural studies of the eRF1 mutants, it is essential in obtaining high yields of pure mutants in the soluble form. The production of eRF1 mutants at cultivation temperature of 25°C yielded sufficient amounts of protein. Purification of mutants using IMAC was applicable for all the different mutants, with comparatively high purity levels required for structural studies. All the mutants were analysed using SDS-PAGE, with the results depicted in **Figure 2.5**. The large scale expression (1 L culture) yielded 4-5 mg of soluble protein suitable for NMR structural studies.

MALDI mass spectrometry was used to confirm the identity of the purified mutants. The mass of the protein showed a single intact species at a mass/charge ratio of 16 607 ± 10 Da, and this is in agreement with the predicted molecular mass of the mutant E55Q of 16 700 Da.

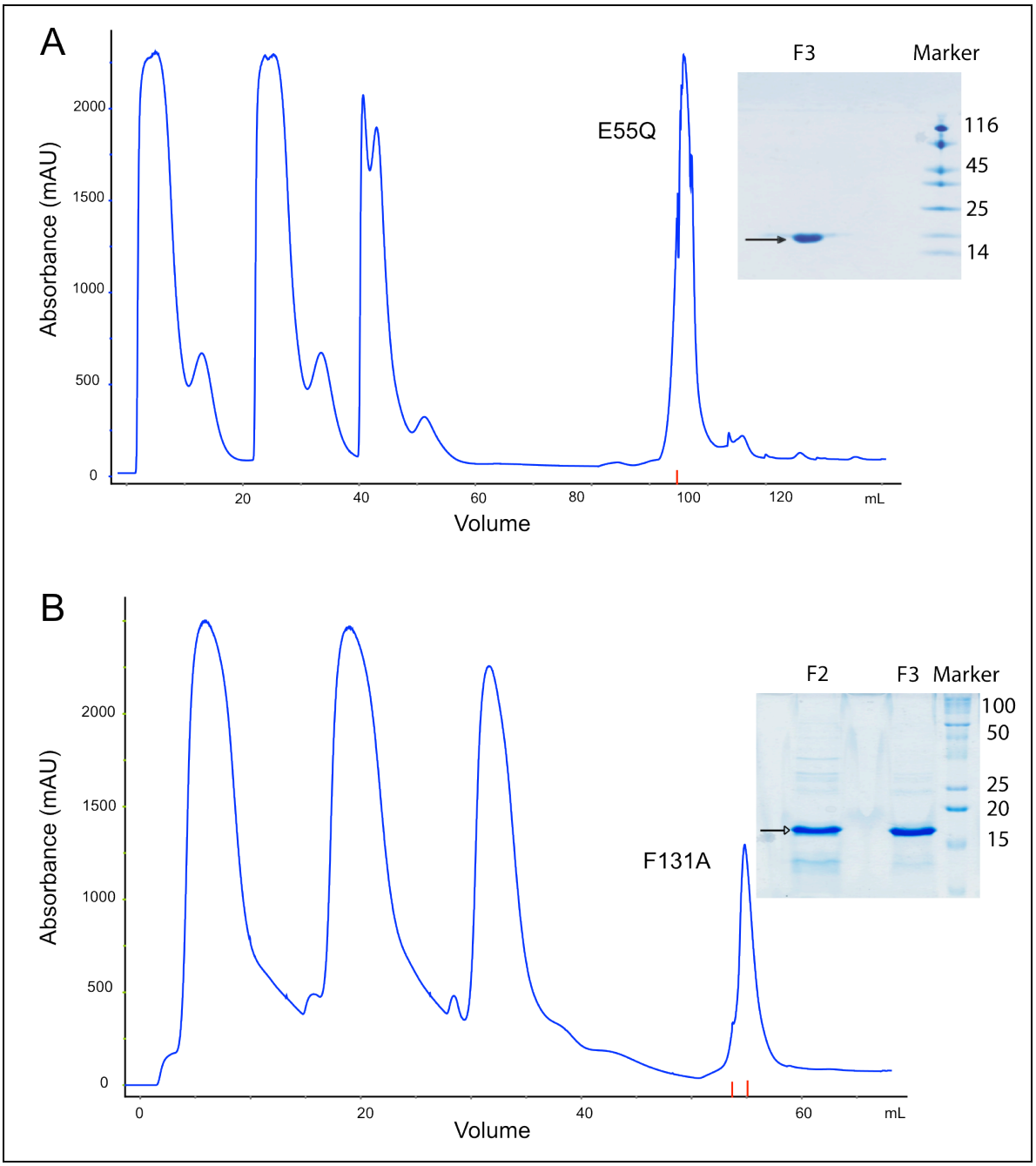


Figure 2.5. Purification of recombinant eRF1 mutants, E55Q and F131A. Marker denotes the lanes loaded with size markers (indicated in kDa). UV absorption curve of A) E55Q ^{15}N labeled and B) F131A ^{15}N labeled, both eluted with a 25–250 mM imidazole gradient in phosphate buffer, pH 7.5, purified using His-Trap column on an Akta Purifier. The inset is the SDS-PAGE (15%) stained with coomassie blue of protein after purification with Hi-Trap desalting column (>97 % pure as evinced by SDS-PAGE).

2.4.2 CD spectroscopy of eRF1 mutants

The secondary structure of eRF1 mutants, E55Q, Y125F and $Q^{122}FM(Y)F^{126}$ was determined through the analysis of circular dichroism, which was measured in the far UV range of 190-240 nm and near UV range of 240-320 nm. **Figure 2.6** illustrates the CD spectra of all the mutants and wildtype overlaid, measured at room temperature. The same NMR sample was used (20 mM MES-K (pH 6.0), 100 mM KCl) for acquiring the spectrum. The temperature was maintained at 25°C and all spectra were corrected against the buffer signal. Both the minima (208 and 222 nm) and the maxima (192) of the near UV spectra are indicative of α-helices present in the protein. The CD spectrum shows clearly that the introduction of mutation(s) to the protein has not, in any way, distorted or changed the secondary structure relative to the wildtype. In contrast, the near UV spectrum **(Fig. 2.6B.)** depicts alterations in the tertiary structure specifically in the environment of aromatic amino acids. The alterations are more significant for $Q^{122}FM(Y)F^{126}$, consistent with the expected effects that these mutations cause on the tryptophan residue.

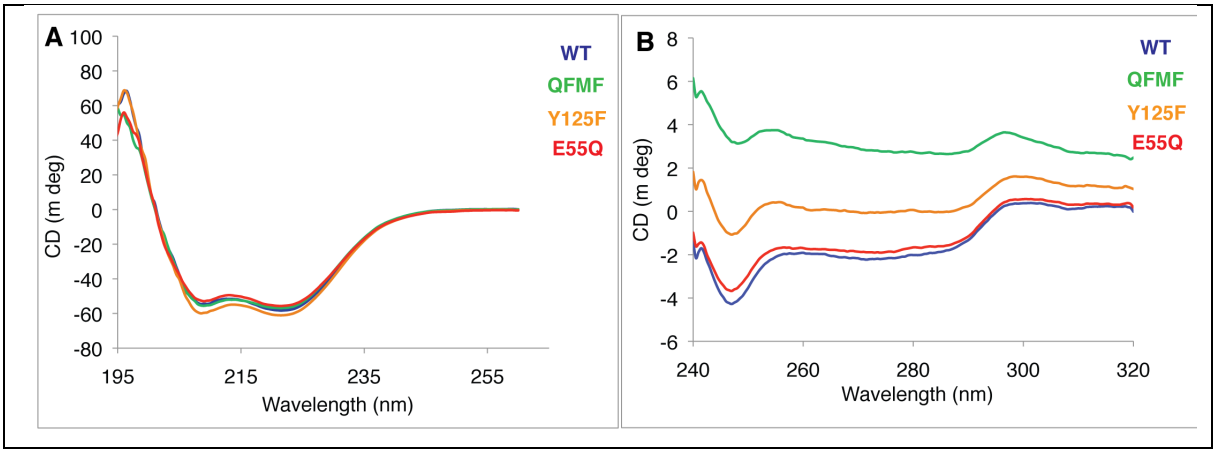


Figure 2.6. CD spectroscopy of wildtype eRF1 and its mutants, E55Q, Y125F and $Q^{122}FM(Y)F^{126}$ were measured at 25°C on a Chirascan spectrophotometer (Applied Biophysics) in A) far UV region and B) near UV region depicting changes in the tertiary fold as a result of mutations. All samples were analysed in NMR buffer (20 mM MES-K, pH 6.0, 100 mM KCl).

2.4.3 Structural Characterization of eRF1 mutant, E55Q

i) NMR backbone assignments

The eRF1 mutants were then expressed using the same protocols in M9 media thus producing ^{15}N-labeled proteins that were concentrated to 400 µL. The 2D [^{15}N,^{1}H]-TROSY spectra of E55Q (**Fig. 2.7**) and Y125F showed good spectral dispersion; most of the ^{1}H,^{15}N and ^{13}C resonances were assigned with reference to the 2D [^{15}N,^{1}H]-TROSY spectra of wildtype eRF1, N-domain at pH 6.0 and 3D-HNCA. ^{13}C-NOESY-HMQC, ^{15}N-NOESY-HSQC and TOCSY were measured to aid the process of side-chain assignment and structure determination via an iterative analysis. The ^{13}C-resolved NOESY spectra contain both aromatic and aliphatic carbons and protons.

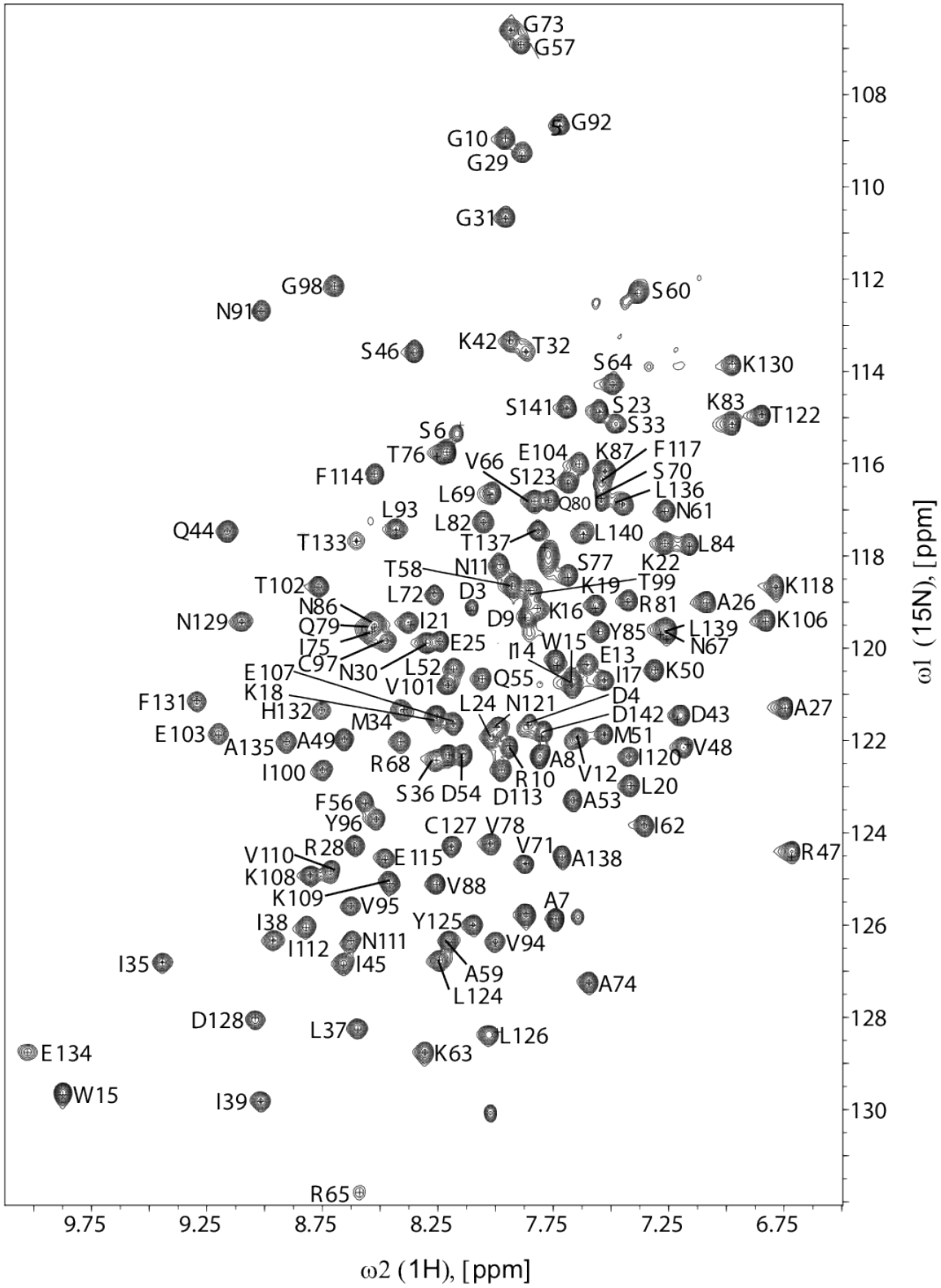


Figure 2.7. 2D [^{15}N,^{1}H]-TROSY of mutant eRF1 E55Q in 20 mM MES-K (pH 6.0), 100 mM KCl, 90% H_2O/10% D_2O at 298K, using 700MHz NMR spectrometer. Residue specific assignments of the amide resonances are labeled.

ii) Structure determination of eRF1 mutant, E55Q

NOE distance restraints for the calculated structures of mutant eRF1, E55Q were obtained from ^{15}N-NOESY-HSQC and ^{13}C-NOESY-HMQC spectra, respectively. Backbone dihedral angle restraints (φ and ψ) were derived from the backbone ^{13}C', ^{13}Cα, ^{13}Cβ, ^{1}Hα, ^{1}Hβ, ^{1}HN chemical shift values using TALOS. Hydrogen bond restraints were obtained from crystal structure, and they were used to help structure determination at the beginning stage. Structure calculations were performed using CYANA 2.1 [93, 94], CNS 1.3 [95, 96] and visualized using PyMOL (Delano Scientific).

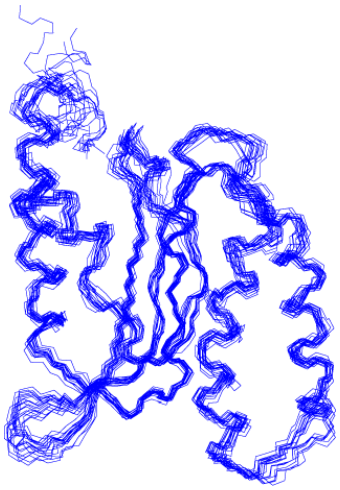


Figure 2.8. Three-dimensional solution structure of mutant eRF1 E55Q. 20 lowest energy conformers are shown in the lines representation.

Table 2.6. Structure statistics for the selected 20 structures of mutant eRF1, E55Q.

NMR restraints	
Total unambiguous distance restraints	2160
Intra residual	1371
Sequential (|i-j| = 1)	294
Medium (2< |i-j| < 5)	1338
Long range (|i-j| > 5)	528
Hydrogen bond restraints	106
Dihedral angle restraints	180
RMSD from average atomic coordinates (residues 6-140, A)	
Backbone atoms	0 .83 +/- 0.2
All heavy atoms	1.33 +/- 0.19
Ramachandran analysis (%)	
Residues in most favored regions	62.50%
Residues in additionally allowed regions	32.40%
Residues in generously allowed regions	5.90%
Residues in disallowed regions	1.20%

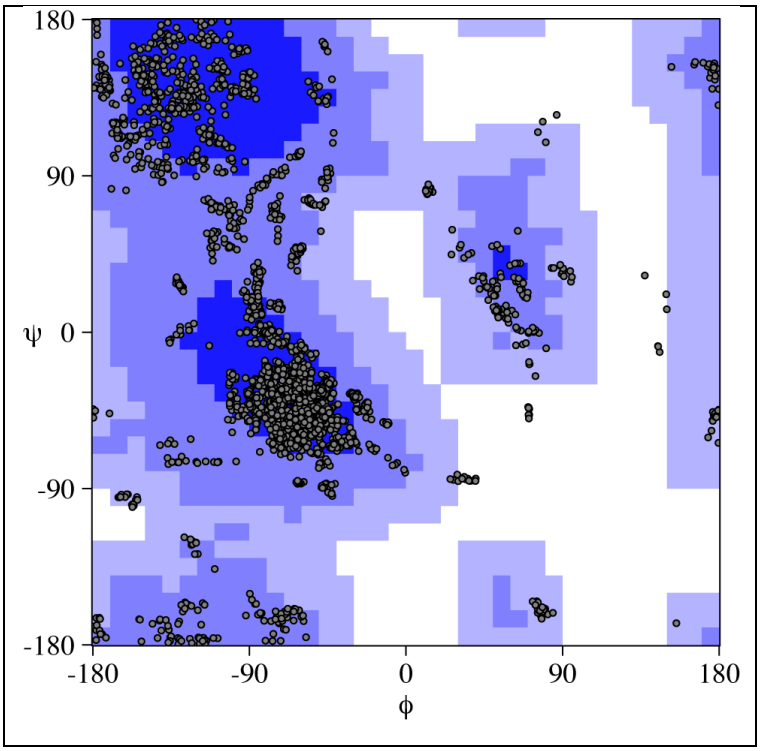


Figure 2.9. Ramachandran plot of residues φ and ψ angles of mutant eRF1, E55Q. 62.5% in most favored regions, 32.4% in additionally allowed region, 5.9% in generously allowed regions and 1.2 %in disallowed regions. This plot is made from CYANA [93].

2.4.4 Chemical shift deviations linked to mutations

The sequence specific structure of eRF1-wt and mutants $Q^{122}FM(Y)F^{126}$, E55Q, and Y125F were defined by ascertaining the backbone NMR resonances of each protein. Each spectra for the corresponding proteins showed well resolved resonances in the 2D [^{15}N,^{1}H]-TROSY. Each mutant has been reported to exhibit different RF activity *in vitro* or *in vivo* [78, 79, 80]. Hence, we strive to extract evidence for the correlation between change in RF activity and structural changes experienced in terms of specific chemical shift changes due to mutations introduced in the protein.

The overall spectral changes observed between eRF1-wt and mutants were minor, however notable changes were observed at the mutation points and the specific regions of N-domain, which have been reported to influence stop codon recognition at the small ribosomal subunit, namely the highly conserved NIKS and YxCxxxF motif [79, 80]. *Stylonychia* eRF1 exhibits UGA-only specificity, which is believed to be caused by the $Q^{122}FM(Y)F^{126}$ pentapeptide sequence. Inserting these mutations into human eRF1 converts its function from omnipotent to unipotent, thus leading to the hypotheses that the residues

from positions 122-132 are not only essential for purine discrimination but also stop codon recognition [73, 74, 81]. Based on this observation, we investigated whether a correspondence exist between the wt-eRF1 and the mutant $Q^{122}FM(Y)F^{126}$ with respect to their [^{15}N,^{1}H]-TROSY spectra. A general trend in the deviation profile is observed for $Q^{122}FM(Y)F^{126}$ compared to eRF1-wt (**Fig. 2.10**). The introduction of mutations caused perturbations to occur throughout the protein with predominantly small CSP (<0.05 ppm). Resonances corresponding to residues Q123, Y125 and F126 were observed to experience significant perturbations ($0.15 < \Delta\delta < 0.3$ ppm) as expected. Interestingly, these mutations also influenced other regions of the protein resulting in distinct deviations specifically for residues G29, G31-S33 (GTS loop), L37-I39, and Q44 in the N-terminal region of the protein. The C-terminal tail of $Q^{122}FM(Y)F^{126}$ experienced deviations at E134 and A135 resulting in shifts of more than 0.1 ppm.

Both E55 and Y125 are associated to the recognition of G in UGA, with Y125 playing the dominant role of recognition. The hydrogen bond formed between these two residues may help in maintaining the spatial proximity of the protein and thus influence RF activity [74] as well as being part of the YxCxxxF motif. The [^{15}N, ^{1}H]-TROSY spectrum of Y125F was contrasted with that of wt-eRF1 and found to have a similar CSP profile as $Q^{122}FM(Y)F^{126}$. In addition to the similar deviations at the GTS loop (positions 29-33), larger deviations were observed at L52, D54, Q55, A59, R65, and N67 (NIKS tetrapeptide motif). Residues L72, T76, G98, T99, I100 and V101 induced smaller deviations within the range of 0.05 to 0.1 ppm, possibly reflecting a change of conformation in the protein. Interestingly, the region of mutation, Y125F, induced deviations for residues Y125-D128 (the YxCxxxF motif) and E134 leaving residues T122-L124 unchanged. From the overall comparison of CSP data, we observed a correlation between the deviations observed and the mutations introduced. $Q^{122}FM(Y)F^{126}$ and Y125F mutually share similarities in the deviation profile relative to eRF1-wt specifically at the GTS loop. Likewise, E55Q's deviation profile is similar to Y125F but not $Q^{122}FM(Y)F^{126}$. However, all three mutants experienced deviations at residue E134 regardless the site of mutation or number of mutations.

The CSP results are reflective of the nature of interactions present within the protein itself. It mirrors the changes that are experienced by the 3D structure of the protein when

mutations are introduced. Removal of either E55 or Y125 from the protein diminishes the hydrogen bonding between them, which have been shown to influence RF activity specifically in recognizing G of UGA. The similarities of CSP profile between Y125 and $Q^{122}FM(Y)F^{126}$ show that residues at the YxCxxxF motif play a role in stop codon recognition via the GTS loop. We show here the correlation between the three sets of mutations at these conserved residues and how they influence the similar changes in the protein.

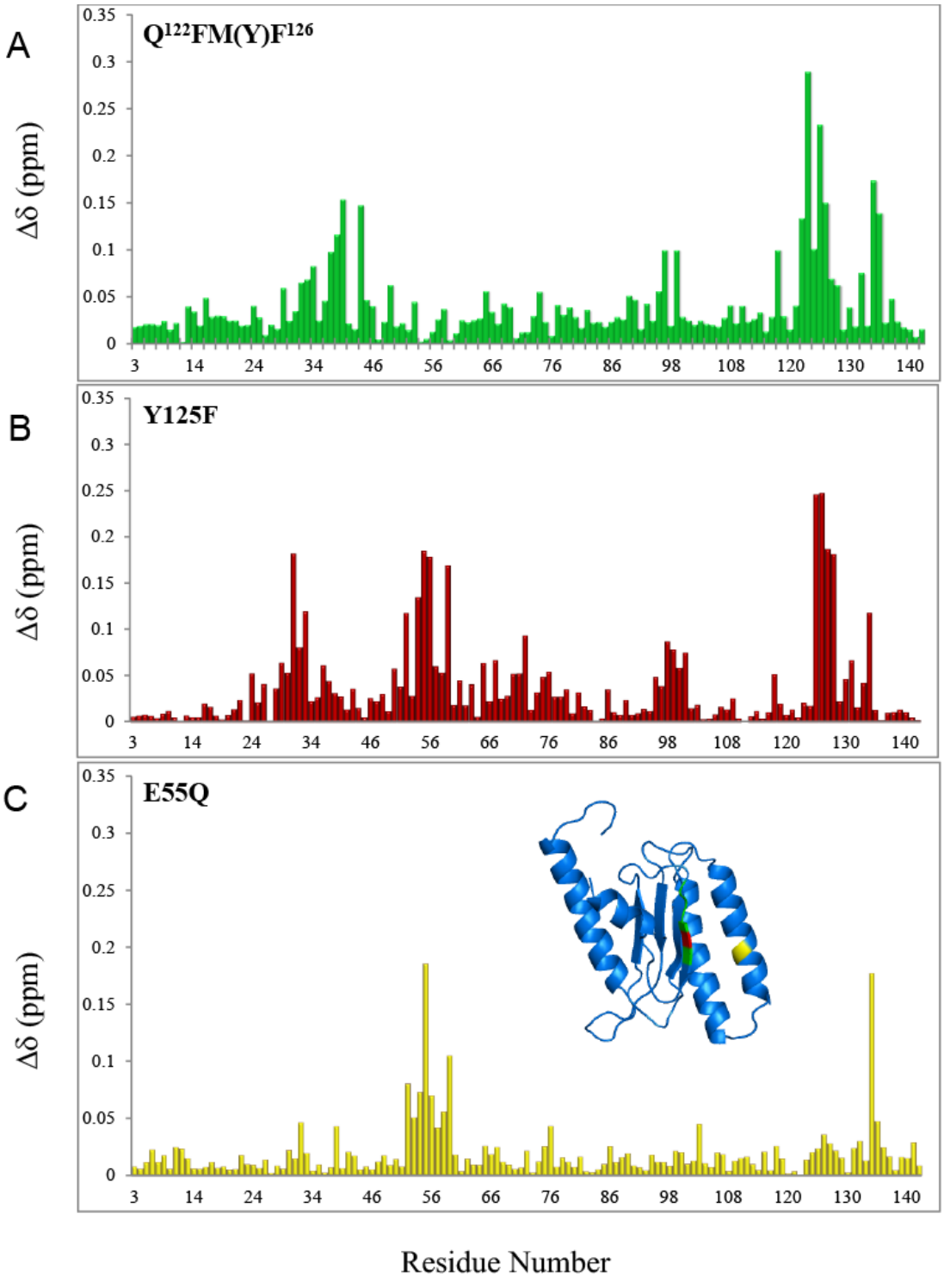


Figure 2.10. Chemical shift perturbation induced to the backbone amide signals between eRF1-wildtype and its mutants. Structure insert shows the location of mutated residues on the wildtype (structure derived from PDB: 1DT9).

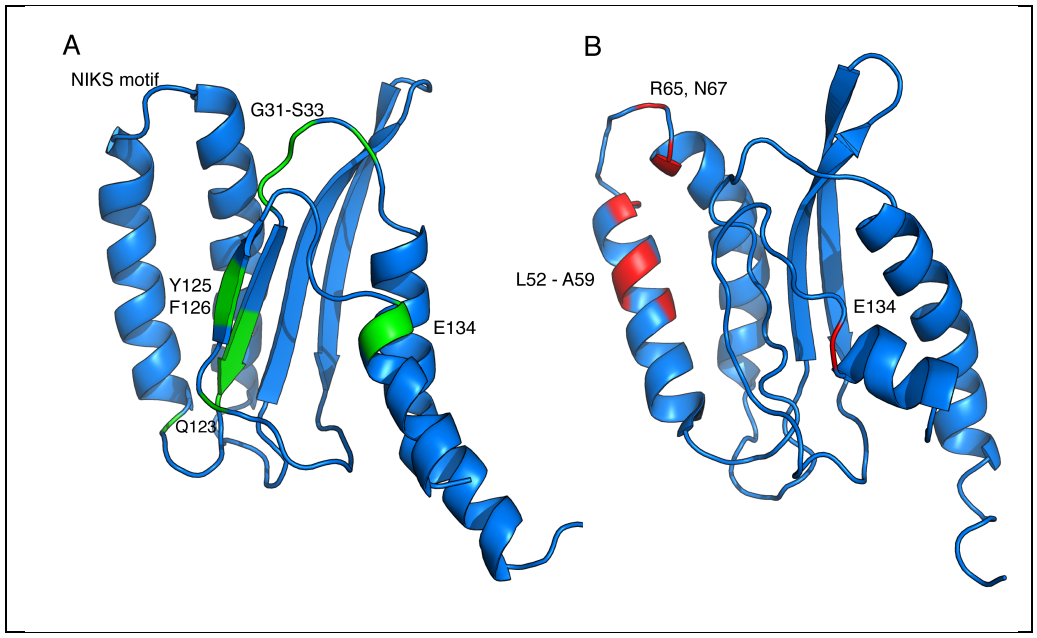


Figure 2.11. Location of residues that experience perturbations upon introduction of mutation a.) $Q^{122}FM(Y)F^{126}$ (structure derived from PDB: 2LGT) b.) E55Q

2.4.5 Residual Dipolar Coupling (RDC)

From the study by Kisselev *et al*., the specificity of stop codon recognition is associated to eRF1 protein itself rather than the ribosome, and the recognition site is observed to be located at the N-domain. Substitutions were introduced in different positions in the N-domain of human eRF1, which to our interest were at $Q^{122}FM(Y)F^{126}$ E55, and Y125 as these positions were deemed as potentially important for stop codon recognition [71, 74].

RDCs provide both short range and long range structural information; due to this dependence on distance, orientation and dynamics, it is therefore an essential restraint in characterizing structure [34]. Two different alignment media was used to determine the residual dipolar coupling; *Pf1* [42] and C8E5 [43]. The reproducible RDC profile for each mutant in both media indicated that there were no drastic perturbations in the ensemble of conformations. The chemical shifts of the wt-eRF1 in the presence of alignment media were the same as the protein alone and similar observations were noted for mutant-eRF1s, suggesting that the anisotropic phase did not perturb the protein. As RDC's reflect the presence of long-range contacts [21], we have used it here to observe and compare the effects mutations have on the conformation of the protein.

The use of different alignment media for a single molecule is beneficial for the structure determination/refinement process to reduce the ambiguities of the translation of dipolar couplings into orientations. This would also allow the application of RDC in the determination of the dynamics of the macromolecule. The distribution of RDC values for eRF1-wt were plotted to determine the suitability of the alignment media to this particular protein. Apart from having a good distribution of RDC values (presence of powder pattern in the histogram-not shown), the peaks in the spectra were well dispersed and well resolved. *Pf1* was then consequently used on all the remaining mutants namely, $Q^{122}FM(Y)F^{126}$, E55Q and Y125F. Another alignment media tried was the polyethylene glycol/alcohol mixtures (C8E5/octanol) mixtures at 5% ratio. This media yielded well resolved peaks and a dispersed spectra. The method was then repeated for the other mutants in order to extract a complete set of RDC values.

Figure 2.12 depicts the results obtained from the SAG method, whereby the protein was introduced into the alignment via two different ways, namely co-polymerization and diffusion. The co-polymerization method did not yield good results, as the peak resolution was low for all the three different percentages of gel used. The main problem faced with this technique is the loss of conformational homogeneity of the protein. We observed a number of cross peaks that are split into several subpopulations of cross-peaks. This is indicative of the presence of long-range dipolar couplings and could be resultant of excessive protein alignment. We then proceeded to try the diffusion method, where the protein was diffused into a dried gel before radial compression to align it. This method yielded best results at a 7% gel. Although the spectra obtained were observed to have all peaks, the peak resolution was still not considered to be sufficiently good enough to extract RDC values. In order to avoid introducing error, we decided not to use this set of data.

The RDC data that was extracted from the different alignment media was then used to estimate the magnitude of the alignment tensor, which comprises the axial (Da) and Rhombic components using PALES [89]. Singular Value Decomposition (SVD) was used in estimating the alignment. They were obtained iteratively by best fitting the alignment tensor to a given structure during the refinement process. The correlations between the experimental and predicted values were determined.

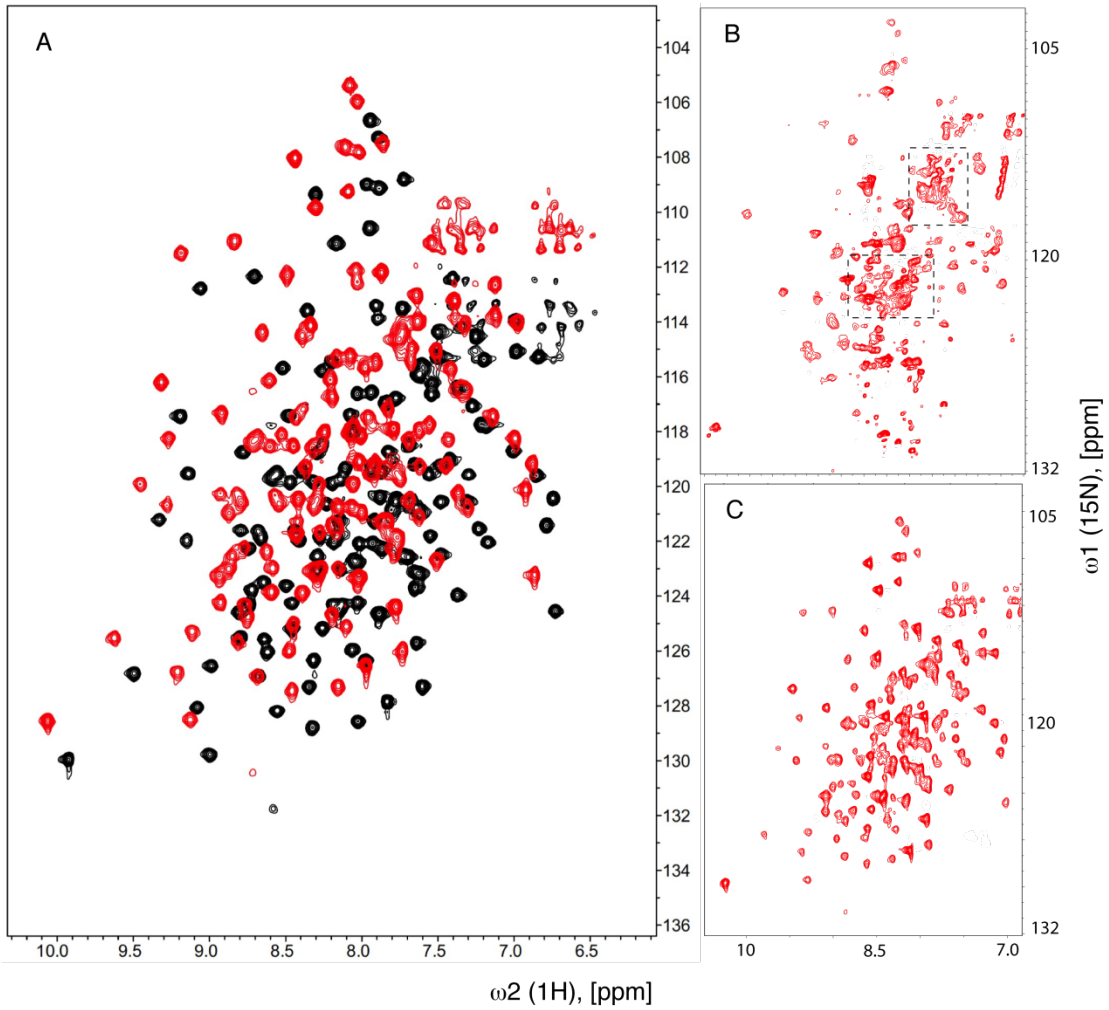


Figure 2.12. 2D [^{15}N,^{1}H]-TROSY spectra for determining dipolar coupling. The anisotropic ^{15}N labeled eRF1-wt in 100 mM phosphate buffer, pH 6.0. The samples were maintained at 25 °C for the measurements performed on a cryoprobe-equipped Bruker 700 MHz spectrometer. A) TROSY (black) and anti-TROSY (red) peaks of anisotropic eRF1-wt in *Pf1* B) anti-TROSY peaks of eRF1-wt in SAG 7% and C) anti-TROSY peaks of eRF1-wt in C8E5 (5%).

2.4.6 eRF1 global conformations perturbed by introduced mutations

Backbone HN RDCs were measured from wildtype eRF1 and its mutants ($Q^{122}FM(Y)F^{126}$, E55Q, Y125F) . These RDCs were analyzed using the crystal structure of wildtype eRF1 and solution structure of $Q^{122}FM(Y)F^{126}$ (PDB 1DT9 and 2LGT respectively).

The structures were analyzed using all the extracted RDC values from the eRF1-wt and each mutant onto the crystal structure of the N-domain eRF1 to evaluate the correlation coefficient, as illustrated in **Figure 2.13** and **2.14**. We observed that the correlation coefficient was low although all RDCs used were accurate. RDCs corresponding to the GTS loop and the TASNIKS region were then removed and the data re-analysed to observe changes, if any. When RDCs not including the GTS loop or TASNIKS regions were fitted, a higher correlation coefficient was observed for the wildtype eRF1 and mutants alike. We then re-analysed the R value using the same alignment tensor but with the addition of the GTS loop and TASNIKS region respectively to observe the deviation of these specific RDCs from those previously calculated.

