UNIVERSITÀ DEGLI STUDI DI MILANO

PhD Course in Molecular and Cellular Biology

XXXIII cycle

STRUCTURAL ANALYSIS OF TRANSCRIPTION FACTOR/DNA COMPLEXES

Michela Lapi

PhD thesis

SSD: BIO/10; BIO/18

Scientific tutor: **Marco Nardini**

Academic year: 2019/2020

Thesis performed at:

Dipartimento di Bioscienze, Università degli Studi di Milano

Contents

ABSTRACT

Binding of transcription factors (TFs) to discrete sequences in gene promoters and enhancers is crucial to the process by which genetic information is transferred to biological functions. TF structural analysis is the key to understanding their DNA-binding mode and for the design of specific inhibitors. In this context, the present PhD project focuses on two TFs: (1) NFIX, a TF with unknown structure that binds to the palindromic motif TTGGC(n5)GCCAA and plays an essential role in skeletal muscle development; and (2) NF-Y, a histone-like TF that binds the CCAAT box in promoters of cell cycle genes.

(1) There is a lack of structural information on NFIX and relative TF family members because of the challenge expression and purification of soluble protein constructs in the amount required for structural characterization. We were able to obtain functional NFIX constructs using *E. coli* cells, and to purify them with a high yield. NFIX constructs were tested for correct folding, stability, and DNA-binding through a series of biochemical and biophysical methods. Furthermore, we managed to produce well-diffracting NFIX crystals, which were used for Single Anomalous Diffraction (SAD) phasing. We collected two different datasets of same NFIX construct at 2.7 Å and 3.5 Å resolution, in two different space groups. Structural analysis of this NFIX construct shed first light on this class of TFs and put the bases for the understanding of its DNA-binding mode.

(2) Genomic data of NF-Y locations at gene promoters indicate that NF-Y plays a key role in oncogenic activation. The knowledge of the 3D structure of NF-Y in complex with its CCAAT box provided the rationale for developing inhibitors able to interfere with DNA-binding. The pipeline used to search for NF-Y inhibitors consisted of *in silico* screenings of compounds that interfere with NF-Y functional trimerization and/or with CCAAT box interaction, followed by *in vitro* biochemical/biophysical confirmation of inhibition and X-ray crystallography validation. The selected compound from the initial screening was suramin, which proved to bind to NF-Y and to functionally inhibit the binding of DNA. We obtained suramin/protein complex co-crystals, which diffracted up to 2.3 Å resolution. The crystal structure of the suramin/NF-Y provides the first evidence of NF-Y inhibition by a small molecule.

RIASSUNTO

Il legame di fattori di trascrizione (FT) a specifiche sequenze di promotori genici è fondamentale per il processo mediante il quale le informazioni genetiche vengono trasferite alle funzioni biologiche. L'analisi strutturale dei FT è la chiave per comprendere la loro modalità di legame al DNA e per analizzare specifiche modalità di inibizione. In questo contesto, il presente progetto di dottorato si concentra su due FT: (1) NFIX, un FT con struttura sconosciuta che si lega alla sequenza palindromica TTGGC(n5)GCCAA e svolge un ruolo essenziale nello sviluppo del muscolo scheletrico; e (2) NF-Y, un FT con folding istonico che lega la sequenza CCAAT nei promotori dei geni del ciclo cellulare.
(1) Vi è una mancanza di informazioni strutturali riguardanti NFIX ed i relativi membri della famiglia di FT a causa della difficoltà nell'espressione e nella purificazione di costrutti proteici solubili nella quantità richiesta per la caratterizzazione strutturale. Siamo stati in grado di ottenere costrutti di NFIX funzionali utilizzando cellule di *E. coli* e di purificarli con un'alta resa. I costrutti di NFIX sono stati testati per il corretto folding, stabilità e legame al DNA attraverso una serie di metodi biochimici e biofisici. Inoltre, siamo riusciti a produrre cristalli di NFIX ben diffrattivi, utilizzando il segnale anomalo per ottenere le fasi sperimentali (SAD). Abbiamo raccolto due diversi set di dati dello stesso costrutto NFIX con una risoluzione di 2.7 Å e 3.5 Å, in due diversi gruppi spaziali. L'analisi strutturale di NFIX ha fatto luce su questa classe di FT e ha posto le basi per la comprensione della sua modalità di legame al DNA.
(2) I dati genomici della localizzazione di NF-Y nei promotori dei vari geni indicano che NF-Y gioca un ruolo chiave nell'attivazione oncogenica. La conoscenza della struttura 3D di NF-Y nel complesso con il suo DNA target CCAAT ha fornito il razionale per lo sviluppo di inibitori in grado di interferire con il legame del DNA. La linea sperimentale utilizzata per la ricerca di inibitori di NF-Y consisteva nello screening *in silico* di composti che potevano interferire con la trimerizzazione funzionale NF-Y e/o con l'interazione con il CCAAT, seguita da una conferma biochimica/biofisica *in vitro* dell'inibizione e validazione cristallografica. Il composto selezionato dallo screening iniziale era la suramina, che ha dimostrato di legarsi ad NF-Y e di inibire funzionalmente il suo legame al DNA. Siamo riusciti a ottenere co-cristalli del complesso suramina/proteina, che hanno diffratto ad una risoluzione di 2.3 Å. La struttura cristallina del complesso suramina/NF-Y fornisce la prima prova dell'inibizione di NF-Y da parte di una piccola molecola.

PART I:

TRANSCRIPTION FACTORS

Transcription factors (TFs) are the readers and effectors of the genetic code. The transcription from DNA to messenger RNA is subordinate to the action of TFs, which finely regulate gene expression in the right cell, at the right time, and in the right amount throughout cell life. TFs act in a coordinated fashion to direct cell division, growth, migration, organization, response to signals, and death (Ulasov *et al*, 2018). Indeed, genes are often flanked by several binding sites for distinct TFs, as they cooperate by promoting (as activators) or blocking (as repressors) gene expression (Latchman, 1997). TFs, like most biological pathways, have multiple layers of control. The mechanisms through which TFs regulate gene expression include:

- stabilizing or impairing RNA polymerase binding to DNA;
- histone acetyltransferase (HAT) or histone deacetylase (HDAC) recruitment. HAT acetylates histone proteins, which weakens the association of DNA with histones, making the DNA more accessible, thereby upregulating transcription. On the contrary, HDAC deacetylates histone proteins, which strengthens the association of DNA with histones, making the DNA less accessible and consequently downregulating transcription (Shen *et al*, 2015);
- recruiting coactivator or corepressor proteins to the TF-DNA complex.

TFs can also regulate themselves at the expression level, nuclear localization, activation upon ligand binding, or through post-translational modifications (Lambert *et al*, 2018b).

I.1 TF STRUCTURE

TFs are modular proteins, typically containing the following structural domains: DNA-binding domain (DBD), trans-activating domain (TAD), and optionally a signal-sensing domain (SSD) (Lambert *et al.*, 2018b). The DBD is the defining feature of TFs because it contains the structural motif that recognizes double- or single-stranded DNA in a sequence-specific manner. Other proteins such as coactivators, chromatin remodels, HDACs, HATs, kinases, and methylases also have a great affinity for DNA, but lack DBD, and therefore are not able to influence genes transcription alone (Zhu *et al*, 2016). Indeed, only residues present in the DBD can make functional contacts with site-specific nucleotides,

thus can “read” the correct DNA sequence. The tertiary structure of DBD is the key element in the classification of TFs. The description of major TFs families began in the early 80s. Initially, the most characterized DBDs were zinc finger (ZF), helix-turn-helix (HTH), basic leucine zipper (bZIP), basic helix-loop-helix (bHLH), and high mobility group (HMG) domains (Johnson & McKnight, 1989) (Appendix IV.1 Figure A1). ZFs constitute the largest individual family in the TF group and more than a thousand distinct sequence motifs have been identified in TFs. The structure of the finger is characterized by a short two-stranded antiparallel β sheet followed by an α helix. Two pairs of conserved histidine and cysteine residues in the α helix and second β strand coordinate a single zinc ion. Protein subunits often contain multiple fingers that wrap round the DNA in a spiral manner. The HTH is a common recognition element formed by two almost perpendicular α helices connected by a four-residue β turn or by a longer loop. The second α helix, commonly known as the recognition, or probe, helix, is inserted in the DNA major groove to contact the nucleotide bases. Supporting contacts with the DNA backbone are mainly made by the linker and the first α helix. The HTH motif is typically found in a bundle of three to six α helices, which provide a stabilizing protein core. The “winged” HTH motif is an extension of the HTH group, characterized by the presence of a third α helix and an adjacent β sheet, which provide additional contacts with the DNA backbone. In the bZIP family, the structure of the protein can be split into two parts: the dimerization region and the DNA-binding region. These TFs are dimers formed by subunits of about 60 amino acids. Dimerization is mediated through the formation of a coiled coil by a 30-amino-acid section (Leu-zipper) at the C-terminal end of each helix. The DNA-binding region, also known as the “basic region”, is found as the N-terminus of the subunits as a direct extension of the dimerization region, with the helices that diverge from the coiled coil and enter the DNA major groove, each binding to half of the target. bHLH TFs are a modification of the continuous helices of the bZIP proteins in which the DNA-binding and dimerization regions are separated by a loop, resulting in a four-helix bundle. By separating the two segments, more flexibility is allowed in positioning the probe helices on the nucleic acid.

Later, more DBDs from eukaryotic TFs were characterized, making the picture even more intricate. There are several families with very different functions but all characterized by an α-helical fold, including high-mobility group HMG, MADS, and histones. Other TFs use a β-sheet fold to bind DNA such, for instance, the TATA-binding protein where the antiparallel β-sheet covers the DNA minor groove intercalating Phe sidechains from either end of the sheet. The β-hairpin/ribbon proteins are different from the TATA box-binding protein in that they use smaller two- or three-stranded β-sheet or hairpin motifs to bind

in either the DNA major or minor groove (Appendix IV.1 Figure A1). For a comprehensive overview of the protein-DNA structure classification see (Luscombe et al. 2000). Notably, some TFs do not contain a canonical DBD, and therefore are obliged to hetero-oligomerize in protein complexes to carry out transcription. This is the case of NF-Y (Nardini *et al*, 2013), which will be explained in the following chapters.

While the DBD is typically used to classify the TFS from the structural viewpoint, TAD is the functional/regulator domain of TFs that binds transcription co-regulators. These elements are required for building the transcription machinery that recruits RNA polymerase to the target DNA sequence (Lambert *et al.*, 2018b). TADs dynamically interact with other co-factors and other TFs, acting as hubs in functional protein–protein interaction networks (Gonzalez-Sandoval & Gasser, 2016). TADs have intrinsically disordered regions, whose structural flexibility and conformational adaptability are required to provide TFs with unique functional capabilities, not achievable by rigidly structured regions of a protein (Tsafou *et al*, 2018).

Finally, SSD is an optional element of TFs. It is a ligand-binding domain, which senses external signals and, in response, transmits these signals to the rest of the transcription complex, resulting in up- or down-regulation of gene expression. In addition, DBD and SSD may reside on separate proteins that associate within the transcription complex (Lambert *et al.*, 2018b).

In eukaryotes, DNA is organized into compact particles called nucleosomes, where sequences of about 147 DNA base pairs (bp) make ~1.65 turns around histone protein octamers. Therefore, DNA underneath nucleosomes is inaccessible to many TFs (Li *et al*, 2007; Luger *et al*, 1997). There are two mechanistic classes of TFs: TFs involved in undoing the nucleosome knot, thus allowing access to the naked DNA, and TFs that interact with DNA through a preformed initiation complex. The vast majority of TFs belong to the second case. Such TFs are unable to gain access to repressed or non-modified chromatin domains, even if high-affinity binding sites are present (Rivera & Ren, 2013). The pioneer TFs, instead, are able to bind their DNA binding sites on the nucleosomal DNA. The main function of pioneer TFs is to establish the right conditions for gene expression either by promoting the binding of non-pioneer TFs through direct cooperativity or by recruiting chromatin remodelling/modifying complexes, which in turn physically provide DNA access to other TFs (Wingender, 2013).

I.2 TF FUNCTIONAL MODULATION

Due to their important roles in manipulating cell fate, some human diseases have been associated with mutations in TFs. Many TFs are either tumour suppressors

or oncogenes and, thus, mutations or aberrant regulation of them is associated with cancer (Ulasov *et al.*, 2018). Current strategies to modulate gene expression, during *e.g*. cancer treatment, indirectly affect TFs activity, since it is usually based on inhibition of upstream activating kinases and, therefore, it does not affect one single TF. Targeting TFs would allow a "transcriptome-specific" therapy, improving specific therapeutic intervention by minimizing side effects (Hagenbuchner & Ausserlechner, 2016).

However, targeting TFs is challenging as protein-protein or protein-DNA interactions often involve extensive protein surfaces and lack well-defined pockets for ligand binding, except for ligand-inducible nuclear receptors (Buchwald, 2010). Indeed, TFs have been considered for a long-time as "undruggable" therapeutic targets (Lambert *et al*, 2018a). However, significant breakthroughs in terms of TFs structure, function (expression, degradation, interaction with co-factors, and other proteins) and dynamics of their DNA-binding mode have recently changed this postulate (Papavassiliou & Papavassiliou, 2016). Three major strategies can be adopted to modulate the activity of TFs with small compounds or peptidomimetics. The first strategy focuses on the inhibition of protein-protein interactions since many TFs act as homo- or heterodimers and depend on co-factors complexes for specific DNA recognition. Protein-protein interactions require large interfacing surfaces that, if appropriate structural data is available, can be targeted through virtual screening approaches (Bouhlel *et al*, 2015). One example is p53 which acts *per se* as a tetramer and cooperates with multiple and known partners. A successful inhibition strategy described the inhibition of the p53/MDM2 interaction with compounds RITA (Issaeva *et al*, 2004) and Nutlin-3 (Secchiero *et al*, 2011) (Table A1, section IV.2). With the same approach, the substance NSC13728 was found to impair MYC/MAX heterodimerization (Jiang *et al*, 2009).

The second strategy targets TFs DNA-binding activity with peptidomimetics or small molecules. This strategy focuses on either preventing the target promoter recognition by occupying essential pockets/surfaces on the DBD that impair DNA binding. Therefore, precise structural knowledge of DBD with and without target DNA is an essential starting point for inhibitors development. A successful compound, derived from virtual docking screening and subsequent *in vivo* verification strategy, is inS3-54 that interacts with the DBD of STAT3, inhibiting its transcriptional activity (Huang *et al.*, 2016).

Most recent approaches target chromatin remodelling/epigenetic reader proteins, which are essential for allowing DNA access by TFs. Blocking the function of these proteins, that recognize specific acetylated lysine residues on histones, provides inhibition of oncogenic TFs (Hagenbuchner & Ausserlechner, 2016). An example is the so-called bromodomain and extraterminal (BET) protein

family. BET protein inhibitors occupy the acetyl-lysine binding pocket of these epigenetic readers and cause TF-specific repression of gene expression. These compounds were proved efficient in blocking MYC transcriptional activity in hematopoietic malignancies (Filippakopoulos *et al*, 2012).
In section IV.2 table A1 are reported examples of TF inhibitors in clinical development or in clinical trials. Strategies include inhibition of TF-cofactor protein-protein interactions, inhibition of regulators of TF expression and proteolysis targeting chimaeras (PROTACs) (Bushweller *et al*, 2019).
TFs modulation with the above-mentioned strategies is a field with increasing future potential. In support of this, continued efforts to understand the molecular mechanisms of TF DNA-binding mode, of transcriptional complex formation and structure, will provide clues for the development of drugs for the treatment of life-threatening diseases.

