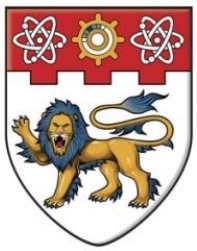


NANYANG TECHNOLOGICAL UNIVERSITY

CHARACTERIZATION OF DNA BINDING AND OLIGOMERIZATION DOMAINS OF STEM CELL TRANSCRIPTION FACTOR SOX9 AND SOX5

SARAVANAN VIVEKANANDAN

SCHOOL OF BIOLOGICAL SCIENCES

NANYANG TECHNOLOGICAL UNIVERSITY

SINGAPORE

2013

Characterization of DNA Binding and Oligomerization Domains of Stem Cell Transcription Factor Sox9 and Sox5

SARAVANAN VIVEKANANDAN

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

2013

"There Is No Greater Object Of Wonder, No Greater Thing Of Beauty, Than The Dynamic Order, The Organized Complexity Of Life."

Ariel G. Loewy & Philip Siekevitz, 1969.

"He (Biochemist) Should Be Bold In Experiment But Cautious In His Claims. His May Not Be The Last word In The Description Of Life, But Without His Help The Last Word Will Never Be Said."

Prof. Hopkins (1931)

"When You Really Want To Do Something, All The Universe Conspires In Helping You To Achieve It"

Paulo Coelho, The Alchemist.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to prof. Julien Lescar for providing me this exciting opportunity and I am grateful to him for his encouragement and timely suggestions which helped me a lot throughout this learning phase. I take this opportunity to express my deep sense of gratitude to him for his concern and guidance.

It was wonderful to be associated with Prof. Prasanna R. Kolatkar. The freedom of thought and action extended by him to pursue my ideas throughout the course of my research was immense, I am grateful to him for his keen interest in my work, his constructive criticisms during the scientific discussions, constant support, for inculcating "good" lab practices and the awe for science.

I am greatful to Prof H.S.Savithri, Prof. N.Appaji Rao and Prof. M.R.N. Murthy, (Indian Institute of Science, Bangalore) for being unsurpassable inspirations, from my IISc Bangalore days. The smile and warmth of Prof. H.S.Savithri, the lively scientific discussions with Prof. N.Appaji Rao, the care and support of Prof. M.R.N. Murthy have been treasured memories fuelling this entire long journey of Ph.D.

My sincere thanks to Dr. Paaventhan, Dr. Pugal, Dr. Bala, and Dr. Kamesh for their constant interest in my progress. Their suggestions in the lab meetings helped me to refine my approach towards any scientific problems. Dr. Ralf deserves special thanks for helping us to adopt various new advancements in our research work.

I would also like to thank Dinesh and Justin for their warmly friendship and wonderful time I have spent with them.

My sincere thanks to Calista, Dr. Nithya and Siew Hua, Dr. Marie for all fun moments in the lab and their tips in various techniques. I owe special thanks to Dr. Kareem, for always being there to clarify my doubts. My sincere thanks to my FYP students Ms.Sathya, Ms.Yulu and Ms. Tung Yu Ting (Mei) for all their help during their stay.

*I would like to acknowledge NTU for my graduate research scholarship. This work is supported by the Agency for Science, Technology and Research (A*STAR) and the Genome Institute of Singapore.*

I am indebted to my family for their selfless love and encouragement during the course of my study. I owe this accomplishment to my parents, brothers Pugal & Nambi and sisters Senthakka & Poovu who have been there patiently supporting me.

I am indebted to my wife Rathi for her marvelous patience, love and support during all my ups and downs. Without her support my dream of Ph.D would have been a difficult reality. I am thankful to my in-laws for their love and support. My special gratitude to Nikita, my lovable daughter, for her innocent understanding and adaptability with my work schedule and Srishti for all her innocent and warmth smile.

Finally, I dedicate this thesis to my dad, M.S.Vivekanandan who took all pain for my pleasure.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I INTRODUCTION

CHAPTER II MATERIALS AND METHODS

PROTEIN – DNA INTERACTION OF SOX TRANSCRIPTION FACTORS

CHAPTER III DNA BINDING HMG DOMAIN OF SOX9

CHAPTER IV IDENTIFICATION AND VALIDATION OF NOVEL SOX9 REGULATORY MOTIF

CHAPTER V THE DNA BINDING HMG DOMAIN OF SOX5

CHAPTER VI CRYSTALLIZATION OF PROTEIN-DNA COMPLEX

PROTEIN – PROTEIN INTERACTION OF SOX TRANSCRIPTION FACTORS

CHAPTER VII DNA DEPENDENT PROTEIN-PROTEIN INTERACTION SOX9 DIMERIZATION DOMAIN

CHAPTER VIII COILED COIL MEDIATED OLIGOMERISATION OF SOX5

LIST OF TABLES

Abbreviations

bHLH	**B**asic-**H**elix-**l**oop-**h**elix
bZIP	**B**asic-Leucine **zip**per factors
CD	**C**ampomelic **D**ysplasia/ **C**ircular **D**ichroism
ChIP-Seq	Chromatin immunoprecipation sequencing
DMSO	Dimethyl sulfoxide
Col2a1	**Co**llagen, type **II**, **a**lpha **1**
Col4a1	**Co**llagen, type **IV**, **a**lpha **1**
DNA	Deoxyribo nucleic acid
EMSA	**E**lectrophoretic **M**obility **S**hift **A**ssay
FAM	Fluorescein amidite
FGF	Fibroblast growth factor
FP	Fluorescence polarization
FRET	Fluorescence resonance energy transfer
HMG	**H**igh **M**obility **G**roup
LEF	**L**ymphoid **E**nhancer-binding **F**actor 1
NLS	**N**uclear **L**ocalization **S**ignal
Oct	**O**ctamer TF, Oct-1 and Oct-2 **U**nc-86 TF
PIC	**P**re**i**nitiation **C**omplex
PQA	**P**roline-**G**lutamine-**A**lanine rich motif
POU	**P**ituitary-specific Pit-1
Q-Box	**Q**lutamin Rich **Box**
SRY	**S**ex-determining **R**egion **Y**
TA Domain	**T**rans **A**ctivation domain

TCF	**T C**ell-specific transcription **F**actors
TEV	**T**obacco mosaic **E**tched **V**irus
TF	**T**ranscription **F**actors

ABSTRACT

Sox (SRY-related HMG box) family of proteins is mammalian stem cell transcription factors, playing a pivotal role in regulation of numerous developmental processes. So far nearly 30 Sox proteins have been identified and categorized into eight subgroups of A to G, based on high sequence similarity of the conserved HMG box domain. Unlike other transcription factors, Sox proteins are unique as they recognize and bind the minor groove of DNA, inducing a strong DNA bend of 70-85°, an essential step in eukaryotic transcriptional regulation. Although the HMG domains of all Sox proteins are highly conserved and bind similar DNA elements of sequence (A/T) (A/T) CAA (A/T), they regulate a wide assortment of genes in diverse developmental processes. The functional specificity of Sox transcription factors depends on (i) subtle nucleotide variations in the DNA sequence; (ii) differential minor groove bend as a consequence of HMG domain mediated DNA interaction (iii) different co-factor recruitment through protein-protein interactions.

In this regard, Sox9 (group SoxE), Sox5 and Sox6 (group SoxD), famously known as "Sox trio", are ideal prototypes with conserved HMG domains and group specific domains for partner recruitment. Sox9 has a DNA dependent dimerization domain and Sox5 has "DNA independent" coiled-coil (CC) domain. The objective of the current study is to comprehend the underlying molecular mechanism of DNA recognition and transcriptional specificity of the Sox trio.

Towards this end, the DNA binding HMG domain of Sox9 and Sox5; protein interacting domains of Sox9 (DNA dependent dimerization domain) and Sox5 (coiled-coil domain) were studied. In the case of protein-DNA interactions, the HMG domains of Sox9 and Sox5 were cloned, purified and their DNA binding abilities were determined by EMSA. Moreover, two novel Sox9 DNA binding motifs have been identified employing ultra-high-throughput DNA sequencing (ChIP-Seq) data. Functional validation of the novel motifs by

EMSA and luciferase assay confirms Sox5 as downstream target of Sox9. Crystalisation of DNA bound Sox9/Sox5 HMG domains have yielded promising outcomes.

Protein-protein mediated oligomerisation was analysed employing the DNA dependent dimerization domain of Sox9 and the “DNA independent” coiled- coil domain of Sox5, as model systems. The study involves the cloning, purification, crystallization, biochemical and biophysical characterization of the corresponding domains. The results presented indicate Sox9 to exclusively exist as dimers in the presence of DNA and contrastingly, Sox5 harboring intact CC domain tetramerizes, indicating a possibility of higher order homo/hetero oligomerisation. The current study is the first comprehensive biochemical validation of Sox protein-protein interacting domains, highlighting the role of Sox homo/hetero oligomers in dictating the transcriptional specificity of Sox transcription factors.

INTRODUCTION

The central dogma of life is that the genetic information of an organism, inherited as genotype (DNA) manifests as corresponding phenotype (proteins) by way of RNA. Thus, transcription of an RNA product from DNA is a pivotal regulating point of gene expression in specific cell types, in response to particular signals. Transcription factors (TF) are key mediators of this complex, yet strictly regulated process. For example (*Fig.1.1*), the same transcription factor X turns "ON" genes A, and C, while it turns "OFF" gene B. Likewise, in a cell at a given time, many transcription factors may recognize and bind the same gene promoter, but may have conflicting or augmenting effects on gene activity [1].

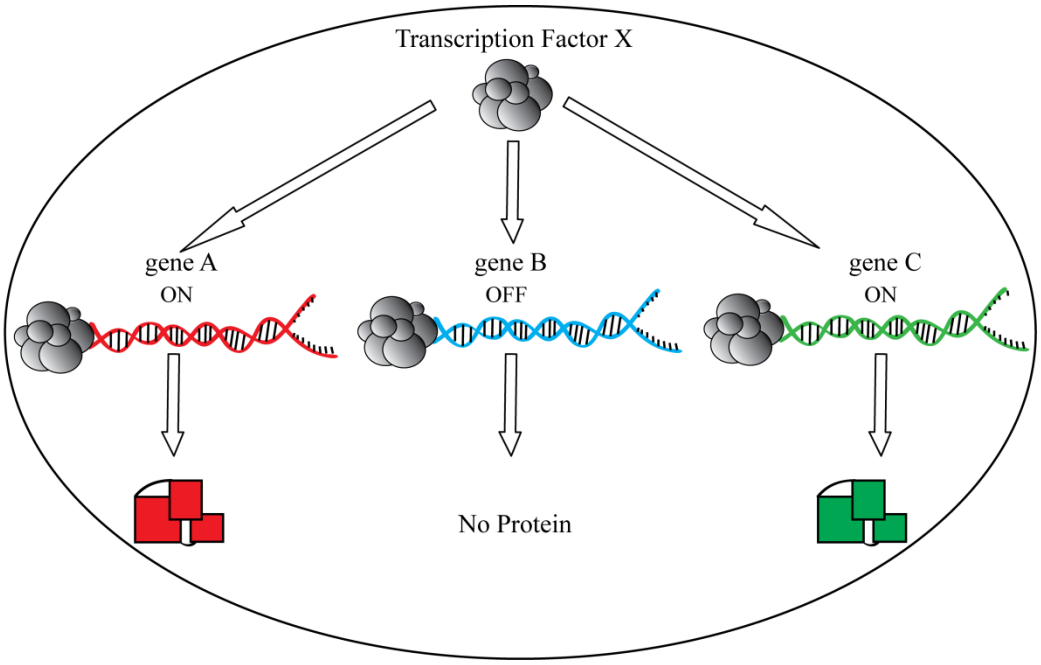


Figure 1.1 Diagrammatic representation of transcriptional gene regulation. Transcription factor X has been denoted as grey circles and proteins as colored squares.

Transcription requires binding of RNA polymerase II at desired specific start sites upstream of the coding sequence. Binding to the promoter region is facilitated by protein factors which either recognize specific DNA sequences within the core promoter region or bind to distal enhancer elements or bridge between distal and proximal promoter elements [1, 2]. While core promoter elements are bound by factors able to initiate transcription at

basal levels [3] gene-specific enhancement or repression is commonly facilitated by factors binding to distal elements. Specificity and regulation of gene transcription is achieved through collaboration of activators and co-activators of the transcription machinery. In the case of a single cell eukaryote *Saccharomyces cerevisiae*, a subset of over hundred transcription activators and repressors interacts with several dozens of transcription factors to regulate the transcription of approximately six thousand genes [3]. The combinatorial possibilities suggest huge diversity in promoter selection and transcription regulation.

1.1 Prokaryotic transcription

In prokaryotes, there is a single RNA polymerase holoenzyme that binds to specific promoter regions, typically found in the -35 and -10 region of the transcription start site, resulting in the initiation of transcription [4]. This consensus promoter sequence is usually of the sequence "TATAAT" referred as the "Pribnow box" structure [4, 5]. Some of the most well studied transcription factors in prokaryotes are those that regulate gene expression in response to a variety of environmental cues like substrate availability, presence of antibiotics, quorum sensing of neighboring bacterial colonies in response to bacteriophage infections. Prokaryotic genes usually occur in clusters that get transcribed in the control of a promoter, referred to as an "Operon" [6]. Genes are transcribed from operons into polycistronic mRNA that carries several open reading frames each of which can be translated into a polypeptide [5]. Classical examples of *Escherichia coli* (*E.coli)* operon transcriptional regulation (lac, trp, arb repressors etc.) usually involves a decision making switch responsible for repression of utilization of various metabolites like lactose, arabinose and galactose in the presence of primary metabolite glucose [5]. In general, the Helix-Turn-Helix (HTH) domain is the most common DNA recognition mode by multi-domain transcription factors in prokaryotes [7]. HTH domains are structures of ~20-30 amino acids that form two α-helices that cross each other at ~120^{o} and base-pair with the DNA primarily via the side-chains extending from the

second-helix of the HTH domain, referred to as the recognition helix (helix E in trp repressor) [7]. Studies on transcriptional mechanisms in *E.coli* have laid the foundation for studying complex gene expression mechanisms in eukaryotes [8].

1.2 Eukaryotic transcription

Transcriptional regulation in eukaryotes is a far more complex event that involves the assembly of multi-protein complexes on core/proximal promoter or upstream enhancer modules and requires the co-operative assembly of co-activators, mediators, general and sequence specific transcription factor complexes [5]. The eukaryotic core promoter is typically a TATA-box of the consensus sequence TATAAA, found in the -25 region. RNA polymerase II and general transcription factors like TFIIA, TFIIB, TFIID, TFIIE and TFIIH constitute the basal transcriptional machinery responsible for transcription from the core promoter region [5]. However, the level of transcription by RNAP II and the general TFs alone is usually low. Enhancers, part of the eukaryotic non-coding genome exert spatial and temporal control over gene expression programs in specific tissues though as far as few kilobases (kb) or Megabases (Mb) away from the transcription start site [9].

Sequence specific transcription factors bind to the proximal promoter or enhancer module and serve to enhance the rate of the transcription of genes under its control. Despite lack of direct structural evidence, it is widely believed that a complex DNA looping event presumably brings the specific transcription machinery lined up on the enhancer region to come into contact with RNA polymerase II and the general transcriptional factors, leading to a synergistic stabilization of a multi-protein pre-initiation complex in the promoter region [5]. Notably, the stem cell transcription factor Sox is unique in the sense that it binds to the minor groove of the DNA through HMG domain and causes a strong bend in the DNA. This DNA bending is an essential step in eukaryotic transcriptional regulation and it may alter the local chromatin structure aiding in assembling the large transcriptional complex on DNA, to

facilitate the interaction between transcription factors and also between regulators and promoters (*Fig. 1.2*) [10]. The model of a multi-protein complex assembly in eukaryotic transcription has received further support after the report of coactivator p300 crystal structure in complex with transcription factor Mef2, denoting p300 scaffold role in the assembly of transcription factors [11].

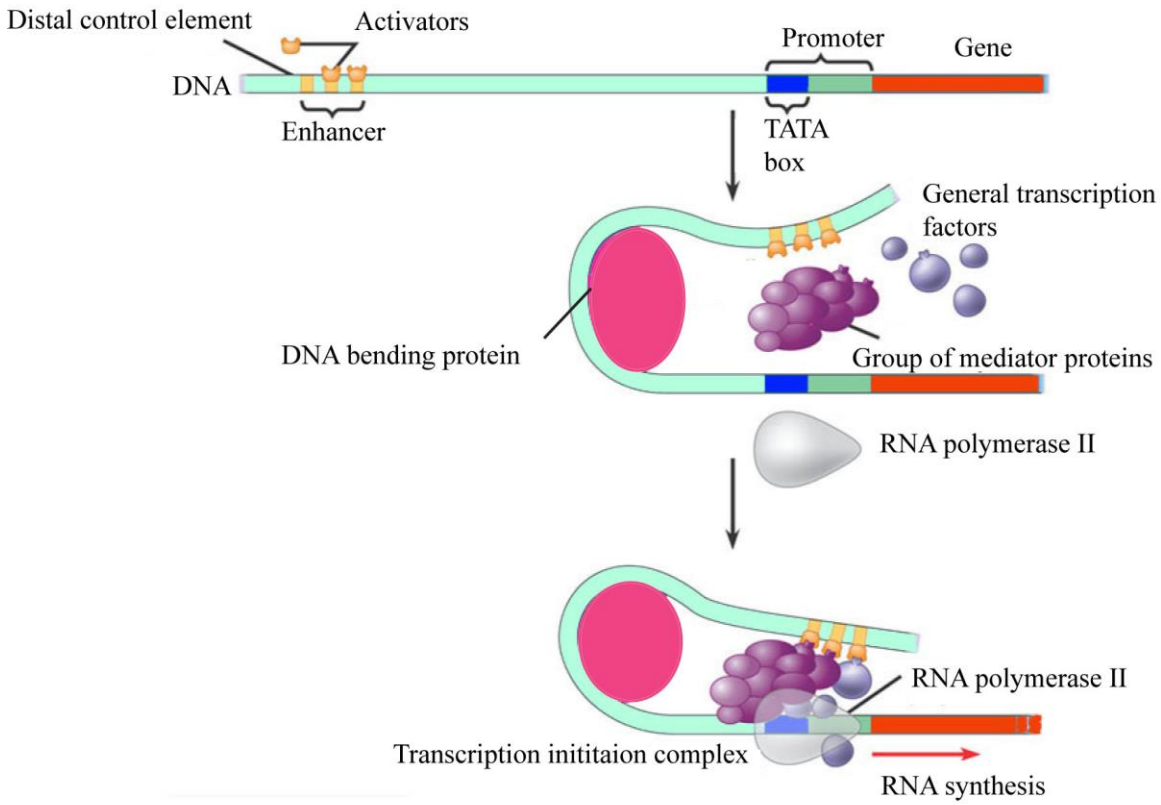


Figure.1.2 DNA bending is an essential step in the transcriptional regulation in eukaryotic gene. A DNA binding and bending proteins facilitate the activators - promoter interaction by reducing the proximity which is an important step for eukaryotic transcription.

1.3 DNA recognition motifs

There are about 2000 to 3000 sequence specific DNA binding transcription factors classified into five major superclass [12],[13] based on the structure of their DNA binding domains:

1. Basic domain
2. Zinc – coordinating

3. Helix-turn-Helix
4. β-scaffold and
5. Others

1.3.1 Basic Domain

The members of this superclass are characterized by large excess of positively charged DNA interacting domains, folded as alpha helical. They are found in close contact with either leucine zipper (ZIP), helix-loop-helix (HLH) or helix-span-helix (HSH) domain. Dimerization is a prerequisite for their DNA binding ability and determines the specificity of DNA binding. In humans, the *basic* superfamily consists of TFs that include notable families like the bZIP (53 members) and the bHLH (110 members) (20).

1.3.2. Zinc-Coordinating Domains

The alpha helix of the well-folded stable structure of Zinc finger proteins interacts with three base pairs of DNA and does not undergo conformational changes upon DNA binding. The domains are typically with two or more zinc fingers with ~25-30 amino acid and has 2 cysteine and 2 histidine residues. These Cys_2His_2 are important residues for the coordination with a zinc atom. The amino acid sequence and the linker between the zinc fingers are responsible for the determination of binding. Zinc finger domains are one of the targets for protein engineering. Eg., C2H2 zinc finger, nuclear receptors (C4 zinc fingers) and the GATA family of TFs (20).

1.3.3. Helix-turn-Helix

Helix-turn-helix (HTH) is a major structural motif capable of binding to DNA and regulating gene expression. Of its two alpha helices one binds with the N-terminal and the other with the C-terminal end of the motif and these two alpha helices are with a short amino acid linker region, one responsible for DNA recognition and other responsible for stabilization of interaction. Most HTH proteins are major groove binding transcription factors through series of hydrogen bonds and vander waals interaction. Besides, there are tri-helical, tetra-helical and winged helix-turn-helix forms of helix-turn-helix transcription factors. Example of members of this family includes, Hox, POU, Fox, IRF, Ets, RFX, HSF and E2F (20).

1.3.4. β-scaffold Factors

Transcription factors that bind DNA using a β-scaffold like structure are called β-scaffold transcription factors. For example the well known p53 tumor suppressoris a β-scaffold proteins and mutation or inactivation of p53 may lead to cancerous situation including apoptosis and therefore is rightly called as and also as "the guardian of the genome". Eg., p53, RHR, NF-κB and the STAT family (20).

1.3.5. Others

Sequence specific transcription factors that do not fall into the above categories are referred to as the unclassified. This family of proteins include Copper fist proteins, HMGI(Y) (HMGA1), Family: HMGI(Y), Pocket domain, E1A-like factors,AP2/EREBP-related factors, Apetala 2 (AP2), Ethylene-responsive element binding protein (EREBP), Auxin response factors (ARF), The Arabidopsis ABA-Insensitive (ABI), Related-to-ABI3/VP1 (RAV) transcription factor (20).

1.4 High Mobility Group (HMG) proteins

The lengthy eukaryotic DNA should be extremely condensed to accommodate into the cell nucleus. This involves coiling of the DNA around histone octamers to form nucleosomes. The nucleosomes are involved in the further compactness and forms higher order chromatin fibers including chromosomes. Inflection of chromatin folding has impacts on the accessability of regulatory factors to their DNA. The loosening or disturbance of the nucleosome structure through DNA bending and unwinding will help in this regard and also affects DNA-histone contacts through histones modification. Most of these changes in structural aspects are interceded by high mobility group (HMG) proteins [14] and are defined as:

1. The HMG-Nucleosome binding family (HMGN)
2. The HMG-AT-hook family (HMGA)
3. The HMG-Box family (HMGB)

1.5 Mammalian High Mobility Group Box family

HMG-box containing proteins are classified into 2 major groups: (i) consisting of HMGB-type with 2 HMG-box domains, non-sequence-specific DNA binding ability with a highly acidic C-terminal; (ii) highly diverse proteins having a single HMG-box and no acidic C-terminal and have sequence-specific DNA binding ability. Nonetheless, there are other architectural proteins which have up to 6 HMG-box domains *vide infra* [15, 16].

Binding on DNA minor groove and bending by HMG proteins leads to unwinding and widening of the minor groove. Generally, it has been postulated that the degree of DNA bending diverges among HMG-boxes and differs with variations in sequences at specific positions [17]. The HMG-box protein binds on the outside of the DNA bend, compressing the major groove [18]. Most of the HMG family members are key regulators of mammalian stem cell development, critical for cellular differentiation and Sox is one such HMG box containing imperative transcription factor [19, 20].

1.6 Sry-related HMG-box (Sox) Transcription Factors

Sox (Sry-related HMG-box) proteins or SOX genes are refered as HMG domain containing proteins or genes which have 50% or greater sequence similarity to the Sry (sex-determining region Y) [21]. They consist of greater than 25 members in mammals [29]. Sox proteins are classified into subfamilies A to J, on the grounds of amino acid sequence similarity of the HMG domains (*Fig.1.3*) [22]. Within each group of Sox subfamily, the identity amino acid sequence remains greater than ninety percentile to around sixty percentile identity between upstage groups and closely related Sox proteins even outside the HMG domain, share similar amino acid sequences.

A neighbor joining tree generated using the online multiple sequence alignment program MAFFT and visualized using splitstree shows different phylogenetic groupings (Groups A-H) of the Sox-HMG domains, reaffirming the HMG domain-based Sox

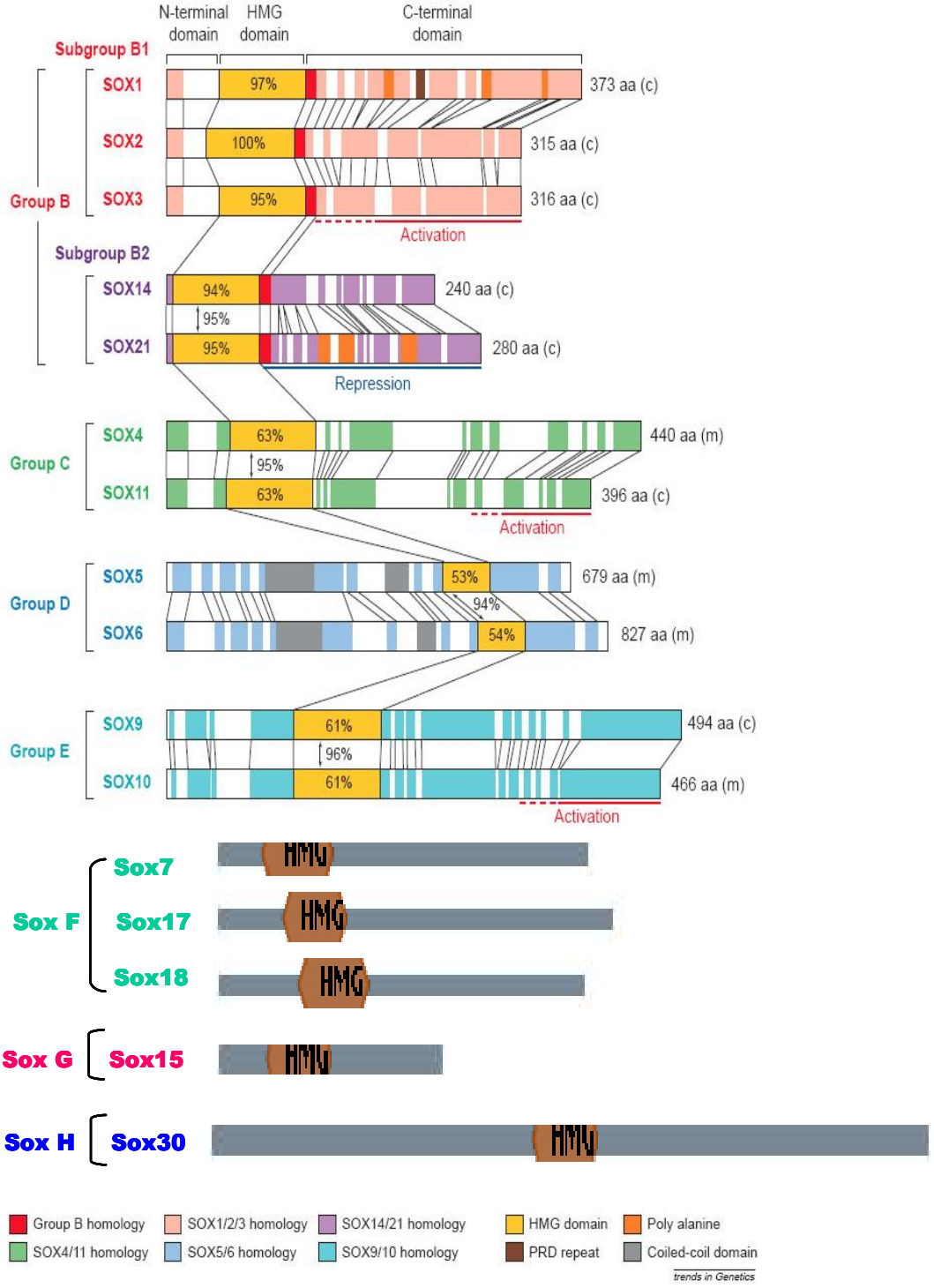


Figure 1.3 Classification of Sox transcription factors based on sequence similarity of the HMG boxes [22](Adapted from ref. *Kamachi,Y. et al., 2000*)

classification (*Fig. 1.4*) [23-25]. The prototypical SOX gene Sry, belongs to Group A Group B1 consists of Sox 1, 2, 3; Group B2 consists of 14, 21, 25; Group C consists of Sox 4, 11, 12, 22, 24; Group D consists of Sox 5, 6,13, 23; Group E consists of Sox 8, 9, 10; Group F consists of Sox 7, 17, 18; Group G consists of Sox 15, 16, 20; Group H consists of Sox 30; (*Fig. 1.3*) Group I consists of Sox 31 and finally Group J consists of Sox 32, and Sox33 [23].

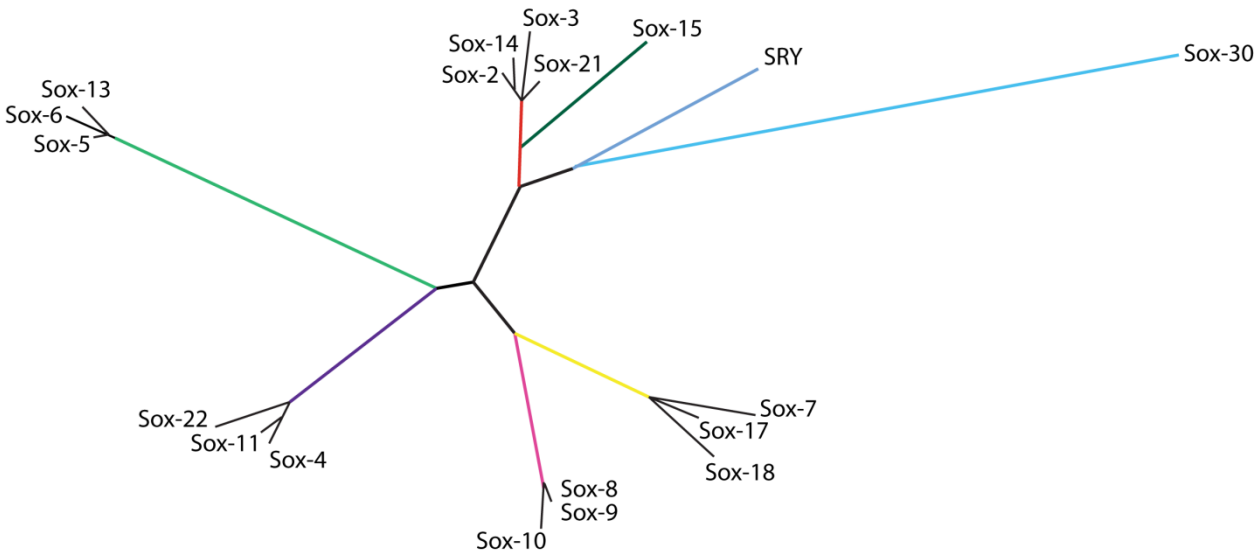


Figure 1.4 A neighbor joining phylogenetic tree generated using MAFFT Visualized using splitstree showing the different groupings of representative human Sox-HMG domain sequences [23-25]

Multiple sequence alignment of HMG domains of proteins from different organisms such as Yeast, *Caenorhabditis elegans*, *Drosophila melanogaster*, mouse and human shows high sequence similarity between HMG domains of non-sequence-specific to sequence-specific HMG domains in the view of amino acid sequence irrespective of species and family [24]. Particularly, in the case of sequence-specific HMG members, the residues, V3, R5, P6, H63, H67, P68, Y70, Y72, R75, and R76, crucial in the ordering of the C-terminal region of the HMG, are highly conserved [26]. For positioning of the C terminus in the minor groove

residues P68, Y72, and P74, are critial role (*Fig.1.5*). Nonethless, these residues are divergent in HMG domains that bind DNA in a non-sequence-specific manner [27, 28].

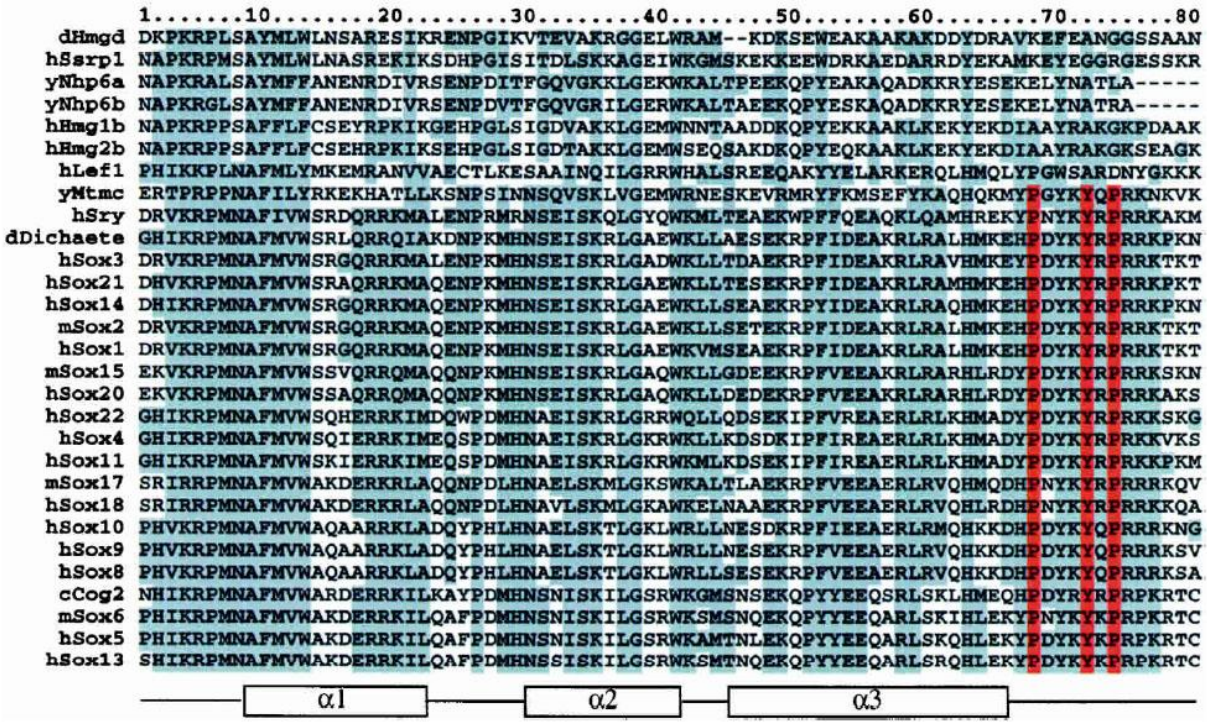


Figure.1.5 Multiple sequence alignment of HMG domains of several proteins from different organisms. (y) Yeast; (c) *Caenorhabditis elegans*; (d) *Drosophila melanogaster*; (m) mouse; (h) human. HMG domains from the first six proteins (dHmgd–hHmg2b) are known to bind DNA in a non-sequence-specific manner, whereas the others (hLef1–hSox13) bind DNA according to a specific sequence. Protein residues that are highly conserved are boxed in gray [26] (Adapted from ref. *Remenyi, A. et al., 2003)*

1.7 General Structural Features of Sox HMG Domain

The Sox HMG domain comprises ~ 79 residue three-helix bundle, exhibiting a characteristic L-shaped arrangement of the helices with an angle of ~80° between the arms. The long arm comprises the extended N-terminal strand and helix III, wheras the short arm comprises helices I and II (*Fig.1.6*). The Sox HMG domain binds to the consensus C(T/A)TTG(T/A)(T/A) motif [29, 30] and inserts the hydrophobic phenylalanine-methionine wedge into TT/AA DNA base pairs inducing ~70º kink (PDB:1GT0) [26, 31-33].

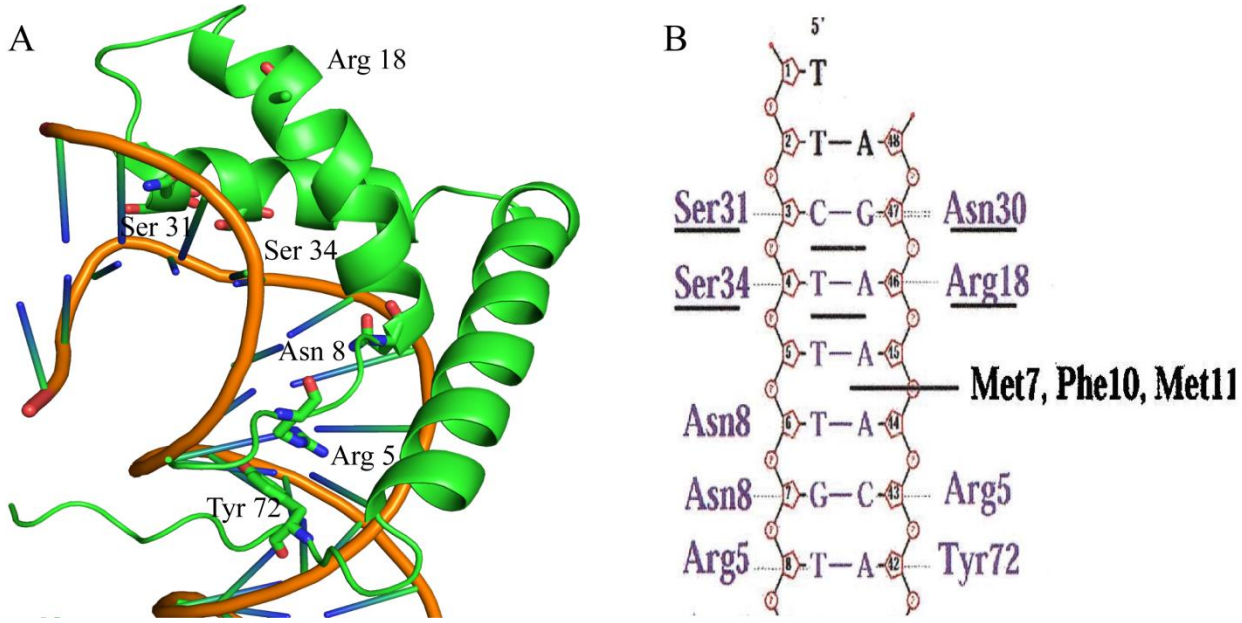


Figure 1.6 A) Crystal structure of the Sox2 HMG domain (in green)-*FGF4* complex; B) Protein–DNA interaction between the HMG and the *FGF4* enhancer. The hydrophobic side chains of M7, F10, and M11 are inserted between base pairs T5 · A45 and T6 · A44, causing a 45° bend of the DNA axis. These and other amino acid residues (represented as sticks in panel A) that play a role in bending the DNA at three different base stack levels are either underlined or written in black. (Adapted from ref. *Remenyi, A, et al., 2003)*

Several studies suggest the functional specificity of Sox proteins are determined by either (i) imparting structural rearrangements to the flexible long and short arms of the HMG domain (ii) by inducing specific kinks to the DNA structure [26, 31, 33, 34] (iii) or by recruiting co-factors. Structural comparison of the single HMG-boxes of Sry [31], Lef-1 [32], HMG-D [35] and Sox2 [24] in complex with DNA reveals that despite similar contacts of the helices 1 and 2 with the DNA minor groove, the regulatory specificity is achieved through subtle disparities in the interaction of the HMG C-terminus with DNA. In the case of Lef-1 HMG domain, the C-terminus lies within the compressed major groove, stabilizing the bent DNA conformation, while the C-terminus of the Sry HMG domain is disordered and is not placed in the minor groove [36]. Contrastingly, the Sox2- HMG C-terminus positioned closely into the compressed minor groove with the company of interacting POU domain (*Fig. 1.6*) [24].

Although, Sox induced DNA kinks have been proposed to regulate the assembly of enhanceosomes, critically dependent on the local shape of the DNA[37, 38], comparison of the structures reveals the DNA bend angles of 111° for HMG-D, 117° for Lef-1, 117° for Sry and 90° for Sox2 to be rather within a narrow range. Nonetheless, the regulatory potential of altered DNA bending angle was vividly demonstrated using circular permutation assay of Sox2 on Fgf4 promoter coupled with transfection [17].

Another prominent attribute of all Sox proteins for their gene activation is their habituation on other transcription factors as partners. [22, 39, 40]. The functional and tissue specific gene expression of the Sox proteins are largely contingent on its differential partnership with other transcriptional regulators. Sox proteins are well known to physically interact with POU or Pax transcription factors which involved in the regulation of eye lens development and stem cell pluripotency [20, 26, 41-44].

1.8 An overview of Sox Group E and Domain Structure of Sox9

Group E of Sox HMG proteins comprises of Sox10, Sox9 and Sox8 (*Fig 1.8A*). The overall identity of amino acid sequence among Sox10 and Sox9 is 54% wheras, Sox8 against Sox9 or Sox10 is 47%. Nonetheless, considering the 79 amino acid HMG domain, Sox8 and Sox9 dissent by one residue, Sox8 and Sox10 by five residues and Sox9 and Sox10 by 4 residues. Moreover, the group specific conserved DNA-dependent dimerization domain, N-terminal to the HMG domain consists of 40 amino acids exhibiting 70-85% identity [45-47] (*Fig 1.7).* Apart from these domains, there are other regions which are conserved in the SoxE group members like C-terminal 20 residues with 75-84% identity and residues 74-82 with 56-71% identity acting as strong and weak transcription activation (TA) domains in Sox8 [48, 49] and Sox10 [50] respectively, but absent in Sox9 (*Fig 1.7).* The C-terminus of Sox9 and Sox10 has a strong TA domain [51] [52]. In the context of chondrocyte differentiation, the Sox9 C-terminal trans activation domain physically interacts with ß-catenin [53] and with

transcriptional co-activators TRAP230/ MED12 and p300/CBP [54]. The Sox9 characteristic proline – glutamine - alanine (PQA)-rich motif, absent in Sox10 and Sox8, further augments its TA domain strength [55]. Two nuclear localization signals (NLSs) and a nuclear export signal (NES) are situated at the end and center of the HMG domain respectively. No function has yet been attributed experimentally to the K2 region between residues 233-306 but has been shown to possess possible transactivation potential in Sox8 [56] (*Fig 1.7*).

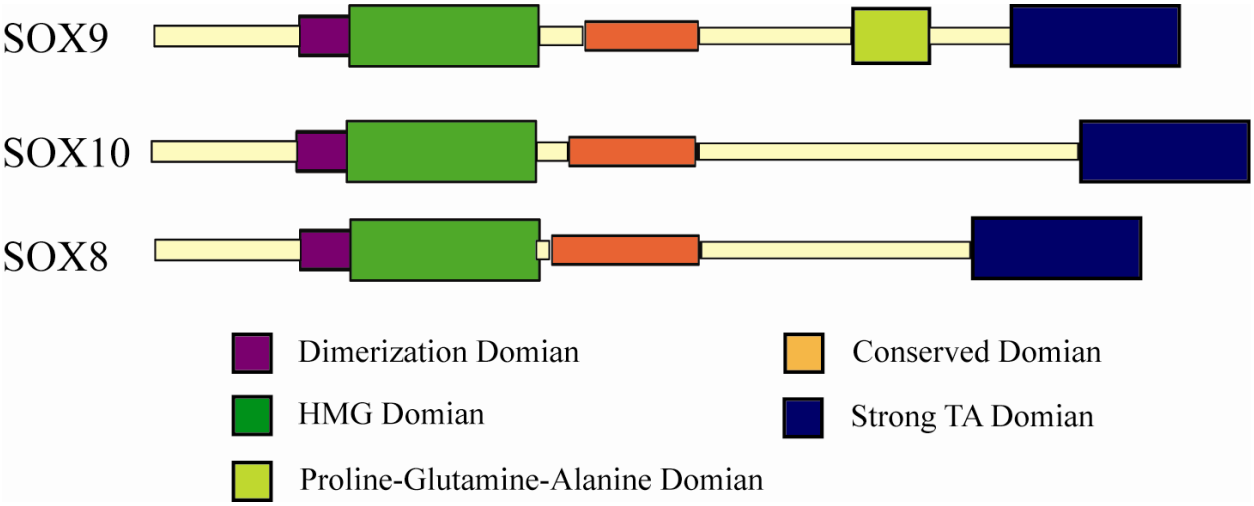


Figure 1.7 Structural and Functional domains of SoxE Group of Transcription Factors.

In humans, Sox10 is essential for various functions in nervous system development, such as neural crest and peripheral nervous system [57, 58] and mutations in the gene are known to be associated with Waardenburg syndrome [59]. Sox10 has been known to interact with Pax3 and MITF in activation of the c-RET enhancer. Sox 10 and Sox 8, as heterodimers influence oligodendrocyte differentiation and development of myelin-forming oligodendrocytes by binding to the myelin basic protein (Mbp) promoter [60]. Heterozygous loss of Sox10 plus loss of Sox8 has been shown to cause significant depletion in differentiated oligodendrocytes [61, 62].

Sox9 is a fundamental sex determining gene, involved in the development of various vital organs like testes, kidney, heart, brain and skeletal development. Sox9 is known to play

a pivotal role in chondrocyte differentiation and has been reported to precede Sox10 expression in glial precursors, suggesting its role in oligodendrocyte development [61, 62]. Sox9 along with Sox5 and Sox6 of Group D is critical for cartilage development. Mutations in the Sox9 gene have been known to cause campomelic dysplasia, a skeletal malformation syndrome (*Fig. 1.8*) [53], [63], [64]. Sox9 gets activated downstream of Sry and triggers testis differentiation by stimulation of sertoli cells [65]. A notable observation of Sox9 activity is that its dimerization domain is necessary for transcriptional regulation of chondrogenesis but not for sex determination [66].

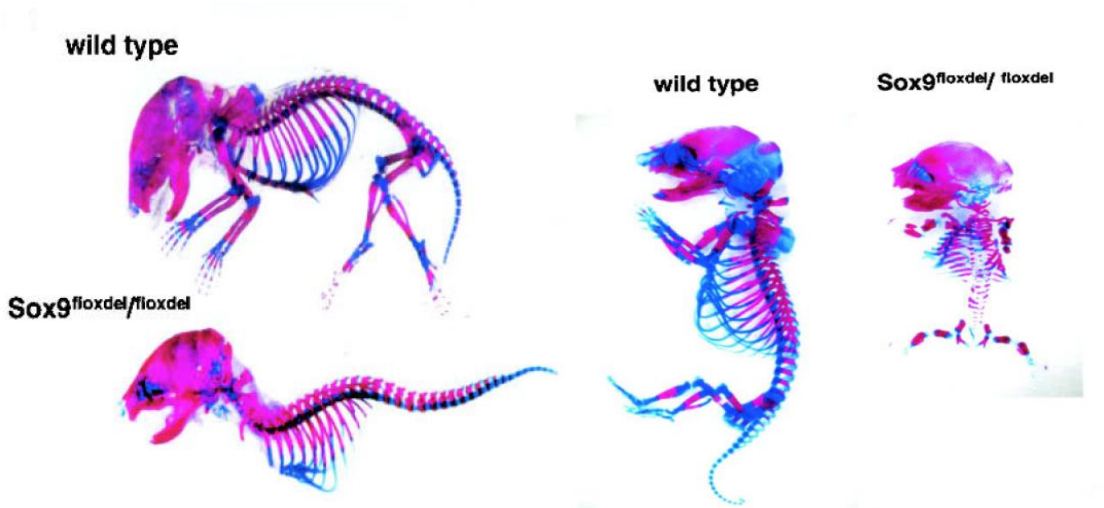


Figure 1.8 Skeletal phenotype of Sox9 mutant mice indicating absence of cartilage and bones in the limbs, characteristic of campomelic dysplasia [53](Adapted from ref. *Akiyama,H.,et al., 2002)*

1.9 Sox9 and Sry in Male Sex Determination

Sex determination in mammals center upon the Y-chromosom of male sex-determining gene Sry, indispensable for testis development. In the absence of Sry, XY humans develop female genitalia [67]. Little is known about Sry gene requlation in spite of its known importance in gonad development. Identification of biologically important mechanism in regulation of Sry function made possible through naturally occurring mutations which caused in XY sex reversal in humans. (*Fig 1.9*).

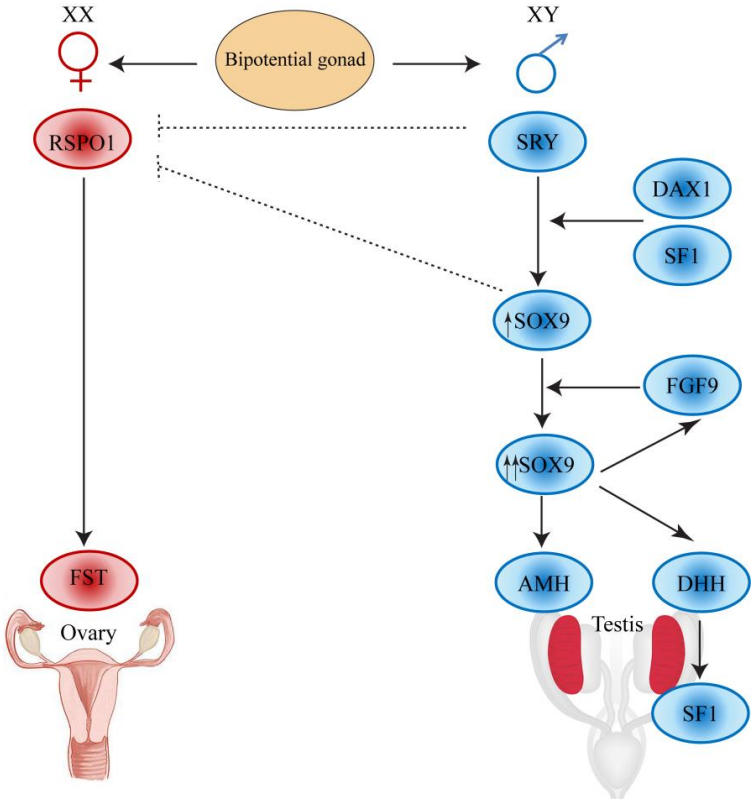


Figure 1.9 Role of Sry in sexual determination indicating feedback loop regulation of Sox9 by Sry and Fgf9 in male testis development.

Sox9 and Sry proteins are functionally important transcription factors in mimicking each other in implicating sex determination In pre-Sertoli cells, Sox*9* is expressed downstream of Sry, and ectopic expression of Sox 9 induces testis development [68, 69] suggesting Sry's predominant role is to activate expression of Sox9 [70]. Likewise, *Fgf9* has been implicated as yet another important regulator of Sox9. Positive feed back loop mechanism works between *Sox9* and *Fgf9* during the expression of FGF9 in XY gonad primordia. So, mice lacking Fgf9 undergoes sex reversal and this idea explains in XY cells the Sry uprequlation of Sox9 leads to loop between Sox9 and Fgf9, which encage male sex determination. (*Fig 1.9*) [70].

1.10 Sox9: Master regulator of Chondrogenesis

Several evidences indicate Sox9 as the master regulator of chondrogenesis, regulating various stages of cartilage development. Chondrogenesis in mesenchymal cells is driven by the overexpression of Sox9 that directly activates transcription of cartilage and extracellular matrix specific genes typified by *Col11a2* and *Col2a1*, and *Aggrecan*. Bone, cartilage development and osteoblasts developments are greately disturbed if Sox9 is carry off from undifferentiated mesenchyme. This clearly proves that osteochondro progenitors requires Sox9. (*Fig 1.10*). Inhibiting osteoblast developmental regulator ß-catenin, by an antagonistic relationship, Sox9 confirms commitment to the chondrocyte lineage (*Fig1.10*) [71]. In humans, haploinsufficiency of Sox-9 manifests as campomelic dysplasia, patients exhibiting skeletal abnormalities, as well as other phenotypes [72]. Runx2, another regulator of osteoblast development, is also inhibited by Sox9, by decreasing Runx2 affinity to target sequences [73].

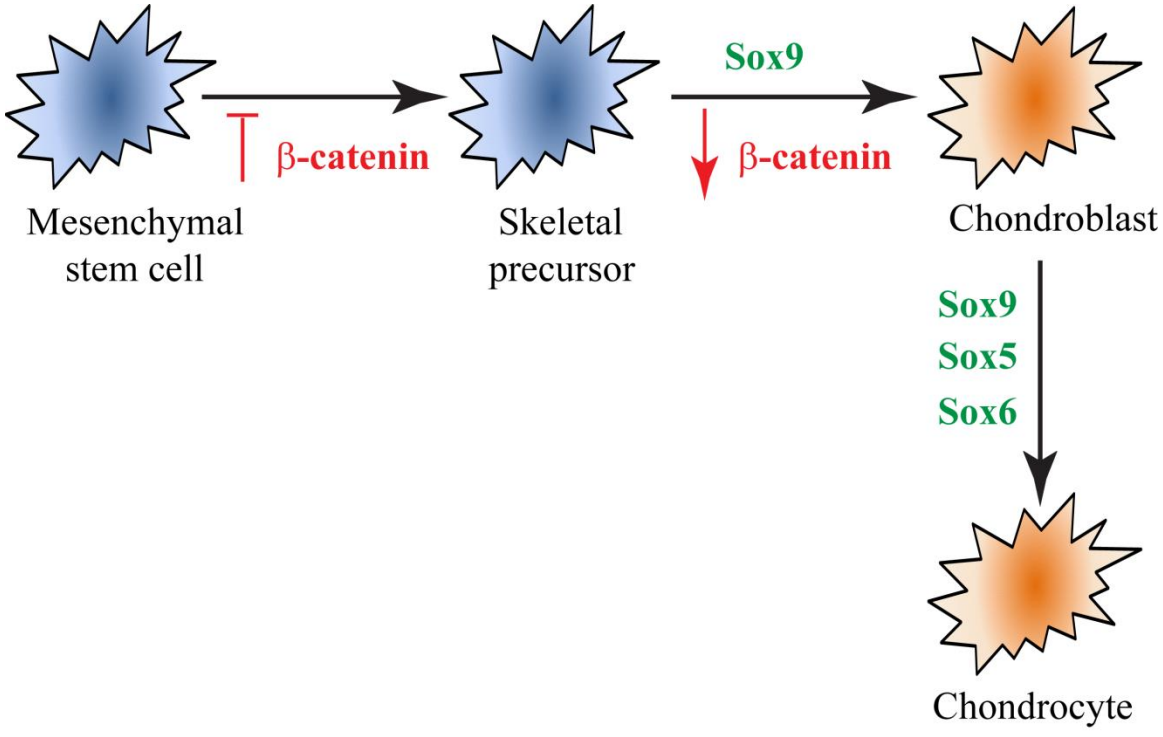


Figure .1.10 Cartoon representation of Sox9 role in chondrogenesis.

Sox9 undergoes numerous posttranslational modifications, significant for the regulation of several signaling pathways e.g., phosphorylation by cGMP-dependent protein kinase II (cGKII) [74], PKA signaling phosphorylation increases Sox9 DNA binding affinity [75] and PIAS1 SUMOylates Sox9, causing stabilization and augmented activation of target, gene *Col2a1* [76]. Deciphering the posttranslational modification code of Sox9 would provide a clear insight into Sox9 mediated chondrogenesis regulation [77, 78].

1.11 *SoxD* Genes Regulate Differentiation, Downstream of *Sox9*.

With ~20 SOX genes in the mammalian genome one and the same Sox protein must participate in many developmental processes [17]. During development, SOX genes are expressed in a wide variety of tissues and possess discrete function in discrete developmental processes. Besides two or more simultaneously expressed Sox proteins may play overlapping or synergistic functions [17]. n case of Sox9, SoxD proteins Sox6 and L-Sox5 have redundant roles in chondrogenesis (*Fig.1.11*) and double mutants of SoxD subgroup genes are severely affected [79]. Sox6 and L-Sox5 expressed in mesenchymal condensations are severely down-regulated in Sox9 conditional mutants (*Fig. 1.12B*). Though LSox5 and Sox6, lack transactivation domains, homo [80, 81] and heterodimerize [82] by means of coiled-coil domains, increases their binding affinities for pair of sites, considered critical for their function.

Co-expression of these two proteins has shown to greatly potentiate transactivation of *Col2a1* by Sox9 [83], likely by altering the architecture of target promoter/enhancers, enabling transactivation by Sox9.