For the eRF1-wt, a correlation coefficient of R = 0.904 was observed for RDCs without the GTS loop and TASNIKS region. R = 0.794 and 0.815 was observed when GTS loop and TASNIKS region was added, separately. Similar lower correlation coefficients were observed for $Q^{122}FM(Y)F^{126}$ mutant. However, for E55Q and Y125F, the TASNIKS region seemed to be more affected, giving a lower correlation coefficient compared to the GTS loop. Referring back to our CSP profile of E55Q and Y125F, we observed that most of the deviations that were observed was in this particular region as well. The lower correlation coefficient could be accounted for as conformational changes experienced by these mutants, suggesting these regions to be highly dynamic in nature and essentially involved in the recognition of stop codons as shown previously [58, 65]. A reduced correlation coefficient for the full structure relative to the selected regions further proves the dynamic characteristic of these regions. The effects of both the GTS loop and the TASNIKS region seem dependent on the mutations introduced thus suggesting that the conformational changes that drive stop codon recognition are a result of many residues as opposed to one single residue.

Although previous studies have shown that these regions in the protein are involved in stop codon recognition, it has always been done separately and without a correlation to

each other at a molecular or structural level. Here we have extended the mutational analysis of the protein at different regions that have been proposed as functional to stop codon recognition to show how these residues affect each other either locally (CSP) or globally (RDC). The RDC data exposes the dynamic nature of the protein specifically at the GTS loop.

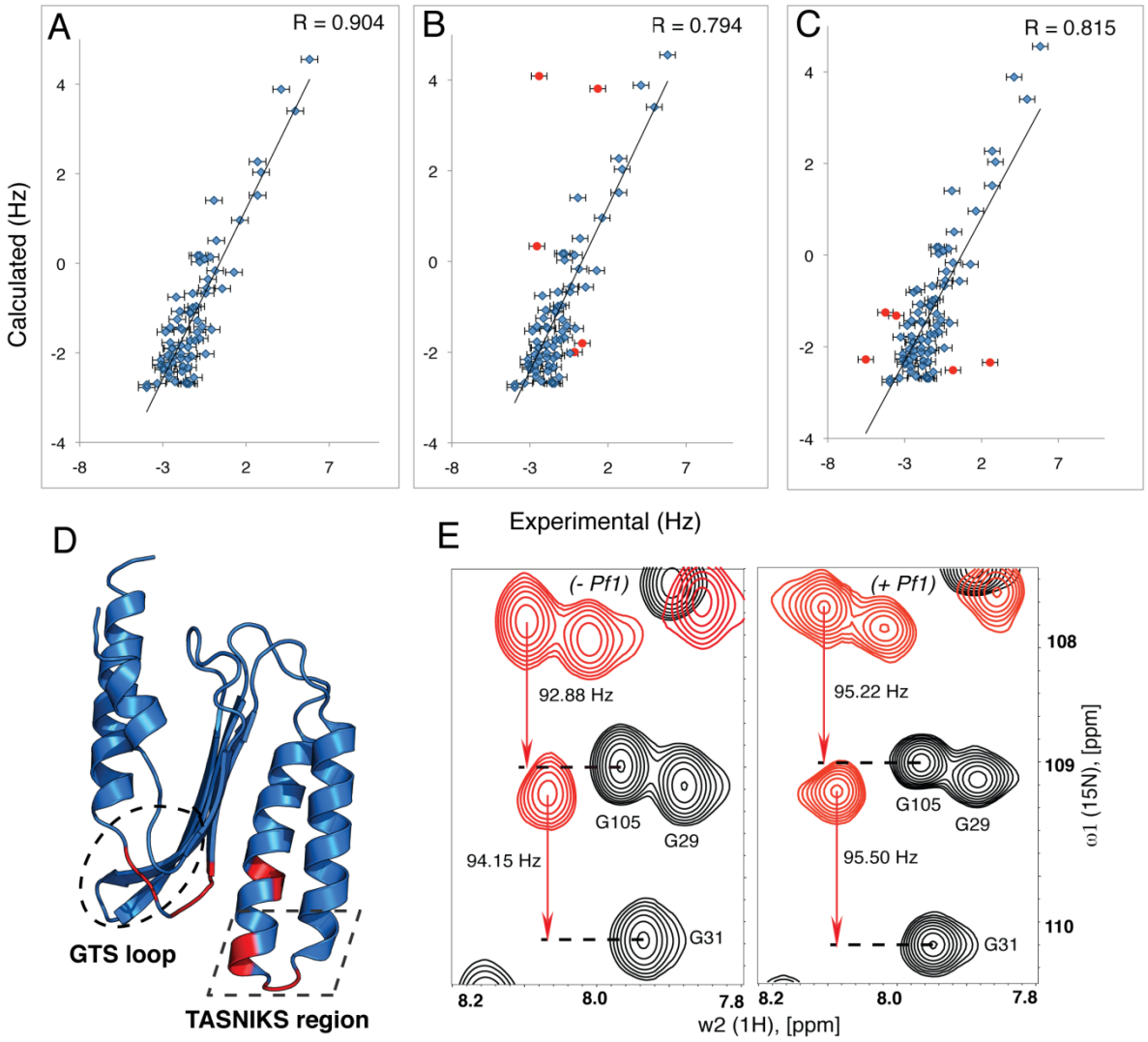


Figure 2.13. RDC experiments with wt-eRF1 and mutants of eRF1.Experimental and calculated Residual Dipolar Coupling for wt-eRF1 and their correlation factors with respect to the full-length A) wt-eRF1 crystal structure without the GTS and TASNIKS regions (blue), B) with the GTS loop included (red) C) with the TASNIKS region included (red) using PALES. D) N-domain of wt-eRF1 derived from the crystal structure (PDB: 1DT9). E) 2D [^{15}N,^{1}H]-TROSY (black) and anti-TROSY (red) spectra illustrating the coupling for isotropic (left) and anisotropic (right) conditions for wt-eRF1 using *Pf1*.

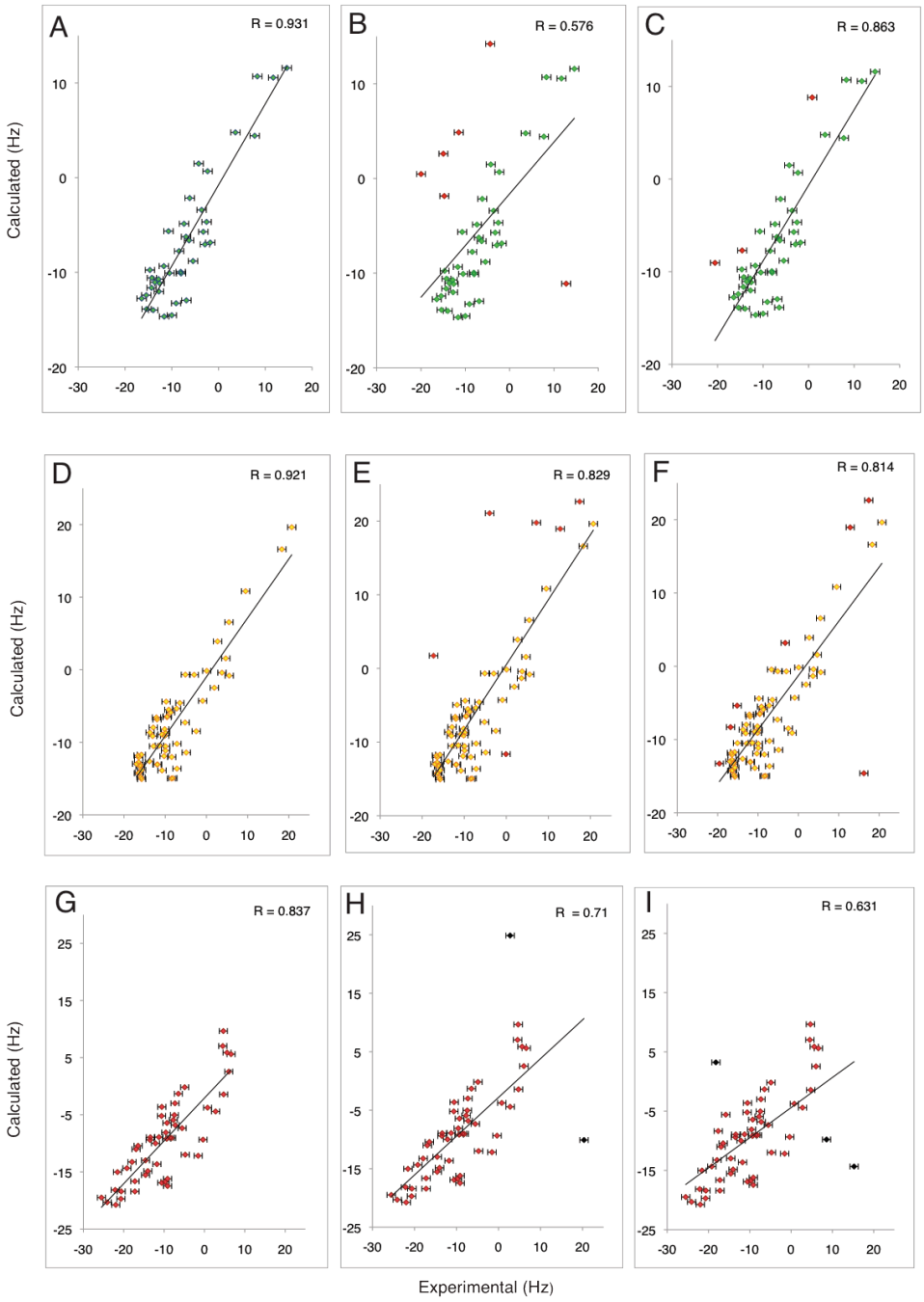


Figure 2.14. Experimental and calculated Residual Dipolar Coupling and their correlation factors, R for A) $Q^{122}FM(Y)F^{1126}$ with respect to the full-length wt-eRF1 crystal structure without the GTS and TASNIKS regions (green), B) with the GTS loop included (red) C) with the TASNIKS region included (red). D) E55Q with respect to the full-length wt-eRF1 crystal structure without the GTS and TASNIKS regions (yellow), E) with the GTS loop included (red) F) with the TASNIKS region included (red). G) Y125Fwith respect to the full-length wt-eRF1 crystal structure without the GTS and TASNIKS regions (red), H) with the GTS loop included (black) I) with the TASNIKS region included (black).

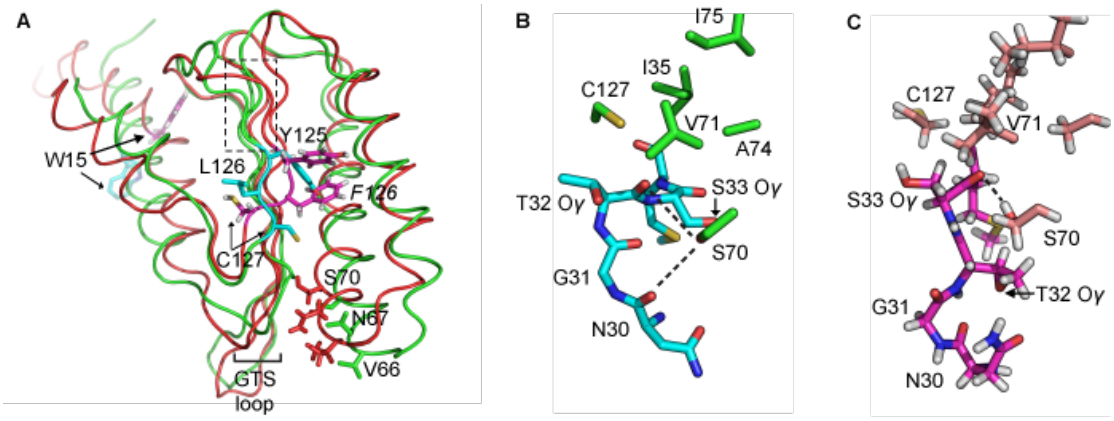


Figure 2.15. Structural comparison between wild-type N-domain and $Q^{122}FM(Y)F^{126}$ Mutant [88]. A) The wild-type crystal structure (green) was superimposed on the average $Q^{122}FM(Y)F^{126}$ solution structure (red) using SuperPose server. The dashed lines enclosed residues at the mutation positions TSL → QFM. Mutation position F126 is in italics. B) Backbone of the GTS loop (N30-M34) in the wildtype structure is traced (cyan), while only side chains are shown for other residues in the vicinity (green). Dashed lines indicate possible hydrogen bonding. C) Backbone of the GTS loop (N30-M34) in the $Q^{122}FM(Y)F^{126}$ structure is traced (magenta), while only side chains are shown for other residues in the vicinity (wheat). Dashed lines indicate possible hydrogen bonding.

2.5 Conclusion

The selectivity of stop codon regulation is governed by the multiple conformations adapted by the strictly conserved GTS loop in wt-eRF1 and its mutants. This loop has been proposed as being involved either directly or indirectly in decoding and interacting with the stop codon [80, 97]. The significance of this finding is the effect that these structural changes might have on stop codon selectivity; significantly influenced by the interactions of amino acids with the stop codon itself (**Appendix A2**).

It is interesting to note that a single point mutation of T32A showed UGA unipotency (E. Alkalaeva, personal communication), which is the same preference as our mutant of interest $Q^{122}FM(Y)F^{126}$. Why does T32A mutant show similar attribute as $Q_{122}FM(Y)F_{126}$? The similarities in preference for the same stop codon between two different mutants has been observed for various other eRF1 mutants; where the point mutations introduced across the N-domain results in the same bias of stop codon selectivity (functional studies). One could surmise that the eRF1 mutant that ceases to interact with one or two out of the 3 nucleotides in a stop codon, could still function as a RF, however with reduced efficiency. Another possibility to explain the above observation is that mutations that give rise to altered stop codon recognition actually influence the conformations of N-domain regions

that are involved in stop codon interaction. Conversely, mutations of C127 (highly conserved) in human, *Euplotes* and *Stylonychia* eRF1s have different responses: the appearance of UGA recognition for *Euplotes* eRF1, and its disappearance for *Stylonychia* eRF1. This leads us to believe that C127 does not in actual fact directly participate in stop codon recognition, and the same could be inferred for the amino acids in this region (residues 122-126).

We propose that mutations that alter the selectivity of stop codon recognition modulate the different conformations of the GTS loop that is required for eRF1 to function efficiently. Mutational studies have been done on the residues constituting the hydrophobic core directly above the GTS loop, namely residues I35, V71, V78 and C127 [60, 67]. A comparison of the width of this hydrophobic core between wt-eRF1 (crystal and solution) and $Q^{122}FM(Y)F^{126}$ depicts a decrease from 7.68A˚ and 7.38±0.18A˚ (crystal and solution respectively) to 6.56±0.13A˚. In light of these observations, we hypothesize that repositioning of helix a3 in N-domain occurs during stop codon recognition as the GTS loop is sampling different configurations. Interestingly, S70 on helix a3 was observed in stabilization of the GTS loop in $Q^{122}FM(Y)F^{126}$ [88]. Residue S70 is critical for UGA-decoding, as a point mutation S70A restricts human eRF1 to recognize UAA and UAG only. At the same time, the single A70S substitution in *Euplotes* eRF1 changes stop codon recognition from UAR-only specificity to an omnipotent one [98]. These data verify the assumption that S70A substitution is associated with UGA reassignment [75]. Helix a2 could also play a role in modulating the selectivity of stop codon recognition. First, M51 and E55 on helix a2 are able to alter stop codon recognition pattern [60, 74]. Secondly, the TASNIKS motif was found to confer distinct requirement of eRF3 upon eRF1 on decoding UAA/UAG and UGA. As T58 in the TASNIKS motif was observed to interact with the 15-mer RNA [88]and the NIKS motif had also been implicated in ribosome binding [61], interactions between helix a2 and the ribosome are highly possible. Furthermore, P41 and P89, which may be critical for the formation of the b-turns that connect the core b-sheet to helices a2 and a3, were also found to affect stop codon recognition [60]. This hints at a higher than anticipated degree of complexity in the stop codon decoding mechanism of eRF1. Based on our current data, we have reinforced that a degree of complexity is involved in the stop codon decoding mechanism that requires the participation of different

regions to modulate conformational changes that serves as a requisite for translation termination to occur.

Understanding of the mechanism of translation termination in eukaryotes will have to rely on piercing together biochemical, structural and genetics data from different studies. Our structural study of N-domain led us to suggest that eRF1 might decode different stop codons by adopting distinct GTS loop conformations, thus implying direct access of the GTS loop to the stop codon. In addition, the data on N-domain-RNA interactions [88] have shown that N-domain potentially interacts with Helix 44 of 18S rRNA.

Based on these results, we propose a structural model that encompasses currently known interactions between N-domain of eRF1 and the A site of eukaryotic ribosome (**Fig. 2.4**). The model shows that it is possible for the GTS loop to contact the stop codon while helix a1 is positioned next to the decoding region of Helix 44. Although helix a1 is not in the exact position to interact with Helix 44, a slight forward movement of the stop codon towards the P site will compensate for this discrepancy. Interestingly, it was reported that a 2-nt toe-print shift occurs when the eRF1·eRF3·GTP complex binds to the pre-TC [99]. On the other side of N-domain, the side-chains of A53, N61, R65, R68 and Q79 are facing 18S rRNA. Residues R65 and R68 affect the binding of eRF1 to the ribosome (14), while each of the point mutants, A53K, N61K and Q79K/ R, was shown to substantially reduce the level of stop codon readthrough in comparison to wild-type, indicating enhanced ribosome binding due to the lysine or arginine substitution [100]. In our model, the GTS loop is close enough to the stop codon to allow photoactivatable cross-linking with the second and third stop codon positions [80]. In the eRF1/pre-TC cross-linking experiments, the KSR loop (positions 63–65) and V66 were suggested to be in contact with the first stop codon position (16,17). Although within margins of cross-linking experiments, in our model these residues are not located in the direct proximity of the uridine of the stop codon, thereby requiring further experiments to resolve. With the orientation of N-domain in our model, a hinge motion between N-domain and M-domain would allow the GGQ motif to reach the 30-CAA tail of P-site tRNA as had been suggested earlier [67, 80], while C-domain would be required to move away from helix a1 (**Fig. 2.4**). Hence, a major domain rearrangement between N-domain and C-domain is likely to occur in translation termination during which N-domain accommodates itself into the A site.

CHAPTER III

CHAPTER III

SRY-Related High Mobility Group (SOX-HMG)

3.1 Introduction

3.1.1 Transcription

Transcription is the process in which an RNA product is produced from the DNA, which would then be translated into proteins. The production of this RNA transcript is essential in ensuring that individual genetic information from the DNA is converted into proteins thus resulting in the phenotype of an individual [101]. Indeed the regulation of this process is essential in development of organisms as it determines the final product/outcome of the transcribed protein. Regulation ensures that the correct genes are transcribed in response to specific signals [102]. In this context, transcription factors are lead players that bind to specific DNA sequences in a gene [103]. In order to execute this, certain features are required of these factors that would enable them to bind DNA sequence specifically [101, 104]. This shall be discussed in the following sections pertaining to specific transcription factors.

3.1.2 Sox transcription factors as possible drug targets

3.1.2.1. SRY-related HMG Box (Sox)

The SRY gene is involved in male sex determination during the developmental process, in particular the HMG box region of the gene [105]. HMG box proteins are categorized into groups based on the number of HMG boxes present in a particular protein, either many or a single [101]. Proteins with many HMG boxes bind DNA based on the structural features rather than sequence specificity, and proteins with single HMG boxes are sequence specific and bind in the minor groove rather than the major and this is accompanied by bending of the DNA helix [106]. These sequence specific interactions are based on hydrophobic interactions between DNA and HMG box of SRY whereby α-helix is inserted into the widened minor groove causing widening [107].

Sox proteins are SRY related HMG box proteins which share the SRY motif similarity in the HMG box domain, which is the DNA binding domain of SRY. **Figure 3.1** depicts the sequences of several Sox proteins and the SRY protein. The sequence similarity

is 60% of the original 79 amino acid in the HMG domain [103, 108]. The ability of Sox proteins to bind to DNA's extends beyond just the binding capacity of its HMG domain [108] (**Fig. 3.2**) but also to its specificity in recognizing the heptameric sequence 5'-(A/T)(A/T)CAA(A/T)G-3' [109]. Based on the amino acid sequence similarity in the HMG box, a similar L shape has been suggested for almost all Sox proteins as well as SRY proteins in the HMG box domain, with high content of α-helices shown through circular dichroism and secondary structure prediction [110], and its hydrophobic core maintains the structure of the protein.

The DNA binding region of the protein is at this hydrophobic core, which is highly conserved [106]. Interestingly, the conformation of the protein remains fairly unchanged upon binding to its target DNA, rather the DNA undergoes changes in the form of bending to allow it to fit into the concave of the binding interface of Sox proteins. It has been observed that this binding occurs in the minor groove of the DNA, followed by widening of the groove [111, 112]. Most transcription factors bind to the DNA via the major groove, the deviation of Sox from the norm suggest that it not only acts singularly but cooperatively with other transcription factors by forming biologically active complexes in order to carry out its function [113].

The DNA binding motif in the HMG box plays an essential role in human development as well as disease, with both the major and minor wings acting as templates. It has been shown that free HMG box proteins and Sox protein exist with partial disorder, and complete protein structure is achieved only upon binding to the DNA of interest [114]. The hypothesis then further argues that this coupled or mutual dependence in achieving a specific architectural structure governs the multiple biological functions of HMG as well as Sox proteins [114, 115].

Of the SRY protein, Sox2 specifically was chosen on the basis of its critical roles in stem cell and cancer biology. Sox2 is required for the maintenance of pluripotency and self renewal of embryonic stem (ES) cells [116, 117]. Consistently, knockdown of Sox2 results in the loss of the undifferentiated state. By featuring in a cocktail of four transcription factors required in generating induced pluripotent stem (iPS) cells, Sox2 gained further prominence [118]. To induce and maintain pluripotency, Sox2 directly interacts with Oct4 when bound to its DNA targets mediated by a small and charged protein interaction

surface. Apart from this, the oncogenic potential of Sox2 was also recognized in a study where after elevated expression levels were detected in several tumors such as lung cancer, gastric carcinoma, malignant glioma, and in breast cancer [119]. It has therefore been hypothesized that the aberrant up-regulation of Sox2 promotes tumorigenesis by stimulating self-renewal, dedifferentiation, proliferation, and cell survival reminiscent of its role in stem cell biology [119].

Transcription factors that constitute DNA binding domains are considered as too difficult to validate as possible drug targets, since they usually lack the hydrophobic and deep ligand binding pockets characteristic for enzymes and cell surface receptors. Instead, transcription factors possess an extended and positively charged recognition interface to dock to polyanionic DNA elements. Moreover, the DNA binding domains are often subjected to structural rearrangements upon DNA binding, whereas the ligand binding pockets of enzymes are largely preformed. Despite these difficulties, several studies demonstrated that small molecules could directly target transcription factors [120, 121].

The processes by which transcription factors locate their target sites are facilitated by three main mechanisms:

ii) sliding,

iii) jumping

iv) and intersegmental transfer [122] [123] [124]

Non-specific binding to the DNA is referred to as the first mechanism, sliding. Jumping entails dissociation of the protein from DNA into the solution followed by re-binding along the DNA. The third mechanism, intersegmental transfer is translocation of the protein between 2 different sites, on two separate DNA fragments, thus forming a ‘bridged’ complex [125].

Protein-DNA complexes regulate transcription through the interaction with downstream transcriptional machinery. In the eukaryotic system, complexes of multiple transcription factors forming a regulatory network, referred to as combinatorial control, accomplish the regulation of transcript [101]. Studies in this field indicate that combinatorial control is achieved by the adaptable characteristic of both the protein and DNA binding surfaces. This adaptability is the driving force toward different transcription factors acting on a single promoter DNA to achieve regulation of the transcription process

[112]. Several HMG boxes have been characterized structurally, with the DNA bound to the protein, where it was observed that the protein takes on a 3-bundle L-shape. Helices 1 and 2 form the major wing and helix 3 the minor wing. These DNA-bound structures of the protein were observed to bind to the minor groove of the DNA, where the DNA bends away from the protein towards the major groove. The DNA in turn is bent through a series of contacts on the inside of the L-shaped HMG box fold and weak interactions with the side chain of adjacent adenines were observed [126].

```
SOX1     MYSMMMETDLHSP-------------------GGAQAPTNLSGPAGARGGGGGGGGGGGG 41
SOX2     MYN-MMETELKPP-------------------GPQQAS-------GGGGGGGNATAAATG 33
SOX3     MYS-LLETELKNPVGPPTPAAGTGVPAAPGAAGKSGANPAGGANAGNGGSGGANGGGGGG 59
SOX14    ------------------------------------------------------------
SRY      ------------------------------------------------------------

SOX1     GGTKANQDRVKRPMNAFMVWSRGQRRKMAQENPKMHNSEISKRLGAEWKVMSEAEKRPFI 101
SOX2     GNQKNSPDRVKRPMNAFMVWSRGQRRKMAQENPKMHNSEISKRLGAEWKLLSETEKRPFI 93
SOX3     GGGGSDQDRVKRPMNAFMVWSRGQRRKMALENPKMHNSEISKRLGADWKLLTDAEKRPFI 119
SOX14    --MSKPSDHIKRPMNAFMVWSRGQRRKMAQENPKMHNSEISKRLGAEWKLLSEAEKRPYI 58
SRY      -----MEGHVKRPMNAFMVWSRGERHKLAQQNPSMQNTEISKQLGCRWKSLTEAEKRPFF 55
         .::**************:*:*:* :**.*:*:****:**. ** ::::****::

SOX1     DEAKRLRALHMKEHPDYKYRPRRKTKT-----------------LLKKDKYSLAGGLLA 143
SOX2     DEAKRLRALHMKEHPDYKYRPRRKTKT-----------------LMKKDKYTLPGGLLA 135
SOX3     DEAKRLRAVHMKEYPDYKYRPRRKTKT-----------------LLKKDKYSLPGGLLP 161
SOX14    DEAKRLRAQHMKEHPDYKYRPRRKPKN-----------------LLKKDRYVFP----L 96
SRY      QEAQRLKILHREKYPNYKYQPHRRAKVSQRSGILQPAVASTKLYNLLQWDRNPHAITYRQ 115
```

Figure 3.1. Multiple sequence alignment of selected Sox proteins domains. Alignment of amino acid sequence of selected Sox-high-mobility group (HMG) domains was prepared using the software ClustalW. The Sox subfamilies are indicated to the left.

3.1.2.2 PAX

Another homeodomain transcription factor of interest in this study is the Pax protein. These factors constitute 422 amino acids that represent conserved DNA binding domains commonly known as the paired domain (**Fig. 3.2**). In addition to this several of the Pax proteins also have the homeodomain and or an 8 amino acid element called the octapeptide. These Pax factors are able to bind to DNA's based on the presence of either a single paired domain or a combination of both the paired domain and homeodomain. The binding affinity is greatly affected by the presence of the domains in the factors [127]. The structural similarity between both domains results in a DNA binding mechanism that is related to one another, via a helix-turn-helix motif.

Pax proteins play a crucial role in the development process especially in the development of the central nervous system and eye development [128]. The activation of the δ-crystalin gene (lens specific gene) is influenced by the cooperative interactions of Sox2 and Pax6. This cooperativity results in the lens placode being formed, with Pax6 being the main determining factor [127]. Pax6, which is the protein of interest in this study, is directly related to the eye development in an organism and mutations present in this protein results in severe eye defects known as aniridia [127]. Studies have shown that Pax6 initiates lens development through a complex formation with Sox2 on DC5. The combination of both these proteins in activating the DC5 element infers that the complex formed is crucial in the process of lens differentiation. The complex formed points towards a cooperativity mechanism where the presence of Sox2 enhances the binding affinity of Pax6 to DC5, resulting in a ternary complex that is stable and with a higher binding affinity in comparison to the binary complex of each protein [127, 129]. The domain of interest is the paired domain as it has been shown that mutations in this region cause the above-mentioned disease.

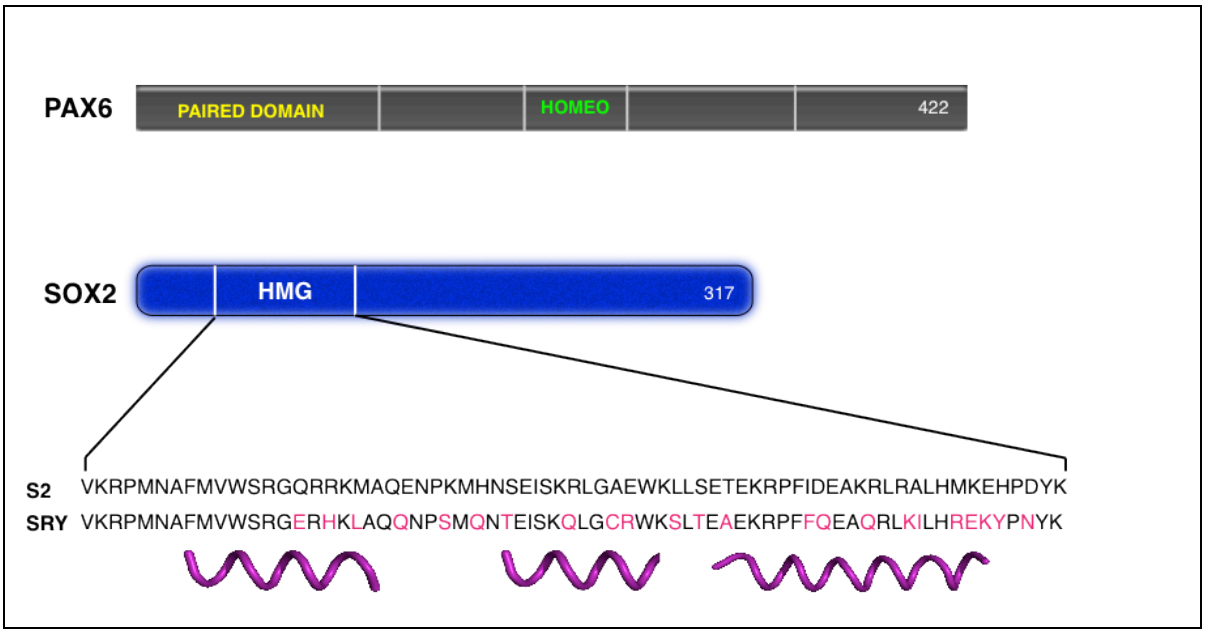


Figure 3.2. Schematic representation of Pax6 and Sox2 structures. The last amino acid is numbered accordingly. The domains of interest in this study are the Paired domain for Pax6 and the HMG domain for Sox2. The purple wires represent helices present in the SRY and Sox2 proteins [129].

3.1.3 Overview of Project

The specificity and functional characteristic of a particular Sox factor is dependent on its co-partnering transcription factor. Recently, many Sox proteins have been found to be involved in human cancers [130, 131]. Chen *et al.* have reported through various techniques of study, that Sox2 promotes cell proliferation and tumorigenesis. This occurs via regulation of the CCND1 gene in breast cancer cells during the G1/S transition phase. Apart from its involvement in tumorigenesis, Sox2 was also discovered to be in a synergistic partnership with β-catenin during the above mentioned transcription regulation [119].

In the light of the importance of Sox2 in stem cell and cancer biology, we envisage small molecule inhibitors of Sox2 to have two areas of application: (i) as anticancer drugs and (ii) as tools to direct differentiation for tissue engineering. As stated above, DNA binding by Sox2 is mediated by 80 residue high mobility group (HMG) domain that binds to the consensus C(T/A)TTG(T/A)(T/A) motif. Its angular inner surface binds to the minor groove of the DNA and inserts a phenylalanine-methionine wedge into DNA base pairs inducing a L-shaped structure composed of flexible major and minor wings that are subject to some structural rearrangements upon DNA binding. This structural flexibility may pose a challenge due to the absence of a preformed small-molecule binding pocket and context-dependent conformational adaptability. Alternatively, we hypothesized that the flexibility and "clamp"-like architecture of the Sox-HMG domain might enable it to wrap around inhibitor molecules in an induced-fit type of mechanism.

3.1.3.1 Sox2_HMG binding with CCND1.

We have proceeded to try and elucidate the binding surface between Sox2 and the CCND1 gene through NMR titration studies. Through this, we have extracted the amino acids affected by the formation of the binary complex as well as observed that the protein achieves structural stability upon binding to the CCND1 gene as previously hypothesized for HMG box proteins [107]. Analysis of the imino proton NMR spectra showed conformational changes experienced by the DNA upon binding with Sox2-HMG. Structural implications of these changes suggest that protein-DNA complex undergo a slow to intermediate exchange, as depicted of the imino region in the 1D-^{1}H NMR spectra.

3.1.3.2. Sox2-HMG binding with Dawson POM

In this study, we identified a Dawson polyoxometalate as a nanomolar inhibitor of the DNA binding activity of Sox2 (**Fig. 3.3**). This binding is direct and selective for HMG domains in comparison to other DNA binding domains. We further mapped the interaction surface of Sox2 with Dawson-POM by NMR and *in silico* docking for Dawson-POM only.

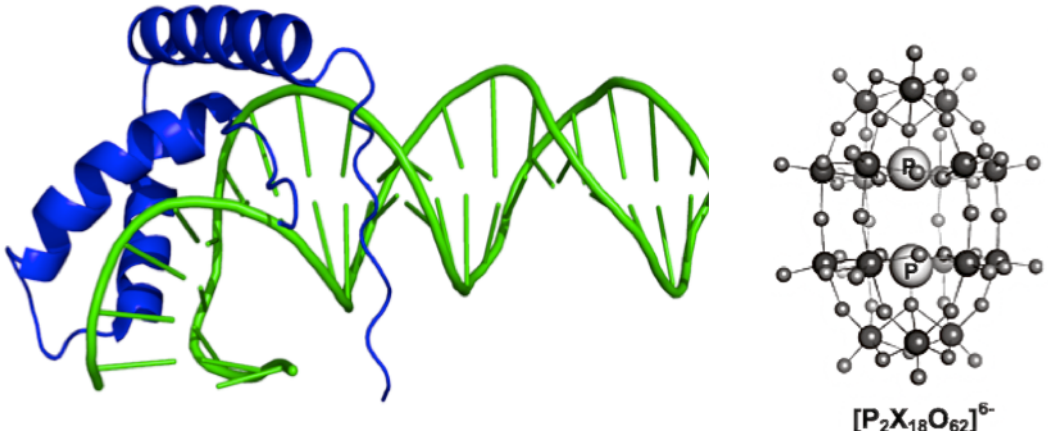


Figure 3.3. Sox2-HMG with Dawson-POM A) Sox2-HMG binding with DNA (PDB: 1GT0) B) Ball and stick representation of the Dawson POMs. Small light gray spheres are oxygen atoms and the bigger dark spheres (X) are transition metals like Mo and W. The central phosphate atoms are labeled.

3.1.3.3. Sox2-HMG_DC5_Pax6 binding

It has been observed that Sox2_DC5_PAX6 cooperate to form a more stable ternary complex in comparison to the individual binary complexes of the protein-DNA. It was also shown that modification on the DC5 element, which increased the distance between the binding sites of the two proteins, affected the affinity by reducing it to a low-mobility form complex. Thus suggesting that the cooperativity mechanism experienced by the complex is DNA sequence dependent. We have observed that the presence of Pax6 in the ternary complex results in stronger binding in the existing binary complex and a further change in the confirmation of DC5 suggest binding of Pax6 to the DC5 element itself.

3.2 Materials

3.2.1 DNA duplexes

The complementary DNA strands were annealed in a buffer containing 20 mM Tris pH 8.0, 50 mM KCl and 50 mM MgCl2 in a PCR thermocycler by initially ramping to a temperature of 95ºC for 5 mins followed by a slow cooling to 4ºC at the rate of (0.5ºC/sec).

CCND1 gene

5′ CGGG**CTTTGAT**CTT 3′

3′ GCCC**GAAACTA**GAA 5′

The red bold highlighted nucleic acids are the Sox2 binding sites.

DC5

5′TT**CATTGTT**GT**TGCTCACCTACCATGGATC**C 3′

3′AA**GTAACAA**CA**ACGAGTGGATGGTACCTAG**G 5′

The red bold highlighted nucleic acids are the Sox2 binding sites, where as blue are the Pax6 binding sites.