PART II:

NFIX

II.1 INTRODUCTION

II.1.1 NFI genes identification and evolution

The Nuclear Factor I (NFI) family of TFs emerge as critical regulator of many aspects of cell biology, being involved in the regulation of cellular gene expression and viral DNA replication (Harris *et al*, 2015). NFIs were first identified through Adenoviral replication studies. The involvement of TFs in the initiation of DNA replication has been described for several viral systems (Dekker *et al*, 1996). In those cases, TFs interact with components of the initiation complex and thereby enhance the activity and/or the assembly of the replication machinery. Indeed, the first described NFI factor was found as a host-encoded protein required for the initiation of Adenovirus replication (Nagata *et al*, 1982). In this context, NFIs are able to recruit the Adenoviral DNA polymerase complex (pTP-pol) to DNA, which leads to pTP-pol-dependent stimulation of replication (Cleat & Hay, 1989).

Four separate genes encode for the NFI family: *nfia*, *nfib*, *nfic* (Rupp *et al*, 1990) and *nfix* (Kruse *et al*, 1991). Homologs of NFI genes have been found in every vertebrate species examined, from *Xenopus* (Puzianowska-Kuznicka & Shi, 1996; Roulet *et al*, 1995), to mice (Chaudhry *et al*, 1997) and human (Apt *et al*, 1994; Kulkarni & Gronostajski, 1996). A single NFI gene has been identified in *Caenorhabditis elegans*, and *Drosophila melanogaster*. NFI genes are absent in any of the sequenced prokaryotic or simple eukaryotic genomes (Fletcher *et al*, 1999).

The porcine *nfic* and rat *nfia* were the first genes for which the genomic structure was determined, showing strong conservation of all their 11 exons (Meisterernst *et al*, 1989; Xu *et al*, 1997). The human *nfix* genomic sequence likewise possesses 11 exons with sizes comparable to those of rat *nfia* and porcine *nfic* (Gronostajski, 2000). The four murine NFI genes also share the same genomic structure and reside on chromosome 4 (*nfia* and *nfib*), 8 (*nfix*), and 10 (*nfic*), suggesting early duplication of the genes and dispersal throughout the genome. The DBD is encoded by a 532 base-pairs exon in all four NFI vertebrate genes (Figure 1a) with identical splice acceptor and donor sites. In contrast, the *C. elegans nfi-1* and *Drosophila nfi* genes

have phased introns interrupting this DBD-encoding exon (Figure 1b, 1c). These additional introns are missing in the single Amphioxus *nfi* gene (Figure 1d), suggesting that the exons were either inserted recently into the nematode gene, or lost from the cephalochordate gene prior to the duplication of the four genes in the vertebrate lineage (Fletcher *et al.*, 1999).

In vertebrates, a significant sequence homology is evident throughout the NFI genes (Appendix VI.2 Figure A2). However, this homology decreases when the vertebrate genes are compared with those from *C. elegans*. In particular, the lack of conservation outside the NFI DBD between *C. elegans* and vertebrate genes may suggest that changes in function have developed during gene family evolution (Gronostajski, 2000). Moreover, since the vertebrate NFIC proteins are the most divergent NFI members, it has been suggested that NFIC represents the earliest vertebrate NFI gene, and that all the others were derived from it (Fletcher *et al.*, 1999).

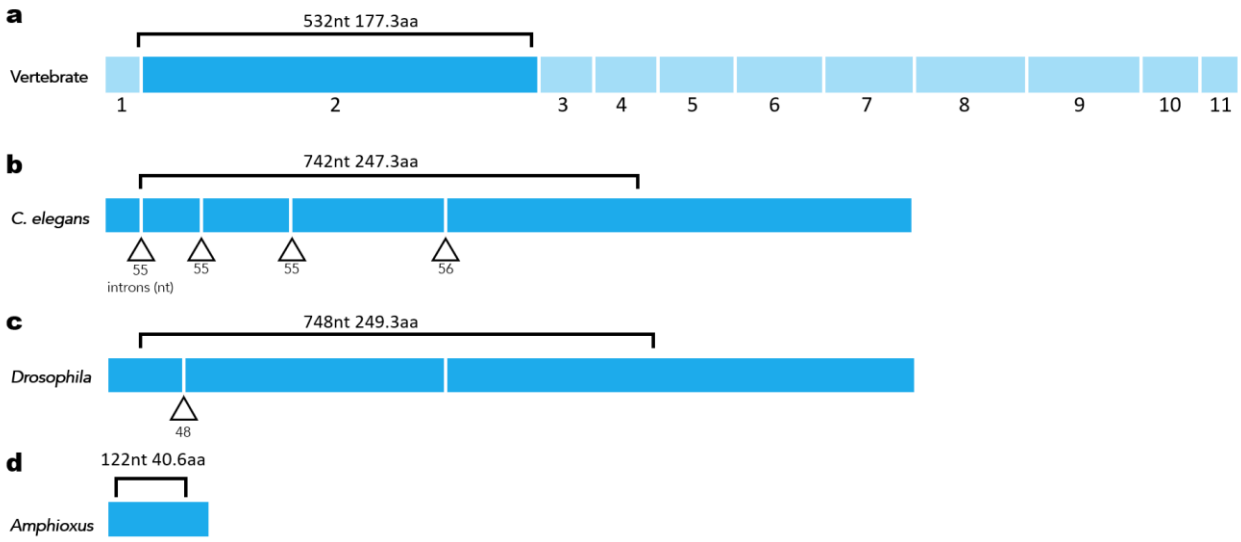


Figure 1. ***Genomic structure of NFI.*** Comparison of NFI genomic structure between **a**) vertebrates, **b**) the nematode *Caenorhabditis elegans*, **c**) *Drosophila melanogaster* and **d**) *Cephalochordata* Amphioxus. The genomic region encoding for the DBD is highlighted in dark blue. The vertebrate NFI exon 2 encodes an uninterrupted 532 nucleotides long DBD. *C. elegans* and *Drosophila* possess larger exons encoding the DBD interrupted by four and one introns for *C. elegans* and *Drosophila*, respectively. Introns within the DBD were marked with triangles. Amphioxus have a single short exon encoding DBD with no intron interruption.

II.1.2 NFI DNA-binding activity

At the molecular level, NFI proteins bind to the palindromic consensus sequence TTGGC(n5)GCCAA on double stranded (ds) DNA as a dimer (Gronostajski, 1986). The binding affinity of the protein dimer for its consensus DNA sequence is in the

nanomolar range (Table 1) (Meisterernst *et al.*, 1989; Rosenfeld *et al*, 1987). Interestingly, NFI monomers can also bind to individual half-sites (TTGGC or GCCAA) with two orders of magnitude less affinity (Table 1) (Meisterernst *et al.*, 1989).

DNA-binding sequence specificity of NFI was challenged by testing a set of oligonucleotides carrying individual or multiple base substitutions. This measurement revealed that single mutations introduced within a half-site result in decreased, but still detectable binding (Table 1). In contrast, when substitutions are spread over both half-sites, DNA-binding activity is abolished (Table 1). Multiple substitutions in the same half-site are generally more deleterious than the corresponding single substitution (Table 1) (Roulet *et al*, 2000).

NFI monomers bind to both DNA half-sites separated by a spacer region of five base-pairs on its consensus sequence. Therefore, similar experiments carried out by varying the (n5) spacer region revealed that the insertion of one or two base-pairs has relatively mild effects on the NFI affinity. Rather, shortening the spacer region by one base-pair lowered the affinity to a value similar to that observed with a single half-site. Arguably, a shorter spacer may impact the simultaneous interaction of NFI monomers with both half-site sequences (Roulet *et al.*, 2000). Additionally, mutations in sequences flanking the consensus site induce minor differences in NFI binding affinity (Gronostajski, 1987).

These data lead to the hypothesis that NFI proteins display flexible DNA-binding rules, allowing the TF to accommodate few base substitutions without losing interactions with neighbouring nucleotides. Therefore, small variations on the canonical sequence may finely tune NFI binding affinity for their target sequence (Roulet *et al.*, 2000).

Dimerization is an integral part of NFI DNA-binding functions. Indeed, protein mutants impairing dimerization activity affect DNA binding. On the other hand, the mutation of sequence-specific nucleotides in the NFI binding site does not impact dimer formation. Thus, NFI dimerization has been proposed to be a specific protein-protein interaction that is the prerequisite for site-specific recognition of DNA (Armentero *et al*, 1994). As the consensus site of NFI is made of three 5bp blocks (5bp of the TTGGC, 5bp spacer, and 5bp of the GCCAA), NFIs contact the DNA through two consecutive major groove turns. Contact-point experiments proved that both NFI monomers accommodate at the same side of the DNA helix (de Vries *et al*, 1987).

Another important property of NFI proteins is their capacity to homo- or hetero-dimerize with other family members (Kruse & Sippel, 1994). Different NFI combinations may exert different roles in regulating gene expression but, so far, the

functional meaning of the preferred NFI dimerization partner was not characterized yet.

Mutations	DNA sequence	Affinity (M^{-1})
NFI consensus sequence	5′-GTCCC**TTGGC**GTGCA**GCCAA**TGCAC-3′	1x10^{9}
Single base substitutions	5′-GTCCC**a**TGGCGTGCA**GCCAA**TGCAC-3′	2x10^{8}
	5′-GTCCC**T**c**GGC**GTGCA**GCCAA**TGCAC-3′	2.5x10^{7}
	5′-GTCCC**TTGG**aGTGCA**GCCAA**TGCAC-3′	1x10^{7}
	5′-GTCCC**T**a**GGC**GTGCA**GCCAA**TGCAC-3′	6.3x10^{6}
	5′-GTCCC**TTGG**gGTGCA**GCCAA**TGCAC-3′	5x10^{6}
	5′-GTCCC**TT**t**GC**GTGCA**GCCAA**TGCAC-3′	1x10^{6}
	5′-GTCCC**TTG**c**C**GTGCA**GCCAA**TGCAC-3′	1x10^{6}
	5′-GTCCC**TT**c**GC**GTGCA**GCCAA**TGCAC-3′	4x10^{5}
Spacing variants (7nt, 6nt, 4nt)	5′-GTCCC**TTGGC**ttgcgca**GCCAA**TGCAC-3′	2.5x10^{7}
	5′-GTCCC**TTGGC**gtgcca**GCCAA**TGCAC-3′	1.3x10^{7}
	5′-GTCCC**TTGGC**gtca**GCCAA**TGCAC-3′	3.2x10^{7}
Half site substitutions	5′-GTCCC**TTGGC**GTGCA**atggt**TGCAC-3′	3.2x10^{7}
	5′-GTCCC**TT**c**GC**GTGCA**atg**A**t**TGCAC-3′	<2x10^{5}
	5′-GTCCC**T**c**G**a**C**GTGCA**a**C**g**A**t**TGCAC-3′	<2x10^{5}
Homologous double substitutions	5′-GTCCC**c**TGGCGTGCA**GCCA**g**TGCAC-3′	6.3x10^{8}
	5′-GTCCC**a**TGGCGTGCA**GCCA**t**TGCAC-3′	3.2x10^{7}
	5′-GTCCC**TTGG**a**GTGCA**t**CCAA**TGCAC-3′	1.3x10^{8}
	5′-GTCCC**TTGG**t**GTGCA**a**CCAA**TGCAC-3′	6.3x10^{6}
	5′-GTCCC**T**g**GGC**GTGCA**GCC**c**A**TGCAC-3′	5x10^{6}
	5′-GTCCC**TTG**c**C**GTGCA**G**g**CAA**TGCAC-3′	<2x10^{5}
Heterologous double substitutions	5′-GTCCC**TTGG**a**GTGCA**GCCA**t**TGCAC-3′	1.3x10^{7}
	5′-GTCCC**TTGG**a**GTGCA**GCCA**g**TGCAC-3′	1x10^{7}
	5′-GTCCC**c**TGG**a**GTGCA**GCCAA**TGCAC-3′	1.6x10^{8}
	5′-GTCCC**a**TGG**a**GTGCA**GCCAA**TGCAC-3′	3.2x10^{7}

Table 1. ***Comparison of NFI affinity to DNA consensus sequence mutations.*** The first row corresponds to the NFI DNA-binding sequence, with the palindromic consensus site marked in bold. Tested DNAs are divided into: oligonucleotides carrying one single base substitution in one half-site, oligonucleotides with distinct spacer lengths, oligonucleotides with substitutions within one or both half-site, oligonucleotides with homologous or heterologous substitutions in both half-sites. Mutated base-pairs are marked with red and low-case letters. For each mutant, binding-affinity data were estimated by EMSA competition (adapted from (Roulet *et al.*, 2000)).

II.1.3 NFI domains

NFI proteins display a modular organization consisting of two functionally distinct domains: the DBD and the TAD, located at the protein N- and C-terminus, respectively (Roulet *et al.*, 1995). A comparison of the NFI amino acid sequences reveals, on one hand, that the N-terminal half represents the most conserved region among the four vertebrate proteins (Appendix VI.2 Figure A2). In facts, the sequence identity of this region is ~90% for NFIA, NFIB, NFIC, and NFIX. On the other hand, the reside conservation in C-terminal half of NFIs significantly diverges (Gounari *et al*, 1990).

The boundaries of the NFI DBD were investigated through the analysis of NFI truncated mutants. The first 240 N-terminal amino acids of NFIC are sufficient to dimerize and bind DNA, and therefore this protein region seems to host both DBD and dimerization domain (Figure 2) (Gounari *et al.*, 1990). Precisely, the region comprising 75-182 residues is necessary for specific DNA base recognition (Dekker *et al.*, 1996), whereas the dimerization domain has been identified within amino acids 170-240 of the protein (Kruse & Sippel, 1994; Mermod *et al*, 1989). Notably, the first 75 N-terminal amino acids contain many basic residues (Figure 2). It has been demonstrated that this region interacts with DNA unspecifically, and therefore it is proposed to make directly contact with DNA through the phosphate backbone at the recognition site or the flanking regions (Gounari *et al.*, 1990).

The NFI proteins contain four cysteine residues (Cys-2, Cys-3, Cys-4, and Cys-5) that are conserved in all NFI DBDs (Figure 2 and Appendix VI.2 Figure A2). The role of these conserved residues was analysed by site-directed mutagenesis and treatment with the oxidizing agent. Mutation of Cys-2, Cys-4, and Cys-5, but not Cys-3 impairs DNA-binding. Treatment of wild-type NFI with an oxidizing agent, diamide, also inactivates DNA binding; but subsequent reduction with the reducing agent dithiothreitol (DTT) restores the binding activity. In contrast, the NFI Cys-3 mutant was the only one resistant to diamide-inactivation. This indicates that the Cys-3 residue is sensitive to oxidation, which leads to inactivation of the protein (Bandyopadhyay & Gronostajski, 1994; Bandyopadhyay *et al*, 1998). Thus, cysteine residues in NFI DBD play an important role in different aspects of protein-DNA interactions. Although Cys-3 is not essential for DNA-binding, it is able to modulate it through its oxidation state. Overall, these studies suggest that redox regulation can modify NFI-DNA interactions (Bandyopadhyay *et al.*, 1998; Dekker *et al.*, 1996).