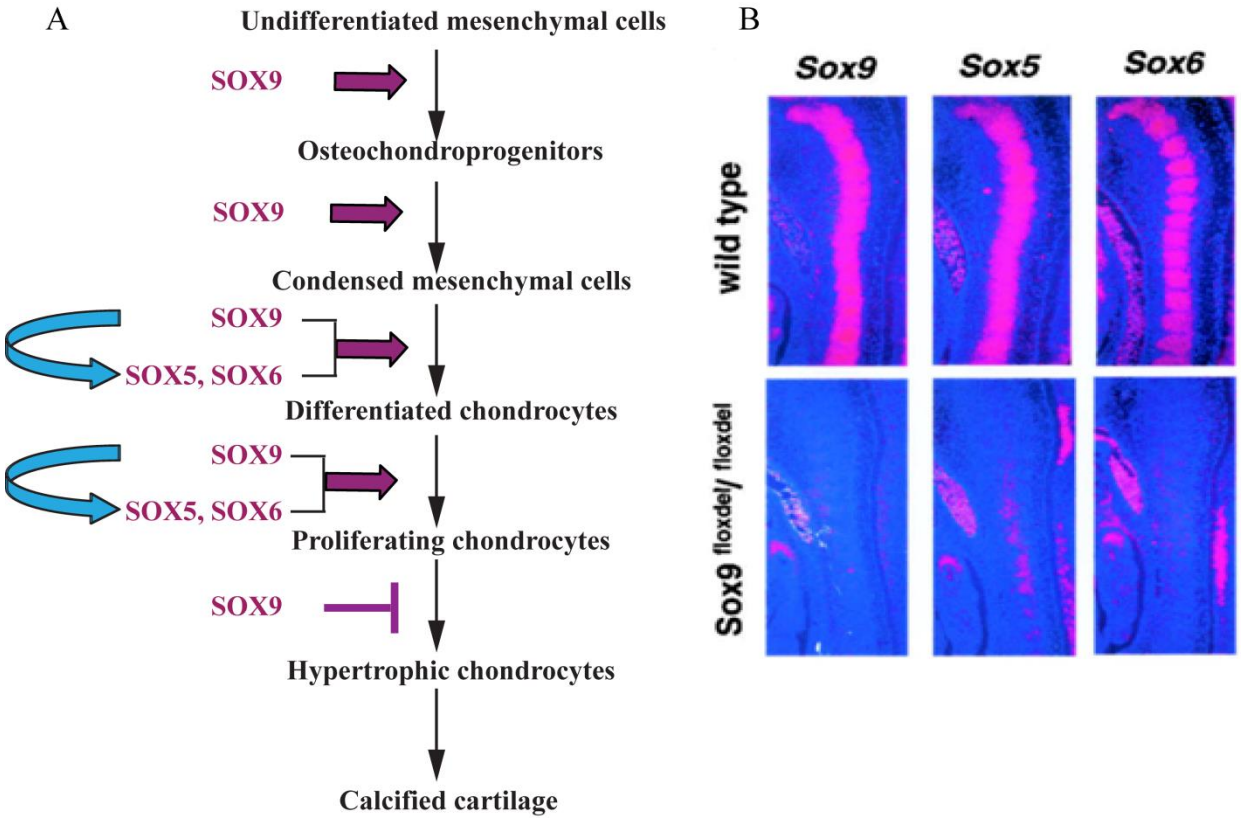


Figure 1.11 Sox9 down stream regulation. Panel A) Interplay of Sox9, Sox6 and Sox5 in chondrocyte differentiation pathway and Panel B) RNA in situ hybrisation of Sox9 mutant mice showing down regualtion of Sox 5 and Sox 6 suggesting these genes to be downstream of Sox9 in chondrogenesis. (Adapted from ref. *Akiyama,H.,et al., 2002)*

1.12 An overview of SoxD proteins

SoxD group of proteins includes Sox5, Sox6, Sox13 and Sox23 (*Fig 1.12*). The human SOX5 and SOX6 genes are found in paralogous chromosomal regions, more closely related to each other than to SOX13. The SOX5 and SOX6 genes having 12-16 coding exons are distributed over 300-400kb of genomic DNA wheras SOX13 is only 12kb dispersed. [84]. In adult testis, both Sox5 and Sox6 are expressed as short transcripts of 2 and 3kb, respectively, however in other tissues are expressed as long transcripts of 6 and 8 kb, respectively. Both Sox6 transcripts transcribe full-length protein. Conversely, the short Sox5 transcript discovered first and named Sox5 transcribes a protein isoform that lacks the N-terminus of the full-length protein [85]. The later identified full-length Sox5 isoform, named

L-Sox5 or Sox5-L (*Fig 1.13*) [82, 86] is structurally and functionally equal to Sox6 and Sox13 [87].

Sox5 and Sox6 are known to antagonize and negatively regulate the SoxE family in oligodendrocyte development [88]. The mechanism behind SoxD negative regulation of SoxE is attributed to two primary reasons. Firstly, SoxD proteins compete with SoxE proteins for the same binding sites as observed in competition between Sox5 and Sox10 to bind to myelin gene promoters [82, 89]. Secondly, as reported in the repression of Sox10 activation of myelin gene promoters, SoxD (Sox5) have been reported to recruit co-repressors of the SoxE protein expression [89]. The prevalence of auto-antibodies against Sox13 in patients suffering from type 1 diabetes has been reported implicating the role of Sox13 in auto-immune diabetes [90]. The expression of Sox13 is controlled by another helix-loop-helix leucine zipper called MITF in melanocytes and osteoclasts [91] [92].

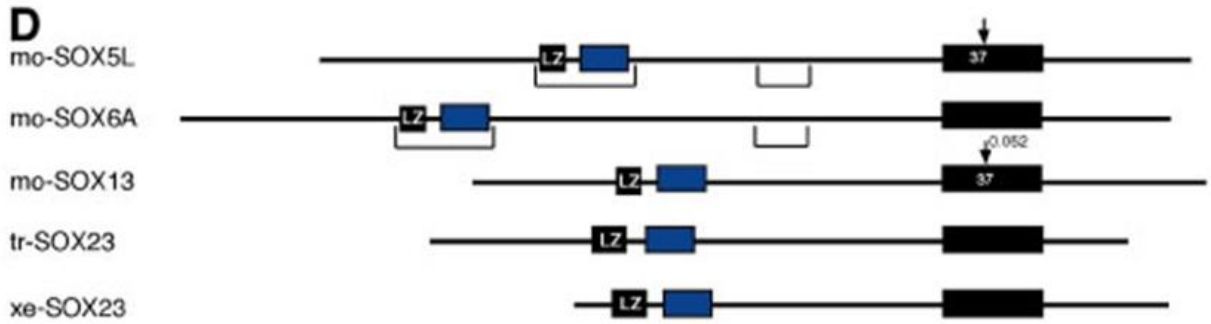


Figure 1.12 Structural and Functional domains domains of SoxD Transcription Factors. C-term Black Box-HMG Domain; LZ-Leucine Zipper; BlueBox-Proline Glutamine Rich Domain; ⌊⎽⎽⎽⌋- Coiled-Coil Domain.

SoxD proteins are the largest proteins of the Sox family encompassing two significantly conserved functional domains. The C-terminally located, conserved DNA-binding HMG domain, (*Fig1.12*) binds DNA with conserved AACAAT motif in gel retardation assays [83]. In the case of human and mouse SoxD proteins the HMG domain is 87% identical while in other Sox proteins the similarity is less than 60. The second important SoxD group specific domains are the pair of coiled-coil domain present in all four Sox group

D proteins. The domains, one present at the N-terminal half and the other at the middle of the protein, share about 76% identity among human and mouse SoxD proteins (*Fig1.12*).

1.13 Sox HMG- DNA interactions

So far around 30 sox proteins have been identified, yet there is no high resolution structure for any full length Sox proteins. Only three crystal structures of HMG domains are available, Sox2 [36], Sox17 [93] and Sox4 [94] belonging to different Sox groups. The three helixes of Sox HMG protein arranged in the characteristic L-shape and the N and C-termini are placed in the same molecular surface. Sox2 is a prototypical representative of the Sox-HMG domain family which is approximately 40% identical to the rest of the family members [22, 30, 95]. The DNA binding interface of Sox2 is predominantly cationic, extended, with no binding pockets and bends DNA at an angle of ~70 ° comparable to other Sox proteins like Sox4 and Sox17 *(Fig. 1.13*) [26, 33, 96]. The binding of Sox17 with *Lama1* DNA, with respect to B-DNA induces topological deformations by a bend of approx 80°. The widening of minor groove and the decrease in the width of major groove at the core of the interaction site after binding are important structural feature for its functional activity. *(Fig. 1.13C).* Sox17-HMG and the *Lama1* DNA interaction interface are mediated by the aminoacid residues of N-terminal of the HMG domain. The minor groove of the DNA is contacted by Arg, Asn, Ser and Tyr [93] and of the nucleotide TATTGTC, with base pairs TATTGT directly contacted by residues of Sox17-HMG *(Fig. 1.13C).* In comparison with Sox17, Sox4 was found to deform *Lama1* DNA by binding to the minor groove (*Fig. 1.13B),* with a helical bend axis of 65.8° identical to Sox17 with a helical axis of 68.9° suggesting differential deformation of DNA has no drastic contribution to the DNA associated conformational changes of Sox4 and Sox17 proteins. The DNA contacts are also identical for both proteins except, Arg18 and Asn30 which show a different pattern of hydrogen bonding.

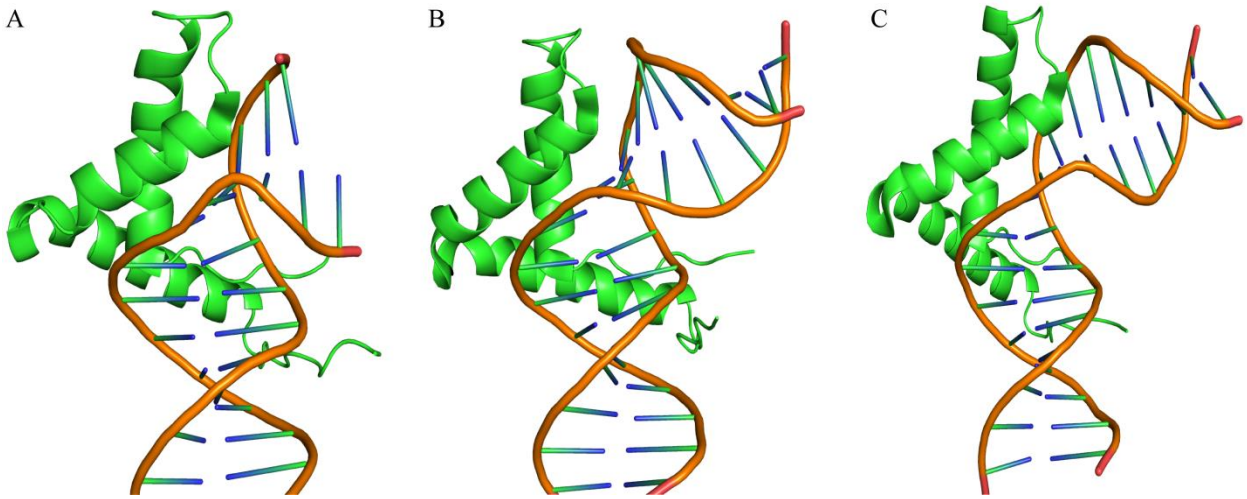


Figure 1.13. Structural comparison of three different group of Sox transcription factors HMG domains highlighting the protein DNA contacts. Crystal structure of A) Sox2 [36], B) Sox4 [94] and C) Sox17 [93] proteins with DNA. (Adapted from ref. *Remenyi, A., et al., 2003, Jauch, R et al., 2011 and Palasingam, P., et al, 2009)*

1.14 Protein-protein interactions: SoxE- DNA dependent Dimerization Domain

DNA dependent dimerization is an important feature in many nuclear receptors such as Androgen receptors, orphan receptors, vitamin D (VDR) etc. and plays a crucial role in facilitating nuclear localization, cofactor binding, DNA binding, and transactivation [97, 98]. The Androgen Receptor (AR) is a modular protein which consists of various functional domains. The androgen bound activated receptor detaches from its cytoplasmic chaperone complex and the DBD induces a conformational change. Thus the receptor monomers oligomerises on specific DNA targets to achieve cooperative dimerization essential for DNA binding [99-101]. Margaret *et al*.[102] demonstrated that AR preferentially binds DNA as a dimer. A telomere-binding protein consisting of alpha and beta subunits exists preponderantly as a monomer in the absence of telomeric DNA. Upon binding to DNA, the alpha and beta subunits interact to form a heterodimer suggesting DNA dependent heterodimer formation. Further analysis confirms the telomeric complex to contain one alpha subunit, one beta subunit, and one DNA molecule [103]. Dimerization with DNA binding is an important feature of many nuclear receptors; thyroid hormone (TR), 1 retinoic acid (RAR), vitamin D

(VDR), eicosanoids (PPAR), and a number of orphan receptors with retinoid X receptor (RXR) [104, 105]. DNA dependent homodimer formation has also been observed in other DNA-binding proteins like glucocorticoid receptor proteins and GAL4 DNA-binding domains [106-109].

As presented above (Section 1.7) the DNA dependent dimerization domain of Sox9 is highly conserved in all members of SoxE transcription factors. Mutations in this dimerization domain selectively reduce the DNA binding affinity and abrogates dimerization, thus, interfering with promoter activation through natural target sites that require binding of Sox9 dimers [46] demonstrating dimerization as an essential component. Mutations in Sox9 cause Campomelic Dysplasia (CD) marked by anomalies of the ribs, vertebral column, bowing of the long bones and importantly male-to-female sex reversal [110] . The initial sequencing report of the entire SOX9 ORF of CD affected patient indicated a heterozygous point mutation in codon 76 of the dimerization domain, leading to mutation of an alanine to a glutamic acid residue (A76E; GCG to GAG) highly conserved in all members of Sox group E [46]. A common feature observed in the binding sequences of Sox enhancers genes like Col11a2, Col9a2, Col2a1, Col9a1, aggrecan and CD-rap [111-115] is the presence of multiple Sox binding sites. Interestingly, genes involved in chondrogenesis possess multimerized Sox binding sites e.g., Sox9, Sox5 and Sox6 bind four consensus sites in the Col2a1 enhancer. While the two known Sox target genes in gonadal and reproductive tract development, SF1 and AM have a single Sox binding site to which Sox9 binds as a monomer. On this basis, it has been hypothesized that the DNA dependent dimerization of Sox9 is prerequisite for bone development however gonadogenesis would require only a monomeric form of Sox9. The other member of SoxE group, Sox10 was shown to bind to DNA as dimer in a cooperative manner [45] only in the presence of the 40 amino acid dimerization domain preceding the N-terminal of the HMG domain. Thus *in vitro* and *in*

vivo studies show that dimerization is important for SoxE group, for specific functions in the physiological context of Sox9 target gene enhancers. The determinants for dimeric or monomeric binding dictate the function of Sox9 during development and thus its functional specificity is achieved by the level of complexity and flexibility of the dimerization domain and additional versatility is achieved through sox9/sox10 heterodimerization.

1.15 Protein-protein interactions: SoxD- Coiled coil domains

Most notably the coiled-coil domains are found in DNA binding protein transcription factors, particularly leucine zippers, for example, bZIP transcription factors fos and jun [116] and yeast transcriptional activator GCN4 [117]. Coiled-coil interactions involving zipper domains are necessary for the dimerization and subsequent DNA binding of numerous transcription factors and can regulate the specificity of dimerization of factors with the basic leucine zipper (bZIP) domain (*Fig1.15*). The coiled-coils can also be found in a large number of other proteins like macrophage scavenger receptor [118]. In such DNA binding coiled-coils, the proteins are active only when they are allowed to form their oligomer status and the DNA binding region is towards the amino terminal end from the zipper domain [119]. Different combination of the subunits of transcription factors are held together by the dimerization of the coiled-coil in basic zipper domain [120] critical in regulating subcellular localization.

Contrasting to classic bZip transcription factors, the CC domain of SoxD proteins has two striking differences. Firstly, the DNA binding HMG domain of SoxD proteins is towards the C –terminus of the CC domain as opposed to the amino terminus in other leucine zipper transcription factors and secondly, the leucine zipper of Sox D coiled-coil domains is not adjacent to a basic region but, is always followed by a Q-Box domain (*Fig.1.15*). Owing to this unique property, dimerization does not lead to new DNA-binding interface as in the case of bZip proteins (*Fig.1.14*). Thus dimerization of SoxD lowers the affinity of the dimer for

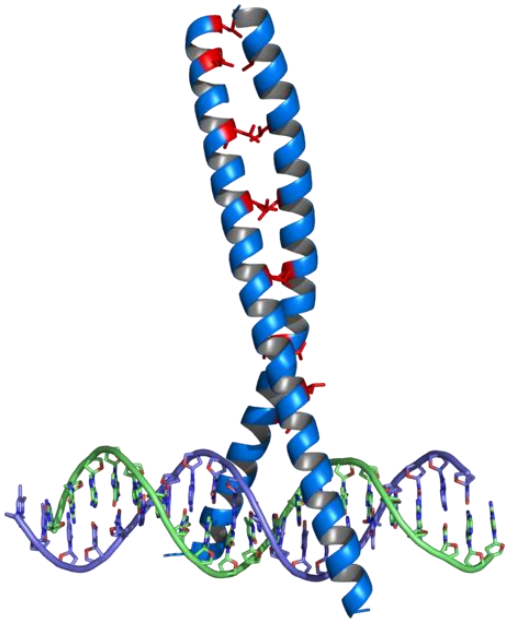


Figure.1.14 Leucine Zipper (blue) bound to DNA. The leucine residues that represent the 'teeth' of the zipper are colored red. (Adapted from ref. *Splettstoesser, T., 2007*)

single binding sites [80, 81] and augments binding to pair of consensus sites. Both *in vitro* and *in vivo*, SoxD proteins competently bind sites harboring one or two mismatches and display low preference specificity for the relative positioning or the length intervening the paired sequence [121]. Thus, contrasting to Sox9 and other Sox proteins, the SoxD proteins are more flexible in binding DNA sequences. Considering the high flexibility, putative SoxD binding sites could be seen in most promoter or DNA regulatory elements.

Although the SoxD proteins lack a transactivation/repression domain, they participate in both activation and repression of transcription by binding to sites discrete from that of Sox9. Specifically, in erythroid cells, Sox6 represses expression of embryonic globin genes while Sox5 and Sox6 induce transrepression by binding to proximal promoter sequences of CtBP2 co-repressor and histone deacetylase Hdac1 [88], indicating SoxD proteins to act either as positive or negative modulators of transcription and likely through diverse mechanisms.

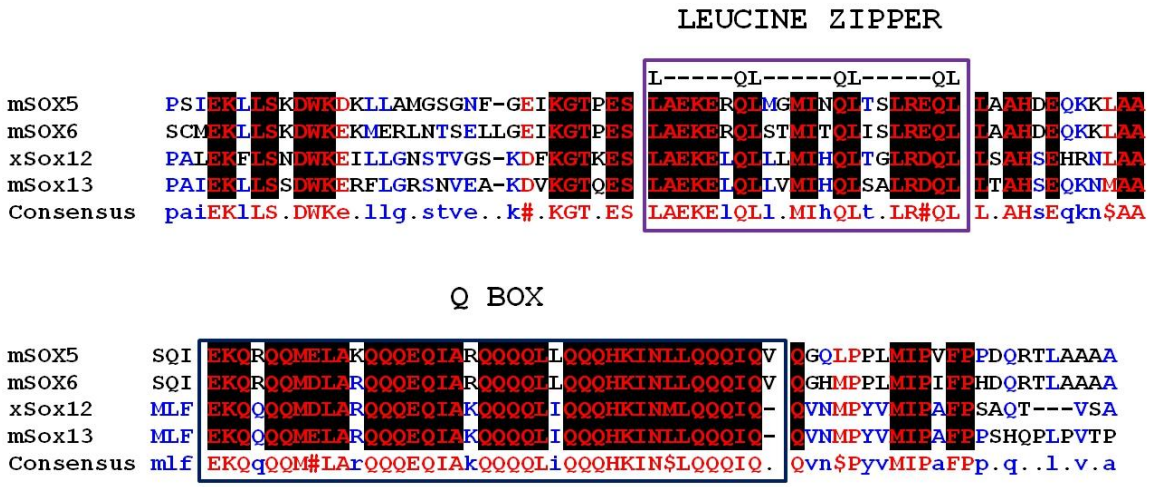


Figure 1.15 Multiple sequence alignment of shows Coiled Coil and Q-Box domain of SoxD proteins. Conserved domains are shaded in black.

1.15 Biological Functions of Coiled-Coils

Coiled-coil proteins dictate key biological functions such as the regulation of gene expression e.g. transcription factors. Typical examples are the oncoproteins c-fos and jun, [116] [122] and the muscle protein, tropomyosin. Coiled-coils (CC) domains are found to be in about 2-3% of naturally occurring proteins [123]. The coiled-coil motifs are best suited model systems for understanding protein folding [124], molecular recognition[125] and *de novo* protein design, [126, 127]. An important property of coiled- coil is its plasticity in structure accommodating changes in oligomeric state, strand polarity, homo- v*ersus* heteromeric association, heptad register of helices, [128, 129]. CC domains are abundant in structural proteins, for example, collagen, Matrilins (skin, bone, connective tissue) [130], keratin (hair, nails) and myocin (muscle protein), components of cytoskeleton (intermediate filaments) [122]. The coiled-coil structural feature is ideal for filamentous structure formation. Fibritin, coiled-coil containing protein of bacteriophage T4 belongs to a class of chaperones involved in the specific phage-assembly and assembly of long tail fibers. Fibritin's sensing ability controls the withdrawal of the long tail fibers in extreme environments and that helps prevention of infection [131]. A network of dynamic protein

filaments, the cytoskeleton, is an important tool to structurally orchestrate the cells both in eukaryotes and prokaryotes. In this regard, the identification of small synthetic peptides representing the N- and C-terminal heptad repeat regions of gp41 is a significant discovery for the design of novel potent HIV entry inhibitors for prophylaxis of HIV infection and AIDS [132].

1.16 Objective of study

From the literature presented it is apparent that Sox transcription factors are key determinants of several developmental processes and mutations in human SOX genes cause complex disease syndromes. Specfically, mutations in SOX5 (GroupD) and SOX9 (GroupE) of the Sox trio result in campomelic dysplasia, a disorder marked by cartilage and bone defects, XY sex reversal, and anomalies of the heart, kidneys, brain, gut and pancreas. It is also evident that GroupD SOX genes are expressed in glioma, prostate, and other types of tumors [133-135]. In particular, SOX5 increases progression of nasopharyngeal carcinoma [136] and SOX13, one of SoxD protein, known as the islet cell auto-antigen 12, induces autoimmune diseases such as type I diabetes and primary biliary cirrhosis [90].

Although all Sox proteins which possess highly conserved HMG domains bind similar DNA elements, they regulate a wide assortment of genes in diverse developmental processes. The functional specificity of Sox transcription factors plausibly depends on (i) subtle nucleotide variations in the DNA sequence; (ii) differential minor groove bend as a consequence of HMG domain mediated DNA interaction (iii) different co-factor recruitment through protein-protein interactions. In this regard, Sox9 (group SoxE), Sox5 and Sox6 (group SoxD), famously known as “Sox trio”, are ideal prototypes with conserved HMG domains and group specific domains for partner recruitment. Accordingly, the current study addresses protein-DNA interaction and protein-protein interactions of Sox9 and Sox5.

Among all Sox sub groups only the groupD and groupE encompass group specific protein interaction domains, “a DNA dependent” dimerisation domain and “a DNA independent” coiled-coil domain respectively. Presence of such oligomerizing domains might possibly provide Group D and Group E proteins with an additional level of transcriptional specificity. Until now, the mechanism of Sox mediated regulation is still at the primeval

stages and the precise role of the group specific protein interacting domains on Sox transcriptional regulation has not been studied.

As the exact contribution of Sox in sex reversal, cancer development and other diseases remains unknown, the objective of the current study is to comprehend the underlying molecular mechanism of DNA recognition and transcriptional specificity of the Sox trio. The objective would be addressed by accomplishing the following specific aims:

1. Biochemical validation of in-vivo identified novel Sox regulatory motifs to understand how subtle variations in the DNA sequence might drive target gene specificity;
2. Characterization of the DNA binding HMG domains of Sox9 and its downstream regulated Sox5;
3. Characterization of the DNA-dependent dimerization domain of Sox9;
4. The role of coiled-coil domain mediated oligomeristion in Sox5.

CHAPTER II

Materials and Methods

2.1. Chemicals

The fine chemicals routinely used in the laboratory for the biochemical and molecular biology experiments such as agarose, ampicillin, IPTG, Tris, NaCl, Imidazole, HCL, Ni-NTA were purchased from Sigma-Aldrich, Fluka and Qiagen. Clonase, TOPO vectors from Invitorogen; polymerases, protein markers and DNA markers were purchased from New England Biolabs. Crystallization screens were obtained from Hampton Research, Qiagen and Innovadyne. The 96-well multicavity plates used for crystallization were from Innovadyne. Most of the other chemicals used in the study were of analytical grade, purchased from Merck.

2.2 Plasmids

pUC-derived gateway destination vectors pETG-20A (Fig.2), pETG-30A, pETG-60A, pDEST-17 and pDEST-HisMBP (Invitrogen) (Fig.2) were used for cloning the gene of interest with a hexa-histidine tag at N-terminal (Table.2). All these designed high-level protein expression vectors include an ATG translation initiation codon, T7 or T7/lac promoter, TEV cleavage site for tag removal and a second optional tag.

2.3 Bacterial strains

Plasmids for cloning were propagated in *E. coli* DH5α strain. Protein expressions were carried out in *E. coli* BL21(DE3) harboring lambda DE3 lysogen (Invitrogen Cat.No.C6000-03), ideal for bacteriophage T7 promoter-based expression systems. BL21 (DE3) yields high level expression of recombinant proteins non-toxic to E. coli upon IPTG induction.

Vector	Promoter	Selection	Tag	Protease cleavage site	origin	Description
pDEST17	T7	amp	N-His	TEV	pBR322	expression vector
pETG-20A	T7/lac	amp	N-His N-Trx	TEV	pBR322	expression vector based on pET-22b
pETG-30A	T7/lac	amp	N-His N-GST	TEV	pBR322	expression vector based on pET-22b
pDEST-HisMBP	T7/lac	amp	N-His N-MBP	TEV	pBR322	expression vector based on pET-22b
pETG-60A	T7/lac	amp	N-His N NusA	TEV	pBR322	expression vector based on pET-22b

Table 2.1: A Gateway destination vectors used in this study for the production of recombinant proteins as fusions

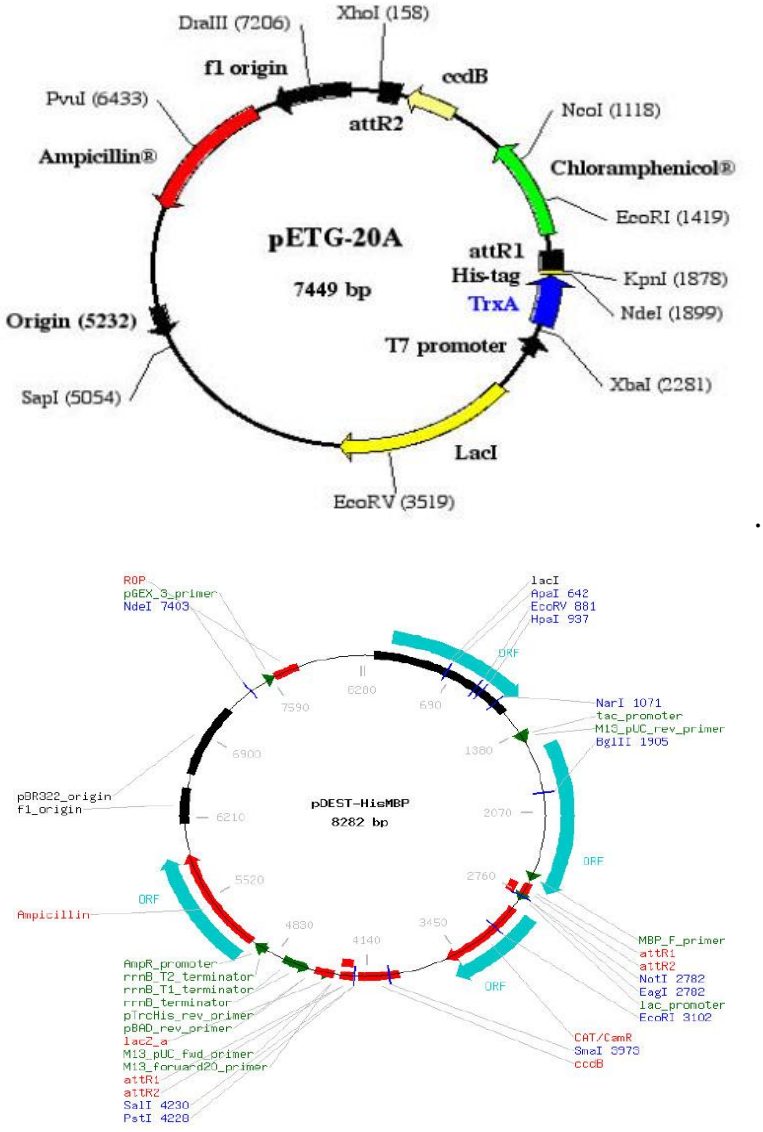


Figure.2.1 Vector maps of pETG-20A (EMBL-AG) (top) & pDEST-HisMBP (Addgene)**(below)**

2.4 Design of PCR primers

To perform directional TOPO cloning, the 4 base pair sequences (CACC) were included at the 5′ end of the forward primer and a stop codon, TTA at the 5′ of reverse primer. Oligonucleotides to be used as PCR primers were designed for a melting temperature (Tm) above 45° C and checked for primer dimerization and secondary priming sites with the programs "Net primer" online software (www.premierbiosoft.com) with default parameters. Primers used in the study are listed (Table 2.2)

mSox9_HMG	Forward	5' CACCCCACACGTCAAGCGACC 3'
	Reverse	5' TTACACCGACTTCCTCCGCCG 3'
mSox6_HMG	Forward	5' CACCCCCCACATCAAGCGACC 3'
	Reverse	5' TTAGCATGTGCGCTTCGGCCG 3'
mSox5_FL	Forward	5' CACCATGCTTACTGACCCTGATTTAC 3'
	Reverse	5' TTATCAGTTGGCTTGTCCCGCAATGTG 3'
mSox5_CC1	Forward	5' CACCACTCCTGAGAGCCTCG 3'
	Reverse	5' TTACATCAATGGCGGCAGCTGA 3'
mSox5_CC1_2	Forward	5' CACCACTCCTGAGAGCCTCG 3'
	Reverse	5' TTATATGCTGTTCACCACGGCC 3'
mSox5_CC1_2_HMG	Forward	5' CACCACTCCTGAGAGCCTCG 3'
	Reverse	5' TTAGGTGCGCTTTGGCCTAGG 3'
mSox5_HMG	Forward	5' CACCCCCCACATAAAGCGTCC 3'
	Reverse	5' TTAACAGGTGCGCTTTGGCCT 3'

Table 2.2 Primers used for Gateway cloning of the different constructs used in the study

2.5 PCR amplification

The polymerase chain reaction (PCR) (Saiki *et al*., 1988) is the basis of the method described here for the cloning of DNA into vectors. In a reaction volume of 50 µl, 10 to 100 ng of plasmid DNA containing the target sequence to be amplified was incubated on ice in a 0.2 ml safe lock Eppendorf tube containing 0.2 µM forward PCR primer, 0.2 µM reverse PCR primer, and Pfu DNA polymerase in reaction buffer (20 mM TrisCl pH 8.8, 10 mM KCl, 10 mM (NH4)2SO4, 2 Mm MgSO4, 0.1% Triton X-100) with 0.2 mM dNTP. Pfu DNA polymerase (2 units) was added and transferred to a thermocycler, pre-warmed to 94° C. The samples were incubated at 94° C for 2 min. Then Amplification was automatically carried out using the following typical scheme: (95° C, 30 seconds > 42° C, 30 seconds > 72° C, 60 seconds per kb) 25-30 cycles ["per kb" means per kilobase of expected PCR product]. A final "polishing" incubation of 2 min at 72° C was applied, followed by incubation at 4° C. DNA products were analyzed by agarose gel. The reaction products were cleaned by PCR purification kit (Qiagen).

2.6 Cloning methods

2.6.1. Directional TOPO Cloning

Directional TOPO® cloning enables cloning of blunt-ended PCR products in a 5′→3′ orientation directly into an expression vector through ligation. Directional TOPO® cloning vector (Table.2.4.1) contains a single-strand GTGG overhang at 5′ end and 3′ blunt end. The four-nucleotide overhang invades the double-strand DNA of the PCR product and anneals to the CACC sequence at 5′of forward primer. Topoisomerase I from *Vaccinia* virus binds to duplex DNA at specific sites (CCCTT) and cleaves the phospho diester backbone in one strand [137-140]. The energy from the broken phospho diester backbone is conserved by formation of a covalent bond between the 3′ phosphate of the cleaved strand and a tyrosyl residue (Tyr-274) of topoisomerase I. The phospho-tyrosyl bond between the DNA and enzyme can subsequently be attacked by the 5′ hydroxyl of the original cleaved strand,

reversing the reaction and releasing topoisomerase. Thus Topoisomerase I ligates the PCR product in the correct orientation (Fig.2.4.1).

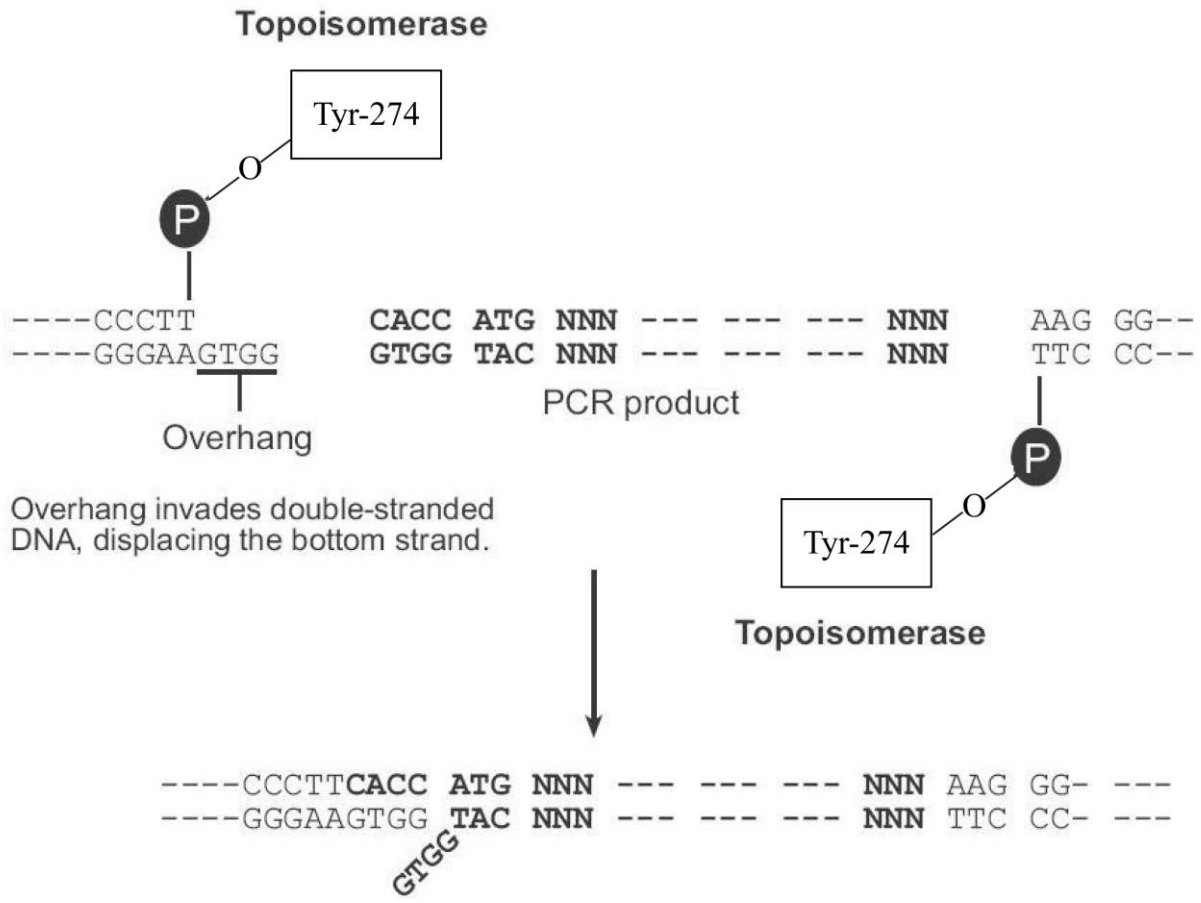


Figure 2.2 Schematic reperesentation of directional TOPO cloning

Vector	Promoter	Selection	Protease cleavage site	origin	Description
pENTR™/TEV/D-TOPO® Cloning vector	T7	Kan	TEV	pUC	Cloning vector

Table.2.3 TOPO Cloning Vector.

2.6.2 Gateway® Technology

The Gateway® Technology is a universal cloning method that takes advantage of bacteriophage lambda's site-specific recombination properties and [141] provides efficient way to clone the desired DNA across multiple destination expression vectors. In the first step, the gene of interest is cloned into the entry vector and the second step involves sub cloning the gene from entry clone into destination vectors (Fig 2.4.2). The main advantage of the

gateway cloning is that once we have made an entry clone, the gene of interest can be easily sub cloned into wide variety of destination vector.

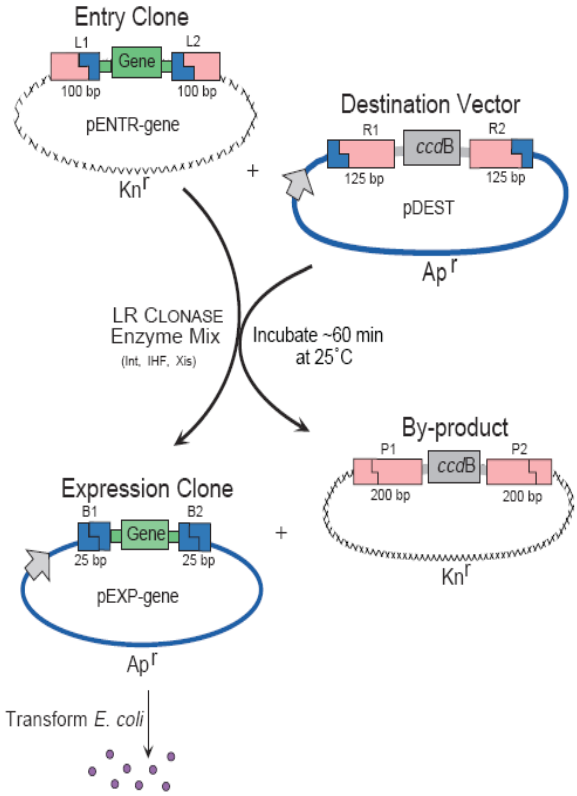


Figure.2.3 Diagrammatic representation of gateway cloning strategy (Adapted from ref. Bioinforx)

The gene of interest was PCR-amplified from a cDNA clone using DNA primers and the PCR product was introduced into the Gateway entry vector pENTR/TEV/d-TOPO by directional TOPO cloning (Invitrogen). The insert was verified by DNA sequencing and introduced into the different Gateway destination vectors by performing a Gateway LR reaction, as per the manufacturer's protocol resulting in the expression plasmid.

2.6.3 Cloning of DNA binding domain of Sox9 and Sox5

The 80 amino acid HMG domain of mSox9 and mSox5 of the full-length mouse protein was PCR-amplified from cDNA clone using gene specific primers (Table 2.4.3). The amplified PCR product was cloned into pENTR™/TEV/D-TOPO® (Invitrogen) vector to generate entry clone [142]. Entry clone was verified by colony PCR and DNA sequencing, indicating the presence of inserts in correct orientation. The mSox9HMG and mSox5HMG

gene in the entry clone was introduced into the Gateway destination vector pETG-20A by performing a Gateway LR reaction, yielding the pETG20A-Sox9HMG and pETG20A-Sox5HMG expression plasmid respectively. Presence of the gene of interest was confirmed by PCR using gene specific primers. Additionally, the presence of TEV cleavage site (ENLYFQG) and absence of mutations were confirmed by DNA sequencing.

2.7 Plasmid transformation

Plasmid DNA (50-100 ng) was added to 50 µl of competent cells which had been thawed on ice in a 1.5ml eppendorf tube. The cell suspension was incubated on ice for 30 min. The cells were incubated at 42° C for 30 seconds as a heat-shock treatment and immediately chilled on ice for 30-120 seconds. 400 µl of SOC media were added and the cells were incubated for 40-60 min. at 37° C in a shaking air incubator at 210 rpm. Half of volume was plated onto LB agar plates containing appropriate antibiotics (100µg/ml ampicillin).

2.8 Colony PCR

A quick way to confirm positive clones of desired genes is by "Colony PCR", designed initially to confirm correct length of vector insert in transformed *E.coli* cells. From LB agar plate with colonies from freshly transformed *E. coli* cells ~5 single colonies were picked with sterile loops which were swirled in 50 µl water in a 1.5 ml Eppendorf tubes. The wet loop was afterwards streaked on an agar plate to preserve the colony for growth later. The cell suspension in the Eppendorf tube was vigorously shaken on a vortex machine for 15 seconds and heated to 90° C for 5 min. A PCR screening reaction mix containing 0.2µM primer, 0.25mM dNTPs, 0.01u/µl Pfu DNA polymerase was prepared and 19µl of it were mixed with 1µl of the sample suspension. The samples were incubated at 95°C for 2min. Amplification was carried out using 25-30 cycles of 30 seconds at 95°C, 30 seconds at a temperature 10°C lower than the lowest melting temperature of the two primers, and

extension at 72 °C for 60 seconds per kilobase of expected PCR product. The reaction product was analyzed by agarose gel electrophoresis.

2.9 Alkaline lysis plasmid preparation (miniprep)

Plasmid preparation on small scale (5ml cultures) was used to quickly obtain small amounts of purified plasmid DNA for cloning and sequencing. The QIAprep miniprep (Qiagen, cat.No. 27104) procedure is based on alkaline lysis of bacterial cells followed by adsorption of DNA onto silica in the presence of high salt. Overnight cultures of *E. coli* (in 5 ml Luria Bertani (LB) media with appropriate antibiotics) grown to a cell density of approximately 3–4 x 109 cells/ml, harboring the desired plasmid were centrifuged (6000 x *g* for 15 min at 4°C, swinging bucket) in 15 ml polypropylene tubes. The pelleted cells were resuspended in 250µl lysis buffer (50 mM Tris·Cl, pH 8.0, 10 mM EDTA, 100µg/ml RNase A) and transferred to 1.5ml Eppendorf tubes. Equal volume of NaOH/SDS solution (200mM NaOH, 1% SDS (w/v)) was added, and the tubes were mixed thoroughly by inverting 4–6 times prior to incubation on ice for 5min. 350µl neutralization buffer N3 (Qiagen, cat.No. 19064) was added to each tube and the tubes were shaken briefly and then incubated on ice for 5 min. After centrifugation of the samples (13,000 rpm, 10 min., RT, microcentrifuge) the supernatant was allowed to enter QIAprep spin column by brief spinning. The column was washed with 500µl of binding buffer, PB(Qiagen, cat.No. 19066) and subsequently with 750µl of wash buffer, PE (Qiagen, cat.No. 19065). The DNA was eluted with water by centrifugation and quantitated by Nanodrop.

2.10.1 Over- expression and solubility

Protein over-expression and solubility test for all clones was performed efficiently using 8 to 14ml of LB media containing appropriate antibiotics. Cultures innoculated with 3-5 colonies of transformed *E.coli* BL21(DE3) cells were grown at 37°C, in an air shaking incubator at 220rpm. The culture was divided up in two vials when it reached an optical

density at 600nm (OD600) of between 0.5 and 0.9. One of the vial was transferred to 30ºC and expression was induced with 0.3 mM IPTG. The OD was measured every 30 or 60 min after induction. After 2-3 hours an aliquot of 1ml culture was taken from both induced and non-induced cultures and centrifuged (8 krpm, 10 min, 4ºC). The cell pellet was resuspended in lysis buffer (50mM tris, 200mM NaCl, pH.8.0) and lysed by sonication. The solubility of the expressed protein was investigated by centrifugation (13 krpm, 10 min., 4ºC, benchtop) of a 500µl sample of the cell lysate and comparing the content of the supernatant to the content of the pellet of cell debris, resuspended in 500µl lysis buffer. Aliquots of the lysate supernatant and the resuspended pellet were mixed with denaturing protein gel loading buffer, boiled for 5 min. and analysed on SDS-PAGE gel.

2.10.2 Large scale expression

Expression of larger quantities of proteins was achieved by increasing the *E.coli* culture volume to 6 or12 liters, divided up in 12 or 24 flasks, each 2 liters flask containing 500ml of medium. Precultures of 100ml media were inoculated with several colonies from a LB Agar plate with transformed *E.coli* cells and grown at 37° C for over night. The 12 liters of medium were then inoculated with 1 % of over night grown preculture. The cultures were grown at 37° C to an optical density at 600nm (OD600) of 0.7 to 0.9 and transferred to 30° C and expression was induced with 0.3mM IPTG. Cells were harvested by centrifugation (8 krpm, 10 min., RT) and pellet was frozen in 250ml polypropylene flasks in liquid nitrogen and stored at –80° C until processed.

2.11.1 Cell lysis by ultrasonication

E.coli cells were lysed by ultrasonication using a Thermo sonicator device with mounted XL head for samples of 100ml volume in glass beakers on ice. Typical procedure of sonication was: 5 pulses for 3 seconds with 3 seconds pauses for 10 min.

2.11.2 Ni Sepharose chromatography

HisTrap HP (GE Biosciences, Cat. No.17-5281-01) 5-ml columns were used for the purification of $(histidine)_6$-tagged proteins. Prepacked Ni-Sepharose High Performance columns, having high binding capacity and low nickel ion leakage were used for target protein purifications. The column was pre equilibrated with binding buffer 50mM Tris, 100mM NaCl, 30mM imidazole pH. 8.0) Presence of 30mM Imidazole in binding buffer facilitated removal of contaminants that can otherwise be co-purified with the tagged target protein. Buffer containing 250mM imidazole was used to elute the bound $(His)_6$ - tagged protein. The column was washed with buffers with high imidazol concentrations before and after use and stored in 20% ethanol.

2.11.3 Ion exchange chromatography

RESOURCE S (GE Biosciences, Cat. No. 17-1180-01) are strong cation exchange columns prepacked with SOURCE™ 15S. The medium is based on rigid, monodisperse 15µm beads made of polystyrene/divinyl benzene. RESOURCE S columns are generally used only with anionic or zwitterionic buffers and cationic detergents are avoided since they bind to the S groups. Protein purification was done with a buffer whose pH was at least 1unit below the pI of the corresponding protein so that they are positively charged and bind to gel matrix. The column was pre-equilbirated with binding buffer (50mM Tris, 100mM NaCl, pH. 8.0) at a flow rate of 1 mL/min. The bound protein was eluted with 50mM Tris, 1.0 M NaCl, pH. 8.0. Column was cleaned with 70% acetonitrile and stored in 20% ethanol.

2.11.4 Size exclusion chromatography

HiLoad 16/60 Superdex 200 pg (GE Biosciences, Cat. No. 17-1069-01) was used for size based separation of protein under native conditions. The column is pre-packed with cross-linked agarose and dextran with average particle size of 34µm. The column was extensively washed with water and one column volume (120ml) of buffer (50mM Tris,

100mM NaCl, pH. 8.0). The buffer flow at loading and elution was typically 1 ml / min. The column was washed with 1.5 CV of water and stored in 20% ethanol.

2.11.5 Endoproteolysis with TEV NIa

The tobacco etch virus (TEV) protease NIa is a site-specific protease recognizing the heptapeptide ENLYFQG, cutting the peptide chain before the glycine. TEV NIa subcloned into pETM-10 as a polyhistidine fusion protein was expressed in BL21 DE3 cells grown in LB medium containing kanamycin and induced with 0.2 mM IPTG. The protease was purified as a $(His)_6$ –tagged protein by HisTrap HP (GE Biosciences, Cat. No.17-5281-01) 5-ml columns and imidazole was removed by desalting (GE Biosciences, Hi-Prep Desalting 26/10). The activity of this recombinantly expressed protease was checked.

2.12 Over-expression and Purification Sox9HMG and Sox5HMG

Sox9HMG and Sox5HMG cloned in expression vector pETG20A was transformed into *E.Coli* cells BL21 (DE3) and plated on ampicillin containing LB agar-plates. A single colony was picked up and grown in ampicillin resistance LB medium. To induce the expression of recombinant protein, the culture was supplemented with IPTG to final concentration of 0.5 mM and purified by Ni-NTA affinity chromatography with imidazole gradient (25-300) in buffer consisting of 50 mM Tris/HCl (pH 8.0) and 100 mM NaCl was used to separate the target proteins from the main contaminating proteins (*Fig 3.2A*). Distinct bands of proteins were identified on an SDS gel. To remove residual imidazole, the protein was applied onto a desalting column. The purified thioredoxin tagged Thx-His6-Sox9HMG was cleaved by TEV protease as mentioned in Materials and Methods and the tag was separated from the protein of interest by ion-exchange chromatography (RESOURCE S, volume 6 ml, GE Healthcare) using a gradient of buffer A, 50 mM Tris/HCl, 100mm NaCl, pH 8.0) and buffer B, 50 mM Tris/HCl, 1 M NaCl, pH 8.0. The protein was purified to homogeneity through size exclusion chromatography HiLoad 16/60 Superdex 75, pg. The

purity of the appropriate protein peak fractions was analysed by SDS gel electrophoresis and pooled fractions were concentrated to 5-10mg/ml.

2.13 Protein analysis methods

2.13.1 SDS polyacrylamide gel electrophoresis (SDS-PAGE)

Proteins were analyzed using a 12% 1:60 bisacrylamide:acrylamide, 0.75 M Tris/Cl pH 8.8, 0.1% SDS separating gel layered with a 5% 1:20 bisacrylamide: acrylamide, 120 mM BisTris/Cl pH 6.8, 0.1% SDS stacking gel. Samples were diluted with an equal volume of TG buffer (125 mM BisTris/Cl pH 6.8, 20% glycerol, 4% SDS, 0.85 M 2-mercapto-ethanol, with bromophenol blue dye as a marker) and boiled for 2 min before loading. Gels were run for 40 - 60 min at 100-120 volts using a running buffer containing 50 mM Tris/Cl, 0.1 M glycine and 0.1% SDS. After SDS PAGE, the gel was washed with hot water for 5 min. and protein bands were visualized by heating the gel for 5 min with "Safe stain" (Biorad) at microwave oven and de-stained by heating to 55° C and shaking water for 5 min For most SDS-PAGE a low molecular weight protein mix (NEB) was loaded parallel to the samples as molecular weight markers.

2.13.2 Mass spectrometry

In-gel digestion coupled with mass spectrometric analysis is a powerful tool for the identification and characterization of proteins. Trypsin is a serine protease that specifically cleaves peptide bonds at the carboxyl side of lysine and arginine residues. The desired coomassie stained protein band was excised and in-gel digestion was carried out as per manufacturer's protocol (Pierce, Cat.No.89871). The sample was subjected to the liquid Chromatographic separation and Electrospray Ionization Mass Spectrometry (LC-ESI-MS).

2.13.3 Circular Dichroism (CD) measurements

All far-UV circular dichroism spectrums were carried out using a Chirascan Circular Dichroism spectrometer (Applied Photophysics Ltd., UK) using a 0.01cm path length quartz

cuvette (Hellma). The data were collected with a spectral bandwidth of 1nm and a time constant of 1 sec. The protein concentration for the experiments was fixed to 50µM in a buffer containing 50mM Tris-HCL, 100mM NaCl, pH 8.0. The spectra were recorded at 25°C from 190 to 240nm using a 2nm bandwidth and three accumulations, after baseline correction. Averaged data was expressed in molar ellipticity.

2.13.4. Dynamic light scattering

Theory of dynamic light scattering

Dynamic Light Scattering (DLS) is used to measure hydrodynamic sizes, polydispersity and aggregation effects of protein samples. These are important parameters for the crystallization of proteins. The technique has proved to be a valuable tool for rapid screening of protein samples crystalizablity. The polydispersity of the sample and the absence of protein aggregates are two crucial parameters necessary for efficient crystallization. Dynamic light scattering, also known as quasi elastic light scattering (Quels) and photon correlation spectroscopy (PCS), measures laser light scattered from dissolved macromolecules or suspended particles. Due to the Brownian motion of molecules and particles in solution fluctuations in the scattering intensity can be observed. A modern DLS instrument uses a compact laser diode and high-end fiber optics, so called single mode fiber optic. Due to the fact that large molecules or particles move slower than small molecules a defined correlation function yields the diffusion coefficient (D) of the molecules by fitting the data. Finally the hydrodynamic radius (Rh) of the particles and molecules can be calculated:

$$D = \frac{kT}{6\pi\eta R_H}$$

T : temperature
η: solvent viscosity

Polydispersity is the relative standard deviation of a sample. Polydispersity of a sample can be cut into three forms: monodisperse, if polydispersity is less than 20%, medium

disperse if in the range of 20% to 30% and it is polydisperse if the polydispersity is more than 30% *(chemeurope.com/articles/e/61632).*

2.13.5 DLS Experiment

Dynamic light scattering was used to determine the aggregation state of a protein or protein/DNA complex sample. Sample of 30µl at concentrations above 2 mg /ml were used. Buffer conditions (salt, pH) and sample preparation (dialysis, centrifugation, filtering) varied among the different samples. A dozen independent measurements were recorded for higher significance of the data.

2.13.6 Thermofluor

Stabilizing proteins in a homogeneous, natively folded and biologically active form is desirable for numerous applications including structural biology. Optimal buffer conditions for individual proteins however, often require tedious empirical characterization. The thermal unfolding of a protein can serve as a read-out for effects of buffer components (pH, ions, ligands, additives) on protein stability. If the melting temperature of a protein exposed to certain chemical environment is shifted upwardly with respect to a reference condition, a structure stabilizing effect is assumed and vice versa. Here we show that thermal unfolding of proteins can be studied by Thermofluor assay using the LightCycler® 480 in a high-throughput setting. An initial 96-well screen was designed and buffer effects on the thermal unfolding of a set of proteins were assessed. This assay employs a fluorescence dye, Sypro-Orange with quenched fluorescence in aqueous solution that increases upon binding to hydrophobic residues. In the thermofluor assay, as temperature increases, the native protein unfolds, exposing its hydrophobic core. Dye binds to this core and as a result the fluorescence intensity increases as a function of temperature. However, after reaching a plateau, the fluorescence intensity decreases due to aggregation of the protein-dye complex. This thermal unfolding is a typical two state model (Fig 2.7.5), with one step sharp transition from folded

to unfolded state where Tm is defined as mid- point temperature of the protein unfolding transition. At Tm both folded and unfolded states are equally populated.

2.13.7 Sample and Screen Preparation

Protein was diluted to 1 mg/ml in 50mM Tris-HCl, 100mM NaCl at pH 8.0. The 96-well screen was prepared as a 2-fold stock and stored at 4°C. Ionic strength, pH buffers (at 100mM in the presence of 100mM NaCl), salts (200mM) and additives were assayed.

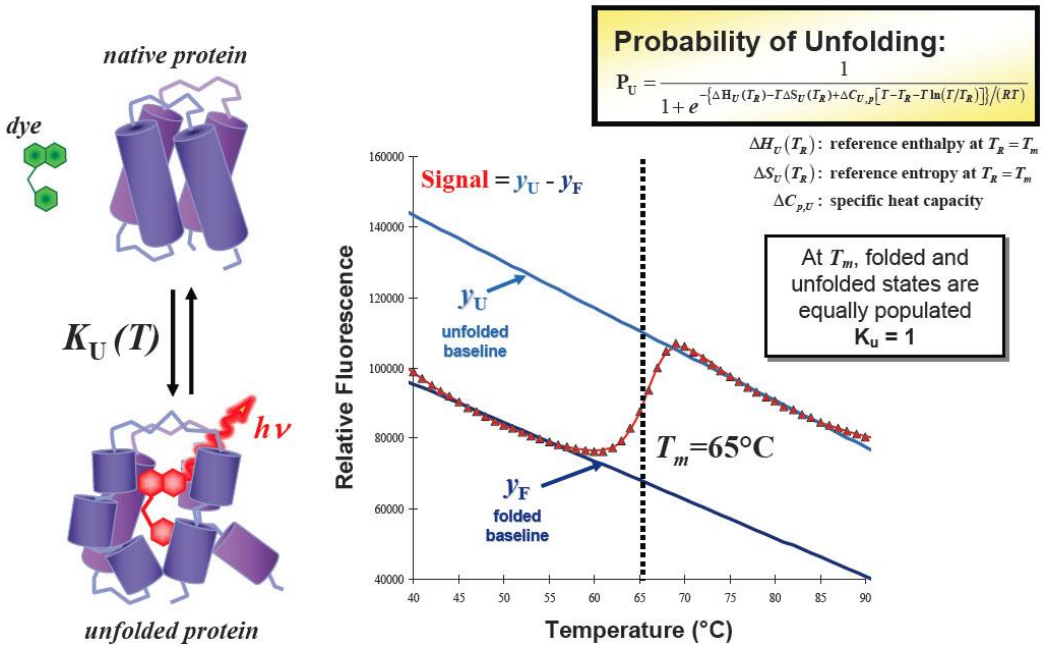


Figure2.4 Principle of Thermofluor assay

2.13.8 Thermal shift assay

In each well of a LightCycler® 480 Multiwell Plate 96 contained 10µl reaction mix obtained after combining 3µl of 100x Sypro Orange (Molecular Probes/Invitrogen S-6651; excitation/emission optima 490/575nm), 5µl of 2X buffer condition and 2µl of 1 mg/ml of the protein. After mixing, plates were sealed with Sealing Foil (Roche) and centrifuged.