3.2.2 Growth media and antibiotics

As described in section 2.2.2

3.3 Methods

3.3.1 Sox2 expression and purification

The HMG domain of Sox2 spanning residues 33-141 (mSox2-HMG) of the full length mouse protein has been cloned into pETG20A expression vector, protein is expressed as His_6thrx-mSox2HMG (*i.e* fusion tag is histidine-thioredoxin). This plasmid was then transformed into an expression strain BL21DE3 using regular transformation protocols. An overnight starter culture was used to inoculate a 10 ml Luria Broth (LB) medium culture supplemented with ampicillin 100 μg/ml, which was incubated at 37 °C/180 rpm. Expression was induced with 0.5 mM IPTG at 25 ºC, overnight. The cells were then harvested by centrifugation at 10 000 rpm (JLA 9.1, Beckmann) at 4ºC for 20 mins and

cells were resuspended in 50mM Tris-Hcl, pH 8.0, 150mM NaCl, lysed by sonication and the lysate was clarified by centrifugation at 18 000 rpm (JA 25.5, Beckmann) for one hour. The lysate was then purified by metal affinity using Ni-NTA resin equilibrated with 50mM Tris-Hcl, pH 8.0, 150mM NaCl. Protein was eluted in the same buffer containing 300mM Imidazole. Eluted protein His_6thrx-mSox2HMG was desalted with 50mM Tris-HCl pH 8, 100mM NaCl on a desalting/GF column in order to remove imidazole prior to TEV digestion. TEV-digestion of Tag was performed at 4ºC for approximately 16 hours (substrate-to-enzyme ratio of 100:1)

Sox2-HMG was purified by cation-exchange chromatography (Resource S column equilibrated with 50mM Tris-HCl pH 8, 100mM NaCl) in order to remove the his_6thrx tag and TEV from the digestion mixture. Sox2-HMG was eluted using a linear gradient ranging from 100mM NaCl to 1.0 M NaCl in Buffer A. The protein was subjected to a final purification step using HiPrep Superdex-75 gel filtration column in the same buffer. Fractions containing Sox2-HMG were pooled and concentrated to ~10 mg/ml. The purity of the collected fraction was verified in a SDS PAGE gel. The protein samples were then concentrated using Sartorius filter tubes with 10 kD cut-off (Sartorius Stedim) as the molecular weight of the protein is 12.6 kD. The NMR samples consisted of 90 % H_2O/10 % D_2O, with 5 % DSS as internal standard.

3.3.2 Circular Dichroism Spectroscopy

All CD experiments were recorded at the pH of the corresponding NMR samples using a Chirascan CD spectrometer (Applied Photophysics, UK). Near and far-UV measurements (200-320 nm) were both performed using a quartz cell with a path length of 0.01cm and the temperature was maintained at 25 °C. CD spectra were acquired using sample concentrations of 100 μM in NMR buffer (50 mM K_2HPO_4/KH_2PO_4, pH of 7.0, 100 mM NaCl buffer). Three replicates were acquired and averaged. All spectra were corrected against the buffer baseline and processing of spectra was done using the available software Chirascan. Similarly, the DNA duplexes were prepared in NMR buffer and recorded independent of the protein. In each experiment, the sample buffer as baseline, protein only, oligonucleotide only and molar ratios of protein: oligonucleotide were acquired to allow for the detection of changes in the CD spectrum, respectively.

3.3.3 NMR sample preparation

Solution state NMR has become one of the most powerful tools in studying the interactions between macromolecules and various ligands. The most fundamental feature is the chemical shifts (resonance position), which are characteristic of the changes that the environment poses on the free and bound states of the protein [132, 133]. The solutions used for this study consisted of 0.65 mM ^{13}C, ^{15}N isotope-labeled Sox2-HMG in 90% H_2O/10% D_2O. All samples used for the 2D [^{15}N,^{1}H]-TROSY , and three dimensional HNCA, HNCACB and CBCAcoNH were adjusted to a pH of 7.0 in 50 mM K_2HPO_4/KH_2PO_4, 100 mM NaCl buffer.

3.3.4 NMR data acquisition and processing

NMR experiments were performed on a Bruker AVANCE II 600 and 700 MHz NMR spectrometer equipped with four RF channels and a 5 mm z-gradient TCI cryoprobe. The spectra were collected at a regulated temperature of 298 K, and sweep widths for ^{1}H and ^{15}N were 9804 and 2412 Hz, respectively. The residual HDO resonance signal was suppressed with presaturation. A combination of experiments was used to derive the assignments of the backbone for Sox2-HMG. ^{1}H and ^{15}N resonances observed from the TROSY experiments were correlated with their corresponding inter- and intraresidue spin systems from 3D experiments, namely, HNCA, HNCACB, and CBCAcoNH, to sequentially correlate the amino acids. Data were processed using TopSpin 2.1, and the chemical shifts were referenced directly (^{1}H) to the frequency of DSS. Peak picking and spectral analysis was done using CARA. Approximately 73% of the residues have been assigned, and work on establishing the complete three-dimensional structure is in progress.

3.3.5 Protein DNA interaction

Sox2-DNA binding

CCND1 was dissolved in the same NMR buffer (50 mM K_2HPO_4/KH_2PO_4, 100 mM NaCl buffer, pH 7.0). The pH of the solutions was adjusted to 7.0 using NaOH. The final concentration of the CCND1 stock solution was determined using NanoDrop at 260 nm. Sox2-HMG was added carefully to the DNA element (14-mer) in a stepwise manner to achieve final molar ratios of Sox2:DNA of 0.25,0.5,0.75 and 1.0, with the final DNA

concentration of 0.2 mM. The 1D spectrum of the 14mer-DNA only was used as a reference to observe the perturbations and binding inflicted by Sox2-HMG on DNA. All experiments were conducted at 308 K, after an incubation period of at least 20 minutes. The same procedure was repeated for DC5.

3.3.6 Protein-ligand Interaction

SOX2 inhibition by a Dawson POM

The methods used in the following experiments were based on protocols obtained from NMR of Macromolecules: A Practical Approach [134]. The same buffer conditions were maintained for the Sox2-Dawson POM binding studies. The Dawson POM was prepared as a 50mM stock solution in dimethylsulfoxide-*d6* (DMSO-$_{d6}$). Addition of POM to the free-ligand solution (Sox2 only) was done stepwise to achieve final concentrations of Sox2: POM ratios of 1:0.25, 1:0.5 and 1:1. The Sox2-HMG sample without POM was used as the reference for calculation of the chemical shift perturbation (Δδ). All experiments were conducted at 298K. The final DMSO-$_{d6}$ concentration in the solution was 2% DMSO-$_{d6}$.

3.4 Results and discussion

3.4.1 Backbone Assignments

3D experiments were measured for Sox2-HMG protein, namely HNCA, HNCACB, CBCAcoNH, ^{15}N-resolved NOESY and TOCSY. Sequence specific assignment of the backbone was achieved using HNCA, HNCACB and CBCAcoNH together with 2D [^{15}N, ^{1}H]-TROSY. **Figure 3.5** illustrate the 2D [^{15}N, ^{1}H]-TROSY of Sox2-HMG in 50 mM phosphate buffer, which was used as the reference spectrum in the sequential backbone assignment.

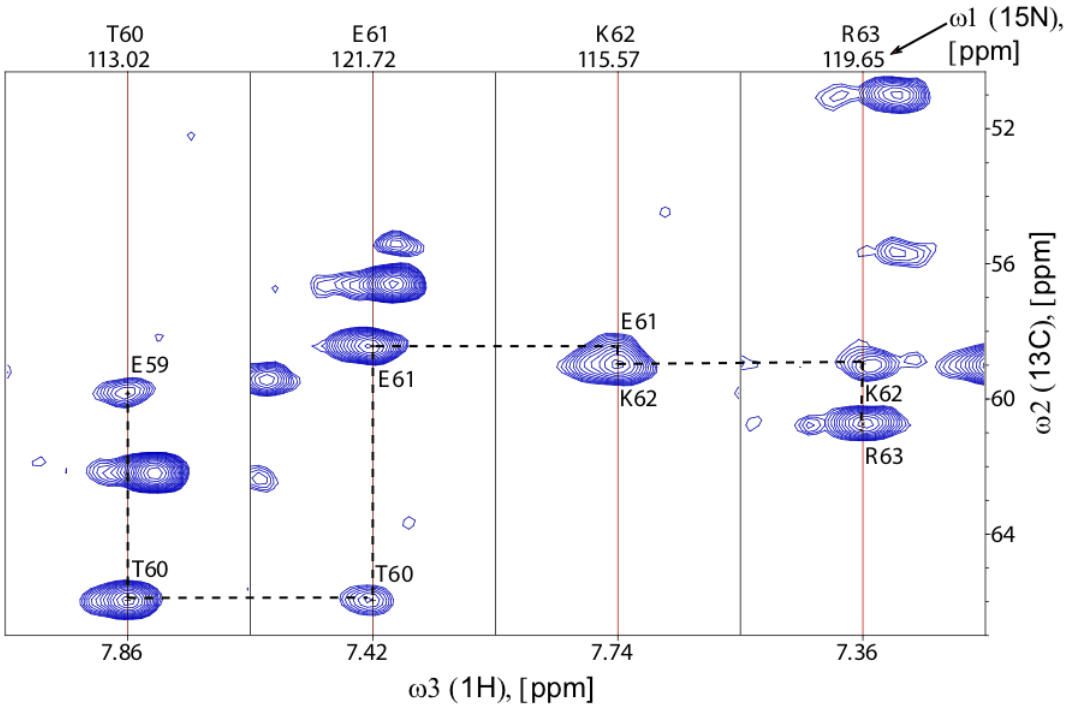


Figure 3.4. [$\omega_2(^{13}C)$, $\omega_3(^{1}H)$] strips from a 3D HNCA spectrum. The strips were taken at the ^{15}N chemical shifts (indicated at the top of the strips) of residues T60–R63 and are centered about the corresponding $^{1}H^{N}$ chemical shifts. At the top of each strip, the sequence-specific assignment is indicated by the one-letter amino acid symbol and the sequence position. Horizontal and vertical lines connect the intraresidual and sequential HNCA connectivities and, thus, outline the sequential assignment pathway.

The HNCA depicts the $^{13}C^{\alpha}$ frequency of the same residue ($^{13}C^{\alpha}_{i}$) and of the preceding residue ($^{13}C^{\alpha}_{i-1}$), corresponding to the amide ^{1}H-^{15}N peak in the ^{1}H-^{15}N plane of HNCA (residue *i*). HNCACB and CBCAcoNH were used together to assign the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ frequencies. In the CBCAcoNH experiment, the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ of the preceding amide residue is observed, whereas in the HNCACB, the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ of the amide pair and its preceding residue are observed (**Fig. 3.4**). Therefore, by combining these 3 experiments, the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ were unambiguously assigned.

The assignment information gathered from here was then combined to determine the sequential relationship of each residue. **Figure 3.5** shows the [^{15}N, ^{1}H]-TROSY spectrum of Sox2-HMG, with the peaks assigned to the residue number in the sequence. Based on the sequence of Sox2-HMG, of the 114 amino acids present, 8 correspond to prolines, which would not be observed in the [^{15}N, ^{1}H]-TROSY spectra. Two additional peaks are expected, corresponding to W13 and W41 side chains (numbering according to

PDB ID:1GT0), 12 peaks corresponding to the asparagine side chains and 6 peaks to glutamine side chains. In total, 126 peaks are expected where 73% of residues have been assigned. The central region of the spectrum is expanded as an inset. It can be seen that peak overlap is predominant and poses as a difficulty in assignment. Secondary structure of proteins can be predicted based on the 13C', $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ deviations from a reference random coil state [135, 136]. The secondary structure of Sox2-HMG was therefore predicted using the changes in chemical shifts of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts from the backbone sequential assignment relative to the reference random coil values using TALOS+ [90].

From the results obtained via TALOS+, 3 α-helices were observed ranging from residue A9-Q23, E32-L43, and E46-K65. However, there were no β-sheets observed. These results were then compared to the existing solution structure of the ternary complex of Sox2-Oct1-DNA. Both proteins show good similarities/overlap. In addition to the conventional restraints, N-H Residual Dipolar Coupling was measured to determine and refine the 3-dimensional structure of this protein (data not included).

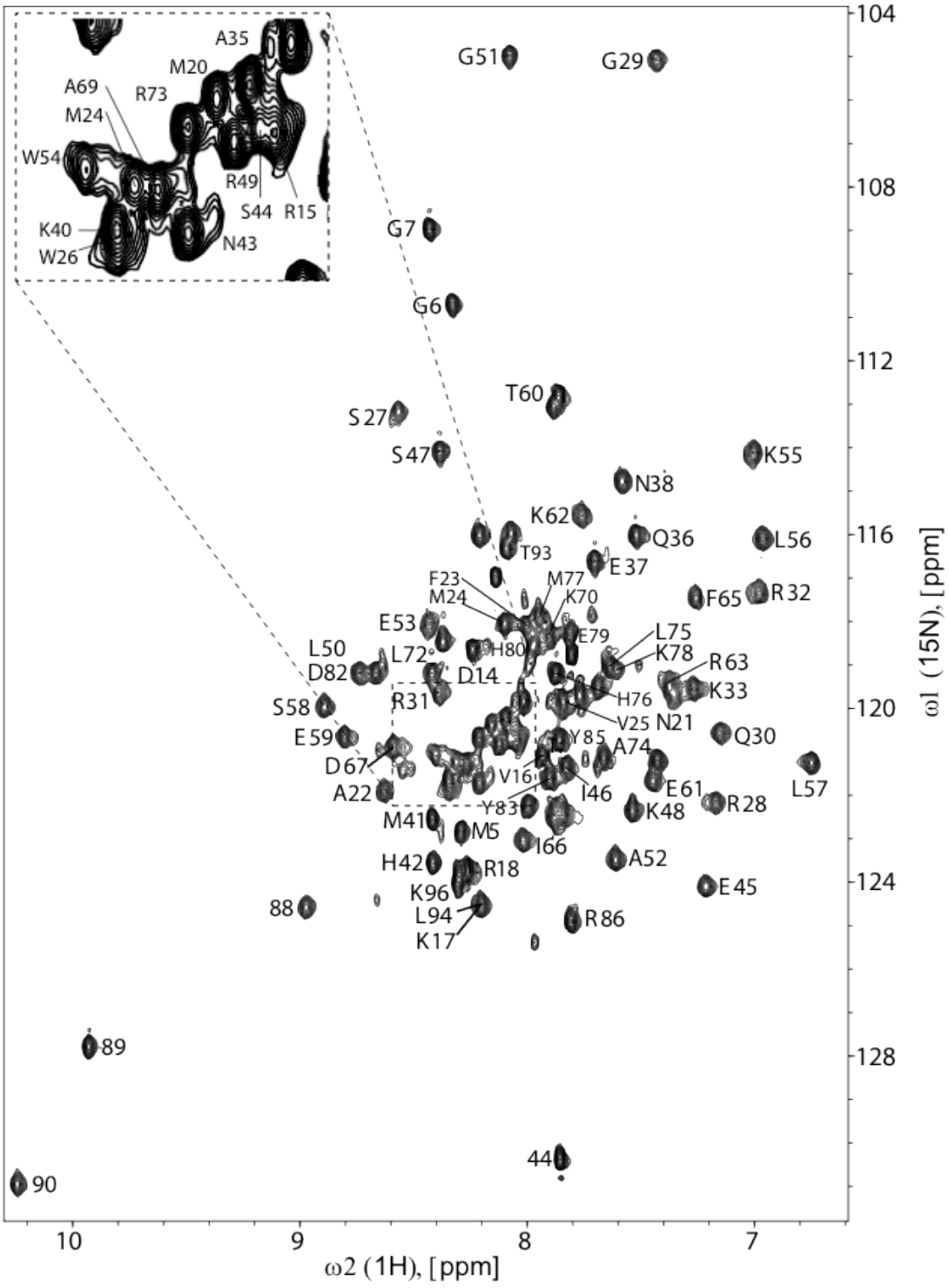


Figure 3.5. 2D [^{15}N,^{1}H]-TROSY of Sox2-HMG in 50 mM Phosphate buffer, pH 7.0, 100 mM KCl, 90% H_2O/10% D_2O at 298K, using 600 MHz NMR spectrometer. Residue specific assignments of the amide resonances are labeled.

3.4.2 Sox2-HMG binding with Dawson POM

i) Dawson-POM directly inhibits Sox2-HMG but Not Unrelated Transcription Factors

Dawson-POM is reported to selectively bind and inhibit the protein kinase CK2 and certain kinesins [137]. This inhibitory potential of POMs was what impelled us to further study the Sox2-HMG POM interactions. Inhibition was studied using various structurally diverse POMs with Sox2 binding. It was thus determined that the originally identified Dawson-POM (K6[P2Mo18O62]) was most potent among all tested POMs [138].

ITC was performed in an attempt to measure the binding affinity of Dawson-POM to Sox2, however titrations with good curvature were not achievable. This loss of binding curvature in ITC experiments is well documented and is known to occur for ligand with very high binding affinity (1-100nM). It is presumable from the IC50 estimation (98.6±22.1nM) that the binding affinity could potentially be on the lower nanomolar range and hence on the tighter binding end for reliable ITC estimation.

Selectivity studies of Dawson-POM for Sox2-HMG and Pax6 were done by our collaborators. Using EMSA, it was determined that the Dawson-POM inhibits Sox2-HMG DNA binding with an IC50 value of 98.6 +- 22.1 nM, Pax6 paired domain activity however was not affected by the POM even at very high concentrations (**Figure 3.6d**). The varying inhibition profiles of the Dawson-POM to unrelated structural classes of transcription factors suggest that the Dawson-POM is selective for certain transcription factor targets. Verification of binding of POM directly to the Sox2-HMG domain was performed via limited proteolysis assay. Trypsin digestion of the Sox2- HMG revealed that Dawson-POM confers sustained resistance to proteolytic digestion (**Figure 3.6a**). A comparison of the thermal unfolding of the Sox2-HMG domain in the absence and presence of the Dawson-POM revealed differences that can be attributed to a stabilization effect of the POM on the melting profile of Sox2-HMG. The addition of increasing concentrations of the Dawson-POM transformed the melting profile and increased the unfolding transition, suggesting that the Dawson-POM directly binds and stabilizes the Sox2- HMG domain (**Figure 3.6b**) [139].

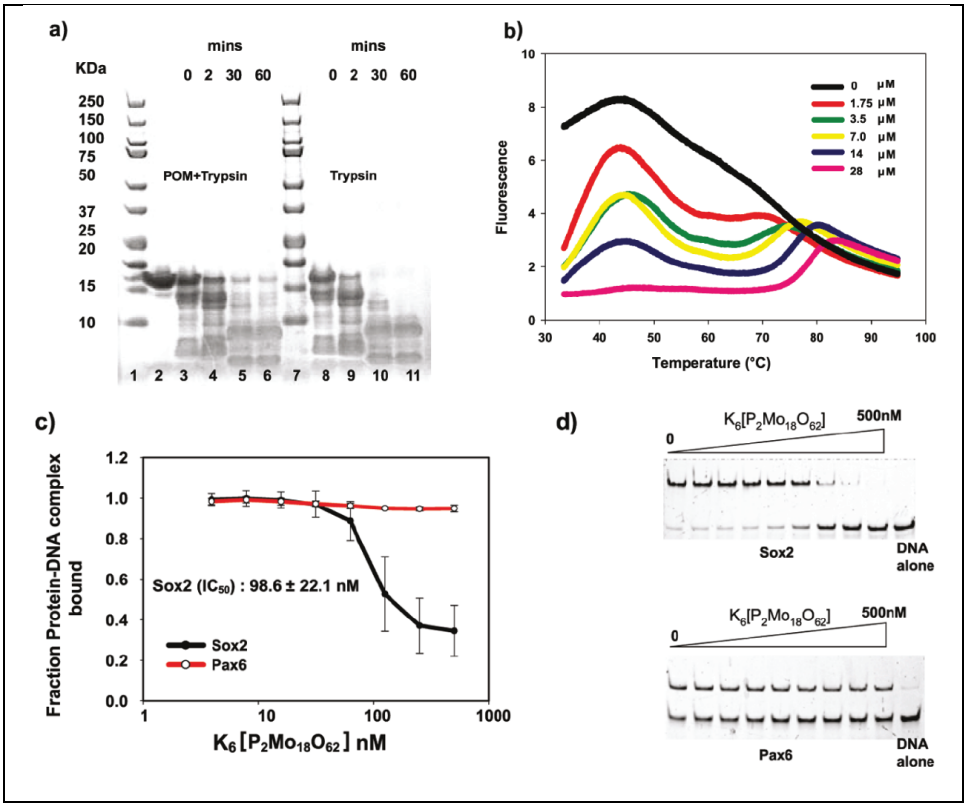


Figure 3.6. Dawson POM interaction with Sox2-HMG domain. (a) Limited proteolysis reveals that interaction of Sox2 with Dawson-POM shows resistance to proteolytic digestion by trypsin. Sox2-HMG was incubated with trypsin in the presence (lanes 3-6) and absence (lanes 8-11) of the Dawson-POM. Reactions were stopped after different time points and analyzed by 4-12% SDS-PAGE. Lanes 1 and 7 contain markers, and lane 2 contains the Sox2-HMG incubated without trypsin. (b) Thermal melting profiles of Sox2-HMG monitored in the presence of sypro-orange at increasing concentrations of the Dawson-POM. As the POM concentration is increased, the initially high fluorescence at lower temperatures (37-50°C) decreases, suggesting POM-mediated stabilization (Figure adapted from [139]).

ii) Sox2-HMG binding of Dawson-POM by NMR

To understand the inhibitory mechanism and to facilitate the follow-up chemistry aimed at optimizing the selectivity and potency of the Dawson POM, we assigned Sox2-HMG backbone resonances using 2D [^{15}N, ^{1}H]-TROSY spectroscopy and tracked the chemical shift perturbation in Sox2 caused by addition of POM.

POM binding to Sox2-HMG was monitored in a series of [^{15}N, ^{1}H]-TROSY spectra recorded in the absence and presence of the Dawson POM. When the Dawson POM was added to 0.65mM protein, backbone resonances of several residues in Sox2-HMG were significantly perturbed indicating direct and specific interactions (**Fig. 3.7**). Chemical shift

perturbation were quantified by using overall weighted chemical shift values ($\Delta\delta = [\Delta\delta^2HN + (0.2\Delta\delta N)^2]^{1/2}$) and displayed on a Sox2 model derived from the Sox2-Oct1/DNA (PDB: 1GT0) structure. Based on the distribution of the weighted chemical shift values, residues were classified into (i) those that undergo significant chemical shift changes ($\Delta\delta \geq 0.065$ ppm), (ii) those that undergo moderate chemical shift changes ($0.04ppm \geq \Delta\delta < 0.065$ ppm) and (iii) those with low chemical shift changes ($\Delta\delta < 0.04$ ppm) corresponding to low Sox2 inhibition by a Dawson POM (**Fig. 3.7**).

We managed to map out the binding sites of Dawson-POM onto the Sox2-HMG; the mode of interaction involves predominantly electrostatic interaction just outside the DNA binding region. The residues Met7, His29, Glu66 and Asp69 shifted significantly (numbering based on PDB ID:1GT0). Among these, Met7 and His29 are directly involved in DNA binding whereas Glu66 and Asp69 are not in direct proximity to the DNA**.** Among the significantly shifted residues Met7, Glu66 and Asp69 are in spatial proximity whereas His29 is located distantly in loop 1 of the HMG major wing. The adequate positioning of the Dawson-POM allows it to compete with the negatively charged DNA backbone [139].

Buried core residues; W13, A22, K27 and I33 which cause the stacking of helices 1 and 2 of the HMG major wing, were found to be least perturbed upon POM binding, suggesting that structural integrity of the HMG domain is retained upon inhibitor binding. Many of Sox2-HMG residues, exhibited moderate chemical shift changes suggesting global structural perturbation spread via entire structure of Sox2 upon POM binding due to allosteric effects [140, 141].

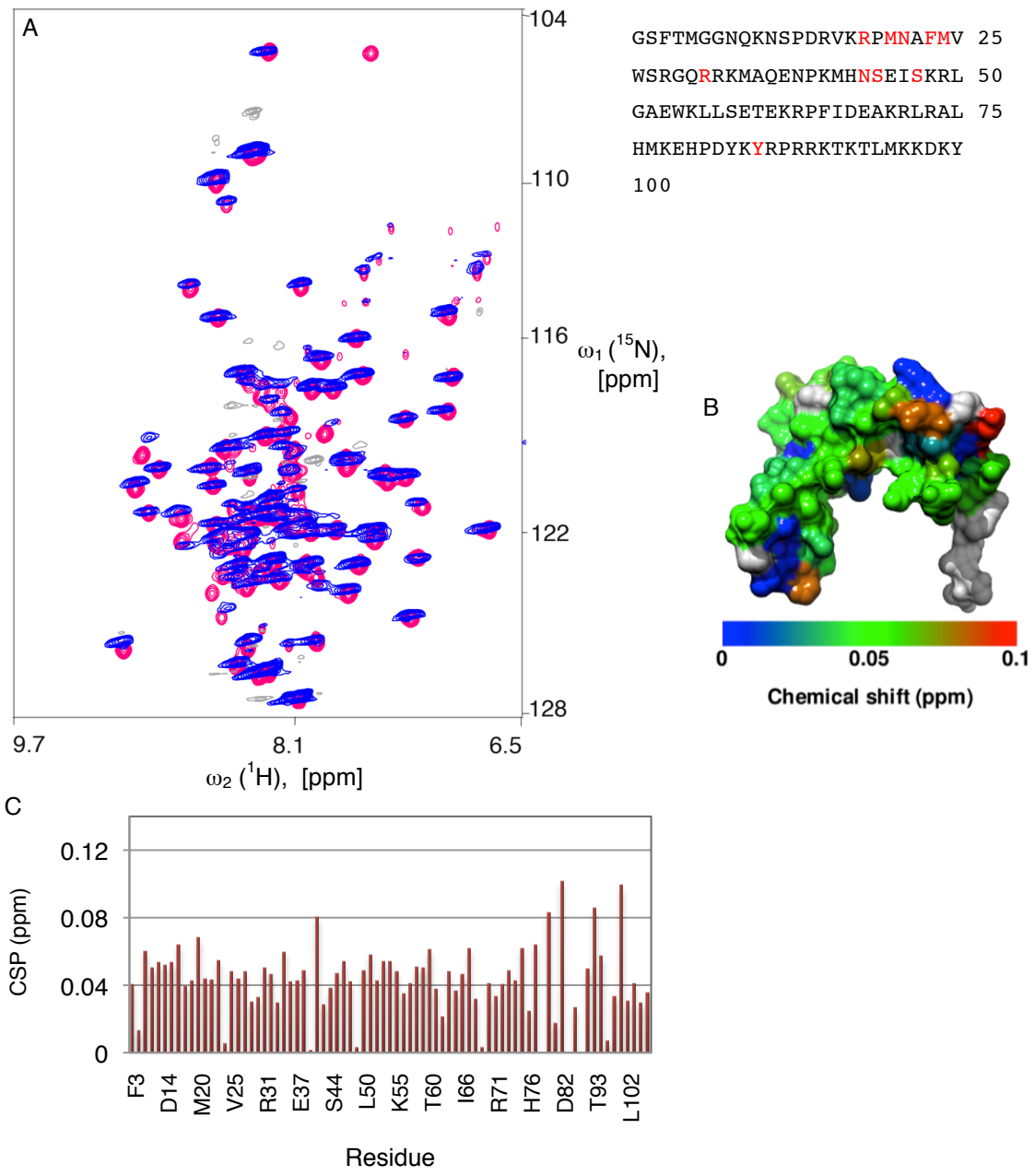


Figure 3.7. Interaction of Sox2-HMG with Dawson-POM analyzed by 2D [^{15}N,^{1}H]-TROSY. A) Superimposed spectra of two-dimensional TROSY spectra of free Sox2 (pink) and Sox2 bound to POM (blue). The peaks that undergo significant shifts upon complex formation namely E79, D82 and H42 are highlighted. B) The weighted change in chemical shift perturbation ($\Delta\delta = [\Delta\delta 2HN + (0.2\Delta\delta N)2]^{1/2}$) obtained from the 2D [^{15}N,^{1}H]-TROSY experiments are mapped on the entire Sox2-HMG surface (1GT0). Residues, which are significantly shifted, are labeled. The colored bar displays the extent of NMR chemical shift perturbation in ppm. Unassigned residues are colored in gray. C) Changes in chemical shift ($\Delta\delta = [\Delta\delta 2HN + (0.2\Delta\delta N)2]^{1/2}$) upon POM Sox2 inhibition by a Dawson POM binding is plotted as a function of the Sox2 amino acid sequence. Sox2-HMG residues involved in DNA binding are colored in red in the amino acid sequence.

iii) Preferential Binding Site of the Dawson-POM on the Sox2-HMG surface

To identify the site of interaction of the Dawson-POM with the Sox2-HMG, autodock searches were undertaken using the docking server methodology using the crystal structure of Sox2-HMG and the CSP data obtained from [^{15}N,^{1}H]-TROSY NMR [142]. As the ligand surface is negatively charged, the calculated docking energy by autodock was found to be very low (-10.6 kcal/mol) corresponding to a binding affinity of ~50 nM (tight binding), which explains the difficulty in obtaining reliable ITC curves as discussed earlier. A cavity formed by the C-terminus of helix-3 and the N-terminal region of the minor wing of Sox2-HMG was identified as a possible interaction site. The positively charged side chains of residues at this cavity, namely Lys4, Arg5, Arg15, His63, and His67 could hydrogen bond or form electrostatic interactions with the negatively charged oxygens of POM. The positively charged residues Lys4, Arg5, and Arg15 exhibit moderate chemical shift changes, whereas His63 and His67 exhibit insignificant perturbation. Solvent accessibility analysis of apo-Sox2-HMG reveals that His63 and His67 are partially buried (13-30% exposure) thus providing a likely reason for the lack of significant chemical perturbation for these residues. Glu66 of the POM binding cavity was most strongly shifted following POM addition (**Figure 3.9**), which is probably caused by unfavorable electrostatic repulsion causing it to structurally reorient with respect to the apo-Sox2-HMG. Asp69 was observed to be strongly perturbed in [^{15}N,^{1}H]-TROSY however missing in the POM docking site. Asp69 is located in the C-terminal of helix 3, which is a region subject to inherent conformational changes due to protein flexibility as shown in NMR studies on apo Sox4 and Sox5 [143, 144]. Upon binding to the DNA, the C-terminus of Sox2-HMG is observed to undergo rearrangement [145], which explains the pronounced perturbation of Asp69 of this region and its structural adjustments accompanying molecular recognition events. The strongly perturbed His29 emanating from loop1 is another residue that is not located in the POM docking. By inspecting the different conformers of the solution structure of Sox17 (PDB: 2YUL), it is evident that the N-terminal region of the minor wing is dynamic adopting conformations where the minor wing approaches loop1 indicating the potential of a cross-talk between the major and minor wings. The Sox2-HMG construct employed here contains an N-terminus extended by 13 amino acids as compared to the homologous Sox17 structure (PDB: 2YUL). Thus, it is conceivable that a longer dynamic Sox2 N-terminus

would come into contact with His29 in loop1 due to conformational reorganization of the HMG wings induced by POM binding. The hydrophobic residues Val3, Met7, Leu59, and Leu62 also shape this cavity of the POM docking site (**Figure 3.8**). Consistently, Met7 exhibited a significant chemical shift in the TROSY experiment while Val3, Leu59, and Leu62 are moderately affected [139].

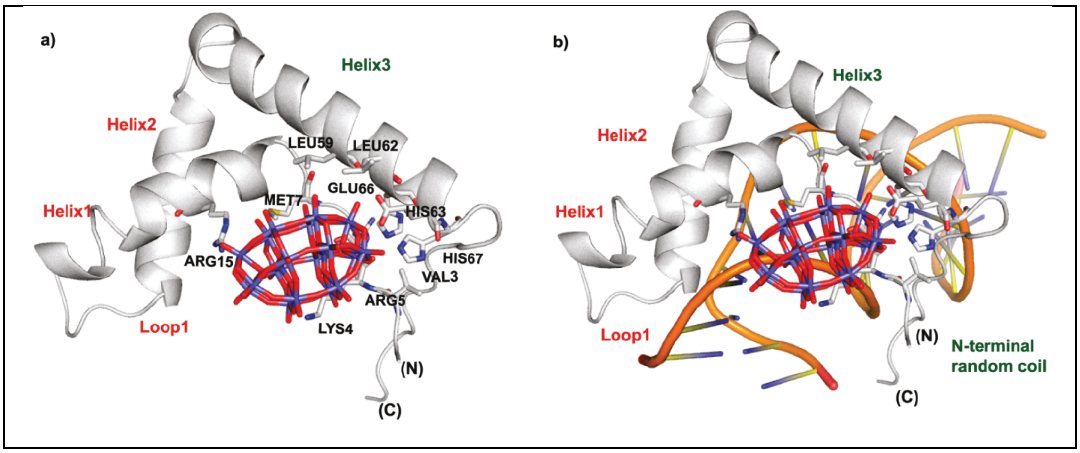


Figure 3.8. Dawson-POM interaction with Sox2-HMG based on docking analyses. A) The lowest energy Sox2-HMG-POM complex structure from autodock searches shows that the POM is positioned within a pocket of the minor wing of Sox2-HMG. Lys4, Arg5, Arg15, His63, and His67 are involved in electrostatic or hydrogen bond interactions. Glu66 could donate hydrogen bonds in a protonated form. Leu52, Leu62, Met7, and Val3 shape the binding cavity. (b) Comparison of the docked model with the Sox2 X-ray structure (PDB: 1GT0) reveals that binding of POM would directly interfere with DNA binding due to charge repulsion Figure adapted from [153]).

iv) Binding Selectivity of POM

The selectivity of POM is most evident when compared against its inhibition against different Sox-HMG members. Multiple sequence alignment reveals consistent differences between the Group F members (Sox7 and Sox18) and Sox2 in at least six out of ten amino-acid positions that has been proposed to bind to the Dawson-POM (**Figure 3.9**), which would account for the observed differential interaction of POM with Sox homologues. It was observed that the group F members Sox7 and Sox18 were the most inert of all Sox-HMG proteins tested against POM, retaining 74 ± 18 % and 61 ± 14 % of their DNA binding activity respectively [138].

Amino acid substitution and its corresponding effect on POM binding with regards to the different Sox-HMG homologues were distinguished. Based on previous NMR and auto-docking studies it was observed that there are consistent differences between Sox7 and Sox18 with Sox2 in six out of ten amino-acid positions known to bind to the Dawson-POM. The substitutions of Val3 and Leu62 (PDB: 1GT0 numbering) in Sox2 by Ile3 and Gln62 in Sox7 and Sox18 led to an altered shape in the hydrophobic Dawson-POM binding cavity. **Figure 3.8** illustrates that Val3 along with Met7, Leu59 and Leu62 are involved in shaping the hydrophobic cavity of Sox2 for the Dawson-POM to bind. Similarly, Lys4 and Arg15 in Sox2 substitution to Arg4 and Lys15 in Sox7 and Sox18 possibly provided an altered electrostatic and hydrogen bonding environment for the POM binding. Thus it can be expected that Dawson-POM is sensitive to binding shape cavity and electrostatics of the Sox-HMG member [138].