Transactivation activities of NFI proteins reside in their C-terminal region where no obvious sequence homology is present. In contrast to the DBD, the C-terminal TAD domain may not possess a well-defined, ordered fold. In fact, as for the majority of

TFs, TADs become structured only upon interaction with TF binding partners (Roulet *et al.*, 1995).

C-terminal characterization of NFI pointed out the presence of a proline-rich region (amino acids 399–499) (Figure 2) able to activate transcription when linked to the DBD of a heterologous protein (Spl) (Mermod *et al.*, 1989). Similarly, the proline-rich activation domain of the human NFIC is a strong transcriptional activator *in vitro* when attached to the DBD of Gal4 (Tanese *et al*, 1991). It has been proposed that the NFIC proline-rich sequence is required for specific interactions with other factor(s) that play a role in the initiation of transcription (Mermod *et al.*, 1989). C-terminal proline-rich regions are conserved among vertebrate NFI proteins (Figure 2). In addition, protein interaction screening assays identified a nucleosome component, the H3 histone, as NFI interaction partner. Thus, TAD of NFI can interact with chromatin components through histone H3 upon TGFβ stimuli. These findings suggest that such interactions may regulate chromatin dynamics in response to growth factor signalling (Alevizopoulos *et al*, 1995). An additional function for the NFI C-terminal TAD is auto-inhibition. Indeed, *X. laevis* full-length NFIX showed weak DNA-binding activity *in vitro*, whereas deletion of the C-terminus 176 amino acids improves DNA-binding approximately 4-fold (Roulet *et al.*, 1995).

The NFI dual function as transcriptional activator or repressor is highly dependent on its interaction partner(s). Such an interaction mainly involves the NFI C-terminal domain, however the nature of the interactors is largely unknown (Piper *et al*, 2019). Overall, since NFI proteins have been shown to activate transcription under one condition, and repress it in another, it is likely that repression and activation by NFI proteins will be both cell-type- and promoter-specific, which makes the picture extremely complex (Gronostajski, 2000).

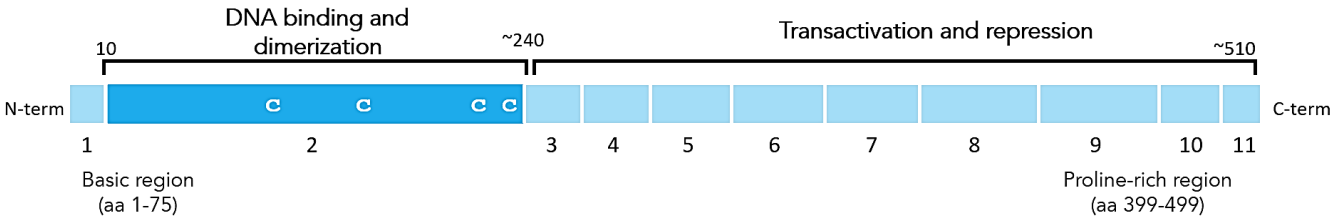


Figure 2. ***Schematic representation of NFI domains.*** NFI vertebrate proteins are composed of 11 exons, whereas exon 2 encodes for DBD at N-terminal half of the proteins. Here, basic residues (amino acids 1-75), involved in directing DNA-binding, are indicated. Also, cysteine residues in the DBD, implied in protein functionality, are marked. Transactivation/repression domain is located at the C-terminal. The proline-rich region (amino acids 399-499) is indicated.

II.1.4 NFI splicing and post translational modifications (PTM)

NFI protein sequences are subjected to considerable alterations, especially in the C-terminal TAD. Alternative splicing is the most widely used mechanism to generate this guided diversity of TFs (Prado *et al*, 2002). All four NFI genes produce several splicing variants, but *nfia* and *nfic* genes give rise to significantly more variations than the *nfib* and *nfix* genes (Grunder *et al*, 2002). Alternative splicing is a finely regulated process, since different relative levels of alternatively spliced NFI transcripts are present in different cell types (Apt *et al.*, 1994; Chaudhry *et al.*, 1997). Changes in the splicing pattern of human NFIC were observed during *in vitro* differentiation of human leukemic cells (Kulkarni & Gronostajski, 1996), whereas NFIX isoform 2 is notable the most expressed splicing variant in mice skeletal muscle (Chaudhry *et al*, 1998). The splicing isoforms of all four NFI transcripts are phylogenetically conserved, suggesting maintained biological meanings (Kruse & Sippel, 1994). However, the functional significance of NFIs variants is still poorly understood.

A less reported NFI modulation, but still significant, is through their PTMs. NFIs undergo several PTMs, mostly occurring in their C-terminal region. To date, most investigations have focused on phosphorylation. Singh et. al. reported that human NFIX phosphorylation affects the intracellular localization of HSF1, a stress-related gene (Singh *et al*, 2009). Furthermore, it has been reported that human NFIB is O-glycosylated in its C-terminal domain. This modification might play a role in the cooperative regulation of *wap* gene transcription by NFIB and STAT5 (Mukhopadhyay & Rosen, 2007). Also, the NFI proline-rich domain contains several stretches of serine and threonine residues, representing a perfect target for certain types of modifications, such as O-linked glycosylation (Jackson & Tjian, 1988) or phosphorylation (Mermod *et al.*, 1989).

Taken together, the widespread (though spatially and temporally unique) expression of the four NFI genes, the alternative splicing of NFI transcripts, their ability to homo- and hetero-dimerize, and their PTMs imply a huge number of NFI regulation mechanisms impacting on their target gene expression (Gronostajski, 2000).

II.1.5 Functional roles of NFI family members

NFIA, NFIB, NFIC, and NFIX regulate a large variety of cellular and viral genes. The analysis of NFIs expression pattern, combined with the generation of *nfi* null mice, showed that NFIs are involved in correct development and post-natal life of tissues. In fact, n*fia* and n*fib* null mice undergo perinatal lethality, whereas *nfix* null mice display very severe developmental defects but they are still viable (Harris *et*

al., 2015). Overall, knock-out experiments demonstrated that the repression of various NFI isoforms results in important changes in a plethora of systems, such as hematopoietic stem cells (Holmfeldt *et al*, 2013), central nervous system (CNS) (Piper *et al*, 2010), skeletal muscle system (Messina *et al*, 2010), lung (Hsu *et al*, 2011), teeth (Park *et al*, 2007), and mammary gland (Nilsson *et al*, 2006). Nowadays, the most characterized role of NFIs regards CNS. NFIA, NFIB, and NFIX all have multifaceted roles in driving the differentiation of stem cells within the developing cerebral cortex and neuronal progenitors within the nascent cerebellum (Piper *et al.*, 2019). All studies converge in the statement that NFIs are important regulators of global stem cell biology. Indeed, most mammalian adult tissues contain a resident stem cell population, which repair and regenerate tissue in response to stress or injury throughout life (Harris *et al.*, 2015). NFIs play a contrasting role during development/post-natal life of tissues. They are involved in correctly promoting stem cell differentiation during tissue development, whereas they promote quiescence and/or survival of stem cells, rather than their differentiation in adult life (Holmfeldt *et al.*, 2013; Martynoga *et al*, 2013). Thus, NFIs are expressed in unique, but overlapping, patterns during embryogenesis and adulthood of several species (Piper *et al.*, 2019).

II.1.5.1 NFIX functions

In recent years, the role of the NFIX TF has been extensively studied, indicating its impact on many aspects of human health. For instance, NFIX is widely expressed within multiple regions of the brain, including the cerebral cortex, cerebellum, and spinal cord (Piper *et al.*, 2019). Moreover, NFIX is expressed in the embryonic mouse telencephalon and within the subventricular zone, suggesting a role for NFIX in nervous system development throughout embryogenesis but also in the homeostasis of neuronal stem cells (Campbell *et al*, 2008). NFIX plays a major role in several aspects of haematopoiesis. Depletion of NFIX significantly reduces the colony-forming potential and repopulating activity of hematopoietic stem and progenitor cells (Holmfeldt *et al.*, 2013). Moreover, loss of NFIX expression affects both myeloid and lymphoid cell differentiation (O'Connor *et al*, 2015). Furthermore, NFIX is linked to multiple human disorders. NFIX is related to several types of cancer, *e.g*. medulloblastoma, squamous cell carcinoma, prostatic cancer, and colorectal cancer (Piper *et al.*, 2019). NFIX variants are the underlying cause of two different syndromes. Malan syndrome is caused by a 21 nucleotide in-frame deletion which implies the loss of 7 amino acids in the DBD of NFIX, resulting in non-functional protein (Priolo *et al*, 2012). While point mutations, mostly clustered in exon 2 encoding for DBD lead to the more severe Marshall–Smith syndrome,

which again involves protein haploinsufficiency or misfunctions (Priolo *et al*, 2018).

Another important role of NFIX lies in muscle tissue. NFIX plays a pivotal role in both skeletal muscle development and post-natal regeneration. In the former case, NFIX leads the transition from primary to secondary myogenesis by both promoting and suppressing specific myogenic genes (Messina *et al.*, 2010). Thus, NFIX-deficient fetuses showed disorganized sarcomerogenesis. Postnatally, NFIX absence induces muscle regeneration delay. Upon muscular damage, NFIX acts through an inhibitory mechanism at the Myostatin promoter in differentiating myoblasts, thereby influencing the commitment of muscle stem cells differentiation during muscle repairing and thus, correct timing of regeneration (Rossi *et al*, 2016). Furthermore, specific deletion of NFIX in macrophages leads to the same muscle regeneration delay, proving that macrophages, as well as muscle stem cells, have a fundamental role upon injury (Saclier *et al*, 2020).

In light of this evidence, NFIX has been recently studied in the pathophysiology of muscular dystrophies. These pathologies comprise a class of genetic diseases that affects the *dystrophin* gene. Dystrophin is required for maintaining muscle fiber architecture during contractions by anchoring the myofiber cytoskeleton to the extracellular matrix surrounding the cells. In muscular dystrophies, dystrophin is absent or dysfunctional, thereby upon every contraction, muscle is damaged. Since the genetic defect is also present in the muscle stem cells, which try to repair the injury, a chronic cycle of degeneration-regeneration is instantiated. This leads in time to irreversible muscle dysfunction and premature death of patients (Gee *et al*, 2017).

Considering these pathological features, the idea is that slowing down muscle regeneration could lag muscle havoc. This was explored through experiments with *nfix* null mice. The regeneration delay, caused by the lack of NFIX, ameliorates the physio-pathological hallmarks of muscular dystrophies. Thus, NFIX absence protects from the progression of the disease by promoting slower exhaustion of muscle regenerative potential and higher protection from damage-induced oxidative stress (Rossi *et al*, 2017).

II.2 AIM

NFIX plays a key role in the regeneration of skeletal muscle tissue and it has been proposed as a pharmacological target for the development of novel therapeutic strategies for muscular dystrophy treatment. Knowledge of the NFIX structure would be an invaluable tool for understanding the molecular mechanism of NFIX, which could be potentially used to identify compounds with inhibitory properties. Hitherto, there is no such piece of information for NFIX nor for other NFI family members. Moreover, there are no updated studies in the literature that offer an extensive characterization of NFIX biochemical and biophysical features.

The aim of the project is to fill this knowledge gap by expressing and purifying NFIX recombinant protein. Then, a biochemical and biophysical study on recombinant NFIX will shed light on NFI-family TF features. Finally, we aim to apply X-ray crystallography techniques to characterize the 3D structure of the TF alone and/or in complex with its target DNA. The NFIX structure would not only be useful in rational drug design for muscular dystrophy treatment, but it can also be a starting point for further studies and characterizations on the other NFI members.

II.3 MATERIALS and METHODS

II.3.1 Bioinformatic analysis

Domain analysis of mice NFIX (splicing variant 2) sequence (Uniprot, identifier: P70257-2) was carried out using Pfam (http://pfam.xfam.org/) (El-Gebali *et al*, 2019) and GlobDoms prediction method by Russell/Linding definition (http://globplot.embl.de/) (Linding *et al*, 2003b). Secondary structure of NFIX N-terminal 240 residues was analysed with SOPMA (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html) (Geourjon & Deleage, 1995), XtalPred (http://xtalpred.godziklab.org/XtalPred-cgi/xtal.pl) (Slabinski *et al*, 2007), Scratch (http://scratch.proteomics.ics.uci.edu/), Jpred (http://www.compbio.dundee.ac.uk/jpred/) (Drozdetskiy *et al*, 2015), INPS-MD (https://inpsmd.biocomp.unibo.it/inpsSuite) (Savojardo *et al*, 2016), YASPIN (https://omictools.com/yaspin-tool) (Lin *et al*, 2005) predictors. The four truncated constructs designed were named NFIX-1 (48-176), NFIX-2 (residues 14-176), NFIX-3 (residues 14-203), NFIX-4 (residues14-240).

II.3.2 Cloning

The cloning protocol was based on the overlap extension PCR method (Bryksin & Matsumura, 2013). Amplification PCR were performed with Phusion® high-fidelity DNA polymerase (Thermofisher). Plasmid template digestion with 20U of DpnI (NEB) was carried out at 37 °C overnight. *E. coli* TOP10 (Thermofisher) cells were used for plasmid amplification. To isolate plasmid DNA, Spin Miniprep kit by Qiagen was used. Cloned vectors sequencing was carried out by Eurofins Genomics. Validation of sequencing products was verified with Pairwise Sequence Alignment EMBOSS Stretcher (https://www.ebi.ac.uk/Tools/psa/emboss_stretcher/). The primers used for construct amplification and the cloning sites, as well as vectors and relative recombinant proteins produced, are schematized in Table 2, all expression cloning vectors tried are listed in section IV.4 figure A4.

	Vector	Cloning sites	Forward primer	Reverse primer
His-NFIX-1	pET-15b	NdeI, BamHI	GCGCGGCAGCCATATGTCAAAGGACGAGG	AACTCAGCTTCCTTTCGGGCTTAGTCCAGTTCTTTGATTGTG
His-NFIX-2	pET-15b	NdeI, BamHI	GCGCGGCAGCCATATGCCGTTTATCGAGGCG	AACTCAGCTTCCTTTCGGGCTTAGTCCAGTTCTTTGATTGTG
His-NFIX-3	pET-15b	NdeI, BamHI	"	AACTCAGCTTCCTTTCGGGCTTAACTATCTGATTGTCCGG
His-NFIX-4	pET-15b	NdeI, BamHI	"	AACTCAGCTTCCTTTCGGGCTTATGATGCAGTCGCAAC
MBP-NFIX-1	pMAL-cRI	EcoRI, HindIII	GGATTTCAGAATTCGGATCCATGTCAAAGGACGAGG	GCCAGTGCCAAGCTTTTAGTCCAGTTCTTTGATTGTG
MBP-NFIX-2	pMAL-cRI	EcoRI, HindIII	GGATTTCAGAATTCGGATCCCCGTTTATCGAGGCGC	"
MBP-NFIX-3	pMAL-cRI	EcoRI, HindIII	"	GCCAGTGCCAAGCTTTTAACTATCTGATTGTCCGG
MBP-NFIX-4	pMAL-cRI	EcoRI, HindIII	"	GCCAGTGCCAAGCTTTTATGATGCAGTCGCAACTGG
MBP-His-NFIX-2	pMAL-cRI	EcoRI, HindIII	GGATTTCAGAATTGGATCCGGCAGCAGCCATCATC	GTGTTAGTTTCTTGACCTGATTTTCGAACCGTGACCG
MBP-His-NFIX-3	pMAL-cRI	EcoRI, HindIII	"	CGGTCACGGTTCGAAAATTGATAGACTAACAGGCC
GST-NFIX-2	pGEX-4t1	EcoRI, XhoI	CGTGGATCCCCGGAATTCCACCCGTTTATCGAGG	GCGGCCGCTCGAGTTAGTCCAGTTCTTTGATTG
GST-NFIX-3	pGEX-4t2	EcoRI, XhoI	"	GCGGCCGCTCGAGTTAACTATCTGATTGTCCGG
LSL-His-NFIX-2	pK-LSLt	EcoRI, HindIII	AACCTGTATTTCCAGGGCGAATTCATGGGCAGCAGCCATC	GTGCGGCCGCAAGCTTTTAGTCCAGTTCTTTGATTGTGAC
LSL-His-NFIX-3	pK-LSLt	EcoRI, HindIII	"	GTGCGGCCGCAAGCTTTTAACTATCTGATTGTCCGGATTCC

Table 2. ***NFIX constructs cloning.*** Recombinant protein, expression vector, cloning sites, forward and reverse primer (5'→3') are listed.