Melting curves were recorded on the LightCycler® 480 using 450/568 filters with continuous heating from ambient temperature to 95°C. The heating rate was adjusted to approximately 1°C/min^{-1} by setting the acquisition rate per minute to 6 and the integration

time to 10. Data were exported in text format and analysed using Microsoft Excel and SigmaPlot. Melting points are defined as half maximal value of the unfolding curve corresponding to the curves inflection point as determined by the minimal value of (dFluorescence/dT) plots.

2.14 Protein/DNA complex methods

2.14.1 Annealing DNA duplexes

PAGE purified, labeled/unlabeled DNA elements were obtained from Sigma-Proligo. The DNA strands were annealed in an annealing buffer with a working composition of 20 mM Tris pH 8.0, 50mM KCl and 50 mM $MgCl_2$ in a PCR thermocycler by initially ramping to a temperature of 95ºC for 5 min followed by a slow cooling to 4ºC at the rate of (0.5ºC/sec).

2.14.2 Electrophoretic mobility shift assays

The interaction of proteins with DNA is central to the control of many cellular processes including DNA replication, recombination and repair, transcription, and viral assembly. One technique that is central to studying gene regulation and determining protein:DNA interactions is the electrophoretic mobility shift assay (EMSA).The EMSA technique is based on the observation that protein:DNA complexes migrate more slowly than free DNA molecules when subjected to non-denaturing polyacrylamide or agarose gel electrophoresis. Due to change in the migration of DNA, either shifted or retarded upon protein binding, the assay is also referred to as a gel shift or gel retardation assay. An advantage of studying DNA:protein interactions by an electrophoretic assay is the ability to resolve complexes of different stoichiometry or conformation. Gel shift assays can be used qualitatively to identify sequence-specific DNA-binding proteins and in conjunction with mutagenesis, to identify the important binding sequences within a given gene’s upstream regulatory region. EMSAs can also be utilized quantitatively to measure thermodynamic and

kinetic parameters. Radioisotope or fluorescence labeling of oligonucleotides lowers the detection limits to femtomolar amounts of macromolecular species on the gel, with appropriate equipment for producing and analyzing an image of the gel.

2.14.3 Parameters for EMSA

The complexes separate during the electrophoresis and depending on the time scale of disruption only "smears" or even only subspecies of the complex can be traced. The experimental parameters for EMSA are critical to the stability of complexes in the electric field and have therefore to be examined carefully before setting up quantitative experiments. Gel composition, temperature, pre-run, time of run and voltage can all affect how sharp and well separated the bands corresponding to different species will be, while pre-incubation time and the concentration range of protein and/or DNA have to be determined in order to obtain accurate and reproducible results. Once these conditions have been optimized, EMSA can be used to determine binding affinity constants, compare the affinities in different buffers or for different protein fragments or DNA sequences, study the stochiometry of the binding reactions, or to simply monitor the activity of proteins binding to DNA.

2.14.4 Titrations of protein versus DNA at nM concentrations

EMSA was the method of choice for activity tests of all purified proteins and oligonucleotides at nanomolar concentrations and for preparation for an extent analysis. The question to be answered by the titrations at nM concentration was therefore, whether the components would bind with high affinity corresponding to in vivo conditions assuring the biological relevance of the investigated complexes.

EMSAs were carried out using DNA probes modified with 5′ cy5 labels (Sigma Proligo). Equimolar amounts of complementary strands were mixed and heated to 95°C followed by gradual cooling to ambient temperature over at least 5 h to anneal the probes. For

binding studies, double-stranded DNA probes at 1nM were mixed with varying concentrations of analyte protein in a buffer containing 20mM Tris–HCl, pH 8.0, 0.1mg ml−1 bovine serum albumin, 50μM ZnCl2, 100mM KCl, 10% glycerol, 0.1% NP-40 and 2 mM β-mercaptoethanol. Binding experiments were carried out by incubating protein with 1nM probe for 1 h at 4°C, in the dark. The 10-μL reaction volume of bound and unbound probes were subsequently separated at 4°C on a pre-run 12% 1× TG polyacrylamide gels and electrophoresed in 1× TG (25mMTris, pH 8.3/192mM glycine) at 200V for 30 min at 4°C. The fluorescence was detected using a Typhoon 9140

The free DNA and bound DNA bands were quantified using the ImageQuant TL software (GE Healthcare), and the bound fraction was plotted against the protein concentration. The dissociation equilibrium constant was determined by non-linear curve fitting in R (http://www.r-project.org/) [143].The concentration of protein was corrected for the active fraction when calculating Kd and cooperativity factors. At least three independent experiments per protein constructs were performed. A curve was fitted, assuming one-site saturation binding using R-Program to estimate the apparent dissociation constant.

2.14.5. Fluorescence anisotropy

Fluorescence anisotropy measurements

The fluorescence anisotropy assay is a spectroscopic technique that measures the tumbling rate of a sample containing a fluorophore [144]. A fluorescently labeled (Fluorescein) DNA element that has been reported to bind to Sox5 was chosen as the DNA substrate [145]. When polarized light excites the (FAM)-DNA element (5’ Fluorescein labeled *cis*-regulatory DNA element), the relatively small (FAM)-DNA element which undergoes rotational diffusion causes depolarization of the emitted light resulting in a low anisotropy measurement. When SoxHMG binds to (FAM)-DNA, the larger size of the protein-DNA complex causes a slower rotation, resulting in a relatively higher polarization/anisotropy of the emitted light (Figure 2.1). This anisotropy assay strategy was

chosen as it can easily be scaled up to a HTS. The fluorescence anisotropy measurements from the microplates are read on a Spectramax M5 microplate reader (Molecular Devices) with excitation at 485 nm, emission at 525 nm and a cut-off filter of 515 nm.

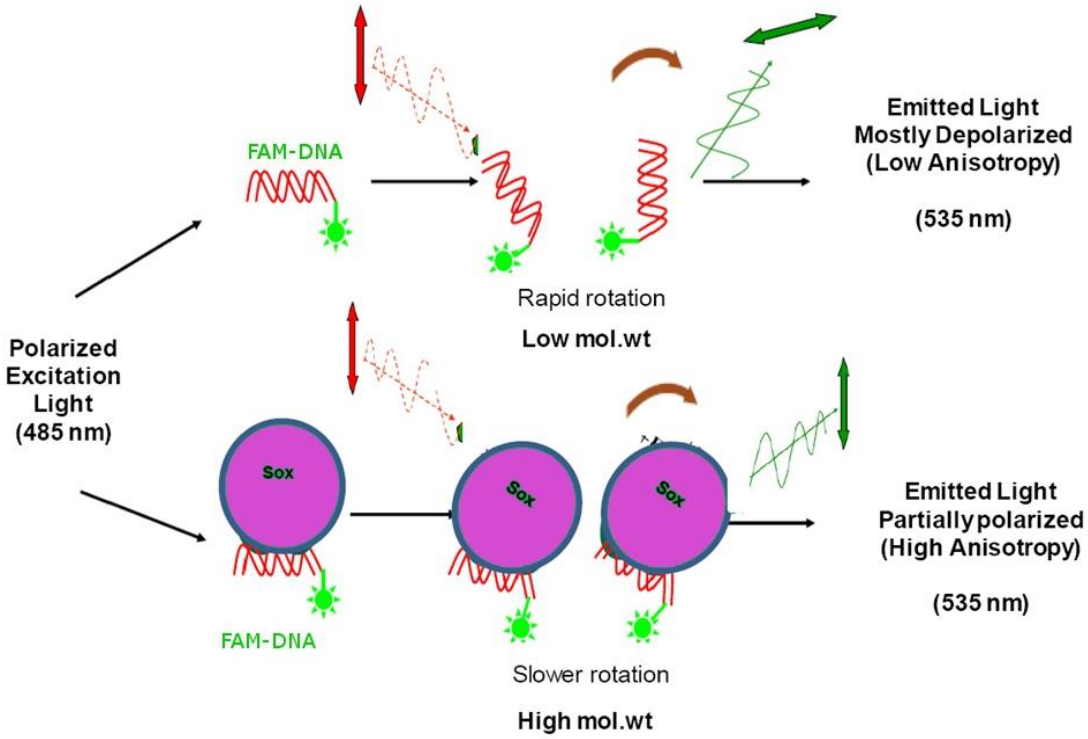


Figure 2.5 The principle of the fluorescence anisotropy. The formed Protein-DNA complex is larger and tumbles more slowly than the unbound nucleotide. The binding affinity for the fluorescein-labelled sequence was determined using fluorescence anisotropy titrations.

2.15 Crystallization and crystal handling

2.15.1 Vapour diffusion crystallization

Protein/DNA complex solutions concentrated to about 10 mg/ml were centrifuged (13 krpm, 10 min, RT/4° C, benchtop) immediately before setting up crystals. Crystallization trials for room temperature incubation were set up at room temperature, crystallization trials for 4°C incubation were set up at room temperature kept at 4°C. All solutions used for crystallization were prepared from chemicals of the highest quality (i.e. Hampton, Sigma and Fluka) using Milli-Q water and filtered through 0.2µm filters (Sartorius). Vapor-diffusion was used with hanging drops or sitting drops. Typically equal volumes of protein and reservoir buffer as precipitant were mixed on the silanized cover slide by pipetting half the

volume up and down the pipette tip several times and allowed to equilibrate with a reservoir solution that had been prepared in the wells. The wells were sealed and never opened but for harvesting of crystals.

2.15.2 Mounting crystals in loops

Fiber loops were made from a single fiber of an ordinary packing cord (diameter of 10μm) glued (with cyanolite fast glue) into a 0.2mm glass capillary or at the tip of a 1.5cm long piece of copper or tin wire. The capillary was mounted on a brass or plastic sleeve convenient for fixing on a goniometer head. The crystals were picked up with these loops from the storage or soaking buffer and transferred to liquid propane or nitrogen within 1-2 sec. Crystals mounted in loops were only used for characterization or data collection at cryo-temperatures.

2.15.3 Storing and mounting cryo-cooled crystals

Cryo-cooled crystals were stored and transported in liquid nitrogen containers. Crystals mounted in loops were transferred into the diffractometer cryo-cooling nitrogen gas stream, Flipping the crystals between liquid propane or nitrogen and the cold nitrogen gas stream (< -170° C) within less than a tenth of a second.

2.16 MultiCoil Program

The MultiCoil program predicts the location of coiled-coil regions in amino acid sequences and classifies the predictions as dimeric or trimeric. The method is based on the Paircoil algorithm which predicts the parallel coiled coil fold from amino acid sequence using pairwise residue probabilitis. The URL of the programme is, http://groups.csail.mit.edu/cb/multicoil/cgi-bin/multicoil.cgi

PROTEIN – DNA INTERACTION

OF SOX TRANSCRIPTION FACTORS

CHAPTER III

DNA Binding HMG domain of Sox9

Results and Discussion

3.1 Cloning of DNA binding domain of Sox9

mSox9HMG protein spanning residues 103-183 was cloned through PCR amplification *(Fig 3.1)* from cDNA clone (IMAGE:5354229) and purified (Fig 3.2) as described in Materials and Methods 2.12

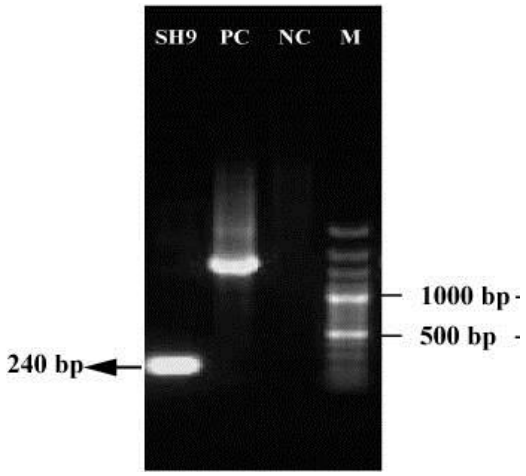


Figure 3.1: PCR Amplification of SOX9 HMG domain. SH9: mSox9HMG; PC: Positive Control; NC:Negative Control; M:DNA ladder. 5 µl of DNA sample was loaded on a 1.0% agarose gel.

The purified DNA binding HMG domain of SOX9 gene encodes 80 amino acid protein corresponding to a calculated molecular mass of 9.71 kDa and was validated by size exclusion chromatography. Physicochemical properties of Sox9HMG were calculated from the primary amino acid sequence. The isoelectric point (pI) of Sox9HMG, the pH value at which the molecule carries no electrical charge was calculated to be 10.43 and the extinction coefficient as 15470 M^{-1} cm^{-1}, at 280 nm measured in water. The extinction coefficient estimation was a convenient preliminary step for following protein purification with a spectrophotometer.

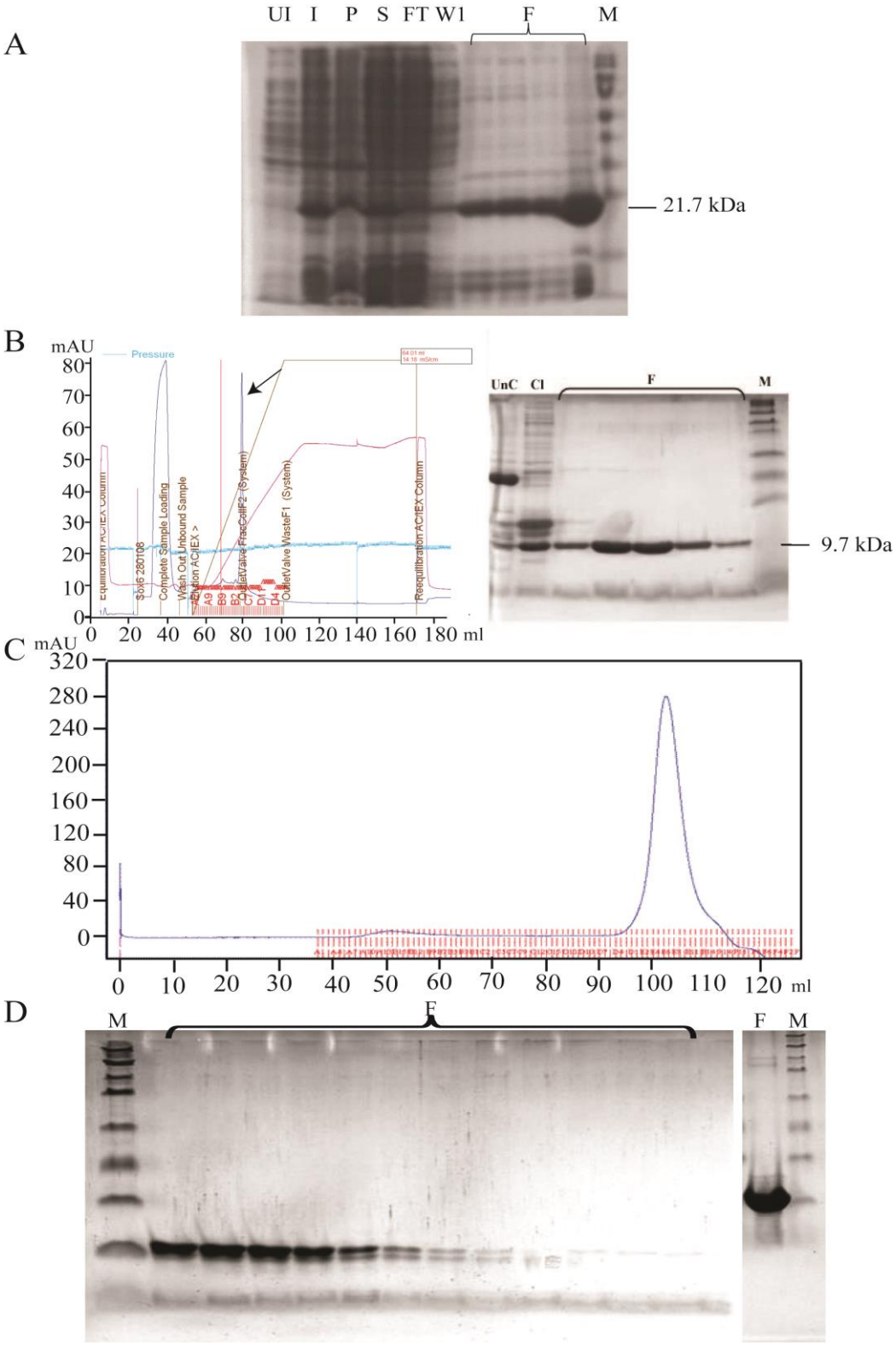


Figure 3.2 Expression and purification of Sox9HMG domain. Panel A. Ni-NTA purification analysed on 12%SDS-PAGE; Panel B. Ion exchange profile showing purity of the expressed protein(Arrow:Eluted Protein band); Panel C. Size exclusion chromatography of the purified protein and Panel D. SDS-PAGE analysis of eluted fractions, (F) pooled and concentrated **.** UI: Un-Induced; I:Induced; P:Pellet; S:Supernatant; W1:Wash 1; F:Fractions of Elution; UnC:Uncleaved; Cl:Cleaved; M:Molecular weight markers (kDa).

3.2 Secondary structure analysis of Sox9 HMG domain

CD spectroscopy in the "far-UV" spectral region from 190-240 nm wavelength, enables determination of protein or peptide secondary structure, as the peptide bonds give rise to signals when located in a regular, folded environment. α-helix, β-sheet, and random coil structures each correspond to a characteristic shape and magnitude in the CD spectrum. Like all spectroscopic techniques, the CD signal reflects an average of the bulk molecular population.

Secondary structure prediction from the primary amino acid sequence was performed using the PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred/) (MCGUFFIN *et al.*, 2000) (*Fig. 3.4*). In order to verify that the purified recombinant Sox9HMG domain was well folded and retained its native structure, circular dichroism (CD) analysis was performed as mentioned in Materials and Methods. The CD spectra of purified HMG domain showed a single positive maximum at 195nm and two negative minima at 208 and 222 nm (*Fig 3.3*), typical of helical structure.

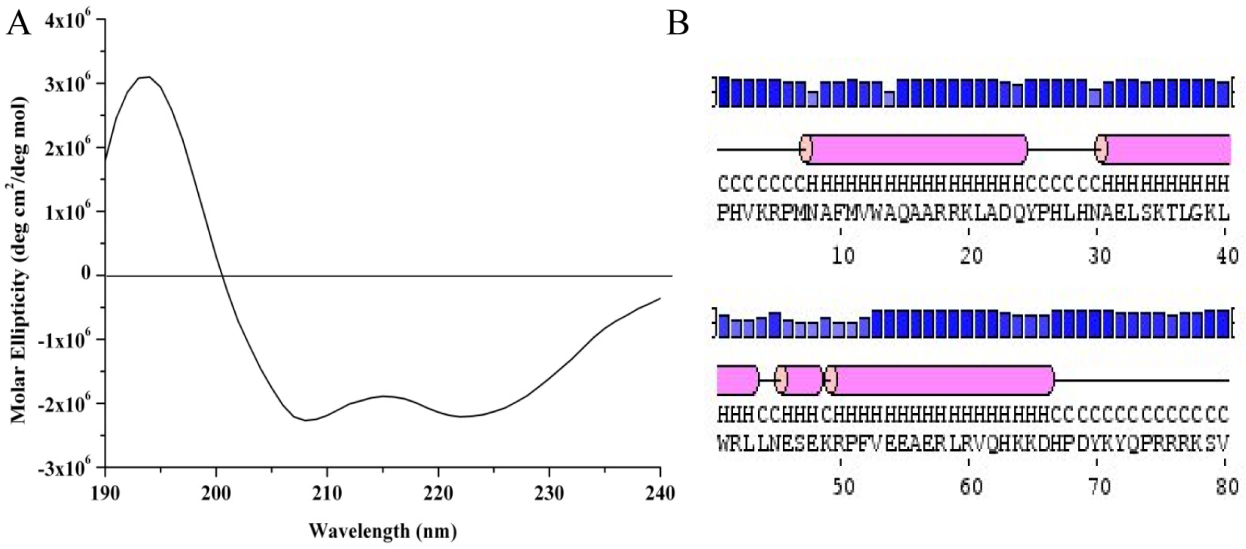


Figure3.3 Secondary structure analysis of Sox9HMG. Panel A Circular dichroism spectroscopy analysis of the purified Sox9HMG domain; Panel B Secondary structure prediction using PSIPRED software [146].

3.3 Mass spectrometry analysis of Tryptic digested Sox9HMG domain

To confirm the authenticity of purified Sox9HMG protein, tryptic digested sample was subjected to Electrospray Ionization – Mass Spectrometry (ESI-MS). The enzyme trypsin cleaves the protein backbone next to a Lysine (K) or Arginine (R) residue except when trailed by a Proline (P) residue, leaving behind peptide chains having none or single Lysine or Arginine residue, a property exploited to study the protein primary structure and identification by analysing the resultant peptides using mass spectrometry (MS). Sox9HMG domain was digested by trypsin leaving the protein as 11 peptides as shown in *Table 3.1*. MS data of the in-gel tryptic digested peptides analysed in MASCOT-DEMAN database.

Start	End	Sequence
107	120	K.RPMNAFMVWAQAAR.R
122	137	R.KLADQYPHLHNAELSK.T
123	137	K.LADQYPHLHNAELSK.T
123	141	K.LADQYPHLHNAELSKTLGK.L
142	151	K.LWRLLNESEK.R
145	151	R.LLNESEK.R
145	160	R.LLNESEKRPFVEEAER.L
152	160	K.RPFVEEAER.L
167	173	K.KDHPDYK.Y
168	173	K.DHPDYK.Y
168	177	K.DHPDYKYQPR.R

Table 3.1 Tryptic digested peptides of Sox9HMG domain analysed in MASCOT-DEMAN database.

3.4 DNA binding affinity of purified Sox9HMG protein

The DNA binding affinity of the purified Sox9HMG protein was assessed using electrophoretic mobility shift assays (EMSAs) with short Sox consensus sequence as described in Materials and Methods. EMSA experiments were performed with purified short oligos of COL2A1 (5'AGCCCCATTCATGAGA3') and COL4A2 (5'CCTTCTTGTTACGG GG3') enhancer elements with Sox binding sites. The purified Sox9HMG protein was incubated with synthesized Cy5 labeled double stranded DNA. Upon incubation with increasing concentration of purified Sox9HMG protein (0 -500 nM) two bands, corresponding to a fast migrating non-shifted free DNA and retarded (shifted) protein-DNA complex band (*Fig.3.4*) were obtained consistently. The concentration of DNA was maintained as 1nM throughout the assay. The apparent dissociation constant (Kd) of Sox9HMG with COL2a1 was determined as ~25nM and with Col4A2 as <50 nM from the protein concentration at which half of the DNA is bound.

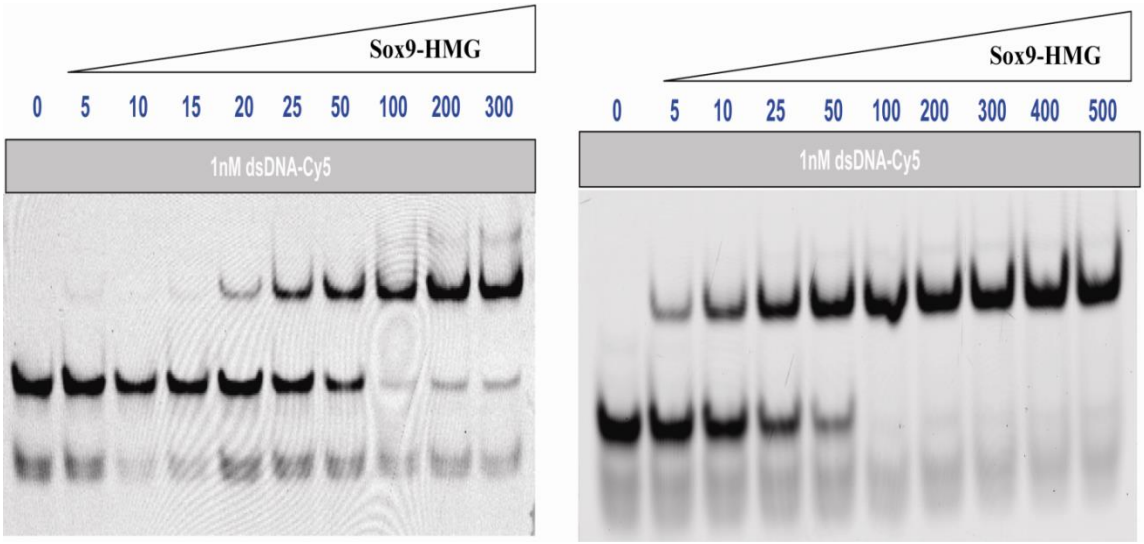


Figure 3.4 Electrophoretic mobility shift assay of Sox9HMG bound to (a) Col4A2 (5'CCTTCTTGTTACGGGG 3') and (b) Col2A1 (5'AGCCCCATTCATGAGA3') enhancer element. Reaction with1nM of dsDNA-Cy5 and increasing concentration of protein

3.4 Discussion

Sox HMG domains encode three helices that fold into characteristic L-shaped structure with a defined DNA binding face. Circular dichroism (CD) spectroscopy revealed the cleaved and purified Sox9HMG to be well folded retaining the compact native alpha helical structure, suitable for further characterization (*Fig 3.3*). The primary sequence authenticity of the purified Sox9HMG was validated using in-gel tryptic digested peptides analysed in MASCOT-DEMAN database (*Table 3.1).*

Mouse COL2A1 gene is an established Sox regulated target gene [111] that transcribes type II collagen, a major consitituent of cartilage. During chondrogenesis, the expression level of Sox9 is comparable to Col2a1 and anomalous regulation of Col2a1 results in campomelic dysplasia, marked by skeletal abnormalities [113]. Sox9 binds COL2A1 and COL4A2 enhancer elements and mutations in the Sox9 binding sequences abrogate chondrogenesis in transgenic mice [111]. The DNA binding affinity of the purified transcription factor, Sox9HMG was analysed with EMSA (*Fig.3.4).* The determined dissociation constant (Kd) of Sox9HMG with COL2a1 as ~25nM and with Col4A2 as <50 nM, indicates the recombinant purified Sox9HMG domain to be in the helical form and possess high affinity for the DNA binding site, a prerequisite for crystallographic structural characterization *vide infra* (*Section 4.1*).

CHAPTER IV

Identification and validation of Novel Sox9 regulatory Motif

The HMG proteins are significant as they play definite and important roles in various developmental processes, nonetheless the DNA-binding sites they recognize are highly degenerate. The HMG proteins trans-activating functions and specificities are highly dependent on the orientation and spacing of their binding sites and the binding sites of other cofactors. Although Sox9 binding motifs have been reported based on computational approach and in-vitro studies, they may not model functional SOX-binding sites precisely and accurately. Therefore, to better model precise *in vivo* functional binding sites of Sox9, immunoprecipitation coupled with ultra-high-throughput DNA sequencing (ChIP-Seq) data, generated from the lab of Prof. Thomas Lufkin, Genome Institute of Singapore was utilized. The ChIP-Seq assay was performed with limb and tail tissues of germline transmitting (GLT) chimeras from *Sox9*$^{+/-(EGFP)}$ embryonic stem cells of mouse using Sox9 antibodies (R & D Systems). Reliable Sox9 binding regions and the respective specific binding peaks from the sequence of immunoprecipitated DNA fragments of ChIP assay, were defined by the method described in Chen et al., 2008 [143].

4.1 Motif Analysis

Repeat-masked sequence from 100bp surrounding the top 200 Sox9 peaks were used as input for MD module program [147] to scan for motifs in the given ChIP-Seq condition. For each potential Sox9 motif, the bound regions were scanned back with e-value cut-off of 0.001 as described earlier [148]. The bound regions were also scanned for the presence of a cluster of Sox9 and other Sox motifs, up to 20 bp distances separating the 2 motifs. The search yielded identification of one Sox9 binding motif, **5'AGAACAAAG 3'** (*Fig. 4.1A*), the

most abundant sequence (35%) from the tissue ChIP-Seq library, consensus with previously known Sox9 motif from Jaspar database **5'**AACAAT **3'** (*Fig 4.1 B*).

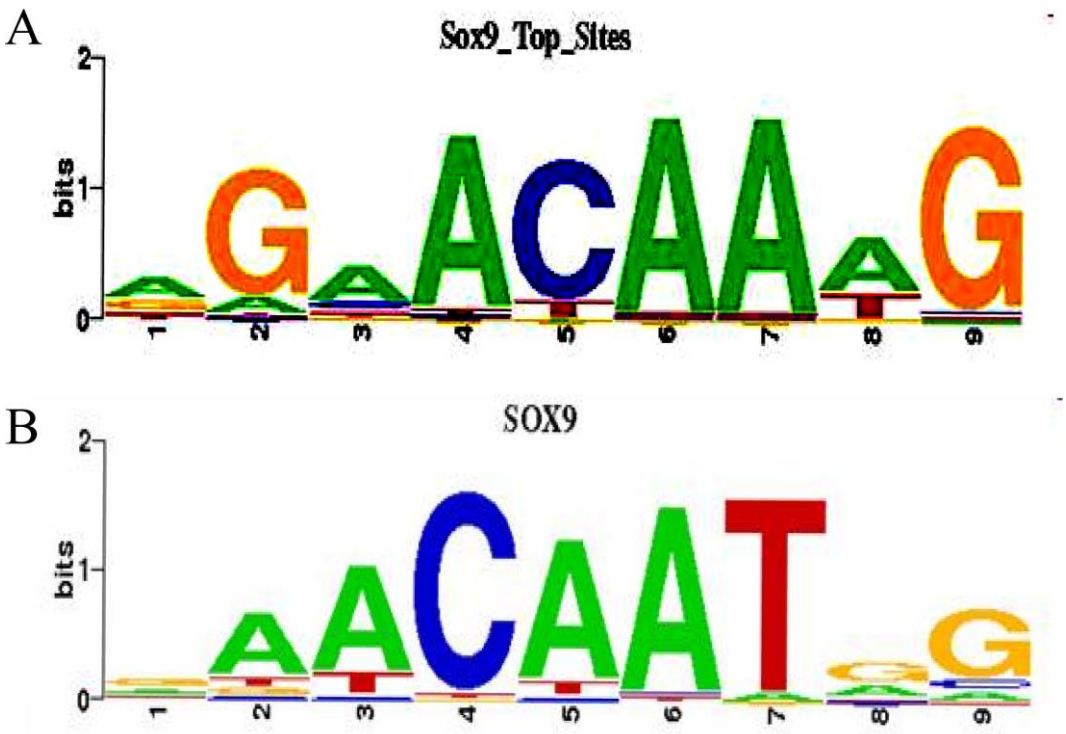


Figure 4.1 Sox9 binding motif analysis. Panel A) ChIP-Seq identified Sox9 motif from the data, AGAACAAAG and Panel B) Reported Sox9 motif from Jaspar database, AACAAT

4.2 Validation of novel Sox9 regulatory motifs

Besides, two other putative Sox9 motifs were identified as **5'ATGAATGGA3'** and **5'CAATGGTC3'** each contributing 16.8% and 13.1% of Sox9 binding sites respectively. Genes harboring these motifs in their promoter sites were taken as representative genes for further analysis: Postn (intron) (5'ATTTATGAACGCTGGGA3') and 4631426J05Rik (5'AGGAATGAATGGATAGA3') as representatives for the first motif; Sox5 (intragenic) (5'CGTTATGAATGGGATCG3'), Myom1 (intragenic) (5'GACACAATGGATCA TA3') for the second motif and FoxP2 promoter for the consensus sequence 5'CAGG**AGAACA AAG**CCTG 3' determined for these motifs. These novel Sox9 motifs were validated using

EMSA of Cy5 labeled double stranded probes harboring the canonical motif and newly identified motifs from representative genes (*Fig 4.2 A,B,C*).

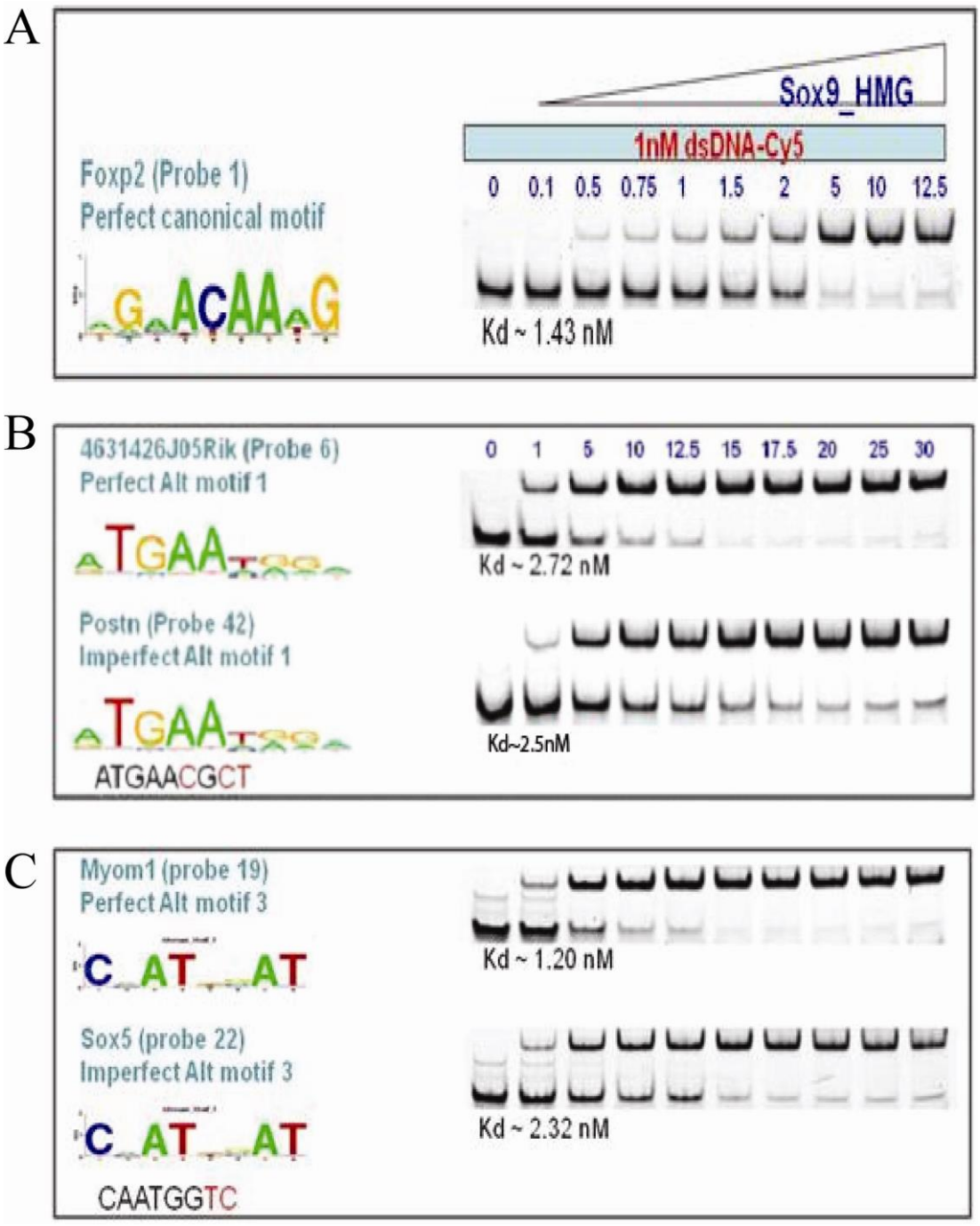


Figure 4.2. ChIP-Seq identified Sox9 binding motifs and EMSA of Sox9-HMG domain with canonical motif of FoxP2 (A), new alternative motif from Rik, Postn (B) and Myom1 and Sox5 (C).

The electrophoretic mobility shift assay (EMSA) is a highly sensitive tool for studying protein-DNA interactions used to determine the binding parameters and relative affinities for one or more DNA sequences or for comparative affinities of different proteins for single site. EMSA is also useful for analyzing ‘shifts’ and 'supershift' of large protein-DNA complexes. All four representatives showed high binding affinity (Kd) in nM range comparable to the canonical motif (*Fig 4.2 A,B,C)*

	Target Gene and Mutant Probe	Cy5 probe seq (5'-3') Forward	Cy5 probe seq (5'-3') Reverse	Kd (nM)
1	**Foxp2**	CAGG**AGAACAAAG**CCTG	CAGG**CTTTGTTCT**CCTG	**1.4**
2	mutant1-1	CAGG**AGACACAAG**CCTG	CAGG**CTTGTGTCT**CCTG	
3	mutant1-2	CAGG**AGCCACCAG**CCTG	CAGG**CTGGTGGCT**CCTG	
4	mutant1-3	CAGG**CTCCACCCT**CCTG	CAGG**AGGGTGGAG**CCTG	
5	mutant1-4	CAGG**AGACACCAG**CCTG	CAGG**CTGGTGTCT**CCTG	
6	**4631426J05Rik**	AGGA**ATGAATGGA**TAGA	TCTA**TCCATTCAT**TCCT	**2.7**
7	mutant2-1	AGGA**ATGCCGGGA**TAGA	TCTA**TCCCGGCAT**TCCT	
8	mutant2-2	AGGA**ATTCCGTGA**TAGA	TCTA**TCACGGAAT**TCCT	
9	mutant2-3	AGGA**CGTCCGTTC**TAGA	TCTA**GAACGGACG**TCCT	
10	**Postn intron, single motif1**	ATTT**ATGAACGCT**GGGA	TCCC**AGCGTTCAT**AAAT	
11	**Myom1 intragenic**	GACA**CAATGGAT**CATA	TATG**ATCCATTG**TGTC	**1.2**
12	**Mutant 5-1**	GACACACGTTATCATA	TATGATAACGTGTGTC	
13	**Mutant 5-2**	GACAACCGTTCGCATA	TATGCGAACGGTTGTC	
14	Sox5 intragenic	**TTTT**CAATGGTC**CATA**	**TATG**GACCATTG**AAAA**	**2.3**

Table 4.1 ChIP-Seq **identified representative gene motifs, their mutant motif sequences and the corresponding Cy5 probes used in EMSA analysis of Sox9HMG binding specificity.**

To further validate the sequence specificity, these motifs were individually mutated by 3–5 bp (CA to GG) and subjected to electromobility gel shift assays (EMSA). Totally four mutants of FoxP2 **(***Fig. 4.3*), three mutants of Rik, and two mutants of Myom1 binding motifs were synthesized as shown in (*Table 4.1*). The mutations resulted in drastic loss of Sox9HMG binding affinity for all the probes tested (*Table 4.1)*

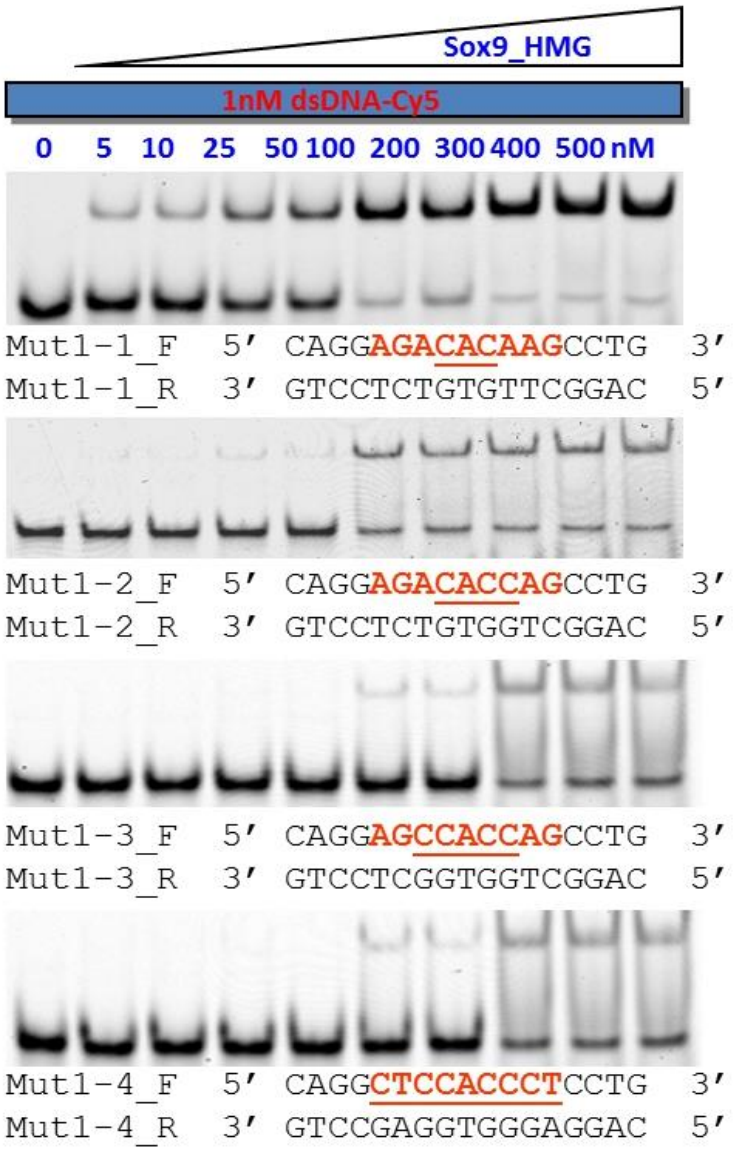


Figure. 4.3 EMSA analysis of Sox9HMG binding affinity with mutated ChIP-Seq identified canonical motifs (Foxp2 5'CAGGAGAACAAAGCCTG3').

4.3 Canonical motif related endogenous binding sequences

The sequence preference of Sox9 for consensus sequence Foxp2, was further tested with the identification of five related endogenous binding sequences (secondary motifs) through ChIP-Seq analysis. The DNA binding ability of Sox9 to the identified related endogenous binding sequences were analysed by EMSA (*Fig 4.4*) as mentioned earlier.

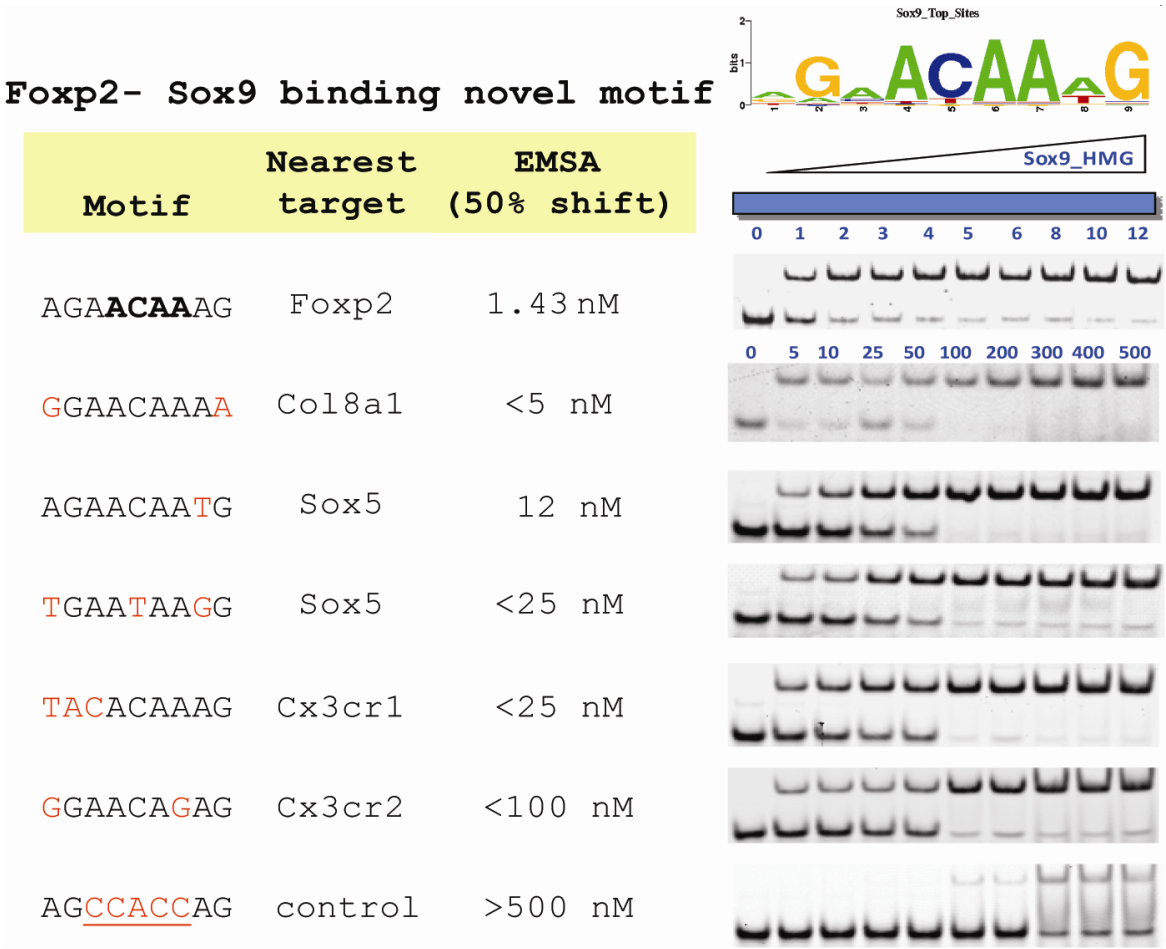


Figure. 4.4 EMSA analysis of Sox9HMG binding affinity to ChIP-Seq identified canonical motif (Foxp2 5'CAGGAGAACAAAGCCTG3') related endogenous sequences.

Col8a1 with intact canonical motif and purine-purine (AG/GA) substitution at the terminal flanking ends had similar binding affinity (<5nM) to that of canonical Foxp2binding site (1.43nM) (*Fig 4.2*). However, Sox5 sequence with a purine to pyrimidine substitution at 8th position showed considerably reduced binding affinity (12nM). Interestingly, Sox9 showed drastic reduction in binding affinity to Cx3cr2, with purine substitution at the 7th

position (<100nM), indicating strong positional preference in the DNA binding sequence. Two other binding sequences Sox5 and Cx3cr1 with alterations in either the core canonical sequence or the flanking ends respectively, also had reduced binding affinity to Sox9HMG domain (<25nM). The control oligo with no match to the canonical motif showed low binding affinity of >500nM.

4.4 Discussion

Despite many advances made so far in the study of DNA binding specificity of transcription factors, understanding the key-principles of transcriptional regulation is still an outstanding puzzle as specific recognition of *cis*-regulatory DNA elements by transcription factors (TF) is determined by a multitude of factors like *in vivo* transcription-factor concentration, the relative affinity of the TF towards specific and non-specific binding sites, co-operativity with other protein-complexes, accessibility of nucleosomal DNA and other aspects like the presence/absence of epigenetic marks such as DNA methylation. Though the DNA binding sequence of all Sox proteins are highly similar, discrete Sox proteins regulate specific target genes. Col2a1 was the first gene proposed as a direct target of regulation of Sox trio comprising Sox9, Sox5 and Sox6. Sox9 activates a 48-bp cartilage-specific enhancer located in the first intron [111, 113] and L-Sox5/Sox6 enhance the activity of Sox9 [82]. However, the enhancer of COL2a1 does not have defined consensus Sox site instead has four sites with as few as five or six Sox9 consensus nucleotides. Sox9 binds the most distal pair of the sites, while L-Sox5/ Sox6 contacts each of the four in vitro [149].

Even though so far several Sox9 binding motifs have been reported based on computational and *in-vitro* studies, *in-vivo* results reflect precise and accurate transcription factor binding site. In this regard, *in vivo* data from immunoprecipitation coupled ultra-high-

throughput DNA sequencing (ChIP-Seq) of chondrogenic limb and tail tissue from living mouse embryos were utilized to identify and validate novel regulatory motifs in cartilage-specific genes. In the current study, two novel regulatory motifs **5'ATGAATGGA3'** an**d 5'CAATGGTC3'** were identified and corresponding promoter sequences of representative genes; Postn and Rik1 (motif1) and Myom1 and Sox 5 (motif2) were biochemically validated (*Fig 4.2*). The representative genes of these newly identified motifs are known to play important roles in skeletal development and development of other vital organs [152].

The gene POSTN (periostin protein) is an osteoblast specific factor that shows expression in preosteoblasts and several chondrocytes and it is required for adhesion of these cells to the extracellular matrix. Periostin activated in alphaV, beta1, beta3 and beta5 integrins located in the cardiomyocyte cell membrane [150] is involved in cell survival and angiogenesis and therefore has become a promising marker for tumor progression in several types of human cancers [151]. 4631426J05Rik, is reported as expressed in all sorts of important organs like bladder, bone, bone marrow, brain, eye, heart, kidney, liver, lung, testis etc. Additionally, a novel Sox-binding motif in the muscle protein Myom1 was identified. Myomesin-1, a skeletal muscle protein encoded by the *MYOM1* gene along with its partner proteins, bridges the major structure of sarcomeres, the M bands and Z discs. Myomesin1 (Myom1) and myomesin2 (Myom2) signifies the principal structural constituent of the M-line. Myom1 shows tissue- and developmental-stage-specific alternative splicing. Sox9 has been shown to downstream regulate Sox5 in the chondrogenesis pathway [118] (*Section 1.12*). Strikingly, the analysis also identified the Sox9 binding motif in Sox5 promoter region, substantiating the downstream regulation of Sox9 and the significance of the Sox5, Sox6, and Sox9 proteins in chondrogenesis [152].

Interestingly, the novel motifs 1 and 2 showed very high binding affinity with disassociation constant (Kd) ranging from 1.2nM - 2.7nM, comparable to the Sox9 canonical motif (1.4nM) (*Fig 4.3*) and contrasting to the known Sox9 regulatory motif, Col2A1 (*Fig 3.5*) with Kd~25nM. Furthermore, the functional validation of both the novel motifs as Sox9-responsive enhancer elements was confirmed employing Luciferase assay (*Collabration, Sook peng et al., unpublished data*). Mutations in any of these gene's promoter Sox-binding sites drasticaly decreased DNA binding affinity in gel retardation assays (*Fig 4.3 and Table 4.1*) as well promoter activity in luciferase reporter assays, further confirming the authenticity of the endogeneous novel genes.

Recently, an extensive protein binding microaray (PBM) based DNA binding study by Bulyk *et al* have revealed Sox proteins to bind alternate (secondary) binding motifs [153]. A primary motif is defined as the highest affinity consensus motif of a transcription factor whereas a secondary motif refersto a population of lower affinity binding sites that considerably differs by more than one base-pair from the primary consensus [94]. Sox proteins were hypothesized to bind such secondary motifs by "positional interdependence" that spanned more than dinucleotides [153]. Most recently, structural evidence for the positional interdependent secondary motif recognition model came from our group, based on the crystal structures of DNA bound to Sox4 and Sox17 HMG domains. A comparison of the structures reveal subtle conformational rearrangements of two interface amino acids at the DNA binding interface to accommodate primary and secondary motifs. Such structural change is presumed to direct altered dinucleotide preferences of Sox4 [94]. Evidence for existence of such secondary binding motifs in other Sox protein subgroups is lacking.

To this end, the identification and subsequent validation of the alternative motifs as direct Sox9 binding motifs evidently suggests presence of secondary binding motifs for the Sox9 protein specifically and for groupE Sox in general (*Fig 4.2*). The novel motifs showed significant albeit slightly lower enhancer activity compared to the canonical motif indicating that Sox dependent enhancer activation could probably be Sox subclass specific and receruitment of specific DNA binding partner recruitement to be dependent on the DNA binding affinity. Furthermore, differences in the DNA binding affinities of Sox9 to canonical motif related endogenous sequences (*Fig 4.4)*, differing at specific flanking or core positions indicates presumable positional interdependence in recognizing secondary binding motifs. However, three-dimensional data of DNA bound Sox9HMG would provide better insight into operation of such DNA recognition model in Sox proteins (*vide infra Section 6.1*). Taken together, the validation of novel Sox9 motifs in the current study is the preliminary biochemical evidence for existence of additional regulatory motifs and will lead to the dissection of new genuine Sox target genes and their respective regulatory mechanisms.

CHAPTER V

DNA Binding HMG domain of Sox5

5.1 Cloning of DNA binding domain of Sox5

The 80 amino acid residues HMG domain of mSox5 of spanning 505-585 was cloned through PCR-amplification (Fig 5.1) from cDNA clone (IMAGE:40047865), purified (Fig 5.2) as described in Materials and Methods 2.12 and concentrated to 5-10mg/ml.

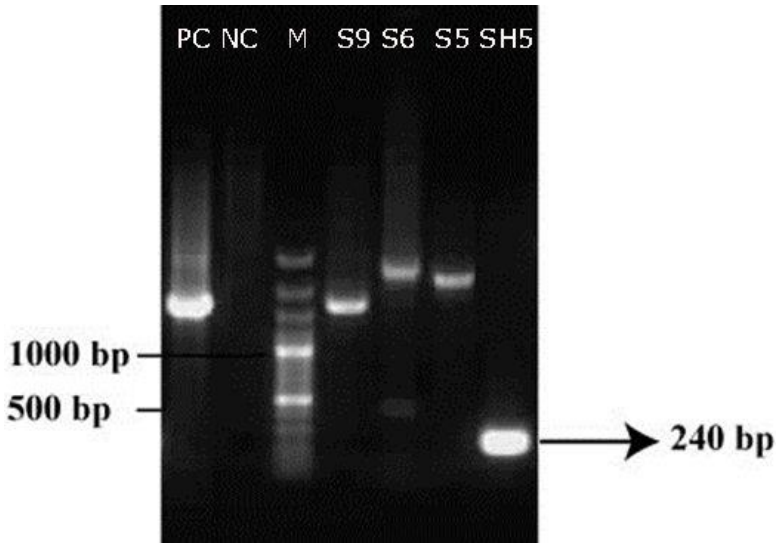


Figure 5.1 Cloning of Sox5HMG domain in pETG20A vector. SH5:mSox5HMG; PC:Positive Control; NC:Negative Control; M:Marker

The purified DNA binding HMG domain of Sox5 encodes 80 amino acid and a calculated molecular mass of 9.7 kDa, confirmed by size exclusion chromatography. Its physicochemical properties were calculated from the protein sequence. The isoelectric point (pI) is the pH value at which the molecule carries no electrical charge or the negative and positive charges are equal. The pI of Sox5HMG was calculated as 9.99. The extinction coefficient of the protein is 18450 M-1 cm-1, at 280 nm measured in water.

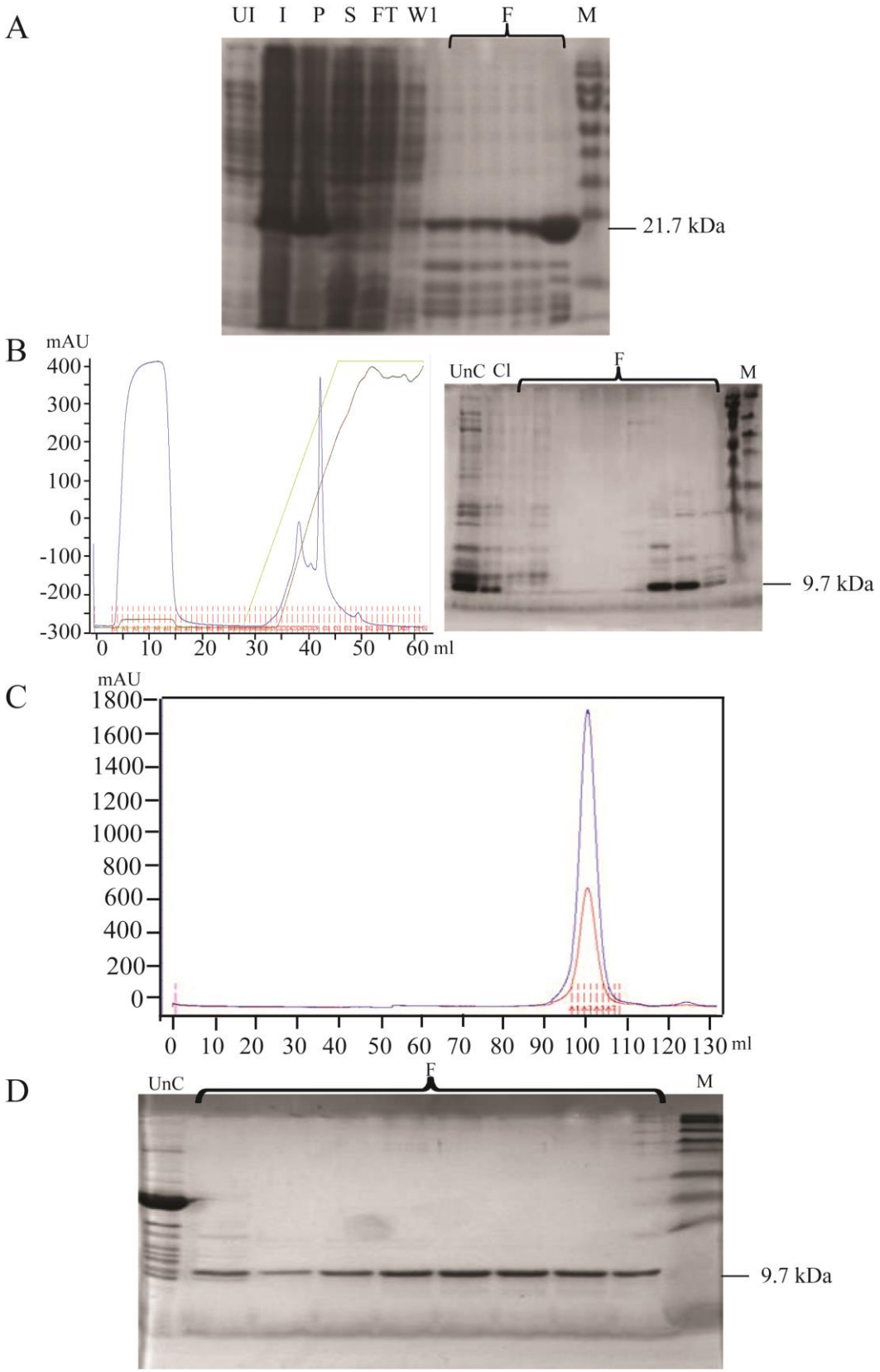


Figure 5.2 Expression and purification of Sox5HMG domain. (A). Ni-NTA purification analysed on 12%SDS-PAGE;(B) Ion exchange profile showing purity of the expressed protein; (C) Size exclusion chromatography of the purified protein and (D) .SDS-PAGE analysis of eluted fractions-F**.** UI: Un-Induced; I:Induced; P:Pellet; S:Supernatant; W1:Wash 1; F:Fractions of Elution; UnC:Uncleaved; Cl:Cleaved; M:Molecular-weight markers (kDa).