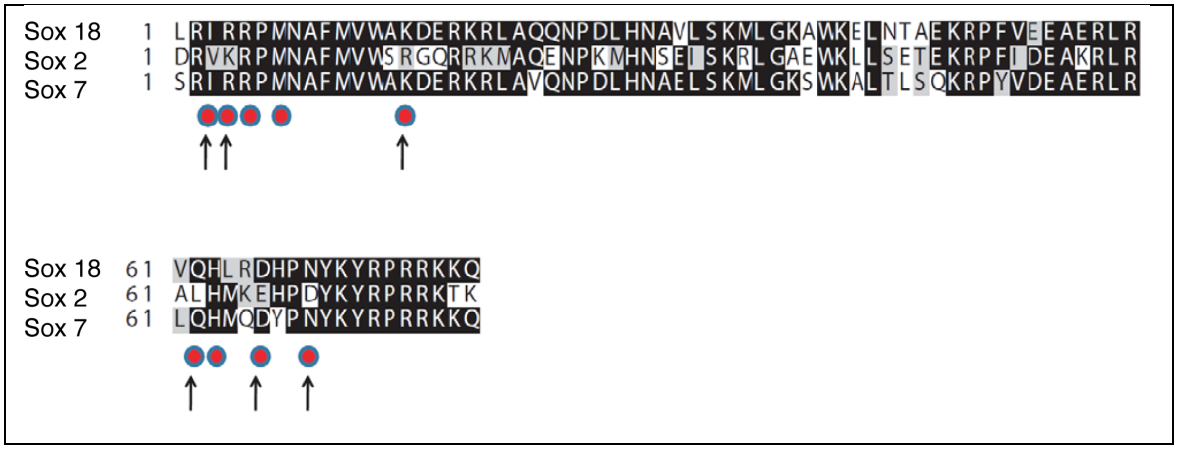


Figure 3.9. Sequence alignment of the core HMG-domain reveals differences between Sox7, Sox18 and Sox2 in 6 out of 10 amino-acid positions proposed to be involved in Dawson-POM binding with Sox2. Residues potentially involved in POM binding based on docking studies with Sox2 are indicated by red dots. Homologous Sox-HMG residue positions involved in POM binding that exhibit consistent differences between the Group F members (Sox7 andSox18) and Sox2 across the sequence alignment are indicated by an arrow. Numbering is based on Sox2 structure from PDB: 1GT0. (Figure adapted [138]).

v) Conclusion

We suggest a model illustrating how the Dawson-POM brings about the inhibition of the Sox2-HMG DNA interaction based on out NMR and docking data. Sox2- POM docked structure with the X-ray structure of the ternary Oct1.Sox2.DNA complex (PDB: 1GT0) illustrates POM binding to this site repels the DNA phosphate backbone while competing

with the Sox2. This binding further induces rearrangements of the N-terminal wing that favors a closed conformation that is incompatible with DNA binding.

A multiple sequence alignment analysis illustrated that majority of the POM contacting residues are conserved within the Sox family. Based on this observation, other Sox proteins (Sox5, Sox7, and Sox17) were tested against Dawson-POM, with a comparable reduction in residual DNA binding observed. Thus, we propose the Dawson-POM as a potent lead scaffold for inhibiting the Sox family and that its selectivity could be fine-tuned by organic modifications. In summary, the inhibitory mechanism of Sox2 by the Dawson-POM demonstrated here could eventually spawn the development of modified classes of POM-based drugs to specifically combat aberrant gene expression.

3.4.3 SOX2-HMG binding with CCND1

i) Selectivity Assay

The selectivity of CCND1 to Sox2-HMG was assessed in a preliminary experiment using electrophoretic mobility shift assays (EMSA) as depicted in **Figure 3.10** (Kamesh Narasimhan *et al*., personal communication). This preliminary study demonstrated that CCND1 binds to Sox2-HMG with a Kd of 11.232 ± 2.47 nM.

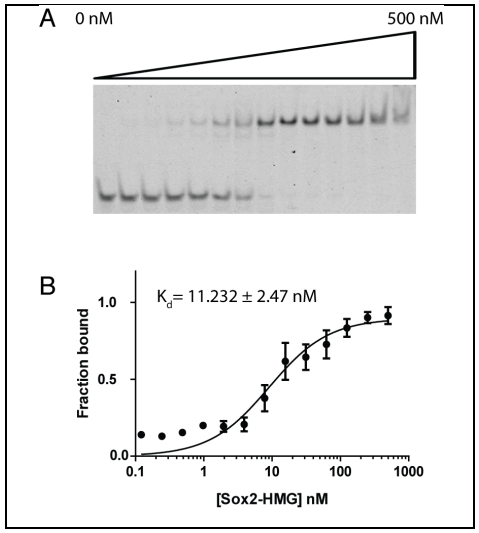


Figure 3.10. A) Varying concentrations of Sox2-HMG were incubated with 1 nM of 5' FAM-labelled CCND1 element to obtain a binding isotherm (protein concentration is zero in the leftmost lane and increases until a final concentration of 500 nM). Bound and unbound probes were separated on a native 12% PAGE gel. B) Apparent dissociation constants estimated by fitting the average bound fraction of at least three experiments using a one-site saturation model in GraphPad Prism and the standard error of the fit are indicated (Kamesh Narasimhan, personal communication).

ii) CD spectroscopy of binary complex

CD spectroscopy was performed on Sox2-HMG to determine its secondary structure as well as possible changes in its secondary structure upon binding to DNA. In this study CCND1 was used as the oligonucleotide of interest to determine specific binding of the protein to DNA. For the protein, both the minima (208 and 222 nm) and the maxima (192) are indicative of α-helices present. In all complex samples, the DNA was slightly in excess to ensure complete binding was achieved. **Figure 3.11** clearly shows that binding is achieved as changes are observed to the protein as well as DNA. As for the oligonucleotide, the positive peak at 275 nm was observed to increase reflecting DNA structural changes, as the protein is not CD active at this wavelength. **Appendix B1** illustrates the CD spectra of the oligonucleotides only and protein only.

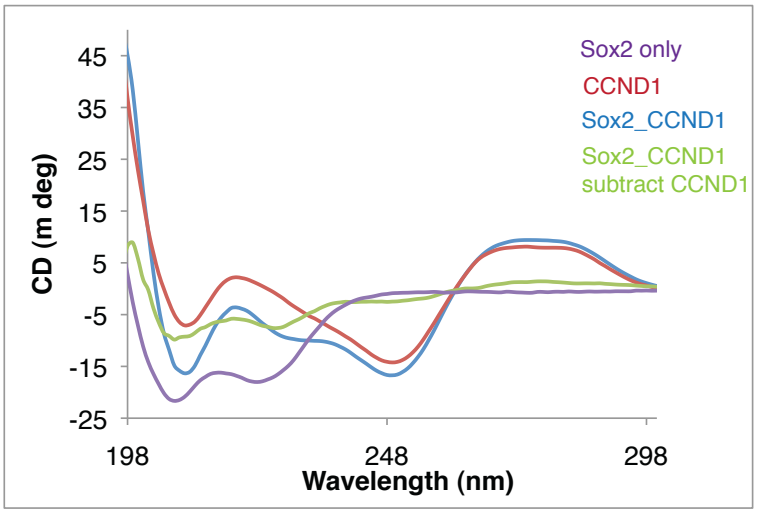


Figure 3.11. Circular dichroism spectroscopy of Sox2-HMG binding to CCND1. Near and far UV spectrum of Sox2-HMG: CCND1 (1:1.2 molar ratio). The lines representing the protein, DNA and complex are color coded on the figure, respectively.

iii) Characterization of the DNA-binding Interface

To further our understanding on the binding and mechanics of the transcription factor Sox2 to the CCND1 gene, NMR monitored titrations were performed using 2D [^{15}N, ^{1}H]-TROSY experiments. Interactions between these two entities were observed via mapping the chemical shift changes that occurred upon addition of protein to the DNA component in increasing amounts. All experiments were recorded using ^{15}N-labeled Sox2 and unlabeled

CCND1 to observe the changes that occur in the protein upon binding; ^{1}H and ^{31}P spectra were used to monitor changes that occurred within the DNA itself. The binding kinetics of Sox2-CCND1 is indicative of being in the slow to intermediate regime as observed by the line broadening. The result of different bound and unbound states are the broadening of peaks, which is observed for much of the spectra besides shifting of several peaks. There were no significant changes when Sox2 was titrated in increasing amounts to the DNA. Therefore, the chemical shift changes presented here represent the difference observed between the free protein and the bound state at a molar ratio of 1:1.2.

A large number of contacts with the DNA are observed involving all regions of the protein, 25 residues in total were affected by the addition of DNA, either via peak broadening or perturbation. **Figure 3.12** depict the extent of changes observed (ppm) for each residue of the transcription factor. The largest chemical shifts (> 0.1 ppm) observed were for K4, G38, K42-L44 and R50 at the random coil of the N-terminal of Sox2 and α-helix 2. However, residues that experience peak broadening are predominantly at the hydrophobic core of the protein comprising the major wing of the protein: R2, A9, F10, Q17, R19, A22. At the turn regions of α-helix 1 and α-helix 2 which are involved in DNA binding, residue N30, S31, I33, K35, L37, A39 were affected. At the minor wing, the only changes were towards the C-terminal region corresponding to K57, A61, M64, and H67. It is interesting to note that the side chain of both the tryptophan residues is affected by the addition of Sox2 to CCND1. Both undergo an upfield shift, however the shift is maintained (negligible shifts) even at increased ratios of CCND1 in the complex. W41 experiences a drastic shift in comparison to W13 of about 0.81 ppm (**Fig. 3.12)**. The perturbations experienced by the side chains indicate that there could be possible rearrangement of the protein following DNA binding to allow interaction to occur. So the pronounced perturbation of the indole peaks most likely illustrates the dynamics of this region and its structural adjustments accompanying molecular recognition. The perturbed residues are depicted in a cartoon diagram on the structure of Sox2 (PDB: 1GT0) **(Fig. 3.12)**, where the changes are color-coded (pink-perturbed, grey-disappeared). Sox2 with only the DNA binding sites highlighted based on the ternary complex 1GT0 and the location of the tryptophan residues relative to this binding site is illustrated in **Figure 3.13**.

The overall changes observed in the [^{15}N, ^{1}H]-TROSY spectra are indicative of

conformational changes that occur upon binding to the DNA. The data presented here illustrates that the perturbed residues experience deviations within the range of 0.1 to 0.2 ppm, and are predominantly in the hydrophobic core of the protein. It is this core that interacts with the DNA by inserting a phenylalanine-methionine wedge into DNA base pairs inducing a L-shaped structure composed of flexible major and minor wings that are subject to some structural rearrangements upon DNA binding [146]. The dispersed CSP observed could be due to structural changes subjected to the protein upon binding to DNA. This observation is supported by earlier studies whereby the Sox2 protein by itself was observed to be rather unstructured and achieves structural stability only upon binding to DNA. This would explain the increased number of residues perturbed instead of just the few expected DNA binding sites. Structural similarities have been reported for SRY and Sox2, and differences between the two proteins lie on the surface amino acid residues. Both proteins were observed to bend DNA to the same extent highlighting the fact that only 4 of DNA-contacting residues are different between the two proteins [112]. **Appendix B2** shows an alignment of the solution structure of Sox2 obtained from NOESY data (PDB: 2LE4) with Sox2 in complex with Oct protein (PDB: 1GTO). As is shown, the conformation of α-helix 1 and α-helix 2 are different in both cases. The perturbed residues observed in our study matches the regions where differences are observed between these two structures. α-Helix 3 is observed to bend in towards the DNA compared to the free protein, which accounts for the change observed.

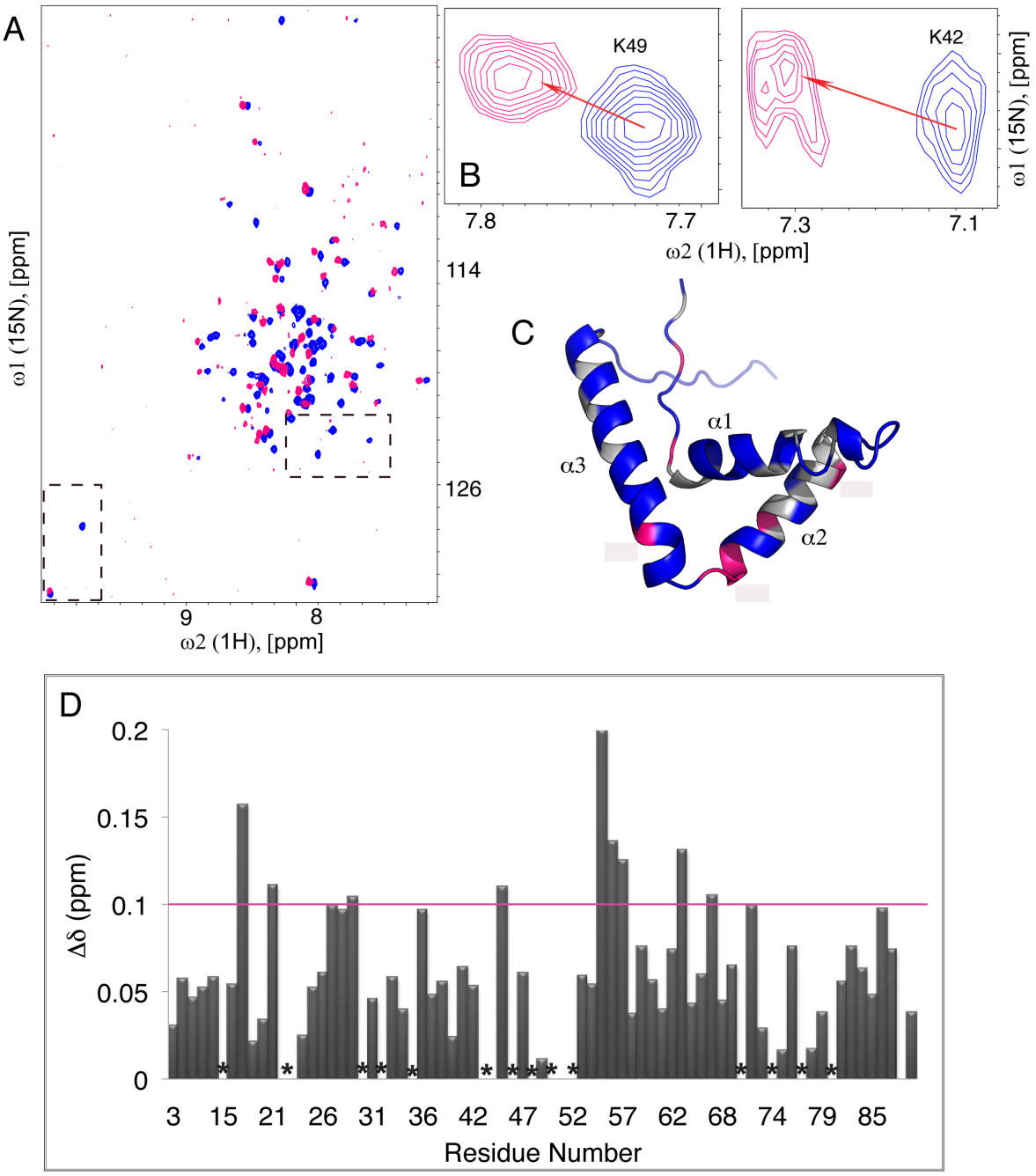


Figure 3.12. Interaction of Sox2-HMG with CCND1 analyzed by 2D [^{15}N,^{1}H]-TROSY. A) Superimposed spectra of two-dimensional TROSY spectra of free Sox2 (blue) and Sox2 bound to CCND1 (pink) at a ratio of 1: 1.2. B) Example of peaks that undergo significant shifts (K42 and K49) are highlighted. C) The weighted change in chemical shift perturbation ($\Delta\delta = [\Delta\delta HN^2 + (0.1\Delta\delta N)^2]^{1/2}$) obtained from the 2D [^{15}N,^{1}H]-TROSY experiments, shifted (pink) and missing (grey) peaks are mapped on the entire Sox2-HMG surface (1GT0). Residues, which are significantly shifted, are labeled. D) Changes in chemical shift ($\Delta\delta = [\Delta\delta HN^2 + (0.1\Delta\delta N)^2]^{1/2}$) upon Sox2_CCND1 binding is plotted as a function of the Sox2 amino acid sequence. The pink line denotes the minimum threshold for significant CSP, and asterix the peaks that broaden upon complex formation.

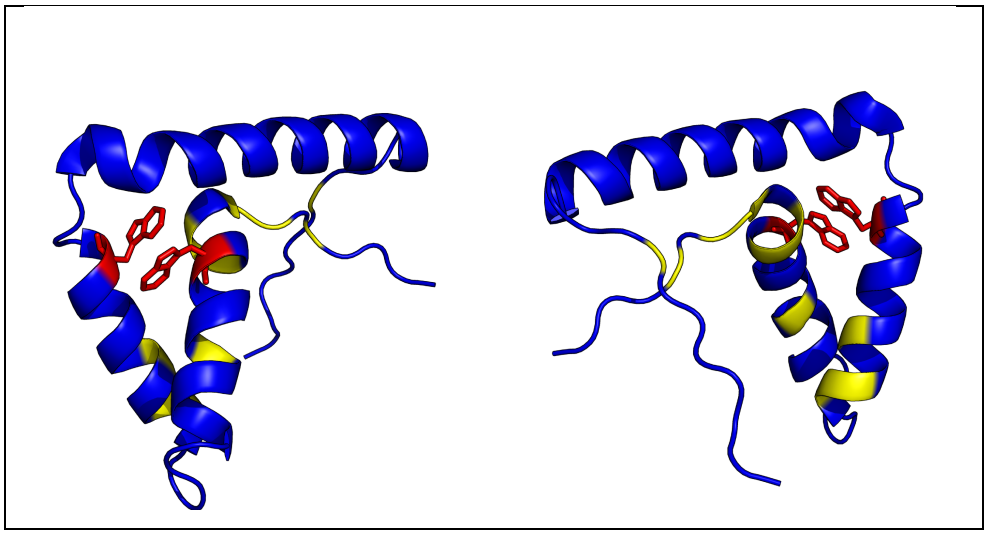


Figure 3.13. The DNA binding residues (yellow) as reported for PDB ID:1GT0 mapped on the entire Sox2-HMG surface with the tryptophan residues W13 and W41 colored in red. Back and front views are depicted.

iv) Characterization of Binding to DNA

Imino protons corresponding to CCND1 were assigned using 2D nuclear Overhauser enhancement spectroscopy (NOESY). Titrations were performed by addition of Sox2-HMG to the DNA to achieve various complex ratios. **Fig. 3.14** depicts the 1D stack plot of the imino proton region in the 1D ^{1}H NMR experiment at various Sox2-HMG : DNA ratios. The imino protons are usually observed as pairs of A-T or G-C in discrete spectral regions. This is because of the characteristic ring currents of these base pairs [147]. Spectral changes are observed within the first point of addition of Sox2 to the DNA (ratio-0.25). Increasing amounts of Sox2-HMG lead to the distinct bound conformation of the protein as can be inferred from the appearance of new peaks (**Fig. 3.14**).

The imino proton region was monitored using 1D ^{1}H-NMR spectrum in the 11-15 ppm range. **Figure 3.14a** depicts the Sox2_CCND1 complex at increasing amounts of protein titrated to the DNA (P/D ratio). As the protein to DNA ratio increased, the imino proton resonances were observed to shift and decrease in intensity and new peaks were observed to appear as well, indicative of conformational changes in the DNA. Looking at the spectra, one observes that significant decrease in peak intensity in observed for all peaks, with several peaks merging into one broadened peak. C12 seems to have moved

considerably to form a rather broad peak with G4 and C5, undergoing further line broadening as the P/D ratio is increased. The peak intensities for A10 and T11 were observed to merge and shift slightly at the P/D ratio of 0.5, and slowly diminish as the concentration of Sox2 increases. Peaks T6, T7, T8 experienced only line broadening with increased ratio without any chemical shift changes. An additional small peak is observed next to peak T6, which increases in intensity as P/D increases leading to a bound-state spectrum of Sox2-CCND1 at 1:1 stoichiometry. The same is inferred for the new peak observed at 14.6 ppm that slowly increases in intensity as Sox2 concentration is increased. Sox2 also caused line broadening for other imino protons without significant chemical shift changes. Of interest is also the intermediate exchange mechanism at the tryptophan side chain (W13, W41). The TROSY and anti-TROSY effect is observed as the linewidth of the peaks in the protein-DNA complex has broadened considerably. This indicates the presence of binding between the DNA and the protein**.**

^{31}P experiments were also performed on the complex of Sox2-CCND1, as it is advantageous for studying conformational changes induced by the binding to proteins. The ^{31}P spectrum of the Sox2_CCND1 binary complex is observed to have the same chemical shift dispersion as that of the free DNA (**Fig. 3.14c**). The line broadening observed in both the ^{1}H imino proton and the ^{31}P spectra of the Sox2-CCND1 complex is a result of its large molecular size in comparison to free CCND1.

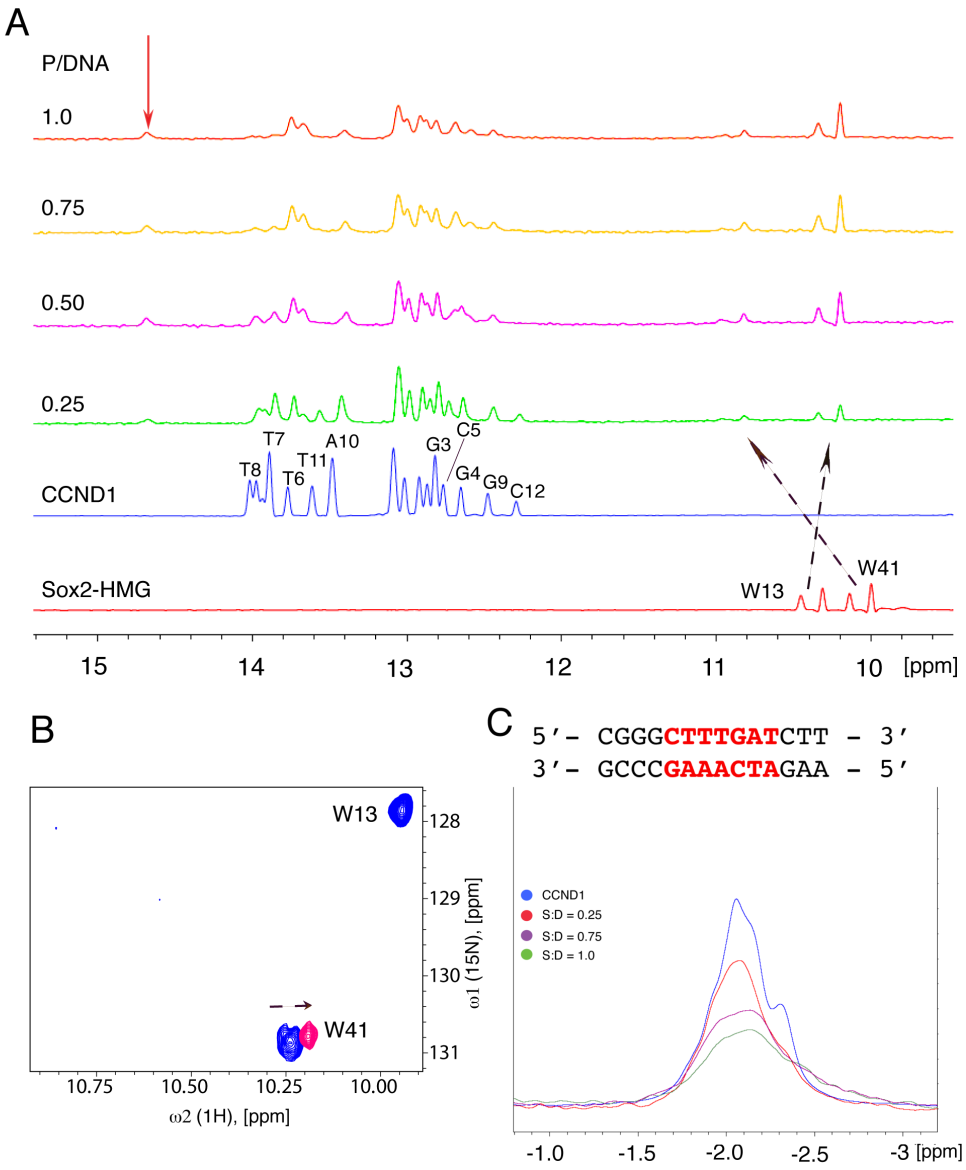


Figure 3.14. Interaction of CCND1 with Sox2-HMG analyzed by the 1D-^{1}H in support of the 2D [^{15}N,^{1}H]-TROSY and ^{31}P NMR **.** A) 1D ^{1}H spectra show increasing stoichiometry of Sox2-HMG in complex with CCND1, from bottom to top. B) The spectral region shown highlights the larger shifts of the tryptophan NE1/HE1 side chain group, indicated by arrows. The chemical shift changes observed is indicative of an intermediate binding between the side chain and the DNA. C) ^{31}P NMR spectra of the 14-mer CCND1 only and in complex with Sox2-HMG at 298 K with increasing stoichiometry.

3.4.4 SOX2-HMG binding with CCND1 and POM

i) Characterization of effect of POM on binary complex

The ligand POM was titrated into the binary complex of Sox2-HMG_CCND1 to observe changes, if any, which were inflicted to the binary complex Sox2_CCND1. Similar spectral changes were observed as that of the binary Sox2_CCND1 spectra, i.e. there was no drastic shift of peaks as observed in the 2D [^{15}N,^{1}H]-TROSY when CCND1 was added into the system. Addition of POM to the binary complex in concentration ratios of 1:0, 1:0.5 and 1:1.2 resulted in attenuated intensities for most of the cross peaks. Several peaks that experienced peak shift in the binary complex were observed to undergo further CSP upon addition of POM to the system. However, cross-peaks corresponding to the free form of Sox2 did not appear, as expected if the Sox2-CCND1 complex were to dissociate. All changes are represented in the graph and structure of Sox2 (PDB:1GT0) in **Figure 3.15.**

As discussed earlier in section 3.4.2 iii, docking analysis illustrated the binding sites that are highly preferred by POM to be in the form of a pocket formed by the C-terminus of helix-3 and the N-terminal region of the minor wing of the Sox2-HMG structure. In accordance with the docking studies, the positively charged residues Lys4, Arg5, and Arg15 exhibit moderate chemical shift changes, whereas His63 and His67 exhibit insignificant perturbation for the Sox2_POM. However these residues were observed to experience increased CSP and line broadening when POM was added to the Sox2_CCND1 binary complex instead (**Fig. 3.15a**).

Although a protonated state of Glu66 could be envisaged to form a hydrogen bond with a terminal POM due to a slightly altered pH microenvironment, it is unlikely at the experimental conditions. Rather, the POM binding might repel Glu66 due to unfavorable electrostatic repulsion causing it to structurally reorient with respect to the apo-Sox2-HMG. Asp82 is a residue strongly perturbed in TROSY (Sox2_POM complex) and Sox2_CCND1_POM complex) however it is not present in the POM docking site. Asp69 is located in the C-terminal of helix 3, which is a region subject to inherent conformational changes due to protein flexibility as shown in NMR studies on apo Sox4 and Sox5 [126]. Following DNA binding, however, the C-terminus is markedly rearranged to participate in the DNA interaction. Thus, the pronounced perturbation detected for Asp69 likely illustrated the dynamics of this region and its structural adjustments accompanying

molecular recognition events which is further illustrated by the NMR results, whereby CSP is significant in the binary complex of Sox2_CCND1 and line broadening in this region is observed to be predominant with the addition of POM to the binary complex.

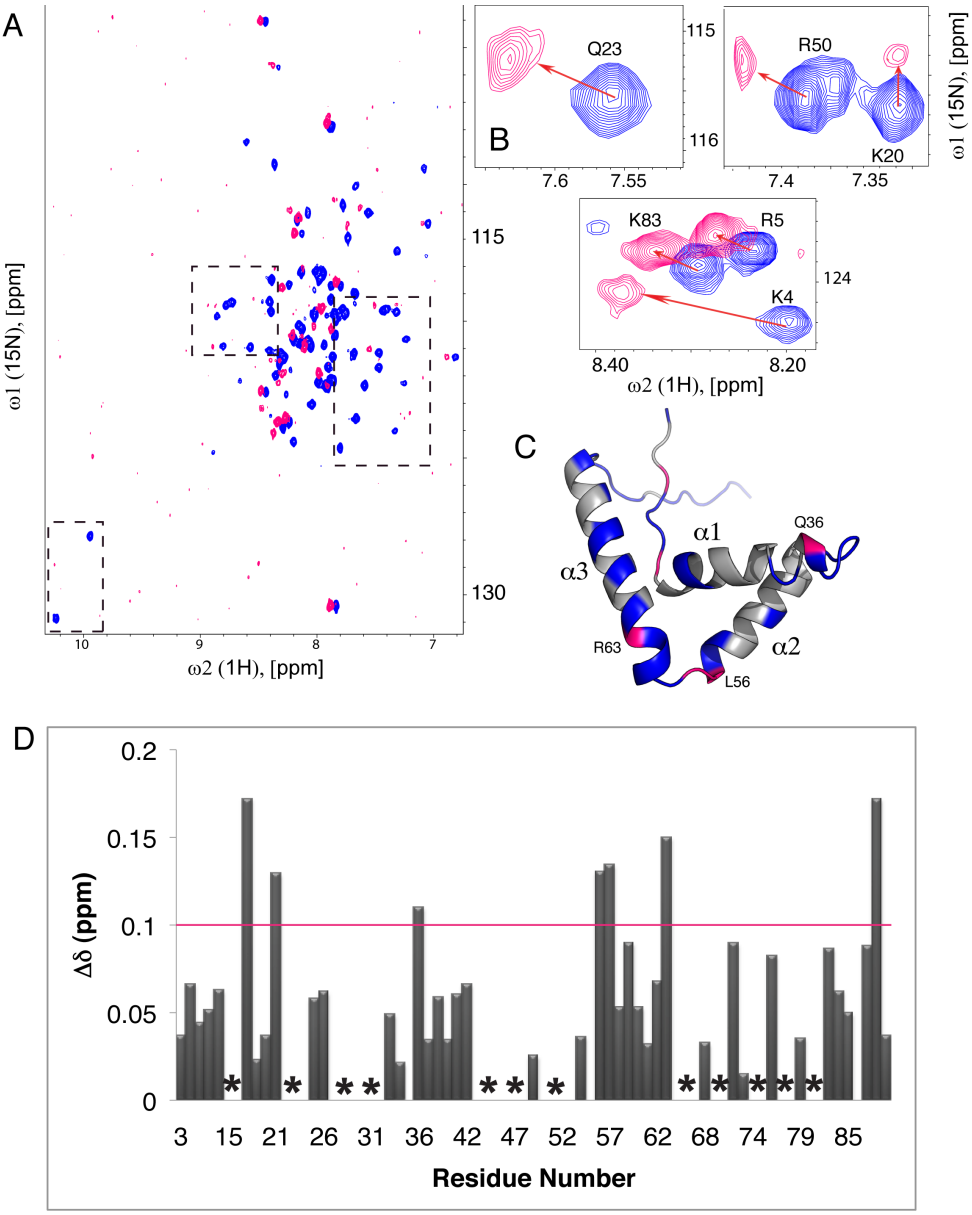


Figure 3.15. Interaction of Sox2-HMG_CCND1_POM analyzed by 2D [^{15}N,^{1}H]-TROSY. A) Superimposed spectra of two-dimensional TROSY spectra of free Sox2 (blue) and Sox2_CCND1_POM (pink) at a ratio of 1: 1.2: 1 B) Example of peaks that undergo significant shifts. C) The weighted change in chemical shift perturbation ($\Delta\delta = [\Delta\delta HN^2 + (0.1\Delta\delta N)^2]^{1/2}$) obtained from the 2D [^{15}N,^{1}H]-TROSY experiments, shifted (pink) and missing (grey) peaks are mapped on the entire Sox2-HMG surface (1GT0). Residues, which are significantly shifted, are labeled. D) Changes in chemical shift ($\Delta\delta = [\Delta\delta HN^2 + (0.1\Delta\delta N)^2]^{1/2}$) upon Sox2_CCND1_POM binding is plotted as a function of the Sox2 amino acid sequence. Asterix denotes broadened peaks.

ii) Conclusion

Stable protein-DNA complexes required for study by NMR is usually within the range of 0.1-1 mM solutions, which is rather challenging to obtain as the highly positively charged state of DNA binding domains contrast significantly with the highly negatively charged nucleic acid phosphate groups. At lower concentrations, these two conditions would not deem as problematic, however at higher concentrations, strong electrostatic interactions between the two interacting partners would occur thus causing precipitation of the complex. Different strategies were applied in this study to overcome this limitation, namely i.) Increasing the salt concentration and ii) using the shortest possible DNA sequence to reduce the number of charges [13]. NMR experiments were performed at 50, 100 mM and 300 mM salt concentrations. The higher salt concentrations increased the solubility and stability of the complex, and since both 100 mM and 300 mM yielded similar results with reference to the CCND1 gene. However, at 50 mM salt concentration, the decrease in signal intensities was too drastic upon the addition of sub stoichiometric amounts of CCND1. Thus, we focused on 100 mM Nacl for all interactions.

Addition of CCND1 in small aliquots to the ^{15}N-labeled Sox2-HMG resulted in broadening and vanishing of a distinct set of peaks as well as perturbations of certain peaks. (In order to preserve the overall quality of the spectra, we performed most of the analysis at low concentration of both the unlabeled binding partner and Sox2-HMG.) Peak broadening upon addition of CCND1 to form the binary complex is indicative of an intermediate exchange of for those residues on the NMR time scale. This intermediate exchange behavior of the binary complex suggests KD values in the low micromolar range (KD ~1–10 μM). Apart from the decrease of peak intensity, residues were also observed to shift upon titration, suggesting that they are in the fast exchange limit. (These strong peaks most likely correspond to residues in flexible tail and loop regions on the periphery of the main protein–protein interface and thus display different dynamic properties). In order to visualize the binding surface on Sox2-HMG, we mapped the changes observed from the [^{15}N, ^{1}H] TROSY onto the Sox2-HMG structure (PDB ID: 1GT0).

The specificity of Sox2-HMG binding to DNA is well documented and this interaction is mediated by numerous base pair- specific hydrogen bonds [148] and involves many DNA contacts with less than half the affected residues resulting in interaction with

the DNA base pairs [146]. It has been reported that the side chains from residues of helices 1 and 2 of the Sox2-HMG domain are inserted between 3-bp stacks of the recognition sequence (C-T-T-T-G-T-T), which leads to unwinding of the DNA helix at the Sox2-binding site [148]. The residues that decrease in intensity for the Sox2-HMG and CCND1 binary complex are predominantly in the binding interface similar to that of the ternary complex (molecular synergistic) and SRY/DNA complex. N30, S31 and R19 could play a role in bending the DNA at three different base stack levels at C.T.T [112]. The additional residues observed at K35, L37, G38 and A39, including the C-termini at helix-3 constituting residues M64, E66 and H67 could be due to the different DNA site preferences. The interacting residues at the C-terminal are similar and in close proximity to those residues identified to be involved to bind electrostatically with POM via NMR titration studies as well as docking analysis. Residues M64, E66 and H67 are at the periphery of helix-3 and are affected by binding to the CCND1 gene and addition of POM affects additional residues in this region, namely A61 and D69. The predominant electrically charged residues in the binding pocket indicate electrostatic interactions between the Sox2-HMG with the negatively charged phosphate backbone and this pattern is consistent with the observed salt dependence of the strength of the interaction between Sox2-HMG and CCND1. The perturbation of positively charged residues at the C-terminus is possibly due to the repulsion experienced between the DNA and Sox2-HMG, as was previously observed between Sox2-HMG and POM. It is interesting to note that the side chain of both the tryptophan residues chain of both the tryptophan residues is affected by the addition of Sox2 to CCND1. Both undergo an upfield shift, however the shift is maintained (negligible shifts) even at increasing ratios of CND1 in the complex. W41 experiences a drastic shift in comparison to W13 of about 0.81 ppm. The C-terminal of Sox2-HMG has been shown to be dynamic and unstructured when in the unbound state [126, 143] and upon binding to DNA, it rearranges itself to better accommodate the DNA partner. Thus the decrease of peak intensities for the residues at this terminus could be due to these rearrangements, specifically upon titration of POM to the binary complex. The residues in the intermediate exchange regime were expected to re-appear upon addition of POM thus signifying the dissociation of the DNA from Sox2-HMG; however it was difficult to detect these peaks possibly due to the high molecular weight of the ternary

complex. Certain residues were observed to shift back to the original chemical shift upon addition of POM to the complex. These residues were V3, S14, R15, G16 and R73. The shifts of these residues are indicative of fast exchange and low binding affinity and the presence of POM in the system inhibits/decreases that fast exchange between the DNA and Sox2 at these regions. Interestingly, these residues are positioned within the binding pocket of Sox2-HMG. Additionally, the concentrations of the complex used in NMR studies in comparison to the EMSA studies are significantly higher. Thus allowing for possibly more interactions to occur and difficult for the complex to dissociate completely.