II.3.3 Expression and purification tests

His-NFIX constructs (Table 2) were used for expression tests on four different *E. coli* strains: BL21 (NEB), Rosetta (Novagen), Origami (Novagen), and SHuffle (NEB). For the above-mentioned cell strains, the following procedures were performed. Each transformation was carried out mixing 100 ng of plasmid DNA with 100 µL of cells, heat shocked at 42 °C for 2', and then placed at 4 °C for 2'. Cells were recovered in 1 mL of Luria-Bertani (LB) broth at 37 °C for 45' in agitation. Cells were plated on a petri dish with LB-agar, supplemented with ampicillin (100 µg/mL final). The plate was incubated at 37 °C overnight. A single colony of cell strain tested was used for expression trials using different culture media: LB, LB 2x, 2xTY (Table 3).

Miller's LB broth	LB 2x broth	2xTY broth	Miller's LB Agar
10 g/L peptone 5 g/L yeast extract 10 g/L NaCl	20 g/L peptone 10 g/L yeast extract 20 g/L NaCl	16 g/L tryptone 10 g/L yeast extract 5 g/L NaCl	10 g/L peptone 5 g/L yeast extract 10 g/L NaCl 15 g/L agar

Table 3. ***E. coli culture media for NFIX expression trials.*** Miller's LB broth (Genespin), LB 2x (Sigma-Aldrich), 2xTY broth (Sigma-Aldrich) and Miller's LB Agar (Sigma Aldrich) formulations.

Cells were incubated on a shaker at 37 °C and 220 rpm until they reached an optical density at 600 nm between 0.6 and 0.8. Cell expression was induced with different Isopropil-β-D-1-tiogalattopiranoside (IPTG) concentrations: 0.2 mM, 0.3 mM and 0.5 mM. Two different time-temperature setups after induction were tested: 20 °C overnight, and 37 °C for 4 hours (h). Cell pellet was harvested at 6500 rpm, at 4 °C for 15' using Beckman Coulter centrifuge. Recombinant

protein expression was verified with a single-step Nickel resin chromatography and SDS-PAGE analysis.

For purification trials on the recombinant tagged proteins, a common purification protocol was applied. Composition of purification resins and relative buffers are listed in Table 4. Cells were resuspended by vortexing them in five volumes of lysis buffer (Table 5) and sonicated with Branson 450 Sonifier (20kHz output of sonicate probe for 6', altering 10" on/off). Cell lysate was clarified by centrifugation with a Sorvall RC 6 Plus Centrifuge (Thermo Scientific) and then filtered. The clarified supernatant was incubated with 1 to 10 mL of slurry beads specific resins (Table 4) for 1 h at 4 ºC with agitation. The supernatant was removed by centrifugation at 500 *rcf* and 4 ºC for 5'. The resin was further washed five times with five volumes of wash buffer (Table 4). Recombinant protein elution was performed in five fractions of 1 mL with relative elution buffer (Table 4). Protein expression yield was evaluated by SDS-PAGE.

	Resin	Wash buffer	Elution buffer
His-NFIX	In batch Ni-NTA His·Bind Resin (Sigma-Aldrich)	50mM Imidazole 300 mM NaCl, 50 mM Tris·HCl pH 8.0, 2 mM DTT buffer	200mM Imidazole, 300 mM NaCl, 50 mM Tris·HCl pH 8.0, 2 mM DTT buffer
MBP-NFIX	In batch Amylose resin (NEB)	300 mM NaCl, 50 mM Tris·HCl pH 8.0, 2 mM DTT buffer	50 mM of Maltose, 300 mM NaCl, 50 mM Tris·HCl pH 8.0, 2 mM DTT buffer
GST-NFIX	In batch Glutathione resin (Sigma-Aldrich)	300 mM NaCl, 50 mM Tris·HCl pH 8.0, 2 mM DTT buffer	50 mM reduced glutathione, 300 mM NaCl, 50 mM Tris·HCl pH 8.0, 2 mM DTT buffer
LSL-His-NFIX	In batch Sepharose resin (Sigma-Aldrich)	300 mM NaCl, 50 mM Tris·HCl pH 8.0, 2 mM DTT buffer	50 mM of Lactose, 300 mM NaCl, 50 mM Tris·HCl pH 8.0, 2 mM DTT buffer

Table 4. ***Resin and buffers used in NFIX recombinant protein purification trials.*** List of His-NFIX, MBP-NFIX, GST-NFIX and LSL-His-NFIX resin, wash buffer and elution buffer used.

II.3.4 MBP-His-NFIX expression and purification

The following protocol was selected to carry out all the biophysical and crystallization methods on recombinant NFIX, which results by the best combination of the expression and purification variables studied.

A single colony of transformed SHuffle cells (NEB) was picked to prepare a starter in LB with the addition of ampicillin, 100 µg/mL final. The starter was

incubated at 37 °C overnight. A 1:100 dilution of the starter in LB, supplemented with 100 µg/mL of ampicillin, was incubated at 37 °C and protein expression was induced at 0.6 optical density at 600 nm with IPTG at 0.3 mM final concentration. After IPTG induction, culture medium was incubated at 20 °C overnight. Cell pellets were harvested at 6500 rpm, at 4 °C for 15' using Beckman Coulter centrifuge. Cell pellet was flash frozen in liquid nitrogen and stored at -20 °C.

Lysis buffer	Binding buffer/ buffer A	Elution buffer/ buffer B	1x Cleavage buffer
200 mM NaCl 50 mM HEPES pH 7.0 2 mM DTT 0.1 mM Phenylmethylsulfonyl fluoride (PMSF) 0.1 mg/mL DNase I	200 mM NaCl 50 mM HEPES pH 7.0 2 mM DTT	2 M NaCl 50 mM HEPES pH 7.0 2 mM DTT	50 mM Tris·HCl, pH 8.0 10 mM $CaCl_2$

Table 5. ***List of buffers used in NFIX purification.*** Lysis buffer, binding buffer, elution buffer and cleavage buffer compositions are reported.

The cell pellet was thawed on ice and resuspended with five volumes of lysis buffer with respect to the cell pellet mass. Cells were sonicated for 6' at 20 kHz frequency, alternating 10" of sonication and 10" of rest (Branson Sonifier 450). Centrifugation at 18000 rpm for 40' at 4 °C with Sorvall R-6 Plus (LSCF) was applied to separate the insoluble lysate fraction. Supernatant was then filtered with a 0.44 µm filter.

The clarified cell lysate was loaded into HiTrap Heparin HP 5 mL column (GE Healthcare) with a peristaltic pump (GE Healthcare) at a flow rate of 0.5 mL/min at RT. Wash and elution were performed with the ÄKTA pure (GE Healthcare) system at 4 °C. The unbound fraction was washed out using binding buffer at a flow rate of 2 mL/min for 30'. The bound fraction was eluted with a gradient between buffer A and B (Gradient setup: 25% of buffer B for 20' at 2 mL/min). Eluted recombinant protein was collected in fractions of 2 mL. Thrombin CleanCleave Kit (Sigma-Aldrich) was used to cleave NFIX from the fusion tags. For the reaction, we incubated 100 µL of thrombin-agarose 50% (v/v) suspension *per* mg of fusion protein at 20 °C overnight in 1x cleavage buffer (Table 5). A second Heparin chromatography step was used to separate NFIX from the uncleaved fusion protein and the MBP-His tag. The sample was firstly diluted with 50 mM HEPES, pH 7.0, and 2 mM DTT buffer to decrease the salt concentration to 200 mM, and then loaded into an HiTrap Heparin 5 mL column

(GE Healthcare). For elution, a 25% buffer B gradient was set to 2 mL/min for 40'. NFIX peak fractions were collected and concentrated with an Amicon protein concentrator (Merck Millipore), 10kDa cut-off, to a final volume of 1 mL, to be subsequently loaded in HiLoad Superdex 75 pg prepacked column (GE Healthcare) using the ÄKTA pure system (GE Healthcare). The protein sample was injected in a 2 mL capillary loop. Size-exclusion chromatography (SEC) was performed using the binding buffer as running buffer and a flow rate of 1.0 mL/min. Eluted NFIX was collected in 2 mL fractions. Protein purity was then evaluated by SDS PAGE (Bio-Rad), concentrated with an Amicon protein concentrator (Merck Millipore), and stored at -20 °C (see section IV.5 for complete NFIX-2 and NFIX-3 purifications).

II.3.5 Circular dichroism (CD) spectroscopy

CD spectroscopy was used to determine the secondary and tertiary structure content of the NFIX-2 and NFIX-3 constructs. CD spectra were recorded with Jasco J-810 instrument equipped with a PFD-425S Peltier temperature controller (Jasco Europe, LC Italy).

Measurements were made in Far-UV spectral region, using 0.1 cm quartz cuvettes and 0.2 mg/mL of protein concentration. Spectra were collected from 260 to 200 nm. For CD measurements, protein buffer only contained 50 mM Tris-HCl, pH 8.0. The CD data were presented in terms of the mean residue ellipticity (mdeg) as a function of wavelength (nm).

The thermodynamics of protein unfolding was investigated by monitoring the ellipticity variations at a single wavelength of 222 nm. Temperature ramp was from 20 to 95 °C, with 1 °C/minute rate of heating. For denaturation ramp experiments, protein buffer contained 50 mM Tris-HCl, pH 8.0, and 200 mM NaCl. Unfolding curves were represented as ellipticity *versus* temperature.

CD measurements were performed in collaboration with prof. Alberto Barbiroli, Dept. of Food, Environmental and Nutritional Sciences, University of Milan.

II.3.6 Nuclear Magnetic Resonance (NMR)

NMR spectra were recorded at 25 °C on a Bruker Advance 600 Ultra Shield TM Plus 600-MHz Spectrometer equipped with triple-resonance cryoprobe and pulsed field gradients. The ^{1}H (proton) NMR measurement on NFIX-2 was carried out using 0.2 mg/mL protein sample concentration in 10% D_20 buffer.

^{1}H-1D NMR spectrum was recorded with an acquisition time of 2 seconds, a recycle time of 6 seconds to minimize peak saturation. ^{1}H-1D NMR spectrum was processed with zero filling to 131,000 points and apodized with an unshifted Gaussian and a 0.5-Hz line broadening exponential. The spectrum was phased and base-plane corrected before peak integration. The global spectrum deconvolution algorithm implemented in the Mnova 9.0 software package of Mestrelab was used to deconvolve and integrate the spectrum. NMR experiments were performed in collaboration with Dr. Giovanna Musco, IRCCS Ospedale San Raffaele, Milan.

II.3.7 Thermal shift assay (Thermofluor)

Thermofluor analysis was carried out on construct NFIX-2 using CFX Real-time PCR Detection system (Bio-Rad) and SYPRO orange dye (Sigma-Aldrich). SYPRO orange stock at 50000x concentration was firstly diluted to 50x. Reactions were carried out in CFX 96-well plate (Bio-Rad) with a final volume of 20 µL *per* well. Reaction mix included protein (final concentration 25 µM), 50x SYPRO orange (final concentration 5x), and a screening buffer (final concentration 1x).

Buffer screening compositions:

Columns 1 to 6 of the plate differ in NaCl concentrations: 0 mM, 100 mM, 200 mM, 300 mM, 400 mM and 500 mM.

Rows A to D differ in pH solution: 50 mM HEPES, pH 6.0; 50 mM HEPES, pH 7.0; 50 mM Tris-HCl, pH 8.0; 50 mM Tris-HCl, pH 9.0.

Each experimental condition was prepared in triplicate; an extra well was used for blank (condition A1 without protein).

Program set up: 1 °C step increases *per* minute from 25 °C to 75 °C, *i.e.* 50 cycles, Exc/em 470-505/540-700 nm.

The CFX Maestro software (Bio-Rad) was used to calculate melting temperature from the derivative of the sigmoidal melting curves.

II.3.8 Dynamic Light Scattering (DLS)

DLS measurements were performed for construct NFIX-2 (shown) and NFIX-3 (not shown) using pUNk (Unchained Labs) instrument. Protein concentration was 1.0 mg/mL. The laser source was an Argon ion laser tuned at 514 nm, with following setups: Laser: 100%, Intensity: 315,140 counts/s, Intercept: 0.879, Attenuator: 57%. Experiments was leaded in a cylindrical quartz cell (Hellma GmbH & Co, Germany) at 20 °C using MW model: Globular Proteins, Radius:

6.10 nm, Standard Deviation: 10.24 nm. The experiment was led in buffer A (see II.3.4 section).

II.3.9 SDS-Polyacrylamide Gel Electrophoresis (PAGE)

For protein electrophoresis we used precast 12% tris-glycine polyacrylamide gel (Genescript) with MOPS buffer 1x (Bio-Rad). About 250 µg of protein sample were mixed with NuPAGE LDS Sample Buffer (4x) (Thermofisher).

The gel was stained with a solution of Coomassie Brilliant Blue R-250 (0.5% [*w/v*] Brilliant Blue R-250, 20% [*v/v*] ethanol and 10% [*v/v*] acetic acid) for 20' Gel destaining was performed with a solution containing 10% acetic acid (*v/v),* 20% ethanol (*v/v*) in H_2O and incubating for 2 h.

II.3.10 Native gel electrophoresis

Non-denaturing electrophoretic gel composition was:

Acrylamide (29:1) 6% *(v/v)* final concentration,
TBE 0.5x final concentration,
Glycerol 1.25% *(v/v)* final concentration,
Ammonium persulfate 9 µM final concentration,
TEMED 13 µM final concentration.

TBE 0.25x (725 mg Tris, 3.37 g Boric acid, 250 µL EDTA 0.5 M pH 8.0) was used as running buffer. Native gels were run at 4 °C, 100 V, for 45'.

II.3.11 Electrophoretic Mobility Shift Assays (EMSA)

A 31bp dsDNA carrying the palindromic NFI consensus sequence, labelled with Cyanidin 5 (Cy5) at the 5' end of the forward strand was used as a probe. Duplex DNA was annealed by heating forward and reverse oligos at 92 °C and cooling down to RT overnight. DNA oligonucleotides were purchased from Eurofins genomics. A list of labeled and unlabeled oligonucleotide probes used in the thesis are reported in figure 3.