5.2 Sox5 HMG domain: secondary structure analysis

As mentioned earlier the secondary structure of the protein can be determined by CD spectroscopy in the "far-UV" spectral region. In order to verify that the purified Sox5-HMG domain was well folded and retained native structure, circular dichroism (CD) analysis was performed as mentioned in Materials and Methods. The purified HMG domain of Sox5 showed typical alpha helical structure with a single positive maximum at 195nm and two negative minima at 208 and 222 nm (*Fig 5.3*) consistent with the amino acid sequence prediction.

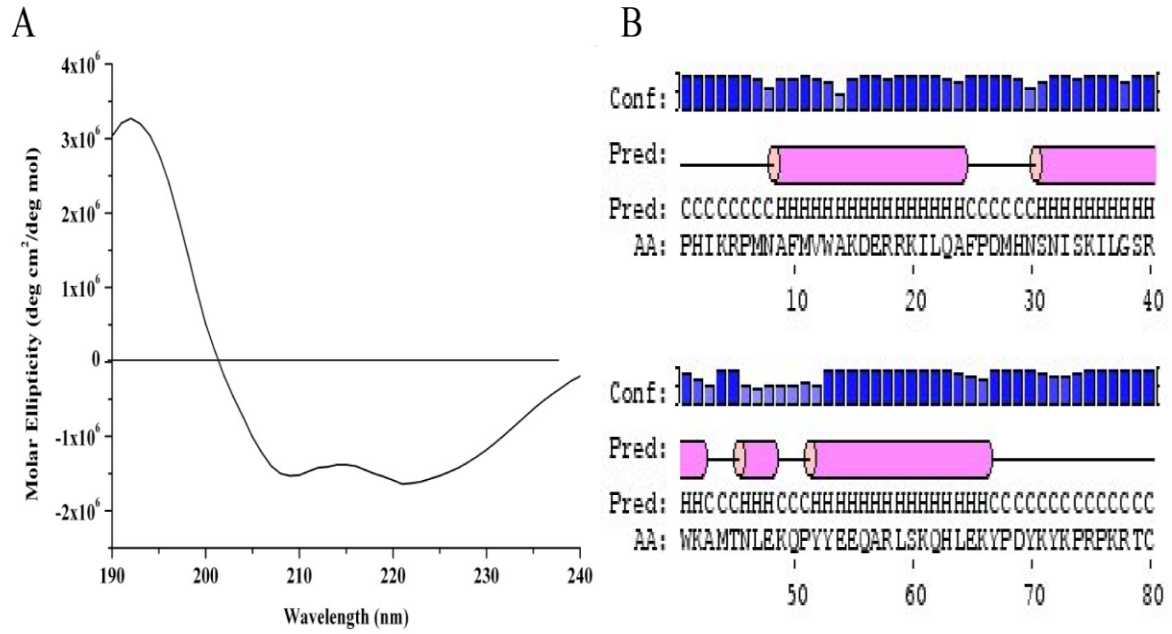


Figure 5.3 Secondary structure analysis of Sox5HMG. Panel A Circular dichroism spectroscopy analysis of the purified Sox9HMG domain; Panel B Secondary structure prediction using PSIPRED software [146].

5.3 Mass spectrometry analysis of Tryptic digested Sox5HMG domain

As described in section 3.4, the Sox5HMG protein was proteolitically digested by trypsin leaving the protein as 7 peptide fragments as shown in *Table 5.1*. The property of trypsin is widely used to authenticate the expressed protein. Mass spectrometry data of the in-gel tryptic digested peptides analysed in MASCOT-DEMAN database.

Start	End	Sequence
505	509	EPHIK.R
505	520	EPHIKRPMNAFMVWAK.D
526	540	K.ILQAFPDMHNSNISK.I
541	545	K.ILGSR.W
548	555	K.AMTNLEK.Q
555	564	K.QPYYEEQAR.L
567	572	K.QHLEK.Y

Table 5.1 Tryptic digested peptides of Sox5HMG domain analysed in MASCOT-DEMAN database.

5.4 DNA binding affinity of purified Sox5HMG protein

The affinity of the purified Sox5HMG protein with DNA was assessed in a preliminary experiment with short Sox consensus sequence using electrophoretic mobility shift assays (EMSAs) as described in Materials and Methods. EMSA experiments were performed with purified short oligo of Col2a1 (5'AGCCCCATTCATGAGA3') and Col4a2 (5'CCTTCTTG TTACGGGG 3') enhancer element run on native gel. A short oligo was designed from the Sox binding site containing Col2A1 and Col4A2 enhancer sequence. The purified Sox5HMG protein was incubated with synthesized Cy5 labeled double stranded DNA. The concentration of DNA was maintained 1nM throughout the study. With increasing concentration of Sox5HMG protein bands corresponding to non-shifted free DNA and the shifted protein-DNA complex band (*Fig. 5.4*) were observed. The apparent dissociation

constant (Kd) of Sox5HMG with COL2a1 and COL4A2 was determined as 25nM as 10 nM respectively, from the protein concentration at which half of the DNA is bound.

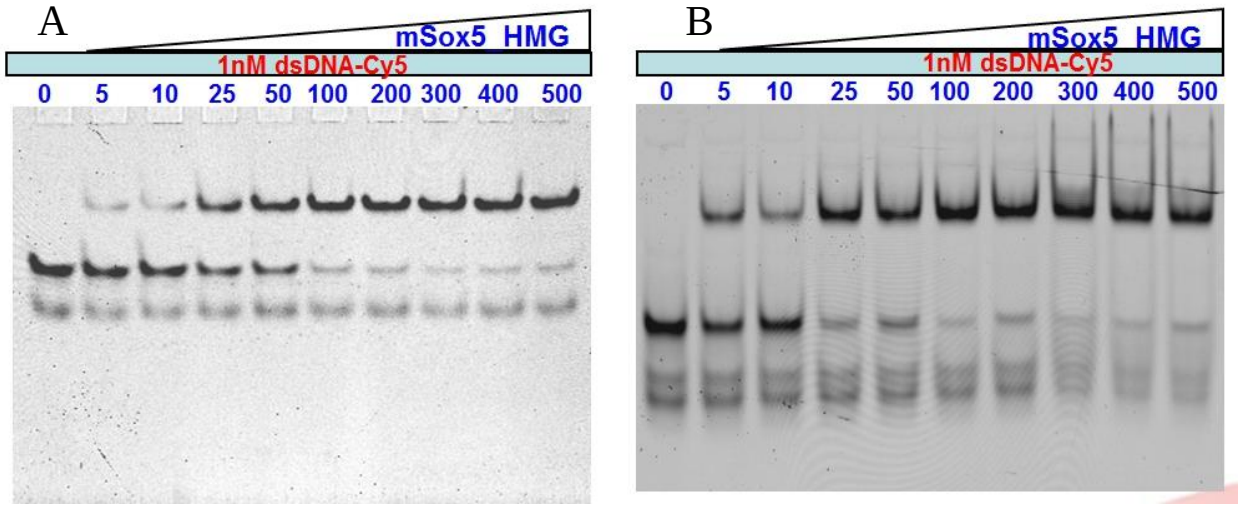


Figure 5.4 Electrophoretic mobility shift assay of Sox5HMG bound to the (a) Col4A2 (5'CCTT CTTGTTACGGGG 3') and (b) Col2A1 (5' AGCCCCATTCATGAGA 3') enhancer element. Reaction with1nM of dsDNA-Cy5 and increasing concentration of protein

5.5 Discussion

Sox-DNA interaction is the consequence of HMG domain binding to DNA. Most DNA binding domains are alpha helical and binding may cause further conformational changes. Circular dichroism (CD) spectroscopy revealed that the purified Sox5HMG comprises of typical alpha helical structure with a single positive maximum at 195nm and two negative minima at 208 and 222 nm (*Fig 5.3),* in agreement with the predicted secondary structure calculated from the primary amino acid sequence analysis (*Fig. 5.3*) using online software package at PSIPRED server. Additionally the result validates the purified protein, owing to its folded compact structure, to be suitable for obtaining quality crystals for further structural characterization. Tryptic digested Sox5HMG domain yielded 7 peptides as shown in Table 4. The specific cleavage property of trypsin has been exploited to analyse the protein primary structure using mass spectrometry (MS). MS data of the in-gel tryptic digested peptides analysed in MASCOT-DEMAN database confirmed the presence of Sox5HMG

domain. The DNA binding domain of the purified transcription factor, Sox5 was analysed with EMSA experiment to assess its binding affinity towards DNA (*Fig. 5.4*). Mouse Col2a1 gene is an established target of regulation by Sox [111], anomalous regulation leads to campomelic dysplasia [113]. Sox5 is shown to bind to Col4A2 comparitively with a higher binding affinity (10nM) than COL2A1 (25nM).

CHAPTER VI

Crystallization of Protein- DNA Complex

6.1 Co-crystallization of Sox9HMG Domain

Crystallization trials were carried out using 96-well Innovadyne sitting drop vapor diffusion plates for homogeneously purified Sox9HMG domain. Collagen2 gene, Col2a1 is *a* well-established direct target of the Sox trio proteins. Sox9 robustly activates a 48-bp cartilage-specific enhancer located in the first intron of Col2a1 and L-Sox5/Sox6 has been shown to potentiate the activity of Sox9 [111]. Sox9HMG and COL2A1 complex formed as mentioned in Materials and Methods (*Section 2.14*) was used for crystallization screening. Though EMSA performed with 16 bp Cy5 labeled double stranded probe from the enhancer region of Col2a1 (5'AGCCCC**ATTCATG**AGA3') showed high binding affinity with Sox9HMG protein (*Fig 3.5*), initial crystallization trials yielded only poorly diffracting crystals. Therefore, several rounds of optimization including variations in composition of condition, temperature, protein concentration, seeding etc., were carried out.

Length of the DNA and unpaired base pairs in flanking region are two key determining factors for crystallization of protein-DNA complexes [154]. Consequently, oligos from Sox binding sites ranging from 16mer to 14mer (*Table 6.1*) and with AT/CG/GC overhangs were also used for co crystallizing Sox9HMG domain. No crystal growth was observed in case of oligos with AT overhangs, but oligos with CG and GC overhangs produced weakly diffracting crystals (~ 8 to 10 Å) (*Fig 6.1 A- F*).

No	Oligos Name	Sequence
1	Sox9F_16	5' GGAAGAACAATGCCCC 3' 5' GGGGCATTGTTCTTCC 3'
2	Sox9_Col2a1_16	5' AGCCCCATTCATGAGA' 3' 5' TCTCATGAATGGGGCT 3'
3	Sox9_Col4a2_14	5' CTTCTTGTTACGGG 3' 5' CCCGTAACAAGAAG 3'
4	Sox9_Col4a2_16	5' CCTTCTTGTTACGGGG 3' 5' CCCCGTAACAAGAAGG 3'
5	Sox9-41F_16	5' CAGAACATTGTCTGCG 3' 5' CGCAGACAATGTTCTG 3'
6	Sox9_CG_14	5' CAGAACATTGTCTG 3' 5' GCAGACAATGTTCT 3'
7	Sox9_AT_15	5' AGAACATTGTCTGCG 3' 5' TCGCAGACAATGTTC 3'
8	Sox9_14	5' AGAACATTGTCTGC 3' 5' GCAGACAATGTTCT 3'
9	Sox9_GC_17	5' GTCAGAACATTGTCTGC 3' 5' CGCAGACAATGTTCTGA 3'
10	Sox9_CG_14	5' CAGAACATTGTCTG 3' 5' GCAGACAATGTTCT 3'

Table 6.1 Oligonucleotides used for protein-DNA complex formation at mM concentration

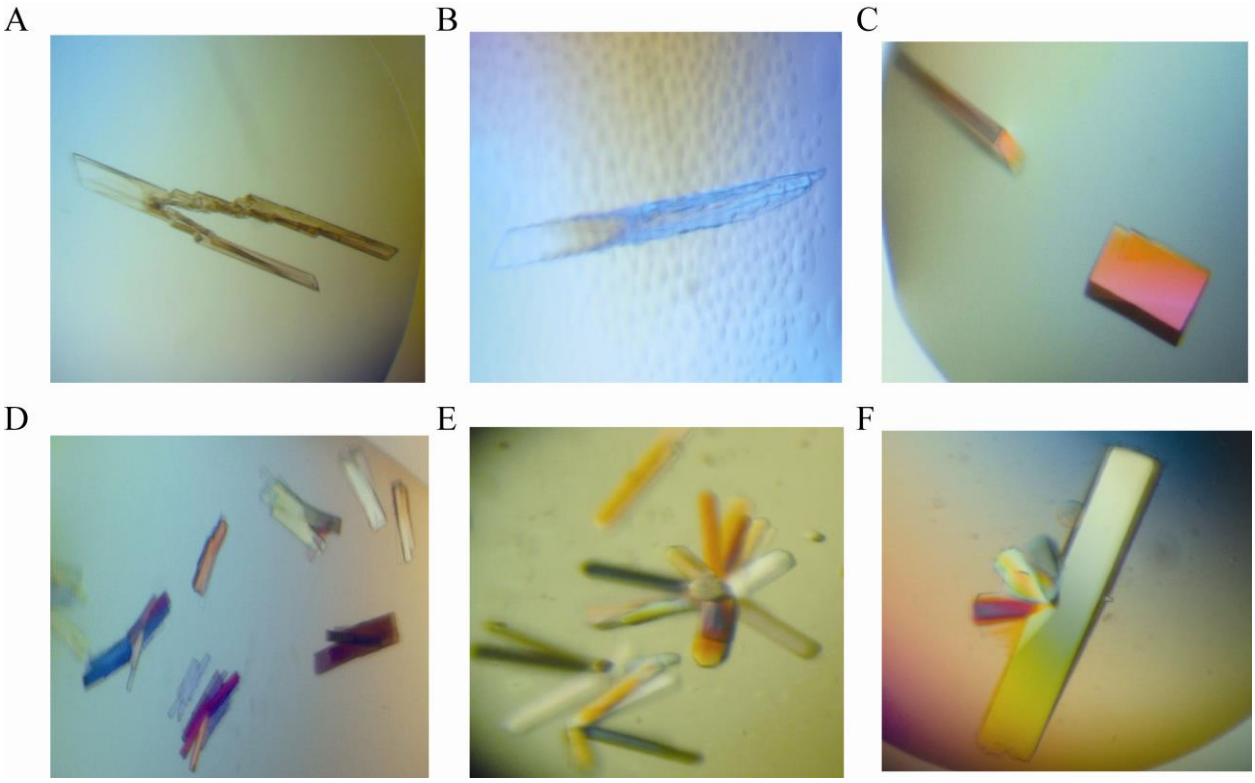


Figure 6.1 Co-crystalization of Sox9HMG domain with of COL2A1 (5'AGCCCCATTCA TGAGA3') DNA element. The complex was formed by incubating protein with DNA at a molar ratio of 1:1.2 and the concentration of protein was maintained as 6 mg/ml. The different conditions which favored crystal growth are A) 10mM Magnesium Acetate; 50mM MES pH 5.6; 2.5M Amso4 B) 0.05 M Na cacodylate pH 6.5, 5 mM CoCl2, 2.5 M KCl C) 100mM Tris pH 8.5; 32% PEG4000; 800mM Lithium Chloride and with 16mer of COL4A2 D) 1M Lithium Chloride; 100mM Sodium Acetate; 30% PEG6000 E) 1.5M Spermine; 10mM Magnesium Chloride; 50mM Sodium cacodylate, pH 6.5; 3M Amso4 F) 100 mM MES pH 6.5; 15 %(w/v) PEG 20000. All crystals diffracted poorly.

6.2 Co-crystallization of Sox9HMG Domain with novel regulatory motif

Co-crystallization of Sox9HMG with oligos of Foxp2 gene promoter sequence was set up. The size (15-17bp) and overhangs (AT, CG and GG) of the oligos were altered to obtain quality crystals. Though, Sox9HMG protein with Foxp2 DNA (5'AGGAGAACA AAGCCTG3') containing GG overhangs yielded better mountable crystals in the presence of tacsimate and PEG 3350, a diffraction lower than 6Å could not be acheived (*Fig 6.2*). Tacsimate, composed of a mixture of titrated organic acid salts contains 1.8305 M Malonic acid, 0.25 M Ammonium citrate tribasic, 0.12 M Succinic acid, 0.3 M DL-Malic acid, 0.4 M Sodium acetate trihydrate, 0.5 M Sodium formate, and 0.16 M Ammonium tartrate dibasic

[155], is a unique crystallization reagent developed exclusively by Hampton Research. Further optimization with variation in pH (pH 5 (*Fig.6.2D*), pH 6 (*Fig.6.2C*), pH 7 (*Fig.6.2B*), pH 8 (*Fig. 6.2A)*), tacimate concentration (from 2 v/v, 4 v/v, 6 v/v and 8 v/v), percentage of PEG3350 and temperature (4, 15, 18 and 25°C) were carried out.

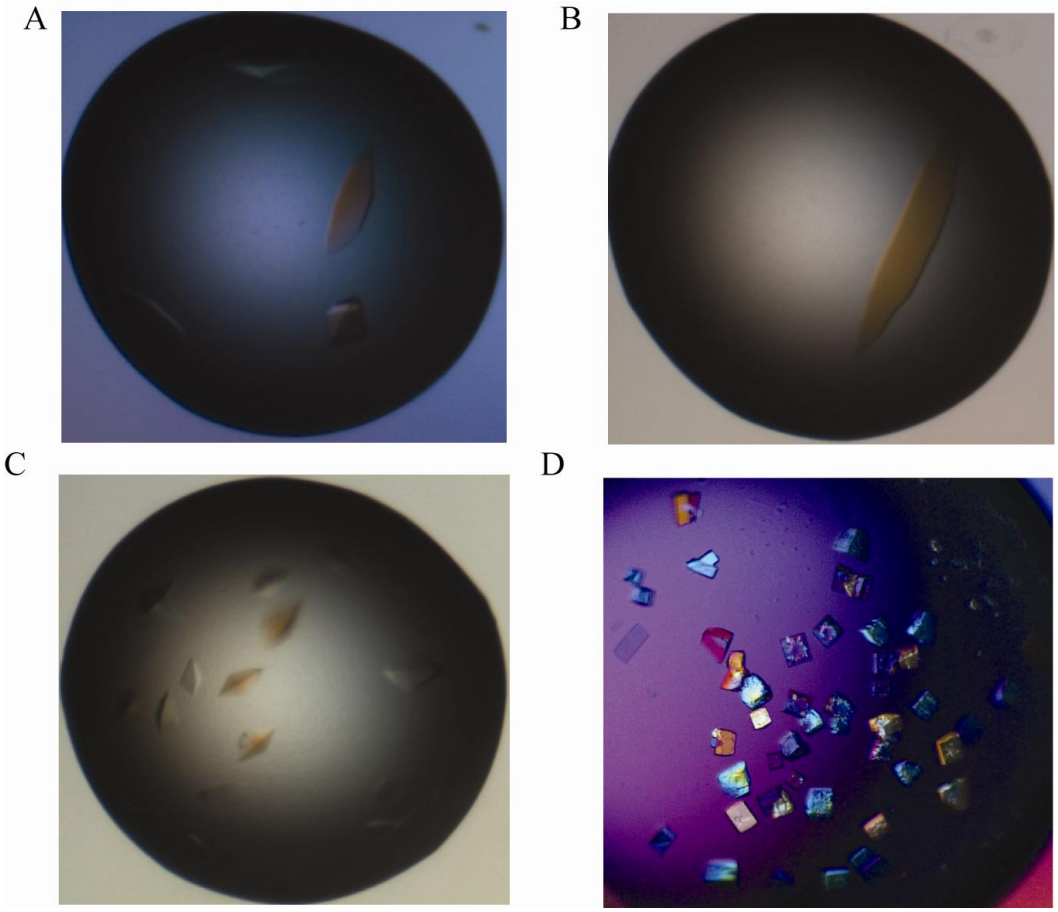


Figure 6.2 Crystals of Sox9HMG domain with Foxp2 DNA (GG overhang) 5' AGGAGA ACAAAGCCTG 3'. The complex was formed by incubating protein with DNA at a molar ratio of 1:1.2 and the concentration of protein was maintained as 3 mg/ml. The different pH conditions that favored crystal growth at 18° C are (A) 20 w/v PEG 3350, 8 v/v Tacsimate, pH 8. (B) 20 w/v PEG 3350, 8 v/v Tacsimate, pH 7 (C) 20 w/v PEG 3350, 8 v/v Tacsimate, pH 6 (D) 16 w/v PEG 3350, 2 v/v Tacsimate, pH 5, 100mM tri-sodium citrate, pH 5.6. The crystals in conditions A,B,C diffracted up to 6.0 Å and condition D diffracted to 4.0 Å.

Finally, better diffracting crystals (~4Å) (*Fig 6.2 D*) were obtained with conditions of 16% W/V PEG 3350, 2V/V Tacsimate, pH 5.0 with 100mM tri-sodium citrate, pH 5.6. Crystals testing for X-ray diffraction showed 10-20% glycerol as cryo-protectant reduced the diffraction quality. The crystals were harvested in 5 days and stored in liquid nitrogen. The

3.0Å native data set was collected at National Synchrotron Light Source, Brookhaven National laboratory (X29) (*Table 6.2*).

Parameters	Values
Detector	CCD ADSC-Q315r
X-ray source	BeamlineX29-54 pole mini-gap undulator
Wavelength (Å)	1.0750
Oscillation angle (˚)	1
Exposure time (Sec.)	40
Space group	$P4_32_12$/ $P4_12_12$
Crystal to detector distance (mm)	300
Unit-cell parameters (Å, ˚)	a = b= 98.581, c =45.919
	α = β = γ = 90
Resolution range (Å)	50-3.0 (3.11-3.0)
No. of reflections	97935
No. of unique reflections	4848
Matthew coefficient [V_m(Å^3/Da)]	2.88
Solvent content (%)	61.83
No. of molecules in asymmetric unit	1
$^{\dagger}R_{sym}$ (%)	13.1
Average redundancy	21.2 (23.3)
Completeness (%)	95.5 (96.4)
Average *I/σ(I)*	21.7 (4.3)

Values in the parenthesis are for highest resolution bin $^{\dagger}R_{sym} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the measured intensity of reflection I and $\langle I(hkl)\rangle$ is the mean intensity.

Table 6.2 Data collection and processing statistics for Sox9HMG

6.3 Co-crystallization of Sox5HMG Domain

Crystallization trials for Sox5HMG were performed using 96-well Innovadyne sitting drop vapor diffusion plates for homogeneously purified Sox5HMG domain. Since Collagen2 gene, *Col2a1* was well established direct target of the Sox trio, the Sox5HMG and Col2a1 complex was formed as mentioned in Materials and Methods and used for crystallization screening. EMSA experiment was carried out with 16 bp Cy5 labeled double stranded probe from the enhancer region of Col2a1 (5' AGCCCC**ATTCATG**AGA 3') with Sox5HMG

protein confirming the binding. However no well diffractable crystals were obtained even after several rounds of optimization including changing the temperature, concentration of protein, varying length of DNA, different over hangs, seeding etc., So altered size of oligos ranging from 16mer to 14mer oligos (*Table 6.1*) with unpaired base pairs at the Sox binding sites were also used. Unpaired base pairs at the flanking region aid in stacking of DNA in the crystal lattice or form symmetry related protein-DNA contacts leading to diffraction quality crystals [156, 157]. Numerous micro crystal growth was observed in case of oligos with AT overhangs. Though oligos with CG and GC overhangs gave mountable crystals diffraction was very weak (~ 15 to 20 Å) (*Fig 6.3 A, B*).

Apart from Collagen gene promoter sequence, EY-Globin gene promoter sequence and Lama1 promoter sequence were also tried with Sox5HMG (*Table 6.3*). Sox5HMG protein with Lama1 DNA (5' CCAGGACAATAGAGA 3') (CG overhang) complex gave better mountable crystals. Most of the conditions with MgCl2 or CaCl2 with PEG3350 or PEG4000 produced crystals and most of them are mountable in size (*Fig 6.4)*. The crystal of Sox5HMG-Lama1 with single base overhangs of C-G grown in the condition of 200mM Magnesium Chloride, 100mM HEPES, pH. 7.5, 15% (w/v) PEG 400 and 100mM Bis-Tris, pH. 6.5 500mM Magnesium Formate dihydrate, produced crystals with ~5-6 Å (*Fig 6.5)*.

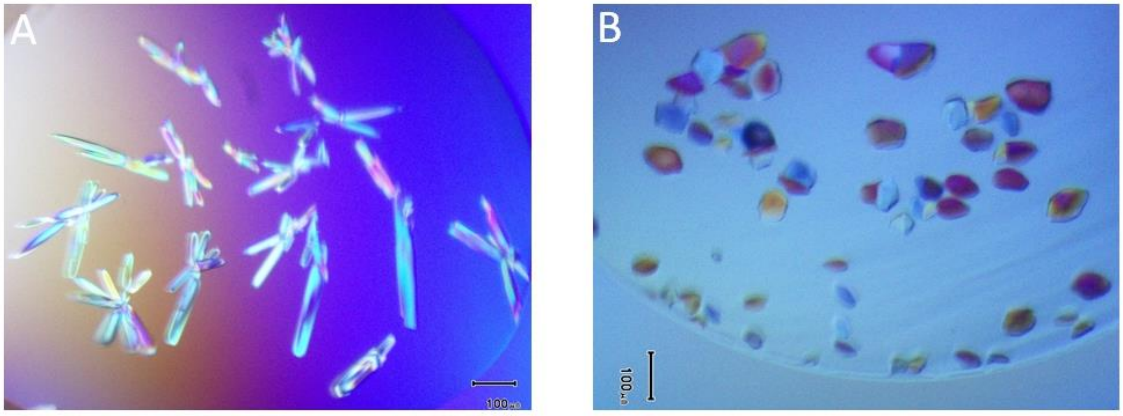


Figure 6.3 Crystal obtained from Sox5HMG-Col2A1. A) CG overhang with the condition of 100mM MES, pH. 6.5, 600mM NaCl, 20% (w/v) PEG 4000 and diffracted 15 Å. B) GC overhangs with the condition of 200mM Calcium Acetate, 100mM HEPES, pH. 7.5, 10% (w/v) PEG 8000 diffracted 20 Å.

Name	Forward	Reverse	Overhang	Remarks
Lama1	CCAGGACAATAGAGA	GTCTCTATTGTCCTG	CG	Bigger/Better Diffraction
Lama1	GCAGGACAATAGAGA	CTCTCTATTGTCCTG	GC	Bigger/poor Diffraction
Lama1	TCAGGACAATAGAGA	ATCTCTATTGTCCTG	TA	Bigger/No Diffraction
Lama1	TCAGGACAATAGAGA	TTCTCTATTGTCCTG	TT	No Crystals
Lama1	CCAGGACAATAGAGAC	GTCTCTATTGTCCTGG	Blunt	Poor Diffraction

Table 6.3. Lama1 DNA element used for Sox5HMG-DNA complex crystallization

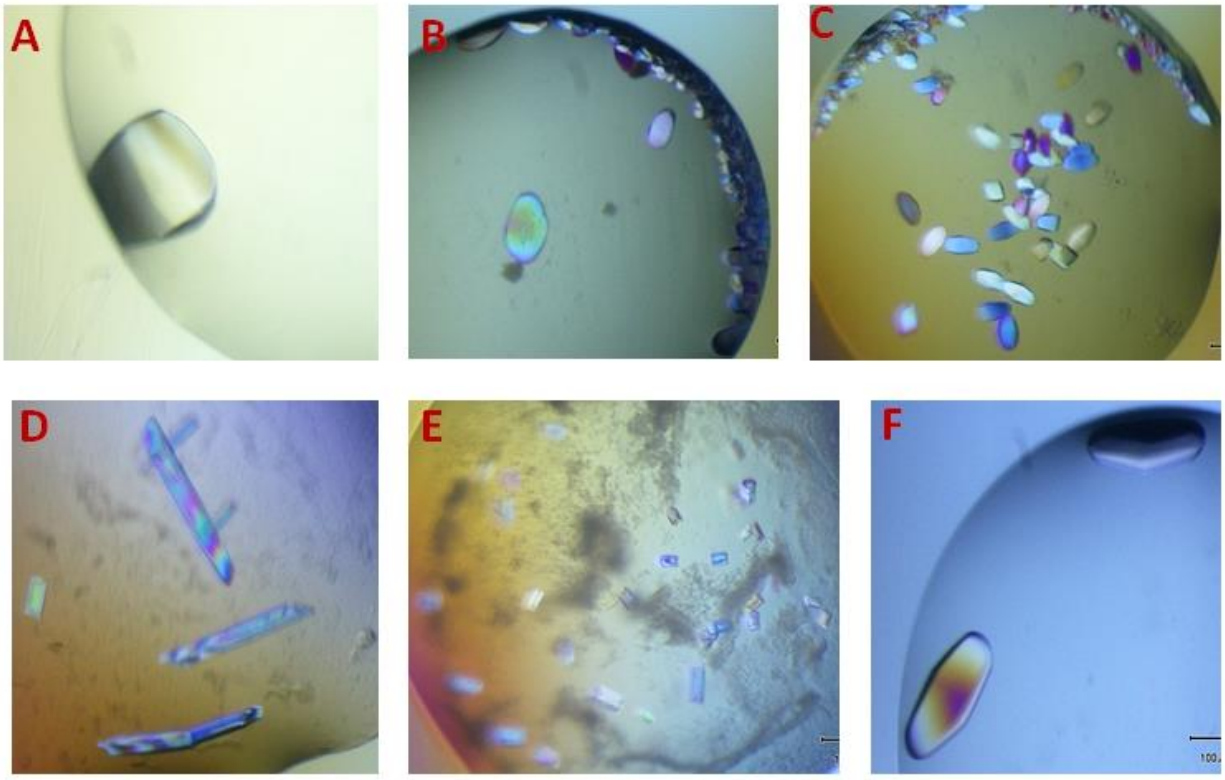


Figure 6.4 Crystal obtained from Sox5HMG-Lama1 DNA with the conditions of A) CG overhang-200mM Magnesium Chloride, 100mM HEPES, pH. 7.5, 15% (w/v) PEG 400 B) TA Overhang-180mM tri-Ammonium Citrate, 20% (w/v) PEG 3350; C) GC overhang-200mM Magnesium Chloride, 100mM MES, pH. 6.5, 10% (w/v)PEG 4000; D) CG overhang-100mM Bis-Tris, pH. 6.5 500mM Magnesium Formate dihydrate; E) With TA Overhanf_ 180mM tri-Ammonium Citrate 20% (w/v) PEG 3350.

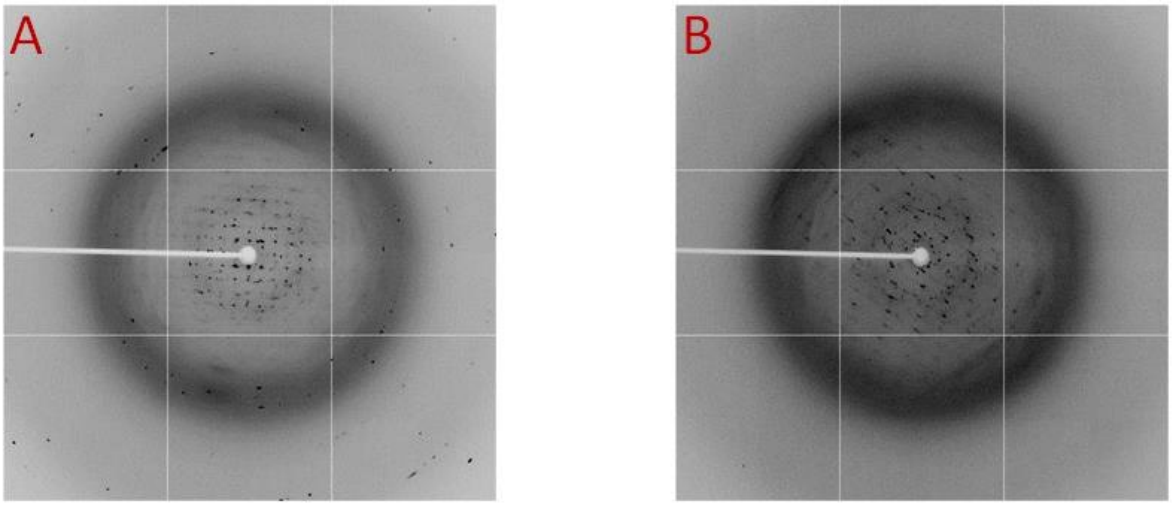


Figure 6.5 The Diffraction Images of crystal obtained from Sox5HMG-Lama1 DNA (CG overhang) with the condition of A) 200mM Magnesium Chloride, 100mM HEPES, pH. 7.5, 15% (w/v) PEG 400 B) 100mM Bis-Tris, pH. 6.5 500mM Magnesium Formate dihydrate

6.4 Discussion

Transcription factors (TFs) and their definite interactions with targets is key for detemining gene expression patterns. In order to gain insight into the transcriptional regulatory networks and mechanism of action it is essential to understand the exact binding motif and the protein DNA specific interactions. Although nearly 30 or more Sox proteins have been identified so far, only three high resolution Sox HMG domains belonging to sub groups B (Sox2), sub group C (Sox4), sub groupF (Sox17) have been obtained in the presence of DNA (*Section 1.12, Fig 1.13*).

Despite the fact that Sox HMG domains display high similarity at the nucleotide level, similar helical axis of the HMG L-shaped helix arrangement, amino acid contacts with the DNA minor groove, transcriptional regulation is highly group specific (*Section 1.12, Fig 1.13*). Moreover, the functional validation results of the novel regulatory motifs identified in the current study (*Fig 4.2)* and other Sox transcriptional regulation studies have clearly denoted that enhancer activation by Sox is highly specific for each subclass of Sox proteins depending on their DNA-binding specificities [94]. Based on these observations we presume that high resolution structural details of Sox trio with cognate DNA would shed insight into

the transcriptional specificity of the Sox9 and its regulatory effect on the downstream Sox5 and Sox6.

Crystallization trials were carried out with homogeneously purified Sox9HMG and Sox5HMG proteins. Collagen2 gene, Col2a1 is well established direct target of the Sox trio proteins; Sox9 robustly activates a 48-bp cartilage-specific enhancer located in the first intron of Col2a1 and L-Sox5/Sox6 have been shown to potentiate the activity of Sox9 [111].

Though Col2a1 motif is a direct target of Sox proteins, and bound Sox9HMG or Sox5HMG at nM concentrations, initial crystallization trials bestow only poorly diffracting crystals despite modifications in the length and overhangs of oligos used (*Table 6.1 and Fig 6.1).* Additionally and importantly, the results indicate that the affinity of Sox9HMG or Sox5HMG towards Foxp2 was comparatively higher than that of Col2a1. This could likely be due to the fact that Col2a1 has only four sites with as few as five or six Sox consensus nucleotides compared to Foxp2. Accordingly, co-crystallization of Sox9HMG with GG overhang oligos of Foxp2 gene promoter sequence element yielded well grown crystals diffractable up to 3.0Å. The data processing indicated that the crystal belongs to primitive tetragonal $P4_32_12$ or $P4_12_12$ with unit cell parameters of a=b= 98.581 c=45.919 Å (*Table 4.3*). The only hindrance for structure solving in the collected data was high mosaicity (1.73-2.26). Mosaicity is angular measure of the degree of long-range order of the unit cells within a crystal. A low mosaicity index denotes ordered crystals of better diffraction quality. Variations in the crystal conditions temperature, prescence of additives might aid to reduce the mosaicity. Glycerol used as an additive to reduce the mosaicity of Sox9HMG-Foxp2 crystals reduced the diffraction quality and annealing (freeze/thaw) the crystal also increase the mosaicity. Further efforts in optimizing the condition would yield better solving crystals. (The solved Sox9HMG-Foxp2 (2.7 Å) crystal structure details are provided in the Appendix I).

In the case of Sox5HMG, though extensive crystallization trials were carried out with several oligos as shown in *Table 6.1*, varying in their DNA sequences, length and overhangs, and Sox gene regulatory promoter sequence gave only poorly diffracting crystals. Apart from Collagen gene promoter sequence, EY-Globin gene promoter sequence and Lama1 promoter sequence were also tried with Sox5HMG. Laminin subunit alpha-1, a protein encoded by the LAMA1 is a major protein in the basal lamina, a protein framework for most cells and organs. ChIP–ChIP analysis demonstrated that Sox17 occupied the regulatory regions upstream of Col4a1, Col4a2, Lama1 and loss of Sox17 resulted in a significant reduction in Lama1, Col4a1, Col4a2. Earlier, our group has reported the crystal structures of lama1 DNA bound Sox17 and Sox4 [94]. Likewise, Sox5HMG protein with Lama1 DNA (CG overhang) complex gave better mountable crystals with the conditions of MgCl2 or CaCl2 with PEG3350 or PEG4000. The crystal of Sox5HMG-Lama1 with single base overhangs of C-G grown produced crystals with ~5-6 Å (*Fig 6.4)*. But TT overhang did not produce any crystals. With the overhang of CG the protein diffracted well but could not help with crystal dying. The crystal growth hurdle requires further optimizing.

PROTEIN – PROTEIN INTERACTION

OF SOX TRANSCRIPTION FACTORS

CHAPTER VII

DNA Dependent protein-protein interaction of Sox9 Dimerization domain

7.1 Dimerization Domain of Sox9

The DNA binding domains of Sox transcription factors, HMG are significant as they play discrete and important roles in varied developmental processes, nevertheless the DNA-binding sites they recognize are dissolute. The binding affinity, specificity and trans-activation of Sox TFs is highly dependent on the orientation, spacing of the binding and the binding sites of other cofactors. Sox proteins are in great degree conceived to interact specifically with other transcription factors as partners to play unique roles in diverse cell types and regulate distinct pathways in the same cell type. Several reports have suggested that SoxE transcription factors bind as homodimer or heterodimer (with other members of SoxE group) through the dimerization domain located at the N-terminus of the HMG domain, functionally important in chondrocytic cells (Han and Lefebvre, 2008). In order to gain insight into the precise role of the DNA dependent oligomerisation and the effect of oligomerisation on the DNA binding affinity of HMG domain, Sox9 was taken as a model.

7.2 Cloning and expression of Sox9HMG encompassing Dimerization domain

The Sox9 dimerization domain along with HMG domain (Sox9DHMG) spanning amino acids 60-181 was PCR-amplified from a cDNA clone (IMAGE;5354229) using the following DNA primers Forward: 5' GGGGACAAGTTTGTACAAAAAAGCAGGCTTCG AAAACCTGTATTTTCAGGGCATGAATCTCCTGGACCCCTTCAT and Reverse: 5' GGGGACCACTTTGTACAAGAAAGCTGGGTTTATCACACCGACTTCCTCCGG 3'.

The PCR product of Sox9DHMG was cloned into the entry vector pDONR221 using Gateway BP technology (Invitrogen) and confirmed by sequencing. The insert was subcloned in the destination vector pETG20A using Gateway LR cloning technology (Invitrogen). The expression plasmid was transformed into Escherichia coli BL21 (DE3) cells (Invitrogen) and grown in Luria–Bertani (LB) broth containing 100ug/ml ampicillin. When an OD600nm of 0.7 was reached, the temperature was lowered to 30° C and protein expression was induced by adding 0.3 mM isopropyl -d-1-thiogalactopyranoside (IPTG). Cells were harvested by centrifugation after 4 hours and stored at -80° C.

7.3 Purification of Sox9HMG-Dimerization domain

The expressed protein of interest was purified from the mixure of protein by His-Trap columns (GE Healthcare) equilibrated with buffer A (50 mM Tris, 100 mM NaCl, 10 % (v/v) glycerol, 0.5 mM TCEP, pH 8). The fusion protein bound to His-Trap column matrix was eluted using buffer B (50 mM Tris, 100 mM NaCl, 300 mM imidazole, 10 % (v/v) glycerol, 0.5 mM TCEP, pH 8) and subsequently desalted into buffer A to remove imidazole using a pre-packed desalting column. The His6Trx fusion tag of Sox9DHMG was removed from the protein of interest through TEV digestion, performed using protease:substrate ratio of 1:50 (w:w) at 277 K for approximately 12 hours (*Fig 7.1*). The cleaved protein was gradually eluted with 1M NaCl through ion-exchange chromatography, Resource S column (GE Healthcare) to remove tag. The formation of disulfide bonds between the two cysteines in the protein was prevented by addition of TCEP in the buffer. The protein was further purified by size-exclusion column (HiLoad 16/60 Superdex 75 pg, GE Healthcare Bioscience) and fractions were concentrated and stored in -80° C. (Fig. 7.1) The authenticity of the protein was verified by mass spectrometry with in-gel tryptic digestion.

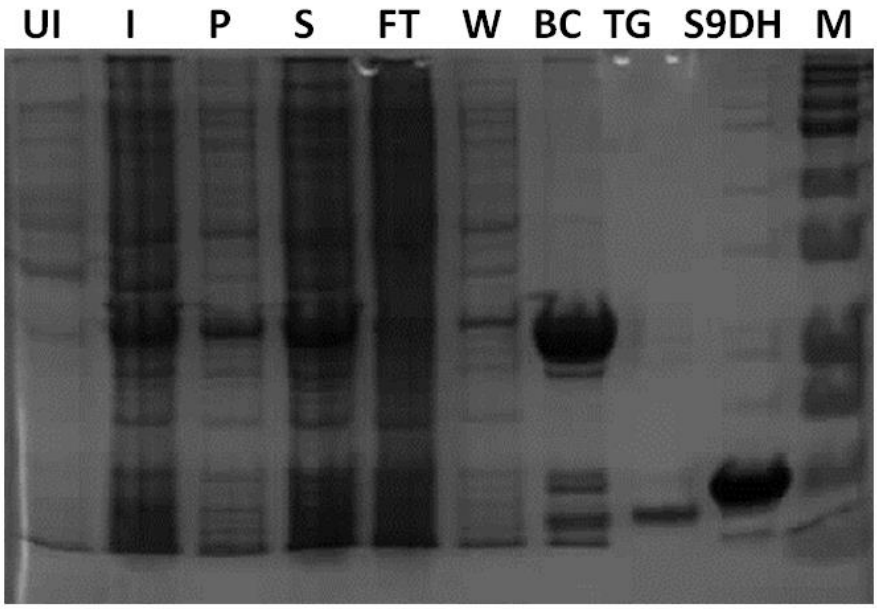


Figure 7.1 Overexpression and purification profile of Sox9DHMG. UI:Un-induced ; I:Induced; P:Pellet; S: Supernatant; FT: Flowthrough; W:Wash; BC:Before cleavage; TG:Tag; S9DH: Sox9DHMG protein. M:Markers

7.4 DNA binding analysis of Sox9HMG-Dimerization domain

In order to determine the orientation and space between the binding sites of the genes involved in cartilage function and regulation, the chromatin immunoprecipitation Sequencing (ChIP-Seq) data was used for electrophoretic mobility assay. Repeat-masked motif analysis of sequence from 100bp surrounding Sox9 peaks was used as input for MD module program as described in Section 4.1. Sox5 promoter sequence has been taken to determine the space requirement between these sites with Sox9DHMG. For electrophoretic mobility assay, the probes were designed by introduction of 2 base pairs to 9 base pairs in between the two binding sites. For analysis purpose, the size of all the Cy5 labelled DNA probes were maintained equalling as 33bp and the introduced flanking sequence and the space sequence were maintained as either C or G an equal proportion, 5’ GCGCGGACAACAATCGGCCA TTGTTCTCGGCGG 3’.

The purified Sox9DHMG protein was incubated with varying spaced Cy5 labelled double stranded DNA in EMSA buffer as referred in Materials and Methods. The concentration of DNA was maintained as 1 nM throughout the studies with increasing concentration of protein. Overall, increasing concentration of Sox9DHMG protein indicated the binding of dsDNA resulting in two higher migrating "shifted" and "super shifted" bands (*Fig 7.2*). Further increase in protein concentration showed higher migrating bands and disturbed migration due to non-specific binding of protein with DNA.

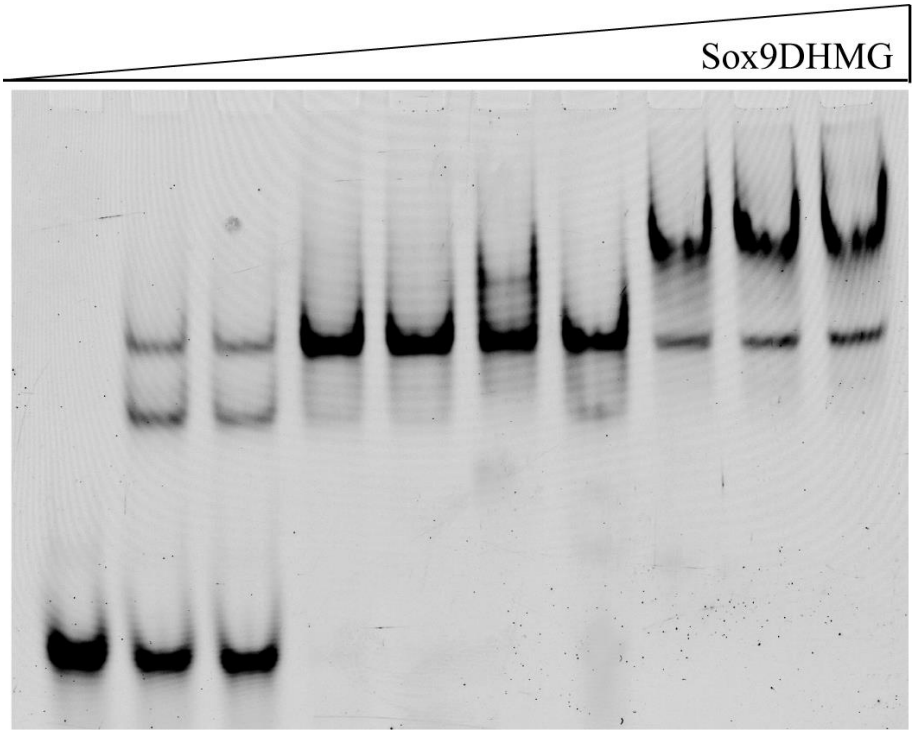


Figure 7.2. Electrophoretic Mobility Shift Assay of Sox9DHMG in complex with Sox5 promoter sequence encompassing two binding sites. The concentration of DNA was maintained as 1 nM with increasing concentration of protein from 0 to 500 nM. 10 µl of Samples were loaded onto 12% native PAGE and bands were detected using a Typhoon 9140 Phosphor Imager

7.5 Oligomeric analysis of Sox9HMG encompassing Dimerization domain

SDS- PAGE analysis of the homogenously purified Sox9DHMG protein migrated as 16.6 kDa was confirmed as a monomer by size exclusion chromatography. Several reports have suggested that Sox9 binds as homodimer or heterodimer (with other members of SoxE group) and is required for chondriogenesis regulation. Through MEME, Motif-based

sequence analysis tool we found a pair of binding sites frequently found in Sox9 Chip-Seq data affirming dimerization of Sox9 to be essential for chondrogenesis regulation. In order to analyse its oligomerization in the presence of DNA, Sox5 promoter sequence with two intact Sox binding site was incubated with Sox9DHMG and subjected to size exclusion chromatography. The molecular weight of size exclusion chromatography corresponds to 60 kDa (*Fig 7.3*).

7.6 Thermal stability: ThermoFluor Assay

The thermal stability of the naked and DNA bound Sox9DHMG DNA was assessed by ThermoFluor Assay using Sypro orange fluorescence. In this experiment the temperature is increased upto 95 °C when protein's native structure is thermodynamically less favourable and the unfolding of protein structure measured as a function of time. Tm is calculated as the midpoint in the thermal progression and defined as the temperature at which the free energy of the native and non-native forms is equivalent. The T_m value was determined by the "R" programme as described in Materials and Methods. Here, Sox9DHMG protein stability was measured in the absence and prescence of preformed DNA protein complex against variations in salt (0 to 1000 mM NaCl) and pH (acidic to basic). The Sox9DHMG without DNA exhibits a sigmoidal curve when the fluorescence intensity is plotted as function of temperature describing as two-state transition whereas DNA bound Sox9DHMG exhibits three state transitions (*Fig 7.4*).

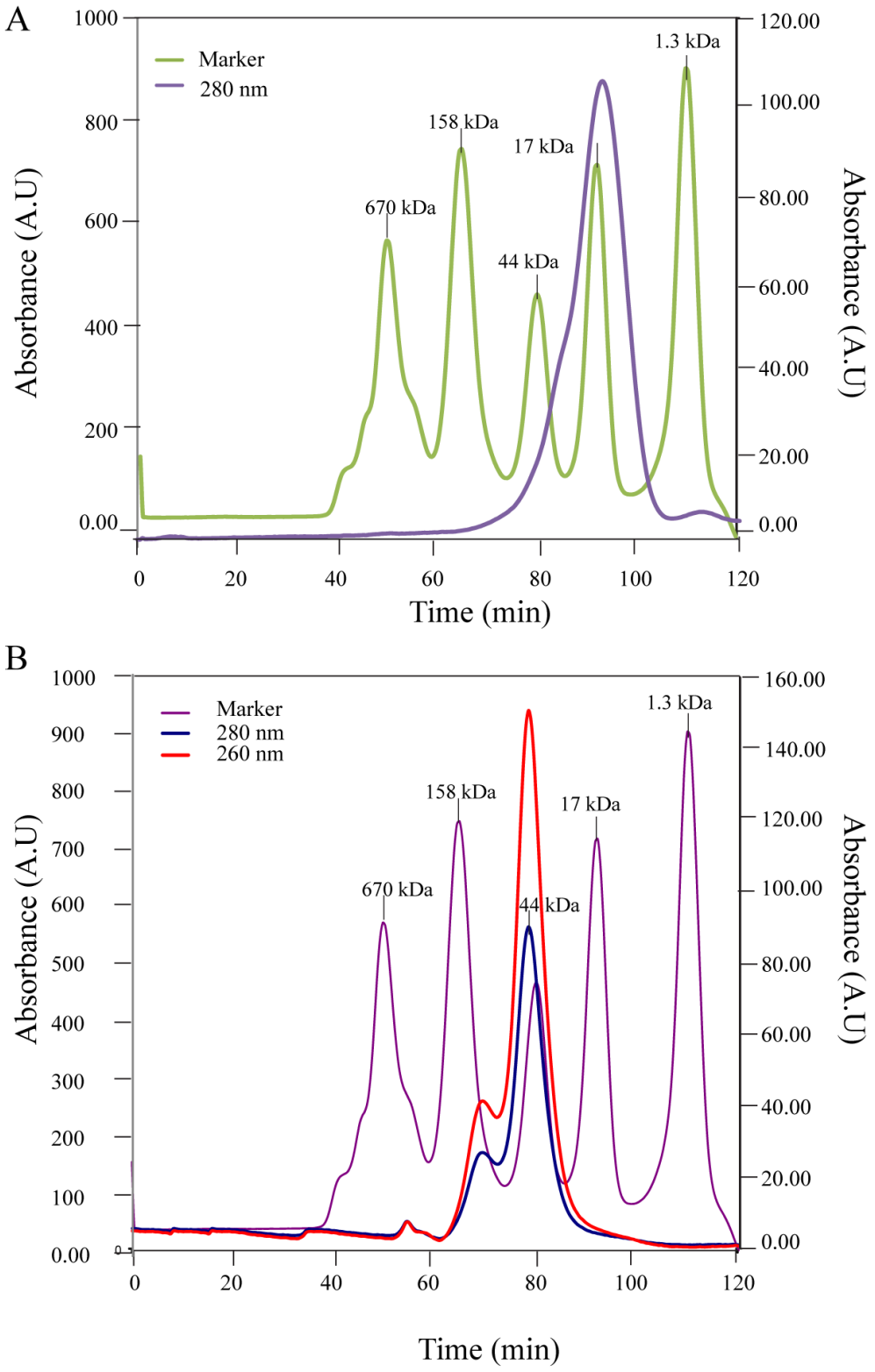


Figure 7.3 Oligomeric analysis of Sox9DHMG. Panel A) Size exclusion chromatography shows the Sox9DHMG eluted as a monomer, overlaid on Marker. Panel B) Eluted as a Dimer in the presence of DNA, overlaid on Marker. Blue line: at 280nm; Redline: at 264nm. Markers (kDa)- 670-158-44-17-1.3kDa.

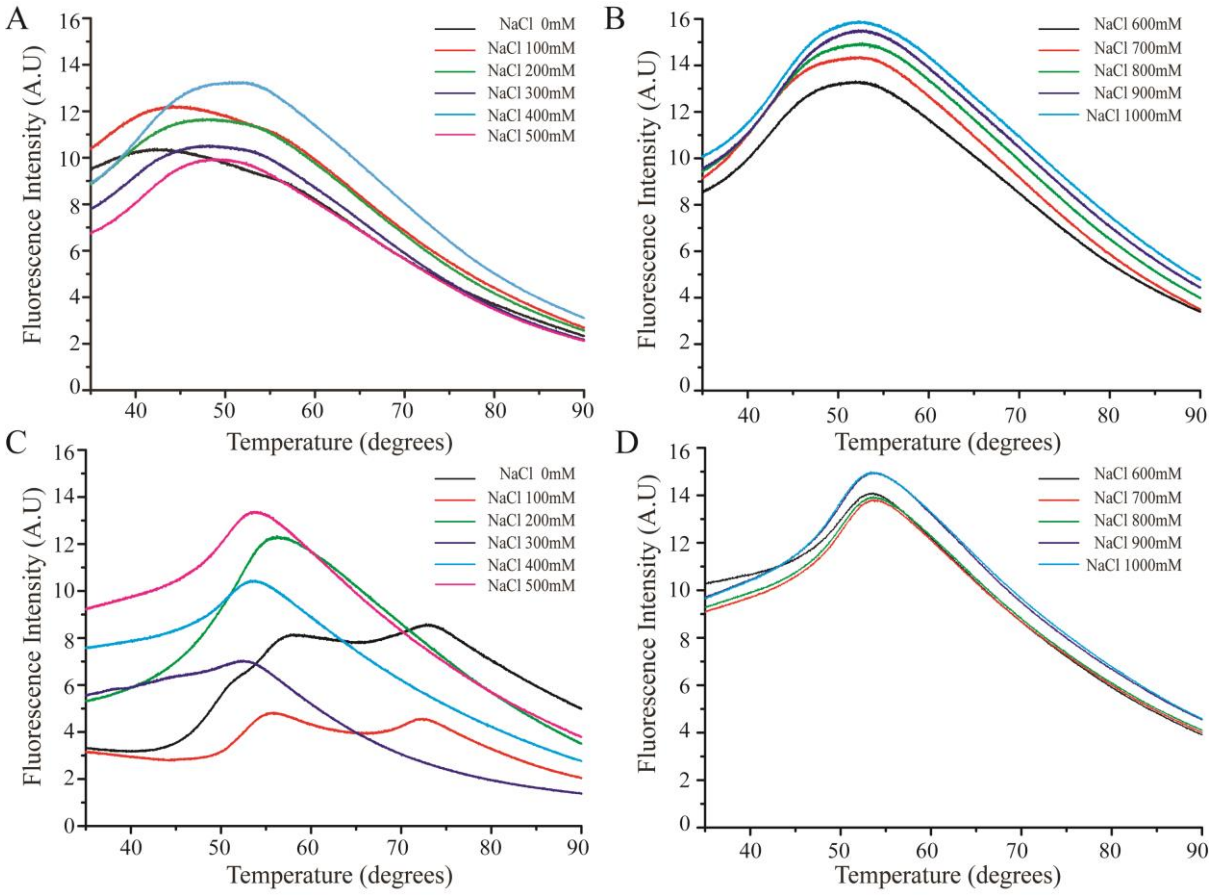


Figure 7.4 Thermal stability of Sox9DHMG and Sox9DHMG-DNA with different salt concentration. A) Sox9DHMG with 0-500mM Nacl B) Sox9DHMG with 600-1000 mM Nacl C) Sox9DHMG+DNA with 0-500 mM Nacl D) Sox9DHMG+DNA with 600-1000 mM Nacl. The Tm for the Sox9DHMG at <100 mM salt concentration is 36° C, while for concentrations > 100 mM is 42° C. Similarly for the two state thermal transition of DNA bound Sox9DHMG indicats two Tm of 52° C and 70° C at salt concentrations <100 mM, wheras at concentrations > 100 mM the transition is single step with Tm 46° C.