3.4.5 Sox2-HMG binding with DC5 and PAX6

i) Cooperativity Assay

The extent of cooperativity (referred by a cooperativity factor ω) between Sox2 and Pax6 on the DC5 enhancer element was quantitiavely inferred by Kamesh.N et.al., (unpublished) using a previously established EMSA based cooperativity assay [149]. From the cooperativity assay, it was established that Sox2 and Pax6 shows a higher cooperativity while binding to the 30bp DC5 enhancer element that is ~50 fold (ω~40-50) higher as compared to their binding to the idealised element where they show no cooperativity (ω ~1). These results provide the first ever quantitative biochemical measurement of cooperativity in accordance with the qualitative assessment of cooperativity made by Kondoh *et al* [150].

ii) CD spectroscopy of ternary complex

In this study DC5 was used as the oligonucleotide of interest to determine specific binding of the protein to DNA. For the protein, both the minima (208 and 222 nm) and the maxima (192) are indicative of α-helices present (Pax6 and Sox2). In all complex samples, the DNA was slightly in excess to ensure complete binding was achieved. **Figure 3.16** clearly shows that binding is achieved as changes are observed to the protein as well as DNA. As for the oligonucleotide, the positive peak at 275 nm was observed to increase reflecting DNA structural changes.

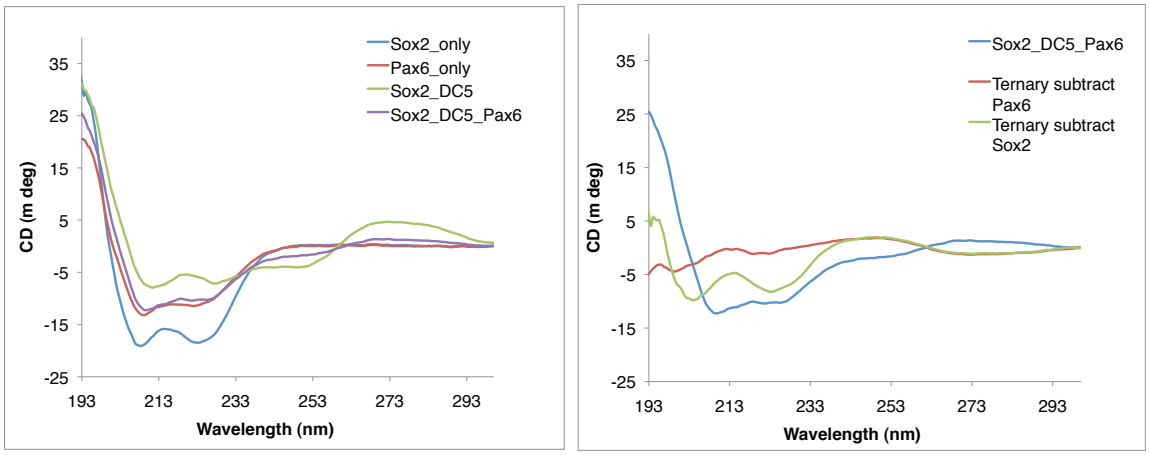


Figure 3.16. CD spectroscopy of Sox2-HMG binding to DC5. Near and far UV spectrum of Sox2-HMG:DC5:Pax6 (1:1.2:1 molar ratio).

iii) Characterization of the DNA-binding Interface

Transverse relaxation optimized spectroscopy (TROSY) spectrum of the binary complex consisting of Sox2-DC5 maintains a well-dispersed spectrum. Interactions between the binary and the final ternary complex were observed via mapping the changes observed in chemical shifts (**Fig 3.17**). Sox2 was added in increment amounts to the large 31-mer DC5. Initial (small) titrations showed similar perturbations as that of the final molar ratio (as was observed for the Sox2-CCND1 binary complex) (**Appendix B4**). As such, the figure shows results extracted from ratios of 0, 1:1.2 (Sox2-DC5) and 1:1.2:1.0 (Sox2-DC5-PAX6). 2D [^{15}N,^{1}H]-TROSY was used to monitor changes in Sox2 and to monitor changes in the DNA, ^{1}H and ^{31}P experiments were recorded. In the course of the titration, the 1D ^{1}H spectra of the imino protons of DC5 broadened indicating a binding interaction between both the protein and DNA (**Fig. 3.17**). The 2D [^{15}N, ^{1}H]-TROSY showed shifts for most peaks and lowered intensity for a significant number of peaks. The binding kinetics of Sox2-DC5 inferred from the 2D spectrum is indicative of being in the slow to intermediate regime as observed by the line broadening.

Significant changes were observed on several areas of Sox2 upon binding to DC5 apart from the specific DNA binding sites. **Figure 3.17** depicts the extent of changes observed (ppm) for each residue of the transcription factor. Similar to the CCND1 binary complex, two forms of changes were observed: peak shift and peak broadening. The largest chemical shifts perturbations (> 0.1 ppm) observed were for K4, G16, K20, E24, N25, S34,

K42, L43, A56 and K65 situated at the random coil of the N-terminal of Sox2 and α-helix 2 and 3. These CSPs could be indicative of fast exchange on the chemical shift scale and a change in the conformation of the protein upon binding to DC5. Residues that experience peak broadening are predominantly at the hydrophobic core of the protein comprising the major wing of the protein: R2, A9, F10, V12, S14, Q17, R18; and at the turn regions of α-helix 1 and α-helix 2 which are involved in DNA binding, residue N30-I33, G38, A39, W41 were affected. At the minor wing, the only changes were towards the C-terminal region corresponding to K57, L59, L62, M64, H67 and R73. Following the binary complex, Pax6 was titrated into the system, as it was our interest to observe the influence of Pax6 on the Sox2/Pax6 complex to DNA [129]. The sensitivity of the resonances corresponding to the binary complex upon this addition was observed to be focused at the major wing. The attenuated intensity of the binary complex 2D [^{15}N,^{1}H]-TROSY spectrum in the presence of Pax6 is consistent with the formation of a high molecular weight complex. The fact that the resonances remain in the same positions as that of the binary complex Sox2_DC5 indicates that the protein Pax6 has not cause dissociation of Sox2 from DC5. The interaction that arises from this complex Sox2_DC5_Pax6 under NMR conditions enables us to determine the sites on the Sox2 bound DC5 that are preferentially affected by Pax6. The series of titrations done showed that the similar changes are observed at the first titrated point until the last point, with significant changes being related to the attenuation of peak intensity.

Peak broadening was observed for these residues: W13, G16, K20, M21, E24 in α-helix 1; and in the turn regions of α-helix 1 and α-helix 2 residue K27, H29, L37, L43, and E48. At the minor wing (α-helix 3) residue F52, I53, R60, A61, and D69 are affected as well. Changes predominantly in peak intensity indicate that the binary complex is still formed upon the addition of Pax6 to the system, and more importantly Pax6 has not caused the dissociation of the binary complex. This is consistent with the formation of high molecular weight ternary complex (42 kDa). The absence of new reemerging resonances suggest that not only is the complex stable, but there are also no dissociated components present.

The side chain tryptophan residue W13 (εH) and W41 (εH) are of unique interest as they are quite affected by DC5 but not so much the protein, Pax6. In the binary complex, addition of DC5 at a molar concentration of 0.25 ratio causes a dramatic shift of W54 that

is maintained at the complete molar ratio of 1:1.2. Addition of Pax6 does not change the positions of both the side chain tryptophan residues, but it does affect their intensity. Figure 3.11b illustrates this better, where resonance corresponding to W41 (εH) is observed to attenuate considerably as opposed to W13 (εH) that remains fairly visible. Possible interactions between Sox2-HMG and Pax6 were determined and found to be not present through NMR titration experiments of Sox2:Pax6 stoichiometry up to 1:2 (**Appendix B3**).

These results suggest that incorporation of Pax6 into the binary system increases the site-specific affinity of Sox2 for DC5. Kinetics of Sox2 interactions with DC5 is in the fast exchange regime on the chemical shift timescale, however in the presence of Pax6 it changes to the intermediate exchange regime (**Fig. 3.17**). Based on this observation of a reduction of the exchange rate in the presence of Pax6, suggests the dissociation of Sox2 from DC5 to be decreased. We postulate that this is the preferred situation for complex formation as the reduced dissociation points towards increased stability of the ternary complex thus promoting transcription activation. *Kamachi et. al.* [129] have shown that Pax6 and Sox2 bind cooperatively to the DC5 gene and this cooperativity plays an essential role in transcription activation. Their *in vivo* studies proved similar observations; cooperativity is important in DC5 activation and the presence of a single factor only is insufficient to associate with the DC5 in the chromatin [129].

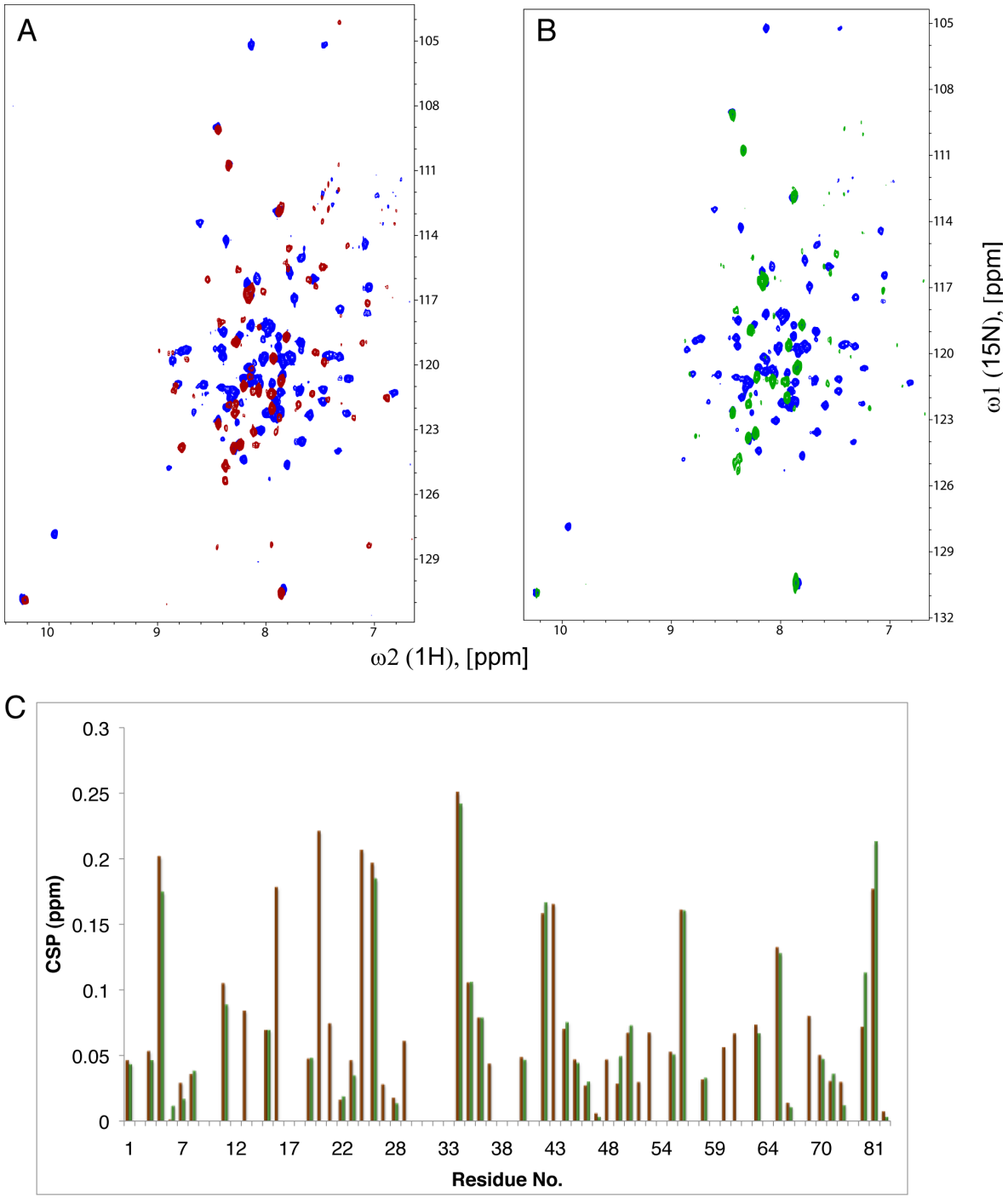


Figure 3.17. Interaction of Sox2-HMG with DC5 by 2D [^{15}N,^{1}H]-TROSY**.** A) Superimposed spectra of two-dimensional TROSY spectra of free Sox2 (blue) and Sox2 bound to DC5 in a series of titrations, final ratio 1: 1.2 (brown) B) Superimposed spectra of two-dimensional TROSY spectra of free Sox2 (blue) and Sox2_DC5_Pax6 in a series of titrations, final ratio of 1: 1.2: 1.0 (green). C) Changes in intensity upon DC5 binding is plotted as a function of the Sox2 amino acid sequence, brown bars-Sox2_DC5 only, green bars-Sox2_DC5_PAX

iv) Characterization of binding to DC5

Similar to the Sox2-CCND1 interaction, the 1D ^{1}H spectra also depict an intermediate exchange mechanism at the tryptophan side chain. The TROSY and anti-TROSY effect is observed as the linewidth of the peaks in the protein-DNA complex has broadened considerably. This indicates the presence of binding between the DNA and the protein.

It is interesting to note that the side chain of both the tryptophan residues is affected by the addition of Sox2 to DC5. Both undergo an upfield shift, however the shift is maintained (negligible shifts) even at increased ratios of DC5 to the protein. W41 experiences a drastic shift in comparison to W13 of about 0.81 ppm (**Fig. 3.18**). Addition of Pax6 to the binary complex causes the W41 side chain to reduce in intensity, indicating a strong interaction within the ternary complex. However W13 still remains observable in the 2D [^{15}N, ^{1}H]-TROSY spectra without further changes.

The 1D ^{1}H spectra of only the 31-mer DC5 gives an idea of the influence the size of DNA has on the peak resolution. Comparing 1D of CCND1 (**Fig 3.14**) and 1D of DC5 (**Fig. 3.18**), it is clear that assignment of the imino protons for the DC5 gene is rather difficult at this stage. However, a general analysis of the spectra upon titration of Sox2 illustrates that almost all resonances experience line broadening with very minimal peak shift. At 14.4 ppm, a new peak is observed to emerge, indicating conformational changes on DC5 to accommodate the bound Sox2. The intensity of this peak increases as the stoichiometry of Sox2_DC5 increases, and it experiences further line broadening when Pax6 is introduced to the system. This severe line broadening could be due to the fact that the total molecular weight of the ternary complex has significantly increased.

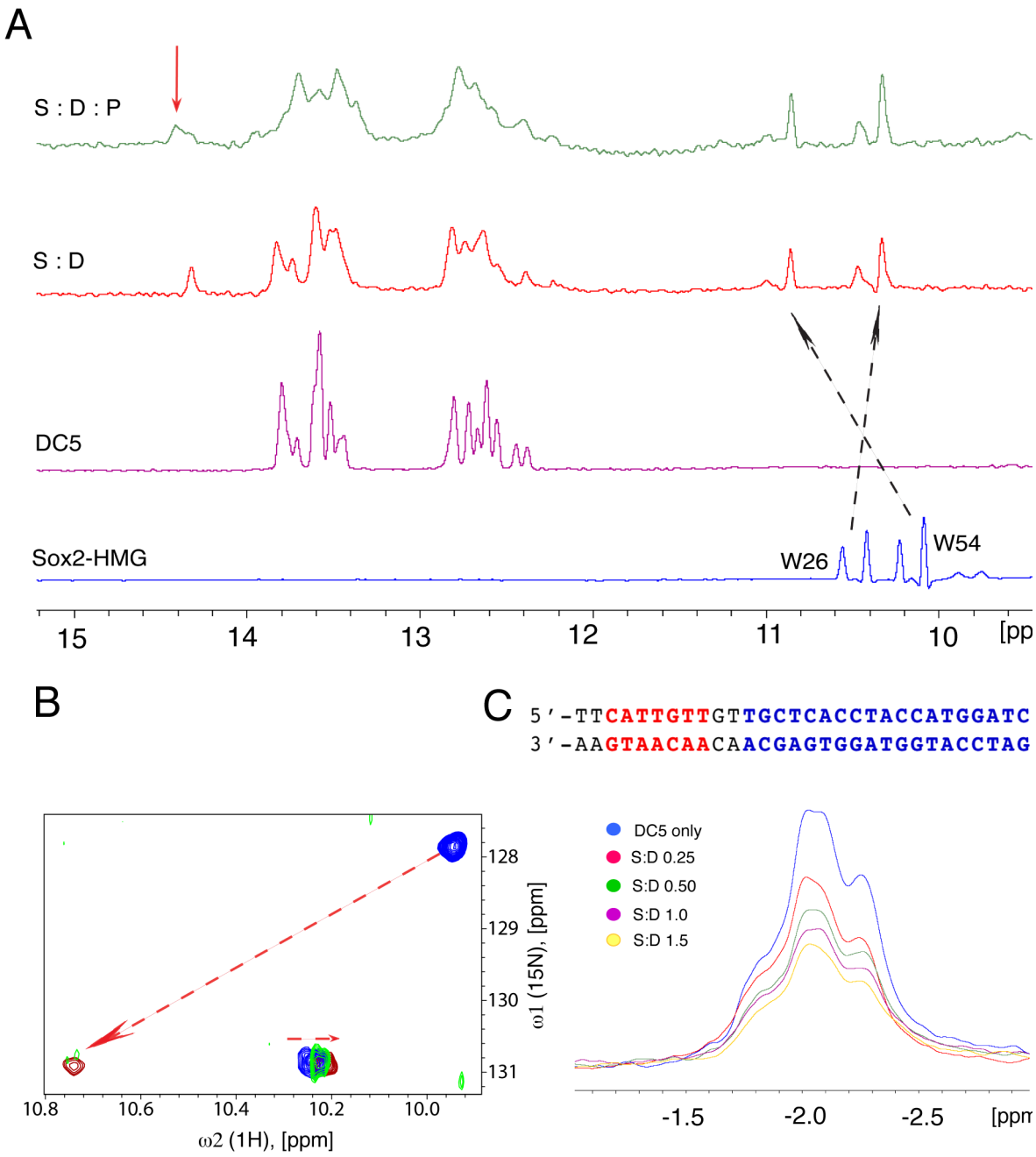


Figure 3.18. Interaction of DC5 with Sox2-HMG analyzed by the 1D-^{1}H in support of the 2D [^{15}N,^{1}H]-TROSY and ^{31}P NMR **.** A) 1D ^{1}H spectra show increasing stoichiometry of Sox2-HMG in complex with DC5, from bottom to top. B) The spectral region shown highlights the larger shifts of the tryptophan NE1/HE1 side chain group, indicated by arrows. The chemical shift changes observed is indicative of an intermediate binding between the side chain and the DNA. C) ^{31}P NMR spectra of the 31-mer DC5 only and in complex with Sox2-HMG at 298 K with increasing stoichiometry, top to bottom.

3.5 Summary

Transcription factors have been considered as too challenging to be studies as possible drug targets, specifically the DNA binding domains. This is rather unfortunate because upregulated transcription factors have been shown to be a major contributor of cancer and other related diseases (1,4,49). In this study we have studied these transcription factors in order to understand better the binding of protein to DNA, and the cooperativity elements involved in enhancing transcription activity as well as identify possible candidates in inhibiting the DNA binding activity of Sox2.

During the course of this study, certain challenges were confronted with regards to studying protein-DNA binding under NMR conditions. This is due to the fact that stable protein-DNA complexes required for study by NMR is usually within the range of 0.1-1 mM solutions, which is rather challenging to obtain as the highly positively charged state of DNA binding domains contrast significantly with the highly negatively charged nucleic acid phosphate groups. At lower concentrations, these two conditions would not deem as problematic, however at higher concentrations, strong electrostatic interactions between the two interacting partners would occur thus causing precipitation of the complex. Different strategies were applied in this study to overcome this limitation, namely i.) increasing the salt concentration and ii) using the shortest possible DNA sequence to reduce the number of charges [13]. Another method of overcoming the precipitation problem would have been to mutate the basic residues on the protein surface away from the DNA recognition areas, however this strategy was not used. The reason being that mutations on the protein have been shown to cause diseases and therefore we strived to avoid such conditions. Stability of the complex at high temperatures (25°C to 35°C) is also an important factor as this allows acquisition of sharp signals, especially for the identification of intermolecular NOEs.

We have mapped in this study the interaction sites of Sox2 and the CCND1 gene as well as identified a Dawson-POM as a potent compound to inhibit the DNA binding activity of Sox2 [139]. The mode of binding is electrostatic in nature and involves predominantly electrostatic interactions at the pocket just outside the DNA binding region, but close enough to inhibit the DNA binding to Sox2-HMG. This inhibitory mechanism could be used to develop modified classes of POM in combating aberrant gene expression [139].

This study also addressed the question pertaining to the importance of Sox2 proteins and their co-DNA binding partners in transcription activation. It has been proposed that Sox2 and Pax6 cooperatively regulate the initiation of lens development, and it is this co-partnering with different DNA binding factors that determine their regulatory targets [129]. We observed from the NMR results that the addition of the co-DNA binding partner Pax6 leads to a cooperative increase in DNA binding affinity for Sox2-HMG. Control of transcriptional activation is dependent on the combinatorial control of transcription factors, and the decreased dissociation of Sox2 incurred by the presence of Pax6 results in a Sox2-HMG_DC5_Pax6 ternary complex that fulfills this requirement.

CHAPTER IV

CHAPTER IV

N-Terminal Domain Histone 4 (NTD-H4)

4.1 Introduction

4.1.1 Chromatin Overview

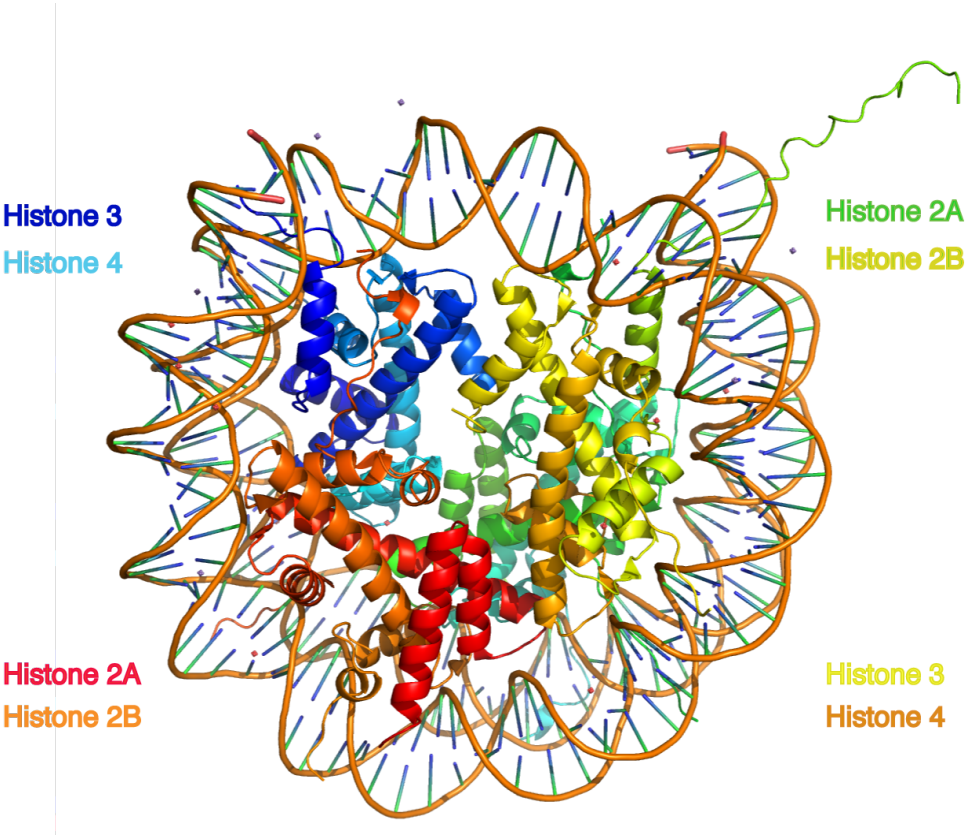


Figure 4.1. Crystal structure of nucleosome core particle (NCP) (PDB:1AOI) [151].

Eukaryotic DNA is organized in the chromatin complex via packaging it in the nucleosomes using histone and non-histone proteins [152]. It is the essential role that chromatin plays by being part of the gene regulation process that demands the understanding of its structure and dynamics. Its organization comprises the nucleosome complex (**Fig. 4.1**), which consists of equal amounts of histone proteins and DNA [153]. The structural arrangement of the nucleosome core particle (NCP) in the chromatin fiber is believed to be a helical packing [154].

Chromatins are 30 nm fibers that are compacted in the cell, which undergo further compaction during the mitosis state in the cell cycle [155] with the overall compaction

during this state is 10, 000 fold. These compacted states are not included in interactions or mechanisms that usually involve the DNA, such as repair and recombination [156, 157]. Such conditions are highly unsuitable for transcription by eukaryotic RNA polymerases and therefore require that these compacted chromatins to be unfolded first (open); also known as 'transcriptionally competent' state. This open state would then lead to the basic unit of the chromatin, which are the nucleosomes [155]. The nucleosome is approximately 147 base pairs of DNA wrapped 1 and 3/4 times around the core histone octamer; this nucleosome is then linked between each other through a linker DNA (~10-90 bp). 75% of the core histone forms the spool that wraps the DNA, and is comprised of 2 heterodimers H2A/H2B (dimer) and H3/H4 (tetramer). The remaining core histones are made up of the structurally undefined 'tail' domains.

These domains are situated at the N-terminal of each histone protein and the C-terminal of H2A. The positioning of the tail domains are either along the minor DNA grooves (H3 and H2B) or a bulk proportion exiting the exterior of the nucleosome causing it to be solvent exposed and accessible to many interactions (**Fig. 4.1**) [158]. The mobility of this tail would allow it to be located in different positions whether in the nucleosome core, mononucleosomes and chromatin fibers. This is closely dependent on the manner in which they are isolated or assembled. This tail domain rearrangement relates very closely to its function in the nucleosome core [159]. Close studies of the tail domain have shown that various posttranslational modifications occur within this domain, modulating their structure and ultimately their function. In order to access the DNA either during the transcription, replication or repair process, the decondensation of the chromatin is required. NMR experiments show that the binding interactions of the tail domain are predominantly electrostatic in nature; forming random coil when the histone is in its free form [158]. Removing the histone tail disrupts the compaction of the nucleosome arrays into 30 nm fibers. This is possibly due to histone-histone interactions which leads to inter-nucleosomal aggregation [160].

4.1.2 Histones

In forming the NCP, the four histones (H2A, H2B, H3 and H4) interact very selectively with each other. Both H2A and H2B form heterodimers, and H3 and H4 form tetramers. It is observed that the C-terminal domains of all 4 histones are very similar, in contrast to their N-terminal domains. The C-domain consists predominantly of helices bordered by β-structured loops on each side. The central long helix is the interface where dimerization occurs. The area of interface is different for each of the corresponding heterodimers; between two H3/H4 is smaller compared to H3/H4 and H2A/H2B. The larger interface of the latter heterodimer allows it solvent accessibility and therefore renders it less stable.

Histones are made up of 2 domains - one being the structured protein, and the latter being an unstructured "tail" domain at the N-terminal end of the protein. It is this tail that undergoes a variety of post-translational modification (PTM) that mediate chromatin structure, assembly, replication and silencing [161, 162]. PTMs that have been studied with regards to histone proteins include acetylation, phosphorylation and methylation [163]. H3 and H4 are both highly conserved, thus it would be natural to suggest that the structural and functional roles of these proteins are essential, however there have been studies that show otherwise as both proteins are expendable respectively, in the cell cycle and chromatin assembly [164]. As for the other two histone proteins, H2A and H2B, the sequences are considerably similar between species and highly conserved.

However, with regards to each other, H3 and H4 have been shown to be more crucial in NCP folding in comparison to the H2A/H2B dimers [165], with the H3 being a repository of regulatory information as opposed to affecting chromatin order [152]. H4 on the contrary is essential for chromatin compaction, which could be manipulated with different concentrations of Mg^{2+}. This indicates the interaction between the former and the latter to be electrostatic in nature [166, 167].

4.1.3 Histone tail importance in chromatin folding

Studies have shown that the histone-DNA interactions are straightforward in its ability to maintain the NCP. Association and dissociation of the NCP can be directed via manipulation of the ionic strength of the system. This clearly indicates the role that electrostatic interactions play in the NCP formation. In the course of dissociating an NCP, the first to go is the H2A/H2B heterodimer at an ionic strength of 0.8 M NaCl, and the H3/H4 tetramer only follows suit at a higher ionic strength of 1.2 M NaCl. The central point of this whole interaction is that it is reversible. This basically means that if one simply dilutes the solution, the core histones would re-assemble themselves with a DNA to form the famed NCPs [168].

The histone tails are the most dynamic regions of the nucleosome core and it plays a vital role in chromatin compaction and transcription [159]. The position of these tail domains in the NCPs, mononucleosomes or chromatin fibers are dependent on the method of isolation or assembly. This inherent dependence gives an insight into the importance of understanding the rearrangement that the tail domains undergo as it affects the nucleosome function. The presence of an abundance of positively charged amino acids creates a suitable environment for the internucleosomal interaction as well as DNA interactions to occur. The natural opinion of the positioning of the histone tails is on the major groove of the interacting DNA. However it has been proven that this is not the case . The crystal structure [151] shows that the tail of H4 (16-25) possibly interacts with the H2A/H2B dimer from an adjacent nucleosome [151]. Dorigo *et al*. further supports the conclusion of H4 being of vital importance, especially the tail region in mediating nucleosome-nucleosome interactions through deletion and mutation studies [152].

H3 and H4 tails interact with the central region of the nucleosomal DNA whereas H2A and H2B interact with the peripheral regions of nucleosomal DNA [169]. It has also been observed that positions of histone tails were altered from non-nucleosomal to nucleosomal when chromatic fibers undergo digestion to produce NCPs [170]. Several conclusions have been reached over the decade regarding the function and implications these tails have on chromatic folding. Firstly, chromatin fiber digestion into NCPs leads to interaction of tails with the nucleosomal DNA, however it is yet to be shown that these tails are bound to the nucleosomal DNA in physiological conditions. Secondly, it has been

observed that the tail domains mediate chromatin folding via protein-protein interactions and tail domain-linker DNA interactions [159].

Apart from its essential role of maintaining protein-protein interactions during compaction, these tail domains also regulate intra-nucleosomal interactions [171]. It was concluded that histone tails play a predominant role in chromatin compaction; a structural role, functional targets for signaling cascades (PTM), shielding for DNA and possibly link other fibers together [152].

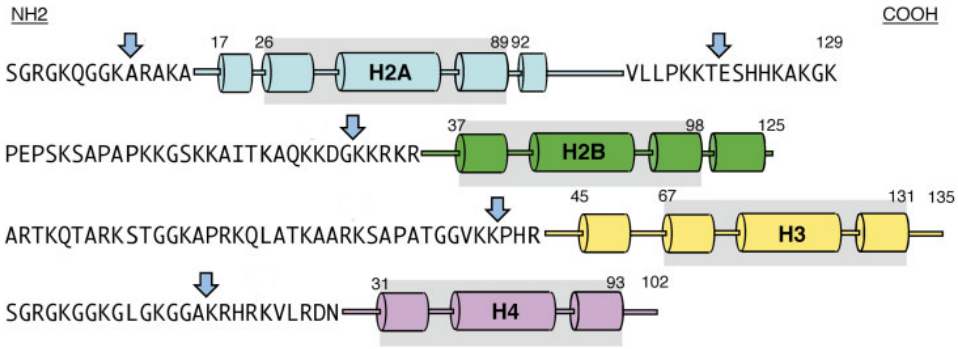


Figure 4.2. Schematic representation of histones with helical regions illustrated as columns. The tail regions are denoted as amino acid sequences. The arrows refer to the nearest point of exit of the tail domains from the nucleosome [172].

4.1.4 Effect of Salt on Array compaction

Nucleosome arrays increase in their compact structure relative to the concentration of Mg^{2+} ions present in the system. The extended "bead-on-a-string" structure is present at concentrations below 0.2 mM Mg^{2+}, and as it increases to 1 mM MG^{2+}, it adopts a compact nucleosomal array that is of interest to us in this present study. At concentrations above 2 mM, the arrays become a highly compact (aggregated) insoluble structure [173]. The different length of linker DNA that comprises 10-80 bp connects NCP, which form the "beads-on-a-string" nucleosome array.

It has been shown *in vitro* that compaction of nucleosome array can be achieved via addition of multivalent cations, namely Mg^{2+} [159].Chromatin folding as well as aggregation has been done using nucleosome arrays constituting the 601-DNA nucleosome positioning sequence [174]. It has been concluded that the factors that modulate chromatin

folding are length of the linker DNA, presence of the linker histone, the location of the histone tails and importantly the valency as well as the concentration of cations in the solution [175].

PTMs that most frequently occur is the lysine acetylation that results in a loss of positive charge [176]. This acetylation of lysine causes the chromatin to unfold by disrupting the nucleosome-nucleosome interactions, specifically the K16 of the H4 tail [177, 178]. Due to the observation that Mg^{2+} induces the compaction of chromatin fiber, it can be concluded that the dominating interactions involved are electrostatic in nature [179]. Lars and colleagues have shown the above to be true in their studies of the oligocation-dependent condensation properties as well as determined the optimum conditions for forming stable and compact nucleosomal arrays in solution using analytical ultracentrifugation (AUC) methods [173, 175]. There have been a few studies that reported inter-nucleosome interactions to occur in the vicinity of the H4-K16 and the 'acidic islet' of the corresponding H2A globular domain [180, 181], with the NCP-NCP stacking being considered as the important structural element required for chromatin compaction [175]. Although the method of chromatin fiber condensation is known *in vitro*, the physical and structural implication of this compaction on histones and their neighboring nucleosomes is still unclear.

4.1.5 Use of chemical shift in characterizing unfolded protein

The chemical shift of nuclei is dependent on the chemical and magnetic environment of an amino acid. As the protein gets denatured, the range of chemical shift dispersion for the particular protein reduces. This crowded and degenerate spectrum makes assignment of resonances rather difficult. When the confirmation of a protein is absent, the chemical shift is then calculated based on 'random-coil' chemical shift values [135]. The chemical shifts of ^{15}N and ^{13}CO have been observed to depend quite largely on the sequence of the protein even in the absence of a well ordered structure, as opposed to the other nuclei in an amino acid, namely $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$ and $^{1}H^{\beta}$ [135, 182]. The latter nuclei however can be useful in extracting information between different amino acid side chains, which inadvertently aids in the assignment process [183]. The $^{13}C^{\alpha}$ chemical shift however is believed to be the most sensitive of all in determining the population of secondary structure of a protein, especially

in unfolded proteins, whereas the dependence of the ^{13}CO resonance on the local sequence makes it unsuitable for such predictions [182]. The 3D heteronuclear experiments that would deem as a suitable combination for determining the backbone resonance assignment for unfolded proteins are HNCO (correlates NH*i*, N*i*, and CO*i*-1), HN(CO)CA (correlates NH*i*, N*i* and CO*i*/*i*-1), HNCACB (correlates NH*i*, N*i*, and C^{α},C^{β} *i*/*i*-1) and CBCA(CO)NH (correlates NH*i*, N*i*, C^{α},C^{β} *i*-1). These experiments result in peaks that are derived from 3 different chemical shifts; the H^N, N^H and the CO, C^{α}, or C^{β} (either for *i* or *i*-1) [184].

The combination of the above experiments would enable the linking of di-peptide systems. These sequences are then matched to form sequence specific assignment by overlapping the matching chemical shifts. For unfolded proteins, the degenerate C^{α} and C^{β} are not as effective in forming links compared to CO. But because it is useful in determining the possible side chains in a residue, this information can be used to infer the position in the sequence (segments). This approach has been automated for well structured proteins in software such as AUTOLINK [185], MARS [186] and many others. However, the manual approach is more suitable when dealing with disordered proteins albeit time consuming.

4.1.6 Hydrogen Exchange Methods

Two common methods of hydrogen exchange are the pulse-labeled H/D and the quenched H/D exchange. The former is initiated by unfolding of the protein of interest in D_2O at low pH to allow H/D exchange to occur at the required time range. This is then followed by refolding of the protein in H_2O buffer under normal refolding conditions such as pH 5.0 – 7.0 and 5 – 20°C. During this process, the folding intermediates formed are usually faster than the exchange rate constant due to the folding conditions. The folding process is then interrupted by introducing fast exchange conditions (pH 10 and 20°C) that enables the H-labeling. Because this is done in short burst of time, its known as pulse H-labeling. This process would allow the non-protected amide deuterons to exchange back to protons, but the protected deuterons remain as ND. The pulse time range can be from milliseconds up to several seconds. The following step in this H/D exchange is lowering the pH of the system to terminate or quench the process, which would allow the protein to re-fold back into

native state and maintain the existing H/D labeled residues. This experiment could either be monitored by 2D NMR or mass spectrometry, where NMR provides information that is residue specific [187].