Dose response experiments were performed by combining 14 µL of a premixed Binding Mix (BM), containing 20 nM NFI Cy5-probe, 20 mM Tris-HCl pH 7.5, 50 mM NaCl, 5 mM $MgCl_2$, 30 mM KCl, 0.5 mM EDTA, 6.5% *(v/v)* glycerol, 2.5 mM DTT, 0.1 mg/mL BSA, 1 ng/mL Poly (dIdC), with 2 µL of protein serial dilutions (protein final concentration in the experiment: 0 nM, 60 nM, 180 nM, 540 nM, 1.6 µM). The mixtures were incubated at 30 °C for 30' in the dark and then loaded on a 6% polyacrylamide native gel.

In EMSA competition experiments, the BM was supplemented with 150 nM of NFIX-3 (final composition: 20 nM NFI Cy5-probe, 20 mM Tris· HCl pH 7.5, 50 mM NaCl, 5 mM $MgCl_2$, 30 mM KCl, 0.5 mM EDTA, 6.5% *(v/v)* glycerol, 2.5 mM DTT, 0.1 mg/mL BSA, 1 ng/mL Poly (dIdC), 150 nM NFIX-3). 2 µL of unlabeled (cold) oligos serial dilutions were mixed with 14 µL of the BM. Two sets of competitors were used: an unrelated 31bp random sequence DNA *versus* the 31bp NFI DNA competitor (Figure 3). The competitor concentrations for the experiment with the first set were 1x, 1.25x, 6.25x, 31.25x respect to the Cy5-probe. The experiment with the second set was prepared with concentrations of cold competitors of 10x and 50x with respect to the Cy5-probe concentration.
The kinetic EMSA assay was carried out with a BM: consisting in 20 nM NFI Cy5-probe, 20 mM Tris· HCl pH 7.5, 50 mM NaCl, 5 mM $MgCl_2$, 30 mM KCl, 0.5 mM EDTA, 6.5% *(v/v)* glycerol, 2.5 mM DTT, 0.1 mg/mL BSA, 1 ng/mL Poly(dIdC). Competitor oligos were used at 10x and 50x final concentration respect to the probe and NFIX-3 final concentration was 150 nM. Two different set-ups for the experiment were used. For the first, 12 µL of binding mix were mixed with 2 µL of competitor and then 2 µL of NFIX-3 were added. The second experiment was performed by mixing 12 µL of BM with 2 µL of NFIX-3, and after 30' of incubation at 30 °C in dark, 2 µL of competitor were added.
Cy5 chemiluminescence was acquired with ChemidocTM MP apparatus (Bio-Rad) with an exposure time of 5".

```
5′[Cy5]-GGGTCTCTTTGGCAGGCAGCCAACCAGCAAA-3′
5′[Cy5]-GGGTCTCTTTGGCAGCAGCCAACCAGCAAA-3′

31bp:5′-GGGTCTCTTTGGCAGGCAGCCAACCAGCAAA-3′
25bp:    5′-TCTCTTTGGCAGGCAGCCAACCAGC-3′
23bp:     5′ CTCTTTGGCAGGCAGCCAACCAG-3′
21bp:      5′-TCTTTGGCAGGCAGCCAACCA-3′
19bp:       5′ CTTTGGCAGGCAGCCAACC-3′
17bp:        5′-TTTGGCAGGCAGCCAAC-3′

Urelated:5′-AGTTATGGTAACCATGGATTCAGGCGGCCAT-3′
```

Figure 3. ***DNA probes used in EMSA experiments*****.** Above the Cy5-labeled probes and below the cold competitors used. Presence of Cy5 in highlighted in yellow. NFI binding-sequence was bolded and the spacer for the labeled oligos was underlined.

II.3.12 DNA-protein complex formation and purification in solution

To form the protein-DNA complex, we mixed freshly purified protein and dsDNA oligo at a 2:1 molar ratio and we decreased salt concentration to 50 mM by adding 50 mM HEPES pH 7.0 buffer containing 2 mM DTT. To isolate the

complex in solution, we performed a SEC with HiLoad Superdex 200 pg prepacked column (GE Healthcare) using binding buffer (Table 5) as running buffer at 1 mL/min flow rate. Complex formation was verified using on a native gel and by measuring the UV absorbance at 260 nm for DNA and 280 nm for protein components. The latter parameter was used to estimate the final concentration of the protein in the complex.

II.3.13 Flame Atomic Absorption Spectrometry (FAAS)

In order to determine the binding of Zn^{2+} to NFIX-2, a homogeneous solution of the protein (0.2 mg/mL) was exposed to wet decomposition. To the sample, 5 mL of nitric acid (65%) were added, and then the solution was gently boiled on a hot plate until the red fumes coming from the beaker terminate (1 h). After cooling, 5 mL of a freshly prepared mixture of 65% nitric acid and 37% hydrochloric acid (1:3) were added and the solution was warmed on a hot plate for 2 h. After cooling, about 3 mL of hydrogen peroxide (30%) were added and the solution was boiled again to evaporate until a small portion remained. After cooling, the resulting clear digested solution was quantitatively diluted to a final volume of 10 mL with 2% nitric acid before being analyzed by FAAS. Nitric acid (65% w/w), hydrochloric acid (37% w/w), hydrogen peroxide (30% w/w) were purchased from Sigma Aldrich.

The calibration curve was based on five standards (including the blank). Working calibration solutions were freshly prepared using appropriate stepwise dilutions of standard Zn stock solution (Ultra grade, 1000 mg/L, 2% HNO_3, Perkin-Elmer). The working standards were as follows: 0.25, 0.5, 1 and 1.4 ppm diluted in 2% nitric acid.

Data were collected by using an atomic absorption spectrometer (PinAAcle 900T, Perkin-Elmer) equipped with deuterium lamp for background correction, air-acetylene flame and zinc hollow-cathode lamp operating at 213.9 nm. The linear range was 0.01-2 mg/L. Using the standard calibration graph, measurements were performed in triplicate and the mean was automatically calculated. FAAS measurements were performed in collaboration with Dr. Silvia Cauteruccio, Dept. Chemistry, University of Milan.

II.3.14 Crystallization experiments

For initial crystallization trials, we used the following commercial screens: Crystal screen I/II, Index and PEG/Ion from Hampton Research; PACT,

MacroSol, Proplex, Morpheus, JCSG and Wizard from Molecular Dimension; JBScreen Classic I, II, III and IV from Jena Biosciences (See section IV.7, table A2). Automated dispensing was carried out using the Oryx8 crystallization robot (Douglas Instruments). Greiner 96-well flat-bottomed CrystalQuick plates and Douglas 96-well Vapor Batch Plates (Douglas Instruments) were used for vapor diffusion (sitting drop) and microbatch crystallization methods, respectively. For sitting drop experiments, we dispensed three different protein:precipitant ratios, 30%, 50% and 70% (*v/v*) that were equilibrated against 100 µl of precipitant in each well. For microbatch experiments, a constant protein:precipitant ratio of 50% (*v/v*) was dispensed. 9 mL of a 1:1 paraffin:silicon oil mixture was pipetted over the microbatch wells.
For the optimization of condition B2 of JCSG screening (0.2 M NaSCN, 20% PEG 3350) the following parameters were screened:

0.2 M NaSCN PEG 3350 **18%** 0.1 M HEPES **pH 6.8**	0.2 M NaSCN PEG 3350 **18%** 0.1 M HEPES **pH 7.0**	0.2 M NaSCN PEG 3350 **18%** 0.1 M HEPES **pH 7.2**
0.2 M NaSCN PEG 3350 **20%** 0.1 M HEPES **pH 6.8**	0.2 M NaSCN PEG 3350 **20%** 0.1 M HEPES **pH 7.0**	0.2 M NaSCN PEG 3350 **20%** 0.1 M HEPES **pH 7.2**
0.2 M NaSCN PEG 3350 **22%** 0.1 M HEPES **pH 6.8**	0.2 M NaSCN PEG 3350 **22%** 0.1 M HEPES **pH 7.0**	0.2 M NaSCN PEG 3350 **22%** 0.1 M HEPES **pH 7.2**
0.2 M NaSCN PEG 3350 **24%** 0.1 M HEPES **pH 6.8**	0.2 M NaSCN PEG 3350 **24%** 0.1 M HEPES **pH 7.0**	0.2 M NaSCN PEG 3350 **24%** 0.1 M HEPES **pH 7.2**

The best crystals were obtained from 0.2 M NaSCN, 0.1 M HEPES pH 7.0, 22% PEG 3350 condition. Crystals were fished with CrystalCap™ SPINE (Hampton), soaked in cryo protectant solution (NaSCN 0.2 M, HEPES pH 7.0 0.1 M, PEG 3350 22% [*v/v*], Glycerol 20% [*v/v*]) and flash frozen in liquid nitrogen.

II.3.15 Data collection and structure determination

X-ray diffraction data collections were performed at the XRD2 beamline of the ELETTRA Synchrotron, Trieste (Italy), equipped with a Pilatus 6M hybrid pixel area detector (Dectris, CH). The best dataset used for phasing was collected at 100 K at a wavelength of 1.2705Å. The NFIX-2 structure has been solved by SAD method in the $P2_1$ spacegroup, exploiting the anomalous signal of bound zinc atoms. Data were indexed, integrated, and scaled using XDS (Kabsch,

2010), and the heavy atom substructure was determined using SHELX C/D/E (Sheldrick, 2010). The program SHELXC provides a statistical analysis of the input data, estimates the marker-atom structure factors F_A and the phase shifts α and sets up the files for the other two programs. SHELXD is used for solving the sub-structure (*i.e*. locating the marker atoms) and SHELXE (Schneider & Sheldrick, 2002) provides iterative phase improvement by density modification. Finally, ARP/wARP was used for automated model building (Langer *et al*., 2008). The structure was refined using REFMAC (Murshudov *et al*, 1997) and Phenix.refine (Adams *et al*, 2010), and multiple rounds of manual model building in Coot (Emsley *et al*, 2010).

Another dataset was collected on a $P4_12_12$ crystal form that diffracted to a maximum resolution of 3.5 Å. These crystals were obtained in a condition identical to the $P2_1$ crystal form. Diffraction data were collected at the i041 beamline at DIAMOND synchrotron (UK), and reduced and scaled using XDS and AIMLESS from the CCP4 package respectively (Winn *et al*, 2011).

The structure was solved by Molecular replacement method with the program Phaser (McCoy *et al*, 2007) by using the $P2_1$ crystal form protein structure as a model. At the end of the refinement, using REFMAC (Murshudov *et al*., 1997), ligands were located through the inspection of difference Fourier maps using Coot (Emsley *et al*., 2010). The programs MolProbity (Chen *et al*, 2010) and PISA (Krissinel, 2015) were used to assess the stereochemical quality and to analyse protein quaternary assembly, respectively.

II.4 RESULTS

II.4.1 NFIX sequence analysis and construct design

The biochemical and structural characterization of a protein requires the availability of high yields of the target in soluble, stable, homogeneous, and biologically-active form. Therefore, we analysed the primary structure of NFIX using bioinformatic tools that identify conserved functional domains, putative disordered regions, secondary structure elements and motifs that could guide us to design the best NFIX constructs for 3D structural studies. It should be noted that recombinant NFI proteins have never been expressed and purified before with yields needed for structural biology experiments, mostly because of their solubility limitations.

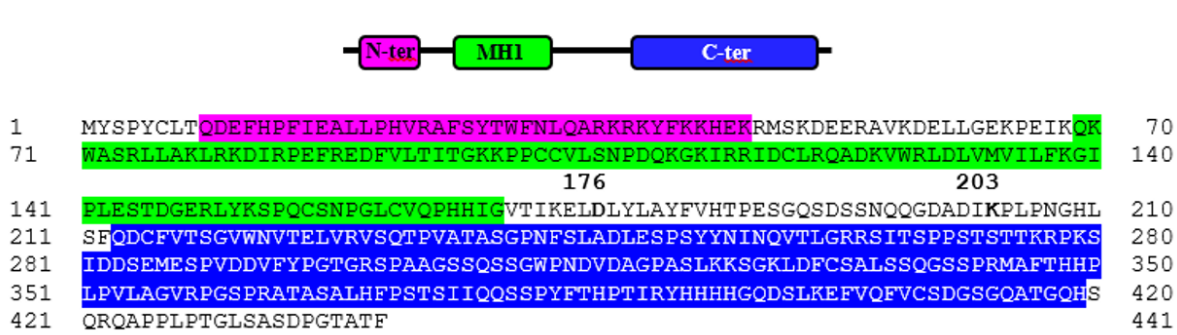
NFIX sequence analysis with Pfam (El-Gebali *et al.*, 2019) identified three main domain regions: a basic-rich region at the N-terminus (residues 9-46), an MH1 (MAD homology 1) DBD (residues 69-169), and a C-terminal regulatory domain (residues 213-419) (Figure 4a). Further analysis in search for disordered/flexible regions with GlobPlot (Linding *et al.*, 2003b) assigned a potential globular domain to the first 145 residues, recognizing a Pfam domain in the 67-175 region, while the C-terminal regulatory domain is mostly characterized by predicted disordered residues (Figure 4b). These conclusions are supported by the analysis performed by the Protein disorder prediction server DisEMBL (Linding *et al*, 2003a), on loops/coils, hot loops (high B-factors), and missing coordinates prediction (Figure 4c).

We then performed a secondary structure prediction (Figure 4d). All predictors suggested a high secondary structure content for the first 185 residues, mostly enriched in α-helices with a few short strands, in opposition to the C-terminal region, which does not contain secondary structure elements. Considering that the Pfam analysis suggested the presence of a MH1 domain, we ran a BLAST search (https://blast.ncbi.nlm.nih.gov/Blast.cgi) against the Protein Data Bank (https://www.rcsb.org/) to search for structure homologs. The only match was identified for the MH1 domain (residues 50-170) with Smad proteins, with an overall sequence identity between 17-19% over a 24-26% residue alignment (Figure 4d). This similarity is low and localized, as expected, only at the MH1 domain (residues 50-170), including the presence of a CCCH motif (Cys103, Cys156, Cys162, and His167) that in the Smad MH1 forms a Zn^{2+}-binding site (Figures 4d and 14).