7.7 Spacing requirement for effective cooperative binding of Sox9DHMG

Sox9 transcription factor, through its HMG domain recognizes and binds specific seven base pair DNA sequence (A or T) (A or T) CAA (A ro T) G. Reports suggest that Sox9 binds as homodimer or heterodimer (with members of SoxE group) through the dimerization domain which precede the HMG domain. The analysis of Sox9 gene in some CD patients revealed that the mutation in dimerization domain leads to skeletal deformation. In the case of protein-protein interaction between any two transcription factors, the orientation of binding sequence and spacing between their binding sites are critical parameters which have effect on

the cooperation. Accordingly, employing (MEME) Motif-based sequence analysis tool, we found that the oppositely oriented paired binding sites are most frequently found in Sox9 Chip-Seq data suggesting a potential role for Sox9 dimerization in regulation of chondrogenesis. In order to determine the orientation and spacer between the binding sites of the genes involved in cartilage function and regulation, the chromatin immunoprecipitation Sequencing (ChIP-Seq) data was used for electrophoretic mobility assay. The predicted Sox5 promoter sequence possessed two consensus Sox binding sites in opposite orientation with the space of three bp. The space requirement for the co-operative binding of Sox9 was analysed using EMSA, the routine method for analyzing protein–DNA interactions. Examination of the orientation and spacing of the predicted sites revealed that e regulatory elements had one set of pair of sites with opposite orientations seperated by a spacer of 2 to 6 base pair. Therefore the designed model motifs were varied similarly with 2-9 bp (Fig. 7.5) and analysed by EMSA experiment. The purified Sox9DHMG protein was incubated with Cy5 labelled double stranded DNA and EMSA buffer as referred in Materials and Methods.

Figure 7.5 Cy5 labelled DNA sequence used to study the variable spacer length for effective cooperative binding of Sox9DHMG. X=space between the sox binding sites.

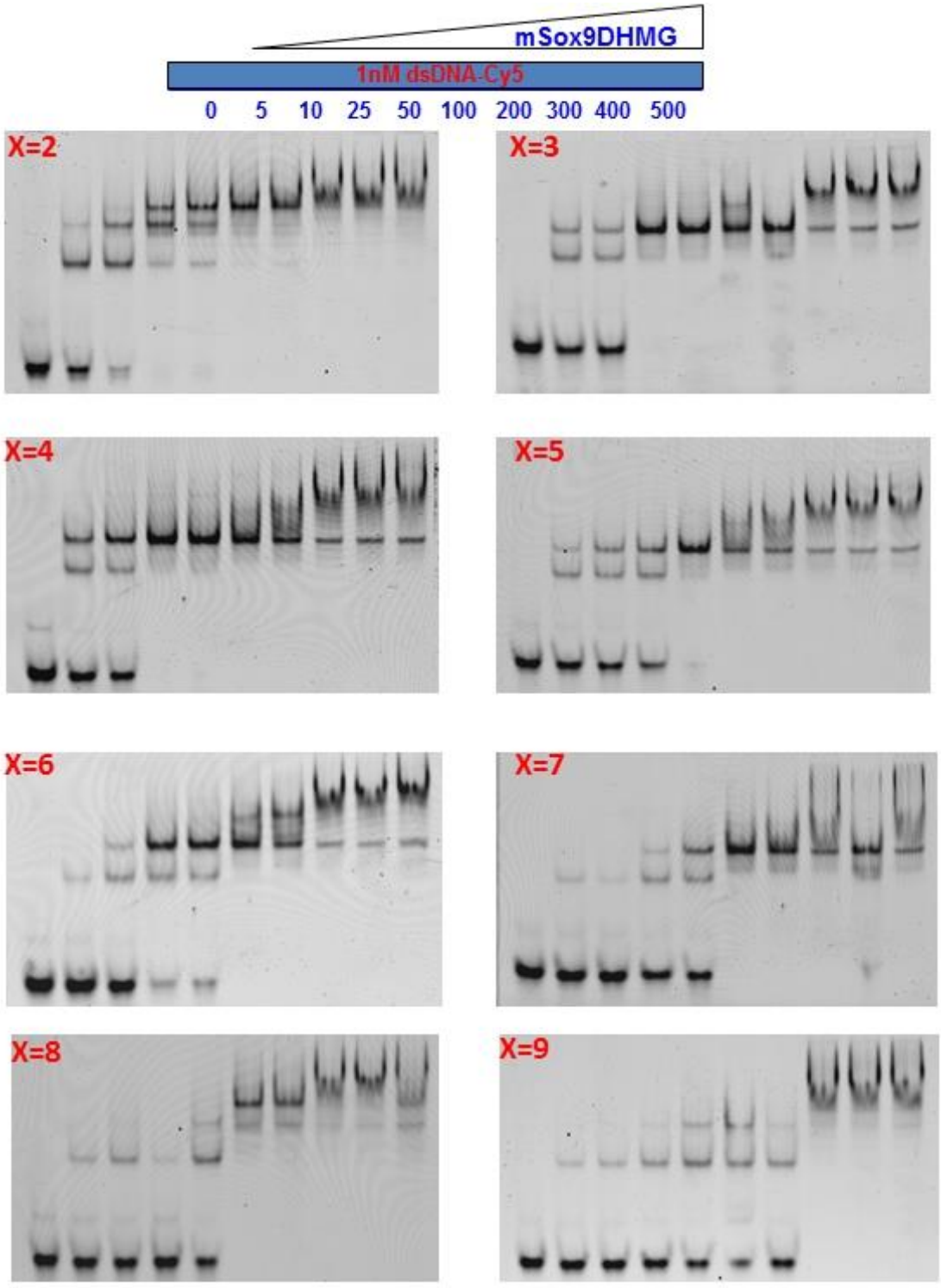


Figure 7.6 EMSA experiment profile with variable spacer length to study the effective co-operative and dimer binding of Sox9DHMG

The concentration of DNA was maintained as 1 nM throughout the studies with increasing concentration of protein. Overall, increasing concentration of Sox9DHMG protein indicated the binding of dsDNA resulted in two higher migrating "shifted" and "super shifted" bands (Fig 7.2). Further increase in protein concentration showed higher migrating bands and disturbed migration due to non-specific binding of protein with DNA. Upon increasing spacer length, the dimer band drastically decreased and the monomer showed very weak binding. DNA binding was completely abrogated with a spacer length of 8 or 9.

7.8 Discussion

Complexity of interaction among transcription factors between protein-protein and protein-DNA guarantees the specificity and efficiency to control the gene expression in eukaryotes. These interactions between multiple protein-protein and protein-DNA interactions enable transcription factors to carry out specific biological functions more precisely than by individual transcription factors. These interactions possibly cause conformational changes, creating specific binding interface on the proteins/DNA, to either activate or repress transcription.

Among Sox transcription factors, only SoxE and SoxD groups have the ability for oligomerization. SoxD proteins possess well known coiled-coil domains which facilitates the oligomerization. Several reports have suggested that SoxE proteins binds to dimeric Sox binding sites functionally important in chondrocytic cells, through a highly conserved region, N-terminal to HMG domain. Consequently, to characterize the conserved domain, we have cloned and purified Sox9HMG with the conserved 40 amino acid at the N-terminal of HMG domain called dimerization domain (Sox9DHMG). The size exclusion chromatography revealed the 16.6 kDa protein corresponds to monomer (*Fig 7.3*) and the protein-DNA complex eluted at size corresponding to 53 kDa, corresponding to dimer complex with DNA, suggesting homodimerization in the presence of DNA and the complex constitutes of 2

molecule of Sox9DHMG on DNA (*Fig 7.3*). Though insilico methods and EMSA experiment suggested the conserved dimerization domain to be DNA dependent, the current study is the first direct molecular size demonstration of the oligomerising domain in an exclusive DNA dependent fashion.

As Sox9DHMG exhibits two state of oligomerization in size exclusion chromatography, monomer and dimer, the stability of the oligomer was studied by ThermoFluor assay. In thermofluor assay, the temperature driven unfolding process exposes the hydrophobic region of proteins, the dye binds to exposed hydrophobic parts of the protein that leads to a significant increase in fluorescence emission, used to monitor the protein-unfolding transition. The fluorescence intensity reaches a maximum and then decreases due to precipitation of the fluorescent probe and denatured protein complex. Protein stability depends on a number of variables such as pH, temperature, ionic strength, buffer, additives etc [158, 159]. Here, the influence of salt and pH on Sox9DHMG thermal stability was studied with and without presence of DNA (*Fig 7.4*). In the absence of DNA, the native Sox9DHMG when challenged by increase in temperature unfolds and exhibits sigmoidal curve of single transition two state model. However, when Sox9DHMG in complex with DNA challenged by increasing temperature unfolds completely and this unfolding correlates to a three-state model with two transitions. In the monomer of Sox9DHMG, the single transition is from F (Folded) state to U (Unfolded) state with the midpoint of 36 °C whereas, in complex with DNA, the dimer of Sox9DHMG, the first transition corresponds to the conversion of F state to the I (Intermediate) state with midpoint of 52.1° C and the second transition, correlates to the unfolding of I state to U state with midpoint occurring at 70°C. In the intermediate state or partially unfolded state of the protein, the protein – DNA complex (in ligand bound state) or protein-protein interaction (in dimer state) might have destabilised. In the case of Sox9DHMG DNA dependent dimerization, forces/factors destabilising the

protein-DNA interaction would also destabilise protein-protein interactions and therefore the protein exists as monomers. The three-state unfolding process consists of two sequential two-state unfolding processes. Influence of ionic strength on thermal stability was studied with increasing concentration of salt. In the absence of DNA, at low salt concentration (<100mM NaCl) the monomer melts with a Tm of 36° C and is more or less stable Tm of 42° C uoto salt concentration of 1000mM. In the case of dimerized Sox9DHMG, the first Tm was 52° C and the second Tm was 70° C. Increase in salt concentration (>100mM NaCl) drifted the denaturation from two-step to one step with a Tm of around 46°C indicating destabilization of the oligomer.

Under physiological conditions, native proteins are stabilized by various weak non-covalent interactions- electrostatic interactions, hydrogen bonding, van der Walls and hydrophobic force, covering full protein molecule. Thus protein conformation is influenced by experimental factors temperature, pH, pressure; presence of destabilizing agents such as surfactant, denaturant alkali and salt and stabilizing agents like metal ions, anions and small organic molecules [15-19]. Native proteins with intrinsically low stabilities are easily susceptible to denaturation and unfold in a highly cooperative manner. Partial unfolding of the structure destabilizes and concominantly collapses to random coil. Reports show unfolding or folding of small globular proteins occurs via a two state process. On the other hand for larger proteins the unfolding or folding is quiet complex and mostly includes the formation of an intermediate [9].

The comparison of the thermal unfolding of the Sox9DHMG domain in the absence (monomer) and presence of DNA (dimer) revealed differences that can be attributed to a stabilization of the protein upon DNA binding. The addition of increasing concentrations of salt or different pH transformed the melting profile, augmenting the unfolding transition of the homodimers, suggesting that Sox9DHMG forms homodimers by directly interacting with

other molecule and stabilizing the complex. Together, the results indicate that Sox9DHMG have direct physical interaction with the DNA increasing the stability of the protein.

Physical interaction between two biological molecules increases the stability of the complex. Sox9, possess both protein-protein interaction domain and DNA binding domain for increased stability, possibly a prerequisite for its functional ability and specificity. During the titration of sox9DHMG against two sites of the Sox5 promoter sequence, the protein binds preferentially to one site and further increase in concentration of protein shows co-operative binding to another binding site which turned to be the super shifted dimer bands.

Cooperative binding happens with both consensus and non-consensus sites and R.I.Peirano et al reported that changing non-consensus sites to consensus does not have any significant impact on co-operative binding ability of Sox9 protein but increases the binding affinity. So consensus or non-consensus is not a determinant of co-operativity of the Sox proteins. The Sox5 promoter binding sites are separated by 3bp. The core sequence of each binding site is one turn helix apart which makes the proteins bind on the same side of the DNA helix. Electrophoretic mobility shows there is a shift and super shift band at a 5nM concentration of protein. The shift corresponds to a monomer bound to a single site of the DNA, and the super shift denotes homodimers bound to two sites of the DNA. In the presence of low concentration of protein, the protein molecules preferentially bind to the single site of the DNA and the monomer band looks prominent over the dimer band. As the concentration of protein increases the protein molecules occupy both binding sites and the dimer band looks prominent. Further titration with higher concentration of protein shows only the dimer band, ie, super shift band. Higher level of protein concentration, force the protein to bind to DNA non-specifically resulting in distorted binding and migration.

It is presumed that the two Sox9 molecules through dimerization domain individually bind to the DNA, physically interacting with each other to form the dimer. But when the

space between these sites has been widened, it disturbs the protein binding on DNA and when the spacer is more than 8bp, it is drastically reduced. It suggests that Sox9DHMG preferentially binds as a dimer to DNA in a co-operative manner and has less affinity for the single binding site. Unlike Sox9 DNA dependent dimerization domain, Sox5 possesses a coiled-coil domain known for its oligomerization ability and "DNA independent" dimerization. Sox5 binds to DNA as a dimer (*vide infra*). Protein-protein interaction through its dimerization domain is particularly important for the transcriptional regulation specificity of Sox9 since Sox proteins bind to similar DNA sequences. An interaction between identical proteins (homodimer) or different proteins (heterodimer) facilitates the complex to bind to different sets of DNA sequences, important criteria for specific gene regulation. Having a DNA binding HMG domain and a dimerization domain in Sox9DHMG suggests that the surfaces for protein-DNA interaction and protein-protein interaction are discrete and co-operative. Our size exclusion chromatography suggests the Sox9DHMG protein exists as a monomer in the absence of DNA and in the presence of DNA they form a dimer. It has been assumed that the protein DNA interaction through HMG domain happens at first followed by protein-protein interaction. The protein-DNA interaction might induce conformational changes on the protein, exposing the protein-protein interaction interface leading to dimer formation. Based on our results, we presume that the dimerising ability of monomeric Sox9 transcription factors could be relatively lower and therefore a DNA-dependent mode of dimerization, wherein the DNA acts as a platform favouring monomeric interactions yield stable dimers. Furthermore, Sox transcription factors are known for bending the DNA and dual bending may change the DNA conformation facilitating the complex to recruit co-factors that further regulate specific biological function.

CHAPTER VIII

Coiled coil mediated oligomerisation of Sox5

8.1 In Silico Sequence Predictions

Besides the conserved HMG domain, Sox transcription factors have other structural and functional domains, whose roles in Sox biological function still remains elusive. The amino acid sequence of Sox5 is shown in *Fig 8.1*. Secondary structure prediction by PSIPRED, indicates that the Sox5 full length is predominantly helical (*Fig. 8.1*) consistent with earlier studies [85]. MultiCoil program[160] locates and predicts coiled-coil in proteins by per-residue scores. The coiled-coil helix forming probability of a given amino acid sequence in the window of its surrounding 21 residue is plotted as a function of the residues position in the primary sequence [161]. In silico analysis of full length protein by the MultiCoil program predicts two segments with a propensity to adopt coiled-coil helices, one at the N-terminus between residues of 193-276 another at the middle of sequence between residues of 400-439 (Appendix A). The N-terminal segment has a high coiled-coil probability factor of 0.81 and the middle segment has a probability factor of 0.66. Further analysis with other methods such as PairCoils [160] and Coils [162] yielded an overall similar prediction. The same analysis was carried out with other Sox D group members comprising Sox6, Sox13 and Sox23. Interestingly, among the four members of SoxD group of transcription factors, only Sox5 and Sox6 possess second coiled-coil domain at the middle of the full-length sequence. The first coiled-coil domain is characterized by the presence of both Leucine Zipper motif and a Q box domain (*Fig1.15 Appendix*). Coiled coil domains favor protein-protein interaction via homo or hetero dimerization. In this context, mSox5 was chosen as a model protein to address the role of its unique domain structure in Sox target gene specificity and other coiled-coil transcription factors in general [85, 88] (*Fig 8.2*).

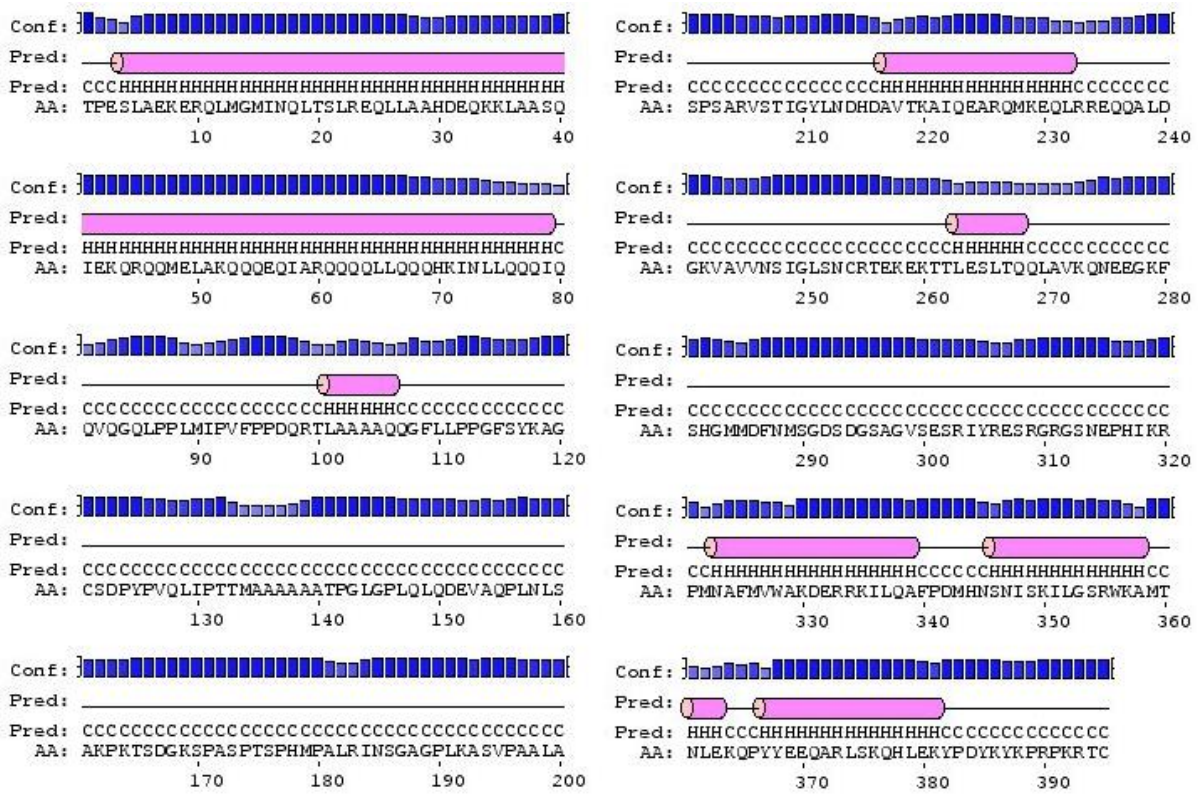


Figure 8.1 Predicted secondary structure of Sox5CC12HMG by online PSIPRED software, which uses a position specific scoring matrices generated by PSI-BLAST for secondary structure calculation by a two stage neural network algorithm (MCGUFFIN *et al.*, 2000).

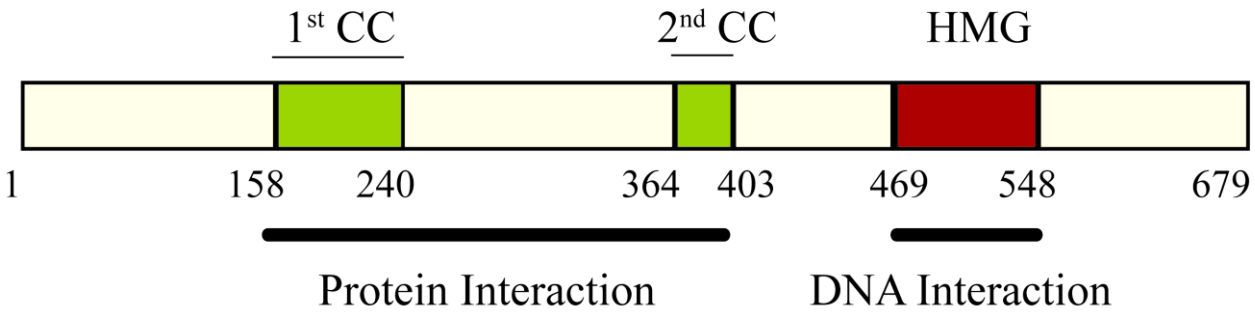


Figure 8.2 Structural and Functional domains of Sox5 Transcription Factor.

8.2 Constructs and Cloning

Various constructs of mSox5, encompassing first coiled-coil domain alone (CC1; 191-280), second coiled-coil domain alone (CC2; 400-439), both coiled-coil domains (CC12; 191-439), second coiled-coil domain with HMG domain (CC2HMG; 400-584), coiled-coil domain 1 and 2 with HMG domain (CC12HMG; 191-584) and a full length (Sox5FL; 1-715) were amplified by using the appropriate primers (as listed on table 2.2.1) from cDNA clone

(IMAGE:40047865) (*Fig 8.3*). Amplified PCR products were cloned into pENTR™/ TEV/D-TOPO® (Invitrogen) cloning vector and cloned products were confirmed by colony PCR and sequencing with gene specific primers and (*Fig 8.4*). Sequencing results for all constructs confirmed the presence of inserts in right orientation. These positive clones were further used as entry clones for Gateway cloning. The gene of interest in the entry clone was introduced into 5 different gateway destination vectors pDEST 17 (HIS tag), pDEST 565 (HIS+GST), pETG20A (HIS+Trx), pETG60A (HIS+Nus A) and pDEST -HIS-MBP (HIS+MBP) by performing a Gateway LR reaction. PCR using gene specific primers confirmed the presence of gene of interest.

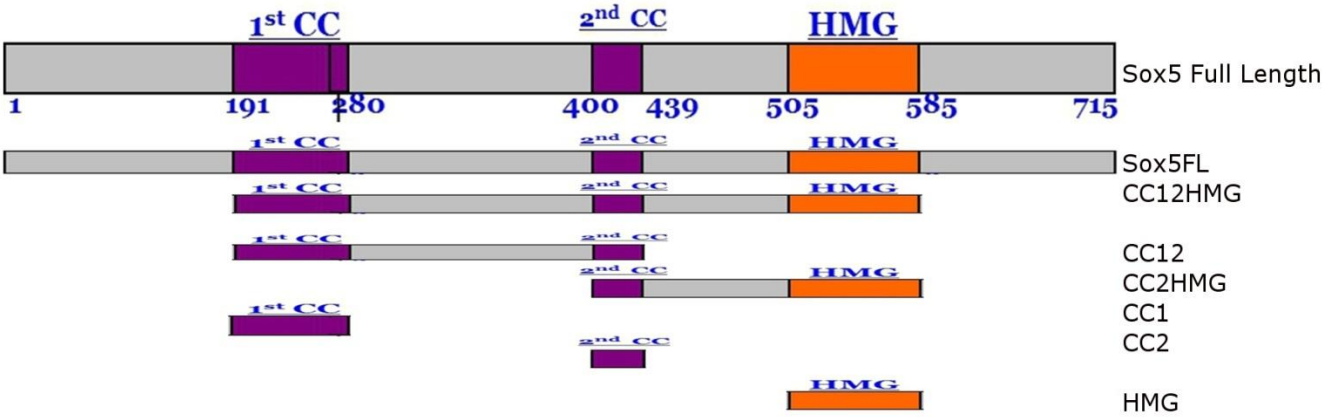


Figure 8.3 Design of Sox 5 constructs encompassing different combination of domains for cloning. Numbers indicate the amino acid position

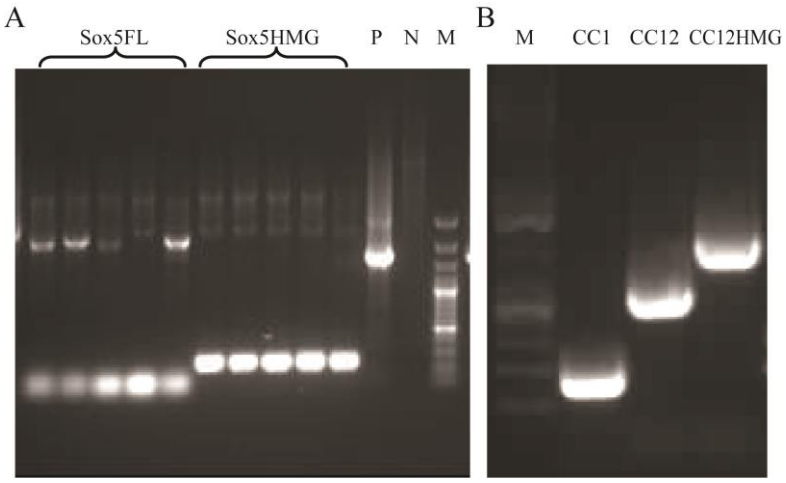


Figure 8.4 PCR amplification of full-length and different domains of mSox5 gene using cDNA clone as template. Lanes labeled "Sox5FL" shows Sox5 Full length (2152 kb); Lanes labeled "Sox5HMG" shows amplified product of Sox5HMG domain (244 kb); Lane labeled "P" shows amplified product of Positive Control (Sox9, 1528 kb) ; Lane labeled "N" shows Negative Control;

Lane labeled "M" shows GeneRuler DNA ladder; Lane labeled "CC1" shows amplified product of coiled-coil domain 1 (274 kb) ; Lane labeled CC12 shows amplified product of coiled-coil domain 1 & 2 (751 kb); Lane labeled CC12HMG shows amplified product of coiled-coil domain 1 & 2 and HMG domain (1186 kb). 10 µl of DNA sample was loaded on a 1 % agarose gel. The exact gene sequences of these constructs were confirmed by DNA sequencing.

8.3 Over-expression and purification

Of the two constructs, the fullength did not express in any of the five different tagged expression vectors. The truncated construct (CC1+CC2+HMG) over-expressed as soluble protein only in pDEST -HisMBP. There was no expression in pETG60A with NUS A tag and all other tags yielded insoluble protein (*Fig 8.5*).

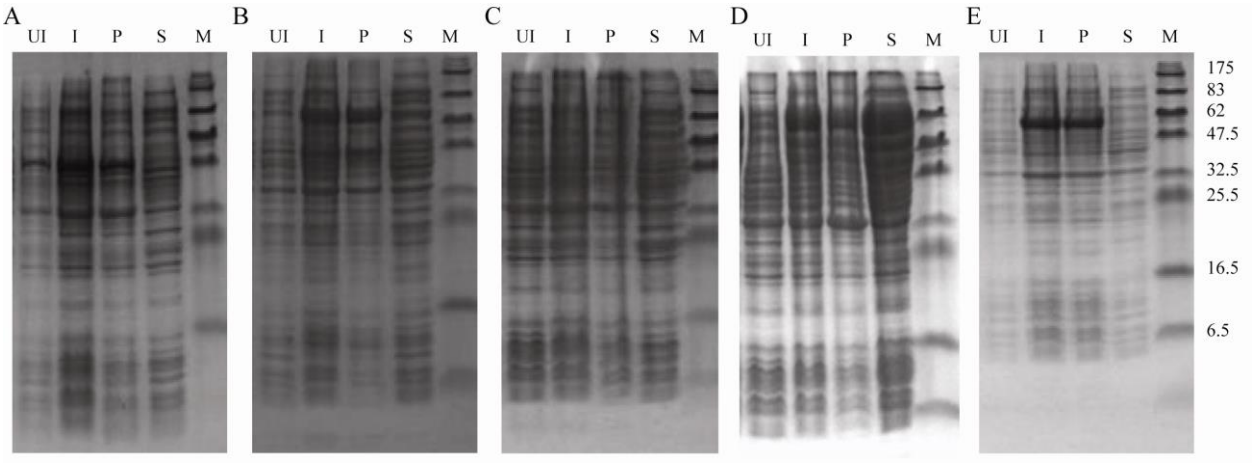


Figure 8.5 Expression of truncated Sox5. Proteins were analyzed on 12% SDS-PAGE and stained with Coomassie blue. UI:Un-Indused; I:Indused; P:Pellet; S:Supernatant; M:Molecular-weight markers (kDa). a-Histag, b-thioredoxin,c-GSTtag, d- HISMBP,e- NUSA

The pDEST -HisMBP harboring truncated Sox5 (CC1+CC2+HMG) was transformed into BL21 (DE3) and purified by Ni-Sepharose affinity chromatography (*Fig 8.5D*). The purified HisMBP tagged truncated Sox5, was cleaved by TEV digestion and purified further by ion-exchange chromatography (*Fig 8.6*). The proteins were purified to homogeneity by size exclusion chromatography, pooled and concentrated to 5-10mg/ml (*Fig 8.7*).

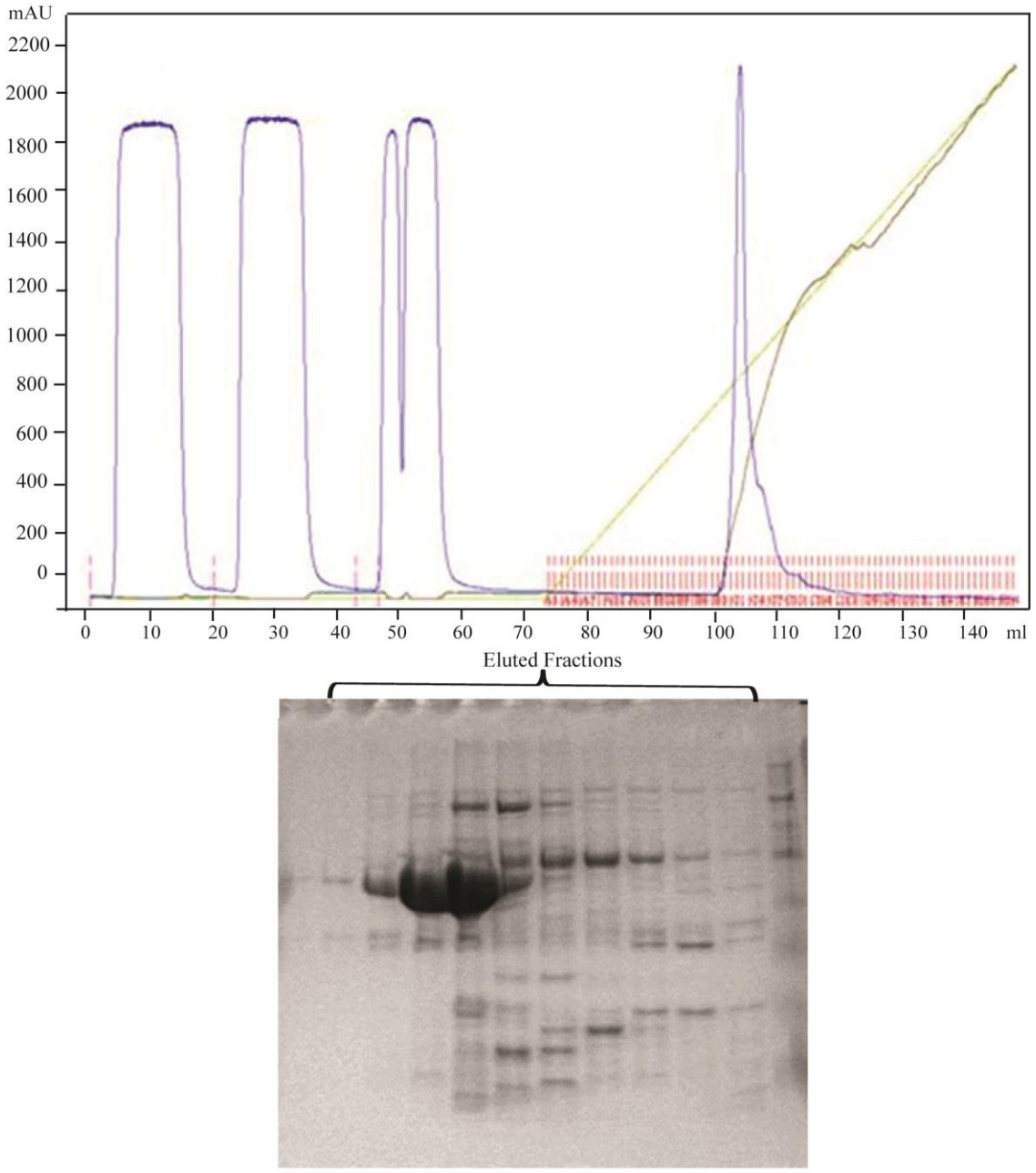


Figure 8.6 Ion-exchange chromatography profile and SDS PAGE analysis of eluted fractions of Sox5CC12HMG. Resource S **(**GE Healthcare) cation-exchange chromatography profile showing unbound protein and the elusion of protein of interest peak with 100 mM to1 M NaCl gradient (light green line). The purity of the eluted protein peak was checked by loading the fractions in 10% SDS-PAGE. The protein of interest was further subjected to size exclusion chromatography to obtain homogenously purified protein.

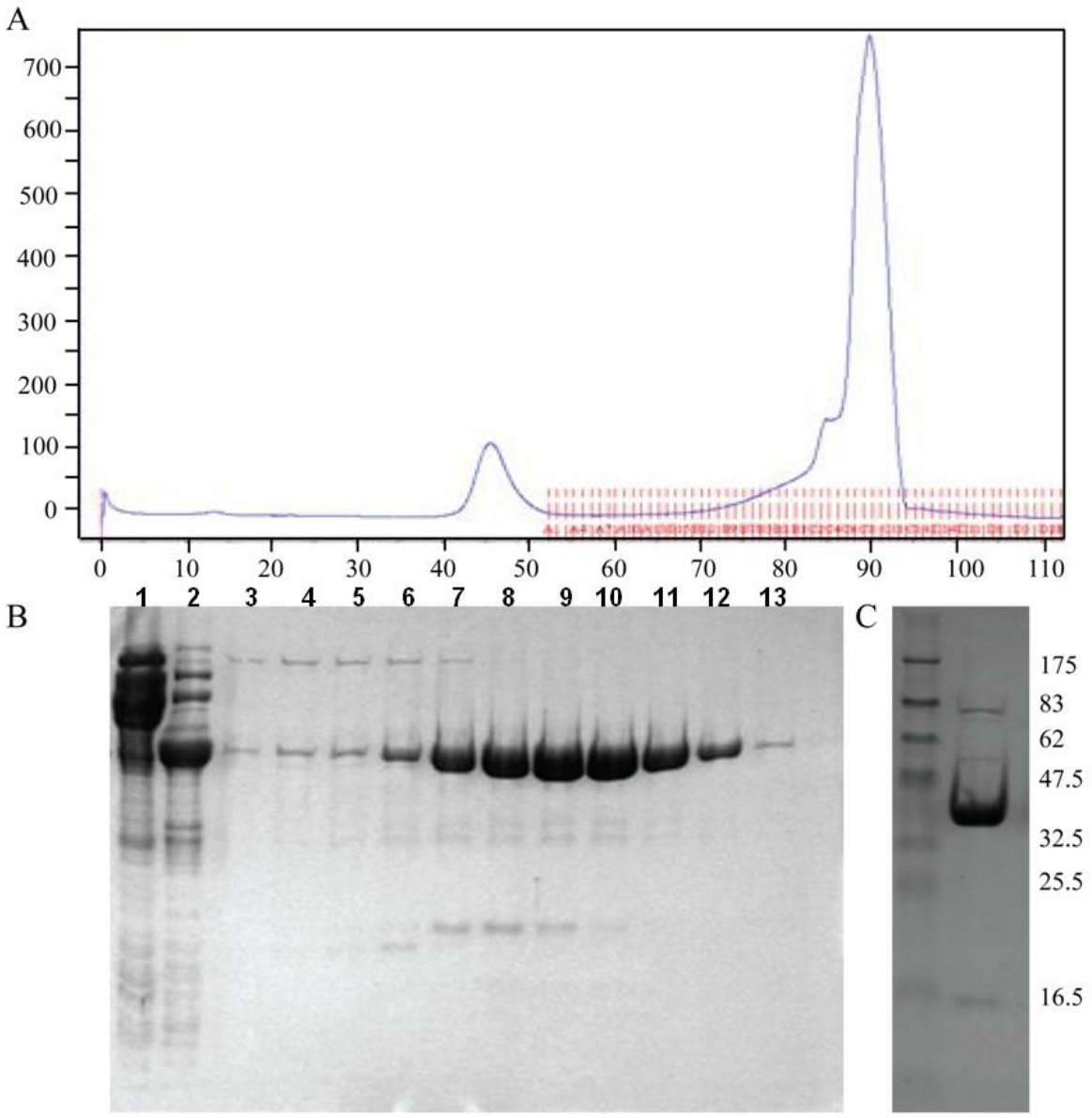


Figure 8.7 Size exclusion chromatography purification profile of Sox5CC12HMG. A. Size exclusion chromatography (Superdex 200 pg, HiLoad 16/60) profile of Sox5CC12HMG protein. B. 10% SDS-PAGE showing the size exclusion chromatography eluted fractions of the protein, Sox5CC12HMG. Lane1: Uncleaved protein; Lane2: TEV protease cleaved Ion-exchange purified fraction. Lane3-13: gel filtration eluted fraction. C. Pooled and concentrated fractions of the protein and molecular weight marker.

8.4 Secondary structure analysis of Sox5 truncated (CC12HMG)

In order to verify the purified truncated protein (Sox5CC12HMG) was well folded and retained its native structure, circular dichroism (CD) analysis was performed as mentioned in Materials and Methods. The purified Sox5CC12HMG showed typical alpha helical structure with single positive maxima at 195nm and two negative minima at 208 and

222 nm (*Fig 8.8*), in agreement with the PSIPRED predicted secondary structure (*Fig 8.1*), indicating that the cleaved and purified protein was suitable for further characterization.

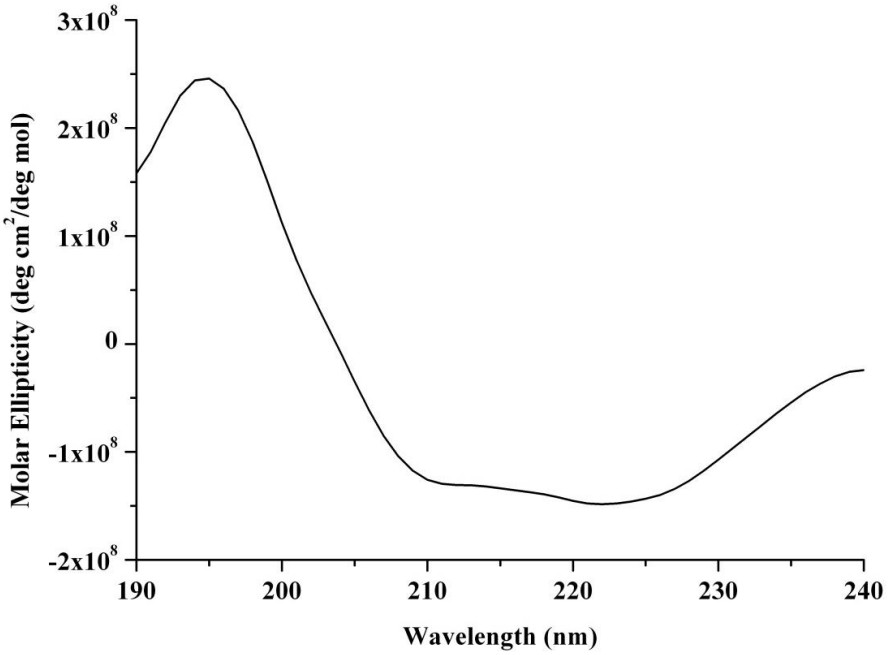


Figure 8.8 Secondary structure analysis by circular dichroism spectroscopy of purified truncated Sox5 protein measured at 50 µM concentration at 25 ºC.

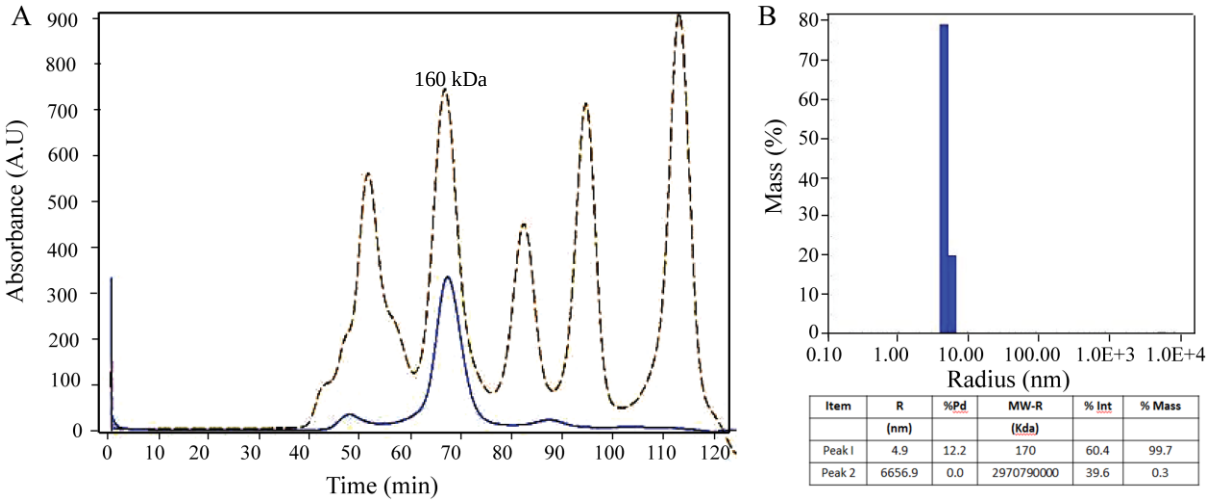
8.5 Oligomeric status of truncated Sox5

SDS- PAGE analysis of the homogenously purified truncated protein migrated as 43.8 kDa monomer (*Fig 8.7C*) as expected. Interestingly, in the size exclusion chromatography it eluted as approximately 175kDa, corresponding to possible tetramer formation (*Fig 8.9A*). Although, coiled-coil domains are well known to mediate homo or heterodimerization, the possibility of tetramerization was rather interesting. To substantiate, DLS confirmed a homogenous species with the molecular mass of approx 140kDa (*Fig 8.9B*).

8.6 The truncated constructs of Sox5

The tetramerization of Sox5 urged futher to understand the precise role of individual structural domains and consequently more constructs with different combination of its structural domains were cloned. Constructs with first coiled-coil domain alone (CC1), second coiled-coil domain alone (CC2), only coiled-coil domains (CC1+CC2) and second coiled-coil

domain with HMG domain (CC2+HMG) were amplified using the appropriate primers (as listed on table 2.2.1) from cDNA clone (IMAGE:40047865). Amplified PCR products were cloned into pENTR™/ TEV/D-TOPO® (Invitrogen) cloning vector and cloned products were



Item	R	%Pd	MW-R	% Int	% Mass
	(nm)		(Kda)		
Peak I	4.9	12.2	170	60.4	99.7
Peak 2	6656.9	0.0	2970790000	39.6	0.3

Figure 8.9 Oligomeric status of Sox5CC12HMG. Panel A Gel filtration chromatogram (Superdex 200 pg, HiLoad 16/60) of Sox5CC12HMG protein. The overlaid purified protein Peak seen corresponding to 160 kDa of molecular weight marker (doted line). Panel B. FPLC fraction was subjected to DLS.

confirmed by colony PCR and sequencing with gene specific primers (*Fig 8.4B*). Sequencing results authenticated the presence of inserts and their right orientation. These positive clones were used as the entry clone for Gateway cloning into 5 different gateway destination vectors pDEST 17 (HIS tag), pDEST 565 (HIS+GST), pETG20A (HIS+Trx), pETG60A (HIS+Nus A) and pDEST -HIS-MBP (HIS+MBP) by performing a Gateway LR reaction as mentioned earlier. PCR using gene specific primers confirmed the presence of gene of interest.

8.7 Over-expression and purification of truncated variants of Sox5

All the positive constructs were transformed into BL21 (DE3) and the protein expression was checked by SDS-PAGE. Constructs with both the coiled-coil domains (CC1+CC2) and second coiled-coil domain (CC2) did not express in any of the five different tags, while all other constructs yielded better protein expression. The proteins were expressed, purified by Ni-NTA affinity chromatography, cleaved by TEV digestion and

purified further by ion-exchange chromatography. The proteins were purified to homogeneity by size exclusion chromatography, pooled and concentrated to 5-10mg/ml (*Fig 8.10*).

8.8 Oligomeric analysis of truncated variants of Sox5

The homogenously purified coiled-coil domain 1 (CC1) migrated as 10.5 kDa (*Fig 8.10A*) band in SDS- PAGE, while eluted in size exclusion chromatography (Superdex 75 pg, HiLoad 16/60) at the molecular weight of 21 kDa (*Fig 8.11A*) corresponding to monomer and dimers respectively. The CC2HMG protein migrated as 21.5 kDa (*Fig 8.10B*) and gel filtration chromatography (Superdex 75 pg, HiLoad 16/60) analysis indicated as 43 kDa (*Fig 8.11C*). DLS confirmed a homogenous species consistent with the gel filtration profile (*Fig 8.11B & D*).

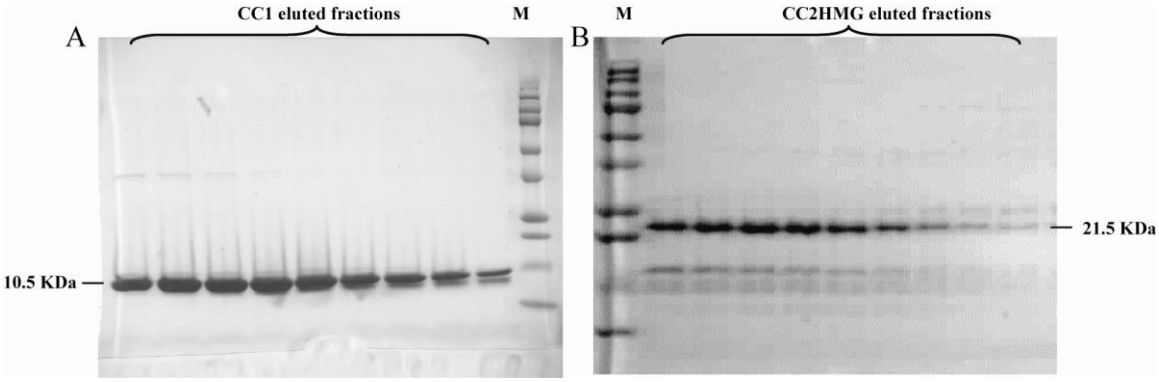


Figure 8.10 Over-expression and purification of truncated variants of Sox5 protein. Gel filtration (Superdex 75 pg, HiLoad 16/60) purified peak fractions of (A) Sox5CC1 and (B) Sox5CC2HMG protein are loaded in to 12% SDS-PAGE to check the purity and monomeric state of the protein. Sox5CC1 protein corresponds to 10.5 kDa and Sox5CC2HMG protein corresponds to 21.5 kDa. M - Molecular weight markers.

8.9 DNA binding anlysis of truncated Sox5

The DNA binding ability of the tetrameric truncated Sox5 interactions was analysed by EMSA. Sox5 is characterized by the presence of coiled-coil domain that favors homo or hetero dimerization and due to the fact that Sox5 has a higher binding affinity for HMG dimer motif than for a single, target gene's have pairs of complementary binding sequence in

their promoters [82, 163]. To analyse the binding affinity of Sox5 with two binding sites, EMSA experiment was performed with the promoter element from EY-Globin gene **(5' CAGAACAAA GGGTCAGAACATTGTCTGC 3'**) having two consensus sequences. The purified Sox5CC12HMG, Sox5CC2HMG and Sox5HMG proteins were individually incubated with synthesized Cy5 labeled double stranded DNA and EMSA buffer as referred in Materials and Methods. The concentration of DNA was maintained as 1nM throughout the studies with increasing concentration of protein. The EMSA experiment with Sox5HMG shows the population of dsDNA to decrease with increasing concentration of protein indicating the binding of dsDNA to protein resulting in two higher migrating "shifted" and "super shifted" bands till 10nM of protein (*Fig 8.12A*). Increasing the protein concentration further shows higher migrating bands and increasing further, resulted in disturbed migration due to non-specific binding of protein with DNA. However, Sox5CC2HMG exhibits a single higher migrating band above the free DNA till 300nM (*Fig 8.12B).*

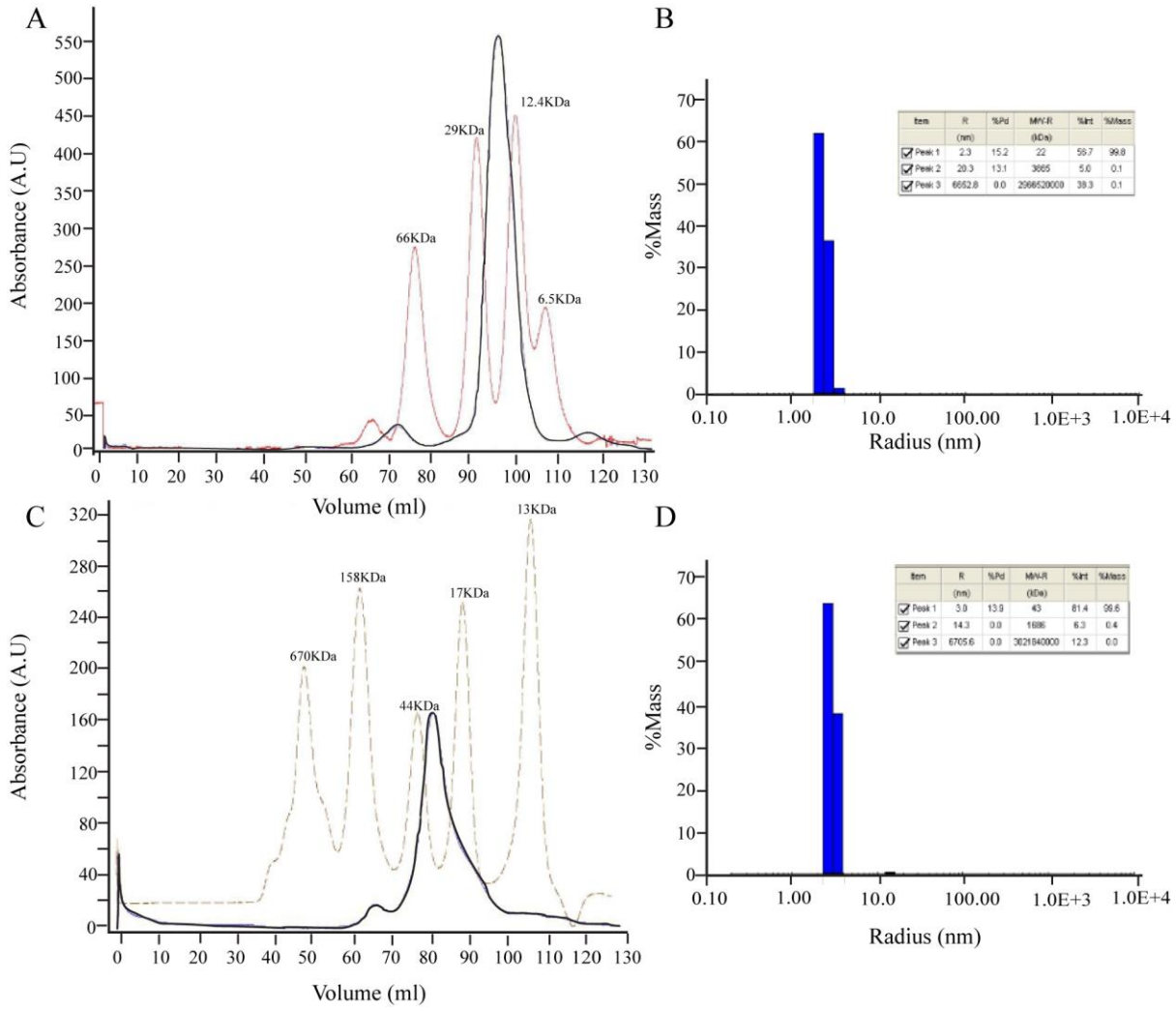


Figure 8.11 Oligomeric states of Sox5CC1 and Sox5CC2HMG. Panel A & C. Gel filtration chromatogram, (Superdex 75 pg, HiLoad 16/60) of Sox5CC1 purified protein corresponding to 21 kDa and Sox5CC2HMG protein corresponding to 42 kDa of molecular weight marker (dotted line) respectively. Panel B & D. Dynamic light scattering analysis confirming the molecular weights of the purified Sox5CC1 and Sox5CC2HMG respectively.

The protein DNA interaction experiment with Sox5CC12HMG through EMSA was not successful as it exhibited distorted binding. The electrophoretic mobility of a protein-nucleic acid complex in EMSA depends on the size of the protein DNA complex. The truncated Sox5 as a tetramer orresponds to a molecular mass of 172 kDa and upon binding to DNA forms a huge complex. Therefore it is a limitation since the migration hindered in native gel. The DNA bend caused by Sox protein further hinders the smooth migration of the

protein-DNA complex in the gel. Consequently, the DNA binding ability of truncated Sox5 was analysed employing fluorescence anisotropy experiment with FAM labeled oligos.

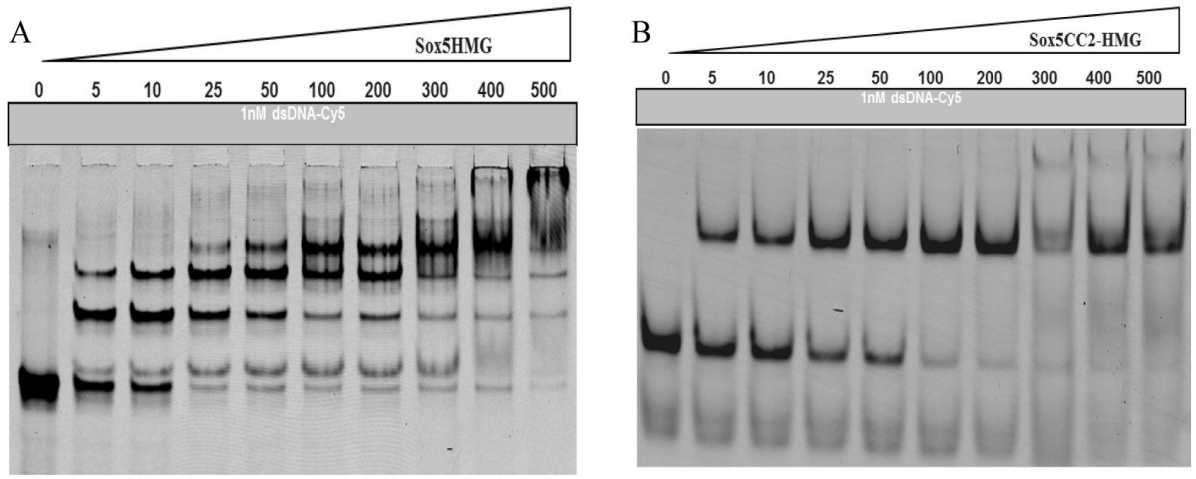


Figure 8.12 DNA binding analysis of truncated variants of Sox5 protein. EMSA experiment was performed by individually incubating Cy5 labeled EY-Globin gene (5' CAGAACAAAGGGTCAG AACATTGTCTGC 3') with (A) Sox5HMG (B) Sox5CC2HMG protein. The concentration of DNA was maintained as 1 nM with increasing concentration of protein from 0 to 500 nM. 10 µl of samples were loaded onto 12% native PAGE and bands were detected using a Typhoon 9140 Phosphor Imager.