The second method of H/D exchange commonly used is quenched H/D, which relies on the fact that a folded protein is stable and undergoes slow amide proton exchange that can be used as structural probes. This method is used to study denatured states as well as complexes, which would otherwise be very difficult to study using either NMR or crystallographic techniques [24]. There are problems related to this method, one of which is the fast back exchange that occurs. In our current study, we have followed a similar approach based on the work by Zhang *et al* [188]. The quench H/D developed uses a mixture of DMSO, H_2O and organic acid to slow down the process of back exchange. They reported that in 95% DMSO-*d6*/4.5% D_2O/0.5% dichloroacetic-*d2* acid at pH 5.5, the H/D exchange rates for model peptides were 100-folds slower compared to a system of 100% D_2O. Apart from preventing back exchange during the acquisition of the NMR experiment, this solvent system also proved to be highly advantageous in dissolving and disassembling amyloid fibrils thus resulting in monomers that were easier to resolve in an NMR spectrum.

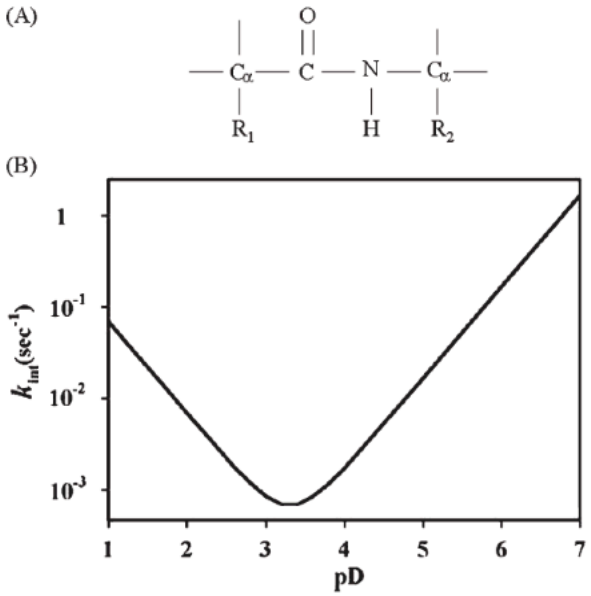


Figure 4.3. Hydrogen exchange for amide protons in an unfolded protein. A) Peptide chemical structure. B) Exchange rates for amide protons in PDLA at 20°C measured as a function of pD* (pD* = pH_{read} + o.4) (adapted from [187])

4.1.7 Overview of Study

As it has been implicated that the structural importance of the NTD lies in the tail domain, it would therefore be of great interest to characterize the conformations of this domain in the nucleosome. The crystal structure of a tetranucleosome has been determined at 9Å resolution; however the conformation of the tail domain was not elucidated [189]. Based on the high functionality of the tail domain, it is therefore our interest to study the N-terminal tail domain (NTD) of H4 in the compact nucleosome array. We would accomplish this by first establishing the sequential analysis of the backbone of H4 in its unfolded state in DMSO-*d6*. This assignment would then be used to identify the solvent exposed residues of NTD-H4 using hydrogen exchange experiments performed on the nucleosomal arrays.

4.2 Methods

4.2.1 Protein Expression and Purification

H4 was expressed and purified in collaboration with Professor Lars Nordenskiold's lab. Only the Histone 4 protein samples were labeled, either ^{15}N or ^{13}C and ^{15}N. The other histones (H2A, H2B, and H3) that were used in forming the nucleosome core particle were unlabelled. The buffer used constituted of 2 M NaCl, 20 mM Tris-HCl, pH 7.5 [173]. The H4 was lyophilized and stored at -80°C until further use.

4.2.2 NMR sample preparation

Solution conditions had to be found to optimize viscosity, sensitivity and most importantly sample stability. The sample was prepared at 0.5 mM and not higher, although higher concentration samples give better signal to noise, however we refrained from preparing such samples as these unfolded proteins have a tendency to aggregate easily. The sample becomes viscous at high concentrations and room temperature (25°C). The lifetime of the sample is approximately 2 weeks (tested so far). A longer period results in degradation, which could be due to the organic solvent, DMSO-*d6* (relatively strong/harsh solvent for biomoecules). The lyophilized H4 was dissolved by adding 0.5% TFA-*d3* to the powder, followed by DMSO-*d6*/D_2O solvent. The methods used in the following experiments were based on the protocols obtained from **NMR of Macromolecules: A Practical Approach** [134] and *Kato et al* [190]. All experiments were conducted at 298K.

4.2.3 Sequential Backbone Analysis

All NMR experiments for backbone sequential assignment were performed as reported in the NMR section of chapters one and two. The backbone resonance assignments were done using 3D HNCA, HNCACB, CBCA(CO)NH, HNCO and HN(CA)CO spectra collected from a Bruker 600 and 700 MHz spectrometer, equipped with Bruker ^{1}H, ^{13}C, ^{15}N triple-resonance probe-heads with three-axis, Ultrashield gradient coils. The spectra were collected at a regulated temperature of 298K.

4.2.4 Hydrogen Exchange

The hydrogen deuterium (H/D) exchange experiment was performed on compacted nucleosome arrays to characterize the conformational states of the N-terminal tail domain of H4. The experiment was first carried out on non-exchanged arrays to determine the suitability of the method for the array system. We were able to reconstitute the H4 in DMSO-*d6*, however at a lower concentration. The experiment was then continued with D_2O at 4 °C to determine the sites of exchange, where samples would be prepared for different time points in the exchange reaction.

Compact nucleosome arrays were produced for hydrogen exchange experiment by incubation of the nucleosome array solution at 23 °C in TK buffer (10 mM Tris, 10 mM KCl, pH 7.0; supplemented with 1.0 mM Mg^{2+}) [173] prior to the deuterium exchange experiment. It is important to note that the concentration at A_{260} of the nucleosome array solution should be <2. To perform the hydrogen exchange, same volume of deuterium exchange buffer (10 mM Tris, 10 mM KCl, pH 7.0 in D_2O; supplemented with 1.0 mM Mg^{2+}) is added to the compact nucleosome arrays and incubated for required time points. To quench the H/D exchange, the solution was rapidly frozen in liquid nitrogen and lyophilized. The lyophilized sample was then dissolved in NMR buffer (in 95% (v/v) DMSO-*d6*, 4.5% (v/v) D_2O and 0.5% trifluoroacetic acid-*d3*). Only the histone proteins from the nucleosome arrays would dissolve in the solvent and the DNA molecules precipitate. The 2D [^{15}N,^{1}H]-TROSY was then measured to observe the residues that undergo H/D exchange. The same procedure was repeated but with H_2O, and D_2O without Mg^{2+} to be used as a reference.

4.3 Results and Discussion

4.3.1 N-Terminal Domain Histone 4 (NTD-H4)

Being that the objective of this study is to identify and characterize the N-terminal tail domain of H4 when it is in the compact nucleosome array form, the first task would be to unambiguously assign the backbone of the protein. The protein sample was first checked using MALDi-TOF MS where the m/z value corresponds to the molecular weight of the protein of interest, 11262 (M^+H), as shown in the table below (**Appendix C1**).

Table 4.1: Mass Spectra of unlabeled and ^{15}N labeled Histone 4

	m/z Ratio	
Histone 4	Calculated	Measured
Unlabeled	11263.12	11262.7
Labeled	11397.93	11364.4

4.3.2 NMR backbone assignments of Histone 4

The ^{15}N, ^{13}C labeled H4 was then analyzed using 2D [^{15}N, ^{1}H]-TROSY experiments (**Fig. 23**). The spectrum of H4 showed moderate spectral dispersion, leading to the measurements of 3-dimensional experiments to aid in the sequence-specific backbone assignment. The 3D experiments measured for NTD-H4 protein were HNCA, HNCACB, CBCA(CO)NH, ^{15}N-resolved NOESY, HNCO and HN(CA)CO. 2D [^{15}N,^{1}H]-TROSY spectrum of H4 in DMSO-*d6* shows a chemical shift range of approximately 1 ppm in the ^{1}H dimension and 20 ppm in the ^{15}N dimension. At 600 MHz, 85 backbone ^{1}H-^{15}N cross-peaks are observed, out of the expected 108 peaks.

The 2D [^{15}N, ^{1}H]-TROSY spectrum of H4 (**Fig. 4.4**) illustrates the low chemical shift dispersion of the unfolded protein; this makes it difficult to unambiguously complete the sequential resonance assignment. Resonance overlap is observed throughout the spectra for both the 2D and 3D experiments. In contrast to folded proteins, unfolded proteins experience poor chemical shift dispersion for $^{13}C^{\alpha}$, $^{1}H^{\alpha}$, and $^{13}C^{\beta}$ resonances, whereas for ^{15}N, ^{1}HN and ^{13}CO the reverse is observed. This reflects the sensitivity of the latter group

of nuclei in their surrounding environment. In unfolded proteins, the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift is very much dependent on the spin system itself and not on the 3D structure (resulting from the absence of it), whereas the 13CO chemical shift is influenced by both the *i* and *i*+1 residues residue attached to it (sequence dependent) [191]. Therefore the chemical shift dispersion of ^{13}CO and ^{15}N is dependent on the amino acid sequence of H4 as opposed to $^{13}C^{\alpha}$ and $^{13}C^{\beta}$, which are dependent on the secondary and tertiary structures of the protein. Thus the regular method of sequential assignment using the standard 3D heteronuclear experiments, namely HNCA, HNCACB and CBCA(CO)NH, would not be suitable for this protein. A different strategy was employed to determine the backbone sequence that made use of the ^{13}CO, ^{15}N and ^{1}H resonances that were observed to have well dispersed chemical shifts [192].

Based on the strategy indicated above [191], the HNCO spectrum was used in distinguishing the interresidue peaks from HN(CA)CO. The HNCACB and CBCA(CO)NH were used for sequential connectivities as well as determining residue type. Apart from this, CBCA(CO)NH gives information regarding the preceding residue, which is then used in identifying tripeptide fragments. Sequence specific assignment for H4 is indicated in **Figure 4.4** for assigned residues.

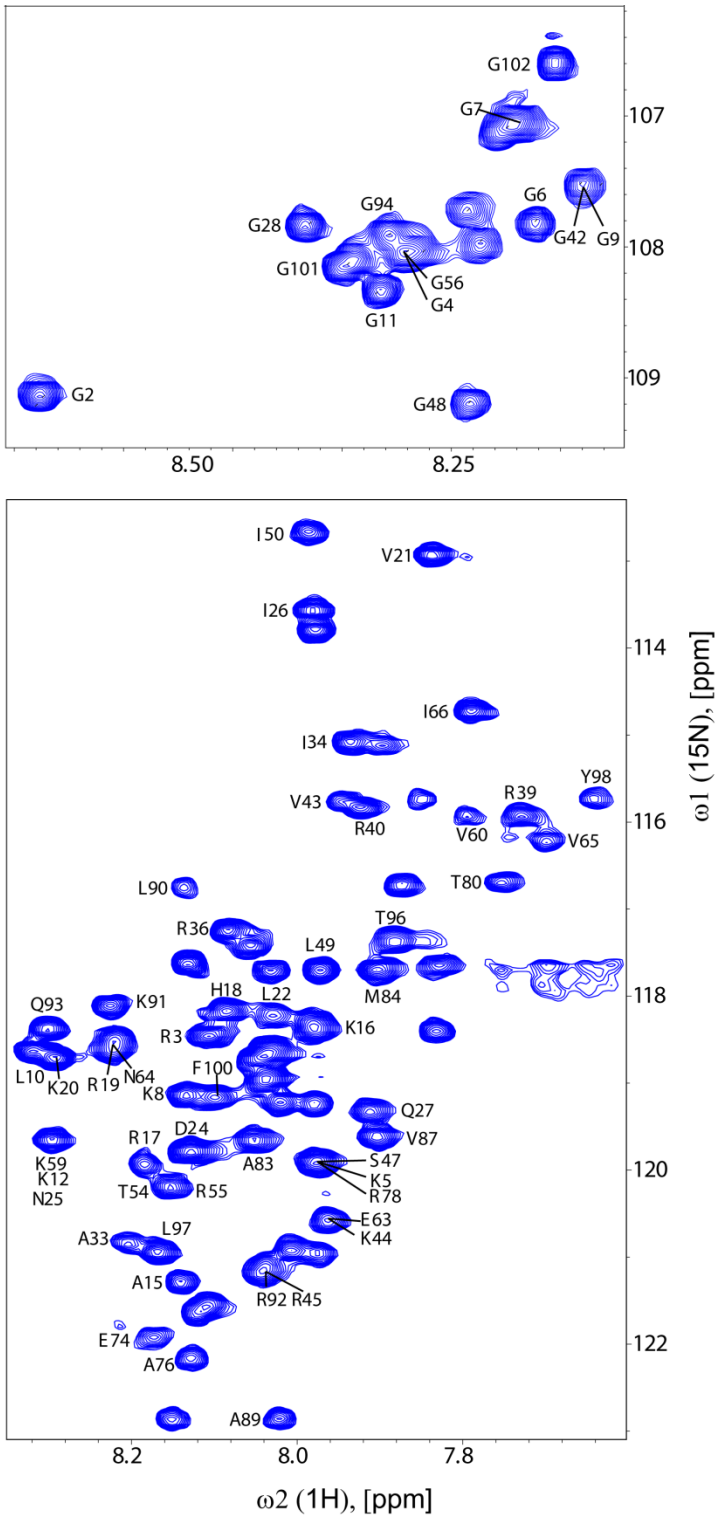


Figure 4.4. 2D [^{15}N, ^{1}H]-TROSY of ^{13}C/^{15}N-labeled NTD-H4 in 95%DMSO-*d6*/5% D_2O, 0.01% TFA-*d3*, at 298K, using 600 MHz NMR spectrometer. The protein is fully protonated in the DMSO/H_2Obuffer and most of the amide protons expected from the sequence of H4 are observed. Assigned residues are denoted with one-letter code for amino acid and residue number according to the primary sequence.

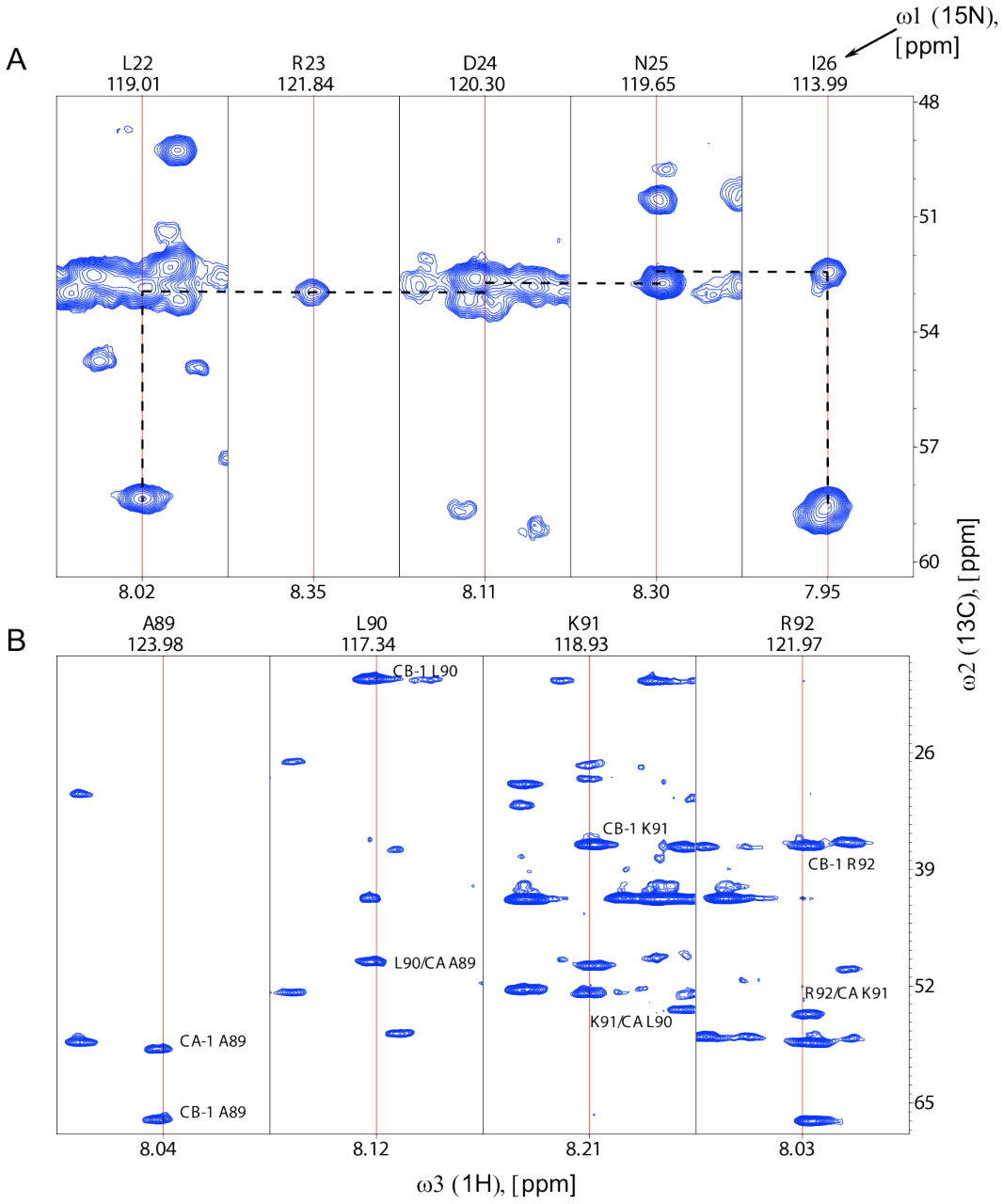


Figure 4.5. Slices taken from [^{13}C, ^{1}H]-planes of the three-dimensional experiments**.** A) HNCA and B) CBCAcoNH spectra measured with uniformly ^{15}N/^{13}C-labeled sample of free NTD-H4 in 95%DMSO-*d6*/5% D_2O, 0.01% TFA-*d3*, at 298K, using 600 MHz NMR spectrometer.

Table 4.2: Assigned residues

Residue	H	N	Residue	H	N
G2	8.718	109.151	S47	7.963	120.008
R3	8.09	119.174	G48	8.218	109.194
G4	8.294	107.909	L49	7.961	118.403
K5	7.984	120.236	I50	7.956	113.122
G6	8.138	107.806	T54	8.146	120.998
G7	8.168	106.931	R55	8.136	120.795
K8	8.151	119.323	G56	8.295	108.105
G9	8.097	107.446	K59	8.293	119.25
L10	8.287	119.006	V60	7.743	116.008
G11	8.337	108.359	E63	7.986	120.505
K12	8.29	119.424	N64	8.206	119.176
A15	8.14	120.999	V65	7.727	116.378
K16	7.981	119.041	I66	7.779	115.085
R17	8.184	120.859	E74	8.156	122.808
H18	8.074	119.026	A76	8.103	122.542
R19	8.209	118.888	R78	8.008	120.148
K20	8.292	119.274	T80	7.761	117.388
V21	7.816	113.196	A83	8.032	120.357
L22	8.02	119.013	M84	7.889	118.237
R23	8.349	121.839	V87	7.901	120.153
D24	8.114	120.297	A89	8.043	123.975
N25	8.283	120.373	L90	8.117	117.342
I26	7.951	113.994	K91	8.207	118.927
Q27	7.897	120.178	R92	8.031	121.972
G28	8.413	107.825	Q93	8.288	119.096
A33	8.166	120.684	G94	8.322	107.91
I34	7.906	115.764	T96	7.881	117.944
R36	8.084	117.928	L97	8.165	120.693
R39	7.684	116.007	Y98	7.565	115.969
R40	7.916	115.99	F100	8.107	119.153
G42	8.082	107.53	G101	8.363	108.149
V43	7.939	116.514	G102	8.125	106.205
K44	7.956	121.453			
R45	8.02	121.981			

4.3.3 Hydrogen Exchange

Hydrogen exchange occurs between amide protons of a protein with their D_2O counterpart and this method has been used extensively to study protein folding and dynamics. The exchange rate of these amides are determined by their function in the protein structure; meaning it would be slowed/delayed if that particular amide proton is involved in hydrogen bond formation or blocked by side chains in a folded structure. α-helices and β-sheets would only exchange upon unfolding of the protein, caused by loss of hydrogen bonds in the main chain. Slow exchange has been observed especially for many DNA-binding proteins during the assembly into higher order complexes (24, 25 replacement paper). Therefore, it is useful in the case of NTD-H4 for identifying the amides that do not participate in such interactions, which are free, as these NTDs of histones play a crucial role in the regulation of high-ordered structures of chromatin apart from regulating gene functions [190]. The hydrogen exchange experiment was carried out on nucleosome arrays to characterize the conformational states of the N-terminal tail domain of H4 and to observe the dynamics of the histone core as it is altered by array folding. H/D exchange was done at different time points on extended arrays and compact arrays to detect changes on the most rapidly exchanging regions.

Hydrogen-deuterium exchange is a method that is currently used to measure the extent to which backbone amide protons exchange with the deuterons that are present in the system. The most vulnerable protons are those that are directly exposed to the solvent, not buried in the core and/or not involved in hydrogen bonds. Exchange also occurs in the α-helices and β-sheets of folded regions when unfolding of the protein take place. We have applied this method to elucidate the dynamics of the histone backbone upon array compaction, by monitoring the H/D exchange of compact and extended nucleosome arrays (1, 2, 10 and 30 min).

The lack of Mg^{2+} cations in the nucleosome array results in a non-compact chromatin structure, and only with the addition of 1-2 mM Mg^{2+} cations can this situation be reversed mediated by internucleosomal contacts [173, 193]. Nucleosomal arrays were assembled via addition of the four histones (H2A, H2B, H3 and H4) at molar ratios of 1:1:1.2:1.2 to the DNA template [175]. The DNA template used in this study consists of 12 tandem repeats of 177 bp of the high affinity 601 sequence [152]. This was then followed

by salt dialysis using KCl instead of the conventional NaCl. KCl was used in this study as it was found to produce more homogenous nucleosomal arrays and the fact K^+ is a major cation in the cell cytoplasm was also a determining factor [173, 194]. TK buffer that consists of 10mM K+ and 10 mM Tris-HCl was used as the final buffer in assembly of the nucleosomal array. This maintained the concentration of monovalent ions to be in the low salt range whereby the nucleosomal arrays exist in extended "bead-on-a-string" conformation. Intranucleosomal interactions which lead to the compact arrays are constrained when cations such as Mg^{2+} is lacking. Addition of divalent cations, Mg^{2+} at 1 mM concentration was done to produce nucleosomal array folding mediated by internucleosomal interactions leading to the formation of a compact array [151, 175]. H/D exchange experiments were then performed on both the extended nucleosomal arrays ("bead-on-a-string") and compact nucleosomal arrays.

Nucleosome array was assembled by addition of histone octamer to purified DNA [173]. We have used molar ratio 1:1:1.2:1.2 for H2A, H2B, H3 and H3 during histone octamer and observed in this ratio has the best yield to get histone octamer [195]. The same procedure as wild type nucleosome array was used to prepare the H4 histone labeled. After the reconstitution, the nucleosome array was purified by addition 5 mM MgCl2 followed by centrifugation at 20000g, re- dissolved in TEK buffer and dialysis to remove of trace amount of Mg2+, then digested by ScaI to check the saturation degree and load to 6% EMSA gel. Also sedimentation velocity behavior of both H4-Labelled nucleosome array is similar to wild type in bead-on-a-string form.

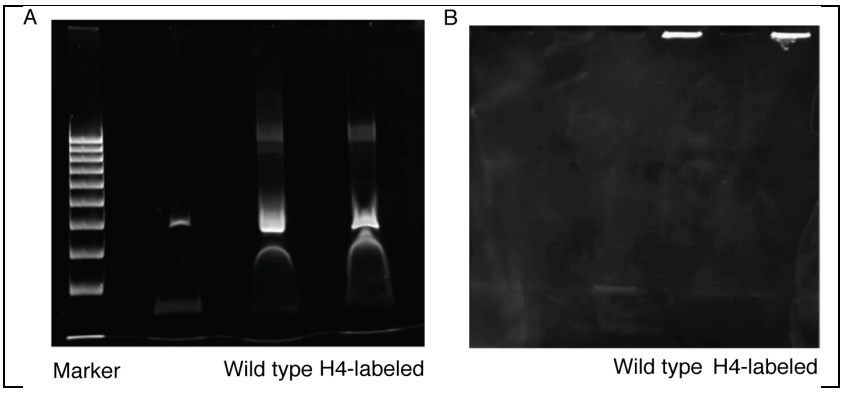


Figure 4.6. A)This gel analyzed the migration and saturation of ScaI digested of both wild type and H4-labled nucleosome , B)This gel analyzed the migration and purity of both wild type and H4-labled nucleosome before digestion by ScaI (Abdollah Allahverdi et al, personal communication)

The experiment was first carried out on non-exchanged extended arrays to determine the suitability of the method for the array system. We were able to reconstitute the H4 in the DMSO-*d6*, however at a lower concentration (**Fig. 4.7**). DMSO is a suitable quench solvent, as it does not have any exchangeable protons and therefore slows the H/D exchange. The lyophilized sample might contain trace amounts of H_2O in them, therefore using DMSO-*d6* as the final NMR buffer would ensure that no further exchange occurs during spectral acquisition. NMR assignments were readily transferred from the H4 only in DMSO-*d6* to the H4-array; however, several peaks experienced shifts and overlap of signals in the H4-array sample.

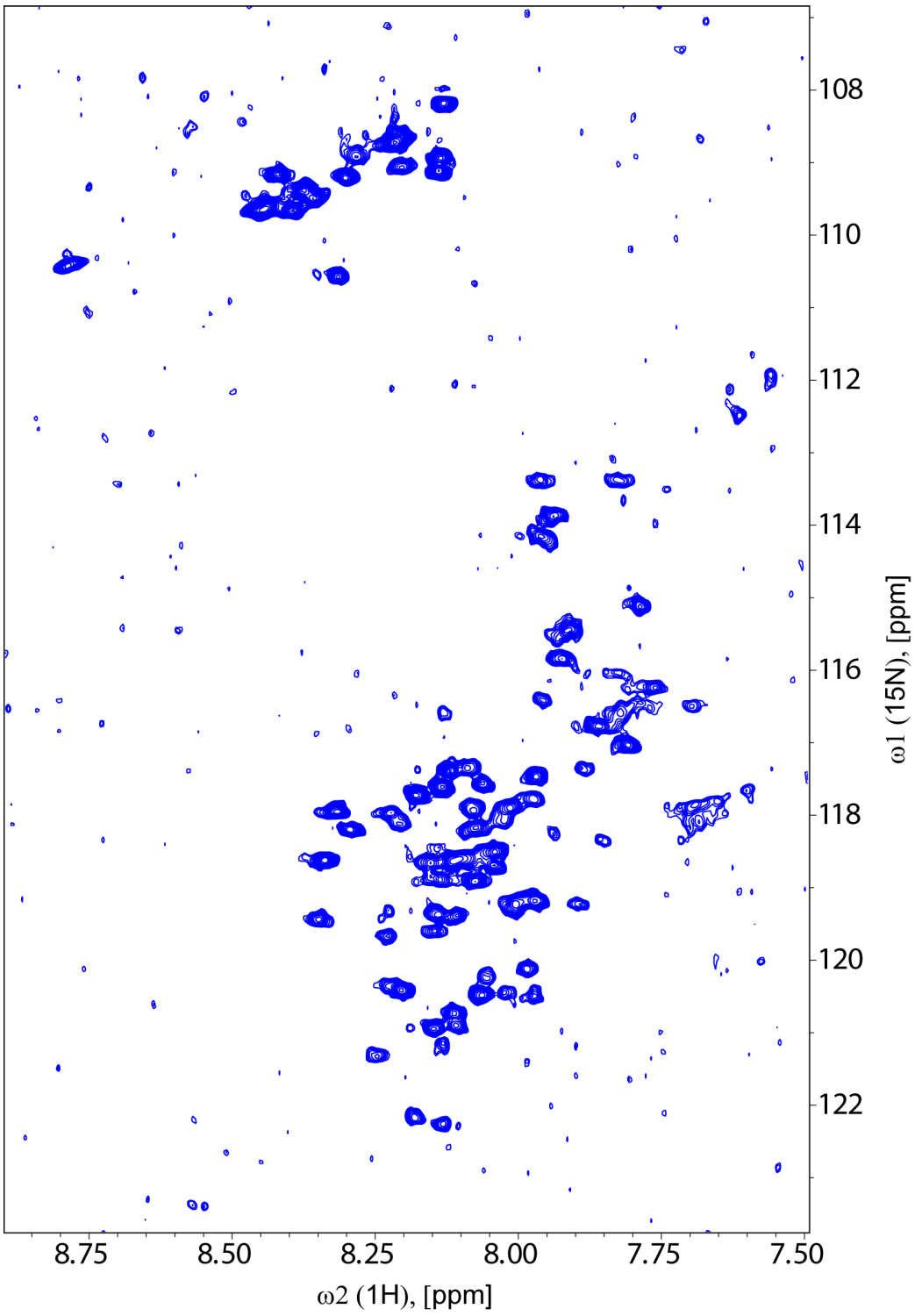


Figure 4.7. 2D [^{15}N,^{1}H]-TROSY of the lyophilized array without H/D exchange (reference spectrum) dissolved in 95%DMSO-*d6*/5% D_2O, 0.01%TFA-*d3*, at 298K, using 700 MHz NMR spectrometer. The nucleosome array dissociates such that only the histone proteins remain in the DMSO-*d6* solvent and DNA molecules would remain as precipitate. Only H4 is ^{15}N-labeled.

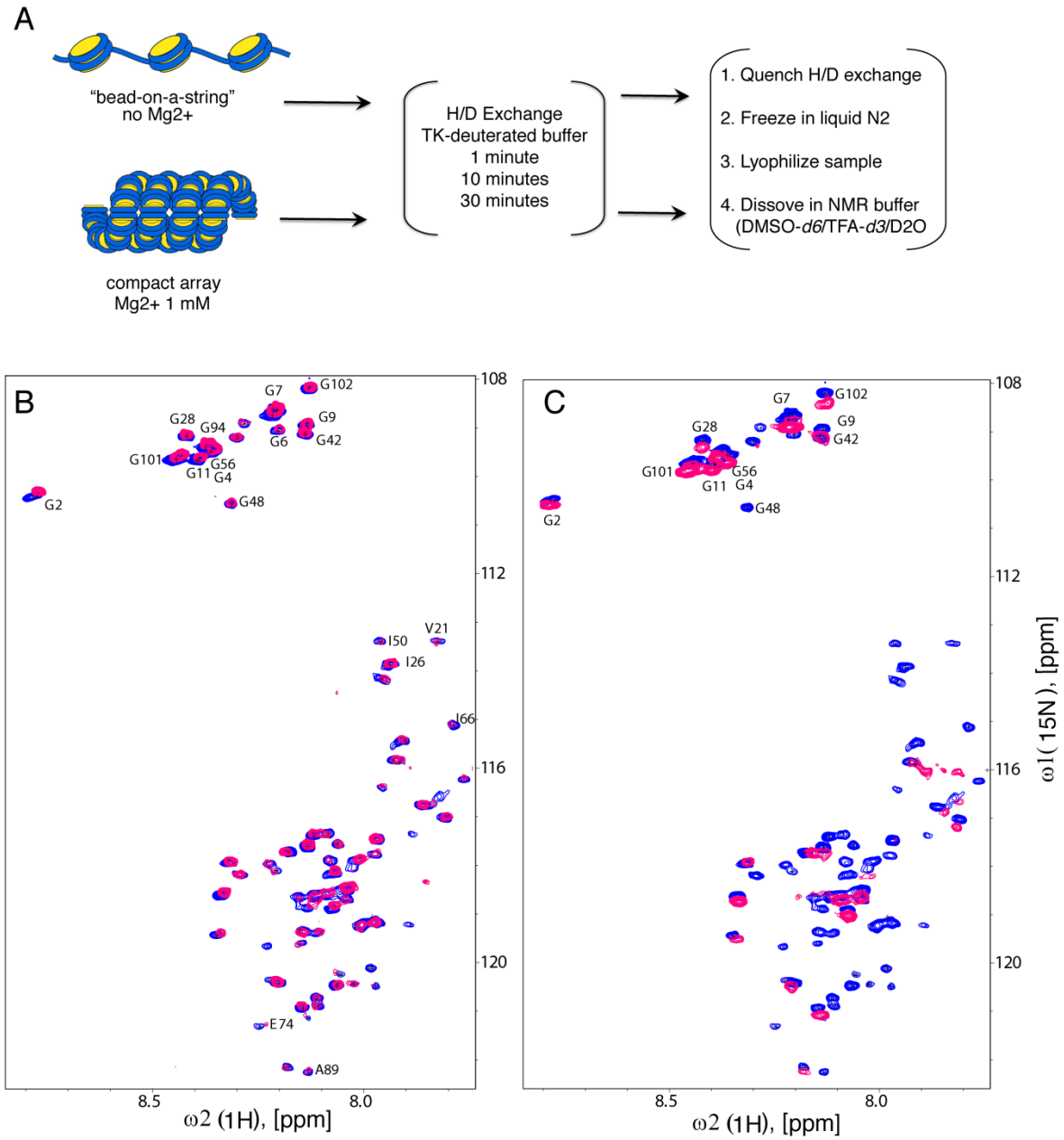


Figure 4.8. H/D exchanged of array A) Experimental scheme for determining the H/D exchange. The 2D [^{15}N,^{1}H]-TROSY of the H/D exchanged lyophilized extended array (bead-on-a-string) dissolved in 95%DMSO-*d6*/5% D_2O, 0.01%TFA-*d3*, at 298K, using 700 MHz NMR spectrometer (reference peak - blue). B) 10 minutes (pink) and C) 30 minutes (pink) time point after initiation of the H/D exchange process.

Figure 4.8b and c depicts the 2D [^{15}N, ^{1}H]-TROSY of array in the exchange buffer conditions but at higher concentrations of extended array (bead-on-a-string). At these conditions, the array was observed to undergo H/D exchange at 10 minutes (most peaks are still observed) and at 30 minutes of incubation time, majority of peaks have experienced exchange. The initial spectrum of non-exchange has 93 picked cross peaks (includes overlapping peaks and unassigned residues), and after 30 minutes, only 38 of these cross peaks are remaining with measurable intensity. The remaining cross peaks represents the amide protons that could possibly be used as probes in the H/D exchange experiment. The bar graph in **Figure 4.9** shows the peak intensities before and after the amide protons undergo exchange rates for 67 of the assigned backbone amide protons that were derived from the previously assigned cross peaks in the 2D [^{15}N,^{1}H]-TROSY spectrum. The region in the spectrum corresponding to glycine residues is observed to have relatively intense peaks compared to the other peaks present, which could be attributed to slow exchanging amides for the residues involved.

Following this, H/D exchange was performed on compact nucleosome arrays (AUC conditions) (**Fig. 4.10**). A compact nucleosome solution was first prepared by the addition of 1 mM Mg^{2+} to Nucleosome array at OD of < 2.0. This solution was then left to incubate for approximately 15 minutes to achieve maximum and overall compaction. Deuterated TK buffer supplemented with 1 mM Mg^{2+} was added to the compact nucleosomal array in equal amounts, so that the final concentration of Mg^{2+} in the solution is 1 mM and the final concentration of the nucleosomal array is at an OD of c.a. 1.0 (AUC conditions) [194]. Throughout the time points of the H/D exchange experiment, the exchange that occurred was very fast and almost the whole protein is affected.

This suggests a lack of folding dependent protection throughout the nucleosome core. One of the contributing factors for this observation could be the lowered concentration of compact arrays present in the exchange buffer as it is the required conditions for producing compact nucleosomal arrays. Although the protection of histones in compact nucleosomal arrays is supposed to be increased upon array folding (lowered H/D exchange), it was observed that the magnitude of protection was less pronounced in the same H4 at this shorter time point. It was observed that the exchange is faster post folding conditions compared to extended array (*i.e.* "beads-on-a-string"; without Mg^{2+}).