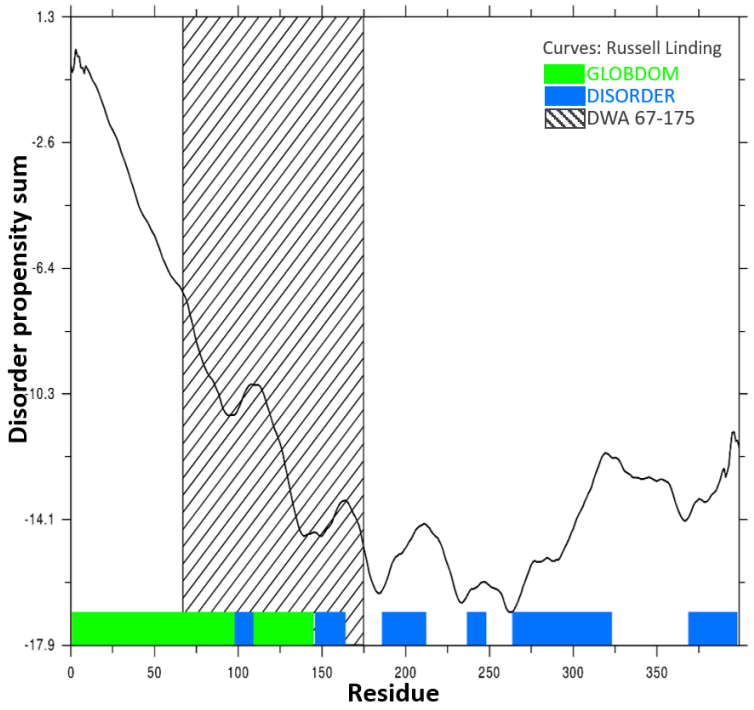
Based on our bioinformatics analyses, we designed four alternative NFIX constructs, named NFIX-1, NFIX-2, NFIX-3, and NFIX-4 (Figure 4e). They all

contain the identified putative DBD and lack partly or completely the predicted disordered C-terminal region, which would likely hinder crystallization. NFIX-1 comprises residues 48-176 that we identified as the minimal DBD construct. NFIX-2 (residues 14-176) contains an extra 34 residues at the N-terminus with respect to NFIX-1. We decided to remove the first 13 N-terminal residues that are predicted not to adopt secondary structure elements and that contain an isolated, potentially solvent-exposed cysteine residue (Cys6), which could induce aggregation in solution by non-specific disulphide-bridge formation. From the literature, it is known that a Cys to Ser substitution of this residue has no effect on binding in NFIC (Novak *et al*, 1992), so we did not expect it to have a DNA-binding role in NFIX. NFIX-3 (residues 14-203) covers 27 additional residues at the C-terminus compared to NFIX-2, to include a predicted extra α-helix that could be important for the DBD folding. Finally, NFIX-4 (residues14-240) was designed to overextend the C-terminus to better explore its solubility/folding limits.

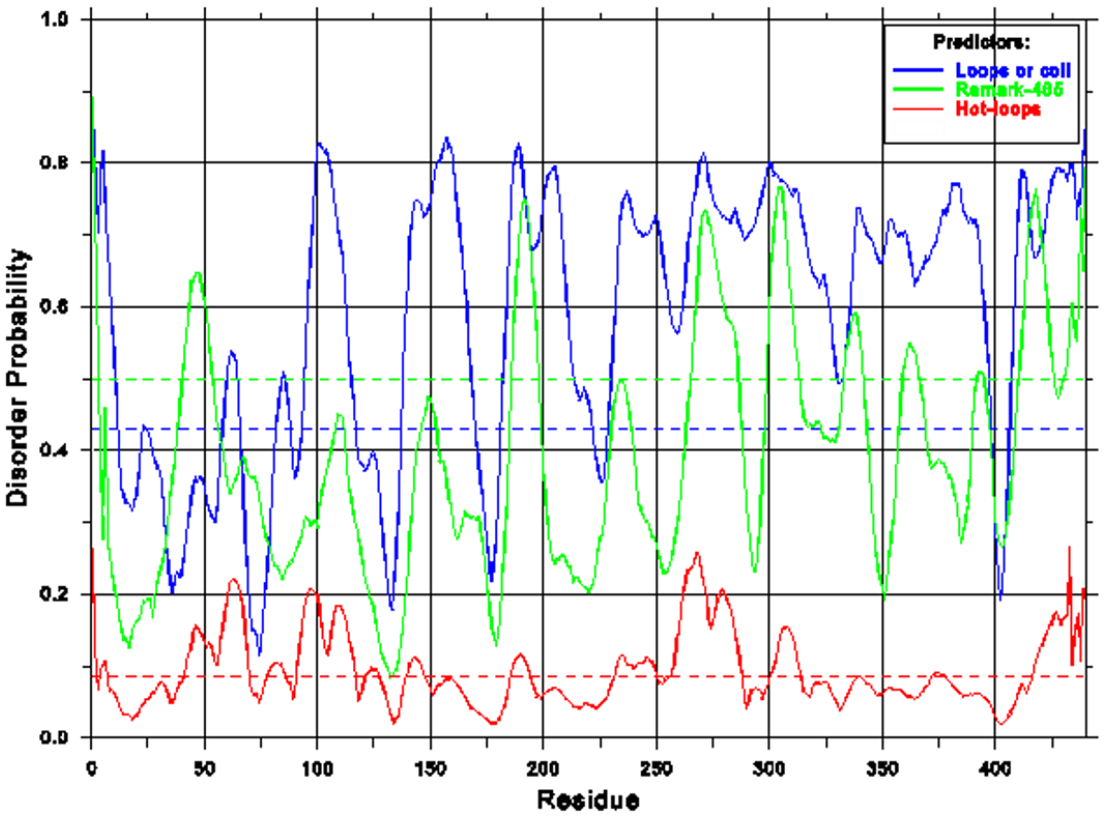
a



b


Curves: Russell Linding
GLOBDOM
DISORDER
DWA 67-175
Disorder propensity sum
1.3
-2.6
-6.4
-10.3
-14.1
-17.9
0
50
100
150
200
250
300
350
Residue

c


Predictors:
Loops or coil
Remark-465
Hot-loops
Disorder Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
350
400
Residue

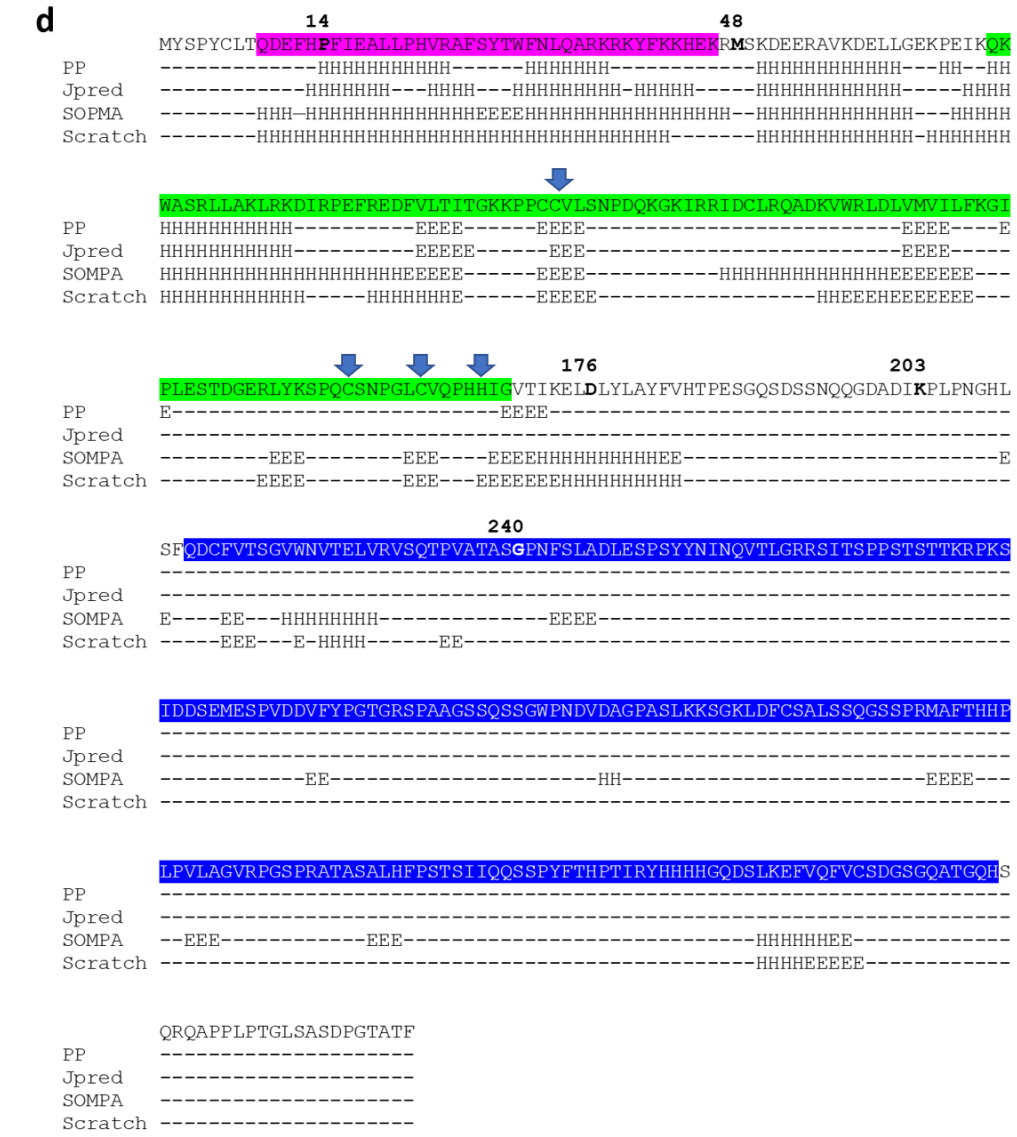

d
14
48
MYSPYCLTQDEFHPFIEALLPHVRAFSYTWFNLQARKRKYFKKHEKRMSKDEERAVKDELLGEKPEIKQK
PP
Jpred
SOPMA
Scratch
WASRLLAKLRKDIRPEFREDFVLTITGKKPPCCVLSNPDQKGKIRRIDCLRQADKVWRLDLVMVILFKGI
PP
Jpred
SOMPA
Scratch
176
203
PLESTDGERLYKSPQCSNPGLCVQPHHIGVTIKELDLYLAYFVHTPESGQSDSSNQQGDADIKPLPNGHL
PP
Jpred
SOMPA
Scratch
240
SFQDCFVTSGVWNVTELVRVSQTPVATASGPNFSLADLESPSYYNINQVTLGRRSITSPPSTSTTKRPKS
PP
Jpred
SOMPA
Scratch
IDDSEMESPVDDVFYPGTGRSPAAGSSQSSGWPNDVDAGPASLKKSGKLDFCSALSSQGSSPRMAFTHHP
PP
Jpred
SOMPA
Scratch
LPVLAGVRPGSPRATASALHFPSTSIIQQSSPYFTHPTIRYHHHHGQDSLKEFVQFVCSDGSGQATGQHS
PP
Jpred
SOMPA
Scratch
QRQAPPLPTGLSASDPGTATF
PP
Jpred
SOMPA
Scratch
e
1 14 48
176 203 240
442
NFIX-fl
N-ter
MH1
C-ter
NFIX-1
NFIX-2
NFIX-3
NFIX-4

Figure 4. ***Sequence analysis and construct design of NFIX*****. a**) Pfam domain analysis. The N-terminal basic region, the MH1 domain, and the C-terminal regulatory region are shown in magenta, green, and blue, respectively. **b**) GlobPlot analysis. The predicted globular and disordered regions are shown in green and blue, respectively. The Pfam domain is shown in dashes. **c**) The output from the DisEMBL web server. The green curve is the prediction for missing coordinates, red for the hot loop network and blue for coil. Horizontal lines correspond to the random expectation level for each predictor. From this plot it is seen that the predictors agree on residues 1-10 and 175-441 as being disordered. **d**) Secondary structure prediction with ProteinPredict, Jpred, SOMPA, and Scratch. Predicted α-helices are indicated as H and β-strands as E. Conserved CCCH motif is indicated with blue arrows. **e)** Schematic representation of the four designed truncated constructs. Colour scheme is the same as in panel a).

II.4.2 NFIX constructs cloning, expression and purification

All four NFIX constructs were cloned into different expression vectors using the overlap extension PCR method (Bryksin & Matsumura, 2013). We initially selected pET-15b (Novagen), which contains an N-terminal His-tag, followed by a thrombin cleavage site (Leu-Val-Pro-Arg-ᵛ-Gly-Ser) to enable successive tag removal.

Expression trials were carried out using different *E. coli* strains: BL21, Rosetta, Origami, and SHuffle. Among them, only SHuffle yielded soluble recombinant protein for constructs NFIX-2 and NFIX-3. All other construct/strain combinations resulted in insoluble protein production. After several trials (see methods in II.3.3 paragraph), the best induction conditions were 0.3 mM IPTG, incubating overnight (16h) at and 20 °C in LB culture medium. This result is in line with previous data for NFIC that suggested that low induction temperatures for the NFI protein may reduce any potential toxicity of the wild-type and mutant NFI proteins in *E. coli* (Bandyopadhyay *et al.*, 1998). Small-scale purification trials of the His-NFIX constructs yielded low amount of protein (<0.1 mg/L of culture) when purified by Nickel affinity chromatography.

We aimed to combine NFIX to a solubility enhancer fusion partner to increase protein yields. NFIX constructs were cloned into the pMAL-cRI vector down-stream from the *malE* gene, which encodes for the maltose-binding protein (MBP). MBP is an affinity tag that binds to amylose resin and promotes the folding of the fusion partner. The pMAL-cRI vector also includes a Factor Xa cleavage site (Ile-Glu/Asp-Gly-Argᵛ). Recombinant MBP-NFIX constructs were all soluble. However, MBP-NFIX-1 aggregated after tag removal and MBP-NFIX-4 was subject to proteolysis (data not shown). Instead, NFIX-2 and NFIX-3 were stable in solution following tag removal. Although the presence of

the MBP tag enhanced protein expression/solubility, tag removal with Factor Xa was unsuccessful. We cloned the above mentioned His-NFIX-2 and His-NFIX-3 constructs into the pMAL-cRI vector. The recombinant protein products MBP-His-NFIX-2 and MBP-His-NFIX-3 contained two tags (MBP and His) that can be used to purify the protein purified with double affinity chromatography, and a Thrombin cleavage site between NFIX and His-tag, that can be used to remove both tags. Both amylose and Nickel affinity chromatography, however, showed low binding capacity for the fusion proteins.

Additional fusion partners, other than MBP, to NFIX-2 and NFIX-3 were tested to improve both soluble expression and purification yield. We tried pGEX4T-1 expression vector constructs to produce GST-NFIX (Glutathione S-transferase) recombinant proteins, and the pKLSL expression vector (Mancheno *et al*, 2005) to produce LSL-His-NFIX (Sepharose-binding Lectin) recombinant proteins (see Appendix IV.4 Figure A4 for all vector maps). Although the GST-NFIX-2, GST-NFIX-3, LSL-His-NFIX-2, and LSL-His-NFIX-3 were all soluble when produced in *E. coli* SHuffle cells, the yield of the pure proteins was not better than the MBP-His-NFIX-2, and MBP-His-NFIX-3 constructs (a maximum of 0.3 mg/L of culture), and therefore still too low to proceed with structural biology experiments. We selected MBP-His-NFIX-2 and MBP-His-NFIX-3 constructs for further purification protocol optimization.

We decided to try the heparin resin for binding NFIX, as an alternative to the affinity MBP tag/His tag binding during the first purification step. Biomimetic affinity ligands, *e.g.* heparin immobilised to a matrix, have been long used for the purification of DNA-binding proteins. Binding to heparin involves both charge and ligand specificity. The structure and the negative charge of heparin enable it to mimic DNA in its overall binding properties. The optimized protocol for the large-scale purification of the MBP-His-NFIX-2 and MBP-His-NFIX-3 includes the binding to a HiTrap Heparin HP resin as the first chromatographic step. Heparin elution was made by increasing buffer ionic strength (Figure 5a). MBP-His-NFIX binding to heparin was 100% efficient, but additional chromatographic steps were required to remove contaminant proteins. MBP-His tags were successively cleaved using thrombin protease, with about 70% cleavage efficacy. A second Heparin chromatography was carried out to separate the cleaved NFIX from the uncleaved fusion protein by exploiting their different heparin-binding affinities (Figure 5b). Finally, SEC was performed to eliminate high molecular weight contaminants (Figure 5c). Final NFIX purity, verified by SDS-PAGE, was >99% (Figure 5d). The optimized NFIX purification protocol yielded up to 10 mg/L and 2 mg/L bacterial culture for NFIX-2 and NFIX-3,

respectively. SDS-PAGE steps of the purification of NFIX-2 and NFIX-3 are showed in figure A5, section IV.5.