8.10 Fluorescence Anisotropy: protein-DNA complex formation

Fluorescence anisotropy is based on the property of fluorescent molecules to emit polarized light. In a dilute solution, small rapidly tumbling flurophores reflect polarized light to greater extent (low anisotropy) than larger, slowly tumbling flurophores (high anisotropy). This principle is exploited in fluorescence anisotropy to study protein DNA interaction, as the complex formed is larger and tumbles more slowly, reflecting light to a lesser extent (high anisotropy) than the unbound nucleotide. The truncated Sox5 showed very low fluorescence anistopic values even at high concentrations substantiating the weak DNA binding affinity observed in EMSA (*Fig 8.13*). Sox5HMG domain-DNA interaction was used as control.

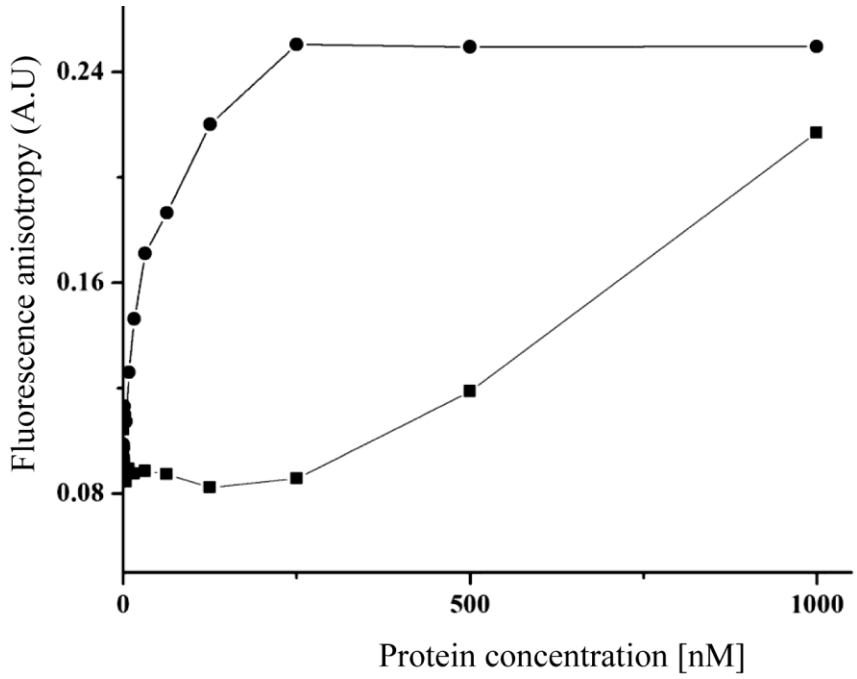


Figure 8.13 Fluorescence Anisotropy profile of Sox5HMG (closed circles) and truncated Sox5CC12HMG (closed squares). Fluorescence anisotropy experiment with FAM labeled EY-Globin gene (5' CAGAACAAAGGGTCAGAACATTGTCTGC 3') was used for the Sox5CC12HMG and Sox5HMG protein. The concentration of DNA was maintained as 1 nM with increasing concentration of protein from 0 to 1000 nM. The fluorescence anisotropy measurements from the microplates are read on a Spectramax M5 microplate reader (Molecular Devices) with excitation at 485 nm, emission at 525 nm and a cut-off filter of 515 nm.

8.11 Thermal stability: ThermoFluor Assay

Apart from homogeneity, stability of a protein directly co-relates to the probability of quality protein crystal formation [164, 165]. Protein stability depends on a number of variables such as pH, temperature, ionic strength, buffer, additives etc [158, 159]. Stability of the truncated Sox5 was studied by ThermoFluor assay. Three-state thermal denaturation was observed, with the first Tm around 57° and second Tm around 70° (*Fig 8.14A*). Influence of ionic strength on thermal stability was studied with increasing concentration of salt. At low salt concentration (<200mM NaCl) the Tm was lowered to 56° and 68.5°. Increase in salt concentration (>200mM NaCl) drifted the denaturation from three state to two state with a

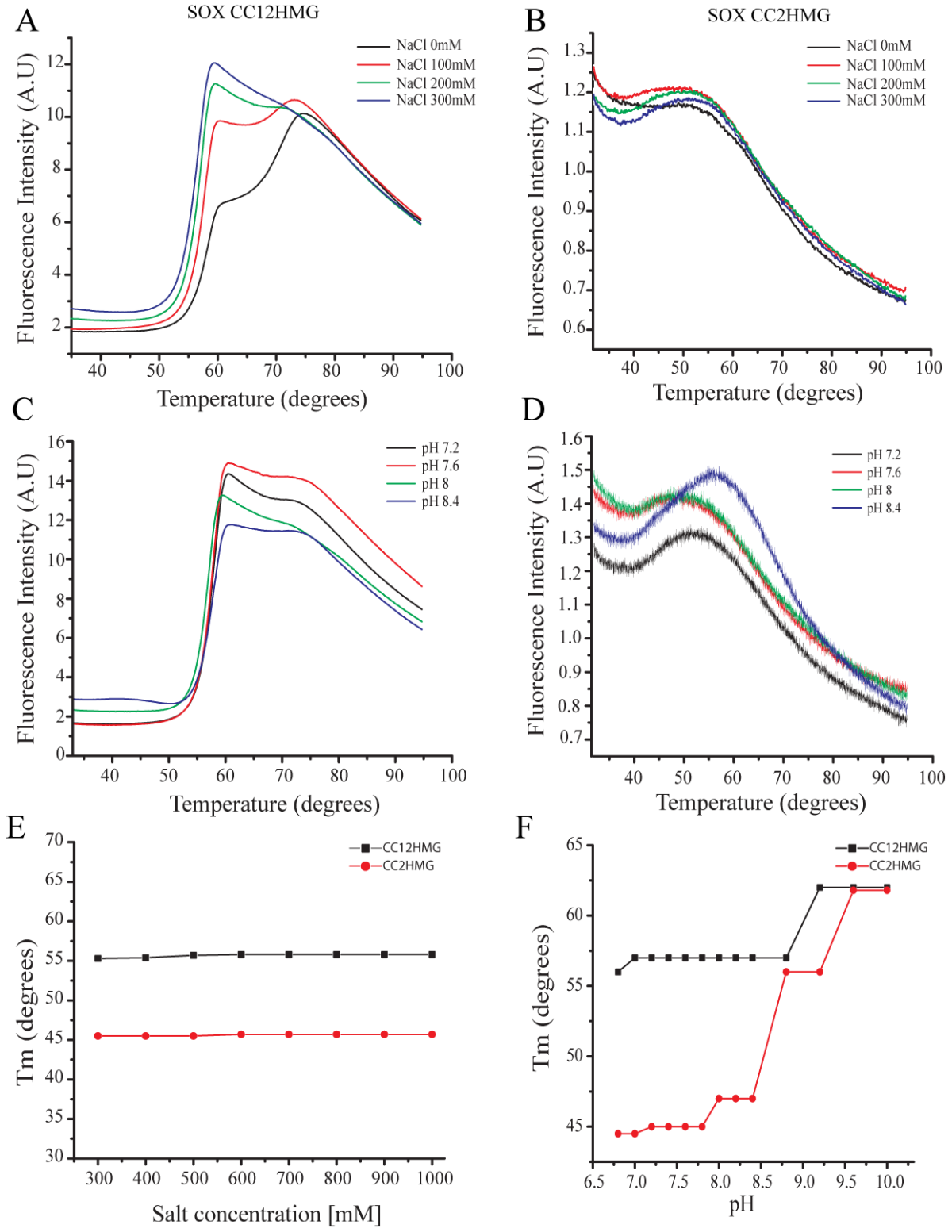


Figure 8.14 Thermal stability analysis of the tetrameric truncated Sox5 and the dimeric SoxCC2HMG proteins. Panels A and C show the influence of salt and pH on the thermal stability of Sox5 and Panels B and C for dimeric Sox CC2HMG respectively. Panels E and F show the transition of melting temperatures for Sox5 and dimeric Sox CC2HMG as a function of salt concentration and pH.

Tm lowered to 54° (*Fig 8.14A*). Influence of pH was analysed with buffers from acidic to basic range with an interval of 0.2. Absence of a well-defined melting curve at acidic pH obviously indicated complete protein destabilization. However, with increase in pH (pH 4.8) a poorly defined two step transition was observed in pH range between 5.2 -6.8 (*Fig 8.14B*). Specifically, at neutral and slightly basic pH, 7.2-8.8, there is a well-defined denaturation curve with sharp shift of Tm (56°) that shifts to very high Tm of 63° at completely basic pH (*Fig 8.14F*).

On the contrary, two state thermal denaturation was observed for CC2HMG with Tm 44.5° unaltered at low salt concentrations (*Fig 8.14B*). Increase in salt with Tm to 45° did not influence the denaturation of SoxCC2HMG (*Fig 8.14E*). Similar to the truncated Sox5, the dimeric SoxCC2HMG was non-stable at acidic pH and with increase in pH to neutral conditions the protein was highly stable with Tm 45° (*Fig 8.14D*). Further, increase in pH above neutral conditions (pH 8-8.5) gradually drifts the Tm to 46° and with drastic shift to 56° at basic pH. At completely basic pH the Tm is 63°, very high and similar to that of the truncated Sox5 (*Fig 8.14F*).

8.12 Crystallization of truncated Sox5

Crystallization trials were carried out exactly as mentioned earlier. Initial levels of screening resulted either in precipitation or clear drops. Based on the thermal stability assay, the protein was purified in a lower ionic strength buffer (50mM Tris, 50mM NaCl pH.8.0) to obtain crystals worth structural analysis. Crystals were obtained in

1. 150mM Malic acid pH7.0, 20% w/v PEG3350
2. 200mM Sodium Malonate, pH.7.0 20% w/v PEG3350
3. 100mM Magnesium formate, 15% w/v PEG3350
4. 100mM $AmSo_4$, 100mM Bis-Tris pH5.5, 17% w/v PEG 10K

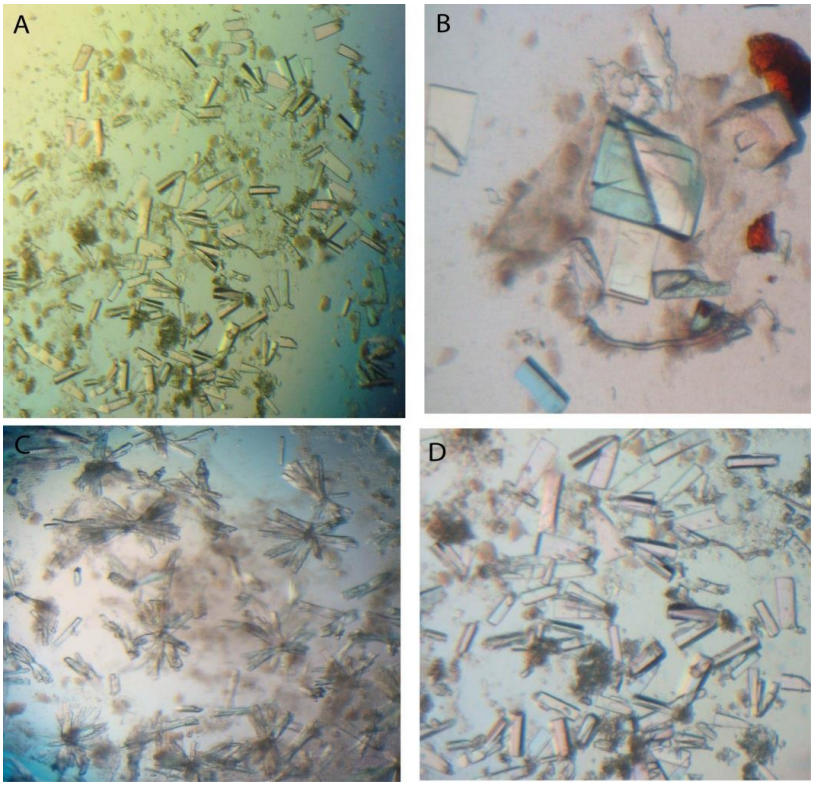


Figure 8.15 Crystal obtained from truncated Sox5 diffracted at 3.8-3.2Å in conditions Panel A 150mM Malic acid pH7.0, 20% w/v PEG3350; Panel B 200mM Sodium Malonate, pH.7.0 20% w/v PEG3350; Panel C 100mM Magnesium formate, 15% w/v PEG3350 and Panel D 100mM $AmSo_4$, 100mM Bis-Tris pH5.5, 17% w/v PEG 10K.

Although the crystals were well grown and mountable, in condition 1, they were fragile. Micro crystals were observed in condition 4 and other two conditions were promising (*Fig 8.15*). Further optimization by varying precipitant percentage and buffer composition favored better crystal formation. Following optimistaion crystals obtained in 200mM Sodium malonate (*Fig 8.15B*), 15% PEG3350 diffracted at 6-4.5Å while crystals obtained in 150mM Magnesium formate, 10-15% PEG3350 diffracted at 3.8-3.2Å (*Fig 8.15C & Fig 8.16*). X-ray diffraction analysis of these crystals indicates that this crystal form belongs to an orthorhombic space group $P2_12_12_1$, with unit-cell parameters a = 36.5, b = 118.4, c = 268.3 Å with the 99.9% completeness.

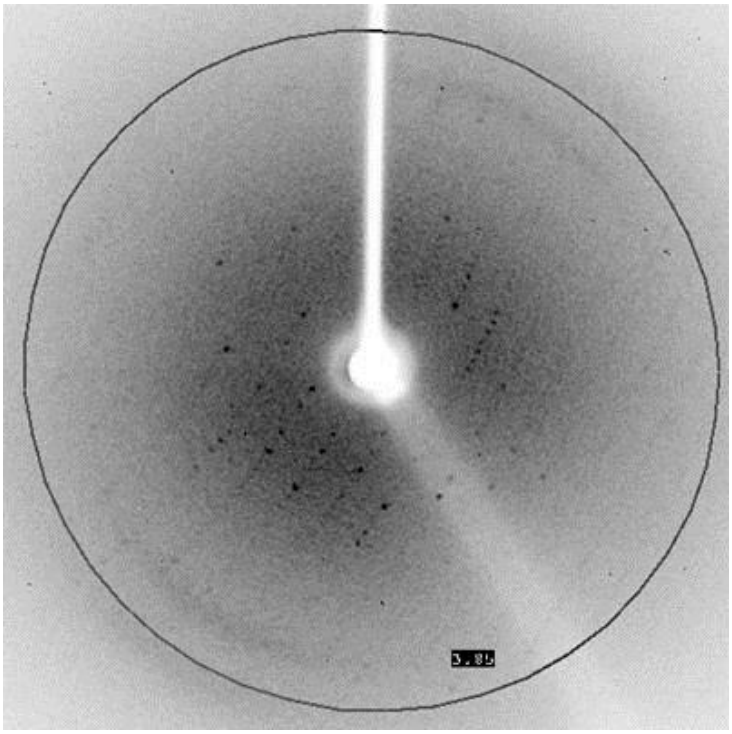


Figure 8.16 Crystal diffraction image shows the pattern of diffraction spots which diffracted 3.8 Å and the crystals belongs to orthorhombic space group $P2_12_12_1$,

8.13 Discussion

Protein-protein interactions arbitrate the association of supra molecular complexes in all vital biological processes of transcription, translation, development, and intercellular communication. The coiled-coil is a prevalent oligomerization domain facilitating specific protein-protein interactions for the assembly of multiprotein complexes. Coiled coils (CC) are bundles of left-handed α- helices, that wrap around each other, possessing motifs with hydrophobic (H) residues alternately spaced three and four residues apart, disconnected by polar (P) residues in the classical heptad repeat pattern $(HPPHPPP)_{n\geq3}$, often denoted as "abcdefg". Over the past few decades extensive studies have underlined the importance of specificity and stability of CC domains in oligomerization. The PV hypothesis of Alber and coworkers [166] outlines three vital guidelines for the formation of specific coiled-coils. 1. The 'a' and'd' positions must be hydrophobic (e.g. leucine, valine, or isoleucine), stabilizing

helix dimerization (or helical interface) through hydrophobic and van der Waals interactions. 2. Residues e and g must be charged (e.g. glutamate or lysine) in order to form interhelical electrostatic interactions. 3. The remaining three positions (b, c, and f) must all be hydrophilic, forming solvent exposed helical surfaces. The interacting surface of the "classical" leucine zipper CC domains with "3-4" or "4-3" heptad repeats are formed by complementary "knobs- into- holes" packing of the a and d side chains (*Fig 1.14*). Deviations from these rule determines the orientation, specificity, and oligomerization state of unique CC domains with novel functions ([167]).

The 33 residue GCN4 leucine zipper CC domain, has been extensively studied as a model system to understand the structural specificity of coiled-coil domains. Lu and co workers have demonstrated that the presence of non polar amino acids at either of the normally charged e and g positions with "3-3-1" heptad repeats of the dimeric GCN4 leucine zipper can direct the formation of stable, four stranded coiled-coils with combined "knobs-against- knobes" and "knobs- into- holes" packing of the three hydrophobic side chains. Similarly, crystal structure of the stable, antiparallel heterotetrameric GCN4 variants with valine substituted at either e or g position reveal differences in the hydrophobic interfaces of hetero versus homotetramers. Likewise, positional preferences of apolar leucine and isoleucine residues at a and d positions, their crucial role in dictating the stochiometry of parallel coiled-coil dimers have been determined as well [168].

As discussed earlier (*Section 1.13*), the Sox D group of proteins are unique among the Sox transcription factors with a protein interaction domain comprising two CC domains, lack the classical basic region of HLH and BZIP transcription factors and encompass a unique Q box domain within the protein interaction domain. The significance of the absence of a basic region and presence of a Q box in the SoxD transcription factors is unclear. Similarly, the effect of the coiled-coil domain mediated oligomerisation on the HMG DNA binding affinity

is yet to be understood. The current study is the first comprehensive qualitative validation of the CC domain mediated oligomerisation of the SoxD transcription factors, thermal stabilities, influence of pH and salt on the stability of the oligomers and their effect on DNA binding with Sox5 as the prototype. As the objective was characterization of Sox5 coiled-coil domains and its DNA binding domain, we utilized the fragment between residues 191-584 for the study.

The occurrence of coiled-coil α-helices in Sox5 sequence was predicted by the MultiCoil, a program that predicts the relative frequency of amino acid residues at each position of the heptad repeat (Appendix A). With respect to the truncated Sox5 construct, two regions of varying lengths, residues between 191-280 with 11 heptad repeats and a much shorter second CC domain between 400-440 with 5 heptad repeats were predicted with high propensity for coiled-coil helix formation. A closer examination of the primary sequence and helical wheel projection of the Sox5 heptad repeats indicate deviations in the presence of residues at hydrophobic 'a-d' and charged 'e-g' positions (*Fig 8.17*). Strikingly, charged glutamine and lysine residues occur at the hydrophobic core forming 'a-d' positions and hydrophobic alanine, isoleucine and leucine residues at 'e-g' positions of both CC1 and CC2 domains (*Fig 8.18*). Moreover, sequence alignment of the two coiled-coil comprising Sox D proteins, Sox 5 and Sox 6 reveals close sequence similarity only for the N terminal CC1 domain, whereas the CC2 appears to be non-conserved (*Appendix*). Therefore it is presumed that the CC1 and CC2 domains might plausibly affect oligomeric specificity differentially.

The CD analysis of the truncated Sox5 was carried out to further validate the helical structure of the coiled-coil. The $[\theta]_{222}/[\theta]_{208}$ ratio in the CD spectrum of α-helical proteins is generally employed to distinguish interacting double stranded coiled-coil helices and non-interacting single stranded helices. Coiled coil helices (inter helical) exhibit a ratio of > 1.0 while single stranded helices (intra helical) have ratio between 0.8 and 0.9. The

difference is due to the amide bond π-π * electronic transition at 208 nm, that polarizes parallel with the helical axis and is sensitive to inter vs intra helical interactions. While the molar ellipticity band at 222 nm, corresponds to the amide n– π * electronic transition, insensitive to the helical interaction remains unaltered. The CD spectra of the Sox5 fulllength with approximately equal molar ellipticities at 208 nm and 222 nm indicated a $[\theta]_{222}/[\theta]_{208}$

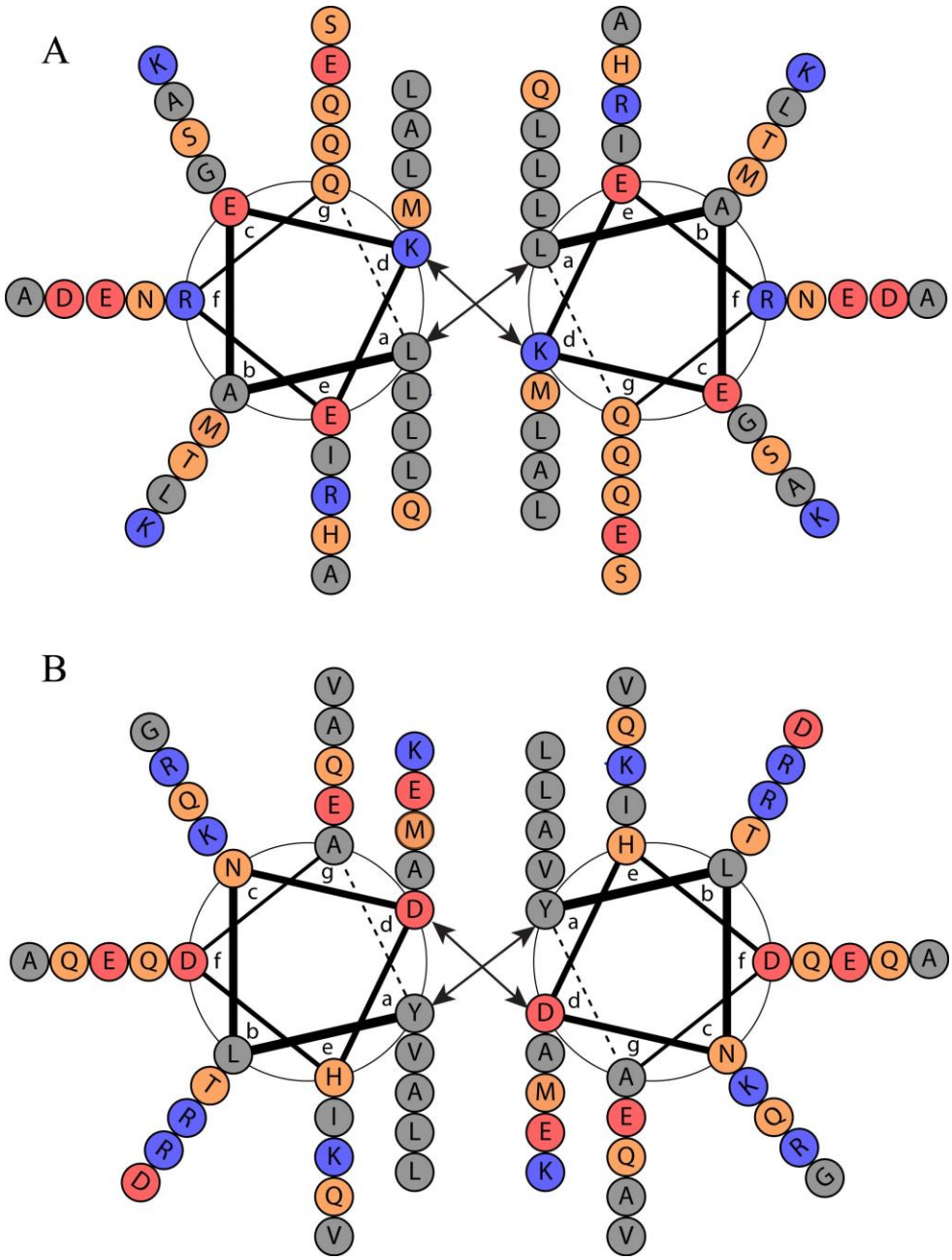


Figure 8.17 Helical wheel projection of the leucine zipper of Panel A coiled-coil domain1 and Panel B coiled-coil of Sox5. Amino acid residues are color coded; Hydrophilic-orange, hydrophobic-grey, negative-red and positive-blue.

ratio of 1.0, confirming interhelical interactions of the two CC domains. Furthermore, intrinsic tryptophan fluorescence of the full length Sox5 in the presence of increasing concentrations of helix stabilizing TFE confirms the tertiary interactions of the coiled-coil domains

As mentioned earlier, most native CC domains such as the classical GCN4 leucine zipper with the "3-4" heptad repeats exist as parallel, two stranded dimeric helices. Tetramers or trimers have been observed in modeled variants with "3-3-1" heptad repeat. Under this perspective, a dimer of the Sox5 full length of molecular mass (87 kDa) was expected upon over expression of the protein. However, a protein of molecular mass (175 Kda) corresponding to a tetramer on the size exclusion chromatography was obtained consistently (*Fig 8.9*). Interestingly, deletion of either the CC1 domain (CC2+HMG construct) or CC2 domain (CC1 construct) abolished the tetramer formation and resulted in dimers (*Fig 8.11).* The molecular masses and oligomeric states of the respective Sox5 proteins were confirmed by Dynamic light scattering measurement and Mass spectrometry. Together, it is evident that the Sox 5 oligomerises as tetramers and presence of both the CC domains is a prerequisite for the tetramerisation of Sox5.

Although the single stranded helical structure is resistant to conformational changes by mutations of aminoacids, the tertiary or quaternary structures associated by coiled-coils are highly sensitive. Apolar residue substitutions at charged e and g positions and charged substitutions at the buried hydrophobic interface have shown to affect the electrostatic/van der walls and hydrophobic forces crucial for the stability of the oligomeric hydrophobic core. Therefore the relative thermal stabilities, the influence of the pH and salt on the Sox5 truncated tetramer and the dimers, were analysed by thermofluor assays. In comparison to the dimeric CC2HMG with two state thermal unfolding ($N_2 \leftrightarrow 2U$) and a single thermal melting

point (Tm) of about 55°C, the truncated Sox5 shows three state thermal unfolding with two Tm.

The first relatively lower Tm of 57°C plausibly corresponds to dissociation of the four stranded tetramer ($N_4 \leftrightarrow 2N_2$) and a second relatively higher Tm of 78°C corresponding to the unfolding of the dimers ($2N_2 \leftrightarrow 4U$), consistent with the oligomeric status (*Fig 8.14*). Besides stabilizing hydrophobic interactions at the dimer interface, electrostatic interactions play a crucial role in stabilizing the oligomers of the GCN4 leucine zipper (O'Shea et al., 1991; Thompson et al., 1993; Krylov et al., 1994). Several reports have shown that the balance between inter helical electrostatic attractions and repulsions is important for determining coiled-coil dimerization both in terms of homo- versus heterodimerization as well as parallel versus antiparallel chain orientation[125, 169-172]. The existence of stabilizing or destabilizing electrostatic effects and their possible role in stabilizing the coiled-coil domain mediated oligomers was investigated by altering the pH and salt dependent stability of the tetramer and dimers.

A characteristic feature is observed upon addition of increased salt concentration in the case of the full length Sox5. The protein indicates three state unfolding upto a salt concentration of 0.2M. Any further addition causes the unfolding to be similar to that of the two state dimeric unfolding and the Tm remains unaffected. This clearly indicates that at salt concentrations less than 0.2M salt has a destabilising effect on the tetramer while at concentrations higher than that when the tetramer is disassociated to dimers it has stabilizing effect. This is consistent with the stabilising effect seen in the case of the dimer too (*Fig 8.14*). The difference in the salt effect could be plausibly attributed to the difference in the ion pairs of the dimers versus the tetramers that might be involved in stabiling the different oligomers. The result is consistent with a destabilizing effect of salt on the GCN4 leucine

zippers up to a concentration of 0.5M. At higher salt concentrations the effect is reversed and a stabilization of the leucine is observed [173].

The possible role of electrostatic interactions in stabilizing the coiled-coil domain mediated oligomers was investigated by altering the pH. Acidic pH displayed obscure thermal melting traces hindering Tm estimations for both the tetramer forming full length Sox5 and dimer forming CC2HMG. Increase in pH results in restoration of the three stages unfolding of the Sox5 full length. At neutral and basic pH the protein remains stable with no alterations in Tm. Similarly the dimeric CC2HMG remains stable across neutral and basic pH (*Fig 8.14*).

The DNA binding property of the different oligomers were analysed through Electrophoretic mobility shift assay and fluorescence anisotrophy experiments with the promoter element from EY-Globin gene **(5'CAGAACAAAGGGTCAGAACATTGTCTG C 3'**) which has two consensus sequences (*Fig 8.12 & Fig 8.13*). The EMSA experiment with EYG gene promoter shows two higher migrating bands as "shift" and "super-shift" up to 10 nM of Sox5HMG protein where the HMG domain freely bound to the two consensus sequence with two different population of complex, single molecule bound and two molecule bound. Further increase in the concentration of protein resulted decreases in the population of single molecule bound and increase in the population two molecules bound. We have also observed increase in non-specific DNA binding with higher migrating distorted bands. The Kd calculated as 5 nM for this Sox5HMG with EY-Globin gene. In the case of Sox5CC2HMG binds DNA as dimer exhibiting shift till 200nM and around 300 nM we observed super-shift (*Fig 8.12)*. The Kd for this Sox5CC2HMG calculated as ~10 nM. Fluorescence anisotropy experiment with FAM labeled EY-Globin oligos was used for the Sox5CC12HMG-DNA interaction. The truncated protein shows no binding up to 250 nM,

further increase in concentration of protein (resulted weak binding up to 1000nM substantiating the EMSA experiment. Contrastingly, Sox5HMG, negative control, showed higher binding affinity with DNA at initial concentration of protein (*Fig 8.13*).

On the basis of these observations and owing to the *in vivo* existence of the different truncated and coiled-coil transcripts of Sox5, we hypothesis a model for its oligomerisation and the plausible functional significance of the different oligomeric states with reduced DNA binding affinities, observed in vitro. It is presumed that the Sox5 full length homo tetramerises (dimer of dimers) through its conserved CC1 domain in a "head-to-head" fashion resulting in an anti-parallel tetramer core, with the C- terminal CC2 and HMG domain extended as two parallel coiled-coil dimer arms. It is presumed that a parallel "head-to-head" interaction through the CC1 domain might guide the hetero oligomeristaion of Sox5 and Sox6 (*Fig 8.18*) as well. There is strong *in vivo* evidence that Sox 5 and Sox 6 oligomerise functionally. NEMO, the regulatory component of the IKK complex has been proved to trimerise through a similar conserved CC2 domain.

Hypothetically, such a "head-to-head" tetramerisation would place the two dimeric HMG domains at two opposite ends, a distance beyond the spacer length required for optimal DNA binding. Moreover, the oligomerisation might confirmationally mask or hinder the efficient DNA recognising/ interacting face of the HMG domain, imparting a possibly architectural role instead of the transcriptional. Contrasting to the other basic HLH and BZIP transcription factors the pressence of the Q box and the placement of the DNA binding HMG domain at the c-terminus has been suggested to reduce the DNA binding affinities as it does not create a new DNA interacting interface. Additionally such hypothetical head to head tetramerisation may lead to further reduction in the DNA binding affinities (*Fig 8.12*).

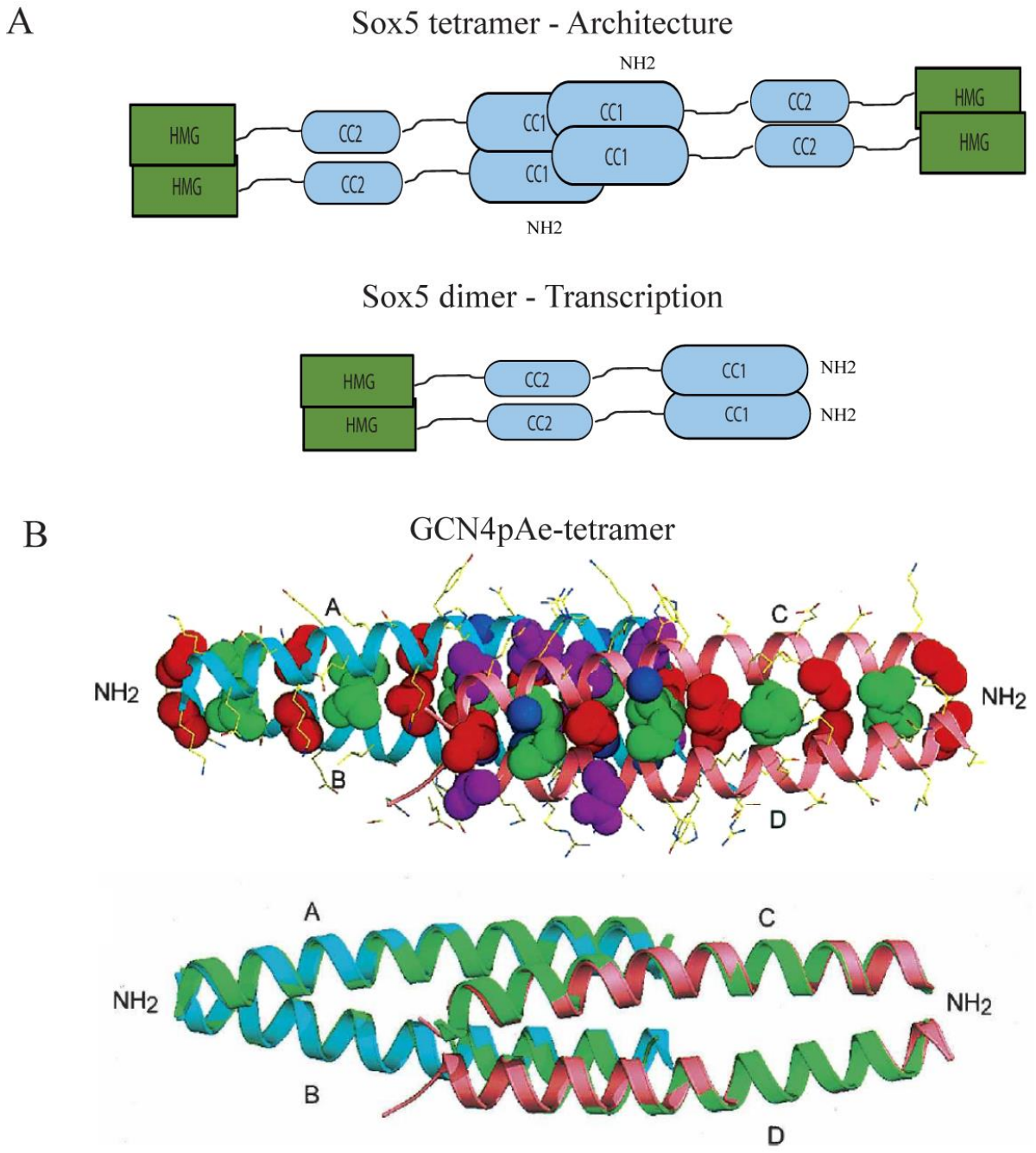


Figure 8.18 Proposed model for the tetramization of Sox5 transcription factor based on the DNA binding affinities of the terameric full length and dimeric truncated coiled-coil domain.
Panel A) Hypothetical head to head tetramerization of Sox5 would place dimeric HMG domains at opposite ends beyond spacer length required for optimal binding, imparting possible architectural role. Panel B) Crystal structure of the GCN4 alanine variants tetramerizing in tail to tail fashion.

Notably, evidence for such kind of unprecedented CC interactions have been documented in the case of GCN4 variants with alanine substitution at three charged leucine zipper 'e' positions yielding tetramers in a "tail- to- tail" fashion (*Fig 8.18B*). A stable seven-helix coiled-coil heptamer was reported upon substitution of all eight 'e' and 'g' positions with hydrophobic alanine residues indicating the extent of coiled-coil plasticity in dictating oligomerisation and the influence of the buried hydrophobic core on the structural specificity. Much has yet to be known in the case of stem cell transcription factors and Sox D in particular and the current study is a significant primary step towards this direction.

CHAPTER IX

CONCLUSION

9.1 Concluding remarks and future directions

Despite compelling evidences on the functional importance of the Sox trio proteins, Sox9, Sox5 and Sox6 in dictating core cellular developmental processes their transcriptional regulation specificities remain elusive. Their trans-activating functions and specificities are conceptualized to largely rely on specific interaction with target DNA and differential partnership with co-factors. As all Sox proteins recognize and bind the same set of consensus sequence, the functional distinctiveness of individual Sox proteins could be explained by sequence preferences of individual Sox proteins and the molecular basis underlying the sequence preference. Subsequently, existence of multiple binding motifs with varied binding affinities would largely favour.

In the current study, identification and validation of high affinity alternative sequences as direct Sox9 binding motifs, evidently suggest prescence of secondary binding motifs for the Sox9 protein specifically and for groupE Sox in general. Differences in the DNA binding affinities of Sox9 to canonical motif related endogenous sequences, varying at specific flanking or core positions indicates presumable positional interdepence in recognizing secondary binding motifs. Recently, structural evidence for the positional interdependent secondary motif recognition model came from our group, based on the crystal structures of DNA bound Sox4 and Sox17 HMG domains revealing subtle conformational rearrangements of two interface amino acids to direct altered dinucleotide preferences of Sox4 [92]. Further experiments and high resolution structural analysis of Sox9/Sox5 DNA complexes would emphasise the existence positional interdependence and secondary binding motifs as a primary mechanism of DNA recognition by Sox proteins. The observed

'secondary motif' trend has not been described earlier for the SoxE group of proteins, and has implications in understanding the DNA interactions of each subclass of Sox proteins and the respective transcriptional specificities.

Yet another means of Sox proteins functional specificity is presumed to be achieved through specific partner interactions that might control target DNA interaction and possibly alter regulatory events. Sox9, being versatile regulator of chondrogenesis and sex reversal achieves functional individuality through the DNA dependent dimerisation domain. The current study provides the prelude biochemical evidence for Sox9 dimerisation domain mediated DNA dependent homodimerisation.

Likewise, in the case of Sox5, the study for the first time provides evidence for coiled -coil domain mediated tetrameristion in Sox proteins. Loss of either of the coiled-coil domains results in dimerisation reinforcing the occurrence of a shorter transcript of Sox5. The observed reduced DNA binding affinities of the oligomers indicate potential roles for Sox5 as architectural protein. Interestingly, the coiled-coil domain mediated hetero/homo oligomerisation provides theoretically large combinatorial possibilities of transcriptional regulation. Apparently, the contrast in the mode of oligomerisation of the two proteins- DNA dependent trait of the Sox9 dimerising domain as opposed to the DNA independent coiled-coil domains of Sox5 reveals prevalance of complexity and richness in the Sox transcriptional regulation.

Future high resolution X-ray structure determination of the Sox9 dimerisation domain and the the Sox5 coiled-coil domains with HMG in the prescence of DNA would provide better insight into the effect of structural plasticity in Sox DNA recognition and the intricate mechanisms underlying the transcriptional specificity of Sox proteins.

Publications Related To This Thesis

Variable Oligomerization Modes In Stem cell Transcription Factor Sox5

Saravanan, V.[1,2] Lescar J.[1] and Kolatkar PR[2].

[1]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551
[2]Laboratory for Structural Biochemistry, Genome Institute of Singapore, Genome, 60 Biopolis Street, Singapore 138672, Singapore. (Manuscript in preparation)

Crystal Structure of DNA bound HMG Domain of the chondriogenesis Master regulator Sox9: Insight into Sox transcriptional specificity.

Saravanan, V.[1,2] Lescar J.[1] and Kolatkar PR[2].

[1]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551
[2]Laboratory for Structural Biochemistry, Genome Institute of Singapore, Genome, 60 Biopolis Street, Singapore 138672, Singapore. (Manuscript in preparation)

Sox9-directed Gene Regulatory Network Controlling Murine Embryonic Skeletal Development

Sook Peng Yap,[1] Siew Lan Lim,[1] Xing Xing,[1] Petra Kraus,[1] V Sivakamasundari,[1] Galih Kunarso,[2] Sumantra Chatterjee,[1] **Saravanan Vivekanandan,**[1] Chi Ho Lin,[2] Lee Hong Justin Tan,[1] Hsiao Yun Chan,[1] Wenqing Jean Lee,[1] Guillaume Bourque,[2] Prasanna R. Kolatkar,[1] and Thomas Lufkin[1]*

[1]Stem Cell and Developmental Biology
[2]Computational and Mathematical Biology

[3] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

(Manuscript under revision)

APPENDIX A

 # Sequence Name: mSox5 Full Length Sequence and the probability of formation coiled-coil domain

#

% Position	Residue	Reg	(Dimer,Trimer)	Probability	Dimer.Prob	TrimerProb
1	M	g	(g,g)	0.000	0.000	0.000
2	L	a	(a,a)	0.000	0.000	0.000
3	T	b	(b,b)	0.000	0.000	0.000
4	D	c	(c,c)	0.000	0.000	0.000
5	P	d	(d,d)	0.000	0.000	0.000
6	D	e	(e,e)	0.000	0.000	0.000
7	L	f	(f,f)	0.000	0.000	0.000
8	P	g	(g,g)	0.000	0.000	0.000
9	Q	a	(a,a)	0.000	0.000	0.000
10	E	b	(b,b)	0.000	0.000	0.000
11	F	c	(c,c)	0.000	0.000	0.000
12	E	d	(d,d)	0.000	0.000	0.000
13	R	e	(e,e)	0.000	0.000	0.000
14	M	f	(f,f)	0.000	0.000	0.000
15	S	g	(g,g)	0.000	0.000	0.000
16	S	a	(a,a)	0.000	0.000	0.000
17	K	b	(b,b)	0.000	0.000	0.000
18	R	c	(c,c)	0.000	0.000	0.000
19	P	d	(d,d)	0.000	0.000	0.000
20	A	e	(e,e)	0.000	0.000	0.000
21	S	f	(f,f)	0.000	0.000	0.000
22	P	g	(g,g)	0.000	0.000	0.000
23	Y	a	(a,a)	0.000	0.000	0.000
24	G	b	(b,b)	0.000	0.000	0.000
25	E	c	(c,c)	0.000	0.000	0.000
26	T	d	(d,d)	0.000	0.000	0.000
27	D	e	(e,e)	0.000	0.000	0.000
28	G	f	(f,f)	0.000	0.000	0.000
29	E	g	(g,g)	0.000	0.000	0.000
30	V	a	(a,a)	0.000	0.000	0.000
31	A	b	(b,b)	0.000	0.000	0.000
32	M	c	(c,c)	0.000	0.000	0.000
33	V	d	(d,d)	0.000	0.000	0.000
34	T	e	(e,e)	0.000	0.000	0.000
35	S	f	(f,f)	0.000	0.000	0.000
36	R	g	(g,g)	0.000	0.000	0.000
37	Q	a	(a,a)	0.000	0.000	0.000
38	K	b	(b,b)	0.000	0.000	0.000
39	V	c	(c,c)	0.000	0.000	0.000
40	E	d	(d,d)	0.000	0.000	0.000
41	E	e	(e,e)	0.000	0.000	0.000
42	E	f	(f,f)	0.000	0.000	0.000
43	E	g	(g,g)	0.000	0.000	0.000
44	S	a	(a,a)	0.000	0.000	0.000
45	E	b	(b,b)	0.000	0.000	0.000
46	R	c	(c,c)	0.000	0.000	0.000
47	L	d	(d,d)	0.000	0.000	0.000
48	P	e	(e,e)	0.000	0.000	0.000
49	A	f	(f,f)	0.000	0.000	0.000
50	F	g	(g,g)	0.000	0.000	0.000
51	H	a	(a,a)	0.000	0.000	0.000
52	L	b	(b,b)	0.000	0.000	0.000
53	P	c	(c,c)	0.000	0.000	0.000
54	L	d	(d,d)	0.000	0.000	0.000

55	H	e (e,e)	0.000	0.000	0.000
56	V	f (f,f)	0.000	0.000	0.000
57	S	g (g,g)	0.000	0.000	0.000
58	F	a (a,a)	0.000	0.000	0.000
59	P	b (b,b)	0.000	0.000	0.000
60	N	c (c,c)	0.000	0.000	0.000
61	K	d (d,d)	0.000	0.000	0.000
62	P	e (e,e)	0.000	0.000	0.000
63	H	f (f,f)	0.000	0.000	0.000
64	S	g (g,g)	0.000	0.000	0.000
65	E	a (a,a)	0.000	0.000	0.000
66	E	b (b,b)	0.000	0.000	0.000
67	F	c (c,c)	0.000	0.000	0.000
68	Q	d (d,d)	0.000	0.000	0.000
69	P	e (e,e)	0.000	0.000	0.000
70	V	f (f,f)	0.000	0.000	0.000
71	S	g (g,g)	0.000	0.000	0.000
72	L	a (a,a)	0.000	0.000	0.000
73	L	b (b,b)	0.000	0.000	0.000
74	T	c (c,c)	0.000	0.000	0.000
75	Q	d (d,d)	0.000	0.000	0.000
76	E	e (e,e)	0.000	0.000	0.000
77	T	f (f,f)	0.000	0.000	0.000
78	C	g (g,g)	0.000	0.000	0.000
79	G	a (a,a)	0.000	0.000	0.000
80	P	b (b,b)	0.000	0.000	0.000
81	R	c (c,c)	0.000	0.000	0.000
82	T	d (d,d)	0.000	0.000	0.000
83	P	e (e,e)	0.000	0.000	0.000
84	T	f (f,f)	0.000	0.000	0.000
85	V	g (g,g)	0.000	0.000	0.000
86	Q	a (a,a)	0.000	0.000	0.000
87	H	b (b,b)	0.000	0.000	0.000
88	N	c (c,c)	0.000	0.000	0.000
89	T	d (d,d)	0.000	0.000	0.000
90	M	e (e,e)	0.000	0.000	0.000
91	E	f (f,f)	0.000	0.000	0.000
92	V	g (g,g)	0.000	0.000	0.000
93	D	a (a,a)	0.000	0.000	0.000
94	G	b (b,b)	0.000	0.000	0.000
95	N	c (c,c)	0.000	0.000	0.000
96	K	d (d,d)	0.000	0.000	0.000
97	V	e (e,e)	0.000	0.000	0.000
98	M	f (f,f)	0.000	0.000	0.000
99	S	g (g,g)	0.000	0.000	0.000
100	S	a (a,a)	0.000	0.000	0.000
101	L	b (b,b)	0.000	0.000	0.000
102	A	c (c,c)	0.000	0.000	0.000
103	P	d (d,d)	0.000	0.000	0.000
104	Y	e (e,e)	0.000	0.000	0.000
105	N	f (f,f)	0.000	0.000	0.000
106	S	g (g,g)	0.000	0.000	0.000
107	S	a (a,a)	0.000	0.000	0.000
108	T	b (b,b)	0.000	0.000	0.000
109	S	c (c,c)	0.000	0.000	0.000
110	P	d (d,d)	0.000	0.000	0.000
111	Q	e (e,e)	0.000	0.000	0.000
112	K	f (f,f)	0.000	0.000	0.000
113	A	g (g,g)	0.000	0.000	0.000
114	E	a (a,a)	0.000	0.000	0.000
115	E	b (b,b)	0.000	0.000	0.000

116	G	c (c,c)	0.000	0.000	0.000
117	G	d (d,d)	0.000	0.000	0.000
118	R	e (e,e)	0.000	0.000	0.000
119	Q	f (f,f)	0.000	0.000	0.000
120	S	g (g,g)	0.000	0.000	0.000
121	G	a (a,a)	0.000	0.000	0.000
122	E	b (b,b)	0.000	0.000	0.000
123	S	c (c,c)	0.000	0.000	0.000
124	V	d (d,d)	0.000	0.000	0.000
125	S	e (e,e)	0.000	0.000	0.000
126	S	f (f,f)	0.000	0.000	0.000
127	A	g (g,g)	0.000	0.000	0.000
128	A	a (a,a)	0.000	0.000	0.000
129	L	b (b,b)	0.000	0.000	0.000
130	G	c (c,c)	0.000	0.000	0.000
131	T	d (d,d)	0.000	0.000	0.000
132	P	e (e,e)	0.000	0.000	0.000
133	E	f (f,f)	0.000	0.000	0.000
134	R	g (g,g)	0.000	0.000	0.000
135	R	a (a,a)	0.000	0.000	0.000
136	K	b (b,b)	0.000	0.000	0.000
137	G	c (c,c)	0.000	0.000	0.000
138	S	d (d,d)	0.000	0.000	0.000
139	L	e (e,e)	0.000	0.000	0.000
140	A	f (f,f)	0.000	0.000	0.000
141	D	g (g,g)	0.000	0.000	0.000
142	V	a (a,a)	0.000	0.000	0.000
143	V	b (b,b)	0.000	0.000	0.000
144	D	c (c,c)	0.000	0.000	0.000
145	T	d (d,d)	0.000	0.000	0.000
146	L	e (e,e)	0.000	0.000	0.000
147	K	f (f,f)	0.000	0.000	0.000
148	Q	g (g,g)	0.000	0.000	0.000
149	R	a (a,a)	0.000	0.000	0.000
150	K	b (b,b)	0.000	0.000	0.000
151	M	c (c,c)	0.000	0.000	0.000
152	E	d (d,d)	0.000	0.000	0.000
153	E	e (e,e)	0.000	0.000	0.000
154	L	f (f,f)	0.000	0.000	0.000
155	I	g (g,g)	0.000	0.000	0.000
156	K	a (a,a)	0.000	0.000	0.000
157	N	b (b,b)	0.000	0.000	0.000
158	E	c (c,c)	0.000	0.000	0.000
159	P	d (d,d)	0.000	0.000	0.000
160	E	e (e,e)	0.000	0.000	0.000
161	D	f (f,f)	0.000	0.000	0.000
162	T	g (g,g)	0.000	0.000	0.000
163	P	a (a,a)	0.000	0.000	0.000
164	S	e (e,e)	0.000	0.000	0.000
165	I	f (f,f)	0.000	0.000	0.000
166	E	g (g,g)	0.000	0.000	0.000
167	K	a (a,a)	0.000	0.000	0.000
168	L	b (b,b)	0.000	0.000	0.000
169	L	c (c,c)	0.000	0.000	0.000
170	S	d (d,d)	0.000	0.000	0.000
171	K	e (e,e)	0.000	0.000	0.000
172	D	f (f,f)	0.000	0.000	0.000
173	W	g (g,g)	0.000	0.000	0.000
174	K	a (a,a)	0.000	0.000	0.000
175	D	b (b,b)	0.000	0.000	0.000
176	K	c (c,c)	0.000	0.000	0.000

177	L	d (d,d)	0.000	0.000	0.000
178	L	e (e,e)	0.000	0.000	0.000
179	A	f (f,f)	0.000	0.000	0.000
180	M	g (g,g)	0.000	0.000	0.000
181	G	a (a,a)	0.000	0.000	0.000
182	S	b (b,b)	0.000	0.000	0.000
183	G	c (c,c)	0.000	0.000	0.000
184	N	d (d,d)	0.000	0.000	0.000
185	F	e (e,e)	0.000	0.000	0.000
186	G	f (f,f)	0.000	0.000	0.000
187	E	g (g,g)	0.000	0.000	0.000
188	I	a (a,a)	0.000	0.000	0.000
189	K	b (b,b)	0.000	0.000	0.000
190	G	c (c,c)	0.000	0.000	0.000
191	T	d (d,d)	0.000	0.000	0.000
192	P	e (e,e)	0.000	0.000	0.000
193	E	f (f,f)	0.286	0.093	0.193
194	S	g (g,g)	0.306	0.102	0.204
195	L	a (a,a)	0.436	0.140	0.296
196	A	b (b,b)	0.436	0.140	0.296
197	E	c (c,c)	0.436	0.140	0.296
198	K	d (d,d)	0.436	0.140	0.296
199	E	e (e,e)	0.552	0.151	0.401
200	R	f (f,f)	0.552	0.151	0.401
201	Q	g (g,g)	0.552	0.151	0.401
202	L	a (a,a)	0.573	0.167	0.406
203	M	b (b,b)	0.573	0.167	0.406
204	G	c (c,c)	0.573	0.167	0.406
205	M	d (d,d)	0.624	0.212	0.412
206	I	e (e,e)	0.771	0.391	0.380
207	N	f (f,f)	0.771	0.391	0.380
208	Q	g (g,g)	0.771	0.391	0.380
209	L	a (a,a)	0.771	0.391	0.380
210	T	b (b,b)	0.783	0.388	0.395
211	S	c (c,c)	0.783	0.388	0.395
212	L	d (d,d)	0.784	0.376	0.408
213	R	e (e,e)	0.783	0.374	0.410
214	E	f (f,f)	0.774	0.352	0.423
215	Q	g (g,g)	0.774	0.352	0.423
216	L	a (a,a)	0.774	0.352	0.423
217	L	b (b,b)	0.774	0.352	0.423
218	A	c (c,c)	0.774	0.351	0.423
219	A	d (d,d)	0.747	0.315	0.432
220	H	e (e,e)	0.747	0.315	0.432
221	D	f (f,f)	0.747	0.315	0.432
222	E	g (g,g)	0.747	0.315	0.432
223	Q	a (a,a)	0.672	0.340	0.332
224	K	b (b,b)	0.672	0.340	0.332
225	K	c (c,c)	0.673	0.342	0.331
226	L	d (d,d)	0.731	0.420	0.311
227	A	e (e,e)	0.539	0.297	0.242
228	A	f (f,e)	0.533	0.290	0.243
229	S	f (f,f)	0.603	0.387	0.216
230	Q	g (g,g)	0.677	0.505	0.171
231	I	a (a,a)	0.677	0.505	0.171
232	E	b (b,b)	0.717	0.564	0.153
233	K	c (c,c)	0.581	0.400	0.181
234	Q	d (d,d)	0.603	0.432	0.172
235	R	e (e,e)	0.603	0.432	0.172
236	Q	f (f,f)	0.603	0.432	0.172
237	Q	g (g,g)	0.603	0.432	0.172

238	M	a (a,a)	0.627	0.424	0.202
239	E	f (f,f)	0.706	0.452	0.254
240	L	g (g,g)	0.721	0.440	0.281
241	A	a (a,a)	0.725	0.448	0.277
242	K	b (b,b)	0.742	0.483	0.259
243	Q	c (c,c)	0.744	0.487	0.258
244	Q	d (d,d)	0.744	0.487	0.258
245	Q	e (e,e)	0.766	0.448	0.319
246	E	f (f,f)	0.785	0.430	0.355
247	Q	g (g,g)	0.814	0.458	0.356
248	I	a (a,a)	0.817	0.462	0.355
249	A	b (b,b)	0.817	0.462	0.355
250	R	c (c,c)	0.817	0.462	0.355
251	Q	d (d,d)	0.817	0.462	0.355
252	Q	e (e,e)	0.817	0.462	0.355
253	Q	f (f,f)	0.817	0.462	0.355
254	Q	g (g,g)	0.817	0.462	0.355
255	L	a (a,a)	0.818	0.485	0.332
256	L	b (b,b)	0.817	0.490	0.327
257	Q	c (c,c)	0.817	0.495	0.322
258	Q	d (d,d)	0.817	0.495	0.322
259	Q	e (e,e)	0.814	0.514	0.301
260	H	f (f,f)	0.814	0.514	0.301
261	K	g (g,g)	0.814	0.514	0.301
262	I	a (a,a)	0.814	0.514	0.301
263	N	b (b,b)	0.814	0.514	0.301
264	L	c (c,c)	0.814	0.514	0.301
265	L	d (d,d)	0.814	0.514	0.301
266	Q	e (e,e)	0.814	0.514	0.301
267	Q	f (f,f)	0.788	0.495	0.293
268	Q	g (g,g)	0.788	0.495	0.293
269	I	a (a,a)	0.787	0.495	0.293
270	Q	b (b,b)	0.784	0.462	0.322
271	Q	c (c,c)	0.777	0.432	0.345
272	V	d (d,d)	0.775	0.424	0.350
273	Q	e (e,e)	0.767	0.409	0.357
274	G	f (f,f)	0.667	0.388	0.279
275	Q	g (g,g)	0.663	0.385	0.278
276	L	a (a,a)	0.632	0.386	0.246
277	P	b (b,b)	0.000	0.000	0.000
278	P	c (c,c)	0.000	0.000	0.000
279	L	d (d,d)	0.000	0.000	0.000
280	M	e (e,e)	0.000	0.000	0.000
281	I	f (f,f)	0.000	0.000	0.000
282	P	g (g,g)	0.000	0.000	0.000
283	V	a (a,a)	0.000	0.000	0.000
284	F	b (b,b)	0.000	0.000	0.000
285	P	c (c,c)	0.000	0.000	0.000
286	P	d (d,d)	0.000	0.000	0.000
287	D	e (e,e)	0.000	0.000	0.000
288	Q	f (f,f)	0.000	0.000	0.000
289	R	g (g,g)	0.000	0.000	0.000
290	T	a (a,a)	0.000	0.000	0.000
291	L	b (b,b)	0.000	0.000	0.000
292	A	c (c,c)	0.000	0.000	0.000
293	A	d (d,d)	0.000	0.000	0.000
294	A	e (e,e)	0.000	0.000	0.000
295	A	f (f,f)	0.000	0.000	0.000
296	Q	g (g,g)	0.000	0.000	0.000
297	Q	a (a,a)	0.000	0.000	0.000
298	G	b (b,b)	0.000	0.000	0.000