This suggest that histone 4 is still flexible in folded chromatin, albeit it being in a compact state. In a compact nucleosomal array, the tail of H4 is believed to be involved in internucleosomal interactions, while the H4 itself is in the core wrapped around the 150 bp DNA. The H/D exchange experiments show that these histone octamers could very well still maintain local flexibility at these conditions.

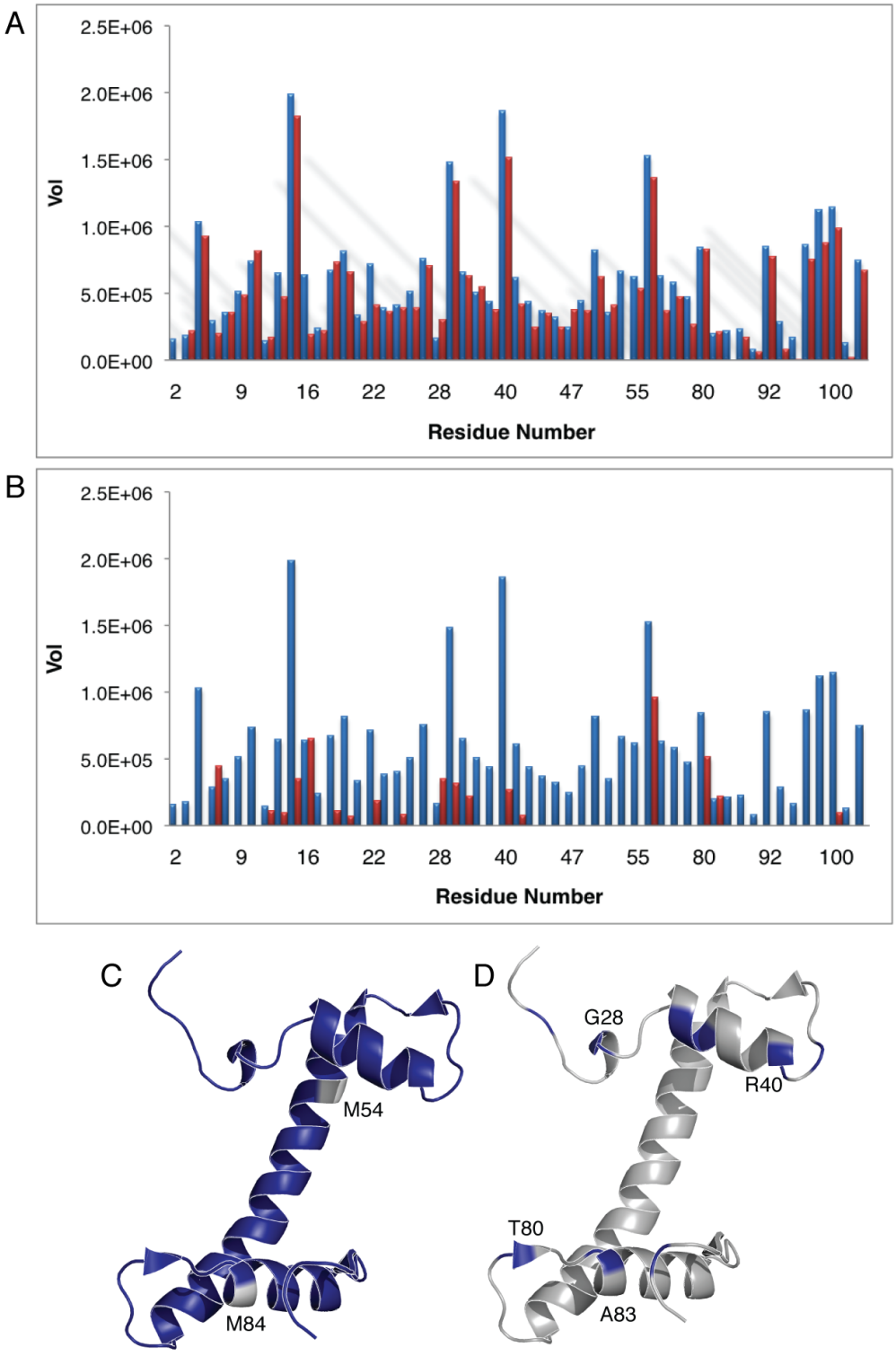


Figure 4.9. H/D exchange of H4 in extended array based on peak intensities A) 10 minutes exchange and B) 30 minutes exchange of individual amide protons of H4 in non-compact array for both time points. Bars indicate the peak intensity (blue- reference, red-after exchange). The lower panel cartoon structure is a representation of H4 derived from the crystal structure of the NCP (PDB:1AOI) with A) 10 mins and B) 30 mins exchange: grey-exchanged, blue-non-exchanged.

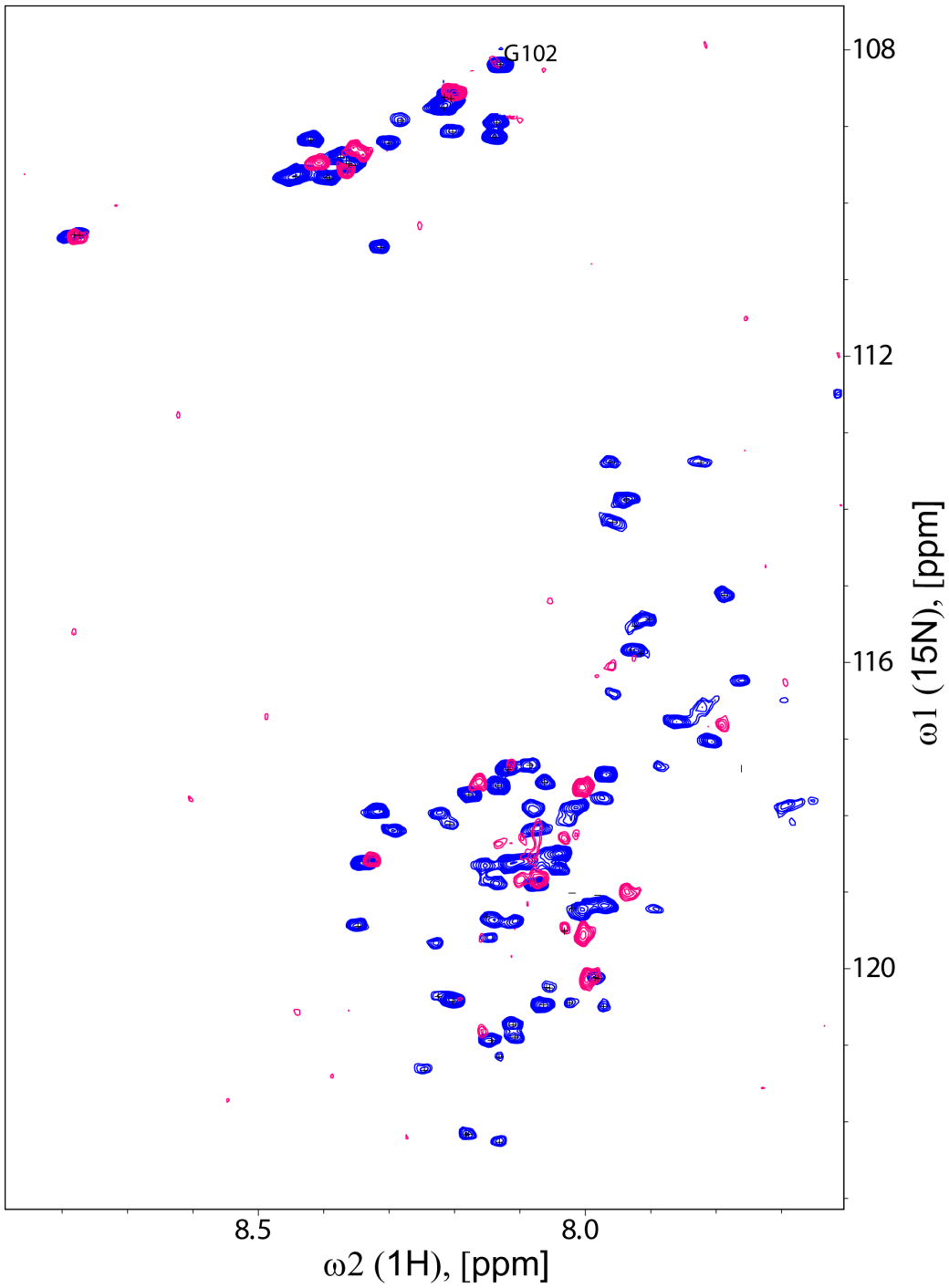


Figure 4.10. 2D [^{15}N,^{1}H]-TROSY of the H/D exchanged lyophilized compact array (pink) overlay reference spectrum (blue), both dissolved in 95%DMSO-*d6*/5% D_2O, 0.01%TFA-*d3*, at 298K, using 700 MHz NMR spectrometer. The H/D exchange was performed on a compact array (with 1 mM Mg^{2+} present in the system at AUC conditions), 1 minute time point after initiation of the H/D exchange process (pink).

4.4 Conclusion

We set out to examine the dynamics of nucleosome arrays in terms of its histone tails. There have been limited studies done on NCP and compact nucleosomal arrays using high-resolution NMR methods. In order to identify and understand the dynamics of H4 at a molecular and structural level, we have employed high resolution NMR in our studies. The elucidation of the backbone amide of H4 prior to any dynamic studies is mandatory as it will enable us to determine the site specific changes that occur. The backbone assignment of unfolded proteins is however a daunting task due to the severe degeneracy of resonances. This challenge was however overcome to a certain extent whereby 67 of the backbone residues have been assigned using selected 3-dimensional experiments.

To study the dynamics of the H4 tails in compact nucleosomal arrays, a comparison of extended (beads-on-a-string) array and compact array was done. Hydrogen exchange experiments were performed on both conditions of arrays and residues that experienced exchange were identified. Our results indicate that the conditions required for producing compact arrays (low concentrations) results in almost complete exchange of H4. The more pronounced exchange results of compact arrays indicate that it is highly probable that the H4 protein is still flexible even in the compact (folded) form as compared to the extended form. This could also be due to the fact that the higher concentration of extended array used in H/D exchange experiments allowed for protein-protein interaction as well as internucleosomal interaction to occur, thus slowing the H/D exchange of the H3/H4 tetramers Interestingly, in the extended protein, the few non-exchanged residues are predominantly in the N-terminal and C-terminal of the protein. Based on the above observations and postulations, it would be of interest to investigate the dynamics of histone octamers in solution in the absence of DNA to elucidate the possible sites of interactions.

More work is needed to answer the questions that we have begun to ask with regards to the dynamics of the NTD-H4 in compact nucleosomal arrays. We have paved the way by establishing a suitable method to study H/D exchange on these systems, which could be further extended in studying arrays with posttranslational modifications (PTMs) that have been shown to have biological significance, in addition to studying the dynamics of the other histone proteins at a molecular and structural level.

CHAPTER V

CHAPTER V
Conclusions

Three bodies of work have been presented in this thesis related to NMR-based studies of protein classes involved in DNA/RNA recognition. My focus was to use the advanced solution NMR methods in studying the interactions that exist between proteins and nucleic acids, specifically translation termination factors and transcription factors.

In the first subproject, structural changes in a stop-codon recognizing protein induced by mutations altering stop-codon recognition specificity were examined. In translation termination, eRF1 is able to decode all three stop codons, however the mechanism of stop codon recognition by eRF1 leading to hydrolysis of the ester bond on the peptidyl tRNA remains obscure. We have proposed a model of the binding mode of eRF1 to pre-translation termination ribosomal complex based on our solution structures of wild-type and several mutants of eRF1's N-domain as well as its interactions with model RNA and M- and C-domains of the same protein. Due to the global character of structural perturbation induced by mutations, Residual Dipolar Coupling (RDC) data using ^{15}N labeled sample aligned with *Pf1* phages and chemical shift perturbation (CSP) based on 2D [^{15}N, ^{1}H]-TROSY were analysed. One could surmise that the eRF1 mutant that ceases to interact with one or two out of the 3 nucleotides in a stop codon could still function as a RF, however with reduced efficiency. Based on this observation, we propose that mutations that alter the selectivity of stop codon recognition modulate the different conformations of the GTS loop that is required for eRF1 to function efficiently. Mutational studies have been done on the residues constituting the hydrophobic core directly above the GTS loop, namely residues I35, V71, V78 and C127 [60, 67]. We postulate from this observation that the sampling of different conformations by the GTS loop results in the repositioning of helix α-3 in the N-domain of eRF1, where S70 plays an interesting role of stabilization of the GTS loop in $Q^{122}FM(Y)F^{126}$. Based on our current data, we have reinforced that a degree of complexity is involved in the stop codon decoding mechanism that requires the participation of different regions to modulate conformational changes that serves as a requisite for translation termination to occur.

In the second project, we collaborated with Dr. Ralf Jauch in order to provide structural basis for activity of novel drug candidates interfering with transcription factor/cognate DNA interactions. The DNA binding domains of transcription factors have so far been considered too impervious to be tackled as drug targets although up-regulated transcription factors are a major cause of cancer and other diseases. Here we identified a Dawson-POM as an unconventional but potent compound to inhibit the DNA binding activity of Sox2. We used NMR to locate binding site of the drug candidates on Sox2. The mode of interaction of the Dawson-POM with the Sox2-HMG domain involves predominantly electrostatic interactions at the pocket just outside of the DNA binding region, but still adequately positioned to compete with the negatively charged DNA backbone. This study also addressed the question pertaining to the importance of Sox2 proteins and their co-DNA binding partners in transcription activation. It has been proposed that Sox2 and Pax6 cooperatively regulate the initiation of lens development, and it is this co-partnering with different DNA binding factors that determines their regulatory targets [129]. We have shown from the NMR results that the addition of the co-DNA binding partner Pax6 leads to a cooperative increase in DNA binding affinity for Sox2-HMG. Control of transcriptional activation is dependent on the combinatorial control of transcription factors, and the decreased dissociation of Sox2 incurred by the presence of Pax6 results in a Sox2-HMG_DC5_Pax6 ternary complex that fulfills this requirement.

The objective of this study is to assess structural flexibility of N-terminal tail domain of Histone 4 when it is in the compact nucleosome array using deuterium/protons exchange experiment in NCP. ^{15}N labeled Histone 4 was prepared and 2D [^{1}H-^{15}N]-TROSY series of experiments were carried out for protein recovered from reconstituted NCPs in H_2O followed by controlled exposure to D_2O. The sequential assignment of the backbone H4 was used to identify the dynamics of the NCP as well as arrays in the compact form. We observed that although in compact states, the magnitude of protection is not as pronounced as expected. This suggests that the NCP is still flexible, albeit being in a compact state.

The critical contribution of NMR spectroscopy is the feasibility to study weak but biologically significant interactions for dynamic processes. We reckon that our approach has successfully shed light on parts of the highly complicated mechanism of translation

termination and transcription activation in eukaryotes, including the physical nature of compact nucleosomal arrays.

BIBLIOGRAPHY

Bibliography

1. Calladine, C.R., *Understanding DNA the molecule & how it works*, 2004, Elsevier Academic Press,: San Diego, CA. p. 1 online resource (xiii, 334 p.
2. Jamin, N. and F. Toma, *NMR studies of protein-DNA interactions.* Progress in Nuclear Magnetic Resonance Spectroscopy, 2001. **38**: p. 83-114.
3. Downing, A.K., *Protein NMR techniques*. 2nd ed. Methods in molecular biology2004, Totowa, N.J.: Humana Press. xi, 487 p.
4. Wong, I. and T.M. Lohman, *A double-filter method for nitrocellulose-filter binding: application to protein-nucleic acid interactions.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(12): p. 5428-32.
5. Lane D, P.P.a.C.M., *Use of gel retardation to analyze protein–nucleic acid interactions.* Microbiological Reviews 1992. **56**: p. 509-528.
6. Heyduk, E. and T. Heyduk, *Mapping protein domains involved in macromolecular interactions: a novel protein footprinting approach.* Biochemistry, 1994. **33**(32): p. 9643-50.
7. Strauch, M.A., *Protein-DNA Interactions:Techniques Used*, in *Encylopedia of Life Sciences*2001, John Wiley & Sons, Ltd.
8. Dey, B., et al., *DNA-protein interactions: methods for detection and analysis.* Molecular and cellular biochemistry, 2012. **365**(1-2): p. 279-99.
9. Wu, C., *An exonuclease protection assay reveals heat-shock element and TATA box DNA-binding proteins in crude nuclear extracts.* Nature, 1985. **317**(6032): p. 84-7.
10. Baldwin, A.S., Jr., M. Oettinger, and K. Struhl, *Methylation and uracil interference assays for analysis of protein-DNA interactions.* Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.], 2001. **Chapter 12**: p. Unit 12 3.
11. Cary, P.D. and G.G. Kneale, *Circular dichroism for the analysis of protein-DNA interactions.* Methods in molecular biology, 2009. **543**: p. 613-24.
12. Kelly, S.M. and N.C. Price, *Circular dichroism to study protein interactions.* Current protocols in protein science / editorial board, John E. Coligan ... [et al.], 2006. **Chapter 20**: p. Unit 20 10.

13. Campagne, S., V. Gervais, and A. Milon, *Nuclear magnetic resonance analysis of protein-DNA interactions.* Journal of the Royal Society, Interface / the Royal Society, 2011. **8**(61): p. 1065-78.
14. Billeter, M., et al., *Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex.* Journal of Molecular Biology, 1993. **234**(4): p. 1084-93.
15. Pervushin, K., et al., *Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(23): p. 12366-71.
16. Gorlach, M., et al., *Interaction of the RNA-binding domain of the hnRNP C proteins with RNA.* The EMBO journal, 1992. **11**(9): p. 3289-95.
17. Howe, P.W., et al., *NMR studies of U1 snRNA recognition by the N-terminal RNP domain of the human U1A protein.* The EMBO journal, 1994. **13**(16): p. 3873-81.
18. Kalodimos, C.G., R. Boelens, and R. Kaptein, *A residue-specific view of the association and dissociation pathway in protein--DNA recognition.* Nature structural biology, 2002. **9**(3): p. 193-7.
19. Fesik, S.W., et al., *Isotope-edited proton NMR study on the structure of a pepsin/inhibitor complex.* Biochemistry, 1988. **27**(22): p. 8297-301.
20. Chen, K. and N. Tjandra, *The Use of Residual Dipolar Coupling in Studying Proteins by NMR.* Topics in current chemistry, 2012. **326**: p. 47-68.
21. Tjandra, N., et al., *Use of dipolar H-1-N-15 and H-1-C-13 couplings in the structure determination of magnetically oriented macromolecules in solution.* Nature Structural Biology, 1997. **4**(9): p. 732-738.
22. Zuiderweg, E.R., *Mapping protein-protein interactions in solution by NMR spectroscopy.* Biochemistry, 2002. **41**(1): p. 1-7.
23. Raschke, T.M. and S. Marqusee, *Hydrogen exchange studies of protein structure.* Current opinion in biotechnology, 1998. **9**(1): p. 80-6.
24. Bai, Y., et al., *Primary structure effects on peptide group hydrogen exchange.* Proteins, 1993. **17**(1): p. 75-86.

25. Gryk, M.R. and O. Jardetzky, *AV77 hinge mutation stabilizes the helix-turn-helix domain of trp repressor.* Journal of Molecular Biology, 1996. **255**(1): p. 204-14.
26. Cavanagh, J., *Protein NMR spectroscopy : principles and practice*. 2nd ed2007, Amsterdam ; Boston: Academic Press. xxv, 885 p.
27. Wider, G., *Structure determination of biological macromolecules in solution using nuclear magnetic resonance spectroscopy.* BioTechniques, 2000. **29**(6): p. 1278-82, 1284-90, 1292 passim.
28. Rule, G.S. and T.K. Hitchens, *Fundamentals of protein NMR spectrosopy*. Focus on structural biology2006, Dordrecht: Springer. xxvi, 530 p.
29. Wuthrich, K., *Determination of three-dimensional protein structures in solution by nuclear magnetic resonance: an overview.* Methods in enzymology, 1989. **177**: p. 125-31.
30. Clore, G.M. and A.M. Gronenborn, *Multidimensional heteronuclear nuclear magnetic resonance of proteins.* Methods in enzymology, 1994. **239**: p. 349-63.
31. Bax, A., *Weak alignment offers new NMR opportunities to study protein structure and dynamics.* Protein Science, 2003. **12**(1): p. 1-16.
32. Takeuchi, K. and G. Wagner, *NMR studies of protein interactions.* Current opinion in structural biology, 2006. **16**(1): p. 109-17.
33. Raman, S., et al., *NMR Structure Determination for Larger Proteins Using Backbone-Only Data.* Science, 2010. **327**(5968): p. 1014-1018.
34. Lipsitz, R.S. and N. Tjandra, *Residual dipolar couplings in NMR structure analysis.* Annual Review of Biophysics and Biomolecular Structure, 2004. **33**: p. 387-413.
35. Tjandra, N. and A. Bax, *Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium.* Science, 1997. **278**(5340): p. 1111-4.
36. Prestegard, J.H., H.M. Al-Hashimi, and J.R. Tolman, *NMR structures of biomolecules using field oriented media and residual dipolar couplings.* Quarterly Reviews of Biophysics, 2000. **33**(4): p. 371-424.
37. Prestegard, J.H., et al., *Determination of protein backbone structures from residual dipolar couplings.* Methods in enzymology, 2005. **394**: p. 175-209.

38. Tolman, J.R. and K. Ruan, *NMR residual dipolar couplings as probes of biomolecular dynamics.* Chemical reviews, 2006. **106**(5): p. 1720-36.
39. Bax, A., G. Kontaxis, and N. Tjandra, *Dipolar couplings in macromolecular structure determination.* Nuclear Magnetic Resonance of Biological Macromolecules, Pt B, 2001. **339**: p. 127-174.
40. Fleming, K. and S. Matthews, *Media for studies of partially aligned states.* Methods in molecular biology, 2004. **278**: p. 79-88.
41. Ding, K. and A.M. Gronenborn, *Measurement of Residual Dipolar Couplings and Applications in Protein NMR.* Modern Magnetic Resonance, 2006: p. 1287-1291.
42. Hansen, M.R., L. Mueller, and A. Pardi, *Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions.* Nature Structural Biology, 1998. **5**(12): p. 1065-1074.
43. Ruckert, M. and G. Otting, *Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments.* Journal of the American Chemical Society, 2000. **122**(32): p. 7793-7797.
44. Tycko, R., F.J. Blanco, and Y. Ishii, *Alignment of biopolymers in strained gels: A new way to create detectable dipole-dipole couplings in high-resolution biomolecular NMR.* Journal of the American Chemical Society, 2000. **122**(38): p. 9340-9341.
45. Hansen, M.R., P. Hanson, and A. Pardi, *Filamentous bacteriophage for aligning RNA, DNA, and proteins for measurement of nuclear magnetic resonance dipolar coupling interactions.* Rna-Ligand Interactions Pt A, 2000. **317**: p. 220-240.
46. Chou, J.J., et al., *A simple apparatus for generating stretched polyacrylamide gels, yielding uniform alignment of proteins and detergent micelles.* Journal of biomolecular NMR, 2001. **21**(4): p. 377-82.
47. Nierhaus, K.H., et al., *Protein synthesis: Mechanism of the elongation of the nascent polypeptide chain on the ribosome.* European Journal of Cell Biology, 1997. **72**: p. 35-35.
48. Schimmel, P., *The protein synthesis world.* Trends in Biochemical Sciences, 2002. **27**(10): p. 505-505.

49. Frank, J. and C.M.T. Spahn, *The ribosome and the mechanism of protein synthesis.* Reports on Progress in Physics, 2006. **69**(5): p. 1383-1417.
50. Sonenberg, N., J.W.B. Hershey, and M. Mathews, *Translational control of gene expression*. 2nd ed. Cold Spring Harbor monograph series,2000, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. x, 1020 p.
51. Lackner, D.H. and J. Bahler, *Translational control of gene expression from transcripts to transcriptomes.* International review of cell and molecular biology, 2008. **271**: p. 199-251.
52. Frolova, L.Y., T.I. Merkulova, and L.L. Kisselev, *Translation termination in eukaryotes: Polypeptide release factor eRF1 is composed of functionally and structurally distinct domains.* Rna-a Publication of the Rna Society, 2000. **6**(3): p. 381-390.
53. Kapp, L.D. and J.R. Lorsch, *The molecular mechanics of eukaryotic translation.* Annu Rev Biochem, 2004. **73**: p. 657-704.
54. Nakamura, Y., K. Ito, and L.A. Isaksson, *Emerging understanding of translation termination.* Cell, 1996. **87**(2): p. 147-150.
55. Kisselev, L., M. Ehrenberg, and L. Frolova, *Termination of translation: interplay of mRNA, rRNAs and release factors?* Embo Journal, 2003. **22**(2): p. 175-182.
56. Nakamura, Y. and K. Ito, *How protein reads the stop codon and terminates translation.* Genes to Cells, 1998. **3**(5): p. 265-278.
57. Nakamura, Y., K. Ito, and M. Ehrenberg, *Mimicry grasps reality in translation termination.* Cell, 2000. **101**(4): p. 349-352.
58. Frolova, L., et al., *A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor.* Nature, 1994. **372**(6507): p. 701-3.
59. Zhouravleva, G., et al., *Termination of Translation in Eukaryotes Is Governed by 2 Interacting Polypeptide-Chain Release Factors, Erf1 and Erf3.* Embo Journal, 1995. **14**(16): p. 4065-4072.
60. Bertram, G., et al., *Terminating eukaryote translation: Domain 1 of release factor eRF1 functions in stop codon recognition.* Rna-a Publication of the Rna Society, 2000. **6**(9): p. 1236-1247.

61. Frolova, L., A. Seit-Nebi, and L. Kisselev, *Highly conserved NIKS tetrapeptide is functionally essential in eukaryotic translation termination factor eRF1.* Rna-a Publication of the Rna Society, 2002. **8**(2): p. 129-136.
62. Song, H.W., et al., *The crystal structure of human eukaryotic release factor eRF1 - Mechanism of stop codon recognition and peptidyl-tRNA hydrolysis.* Cell, 2000. **100**(3): p. 311-321.
63. Kononenko, A.V., et al., *Role of the individual domains of translation termination factor eRF1 in GTP binding to eRF3.* Proteins-Structure Function and Bioinformatics, 2008. **70**(2): p. 388-393.
64. Fan-Minogue, H., et al., *Distinct eRF3 requirements suggest alternate eRF1 conformations mediate peptide release during eukaryotic translation termination.* Molecular Cell, 2008. **30**(5): p. 599-609.
65. Merkulova, T.I., et al., *C-terminal domains of human translation termination factors eRF1 and eRF3 mediate their in vivo interaction.* Febs Letters, 1999. **443**(1): p. 41-47.
66. Ebihara, K. and Y. Nakamura, *C-terminal interaction of translational release factors eRF1 and eRF3 of fission yeast: G-domain uncoupled binding and the role of conserved amino acids.* Rna-a Publication of the Rna Society, 1999. **5**(6): p. 739-750.
67. Cheng, Z., et al., *Structural insights into eRF3 and stop codon recognition by eRF1.* Genes & Development, 2009. **23**(9): p. 1106-1118.
68. Inagaki, Y., et al., *Convergence and constraint in eukaryotic release factor 1 (eRF1) domain 1: the evolution of stop codon specificity.* Nucleic Acids Research, 2002. **30**(2): p. 532-544.
69. Mantsyzov, A.B., et al., *NMR solution structure and function of the C-terminal domain of eukaryotic class 1 polypeptide chain release factor.* The FEBS journal, 2010. **277**(12): p. 2611-27.
70. Frolova, L.Y., et al., *Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis.* Rna-a Publication of the Rna Society, 1999. **5**(8): p. 1014-1020.

71. Seit-Nebi, A., et al., *Class-1 translation termination factors: invariant GGQ minidomain is essential for release activity and ribosome binding but not for stop codon recognition.* Nucleic Acids Research, 2001. **29**(19): p. 3982-3987.
72. Kobayashi, T., et al., *The GTP-binding release factor eRF3 as a key mediator coupling translation termination to mRNA decay.* Journal of Biological Chemistry, 2004. **279**(44): p. 45693-45700.
73. Seit-Nebi, A., L. Frolova, and L. Kisselev, *Conversion of omnipotent translation termination factor eRF1 into ciliate-like UGA-only unipotent eRF1.* Embo Reports, 2002. **3**(9): p. 881-6.
74. Kolosov, P., et al., *Invariant amino acids essential for decoding function of polypeptide release factor eRF1.* Nucleic Acids Research, 2005. **33**(19): p. 6418-6425.
75. Liang, H., et al., *Decoding the decoding region: Analysis of eukaryotic release factor (eRF1) stop codon-binding residues.* Journal of Molecular Evolution, 2005. **60**(3): p. 337-344.
76. Salas-Marco, J. and D.M. Bedwell, *GTP hydrolysis by eRF3 facilitates stop codon decoding during eukaryotic translation termination.* Mol Cell Biol, 2004. **24**(17): p. 7769-78.
77. Kapp, L.D. and J.R. Lorsch, *The molecular mechanics of eukaryotic translation.* Annual Review of Biochemistry, 2004. **73**: p. 657-704.
78. Ma, B.Y. and R. Nussinov, *Release factors eRF1 and RF2 - A universal mechanism controls the large conformational changes.* Journal of Biological Chemistry, 2004. **279**(51): p. 53875-53885.
79. Chavatte, L., et al., *The invariant uridine of stop codons contacts the conserved NIKSR loop of human eRF1 in the ribosome.* The EMBO journal, 2002. **21**(19): p. 5302-11.
80. Bulygin, K.N., et al., *Three distinct peptides from the N domain of translation termination factor eRF1 surround stop codon in the ribosome.* Rna-a Publication of the Rna Society, 2010. **16**(10): p. 1902-1914.
81. Lekomtsev, S., et al., *Different modes of stop codon restriction by the stylonychia and paramecium eRF1 translation termination factors.* Proceedings of the National

Academy of Sciences of the United States of America, 2007. **104**(26): p. 10824-10829.

82. Ito, K., M. Uno, and Y. Nakamura, *A tripeptide 'anticodon' deciphers stop codons in messenger RNA.* Nature, 2000. **403**(6770): p. 680-4.
83. Nakamura, Y. and K. Ito, *A tripeptide discriminator for stop codon recognition.* FEBS letters, 2002. **514**(1): p. 30-3.
84. Bidou, L., et al., *Sense from nonsense: therapies for premature stop codon diseases.* Trends in molecular medicine, 2012. **18**(11): p. 679-88.
85. Maquat, L.E., *NMD in Mammalian Cells: A Splicing-dependent Means to Down-regulate the levels of mRNAs That Prematurely Terminate Translation*, in *Translation control of gene expression*, N. Sonenberg, J. Hershey, and M.B. Mathews, Editors. 2000, Cold Spring Harbor Laboratory Press: New York. p. 849-868.
86. Linde, L. and B. Kerem, *Introducing sense into nonsense in treatments of human genetic diseases.* Trends in Genetics, 2008. **24**(11): p. 552-563.
87. Mort, M., et al., *A meta-analysis of nonsense mutations causing human genetic disease.* Human mutation, 2008. **29**(8): p. 1037-47.
88. Wong, L.E., et al., *Selectivity of stop codon recognition in translation termination is modulated by multiple conformations of GTS loop in eRF1.* Nucleic Acids Research, 2012. **40**(12): p. 5751-65.
89. Zweckstetter, M., *NMR: prediction of molecular alignment from structure using the PALES software.* Nature Protocols, 2008. **3**(4): p. 679-690.
90. Shen, Y., et al., *TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts.* Journal of biomolecular NMR, 2009. **44**(4): p. 213-23.
91. Brunger, A.T., *Version 1.2 of the Crystallography and NMR system.* Nature protocols, 2007. **2**(11): p. 2728-33.
92. Brunger, A.T., et al., *Crystallography & NMR system: A new software suite for macromolecular structure determination.* Acta crystallographica. Section D, Biological crystallography, 1998. **54**(Pt 5): p. 905-21.

93. Guntert, P., C. Mumenthaler, and K. Wuthrich, *Torsion angle dynamics for NMR structure calculation with the new program DYANA.* Journal of Molecular Biology, 1997. **273**(1): p. 283-298.

94. Herrmann, T., P. Guntert, and K. Wuthrich, *Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA.* Journal of Molecular Biology, 2002. **319**(1): p. 209-227.

95. Brunger, A.T., et al., *Crystallography & NMR system: A new software suite for macromolecular structure determination.* Acta Crystallographica Section D-Biological Crystallography, 1998. **54**: p. 905-921.

96. Brunger, A.T., *Version 1.2 of the Crystallography and NMR system.* Nature protocols, 2007. **2**(11): p. 2728-2733.

97. Bulygin, K.N., et al., *Adenine and guanine recognition of stop codon is mediated by different N domain conformations of translation termination factor eRF1.* Nucleic Acids Research, 2011. **39**(16): p. 7134-7146.

98. Eliseev, B., et al., *A single amino acid change of translation termination factor eRF1 switches between bipotent and omnipotent stop-codon specificity.* Nucleic Acids Research, 2011. **39**(2): p. 599-608.

99. Alkalaeva, E.Z., et al., *In vitro reconstitution of eukaryotic translation reveals cooperativity between release factors eRF1 and eRF3.* Cell, 2006. **125**(6): p. 1125-1136.

100. Hatin, I., et al., *Molecular dissection of translation termination mechanism identifies two new critical regions in eRF1.* Nucleic Acids Research, 2009. **37**(6): p. 1789-1798.

101. Latchman, D.S., *Eukaryotic transcription factors*. 4th ed2004, Amsterdam ; Boston: Elsevier/Academic Press. xxvi, 360 p.

102. Yang, V.W., *Eukaryotic transcription factors: identification, characterization and functions.* The Journal of nutrition, 1998. **128**(11): p. 2045-51.

103. Prior, H.M. and M.A. Walter, *SOX genes: architects of development.* Molecular medicine, 1996. **2**(4): p. 405-12.

104. Latchman, D.S., *Eukaryotic transcription factors.* The Biochemical journal, 1990. **270**(2): p. 281-9.
105. Koopman, P., et al., *Male development of chromosomally female mice transgenic for Sry.* Nature, 1991. **351**(6322): p. 117-21.
106. van Houte, L.P., et al., *Solution structure of the sequence-specific HMG box of the lymphocyte transcriptional activator Sox-4.* The Journal of biological chemistry, 1995. **270**(51): p. 30516-24.
107. King, C.Y. and M.A. Weiss, *The SRY high-mobility-group box recognizes DNA by partial intercalation in the minor groove: a topological mechanism of sequence specificity.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(24): p. 11990-4.
108. Wegner, M., *From head to toes: the multiple facets of Sox proteins.* Nucleic acids research, 1999. **27**(6): p. 1409-20.
109. Harley, V.R., R. Lovell-Badge, and P.N. Goodfellow, *Definition of a consensus DNA binding site for SRY.* Nucleic acids research, 1994. **22**(8): p. 1500-1.
110. van Houte, L., et al., *The sequence-specific high mobility group 1 box of TCF-1 adopts a predominantly alpha-helical conformation in solution.* The Journal of biological chemistry, 1993. **268**(24): p. 18083-7.
111. Connor, F., et al., *DNA binding and bending properties of the post-meiotically expressed Sry-related protein Sox-5.* Nucleic acids research, 1994. **22**(16): p. 3339-46.
112. Williams, D.C., Jr., M. Cai, and G.M. Clore, *Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1.Sox2.Hoxb1-DNA ternary transcription factor complex.* The Journal of biological chemistry, 2004. **279**(2): p. 1449-57.
113. Werner, M.H. and S.K. Burley, *Architectural transcription factors: proteins that remodel DNA.* Cell, 1997. **88**(6): p. 733-6.
114. Weiss, M.A., *Floppy SOX: mutual induced fit in hmg (high-mobility group) box-DNA recognition.* Molecular endocrinology, 2001. **15**(3): p. 353-62.