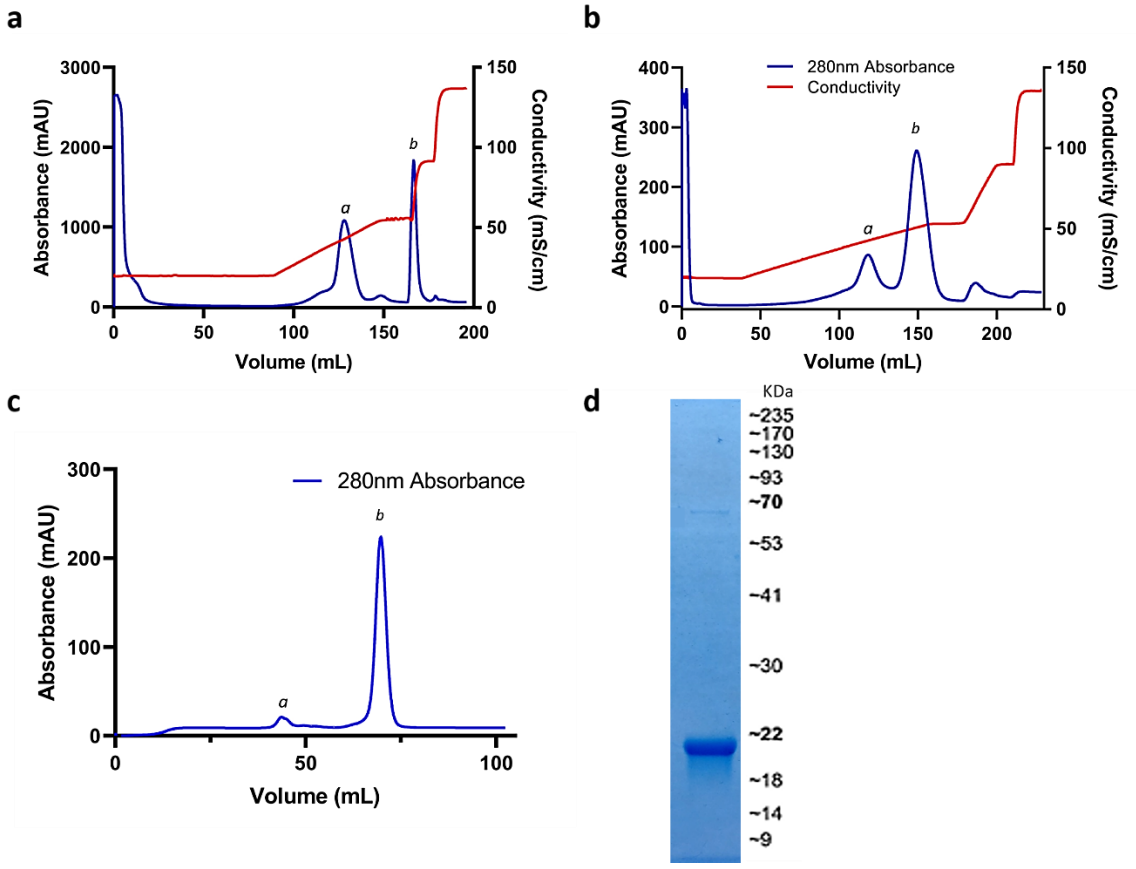


Figure 5. ***Recombinant NFIX-2 purification*****. a**) First heparin chromatography step to isolate MBP-His-NFIX-2 fusion protein (peak *a*: MBP-His-NFIX-2, peak *b*: contaminants). The conductivity and the Absorbance at 280 nm are displayed in red and blue, respectively. **b**) Second heparin chromatography to separate the uncleaved fusion protein, MBP-His-NFIX-2 (peak *a*) and NFIX-2 (peak *b*). **c**) SEC to separate high MW contaminants remained (peak *a*) from NFIX-2 (peak *b*). **d**) NFIX-2 purity is verified by SDS-PAGE. NFIX-2 construct is >99% pure. The same result was obtained for NFIX-3 (not shown).

II.4.3 NFIX DBD biophysical characterization

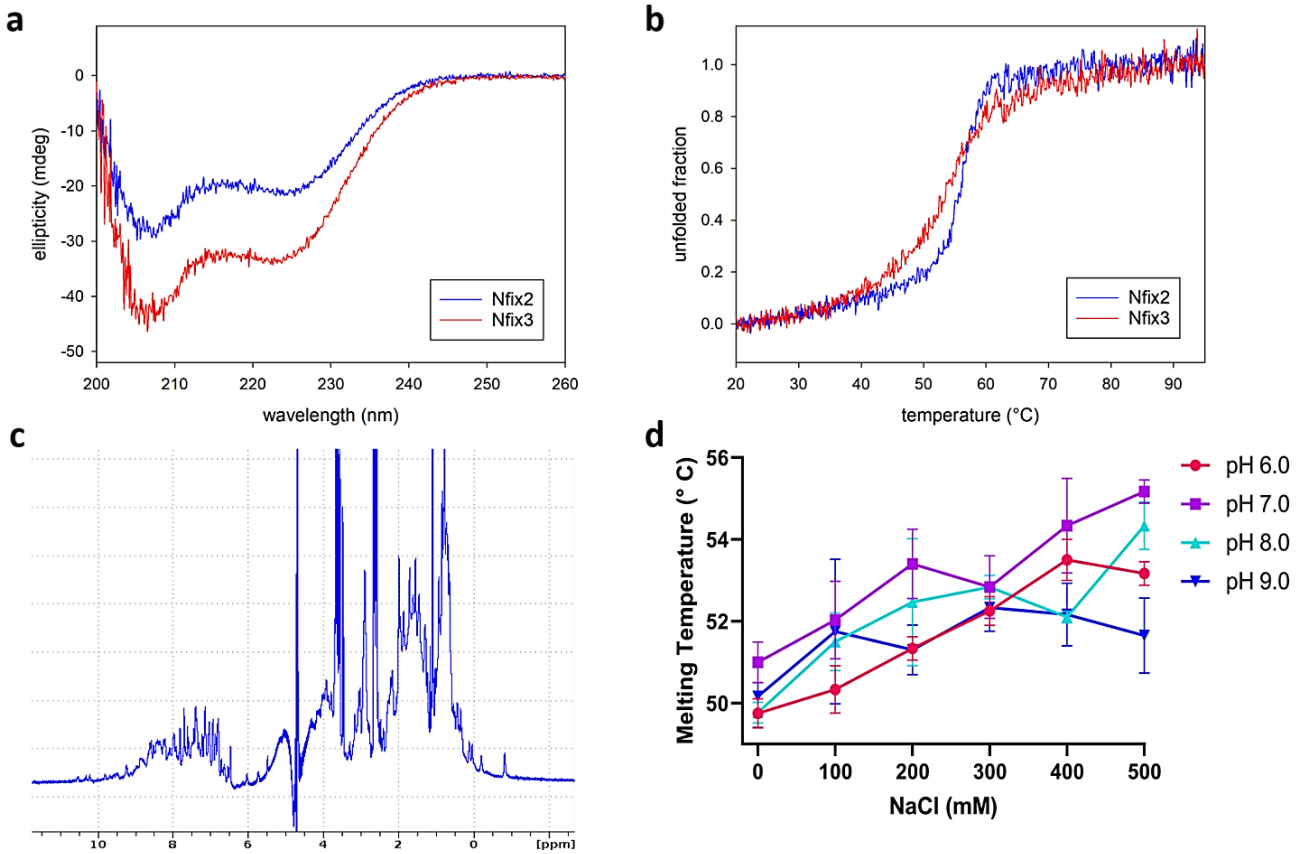
Circular dichroism (CD) was used to assess the secondary structure content and stability of the NFIX-2 and NFIX-3 constructs in solution. Despite a slight difference in ellipticity, due to sample concentration variability, they both display the typical CD spectrum for α-helix-enriched proteins, with peak minima at 222 nm and 208 nm (Figure 6a). Therefore, the secondary structure content is similar for both DBD-containing protein constructs. Thermal stability of the proteins was investigated by monitoring the temperature-dependence of the CD signal at fixed wavelength (222 nm). NFIX-2 showed a sharp transition from

folded to unfolded structure at 58 °C (Figure 6b, blue). The sharp transition indicates that protein denaturation occurs cooperatively as for a single domain. NFIX-3 transition is less sharp, and the melting temperature (T_M) estimated is about 55 °C (Figure 6b, red).

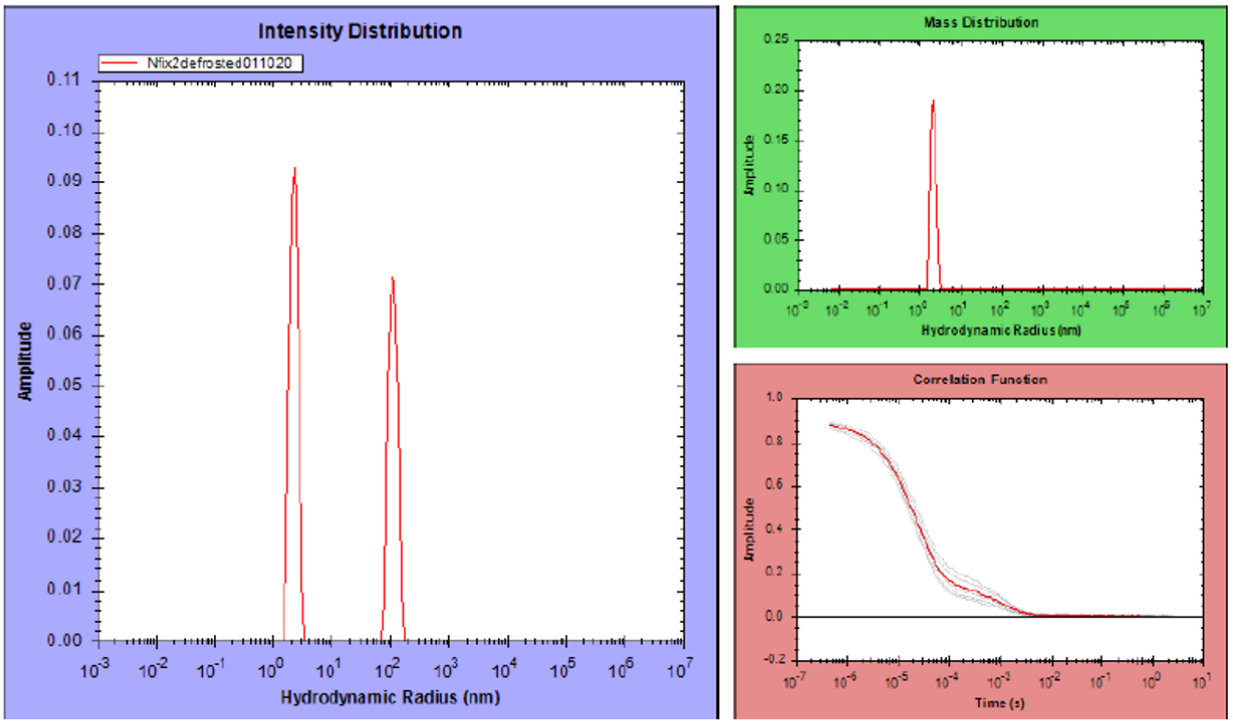
NFIX-2 folding was also assessed by Nuclear Magnetic Resonance (NMR) spectroscopy. NMR records the local molecular environment, providing a map of how atoms are chemically and spatially linked. We recorded mono-dimensional, ^{1}H (proton), NMR spectra of unlabelled NFIX-2. We selected this construct for its smaller size, that permits the collection of improved spectra. NFIX-2 displayed sharp spectral lines and all residues (HN, aromatic, aliphatic, methyl) were packed into defined chemical environments (Figure 6c). This supports that the construct is well-folded.

The protein construct stability under different buffer conditions was tested with fluorescence-based thermal-shift assay (Thermofluor). This assay follows the increase in fluorescence emission (excitation wavelength = 470-505 nm; emission =540 nm-700 nm) of the fluorophore, SYPRO® orange, that binds to core hydrophobic residues becoming exposed upon thermal denaturation. We screened NFIX-2 stability, evaluating two crucial buffer parameters: pH and salt concentration. As represented in figure 6d, there was a clear salt-dependent trend in protein stability. Indeed, the higher the salt concentration, the higher the T_M and thus stability of the protein. The optimum pH was observed to be pH 7.0.

Finally, Dynamic Light Scattering (DLS) was used to investigate the homogeneity of NFIX samples. In DLS, a laser is fired through an attenuator and onto a sample. The diffracted light from the molecules in solution is analysed by an autocorrelator that compares the intensity of light at each spot over time. The output gives information about particles size. We tested NFIX-2 and NFIX-3 behaviour in solution. Here, we report the results of NFIX-2 only; profiles for NFIX-3 were identical. The hydrodynamic radius of NFIX-2 corresponded to an estimated MW of 24.04 kDa, in line with its calculated MW (19746.13 Da) plus the weight of the hydration shell (Figure 6e). DLS confirmed that both NFIX constructs are monomers in solution.



e



Peak #	Mean Rh (nm)	Mode Rh (nm)	Std. Dev. (nm)	Polydisp. (%)	Est. MW. (kDa)	Intensity (%)	Mass (%)	Volume (%)	Number (%)
1	2.30	2.39	0.63	27.47	24.04	53.83	99.99	100.00	100.00
2	113.29	112.56	32.28	28.49	Out of Range	46.17	0.01	0.00	0.00

Figure 6. ***NFIX biophysical characterization*****. a**) CD spectra of NFIX-2 (blue) and NFIX-3 (red) **b**) Thermal denaturation ramp of NFIX-2 (blue) and NFIX-3 (red) followed by absorbance at 222 nm. **c**) NFIX-2 mono-dimensional NMR spectrum. **d**) NFIX-2 T_M at various salt concentrations and pH. Error bars are represented. **e**) DLS measurements on NFIX-2 at 1 mg/mL. Blue panel represents intensity distributions of the species in solution, it reveals the presence of NFIX-2 (53.83% of intensity) and an aggregated specie (46.17% of intensity). Green panel represents mass distribution of the species in solution, NFIX-2 owns the 99.99% of the mass in solution. Pink panel represents correlation function of the sample exponential decays over time. Since instrument reveal modest polydispersity in the sample, correlation function is an average of two exponentials, due to NFIX-2 and the aggregated particles in solution. Last panel reports the species features. Estimated MW of NFIX-2 hydrodynamic radius is 24.04 kDa (std. dev. 0.63).

II.4.4 Functional assays

The DNA-binding potential of NFIX-2 and NFIX-3 was investigated by Electrophoretic Mobility Shift Assay (EMSA). This assay is used to study protein-DNA interactions in a non-denaturing PAGE setup. In this assay, the band corresponding to the fluorescently labelled DNA probe is shifted to a higher molecular weight when the DNA-protein complex is assembled, relative to the free probe. In a dose-response EMSA experiment, we incubated a fixed quantity of Cy5-labelled DNA with increasing concentrations of NFIX. Under these experimental conditions, the concentration of the protein-DNA complex should raise in function of increased protein concentration. The DNA-binding capacity of NFIX-2 and NFIX-3 was evaluated and compared in Figure 7. Both constructs bind the target DNA with an affinity in the nM range. Notably, NFIX-3 affinity is about 10-fold higher than NFIX-2 (Figure 7a).