299	F	c (c,c)	0.000	0.000	0.000
300	L	d (d,d)	0.000	0.000	0.000
301	L	e (e,e)	0.000	0.000	0.000
302	P	f (f,f)	0.000	0.000	0.000
303	P	g (g,g)	0.000	0.000	0.000
304	G	a (a,a)	0.000	0.000	0.000
305	F	b (b,b)	0.000	0.000	0.000
306	S	c (c,c)	0.000	0.000	0.000
307	Y	d (d,d)	0.000	0.000	0.000
308	K	e (e,e)	0.000	0.000	0.000
309	A	f (f,f)	0.000	0.000	0.000
310	G	g (g,g)	0.000	0.000	0.000
311	C	a (a,a)	0.000	0.000	0.000
312	S	b (b,b)	0.000	0.000	0.000
313	D	c (c,c)	0.000	0.000	0.000
314	P	d (d,d)	0.000	0.000	0.000
315	Y	e (e,e)	0.000	0.000	0.000
316	P	f (f,f)	0.000	0.000	0.000
317	V	g (g,g)	0.000	0.000	0.000
318	Q	a (a,a)	0.000	0.000	0.000
319	L	b (b,b)	0.000	0.000	0.000
320	I	c (c,c)	0.000	0.000	0.000
321	P	d (d,d)	0.000	0.000	0.000
322	T	e (e,e)	0.000	0.000	0.000
323	T	f (f,f)	0.000	0.000	0.000
324	M	g (g,g)	0.000	0.000	0.000
325	A	a (a,a)	0.000	0.000	0.000
326	A	b (b,b)	0.000	0.000	0.000
327	A	c (c,c)	0.000	0.000	0.000
328	A	d (d,d)	0.000	0.000	0.000
329	A	e (e,e)	0.000	0.000	0.000
330	A	f (f,f)	0.000	0.000	0.000
331	T	g (g,g)	0.000	0.000	0.000
332	P	a (a,a)	0.000	0.000	0.000
333	G	b (b,b)	0.000	0.000	0.000
334	L	c (c,c)	0.000	0.000	0.000
335	G	d (d,d)	0.000	0.000	0.000
336	P	e (e,e)	0.000	0.000	0.000
337	L	f (f,f)	0.000	0.000	0.000
338	Q	g (g,g)	0.000	0.000	0.000
339	L	a (a,a)	0.000	0.000	0.000
340	Q	b (b,b)	0.000	0.000	0.000
341	D	c (c,c)	0.000	0.000	0.000
342	E	d (d,d)	0.000	0.000	0.000
343	V	e (e,e)	0.000	0.000	0.000
344	A	f (f,f)	0.000	0.000	0.000
345	Q	g (g,g)	0.000	0.000	0.000
346	P	a (a,a)	0.000	0.000	0.000
347	L	b (b,b)	0.000	0.000	0.000
348	N	c (c,c)	0.000	0.000	0.000
349	L	d (d,d)	0.000	0.000	0.000
350	S	e (e,e)	0.000	0.000	0.000
351	A	f (f,f)	0.000	0.000	0.000
352	K	g (g,g)	0.000	0.000	0.000
353	P	a (a,a)	0.000	0.000	0.000
354	K	b (b,b)	0.000	0.000	0.000
355	T	c (c,c)	0.000	0.000	0.000
356	S	d (d,d)	0.000	0.000	0.000
357	D	e (e,e)	0.000	0.000	0.000
358	G	f (f,f)	0.000	0.000	0.000
359	K	g (g,g)	0.000	0.000	0.000

360	S	a (a,a)	0.000	0.000	0.000
361	P	b (b,b)	0.000	0.000	0.000
362	A	c (c,c)	0.000	0.000	0.000
363	S	d (d,d)	0.000	0.000	0.000
364	P	e (e,e)	0.000	0.000	0.000
365	T	f (f,f)	0.000	0.000	0.000
366	S	g (g,g)	0.000	0.000	0.000
367	P	a (a,a)	0.000	0.000	0.000
368	H	b (b,b)	0.000	0.000	0.000
369	M	c (c,c)	0.000	0.000	0.000
370	P	d (d,d)	0.000	0.000	0.000
371	A	e (e,e)	0.000	0.000	0.000
372	L	f (f,f)	0.000	0.000	0.000
373	R	g (g,g)	0.000	0.000	0.000
374	I	a (a,a)	0.000	0.000	0.000
375	N	b (b,b)	0.000	0.000	0.000
376	S	c (c,c)	0.000	0.000	0.000
377	G	d (d,d)	0.000	0.000	0.000
378	A	e (e,e)	0.000	0.000	0.000
379	G	f (f,f)	0.000	0.000	0.000
380	P	g (g,g)	0.000	0.000	0.000
381	L	a (a,a)	0.000	0.000	0.000
382	K	b (b,b)	0.000	0.000	0.000
383	A	c (c,c)	0.000	0.000	0.000
384	S	d (d,d)	0.000	0.000	0.000
385	V	e (e,e)	0.000	0.000	0.000
386	P	f (f,f)	0.000	0.000	0.000
387	A	g (g,g)	0.000	0.000	0.000
388	A	a (a,a)	0.000	0.000	0.000
389	L	b (b,b)	0.000	0.000	0.000
390	A	c (c,c)	0.000	0.000	0.000
391	S	d (d,d)	0.000	0.000	0.000
392	P	e (e,e)	0.000	0.000	0.000
393	S	c (c,c)	0.000	0.000	0.000
394	A	d (a,d)	0.000	0.000	0.000
395	R	e (b,e)	0.000	0.000	0.000
396	V	f (c,f)	0.001	0.000	0.001
397	S	g (d,g)	0.001	0.000	0.001
398	T	a (a,a)	0.002	0.000	0.002
399	I	b (b,b)	0.003	0.000	0.003
400	G	c (c,c)	0.010	0.000	0.010
401	Y	d (d,d)	0.020	0.001	0.019
402	L	e (e,e)	0.088	0.015	0.073
403	N	f (f,f)	0.259	0.067	0.192
404	D	g (g,g)	0.259	0.067	0.192
405	H	a (a,a)	0.351	0.064	0.287
406	D	b (b,f)	0.536	0.178	0.357
407	A	g (c,g)	0.536	0.178	0.357
408	V	a (d,a)	0.536	0.178	0.357
409	T	b (e,b)	0.536	0.178	0.357
410	K	c (f,c)	0.562	0.209	0.353
411	A	d (g,d)	0.633	0.217	0.416
412	I	e (a,e)	0.654	0.219	0.436
413	Q	f (b,f)	0.667	0.219	0.448
414	E	g (c,g)	0.667	0.219	0.448
415	A	a (d,a)	0.667	0.219	0.448
416	R	b (e,b)	0.667	0.219	0.448
417	Q	c (f,c)	0.667	0.219	0.448
418	M	d (g,d)	0.667	0.219	0.448
419	K	e (a,e)	0.667	0.219	0.448
420	E	f (b,f)	0.667	0.219	0.448

421	Q	g (c,g)	0.667	0.219	0.448
422	L	a (d,a)	0.667	0.219	0.448
423	R	b (e,b)	0.667	0.219	0.448
424	R	c (f,c)	0.667	0.219	0.448
425	E	d (g,d)	0.667	0.219	0.448
426	Q	e (a,e)	0.667	0.219	0.448
427	Q	f (b,f)	0.667	0.219	0.448
428	A	g (c,g)	0.667	0.219	0.448
429	L	a (d,a)	0.667	0.219	0.448
430	D	b (e,b)	0.667	0.219	0.448
431	G	c (f,c)	0.667	0.219	0.448
432	K	d (g,d)	0.667	0.219	0.448
433	V	e (a,e)	0.667	0.224	0.443
434	A	f (b,f)	0.652	0.184	0.468
435	V	g (c,g)	0.537	0.142	0.395
436	V	a (d,a)	0.481	0.071	0.409
437	N	b (e,b)	0.427	0.046	0.381
438	S	c (f,c)	0.422	0.039	0.383
439	I	g (g,d)	0.373	0.021	0.352
440	G	e (e,e)	0.182	0.008	0.174
441	L	f (f,f)	0.104	0.006	0.098
442	S	c (g,c)	0.102	0.007	0.096
443	N	d (e,d)	0.037	0.003	0.034
444	C	e (f,e)	0.013	0.001	0.012
445	R	f (g,f)	0.015	0.001	0.014
446	T	a (a,a)	0.015	0.001	0.014
447	E	b (b,b)	0.015	0.001	0.014
448	K	c (c,c)	0.012	0.001	0.011
449	E	d (d,d)	0.012	0.001	0.011
450	K	e (e,e)	0.006	0.000	0.006
451	T	f (f,f)	0.005	0.000	0.005
452	T	g (g,g)	0.005	0.000	0.005
453	L	a (a,a)	0.004	0.000	0.004
454	E	b (b,b)	0.004	0.000	0.004
455	S	c (c,c)	0.004	0.000	0.004
456	L	d (d,d)	0.004	0.000	0.004
457	T	e (e,e)	0.004	0.000	0.004
458	Q	f (f,f)	0.004	0.000	0.004
459	Q	g (g,g)	0.004	0.000	0.004
460	L	a (a,a)	0.004	0.000	0.004
461	A	b (b,b)	0.004	0.000	0.004
462	V	c (c,c)	0.004	0.000	0.004
463	K	d (d,d)	0.004	0.000	0.004
464	Q	e (e,e)	0.004	0.000	0.004
465	N	f (f,f)	0.003	0.000	0.003
466	E	g (g,g)	0.003	0.000	0.003
467	E	a (a,a)	0.003	0.000	0.003
468	G	b (b,b)	0.003	0.000	0.003
469	K	c (c,c)	0.002	0.000	0.002
470	F	d (d,d)	0.001	0.000	0.001
471	S	e (e,e)	0.001	0.000	0.001
472	H	f (f,f)	0.001	0.000	0.001
473	G	g (g,g)	0.001	0.000	0.001
474	M	a (a,a)	0.001	0.000	0.001
475	M	b (b,b)	0.000	0.000	0.000
476	D	c (c,c)	0.000	0.000	0.000
477	F	d (d,d)	0.000	0.000	0.000
478	N	e (e,e)	0.000	0.000	0.000
479	M	f (f,f)	0.000	0.000	0.000
480	S	g (g,g)	0.000	0.000	0.000
481	G	a (a,a)	0.000	0.000	0.000

482	D	b (b,b)	0.000	0.000	0.000
483	S	c (c,c)	0.000	0.000	0.000
484	D	e (e,e)	0.000	0.000	0.000
485	G	f (f,f)	0.000	0.000	0.000
486	S	g (g,g)	0.000	0.000	0.000
487	A	a (a,a)	0.000	0.000	0.000
488	G	b (b,b)	0.000	0.000	0.000
489	V	c (c,c)	0.000	0.000	0.000
490	S	d (d,d)	0.000	0.000	0.000
491	E	e (e,e)	0.000	0.000	0.000
492	S	f (f,f)	0.000	0.000	0.000
493	R	g (g,g)	0.000	0.000	0.000
494	I	a (a,a)	0.000	0.000	0.000
495	Y	b (b,b)	0.000	0.000	0.000
496	R	c (c,c)	0.000	0.000	0.000
497	E	d (d,d)	0.000	0.000	0.000
498	S	e (e,e)	0.000	0.000	0.000
499	R	f (f,f)	0.000	0.000	0.000
500	G	g (g,g)	0.000	0.000	0.000
501	R	a (a,a)	0.000	0.000	0.000
502	G	b (b,b)	0.000	0.000	0.000
503	S	c (c,c)	0.000	0.000	0.000
504	N	d (d,d)	0.000	0.000	0.000
505	E	e (e,e)	0.000	0.000	0.000
506	P	f (f,f)	0.000	0.000	0.000
507	H	g (g,g)	0.000	0.000	0.000
508	I	a (a,a)	0.000	0.000	0.000
509	K	b (b,b)	0.000	0.000	0.000
510	R	c (c,c)	0.000	0.000	0.000
511	P	d (d,d)	0.000	0.000	0.000
512	M	e (e,e)	0.000	0.000	0.000
513	N	f (f,f)	0.000	0.000	0.000
514	A	g (g,g)	0.000	0.000	0.000
515	F	a (a,a)	0.000	0.000	0.000
516	M	b (b,b)	0.000	0.000	0.000
517	V	c (c,c)	0.000	0.000	0.000
518	W	d (d,d)	0.000	0.000	0.000
519	A	e (e,e)	0.000	0.000	0.000
520	K	f (f,f)	0.000	0.000	0.000
521	D	g (g,g)	0.000	0.000	0.000
522	E	a (a,a)	0.000	0.000	0.000
523	R	b (b,b)	0.000	0.000	0.000
524	R	c (c,c)	0.000	0.000	0.000
525	K	d (d,d)	0.000	0.000	0.000
526	I	e (e,e)	0.000	0.000	0.000
527	L	f (f,f)	0.000	0.000	0.000
528	Q	g (g,g)	0.000	0.000	0.000
529	A	a (a,a)	0.000	0.000	0.000
530	F	b (b,b)	0.000	0.000	0.000
531	P	c (c,c)	0.000	0.000	0.000
532	D	d (d,d)	0.000	0.000	0.000
533	M	e (e,e)	0.000	0.000	0.000
534	H	f (f,f)	0.000	0.000	0.000
535	N	g (g,g)	0.000	0.000	0.000
536	S	a (a,a)	0.000	0.000	0.000
537	N	b (b,b)	0.000	0.000	0.000
538	I	c (c,c)	0.000	0.000	0.000
539	S	d (d,d)	0.000	0.000	0.000
540	K	e (e,e)	0.000	0.000	0.000
541	I	f (f,f)	0.000	0.000	0.000
542	L	g (g,g)	0.000	0.000	0.000

543	G	a (a,a)	0.000	0.000	0.000
544	S	b (b,b)	0.000	0.000	0.000
545	R	c (c,c)	0.000	0.000	0.000
546	W	d (d,d)	0.000	0.000	0.000
547	K	e (e,e)	0.000	0.000	0.000
548	A	f (f,f)	0.000	0.000	0.000
549	M	g (g,g)	0.000	0.000	0.000
550	T	a (a,a)	0.000	0.000	0.000
551	N	b (b,b)	0.000	0.000	0.000
552	L	c (c,c)	0.000	0.000	0.000
553	E	d (d,d)	0.000	0.000	0.000
554	K	e (e,e)	0.000	0.000	0.000
555	Q	f (f,f)	0.000	0.000	0.000
556	P	g (g,g)	0.000	0.000	0.000
557	Y	a (a,a)	0.000	0.000	0.000
558	Y	b (b,b)	0.000	0.000	0.000
559	E	c (c,c)	0.000	0.000	0.000
560	E	d (d,d)	0.000	0.000	0.000
561	Q	e (e,e)	0.000	0.000	0.000
562	A	f (f,f)	0.000	0.000	0.000
563	R	g (g,g)	0.000	0.000	0.000
564	L	a (a,a)	0.000	0.000	0.000
565	S	b (b,b)	0.000	0.000	0.000
566	K	c (c,c)	0.000	0.000	0.000
567	Q	d (d,d)	0.000	0.000	0.000
568	H	e (e,e)	0.000	0.000	0.000
569	L	f (f,f)	0.000	0.000	0.000
570	E	g (g,g)	0.000	0.000	0.000
571	K	a (a,a)	0.000	0.000	0.000
572	Y	b (b,b)	0.000	0.000	0.000
573	P	c (c,c)	0.000	0.000	0.000
574	D	d (d,d)	0.000	0.000	0.000
575	Y	e (e,e)	0.000	0.000	0.000
576	K	f (f,f)	0.000	0.000	0.000
577	Y	g (g,g)	0.000	0.000	0.000
578	K	a (a,a)	0.000	0.000	0.000
579	P	b (b,b)	0.000	0.000	0.000
580	R	c (c,c)	0.000	0.000	0.000
581	P	d (d,d)	0.000	0.000	0.000
582	K	c (c,c)	0.000	0.000	0.000
583	R	d (d,d)	0.000	0.000	0.000
584	T	e (e,e)	0.000	0.000	0.000
585	C	f (f,f)	0.000	0.000	0.000
586	L	g (g,g)	0.001	0.000	0.001
587	V	a (a,a)	0.001	0.000	0.001
588	D	b (b,b)	0.001	0.000	0.001
589	G	c (c,c)	0.001	0.000	0.001
590	K	d (d,d)	0.001	0.000	0.001
591	K	e (e,e)	0.001	0.000	0.001
592	L	f (f,f)	0.002	0.000	0.002
593	R	g (g,g)	0.002	0.000	0.002
594	I	a (a,a)	0.002	0.000	0.002
595	G	b (b,b)	0.002	0.000	0.002
596	E	c (c,c)	0.002	0.000	0.002
597	Y	d (d,d)	0.002	0.000	0.002
598	K	e (e,e)	0.002	0.000	0.002
599	A	f (f,f)	0.002	0.000	0.002
600	I	g (g,g)	0.002	0.000	0.002
601	M	a (a,a)	0.002	0.000	0.002
602	R	b (b,b)	0.002	0.000	0.002
603	N	c (c,c)	0.002	0.000	0.002

604	R	d (d,d)	0.002	0.000	0.002
605	R	e (e,e)	0.002	0.000	0.002
606	Q	f (f,f)	0.002	0.000	0.002
607	E	g (g,g)	0.002	0.000	0.002
608	M	a (a,a)	0.002	0.000	0.002
609	R	b (b,b)	0.002	0.000	0.002
610	Q	c (c,c)	0.002	0.000	0.002
611	Y	d (d,d)	0.002	0.000	0.002
612	F	e (e,e)	0.001	0.000	0.001
613	N	f (f,f)	0.001	0.000	0.001
614	V	g (g,g)	0.001	0.000	0.001
615	G	a (a,a)	0.001	0.000	0.001
616	Q	b (b,b)	0.001	0.000	0.001
617	Q	c (c,c)	0.001	0.000	0.001
618	A	d (d,d)	0.001	0.000	0.001
619	Q	e (e,e)	0.001	0.000	0.001
620	I	f (f,f)	0.000	0.000	0.000
621	P	g (g,g)	0.000	0.000	0.000
622	I	a (a,a)	0.000	0.000	0.000
623	A	b (b,b)	0.000	0.000	0.000
624	T	c (c,c)	0.000	0.000	0.000
625	A	d (d,d)	0.000	0.000	0.000
626	G	e (e,e)	0.000	0.000	0.000
627	V	f (f,f)	0.000	0.000	0.000
628	V	g (g,g)	0.000	0.000	0.000
629	Y	a (a,a)	0.000	0.000	0.000
630	P	b (b,b)	0.000	0.000	0.000
631	S	c (c,c)	0.000	0.000	0.000
632	A	d (d,d)	0.000	0.000	0.000
633	I	e (e,e)	0.000	0.000	0.000
634	A	f (f,f)	0.000	0.000	0.000
635	M	g (g,g)	0.000	0.000	0.000
636	A	a (a,a)	0.000	0.000	0.000
637	G	b (b,b)	0.000	0.000	0.000
638	M	c (c,c)	0.000	0.000	0.000
639	P	d (d,d)	0.000	0.000	0.000
640	S	e (e,e)	0.000	0.000	0.000
641	P	f (f,f)	0.000	0.000	0.000
642	H	g (g,g)	0.000	0.000	0.000
643	L	a (a,a)	0.000	0.000	0.000
644	P	b (b,b)	0.000	0.000	0.000
645	S	c (c,c)	0.000	0.000	0.000
646	E	d (d,d)	0.000	0.000	0.000
647	H	e (e,e)	0.000	0.000	0.000
648	S	f (f,f)	0.000	0.000	0.000
649	S	g (g,g)	0.000	0.000	0.000
650	V	a (a,a)	0.000	0.000	0.000
651	S	b (b,b)	0.000	0.000	0.000
652	S	c (c,c)	0.000	0.000	0.000
653	S	d (d,d)	0.000	0.000	0.000
654	P	e (e,e)	0.000	0.000	0.000
655	E	f (f,f)	0.000	0.000	0.000
656	P	g (g,g)	0.000	0.000	0.000
657	G	a (a,a)	0.000	0.000	0.000
658	M	b (b,b)	0.000	0.000	0.000
659	P	c (c,c)	0.000	0.000	0.000
660	V	d (d,d)	0.000	0.000	0.000
661	I	e (e,e)	0.000	0.000	0.000
662	Q	f (f,f)	0.000	0.000	0.000
663	S	g (g,g)	0.000	0.000	0.000
664	T	a (a,a)	0.000	0.000	0.000

665	Y	b (b,b)	0.000	0.000	0.000
666	G	c (c,c)	0.000	0.000	0.000
667	A	d (d,d)	0.000	0.000	0.000
668	K	e (e,e)	0.000	0.000	0.000
669	G	f (f,f)	0.000	0.000	0.000
670	E	g (g,g)	0.000	0.000	0.000
671	E	a (a,a)	0.000	0.000	0.000
672	P	b (b,b)	0.000	0.000	0.000
673	H	c (c,c)	0.000	0.000	0.000
674	I	d (d,d)	0.000	0.000	0.000
675	K	e (e,e)	0.000	0.000	0.000
676	E	f (f,f)	0.000	0.000	0.000
677	E	g (g,g)	0.000	0.000	0.000
678	I	a (a,a)	0.000	0.000	0.000
679	Q	b (b,b)	0.000	0.000	0.000
680	A	c (c,c)	0.000	0.000	0.000
681	E	d (d,d)	0.000	0.000	0.000
682	D	e (e,e)	0.000	0.000	0.000
683	I	f (f,f)	0.000	0.000	0.000
684	N	g (g,g)	0.000	0.000	0.000
685	G	a (a,a)	0.000	0.000	0.000
686	E	b (b,b)	0.000	0.000	0.000
687	I	c (c,c)	0.000	0.000	0.000
688	Y	d (d,d)	0.000	0.000	0.000
689	E	e (e,e)	0.000	0.000	0.000
690	E	f (f,f)	0.000	0.000	0.000
691	Y	g (g,g)	0.000	0.000	0.000
692	D	a (a,a)	0.000	0.000	0.000
693	E	b (b,b)	0.000	0.000	0.000
694	E	c (c,c)	0.000	0.000	0.000
695	E	d (d,d)	0.000	0.000	0.000
696	E	e (e,e)	0.000	0.000	0.000
697	D	f (f,f)	0.000	0.000	0.000
698	P	g (g,g)	0.000	0.000	0.000
699	D	a (a,a)	0.000	0.000	0.000
700	V	b (b,b)	0.000	0.000	0.000
701	D	c (c,c)	0.000	0.000	0.000
702	Y	d (d,d)	0.000	0.000	0.000
703	G	e (e,e)	0.000	0.000	0.000
704	S	f (f,f)	0.000	0.000	0.000
705	D	g (g,g)	0.000	0.000	0.000
706	S	a (a,a)	0.000	0.000	0.000
707	E	b (b,b)	0.000	0.000	0.000
708	N	c (c,c)	0.000	0.000	0.000
709	H	d (d,d)	0.000	0.000	0.000
710	I	e (e,e)	0.000	0.000	0.000
711	A	f (f,f)	0.000	0.000	0.000
712	G	g (g,g)	0.000	0.000	0.000
713	Q	a (a,a)	0.000	0.000	0.000
714	A	b (b,b)	0.000	0.000	0.000
715	N	c (c,c)	0.000	0.000	0.000

APPENDIX B

mSox5 Transcription Factor Coiled-Coil 1 Domain and its position

```
Res.191 12345678901234567890123456789012345678901234567890
Sequenc TPESLAEKERQLMGMINQLTSLREQLLAAHDEQKKLAASQIEKQRQQMEL
Coil    defgabcdefgabcdefgabcdefgabcdefgabcdefFgabcdefgaFg

Res.261 01234567890123456789123456789012345678901
Sequenc AKQQQEQIARQQQQLLQQQHKINLLQQQIQQVQGQLPPLM
Coil    abcdefgabcdefgabcdefgabcdefgabcdefgabcde
```

mSox5 Transcription Factor Coiled Coil 2 Domain and its position

```
Res  400 12345678901234567890123456789012345678901234567890
Sequence GYLNDHDAVTKAIQEARQMKEQLRREQQALDGKVAVVNSI
Coil     gabcdefgabcdefgabcdefgabcdefgabcdefgabcd
```

Appendix C

Sequence Alignment of Coiled-Coil Domain of mSox5 and mSox6

Coiled Coil Domain 1

```
Sox5Coiled Coil 1   TPESLAEKER QLMGMINQLT SLREQLLAAH DEQKKLAASQ IEKQRQQMEL
Sox6Coiled Coil 1   TPESLAEKER QLSTMITQLI SLREQLLAAH DEQKKLAASQ IEKQRQQMDL
Consensus           TPESLAEKER QLmgMInQLi SLREQLLAAH DEQKKLAASQ IEKQRQQM#L

Sox5Coiled Coil 1   AKQQQEQIAR QQQQLLQQQH KINLLQQQIQ QVQGQLPPLM
Sox6Coiled Coil 1   ARQQQEQIAR QQQQLLQQQH KINLLQQQIQ QVQGHMPPLM
Consensus           ArQQQEQIAR QQQQLLQQQH KINLLQQQIQ QVQGq$PPLM
```

Coiled Coil Domain 2

```
Sox5Coiled Coil2   GYLNDHDAVT KAIQEARQMK EQLRREQQ----ALDGKVAVVNSI
Sox6Coiled Coil2   ALFGDQDTVM KAIQEARKMR EQIQREQQQQPHGVDGKLSSMNNI
Consensus          allnDqDaVm KAIQEARqMr EQirREQQ....alDGKlasmNnI
```

APPENDIX D

>mSOX5_isoform_a_165932306_NP_035574.2

>mSOX5_isoform_b_341926306_NP_001107031.2

>mSOX5_isoform_c_341926308_NP_001230092.1

1ST Coiled-Coil Domain
2nd Coiled-Coil Domain
HMG Domain

```
           1
mSOX5_isoA  MLTDPDLPQE FERMSSKRPA SPYGETDGEV AMVTSRQKVE EEESERLPAF HLPLHVSFPN KPHSEEFQPV SLLTQETCGP RTPTVQHNTM EVDGNKVMSS LAPYNSSTSP QKAEEGGRQS GESVSSAALG
mSOX5_isoB  MLTDPDLPQE FERMSSKRPA SPYGETDGEV AMVTSRQKVE EEESERLPAF HLPLHVSFPN KPHSEEFQPV SLLTQETCGP RTPTVQHNTM EVDGNKVMSS LAPYNSSTSP QKAEEGGRQS GESVSSAALG
mSOX5_isoC  MLTDPDLPQE FERMSSKRPA SPYGETDGEV AMVTSRQKVE EEESERLPAF HLPLH----- ---------- ---------- ---------- EVDGNKVMSS LAPYNSSTSP QKAEEGGRQS GESVSSAALG
 Consensus  MLTDPDLPQE FERMSSKRPA SPYGETDGEV AMVTSRQKVE EEESERLPAF HLPLHvsfpn kphseefqpv slltqetcgp rtptvqhntm EVDGNKVMSS LAPYNSSTSP QKAEEGGRQS GESVSSAALG

         131                                                                                                                                         260
mSOX5_isoA  TPERRKGSLA DVVDTLKQRK MEELIKNEPE DTPSIEKLLS KDWKDKLLAM GSGNFGEIKG TPESLAEKER QLMGMINQLT SLREQLLAAH DEQKKLAASQ IEKQRQQMEL AKQQQEQIAR QQQQLLQQQH
mSOX5_isoB  TPERRKGSLA DVVDTLKQRK MEELIKNEPE DTPSIEKLLS KDWKDKLLAM GSGNFGEIKG TPESLAEKER QLMGMINQLT SLREQLLAAH DEQKKLAASQ IEKQRQQMEL AKQQQEQIAR QQQQLLQQQH
mSOX5_isoC  TPERRKGSLA DVVDTLKQRK MEELIKNEPE DTPSIEKLLS KDWKDKLLAM GSGNFGEIKG TPESLAEKER QLMGMINQLT SLREQLLAAH DEQKKLAASQ IEKQRQQMEL AKQQQEQIAR QQQQLLQQQH
 Consensus  TPERRKGSLA DVVDTLKQRK MEELIKNEPE DTPSIEKLLS KDWKDKLLAM GSGNFGEIKG TPESLAEKER QLMGMINQLT SLREQLLAAH DEQKKLAASQ IEKQRQQMEL AKQQQEQIAR QQQQLLQQQH

         261                                                                                                                                         390
mSOX5_isoA  KINLLQQQIQ -VQGQLPPLM IPVFPPDQRT LAAAAQQGFL LPPGFSYKAG CSDPYPVQLI PTTMAAAAAA TPGLGPLQLQ QFYAAQLAAM QVSPGGKLLG LPQGNLGAAV SPTSIHTDKS TNSPPPKSKD
mSOX5_isoB  KINLLQQQIQ QVQGQLPPLM IPVFPPDQRT LAAAAQQGFL LPPGFSYKAG CSDPYPVQLI PTTMAAAAAA TPGLGPLQLQ ---------- ---------- ---------- ---------- ---------D
mSOX5_isoC  KINLLQQQIQ -VQGQLPPLM IPVFPPDQRT LAAAAQQGFL LPPGFSYKAG CSDPYPVQLI PTTMAAAAAA TPGLGPLQLQ ---------- ---------- ---------- ---------- ---------D
 Consensus  KINLLQQQIQ .VQGQLPPLM IPVFPPDQRT LAAAAQQGFL LPPGFSYKAG CSDPYPVQLI PTTMAAAAAA TPGLGPLQLQ .......... .......... .......... .......... .........D

         391                                                                                                                                         520
mSOX5_isoA  EVAQPLNLSA KPKTSDGKSP ASPTSPHMPA LRINSGAGPL KASVPAALAS PSARVSTIGY LNDHDAVTKA IQEARQMKEQ LRREQQALDG KVAVVNSIGL SNCRTEKEKT TLESLTQQLA VKQNEEGKFS
mSOX5_isoB  EVAQPLNLSA KPKTSDGKSP ASPTSPHMPA LRINSGAGPL KASVPAALAS PSARVSTIGY LNDHDAVTKA IQEARQMKEQ LRREQQALDG KVAVVNSIGL SNCRTEKEKT TLESLTQQLA VKQNEEGKFS
mSOX5_isoC  EVAQPLNLSA KPKTSDGKSP ASPTSPHMPA LRINSGAGPL KASVPAALAS PSARVSTIGY LNDHDAVTKA IQEARQMKEQ LRREQQALDG KVAVVNSIGL SNCRTEKEKT TLESLTQQLA VKQNEEGKFS
 Consensus  EVAQPLNLSA KPKTSDGKSP ASPTSPHMPA LRINSGAGPL KASVPAALAS PSARVSTIGY LNDHDAVTKA IQEARQMKEQ LRREQQALDG KVAVVNSIGL SNCRTEKEKT TLESLTQQLA VKQNEEGKFS

         521                                                                                                                                         650
mSOX5_isoA  HGMMDFNMSG DSDGSAGVSE SRIYRESRGR GSNEPHIKRP MNAFMVWAKD ERRKILQAFP DMHNSNISKI LGSRWKAMTN LEKQPYYEEQ ARLSKQHLEK YPDYKYKPRP KRTCLVDGKK LRIGEYKAIM
mSOX5_isoB  HGMMDFNMSG DSDGSAGVSE SRIYRESRGR GSNEPHIKRP MNAFMVWAKD ERRKILQAFP DMHNSNISKI LGSRWKAMTN LEKQPYYEEQ ARLSKQHLEK YPDYKYKPRP KRTCLVDGKK LRIGEYKAIM
mSOX5_isoC  HGMMDFNMSG DSDGSAGVSE SRIYRESRGR GSNEPHIKRP MNAFMVWAKD ERRKILQAFP DMHNSNISKI LGSRWKAMTN LEKQPYYEEQ ARLSKQHLEK YPDYKYKPRP KRTCLVDGKK LRIGEYKAIM
 Consensus  HGMMDFNMSG DSDGSAGVSE SRIYRESRGR GSNEPHIKRP MNAFMVWAKD ERRKILQAFP DMHNSNISKI LGSRWKAMTN LEKQPYYEEQ ARLSKQHLEK YPDYKYKPRP KRTCLVDGKK LRIGEYKAIM

         651                                                                764
mSOX5_isoA  RNRRQEMRQY FNVGQQAQIP IATAGVVYPS AIAMAGMPSP HLPSEHSSVS SSPEPGMPVI QSTYGAKGEE PHIKEEIQAE DINGEIYEEY DEEEEDPDVD YGSDSENHIA GQAN
mSOX5_isoB  RNRRQEMRQY FNVGQQAQIP IATAGVVYPS AIAMAGMPSP HLPSEHSSVS SSPEPGMPVI QSTYGAKGEE PHIKEEIQAE DINGEIYEEY DEEEEDPDVD YGSDSENHIA GQAN
mSOX5_isoC  RNRRQEMRQY FNVGQQAQIP IATAGVVYPS AIAMAGMPSP HLPSEHSSVS SSPEPGMPVI QSTYGAKGEE PHIKEEIQAE DINGEIYEEY DEEEEDPDVD YGSDSENHIA GQAN
 Consensus  RNRRQEMRQY FNVGQQAQIP IATAGVVYPS AIAMAGMPSP HLPSEHSSVS SSPEPGMPVI QSTYGAKGEE PHIKEEIQAE DINGEIYEEY DEEEEDPDVD YGSDSENHIA GQAN
```

APPENDIX E

>hSOX5_isoform_**a**_23308713_NP_008871.3

>hSOX5_isoform_**b**_23308715_NP_694534.1

>hSOX5_isoform_**c**_30061558_NP_821078.1

>hSOX5_isoform_**d**_387157920_NP_001248343.1

>hSOX5_isoform_**e**_387157922_NP_001248344.1

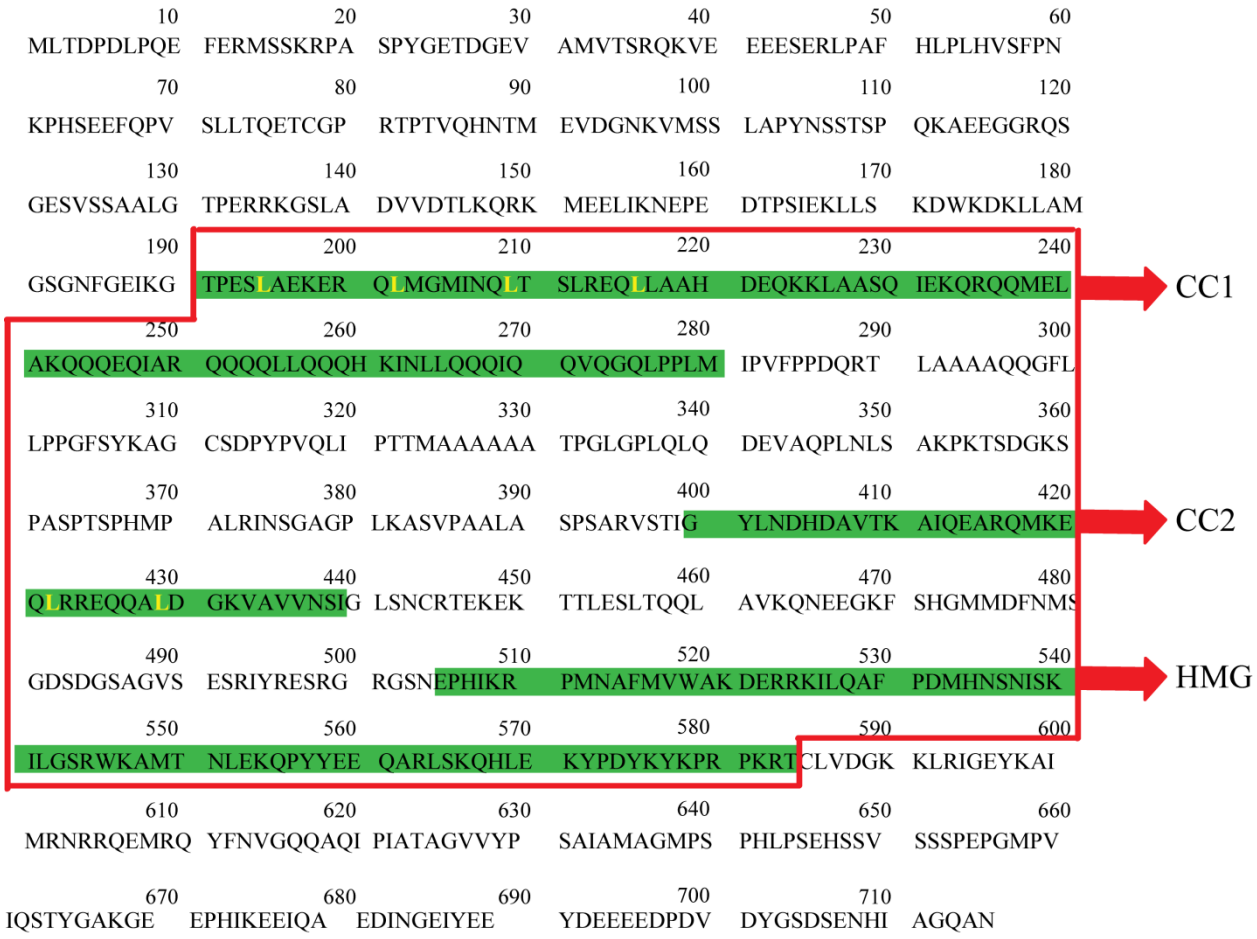
1ST Coiled-Coil Domain

2nd Coiled-Coil Domain

HMG Domain

```
          1                                                                                                  130
hSOX5_isoA  MLTDPDLPQE FERMSSKRPA SPYGEADGEV AMVTSRQKVE EEESDGLPAF HLPLHVSFPN KPHSEEFQPV SLLTQETCGH RTPTSQHNTM EVDGNKVMSS FAPHNSSTSP QKAEEGGRQS GESLSSTALG
hSOX5_isoB                MSSKRPA SPYGEADGEV AMVTSRQKVE EEESDGLPAF HLPLHVSFPN KPHSEEFQPV SLLTQETCGH RTPTSQHNTM EVDGNKVMSS FAPHNSSTSP QKAEEGGRQS GESLSSTALG
hSOX5 isoC
hSOX5_isoD                MSSKRPA SPYGEADGEV AMVTSRQKVE EEESDGLPAF HLPLHVSFPN KPHSEEFQPV SLLTQETCGH RTPTSQHNTM EVDGNKVMSS FAPHNSSTSP QKAEEGGRQS GESLSSTALG
hSOX5_isoE             MSVMSSKRPA SPYGEADGEV AMVTSRQKVE EEESDGLPAF HLPLHVSFPN KPHSEEFQPV SLLTQETCGH RTPTSQHNTM EVDGNKVMSS FAPHNSSTSP QKAEEGGRQS GESLSSTALG
Consensus   ......... ...msskrpa spygeadgev amvtsrqkve eeesdglpaf hlplhvsfpn kphseefqpv slltqetcgh rtptsqhntm evdgnkvmss faphnsstsp qkaeeggrqs geslsstalg

          131                                                                                                260
hSOX5_isoA  TPERRKGSLA DVVDTLKQRK MEELIKNEPE ETPSIEKLLS KDWKDKLLAM GSGNFGEIKG TPESLAEKER QLMGMINQLT SLREQLLAAH DEQKKLAASQ IEKQRQQMEL AKQQQEQIAR QQQQLLQQQH
hSOX5_isoB  TPERRKGSLA DVVDTLKQRK MEELIKNEPE ETPSIEKLLS KDWKDKLLAM GSGNFGEIKG TPESLAEKER QLMGMINQLT SLREQLLAAH DEQKKLAASQ IEKQRQQMEL AKQQQEQIAR QQQQLLQQQH
hSOX5_isoC
hSOX5_isoD  TPERRKGSLA DVVDTLKQRK MEELIKNEPE ETPSIEKLLS KDWKDKLLAM GSGNFGEIKG TPESLAEKER QLMGMINQLT SLREQLLAAH DEQKKLAASQ IEKQRQQMEL AKQQQEQIAR QQQQLLQQQH
hSOX5_isoE  TPERRKGSLA DVVDTLKQRK MEELIKNEPE ETPSIEKLLS KDWKDKLLAM GSGNFGEIKG TPESLAEKER QLMGMINQLT SLREQLLAAH DEQKKLAASQ IEKQRQQMEL AKQQQEQIAR QQQQLLQQQH
Consensus   tperrkgsla dvvdtlkqrk meeliknepe etpsieklls kdwkdkllam gsgnfgeikg tpeslaeker qlmgminqlt slreqllaah deqkklaasq iekqrqqmel akqqqeqiar qqqqlLqqqh

          261                                                                                                390
hSOX5_isoA  KINLLQQQIQ VQGQLPPLMI PVFPPDQRTL AAAAQQGFLL PPGFSYKAGC SDPYPVQLIP TTMAAAAAAT PGLGPLQLQQ LYAAQLAAMQ VSPGGKLPGI PQGNLGAAVS PTSIHTDKST NSPPPKSKDE
hSOX5_isoB  KINLLQQQIQ VQGQLPPLMI PVFPPDQRTL AAAAQQGFLL PPGFSYKAGC SDPYPVQLIP TTMAAAAAAT PGLGPLQLQQ LYAAQLAAMQ VSPGGKLPGI PQGNLGAAVS PTSIHTDKST NSPPPKSKDE
hSOX5_isoC                                                                                                 MHDE
hSOX5_isoD  KINLLQQQIQ VQGQLPPLMI PVFPPDQRTL AAAAQQGFLL PPGFSYKAGC SDPYPVQLIP TTMAAAAAAT PGLGPLQLQQ LYAAQLAAMQ VSPGGKLPGI PQGNLGAAVS PTSIHTDKST NSPPPKSK--
hSOX5_isoE  KINLLQQQIQ VQGQLPPLMI PVFPPDQRTL AAAAQQGFLL PPGFSYKAGC SDPYPVQLIP TTMAAAAAAT PGLGPLQLQQ LYAAQLAAMQ VSPGGKLPGI PQGNLGAAVS PTSIHTDKST NSPPPKSKDE
Consensus   kinllqqqiq vqgqlpplmi pvfppdqrtl aaaaqqgfll ppgfsykagc sdpypvqlip ttmaaaaaat pglgplqlqq lyaaqlaamq vspggklpgi pqgnlgaavs ptsihtdkst nspppkskde

          391                                                                                                520
hSOX5_isoA  VAQPLNLSAK PKTSDGKSPT SPTSPHMPAL RINSGAGPLK ASVPAALASP SARVSTIGYL NDHDAVTKAI QEARQMKEQL RREQQVLDGK VAVVNSLGLN NCRTEKEKTT LESLTQQLAV KQNEEGKFSH
hSOX5_isoB  VAQPLNLSAK PKTSDGKSPT SPTSPHMPAL RINSGAGPLK ASVPAALASP SARVSTIGYL NDHDAVTKAI QEARQMKEQL RREQQVLDGK VAVVNSLGLN NCRTEKEKTT LESLTQQLAV KQNEEGKFSH
hSOX5_isoC  VAQPLNLSAK PKTSDGKSPT SPTSPHMPAL RINSGAGPLK ASVPAALASP SARVSTIGYL NDHDAVTKAI QEARQMKEQL RREQQVLDGK VAVVNSLGLN NCRTEKEKTT LESLTQQLAV KQNEEGKFSH
hSOX5_isoD  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ------EKTT LESLTQQLAV KQNEEGKFSH
hSOX5_isoE  VAQPLNLSAK PKTSDGKSPT SPTSPHMPAL RINSGAGPLK ASVPAALASP SARVSTIGYL NDHDAVTKAI QEARQMKEQL RREQQVLDGK VAVVNSLGLN NCRTEKEKTT LESLTQQLAV KQNEEGKFSH
Consensus   vaqplnlsak pktsdgkspt sptsphmpal rinsgagplk asvpaalasp sarvstigyl ndhdavtkai qearqmkeql rreqqvldgk vavvnslgln ncrtekEKTT LESLTQQLAV KQNEEGKFSH

          521                                                                                                650
hSOX5_isoA  AMMDFNLSGD SDGSAGVSES RIYRESRGRG SNEPHIKRPM NAFMVWAKDE RRKILQAFPD MHNSNISKIL GSRWKAMTNL EKQPYYEEQA RLSKQHLEKY PDYKYKPRPK RTCLVDGKKL RIGEYKAIMR
hSOX5_isoB  AMMDFNLSGD SDGSAGVSES RIYRESRGRG SNEPHIKRPM NAFMVWAKDE RRKILQAFPD MHNSNISKIL GSRWKAMTNL EKQPYYEEQA RLSKQHLEKY PDYKYKPRPK RTCLVDGKKL RIGEYKAIMR
hSOX5_isoC  AMMDFNLSGD SDGSAGVSES RIYRESRGRG SNEPHIKRPM NAFMVWAKDE RRKILQAFPD MHNSNISKIL GSRWKAMTNL EKQPYYEEQA RLSKQHLEKY PDYKYKPRPK RTCLVDGKKL RIGEYKAIMR
hSOX5_isoD  AMMDFNLSGD SDGSAGVSES RIYRESRGRG SNEPHIKRPM NAFMVWAKDE RRKILQAFPD MHNSNISKIL GSRWKAMTNL EKQPYYEEQA RLSKQHLEKY PDYKYKPRPK RTCLVDGKKL RIGEYKAIMR
hSOX5_isoE  AMMDFNLSGD SDGSAGVSES RIYRESRGRG SNEPHIKRPM NAFMVWAKDE RRKILQAFPD MHNSNISKIL GSRWKAMTNL EKQPYYEEQA RLSKQHLEKY PDYKYKPRPK RTCLVDGKKL RIGEYKAIMR
Consensus   AMMDFNLSGD SDGSAGVSES RIYRESRGRG SNEPHIKRPM NAFMVWAKDE RRKILQAFPD MHNSNISKIL GSRWKAMTNL EKQPYYEEQA RLSKQHLEKY PDYKYKPRPK RTCLVDGKKL RIGEYKAIMR

          651                                                                                763
hSOX5_isoA  NRRQEMRQYF NVGQQAQIPI ATAGVVYPGA IAMAGMPSPH LPSEHSSVSS SPEPGMPVIQ STYGVKGEEP HIKEEIQAED INGEIYDEYD EEEDDPDVDY GSDSENHIAG QAN
hSOX5_isoB  NRRQEMRQYF NVGQQAQIPI ATAGVVYPGA IAMAGMPSPH LPSEHSSVSS SPEPGMPVIQ STYGVKGEEP HIKEEIQAED INGEIYDEYD EEEDDPDVDY GSDSENHIAG QAN
hSOX5_isoC  NRRQEMRQYF NVGQQAQIPI ATAGVVYPGA IAMAGMPSPH LPSEHSSVSS SPEPGMPVIQ STYGVKGEEP HIKEEIQAED INGEIYDEYD EEEDDPDVDY GSDSENHIAG QAN
hSOX5_isoD  NRRQEMRQYF NVGQQAQIPI ATAGVVYPGA IAMAGMPSPH LPSEHSSVSS SPEPGMPVIQ STYGVKGEEP HIKEEIQAED INGEIYDEYD EEEDDPDVDY GSDSENHIAG QAN
hSOX5_isoE  NRRQEMRQYF NVGQQAQIPI ATAGVVYPGA IAMAGMPSPH LPSEHSSVSS SPEPGMPVIQ STYGVKGEEP HIKEEIQAED INGEIYDEYD EEEDDPDVDY GSDSENHIAG QAN
Consensus   NRRQEMRQYF NVGQQAQIPI ATAGVVYPGA IAMAGMPSPH LPSEHSSVSS SPEPGMPVIQ STYGVKGEEP HIKEEIQAED INGEIYDEYD EEEDDPDVDY GSDSENHIAG QAN
```

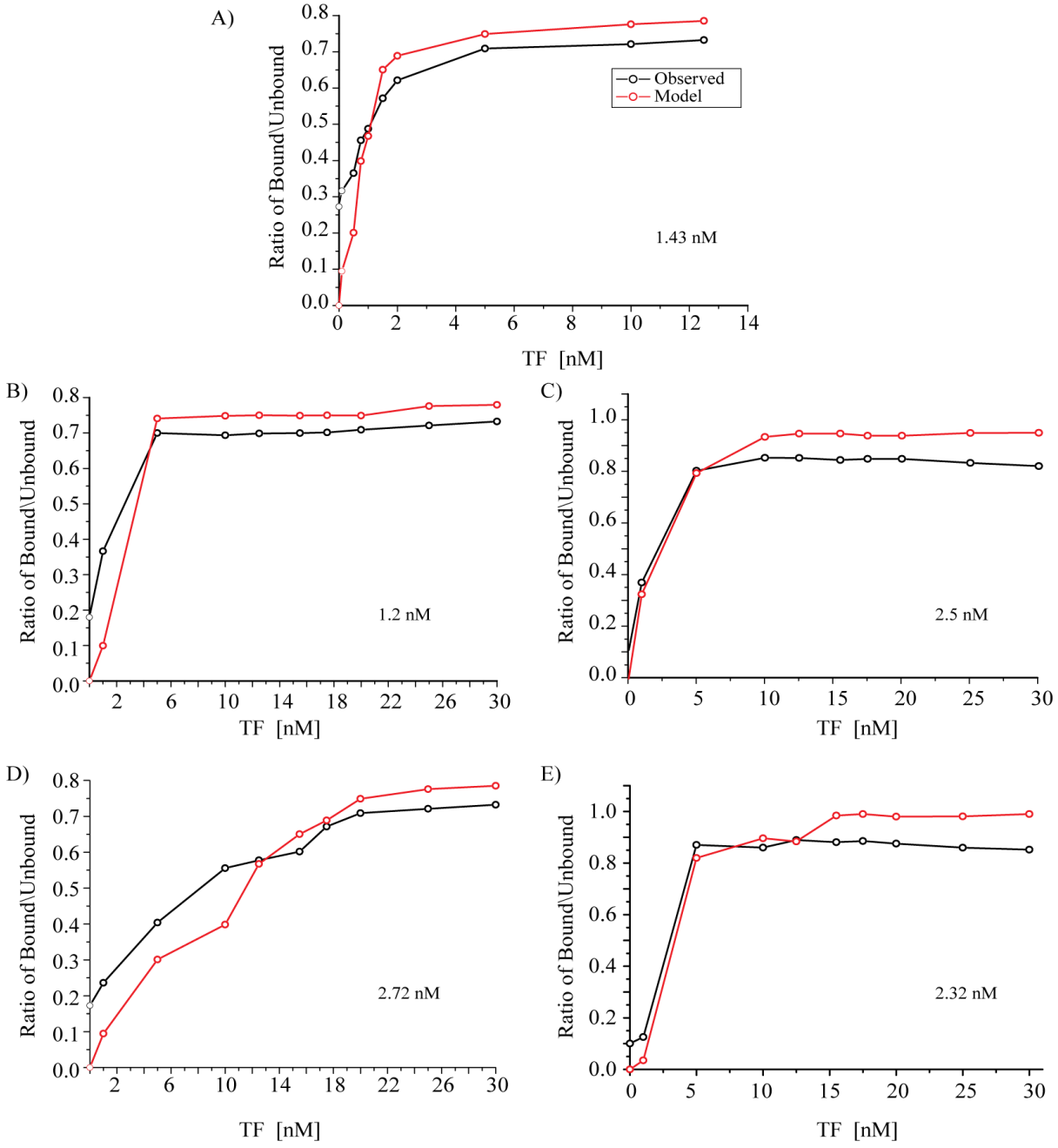
APPENDIX F



Sox5 Full Length Amino Acid Sequence (IMAGE:40047865)

APPENDIX G

APPENDIX H



Non-linear curve fitting of Sox9HMG EMSA using ChIP-Seq identified Sox9 binding motifs. Sox9-HMG domain with canonical motif of FoxP2 Panel (A), new alternative motif from Myom1Panal (B), Postn Panal (C), Rik Panal (D) and Sox5 Panal (E).

APPENDIX I

Crystal Structure of DNA Bound HMG Domain Of The Chondriogenesis Master Regulator, Sox9: Insight Into Sox Transcriptional Specificity

ABSTRACT

Sox9 is a fundamental sex-determining gene and the master regulator of chondrogenesis, involved in the development of various vital organs like testes, kidney, heart, brain and skeletal development. Like other known Sox transcription factors, Sox9 recognizes and binds DNA with consensus sequence C(T/A)TTG(T/A)(T/A) through the highly conserved HMG domain. Nonetheless, the molecular basis of Sox9 functional specificity in key developmental processes is still unclear. As a foremost step towards a mechanistic understanding of Sox9 transcriptional regulation, the current work describes the details of purification of mouse Sox9 HMG domain, its crystallization in complex with a ChIP-Seq identified 16-mer *FOXP2* DNA and the preliminary X-ray diffraction data analysis of this complex. The mSox9HMG-DNA complex was crystallized by the hanging-drop vapor diffusion method using 200 mM Sodium/potassium phosphate, 100 mM Bis Tris propane at pH 8.5, 20% (w/v) PEG 3350. The crystals belong to the tetragonal system, have space group $P4_12_12$ with unit cell parameters a =b= 99.49, c = 45.89 and diffract X-rays to a resolution of 2.7 Å. Analysis of the diffraction data reveals a single molecule in the crystallographic asymmetric unit with an estimated solvent content of 64%.

1.INTRODUCTION

Sox [**S**ex-determining region on the Y chromosome (SRY)-b**ox**] transcription factors contain highly conserved Sry-related High-Mobility Group (HMG) domain of ~80 amino acid, known for binding and bending DNA. Sox proteins are minor groove binding [31], sequence-specific transcription factors that regulate several key developmental processes. The DNA-binding specificities of the 20 mammalian Sox proteins, identified thus far reveal

that Sox transcription factors recognize and bind DNA with C(T/A)TTG(T/A)(T/A) consensus sequence, with similar binding preferences [29, 30].

Sox proteins are grouped into subfamilies A to J based on the amino acid sequence similarity of the HMG domains [22]. Sox9, Sox8 and Sox10 belong to group E. Of these, Sox9 is a fundamental sex-determining gene[67], involved in the development of various vital organs like testes, kidney, heart, brain and skeletal development. Sox9, partnering with Sox5 and Sox6 of Group D, plays a pivotal role as the master regulator of chondrogenesis, regulating multiple stages of cartilage development. Mutations in the Sox9 gene are known to cause campomelic dysplasia, a skeletal malformation syndrome [53], [63], [64]. Despite belonging to the highly conserved HMG domain, binding to the degenerate DNA-binding sites, Sox proteins regulate functionally discrete developmental processes. Sox proteins are believed to achieve functional specificity through either structural rearrangement of the HMG domain arms or by inducing specific kinks to the DNA. Specificity of Sox proteins might also be achieved via bending DNA to distinctive degrees, that might subsequently lead to recruitment of Sox protein specific cofactors. Interestingly, a comparison of the three published DNA-bound crystal structures of Sox2 [36], Sox17 [93] and Sox4 HMG [94] domains, belonging to Sox subgroup B, F, and C respectively, reveals that these transcription factors bend DNA to similar extents (~65°), with comparable helical bend axis and preserve their characteristic L-shaped fold of helices with least structural rearrangements. The few available Sox HMG structures limit our understanding of the functional specificity of the Sox transcription factors. Therefore high-resolution structure determination of the various Sox subgroup HMG domains would provide a comprehensive insight into the mechanism of Sox transcriptional regulation.

To this end, we have attempted to determine the DNA-bound HMG domain structure of mouse Sox9, the master regulator of chondrogenesis. Foregoing, Sox HMG domain crystal

structures have employed DNA elements derived from known Sox enhancer elements such as LAMA1 and FGF4. Alternatively, in the current work, in order to better model the precise *in vivo* functional binding sites of Sox9, we employed a Sox9-specific DNA element derived from the FOXP2 gene promoter, identified through immune precipitation coupled with ultra-high-throughput DNA sequencing (ChIP-Seq) (Moovarkumudalvan et al., unpublished). Here, we present the protein purification, crystallization and preliminary diffraction data of the mSox9HMG domain bound to a FOXP2-derived 16 mer DNA element and the effect of varying overhangs to obtain diffraction-quality crystals.