115. Wilson, M. and P. Koopman, *Matching SOX: partner proteins and co-factors of the SOX family of transcriptional regulators.* Current opinion in genetics & development, 2002. **12**(4): p. 441-6.
116. Avilion, A.A., et al., *Multipotent cell lineages in early mouse development depend on SOX2 function.* Genes & development, 2003. **17**(1): p. 126-40.
117. Graham, V., et al., *SOX2 functions to maintain neural progenitor identity.* Neuron, 2003. **39**(5): p. 749-65.
118. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.* Cell, 2006. **126**(4): p. 663-76.
119. Chen, Y., et al., *The molecular mechanism governing the oncogenic potential of SOX2 in breast cancer.* The Journal of biological chemistry, 2008. **283**(26): p. 17969-78.
120. Ng, P.Y., et al., *Synthesis of diverse lactam carboxamides leading to the discovery of a new transcription-factor inhibitor.* Angewandte Chemie, 2007. **46**(28): p. 5352-5.
121. Koehler, A.N., A.F. Shamji, and S.L. Schreiber, *Discovery of an inhibitor of a transcription factor using small molecule microarrays and diversity-oriented synthesis.* Journal of the American Chemical Society, 2003. **125**(28): p. 8420-1.
122. Zhou, H.X. and A. Szabo, *Enhancement of association rates by nonspecific binding to DNA and cell membranes.* Physical review letters, 2004. **93**(17): p. 178101.
123. Halford, S.E. and M.D. Szczelkun, *How to get from A to B: strategies for analysing protein motion on DNA.* European biophysics journal : EBJ, 2002. **31**(4): p. 257-67.
124. Berg, O.G. and P.H. von Hippel, *Diffusion-controlled macromolecular interactions.* Annual review of biophysics and biophysical chemistry, 1985. **14**: p. 131-60.
125. Doucleff, M. and G.M. Clore, *Global jumping and domain-specific intersegment transfer between DNA cognate sites of the multidomain transcription factor Oct-1.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(37): p. 13871-6.

126. Cary, P.D., et al., *Solution structure and backbone dynamics of the DNA-binding domain of mouse Sox-5.* Protein science : a publication of the Protein Society, 2001. **10**(1): p. 83-98.
127. Hever, A.M., K.A. Williamson, and V. van Heyningen, *Developmental malformations of the eye: the role of PAX6, SOX2 and OTX2.* Clinical genetics, 2006. **69**(6): p. 459-70.
128. Glaser, T., D.S. Walton, and R.L. Maas, *Genomic structure, evolutionary conservation and aniridia mutations in the human PAX6 gene.* Nature genetics, 1992. **2**(3): p. 232-9.
129. Kamachi, Y., et al., *Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development.* Genes & development, 2001. **15**(10): p. 1272-86.
130. Castillo, S.D. and M. Sanchez-Cespedes, *The SOX family of genes in cancer development: biological relevance and opportunities for therapy.* Expert opinion on therapeutic targets, 2012. **16**(9): p. 903-19.
131. Dong, C., D. Wilhelm, and P. Koopman, *Sox genes and cancer.* Cytogenetic and genome research, 2004. **105**(2-4): p. 442-7.
132. Clarkson, J. and I.D. Campbell, *Studies of protein-ligand interactions by NMR.* Biochemical Society transactions, 2003. **31**(Pt 5): p. 1006-9.
133. Goldflam, M., et al., *NMR studies of protein-ligand interactions.* Methods in molecular biology, 2012. **831**: p. 233-59.
134. Roberts, G.C.K., *NMR of macromolecules : a practical approach*. The Practical approach series1993, Oxford ; New York: IRL Press at Oxford University Press. xviii, 399 p.
135. Wishart, D.S., et al., *1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects.* Journal of biomolecular NMR, 1995. **5**(1): p. 67-81.
136. Berjanskii, M.V. and D.S. Wishart, *Application of the random coil index to studying protein flexibility.* Journal of biomolecular NMR, 2008. **40**(1): p. 31-48.
137. Prudent, R., et al., *Identification of polyoxometalates as nanomolar noncompetitive inhibitors of protein kinase CK2.* Chemistry & Biology, 2008. **15**(7): p. 683-692.

138. Narasimhan, K., *Discovery, structure-function characterization and assessment of polyoxometalates as modulators of the DNA binding activity of the Sox-HMG family of transcription factors*, in *DEPARTMENT OF BIOLOGICAL SCIENCES*2012, National University of Singapore: Singapore. p. 140.
139. Narasimhan, K., et al., *Identification of a polyoxometalate inhibitor of the DNA binding activity of Sox2.* Acs Chemical Biology, 2011. **6**(6): p. 573-81.
140. Gao, G., J.G. Williams, and S.L. Campbell, *Protein-protein interaction analysis by nuclear magnetic resonance spectroscopy.* Methods in molecular biology, 2004. **261**: p. 79-92.
141. Akabayov, S.R., et al., *NMR mapping of the IFNAR1-EC binding site on IFNalpha2 reveals allosteric changes in the IFNAR2-EC binding site.* Biochemistry, 2010. **49**(4): p. 687-95.
142. Bikadi, Z. and E. Hazai, *Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock.* Journal of Cheminformatics, 2009. **1**.
143. vanHoute, L.P.A., et al., *Solution structure of the sequence-specific HMG box of the lymphocyte transcriptional activator Sox-4.* Journal of Biological Chemistry, 1995. **270**(51): p. 30516-30524.
144. Cary, P.D., et al., *Solution structure and backbone dynamics of the DNA-binding domain of mouse Sox-5.* Protein Science, 2001. **10**(1): p. 83-98.
145. Palasingam, P., et al., *The Structure of Sox17 Bound to DNA Reveals a Conserved Bending Topology but Selective Protein Interaction Platforms.* Journal of Molecular Biology, 2009. **388**(3): p. 619-630.
146. Murphy, E.C., et al., *Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation.* Journal of Molecular Biology, 2001. **312**(3): p. 481-99.
147. Wüthrich, K., *NMR of proteins and nucleic acids*. The George Fisher Baker non-resident lectureship in chemistry at Cornell University1986, New York: Wiley. xv, 292.

148. Remenyi, A., et al., *Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers.* Genes & Development, 2003. **17**(16): p. 2048-2059.
149. Ng, C.K., et al., *Deciphering the Sox-Oct partner code by quantitative cooperativity measurements.* Nucleic Acids Res. **40**(11): p. 4933-41.
150. Kamachi, Y., et al., *Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development.* Genes Dev, 2001. **15**(10): p. 1272-86.
151. Luger, K., et al., *Crystal structure of the nucleosome core particle at 2.8 A resolution.* Nature, 1997. **389**(6648): p. 251-60.
152. Dorigo, B., et al., *Chromatin fiber folding: requirement for the histone H4 N-terminal tail.* Journal of Molecular Biology, 2003. **327**(1): p. 85-96.
153. Olins, D.E. and A.L. Olins, *Chromatin history: our view from the bridge.* Nature reviews. Molecular cell biology, 2003. **4**(10): p. 809-14.
154. Beard, D.A. and T. Schlick, *Computational modeling predicts the structure and dynamics of chromatin fiber.* Structure, 2001. **9**(2): p. 105-14.
155. Robinson, P.J., et al., *EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(17): p. 6506-11.
156. Poirier, M.G., et al., *Dynamics and function of compact nucleosome arrays.* Nat Struct Mol Biol, 2009. **16**(9): p. 938-44.
157. Richmond, T.J. and C.A. Davey, *The structure of DNA in the nucleosome core.* Nature, 2003. **423**(6936): p. 145-50.
158. Zheng, C. and J.J. Hayes, *Structures and interactions of the core histone tail domains.* Biopolymers, 2003. **68**(4): p. 539-46.
159. Hansen, J.C., *Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions.* Annu Rev Biophys Biomol Struct, 2002. **31**: p. 361-92.
160. Peterson, C.L. and M.A. Laniel, *Histones and histone modifications.* Current Biology, 2004. **14**(14): p. R546-R551.
161. Grant, P.A., *A tale of histone modifications.* Genome Biology, 2001. **2**(4).

162. Rice, J.C. and C.D. Allis, *Histone methylation versus histone acetylation: new insights into epigenetic regulation.* Current Opinion in Cell Biology, 2001. **13**(3): p. 263-73.
163. Wu, J. and M. Grunstein, *25 years after the nucleosome model: chromatin modifications.* Trends in biochemical sciences, 2000. **25**(12): p. 619-23.
164. Ling, X., et al., *Yeast histone H3 and H4 amino termini are important for nucleosome assembly in vivo and in vitro: redundant and position-independent functions in assembly but not in gene regulation.* Genes & development, 1996. **10**(6): p. 686-99.
165. Krajewski, W.A. and J. Ausio, *Modulation of the higher-order folding of chromatin by deletion of histone H3 and H4 terminal domains.* The Biochemical journal, 1996. **316 (Pt 2)**: p. 395-400.
166. Davey, C.A., et al., *Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution.* Journal of Molecular Biology, 2002. **319**(5): p. 1097-113.
167. Davey, C.A. and T.J. Richmond, *DNA-dependent divalent cation binding in the nucleosome core particle.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(17): p. 11169-74.
168. Wolffe, A.P., *Nucleosomes: Detailed Structure and Mutations.* ENCYCLOPEDIA OF LIFE SCIENCES, 2001.
169. Baneres, J.L., A. Martin, and J. Parello, *The N tails of histones H3 and H4 adopt a highly structured conformation in the nucleosome.* Journal of Molecular Biology, 1997. **273**(3): p. 503-8.
170. Angelov, D., et al., *Preferential interaction of the core histone tail domains with linker DNA.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(12): p. 6599-6604.
171. Hansen, J.C., C. Tse, and A.P. Wolffe, *Structure and function of the core histone N-termini: more than meets the eye.* Biochemistry, 1998. **37**(51): p. 17637-17641.
172. Zheng, C.Y. and J.J. Hayes, *Structures and interactions of the core histone tail domains.* Biopolymers, 2003. **68**(4): p. 539-546.

173. Korolev, N., et al., *Electrostatic origin of salt-induced nucleosome array compaction.* Biophysical journal, 2010. **99**(6): p. 1896-905.
174. Lowary, P.T. and J. Widom, *New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning.* Journal of Molecular Biology, 1998. **276**(1): p. 19-42.
175. Allahverdi, A., et al., *The effects of histone H4 tail acetylations on cation-induced chromatin folding and self-association.* Nucleic acids research, 2011. **39**(5): p. 1680-91.
176. Wolffe, A.P. and J.J. Hayes, *Chromatin disruption and modification.* Nucleic acids research, 1999. **27**(3): p. 711-20.
177. Peterson, C.L. and M.A. Laniel, *Histones and histone modifications.* Current biology : CB, 2004. **14**(14): p. R546-51.
178. Shogren-Knaak, M., et al., *Histone H4-K16 acetylation controls chromatin structure and protein interactions.* Science, 2006. **311**(5762): p. 844-7.
179. Clark, D.J. and T. Kimura, *Electrostatic mechanism of chromatin folding.* Journal of Molecular Biology, 1990. **211**(4): p. 883-96.
180. Chodaparambil, J.V., et al., *A charged and contoured surface on the nucleosome regulates chromatin compaction.* Nature structural & molecular biology, 2007. **14**(11): p. 1105-7.
181. Zhou, J., et al., *The nucleosome surface regulates chromatin compaction and couples it with transcriptional repression.* Nature structural & molecular biology, 2007. **14**(11): p. 1070-6.
182. Wishart, D.S. and B.D. Sykes, *The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data.* Journal of biomolecular NMR, 1994. **4**(2): p. 171-80.
183. Grzesiek, S. and A. Bax, *Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins.* Journal of biomolecular NMR, 1993. **3**(2): p. 185-204.
184. Eliezer, D., *Characterizing residual structure in disordered protein States using nuclear magnetic resonance.* Methods in molecular biology, 2007. **350**: p. 49-67.

185. Masse, J.E. and R. Keller, *AutoLink: Automated sequential resonance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic.* Journal of magnetic resonance, 2005. **174**(1): p. 133-151.
186. Jung, Y.S. and M. Zweckstetter, *Mars -- robust automatic backbone assignment of proteins.* Journal of biomolecular NMR, 2004. **30**(1): p. 11-23.
187. Muñoz, V., *Protein folding, misfolding and aggregation : classical themes and novel approaches*. RSC biomolecular sciences2008, Cambridge, UK: Royal Society of Chemistry. xv, 272 p.
188. Zhang, Y.Z., Y. Paterson, and H. Roder, *Rapid amide proton exchange rates in peptides and proteins measured by solvent quenching and two-dimensional NMR.* Protein science : a publication of the Protein Society, 1995. **4**(4): p. 804-14.
189. Schalch, T., et al., *X-ray structure of a tetranucleosome and its implications for the chromatin fibre.* Nature, 2005. **436**(7047): p. 138-41.
190. Kato, H., et al., *Characterization of the N-Terminal Tail Domain of Histone H3 in Condensed Nucleosome Arrays by Hydrogen Exchange and NMR.* Journal of the American Chemical Society, 2009. **131**(42): p. 15104-+.
191. Yao, J., H.J. Dyson, and P.E. Wright, *Chemical shift dispersion and secondary structure prediction in unfolded and partly folded proteins.* FEBS letters, 1997. **419**(2-3): p. 285-9.
192. Zhang, O., et al., *Triple-resonance NOESY-based experiments with improved spectral resolution: applications to structural characterization of unfolded, partially folded and folded proteins.* Journal of biomolecular NMR, 1997. **9**(2): p. 181-200.
193. Schwarz, P.M. and J.C. Hansen, *Formation and stability of higher order chromatin structures. Contributions of the histone octamer.* J Biol Chem, 1994. **269**(23): p. 16284-9.
194. Korolev, N., et al., *A universal description for the experimental behavior of salt-(in)dependent oligocation-induced DNA condensation.* Nucleic acids research, 2009. **37**(21): p. 7137-50.
195. Allahverdi, A., et al., *The effects of histone H4 tail acetylations on cation-induced chromatin folding and self-association.* Nucleic Acids Res., 2011. **39**: p. 1680–1691

Publications

1. Wong, L.E., Li, Y., **Pillay, S**., Frolova, L., Pervushin, K., *Selectivity of stop codon recognition in translation termination is modulated by multiple conformations of GTS loop in eRF1.* Nucleic Acids Research, 2012. **40**(12): p. 5751-65.

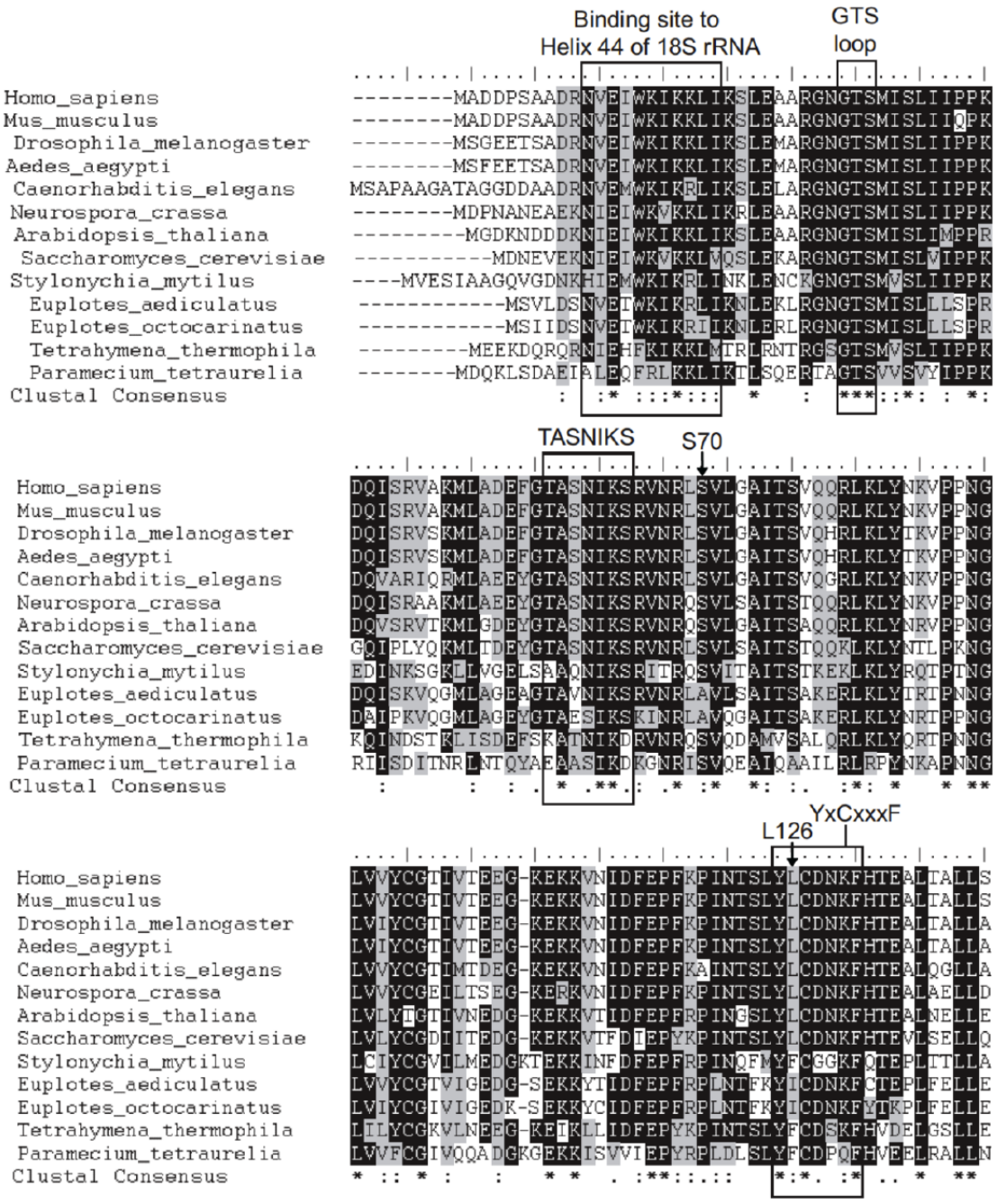
2. Narasimhan, K., **Pillay, S**., Bin Ahmad, N. R., Bikadi, Z., Hazai, E.,Yan, L., Kolatkar, P. R., Pervushin, K., Jauch, R., *Identification of a polyoxometalate inhibitor of the DNA binding activity of Sox2.* Acs Chemical Biology, 2011. **6**(6): p. 573-81.

3. **Pillay, S.,** Li, Y., Wong, L.E., Pervushin, K., *Stop codon recognition mediated by conformational changes of eRF1: structural insights of eRF1 mutants.* Manuscript in preparation.

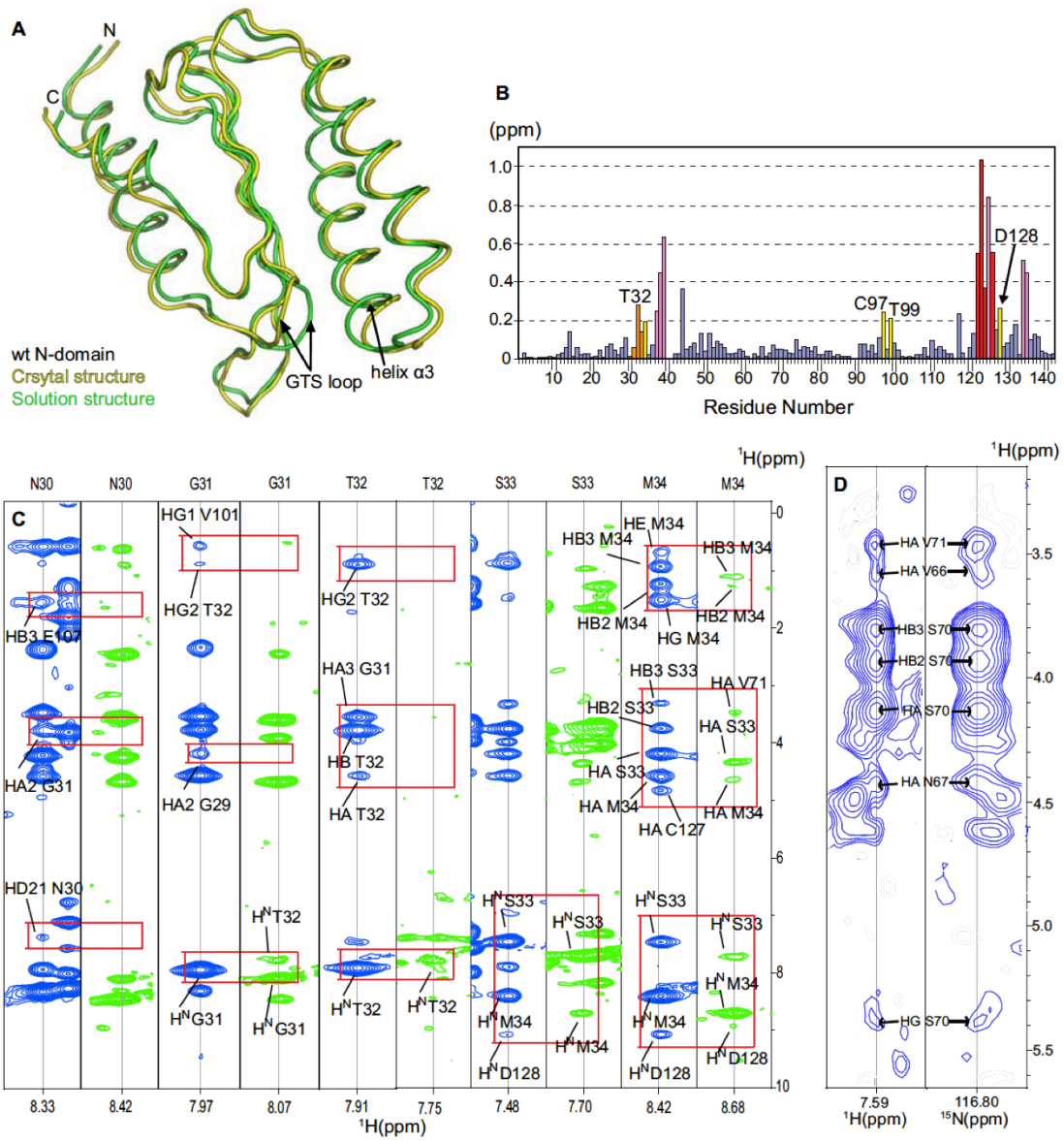
Poster

25th ICMRBS (International Conference on Magnetic Resonance in Biological Systems), Lyon, France from 19th- 24th August 2012.
Structural insights of eRF mutants and its correlation to stop codon recognition.

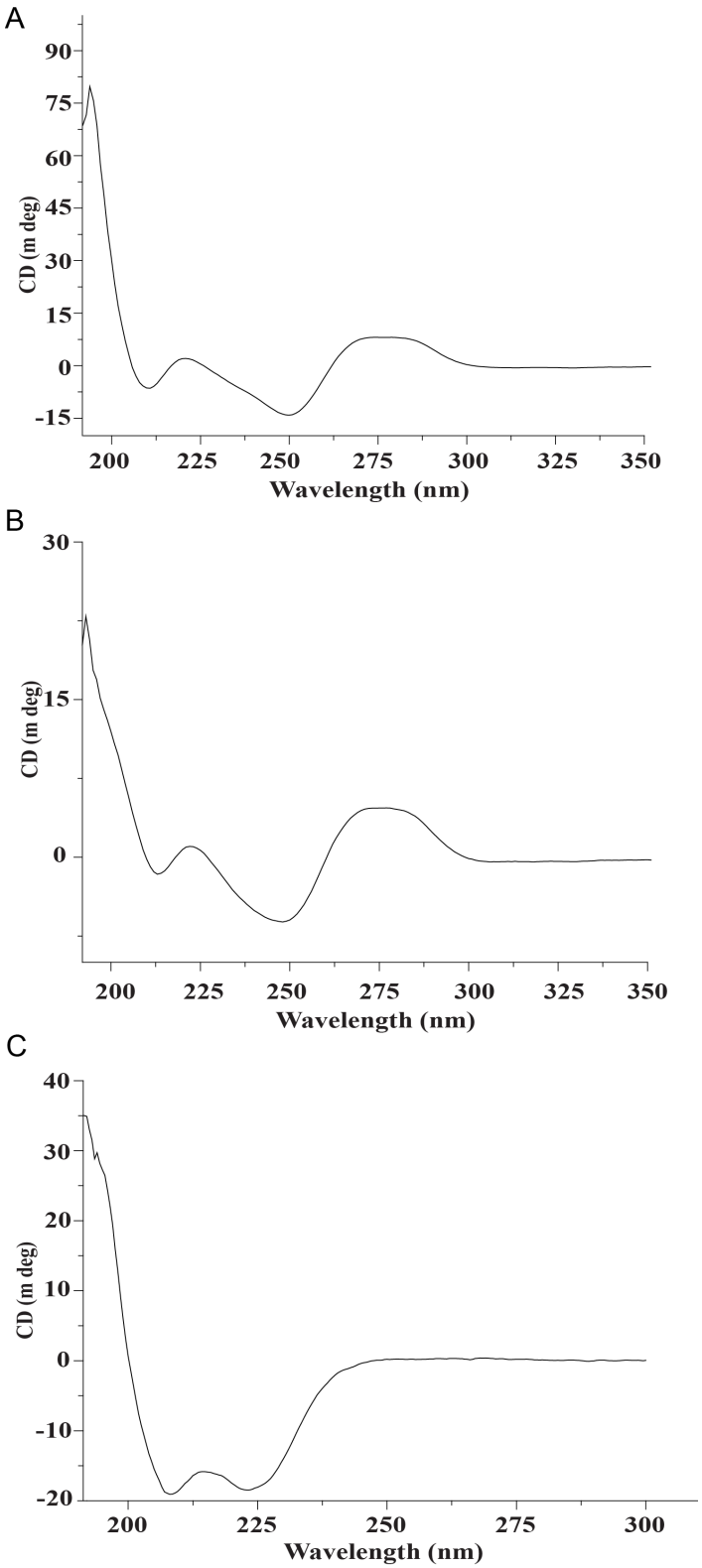
Appendix



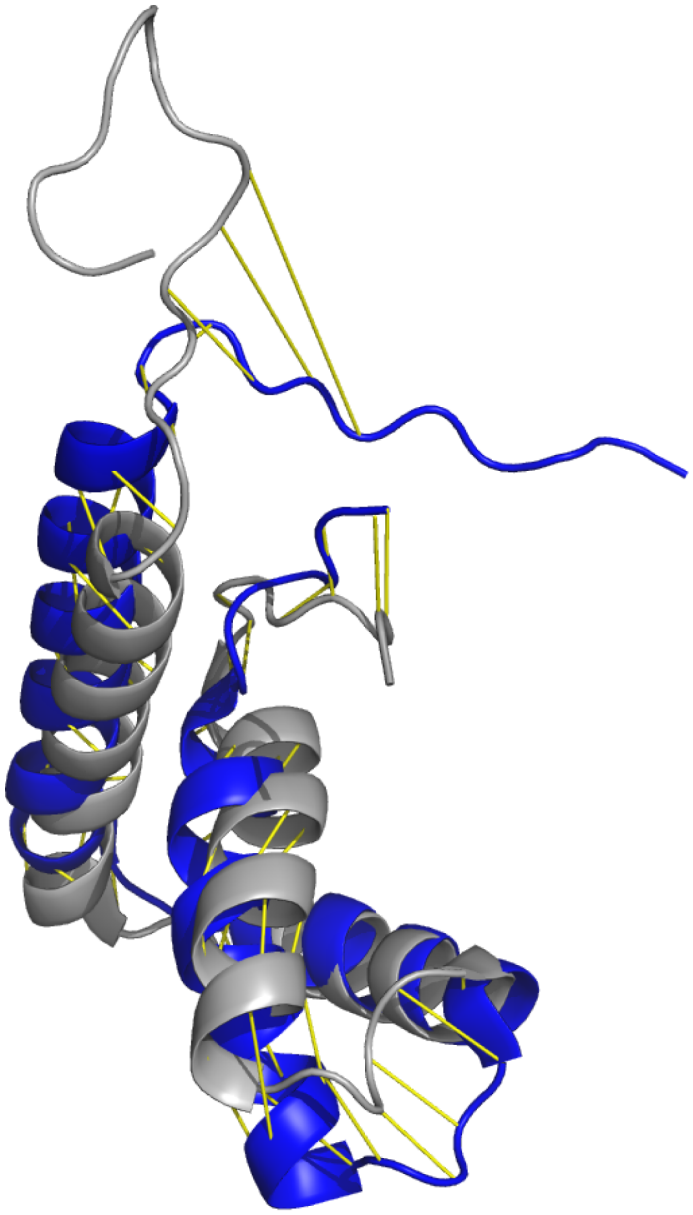
Appendix A1. Sequence conservation of the GTS loop. Multiple sequence alignment of the amino acid sequences of eRF1 N-domain from different eukaryotic organisms using ClustalW2 online server (http://www.ebi.ac.uk/Tools/msa/clustalw2/). Numbering of the residues is according to the human's. The GTS loop is strictly conserved. Substitution S70A occurs in *E. aediculatus* and *E. octocarinatus*, which are variant-code organisms with UAR as stop codons. Conversely, substitution L126F occurs in *S. mytilus*, *T. thermophila*, and *P. tetraurelia*, all of which are unipotent for UGA. Figure adapted from [88].



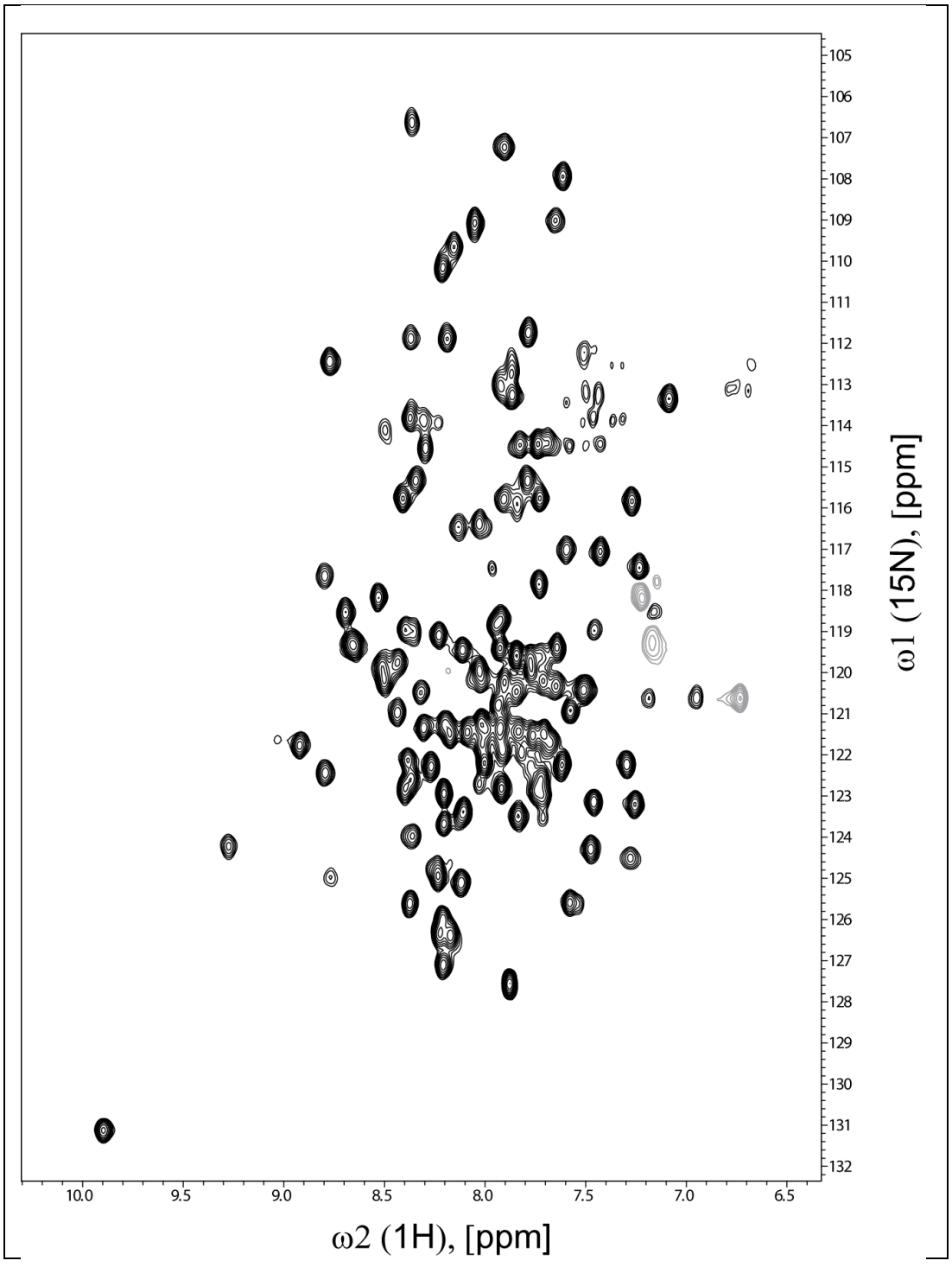
Appendix A2. The GTS loop in Q122FM(Y)F126 has a conformation distinct from wt N-domain. (A**)** Superposition of the crystal structure (yellow) and solution structure (green) of wt Ndomain, showing distinct conformations of the GTS loop and spatial deviation of helix α3. **(B)**Plot of the differences in amide chemical shift between wt N-domain and Q122FM(Y)F126 as calculated by $\Delta\delta NH = [(\Delta\delta H)2 + (0.14*\Delta\delta N)2]1/2$. Specific residues are highlighted as follows: point mutations (red), residues close to the point mutations in space (pink), the GTS loop (orange), and residues close to the GTS loop in space (yellow) (Figure 1B). **(C)** Differences in the NOE patterns as observed from the amides of N30 to M34 between wt N-domain (blue) and Q122FM(Y)F126 (green) show that their GTS loops have distinct conformations. **(D)** The NOE cross-peak of HG S70 as observed from the amide of S70 indicates participation of the hydroxyl group in hydrogen bonding. This NOE cross-peak was not observed in wt Ndomain. Figure adapted from [88].



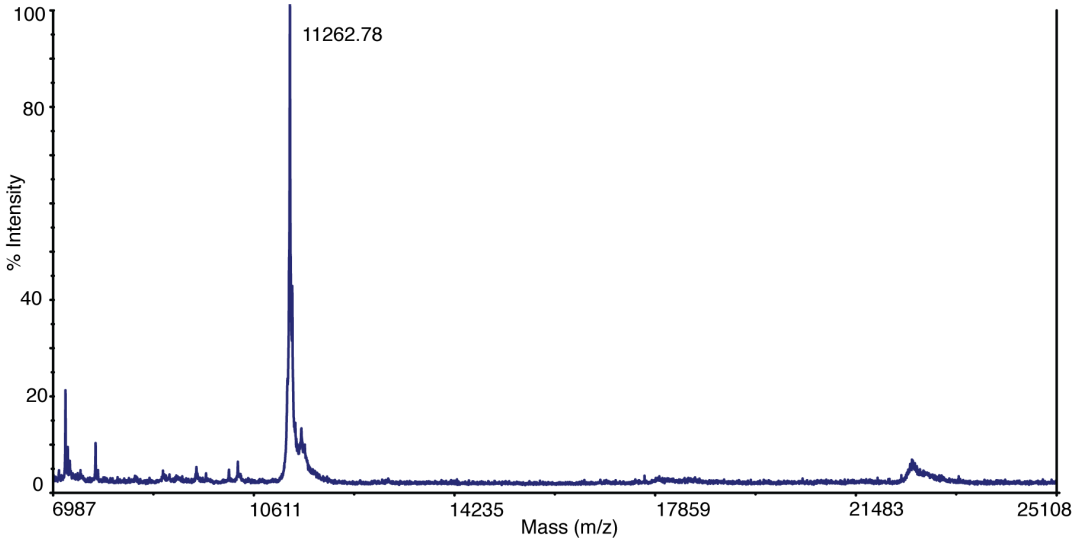
Appendix B1. CD spectra of A) CCND1, B) DC5 and C) Sox2-HMG



Appendix B2. Aligned solution structures of apo Sox2 (grey) (PDB 2LE4) and Sox2 ternary complex (blue) (PDB 1GT0). The RMS between both structures were calculated to be 5.2 using MacPymol.



APPENDIX B3.. 2D [^{15}N,^{1}H]-TROSY of Pax6 in 50 mM Phosphate buffer, pH 7.0, 100 mM KCl, 90% H_2O/10% D_2O at 298K, using 600 MHz NMR spectrometer.



Appendix C1. Mass spectrum of H4 showing that the calculated molecular weight of unlabeled sample was 11263 and observed is as shown in the figure above (11262.78).