2. MATERIALS AND METHODS

2.1 Cloning and expression

The 80 amino acid HMG domain of mSox9, spanning residues 103-183 of the full-length protein was PCR-amplified from cDNA clone IMAGE: 5354229 using gene specific primers 5'-CACCCCACACGTCAAGCGACC-3' and 5'-TTACACCGACTTCCTCCGCCG-3'. The amplified PCR product was cloned into pENTR™/TEV/D-TOPO® by directional TOPO cloning (Invitrogen) vector to generate entry clones, further verified by colony PCR and DNA sequencing. The mSox9HMG gene in the entry clone was introduced into the Gateway destination vector pETG20A by performing a Gateway LR reaction, yielding the pETG20A-Sox9HMG expression plasmid and the presence of the gene was validated by PCR using gene specific primers. The expression plasmid pETG20A-mSox9-HMG, thus obtained was transformed into *Escherichia coli* BL21 (DE3) cells (Invitrogen) and were grown in Luria-Bertani (LB) broth containing 100 μg/ml ampicillin and 0.2% glucose at 310 K to an OD_{600} of 0.7. Further, the temperature was lowered to 303 K and protein expression was induced by addition of 0.3 mM IPTG. Cells were harvested by centrifugation after 4 h and stored at 193 K.

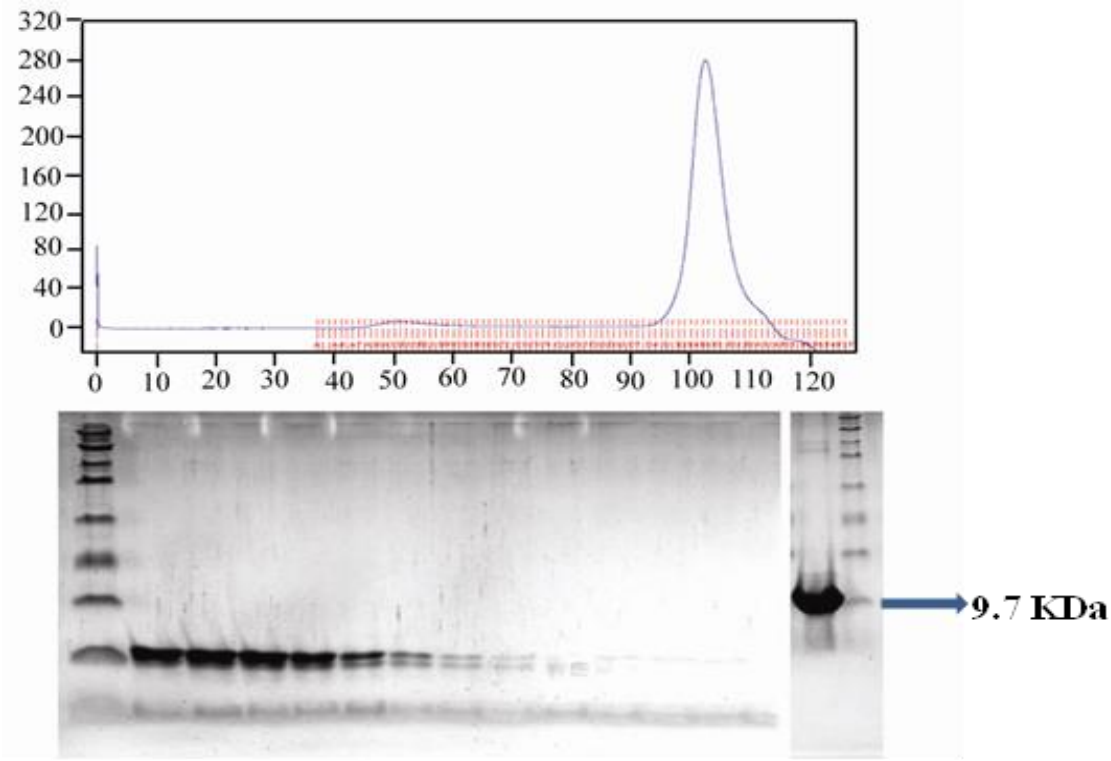


Figure 1. Purification of Sox9HMG domain. Size exclusion chromatography of the purified protein and SDS-PAGE analysis of eluted fractions, pooled and concentrated.

2.2 Protein purification

The harvested cells were thawed, re-suspended in buffer *A* (20 mM Hepes, pH 7.0, 1 mM EDTA, 100 mM NaCl) and homogenized for 20 min at 4°C. The 6x-His and thioredoxin tagged mSox9 HMG was purified through Ni-NTA affinity chromatography in buffer *A* with an imidazole gradient (25-300 mM). The purified, thioredoxin-tagged Thx-His6-Sox9HMG was subjected to cleavage by the TEV protease and purified by ion-exchange chromatography (RESOURCE S, volume 6 ml, GE Healthcare) using the gradient of buffer *B*, 20 mM Hepes at pH 7.0, 1 mM EDTA, 1 M NaCl. The mSox9HMG protein was further purified to homogeneity through size-exclusion chromatography using a HiLoad 16/60 Superdex 75 pg column. The purity of the appropriate protein peak fractions was assessed by matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF

MS) and SDS gel electrophoresis. Pooled fractions were concentrated to 5-10 mg/ml as estimated by standard protein absorbance (A_{280}) using a Thermo Scientific NanoDrop® ND-1000 spectrophotometer.

2.3 Crystallization of mSox9 with DNA

Single-stranded DNA oligonucleotides, purified via Polyacrylamide Gel Electrophoresis, with varying overhangs were commercially obtained at 1 mM concentration (Proligo, Sigma-Aldrich). The complementary oligonucleotides were mixed at equimolar concentration, annealed by heating to 95°C and gradually cooling to ambient temperature. The purified Sox9HMG and the double-stranded DNA were mixed at a molar ratio of 1:1.2 and incubated further on ice for 2 hours. The mSox9HMG-DNA complex thus formed was subjected to crystallization trials at a protein concentration of ~320 μM. Optimal crystal growth conditions were screened using commercially available Hampton Research Crystal and Qiagen Screens, using a liquid-dispensing robot (Innovadyne). Crystallization trials were carried-out using the sitting-drop vapour diffusion method by combining equal volumes of protein solution and precipitating buffer. Optimization of conditions were carried out by varying the length of the DNA, of the overhangs, and by using different ratio of protein and precipitating buffer volume.

2.4 X-ray data collection and processing

Crystals were flash-frozen in liquid nitrogen and a 2.7 Å native data set was collected on the beamline X29 at the Brookhaven National Synchrotron Light Source (NSLS, New York), set at a wavelength of 1.0750 Å. A total of 360 images were collected each with an oscillation angle of 1°. Diffraction intensities obtained using an ADSC Quantum-315r detector were processed and scaled using the HKL-2000 program [108]. Molecular replacement was performed using *Phaser [174]* The Sox17HMG-LAMA1 DNA complex (PDB code:3F27) was utilized as a search probe for molecular replacement with *Phaser*

[174]. Automated model building was initiated using Buccaneer [175] and completed manually using the graphics program Coot [176]. Refinement was carried-out using refmac5 from the CCP4 suite [177]. Further refinement used the *Phenix Refine* pachage *[178]* and simulated-annealing with a starting temperature of 10000 K.

2.5 Electrophoretic Mobility-Shift Assay

EMSA experiments were performed by incubating 5′ Cy5-labeled (Sigma Proligo) 16mer dsDNA FOXP2 promoter elements with mSox9HMG in a binding buffer containing 20 mM Tris–HCl at pH 8.0, 0.1 mg ml^{-1} bovine serum albumin, 50 μM $ZnCl_2$, 100 mM KCl, 10% glycerol, 0.1% NP-40 and 2 mM β-mercaptoethanol. The protein-DNA complex was formed by incubating 0.1, 0.5, 0.75, 1, 1.5, 2, 5, 10 and 12.5 nM protein with 1 nM probe for 1 h at 4 °C, in the dark, in a 10-μL reaction volume. Samples were loaded onto a 12% 1× TG native polyacrylamide gels and electrophoresed in 1×TG (25 mMTris at pH 8.3, 192 mM glycine) at 200 V for 30 min at 4 °C. Bands were detected using a Typhoon 9140 Phosphor Imager (GE Healthcare).

3.Results and discussion

3.1. Protein preparation and protein-DNA complex formation

mSox9HMG protein was over-expressed in a bacterial expression system at its optimal temperature, 30°C after induction with 0.3 mM IPTG and was purified in a soluble form to homogeneity, with typical yields of ~3.5 mg $litre^{-1}$. The pure mSox9HMG protein thus eluted as a monomer with an apparent molecular mass of 9.7 kDa, from a gel filtration column and SDS-PAGE analysis shows >98% purity (Fig. 1).

The *in-vivo* data generated by immuno-precipitation coupled with ultra-high-throughput DNA sequencing (ChIP-Seq) of chondrogenic limb and tail tissues of germline transmitting (GLT) chimeras from $Sox9^{+/-(EGFP)}$mouse (Prof Thomas Lufkin, Genome Institute Singapore, private communication), utilized to identify and validate regulatory

motifs in cartilage-specific genes, yielded a novel Sox9 consensus binding sequence of 5'-AGAACAAAG-3', corresponding to the Foxp2 gene promoter sequence. EMSA using Cy5-labeled dsDNA harboring the FOXP2 motif sequence and Sox9HMG revealed a very high binding affinity with a dissociation constant (Kd) of ~1.4 nM. (Fig. 2). In contrast to Sox9 binding motifs based on computational and *in-vitro* approaches reported earlier, the *in-vivo* results reported here identify a precise and reliable transcription factor binding site and henceforth justify the use of the *FOXP2* DNA element for crystallization with Sox9HMG.

3.2. Crystallization

Crystallization trials were set-up using the sitting drop vapor diffusion method for the homogenously purified Sox9HMG-FOXP2 complex. Initial co-crystallization of Sox9HMG domain set-up with blunt ended 16-mer derived from *FOXP2* gene promoter sequence yielded crystals that were either fragile or of poor diffraction quality. The length of the DNA element and the number of unpaired base pairs in the flanking region are two parameters that are routinely varied for obtaining quality crystals of protein-DNA complexes. Consequently, various *FOXP2* oligonucleotides, ranging from 15-mer to 17-mer and with AT/CG/GC/GG/CC overhangs were utilized for DNA-protein complex formation and the effect on crystal formation was analysed (Table 1). Of all the variants, crystals formed using blunt-end DNA diffracted only up to 9 Å resolution, the use of AT overhangs did not give crystals, crystals with CG and GC diffracted to a maximum of 6 Å resolution, and only GG overhangs yielded better crystals, in the presence of 16% (w/v) PEG 3350, 2% (v/v) tacsimate at pH 5.0 with 100 mM tri-sodium citrate at pH 5.6. The latter crystals diffracted to 3 Å resolution (Fig. 3A). However, data processing was hindered owing to high mosacity. Finally, high quality crystals of a complex with a 16 mer FOXP2 DNA (5'-AGGAGAACAAAGCCTG-3') containing GG overhangs was obtained at 18 ˚C in the following conditions: 200 mM Sodium/potassium phosphate, 100 mM Bis Tris propane at pH

8.5, 20% (w/v) PEG 3350, with a ratio of mother liquor:protein-DNA complex of 2:1 and a protein concentration of 257 μM (Fig. 3B). The crystals were harvested after 25 days, flash frozen and stored in liquid nitrogen. Crystals dissolved in mother liquor and subjected to SDS-PAGE and agarose gel analysis revealed the presence of both DNA and protein.

3.3. Data collection and processing

The optimised crystals of the mSox9HMG-FOXP2 DNA complex belong to the tetragonal system; space group $P4_12_12$ or to its enantiomorph $P4_32_12$, with unit-cell parameters a = b= 99.49, c = 45.89 and one complex per asymmetric unit (Matthews coefficient of 2.8 $Å^3$ Da^{-1} and a solvent content of 64 %). A complete diffraction data set to a resolution of 2.7 Å was collected (Table 2). The Sox17HMG-LAMA1 DNA complex (PDB ID:3F27) was utilized for molecular replacement giving an unambiguous solution. Initial refinement of the structure of Sox9HMG-FOXP2 DNA complex yielded R-factor/R_{free} (21.3/26.8 %) (Table 2).

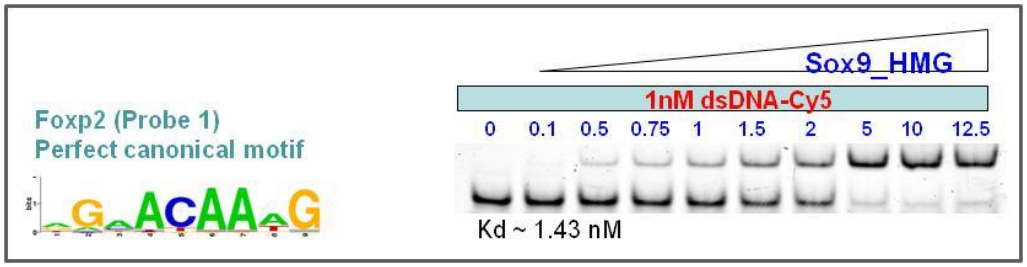


Figure 2. ChIP-Seq identified Sox9 binding motifs and EMSA of Sox9-HMG domain with canonical motif of FoxP2

Figure 3. **mSox9HMG–DNA complex crystals.** Crystals of mSox9HMG–GG-overhang Foxp2 DNA grown in a buffer containing 200 mM Sodium/potassium phosphate 100 mM Bis Tris propane pH 8.5 20% (w/v) PEG 3350 that diffracted to 2.7 A ° resolution.

Parameters	**SOX9**
Data collection	
Source / Wavelength	NSLS X29A beamline / 1.0750
Space group	$P4_12_12$
Unit-cell parameters (Å, ˚)	a = b = 99.492, c = 45.894; α= β =γ = 90
Resolution range (Å)	50.0-2.70 (2.80-2.70)
Total no. of reflections / Unique reflections	203327 / 6721
No. of molecules in asymmetric unit	1
$^{\dagger}R_{merge}$ (%)	13.3 (65.4)
Average redundancy	30.3 (28.2)
Completeness (%)	99.9 (100.0)
Average *I/σ(I)*	19.0 (6.9)
Refinement	
Resolution range (Å)	25.0-2.70
Reflections used	6675
R-factor/R_{free} (%)	21.3/26.8
Mean B values ($Å^2$)	85.98
R.M.S deviations from ideals	
Bond length (Å)	0.008
Bond angle (˚)	1.49

Ramachandran plot	
Most favored region (%)	95.95
Allowed region (%)	4.05

$^{\dagger}R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl) \rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the measured intensity of reflection I and $\langle I(hkl) \rangle$ is the mean intensity. (Values in the parenthesis are for highest resolution bin)

Table.1 Data collection and refinement statistics for Sox9HMG-Foxp2

3.4 The Overall Structure.

The structure was determined by molecular replacement and refined using diffraction data to 2.7 Å resolutions while maintaining good stereochemistry (see Table 1). The final model of Sox9HMG contains 75 residues and the entire 16-bp double-stranded DNA. The electron density is well defined the protein–DNA interface which are functionally and structurally important.The N-terminal four residues, GSFT tetrapeptide which is result of TOPO cloning vector and three C-terminal residues could not be modeled because of structural disorder. The structure of the Sox9HMG exhibits the characteristic L-shaped arrangement of three helices starting from 73-90, 94-108 and 110-132 with N and C-termini positioned at the same molecular surface (Fig. 4).

$^{\dagger}R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl) \rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the measured intensity of reflection I and $\langle I(hkl) \rangle$ is the mean intensity.

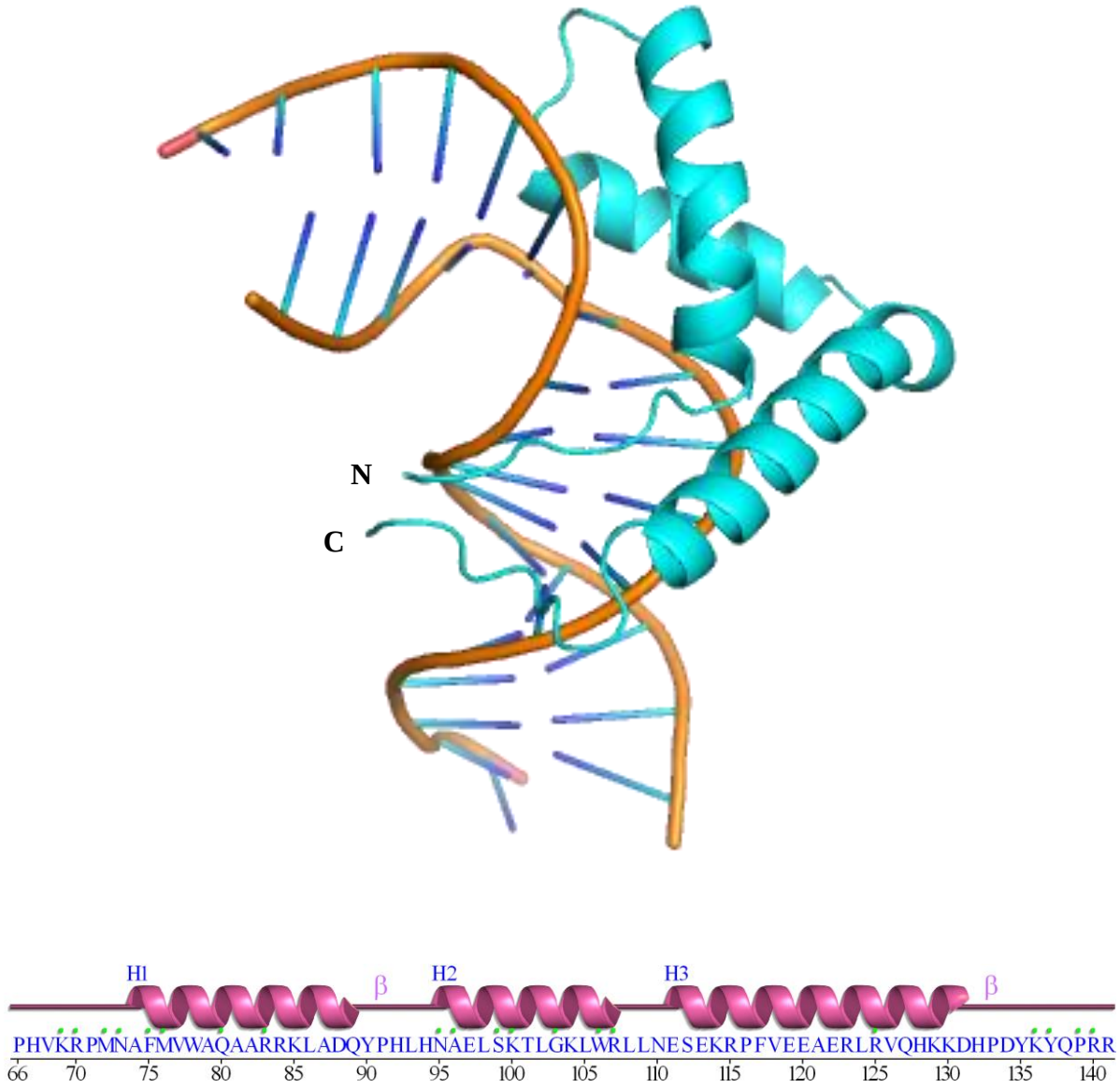


Figure 4. The structure of Sox9HMG protein bound with Foxp2 DNA. The Sox9HMG domain exhibiting a characteristic three-helix bundle, L-shaped arrangement.

The Sox HMG domain amino acid sequences are highly conserved specially the DNA contact residues. The interaction interface between Sox9HMG and the DNA Foxp2 with the majority of DNA contacts mediated by residues extending from the N-terminal half of the HMG domain (Fig. 5). The DNA is contacted from the minor groove side and most DNA contacts are polar with the notable exception of the intercalating Phe75-Met76 dipeptide. Prominent base contacts are mediated by Arg70, Asn73, Ser99 and Trp106 that engage in

hydrogen-bonding interactions with either the carbonyl O2 of pyrimidines or the N3 of purines (Fig. 5). HMG domains are known to bend DNA have been shown to bind the minor groove and to induce prominent topological deformations with respect to standard B-DNA. The binding of Sox9HMG to the Foxp2 promoter element diverts the double helix by introducing a bend of approximately 59°. The minor groove is widened, whereas the major groove width decreases at the core of the interaction site.

Although all Sox proteins possess highly conserved HMG domains, bind similar DNA elements, they regulate wide assortment of genes in diverse developmental processes. The functional specificity of Sox transcription factors plausibly depends on (i) subtle nucleotide variations in the DNA sequence; (ii) differential minor groove bend as a consequence of HMG domain mediated DNA interaction (iii) different co-factor recruitment through protein-protein interactions.

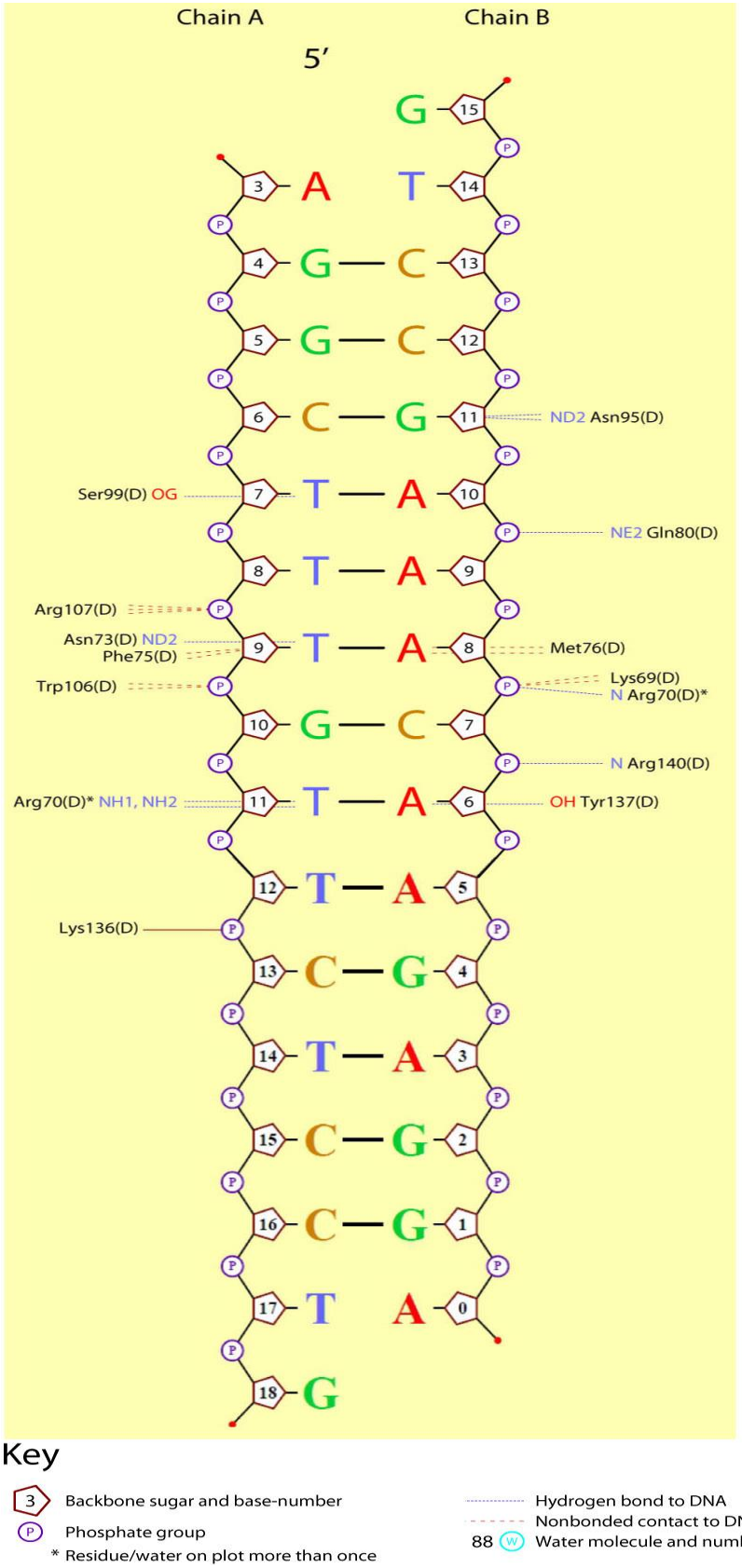


Figure 5. Foxp2 DNA sequence used for co-crystallization with Sox9HMG. Residues interacting with DNA are indicated.

Despite many advances made so far in DNA binding specificity of transcription factors, understanding the key-principles of transcriptional regulation is still an outstanding puzzle as specific recognition of cis-regulatory DNA elements by transcription factors (TF) is determined by multitude of factors like in vivo transcription-factor concentration, the relative affinity of the TF towards specific and non-specific binding sites, co-operativity with other protein-complexes, accessibility of nucleosomal DNA and other aspects like the presence/absence of epigenetic marks such as DNA methylation. Although the binding sequences of all known Sox proteins are barely distinguishable from each other, each Sox protein regulates a distinct set of target genes. Col2a1 was the first and established gene proposed as a direct target of regulation of Sox trio comprising Sox9, Sox5 and Sox6. Sox9 powerfully activates a 48-bp cartilage-specific enhancer located in its first intron [7, 8] and L-Sox5/Sox6 enhance the activity of Sox9 [9]. However, the enhancer of COL2a1 has no defined consensus Sox site but has four sites with as few as five or six Sox9 consensus nucleotides. Sox9 binds the most distal pair of them and L-Sox5/ Sox6 contacts each of them in vitro [10].

Even though so far several Sox9 binding motifs have been reported based on computational and in-vitro studies, in-vivo results reflect precise and accurate transcription factor binding site. In this regard, the study with in vivo data from immunoprecipitation coupled ultra-high-throughput DNA sequencing (ChIP-Seq) of chondrogenic limb and tail tissue from living mouse embryos may reflects the in vivo biological condition.

The three dimensional data of DNA bound Sox9HMG would provide better insight into operation of DNA recognition model in Sox9 proteins specifically and comparison of the known sox structures of other sox proteins's interface amino acids at the DNA binding interface will reveal subtle conformational rearrangements and shed light on mechanism of action and DNA binding specificity of Sox transcription factor as whole.

REFERENCES

1. Bjorklund, S. and Y.J. Kim, *Mediator of transcriptional regulation.* Trends in biochemical sciences, 1996. **21**(9): p. 335-7.
2. Kaiser, K. and M. Meisterernst, *The human general co-factors.* Trends Biochem Sci, 1996. **21**(9): p. 342-5.
3. Orphanides, G., T. Lagrange, and D. Reinberg, *The general transcription factors of RNA polymerase II.* Genes Dev, 1996. **10**(21): p. 2657-83.
4. Pope, A., Thor, M.,, *Polyoxometalates: From Platonic Solids to Anti-retroviral Activity*. 3 ed1994: Springer.
5. Lodish, H., Berk, A., Kaiser, C.A.,Krieger, M., Scott, M.P., Bretscher, A., Ploegh, H., Matsudaira, P.,, *Molecular Cell Biology*2007: W.H.Freeman.
6. Jacob, F., et al., *[The operon: a group of genes with expression coordinated by an operator. C.R.Acad. Sci. Paris 250 (1960) 1727-1729]*. C R Biol, 2005. **328**(6): p. 514-20.
7. Huffman, J.L. and R.G. Brennan, *Prokaryotic transcription regulators: more than just the helix-turn-helix motif.* Curr Opin Struct Biol, 2002. **12**(1): p. 98-106.
8. GM, C., *The Cell: A Molecular Approach*. 2nd edition ed. Transcription in Prokaryotes2000: Sunderland (MA): Sinauer Associates.
9. Prabhakar, S., et al., *Human-specific gain of function in a developmental enhancer.* Science, 2008. **321**(5894): p. 1346-50.
10. Nicerweb.com, *Bio1903, ch18-Activator*, 2009, Nicerweb.com.
11. He, J., et al., *Structure of p300 bound to MEF2 on DNA reveals a mechanism of enhanceosome assembly.* Nucleic Acids Res, 2011. **39**(10): p. 4464-74.
12. Weirauch, M.T. and T.R. Hughes, *A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution.* Subcell Biochem, 2011. **52**: p. 25-73.
13. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
14. Caretti, G., M.C. Motta, and R. Mantovani, *NF-Y associates with H3-H4 tetramers and octamers by multiple mechanisms.* Mol Cell Biol, 1999. **19**(12): p. 8591-603.
15. Xu, Y., et al., *Solution structure of the first HMG box domain in human upstream binding factor.* Biochemistry, 2002. **41**(17): p. 5415-20.
16. Yang, W., et al., *Solution structure and DNA binding property of the fifth HMG box domain in comparison with the first HMG box domain in human upstream binding factor.* Biochemistry, 2003. **42**(7): p. 1930-8.
17. Kamachi, Y., K.S. Cheah, and H. Kondoh, *Mechanism of regulatory target selection by the SOX high-mobility-group domain proteins as revealed by comparison of SOX1/2/3 and SOX9.* Mol Cell Biol, 1999. **19**(1): p. 107-20.
18. Stros, M., D. Launholt, and K.D. Grasser, *The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins.* Cell Mol Life Sci, 2007. **64**(19-20): p. 2590-606.
19. Stefanovic, S., et al., *Interplay of Oct4 with Sox2 and Sox17: a molecular switch from stem cell pluripotency to specifying a cardiac fate.* J Cell Biol, 2009. **186**(5): p. 665-73.
20. Wilson, M. and P. Koopman, *Matching SOX: partner proteins and co-factors of the SOX family of transcriptional regulators.* Curr Opin Genet Dev, 2002. **12**(4): p. 441-6.

21. Pevny, L.H. and R. Lovell-Badge, *Sox genes find their feet.* Curr Opin Genet Dev, 1997. **7**(3): p. 338-44.
22. Kamachi, Y., M. Uchikawa, and H. Kondoh, *Pairing SOX off: with partners in the regulation of embryonic development.* Trends Genet, 2000. **16**(4): p. 182-7.
23. Kiefer, J.C., *Back to basics: Sox genes.* Dev Dyn, 2007. **236**(8): p. 2356-66.
24. Huson, D.H., *SplitsTree: analyzing and visualizing evolutionary data.* Bioinformatics, 1998. **14**(1): p. 68-73.
25. Katoh, K., et al., *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Res, 2002. **30**(14): p. 3059-66.
26. Remenyi, A., et al., *Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers.* Genes Dev, 2003. **17**(16): p. 2048-59.
27. Dow, L.K., et al., *Structural studies of the high mobility group globular domain and basic tail of HMG-D bound to disulfide cross-linked DNA.* Biochemistry, 2000. **39**(32): p. 9725-36.
28. Murphy, F.V.t. and M.E. Churchill, *Nonsequence-specific DNA recognition: a structural perspective.* Structure, 2000. **8**(4): p. R83-9.
29. van Houte, L.P., et al., *Solution structure of the sequence-specific HMG box of the lymphocyte transcriptional activator Sox-4.* J Biol Chem, 1995. **270**(51): p. 30516-24.
30. Boyer, L.A., et al., *Core transcriptional regulatory circuitry in human embryonic stem cells.* Cell, 2005. **122**(6): p. 947-56.
31. Werner, M.H., et al., *Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex.* Cell, 1995. **81**(5): p. 705-14.
32. Love, J.J., et al., *Structural basis for DNA bending by the architectural transcription factor LEF-1.* Nature, 1995. **376**(6543): p. 791-5.
33. Palasingam, P., et al., *The structure of Sox17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms.* J Mol Biol, 2009. **388**(3): p. 619-30.
34. Weiss, M.A., *Floppy SOX: mutual induced fit in hmg (high-mobility group) box-DNA recognition.* Mol Endocrinol, 2001. **15**(3): p. 353-62.
35. Murphy, F.V.t., R.M. Sweet, and M.E. Churchill, *The structure of a chromosomal high mobility group protein-DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition.* The EMBO journal, 1999. **18**(23): p. 6610-8.
36. 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. 2048-59.
37. 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.
38. Scaffidi, P. and M.E. Bianchi, *Spatially precise DNA bending is an essential activity of the sox2 transcription factor.* The Journal of biological chemistry, 2001. **276**(50): p. 47296-302.
39. Wegner, M., *Secrets to a healthy Sox life: lessons for melanocytes.* Pigment Cell Res, 2005. **18**(2): p. 74-85.
40. Wissmuller, S., et al., *The high-mobility-group domain of Sox proteins interacts with DNA-binding domains of many transcription factors.* Nucleic Acids Res, 2006. **34**(6): p. 1735-44.

41. Kuhlbrodt, K., et al., *Cooperative function of POU proteins and SOX proteins in glial cells.* J Biol Chem, 1998. **273**(26): p. 16050-7.
42. Tsukamoto, T., et al., *Down-regulation of a gastric transcription factor, Sox2, and ectopic expression of intestinal homeobox genes, Cdx1 and Cdx2: inverse correlation during progression from gastric/intestinal-mixed to complete intestinal metaplasia.* J Cancer Res Clin Oncol, 2004. **130**(3): p. 135-45.
43. 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.
44. Kondoh, H. and Y. Kamachi, *SOX-partner code for cell specification: Regulatory target selection and underlying molecular mechanisms.* Int J Biochem Cell Biol, 2010. **42**(3): p. 391-9.
45. Peirano, R.I. and M. Wegner, *The glial transcription factor Sox10 binds to DNA both as monomer and dimer with different functional consequences.* Nucleic acids research, 2000. **28**(16): p. 3047-55.
46. Sock, E., *Loss of DNA-dependent dimerization of the transcription factor SOX9 as a cause for campomelic dysplasia.* Human Molecular Genetics, 2003. **12**(>12): p. 1439-1447.
47. Bernard, P., *Dimerization of SOX9 is required for chondrogenesis, but not for sex determination.* Human Molecular Genetics, 2003. **12**(14): p. 1755-1765.
48. Pfeifer, D., et al., *The SOX8 gene is located within 700 kb of the tip of chromosome 16p and is deleted in a patient with ATR-16 syndrome.* Genomics, 2000. **63**(1): p. 108-16.
49. Schepers, G., et al., *SOX8 is expressed during testis differentiation in mice and synergizes with SF1 to activate the Amh promoter in vitro.* The Journal of biological chemistry, 2003. **278**(30): p. 28101-8.
50. Schreiner, S., et al., *Hypomorphic Sox10 alleles reveal novel protein functions and unravel developmental differences in glial lineages.* Development, 2007. **134**(18): p. 3271-81.
51. Sudbeck, P., et al., *Sex reversal by loss of the C-terminal transactivation domain of human SOX9.* Nature genetics, 1996. **13**(2): p. 230-2.
52. Pusch, C., et al., *The SOX10/Sox10 gene from human and mouse: sequence, expression, and transactivation by the encoded HMG domain transcription factor.* Human genetics, 1998. **103**(2): p. 115-23.
53. Akiyama, H., et al., *The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6.* Genes Dev, 2002. **16**(21): p. 2813-28.
54. Wegner, M. and C.C. Stolt, *From stem cells to neurons and glia: a Soxist's view of neural development.* Trends Neurosci, 2005. **28**(11): p. 583-8.
55. McDowall, S., et al., *Functional and structural studies of wild type SOX9 and mutations causing campomelic dysplasia.* J Biol Chem, 1999. **274**(34): p. 24023-30.
56. Schepers, G.E., et al., *Cloning and characterisation of the Sry-related transcription factor gene Sox8.* Nucleic acids research, 2000. **28**(6): p. 1473-80.
57. Pingault, V., et al., *SOX10 mutations in patients with Waardenburg-Hirschsprung disease.* Nat Genet, 1998. **18**(2): p. 171-3.
58. Kim, J., et al., *SOX10 maintains multipotency and inhibits neuronal differentiation of neural crest stem cells.* Neuron, 2003. **38**(1): p. 17-31.
59. Bondurand, N., et al., *Interaction among SOX10, PAX3 and MITF, three genes altered in Waardenburg syndrome.* Hum Mol Genet, 2000. **9**(13): p. 1907-17.

60. Stolt, C.C., et al., *Transcription factors Sox8 and Sox10 perform non-equivalent roles during oligodendrocyte development despite functional redundancy.* Development, 2004. **131**(10): p. 2349-58.
61. Stolt, C.C., et al., *Terminal differentiation of myelin-forming oligodendrocytes depends on the transcription factor Sox10.* Genes Dev, 2002. **16**(2): p. 165-70.
62. Britsch, S., et al., *The transcription factor Sox10 is a key regulator of peripheral glial development.* Genes Dev, 2001. **15**(1): p. 66-78.
63. Ikeda, T., et al., *The combination of SOX5, SOX6, and SOX9 (the SOX trio) provides signals sufficient for induction of permanent cartilage.* Arthritis Rheum, 2004. **50**(11): p. 3561-73.
64. Foster, J.W., et al., *Campomelic dysplasia and autosomal sex reversal caused by mutations in an SRY-related gene.* Nature, 1994. **372**(6506): p. 525-30.
65. Kent, J., et al., *A male-specific role for SOX9 in vertebrate sex determination.* Development, 1996. **122**(9): p. 2813-22.
66. Bernard, P., et al., *Dimerization of SOX9 is required for chondrogenesis, but not for sex determination.* Hum Mol Genet, 2003. **12**(14): p. 1755-65.
67. Clarkson, M.J. and V.R. Harley, *Sex with two SOX on: SRY and SOX9 in testis development.* Trends Endocrinol Metab, 2002. **13**(3): p. 106-11.
68. Chaboissier, M.C., et al., *Functional analysis of Sox8 and Sox9 during sex determination in the mouse.* Development, 2004. **131**(9): p. 1891-901.
69. Vidal, V.P., et al., *Sox9 induces testis development in XX transgenic mice.* Nat Genet, 2001. **28**(3): p. 216-7.
70. Kim, Y., et al., *Fgf9 and Wnt4 act as antagonistic signals to regulate mammalian sex determination.* PLoS Biol, 2006. **4**(6): p. e187.
71. Chew, L.J. and V. Gallo, *The Yin and Yang of Sox proteins: Activation and repression in development and disease.* Journal of neuroscience research, 2009. **87**(15): p. 3277-87.
72. Wagner, T., et al., *Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene SOX9.* Cell, 1994. **79**(6): p. 1111-20.
73. Schroeder, T.M., E.D. Jensen, and J.J. Westendorf, *Runx2: a master organizer of gene transcription in developing and maturing osteoblasts.* Birth Defects Res C Embryo Today, 2005. **75**(3): p. 213-25.
74. Chikuda, H., et al., *Cyclic GMP-dependent protein kinase II is a molecular switch from proliferation to hypertrophic differentiation of chondrocytes.* Genes Dev, 2004. **18**(19): p. 2418-29.
75. de Crombrugghe, B., et al., *Transcriptional mechanisms of chondrocyte differentiation.* Matrix Biol, 2000. **19**(5): p. 389-94.
76. Hattori, T., et al., *Interactions between PIAS proteins and SOX9 result in an increase in the cellular concentrations of SOX9.* J Biol Chem, 2006. **281**(20): p. 14417-28.
77. Malki, S., et al., *Prostaglandin D2 induces nuclear import of the sex-determining factor SOX9 via its cAMP-PKA phosphorylation.* EMBO J, 2005. **24**(10): p. 1798-809.
78. Privalov, P.L., *DNA Binding and Bending by HMG Boxes:*
Energetic Determinants of Specificity. J. Mol. Biol., 2004. **343**: p. 371-393.
79. Smits, P., et al., *Sox5 and Sox6 are needed to develop and maintain source, columnar, and hypertrophic chondrocytes in the cartilage growth plate.* J Cell Biol, 2004. **164**(5): p. 747-58.
80. Yamashita, A., et al., *cDNA cloning of a novel rainbow trout SRY-type HMG box protein, rtSox23, and its functional analysis.* Gene, 1998. **209**(1-2): p. 193-200.

81. Takamatsu, N., et al., *A gene that is related to SRY and is expressed in the testes encodes a leucine zipper-containing protein.* Mol Cell Biol, 1995. **15**(7): p. 3759-66.
82. Lefebvre, V., P. Li, and B. de Crombrugghe, *A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene.* EMBO J, 1998. **17**(19): p. 5718-33.
83. Connor, F., et al., *DNA binding and bending properties of the post-meiotically expressed Sry-related protein Sox-5.* Nucleic Acids Res, 1994. **22**(16): p. 3339-46.
84. Cohen-Barak, O., et al., *Stem cell transplantation demonstrates that Sox6 represses epsilon y globin expression in definitive erythropoiesis of adult mice.* Exp Hematol, 2007. **35**(3): p. 358-67.
85. Ikeda, T., et al., *Identification and characterization of the human long form of Sox5 (L-SOX5) gene.* Gene, 2002. **298**(1): p. 59-68.
86. Hiraoka, Y., et al., *The mouse Sox5 gene encodes a protein containing the leucine zipper and the Q box.* Biochim Biophys Acta, 1998. **1399**(1): p. 40-6.
87. Ikeda, T., et al., *Distinct roles of Sox5, Sox6, and Sox9 in different stages of chondrogenic differentiation.* J Bone Miner Metab, 2005. **23**(5): p. 337-40.
88. Stolt, C.C., et al., *SoxD proteins influence multiple stages of oligodendrocyte development and modulate SoxE protein function.* Dev Cell, 2006. **11**(5): p. 697-709.
89. Stolt, C.C., et al., *The transcription factor Sox5 modulates Sox10 function during melanocyte development.* Nucleic Acids Res, 2008. **36**(17): p. 5427-40.
90. Kasimiotis, H., et al., *Sex-determining region Y-related protein SOX13 is a diabetes autoantigen expressed in pancreatic islets.* Diabetes, 2000. **49**(4): p. 555-61.
91. Hoek, K.S., et al., *Novel MITF targets identified using a two-step DNA microarray strategy.* Pigment Cell Melanoma Res, 2008. **21**(6): p. 665-76.
92. Lefebvre, V., *The SoxD transcription factors--Sox5, Sox6, and Sox13--are key cell fate modulators.* Int J Biochem Cell Biol, 2010. **42**(3): p. 429-32.
93. 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. 619-30.
94. Jauch, R., et al., *The crystal structure of the Sox4 HMG domain-DNA complex suggests a mechanism for positional interdependence in DNA recognition.* The Biochemical journal, 2012. **443**(1): p. 39-47.
95. Francois, M., et al., *Sox18 induces development of the lymphatic vasculature in mice.* Nature, 2008. **456**(7222): p. 643-7.
96. Jauch, R., et al., *Crystal structure of the Sox4 HMG/DNA complex suggests a mechanism for the positional interdependence in DNA recognition.* Biochem J, 2011.
97. Glass, C.K., *Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers.* Endocrine reviews, 1994. **15**(3): p. 391-407.
98. Forman, B.M. and H.H. Samuels, *Dimerization among nuclear hormone receptors.* New Biol, 1990. **2**(7): p. 587-94.
99. Tsai, S.Y., et al., *Molecular interactions of steroid hormone receptor with its enhancer element: evidence for receptor dimer formation.* Cell, 1988. **55**(2): p. 361-9.
100. Rastinejad, F., et al., *Structural determinants of nuclear receptor assembly on DNA direct repeats.* Nature, 1995. **375**(6528): p. 203-11.
101. Verrijdt, G., A. Haelens, and F. Claessens, *Selective DNA recognition by the androgen receptor as a mechanism for hormone-specific regulation of gene expression.* Molecular genetics and metabolism, 2003. **78**(3): p. 175-85.
102. Centenera, M.M., et al., *The contribution of different androgen receptor domains to receptor dimerization and signaling.* Molecular endocrinology, 2008. **22**(11): p. 2373-82.

103. Fang, G. and T.R. Cech, *Oxytricha telomere-binding protein: DNA-dependent dimerization of the alpha and beta subunits.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(13): p. 6056-60.
104. Arai, N., K. Arai, and A. Kornberg, *Complexes of Rep protein with ATP and DNA as a basis for helicase action.* The Journal of biological chemistry, 1981. **256**(10): p. 5287-93.
105. Arai, N. and A. Kornberg, *Rep protein as a helicase in an active, isolatable replication fork of duplex phi X174 DNA.* The Journal of biological chemistry, 1981. **256**(10): p. 5294-8.
106. Marmorstein, R., et al., *DNA recognition by GAL4: structure of a protein-DNA complex.* Nature, 1992. **356**(6368): p. 408-14.
107. Schwabe, J.W., D. Neuhaus, and D. Rhodes, *Solution structure of the DNA-binding domain of the oestrogen receptor.* Nature, 1990. **348**(6300): p. 458-61.
108. Luisi, B.F., et al., *Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA.* Nature, 1991. **352**(6335): p. 497-505.
109. Hard, T., et al., *Solution structure of the glucocorticoid receptor DNA-binding domain.* Science, 1990. **249**(4965): p. 157-60.
110. Houston, C.S., et al., *The campomelic syndrome: review, report of 17 cases, and follow-up on the currently 17-year-old boy first reported by Maroteaux et al in 1971.* Am J Med Genet, 1983. **15**(1): p. 3-28.
111. Bell, D.M., et al., *SOX9 directly regulates the type-II collagen gene.* Nature genetics, 1997. **16**(2): p. 174-8.
112. Lefebvre, V., et al., *SOX9 is a potent activator of the chondrocyte-specific enhancer of the pro alpha1(II) collagen gene.* Molecular and cellular biology, 1997. **17**(4): p. 2336-46.
113. Ng, L.J., et al., *SOX9 binds DNA, activates transcription, and coexpresses with type II collagen during chondrogenesis in the mouse.* Developmental biology, 1997. **183**(1): p. 108-21.
114. Bridgewater, L.C., V. Lefebvre, and B. de Crombrugghe, *Chondrocyte-specific enhancer elements in the Col11a2 gene resemble the Col2a1 tissue-specific enhancer.* The Journal of biological chemistry, 1998. **273**(24): p. 14998-5006.
115. Xie, W.F., et al., *Trans-activation of the mouse cartilage-derived retinoic acid-sensitive protein gene by Sox9.* Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research, 1999. **14**(5): p. 757-63.
116. Glover, J.N. and S.C. Harrison, *Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA.* Nature, 1995. **373**(6511): p. 257-61.
117. Kammerer, R.A., et al., *An autonomous folding unit mediates the assembly of two-stranded coiled coils.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(23): p. 13419-24.
118. Suzuki, K., T. Yamada, and T. Tanaka, *Role of the buried glutamate in the alpha-helical coiled coil domain of the macrophage scavenger receptor.* Biochemistry, 1999. **38**(6): p. 1751-6.
119. Branden, C.T.a.T.J., *Introduction to Protein Structure*, in *Garland Publishing Inc.*1991.
120. Hu, J.C., *A guided tour in protein interaction space: coiled coils from the yeast proteome.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(24): p. 12935-6.

121. Han, Y. and V. Lefebvre, *L-Sox5 and Sox6 drive expression of the aggrecan gene in cartilage by securing binding of Sox9 to a far-upstream enhancer.* Mol Cell Biol, 2008. **28**(16): p. 4999-5013.
122. Lupas, A., *Coiled coils: new structures and new functions.* Trends in biochemical sciences, 1996. **21**(10): p. 375-82.
123. Burkhard, P., J. Stetefeld, and S.V. Strelkov, *Coiled coils: a highly versatile protein folding motif.* Trends in cell biology, 2001. **11**(2): p. 82-8.
124. Zitzewitz, J.A., et al., *Probing the folding mechanism of a leucine zipper peptide by stopped-flow circular dichroism spectroscopy.* Biochemistry, 1995. **34**(39): p. 12812-9.
125. O'Shea, E.K., K.J. Lumb, and P.S. Kim, *Peptide 'Velcro': design of a heterodimeric coiled coil.* Current biology : CB, 1993. **3**(10): p. 658-67.
126. Schnepf, R., et al., *De novo design and characterization of copper centers in synthetic four-helix-bundle proteins.* Journal of the American Chemical Society, 2001. **123**(10): p. 2186-95.
127. DeGrado, W.F., et al., *De novo design and structural characterization of proteins and metalloproteins.* Annual review of biochemistry, 1999. **68**: p. 779-819.
128. Lupas, A.N. and M. Gruber, *The structure of alpha-helical coiled coils.* Advances in protein chemistry, 2005. **70**: p. 37-78.
129. Kunjithapatham, R., et al., *Role for the alpha-helix in aberrant protein aggregation.* Biochemistry, 2005. **44**(1): p. 149-56.
130. Frank, S., et al., *Characterization of the matrilin coiled-coil domains reveals seven novel isoforms.* The Journal of biological chemistry, 2002. **277**(21): p. 19071-9.
131. Tao, Y., et al., *Structure of bacteriophage T4 fibritin: a segmented coiled coil and the role of the C-terminal domain.* Structure, 1997. **5**(6): p. 789-98.
132. Jiang, S. and A.K. Debnath, *Development of HIV entry inhibitors targeted to the coiled-coil regions of gp41.* Biochemical and biophysical research communications, 2000. **269**(3): p. 641-6.
133. Ueda, R., et al., *Immunohistochemical analysis of SOX6 expression in human brain tumors.* Brain Tumor Pathol, 2004. **21**(3): p. 117-20.
134. Ueda, R., et al., *Expression of a transcriptional factor, SOX6, in human gliomas.* Brain Tumor Pathol, 2004. **21**(1): p. 35-8.
135. Ma, M., et al., *Decreased Cofilin1 Expression Is Important for Compaction During Early Mouse Embryo Development.* Biochim Biophys Acta, 2009.
136. Huang, X., et al., *Cloning and characterization of a novel deletion mutant of heterogeneous nuclear ribonucleoprotein M4 from human dendritic cells.* Sci China C Life Sci, 2000. **43**(6): p. 648-54.
137. Shuman, S., *Site-specific DNA cleavage by vaccinia virus DNA topoisomerase I. Role of nucleotide sequence and DNA secondary structure.* J Biol Chem, 1991. **266**(3): p. 1796-803.
138. Shuman, S., *Site-specific interaction of vaccinia virus topoisomerase I with duplex DNA. Minimal DNA substrate for strand cleavage in vitro.* J Biol Chem, 1991. **266**(17): p. 11372-9.
139. Shuman, S., *Site-specific interaction of vaccinia virus topoisomerase I with duplex DNA. Minimal DNA substrate for strand cleavage in vitro.* J Biol Chem, 1991. **266**(30): p. 20576-7.
140. Shuman, S., *Recombination mediated by vaccinia virus DNA topoisomerase I in Escherichia coli is sequence specific.* Proc Natl Acad Sci U S A, 1991. **88**(22): p. 10104-8.

141. Landy, A., *Dynamic, structural, and regulatory aspects of lambda site-specific recombination.* Annu Rev Biochem, 1989. **58**: p. 913-49.
142. Hunt, I., *From gene to protein: a review of new and enabling technologies for multi-parallel protein expression.* Protein Expr Purif, 2005. **40**(1): p. 1-22.
143. BabuRajendran, N., et al., *Structure of Smad1 MH1/DNA complex reveals distinctive rearrangements of BMP and TGF-beta effectors.* Nucleic acids research, 2010. **38**(10): p. 3477-88.
144. Lundblad, J.R., M. Laurance, and R.H. Goodman, *Fluorescence polarization analysis of protein-DNA and protein-protein interactions.* Mol Endocrinol, 1996. **10**(6): p. 607-12.
145. Chen, Y., et al., *The molecular mechanism governing the oncogenic potential of SOX2 in breast cancer.* J Biol Chem, 2008. **283**(26): p. 17969-78.
146. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
147. Conlon, E.M., et al., *Integrating regulatory motif discovery and genome-wide expression analysis.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(6): p. 3339-44.
148. Loh, Y.H., et al., *The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells.* Nat Genet, 2006. **38**(4): p. 431-40.
149. Han, Y. and V. Lefebvre, *L-Sox5 and Sox6 drive expression of the aggrecan gene in cartilage by securing binding of Sox9 to a far-upstream enhancer.* Molecular and cellular biology, 2008. **28**(16): p. 4999-5013.
150. Emans, P.J., et al., *A novel in vivo model to study endochondral bone formation; HIF-1alpha activation and BMP expression.* Bone, 2007. **40**(2): p. 409-18.
151. Tilman, G., et al., *Human periostin gene expression in normal tissues, tumors and melanoma: evidences for periostin production by both stromal and melanoma cells.* Mol Cancer, 2007. **6**: p. 80.
152. Lefebvre, V., R.R. Behringer, and B. de Crombrugghe, *L-Sox5, Sox6 and Sox9 control essential steps of the chondrocyte differentiation pathway.* Osteoarthritis and Cartilage, 2001. **9**: p. S69-S75.
153. Badis, G., et al., *Diversity and complexity in DNA recognition by transcription factors.* Science, 2009. **324**(5935): p. 1720-3.
154. Jordan, S.R., et al., *Systematic variation in DNA length yields highly ordered repressor-operator cocrystals.* Science, 1985. **230**(4732): p. 1383-5.
155. McPherson, A. and B. Cudney, *Searching for silver bullets: an alternative strategy for crystallizing macromolecules.* Journal of structural biology, 2006. **156**(3): p. 387-406.
156. Chayen, N.E., *Methods for separating nucleation and growth in protein crystallisation.* Prog Biophys Mol Biol, 2005. **88**(3): p. 329-37.
157. Rice, P.A., et al., *Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn.* Cell, 1996. **87**(7): p. 1295-306.
158. Ericsson, U.B., et al., *Thermofluor-based high-throughput stability optimization of proteins for structural studies.* Anal Biochem, 2006. **357**(2): p. 289-98.
159. DeLucas, L.J., et al., *Efficient protein crystallization.* J Struct Biol, 2003. **142**(1): p. 188-206.
160. Berger, B., et al., *Predicting coiled coils by use of pairwise residue correlations.* Proceedings of the National Academy of Sciences of the United States of America, 1995. **92**(18): p. 8259-63.

161. Wolf, E., P.S. Kim, and B. Berger, *MultiCoil: a program for predicting two- and three-stranded coiled coils.* Protein science : a publication of the Protein Society, 1997. **6**(6): p. 1179-89.
162. Lupas, A., *Prediction and analysis of coiled-coil structures.* Methods Enzymol, 1996. **266**: p. 513-25.
163. Murakami, A., et al., *SOX6 binds CtBP2 to repress transcription from the Fgf-3 promoter.* Nucleic acids research, 2001. **29**(16): p. 3347-55.
164. Chayen, N.E. and E. Saridakis, *Protein crystallization: from purified protein to diffraction-quality crystal.* Nat Methods, 2008. **5**(2): p. 147-53.
165. Delucas, L.J., et al., *Protein crystallization: virtual screening and optimization.* Prog Biophys Mol Biol, 2005. **88**(3): p. 285-309.
166. Arndt, K.M., et al., *Comparison of in vivo selection and rational design of heterodimeric coiled coils.* Structure, 2002. **10**(9): p. 1235-48.
167. Mason, J.M. and K.M. Arndt, *Coiled Coil Domains: Stability, Specificity, and Biological Implications.* Chembiochem : a European journal of chemical biology, 2004. **5**(2): p. 170-176.
168. Harbury, P.B., et al., *High-resolution protein design with backbone freedom.* Science, 1998. **282**(5393): p. 1462-7.
169. Baxevanis, A.D. and C.R. Vinson, *Interactions of coiled coils in transcription factors: where is the specificity?* Current opinion in genetics & development, 1993. **3**(2): p. 278-85.
170. Graddis, T.J., D.G. Myszka, and I.M. Chaiken, *Controlled formation of model homo- and heterodimer coiled coil polypeptides.* Biochemistry, 1993. **32**(47): p. 12664-71.
171. Monera, O.D., et al., *Comparison of antiparallel and parallel two-stranded alpha-helical coiled-coils. Design, synthesis, and characterization.* The Journal of biological chemistry, 1993. **268**(26): p. 19218-27.
172. Schuermann, M., et al., *Non-leucine residues in the leucine repeats of Fos and Jun contribute to the stability and determine the specificity of dimerization.* Nucleic acids research, 1991. **19**(4): p. 739-46.
173. Kenar, K.T., B. Garcia-Moreno, and E. Freire, *A calorimetric characterization of the salt dependence of the stability of the GCN4 leucine zipper.* Protein science : a publication of the Protein Society, 1995. **4**(9): p. 1934-8.
174. McCoy, A.J., et al., *Likelihood-enhanced fast translation functions.* Acta Crystallogr D Biol Crystallogr, 2005. **61**(Pt 4): p. 458-64.
175. Cowtan, K., *The Buccaneer software for automated model building. 1. Tracing protein chains.* Acta Crystallogr D Biol Crystallogr, 2006. **62**(Pt 9): p. 1002-11.
176. Emsley, P. and K. Cowtan, *Coot: model-building tools for molecular graphics.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 12 Pt 1): p. 2126-32.
177. Murshudov, G.N., A.A. Vagin, and E.J. Dodson, *Refinement of macromolecular structures by the maximum-likelihood method.* Acta Crystallogr D Biol Crystallogr, 1997. **53**(Pt 3): p. 240-55.
178. Afonine, P.V., Grosse-Kunstleve, R. W. and Adams, P. D., *The Phenix refinement framework.* . CCP4 News, http://www.ccp4.ac.uk/newsletters/newsletter42/ articles/Afonine_GrosseKunstleve_Adams_18JUL2005.doc), 2005. **42**.