To further investigate the DNA-binding properties of NFIX, we tested two DNA probes with different spacer lengths. The first probe contained the canonical 5bp spacer, the second was shortened to 4bp. NFIX-2 and NFIX-3 binding to the probe with a 4bp spacer was barely detectable, while the complex with the 5bp spacer probe migrated correctly (Figure 7b). Therefore, the 5bp spacer in the NFIX consensus sequence is necessary to reach a successful DNA-binding.

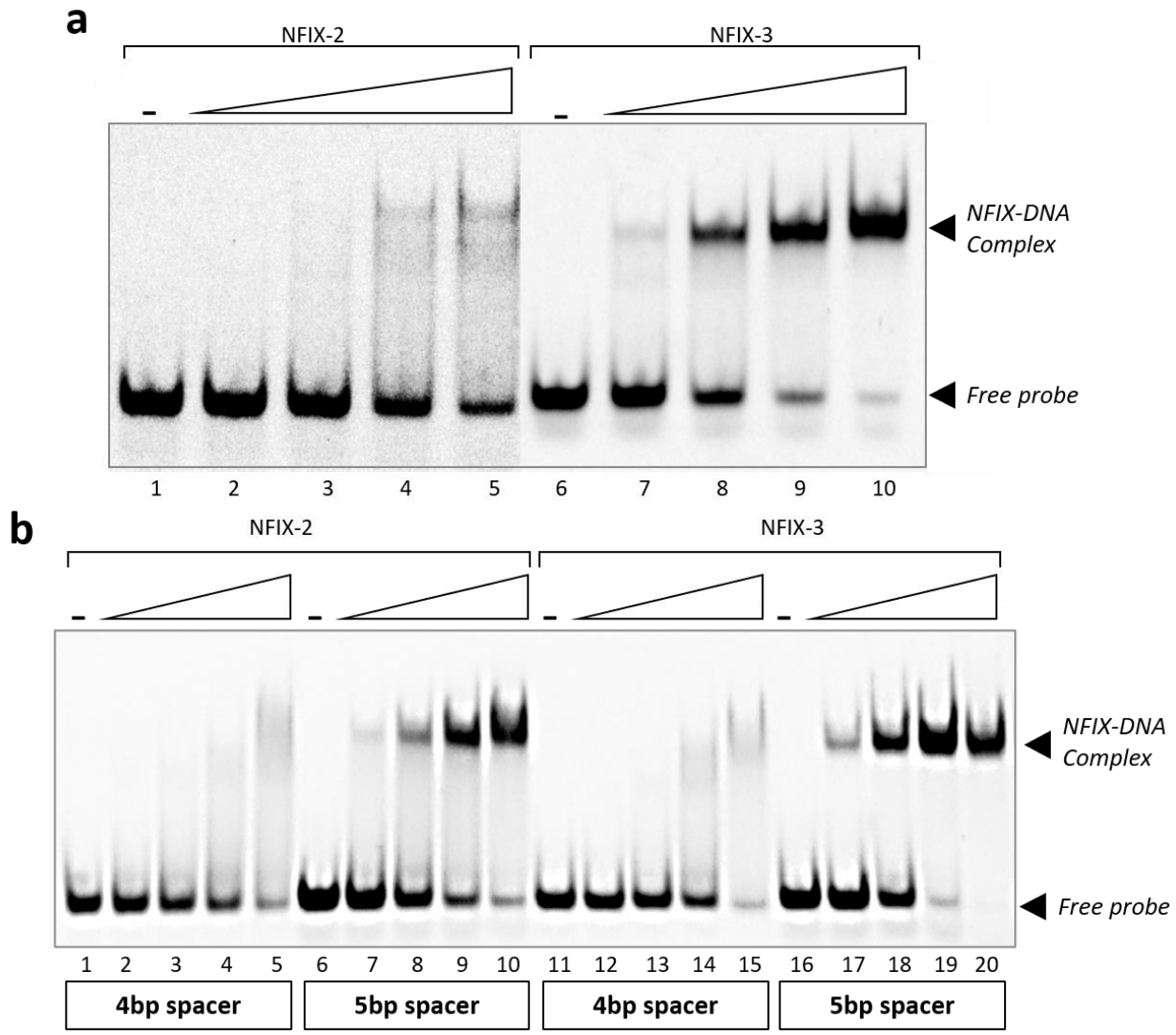


Figure 7. ***NFIX binding to DNA.*** **a**) Dose-response EMSA of NFIX-2 and NFIX-3 (0 nM, 60 nM, 180 nM, 540 nM, 1.6 µM) were incubated with 20nM of DNA probe. Band corresponding to the free probe or the protein/DNA complex were labelled. **b**) Dose-response EMSA of NFIX-2 and NFIX-3 with a 4bp and a 5bp of spacer probes. (NFIX concentrations are 0 nM, 60 nM, 180 nM, 540 nM) Arrows corresponding to the free probe and the protein/DNA complex are labelled. Each lane is indicated at the bottom.

To verify the sequence specificity of NFIX, an EMSA competition experiment was carried out. Specific binding to the NFIX consensus sequence was assessed using two different, unlabelled (cold) competitor DNA probes of the same length: an unrelated sequence and the NFIX consensus sequence. We used NFIX-3, given its improved DNA-binding affinity (Figure 7a). As shown in figure 8a, the random probe does not impact NFIX-3 binding to the labelled probe, whereas the unlabelled NFIX-3 consensus sequence did, confirming that NFIX-3 binding is sequence-specific.

We also explored the kinetics of NFIX association/dissociation through EMSA competition experiments. The aim of the experiment was to evaluate whether NFIX dissociates from a preformed DNA-protein complex to bind freshly added DNA, or if the preformed complex remains stable over a certain range of time. Therefore, in one experiment, NFIX-3 was added to a mixture of both unlabelled and labelled probes and, in a second experiment a complex between NFIX-3 and the labelled probe was preformed, before challenge with the unlabelled probe (Figure 8b). The competition of the unlabelled probe with the preformed complex gave the same results as observed for the experiment in which NFIX-3 was added to the probe mix. Therefore, we can conclude that NFIX-3 is able to exchange its DNA target at a time rate in the order of few minutes (experiment duration was 1 h).

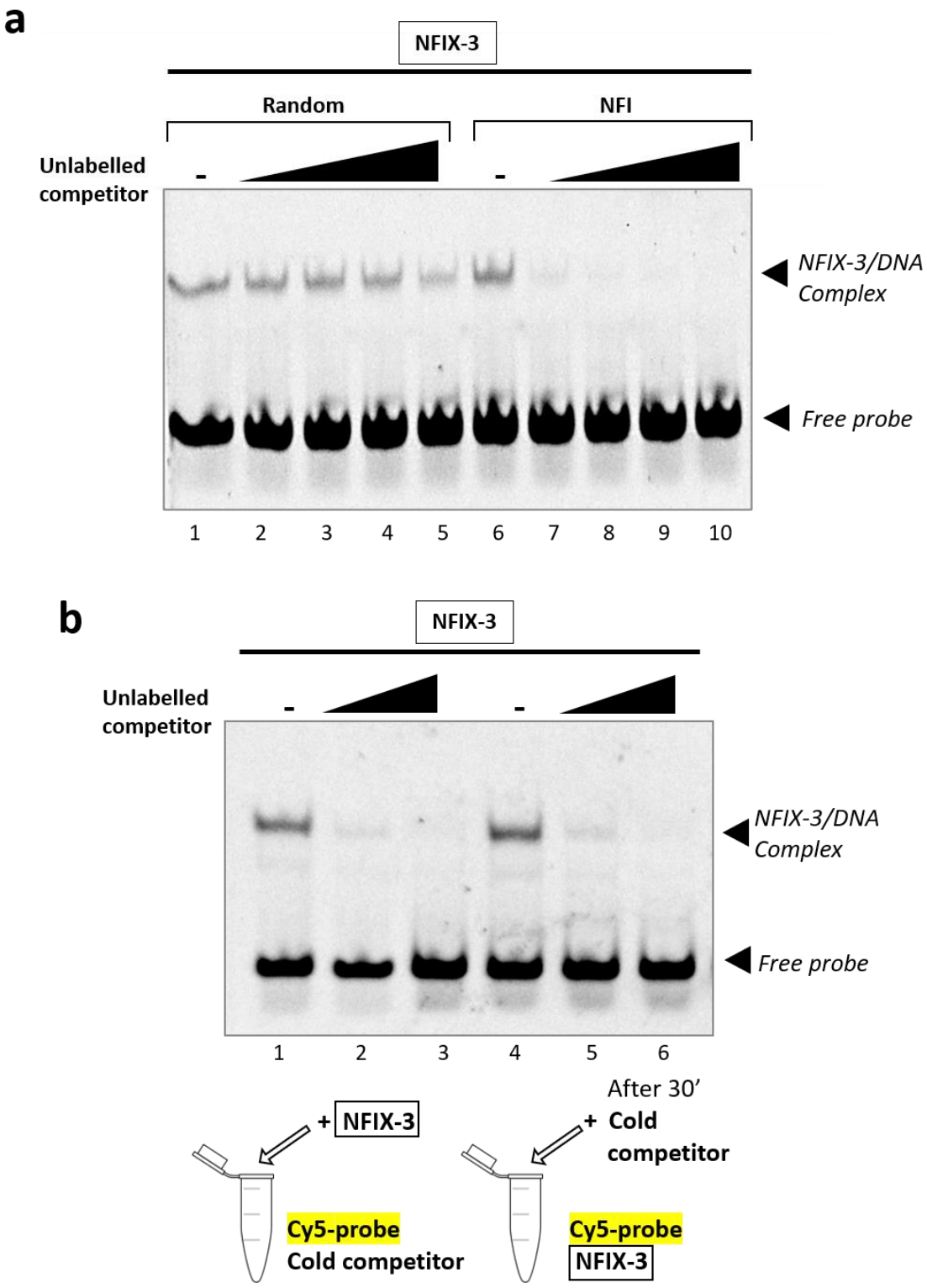


Figure 8. ***NFIX DNA sequence specificity and kinetic analysis.*** **a**) Sequence specificity EMSA competition. Fixed 50 nM of NFIX-3 and 20 nM of NFI Cy5-probe were incubated with increasing concentrations of unlabelled competitors. The lanes represent different competitor concentrations: no competitor, 6.25x, 12.5x, 25x and 50x Cy5-probe. Unrelated and NFI sequence competition were displayed. **b**) NFIX-3 DNA-binding kinetics explored through EMSA competition. Two set ups of the competition experiment were tested. NFIX-3

and Cy5-probe concentrations are fixed at 150 nM and 20 nM, respectively. Below, a scheme showing the experimental set up was added for clarity. Each lane is indicated at the bottom.

II.4.5 Determination of the Zn^{2+} presence by Flame Atomic Absorption Spectrometry (FAAS)

The presence of zinc was assessed in NFIX-2 by calibrated FAAS. Atomic absorption spectrometry (AAS) is one of the most popular techniques for element determination. It is a single-element method used for trace metal analysis with high precision, high sensitivity and easy operation. Quantification by AAS is based on determination of the weakening of the emission light from the radiation source, which has been absorbed by the ground state atoms of the element of interest. The value of the decreased radiation, at a particular wavelength, measured by detector depends on the metal concentration following the Beer's Law equation. When spectrometric techniques with conventional pneumatic nebulization of the sample are employed, as in flame atomic absorption spectrometry (FAAS), the sample must be digested to form a solution in which the analyte is distributed homogeneously. This wet decomposition is produced using combination of oxidizing acids (HNO_3, $HClO_4$, H_2SO_4), non-oxidizing acids (HCl, HF, H_3PO_4) and hydrogen peroxide.

Prior to the metal analysis, a homogeneous solution of 0.2 mg/mL NFIX-2 was completely digested with a solution of nitric acid (65%) and chloric acid (37%). Then, H_2O_2 (30%) was added to the mixture in an open vessel acid digestion system. The determined correlation coefficient for the calibration curve of NFIX-2 sample was 0.995. The amount of zinc detected in NFIX-2 was 0.415 mg/L (RSD: 4.7%), indicating that Zn^{2+} is certainly present in the protein sample (Figure 9).

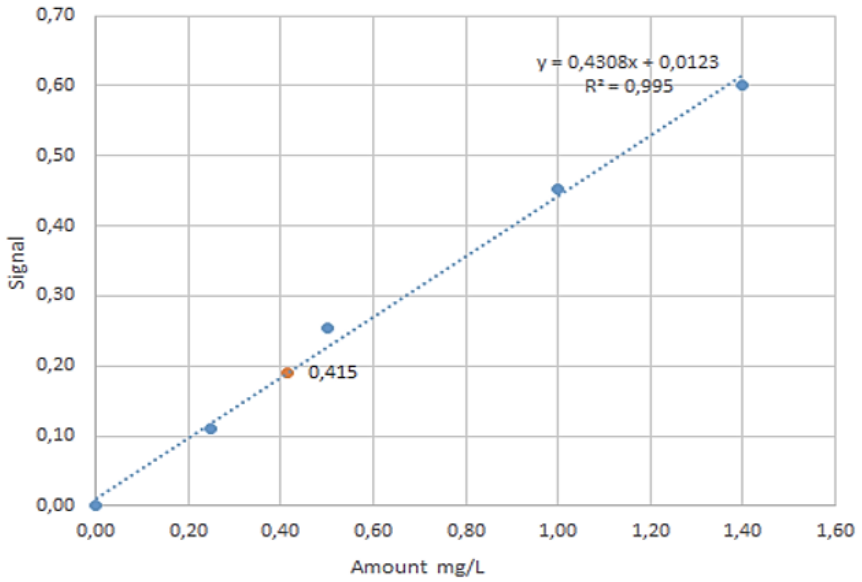


Figure 9. ***Zn Calibration curve.*** Correlation coefficient of zinc calibration curve and the relative standard deviation by FAAS analysis. The amount of zinc in NFIX-2 sample was estimated to be 0.415 mg/L (RSD: 4.7%).

II.4.6 NFIX-2 crystallization

The prerequisite for 3D structure determination of a macromolecule by X-ray crystallography requires the growth of well-diffracting single protein crystals. In general, macromolecular crystallization is empirical and involves screening several parameters that can influence crystal formation, and then optimizing the individual variables to obtain the best possible crystals. This is usually achieved by carrying out an extensive series of crystallization trials, which in our case included the investigation in parallel of NFIX alone and in complex with DNA conditions. Indeed, successful crystallization of TF-DNA complexes is often affected by the choice of DNA length and structure. Therefore, we primarily investigated by EMSA the minimal length of DNA that can still bind NFIX. We focussed on NFIX-2, which is the construct with a lower affinity for DNA but with a higher purification yield. Starting from the palindromic consensus sequence composed of 31bp, we set up EMSA competition experiments to determine the minimum length with preserved DNA-binding capacity. We generated a set of oligos ranging from 25bp to 17bp, by symmetrically removing the terminal nucleotides to the 31bp consensus sequence. Predictably, the shorter the oligos, the weaker the competition and, therefore, the affinity (Figure 10a). From this analysis, we selected a 21bp oligo as a promising candidate for initial crystallization trials since this oligo is not excessively long and it has a sufficient protein-binding affinity.

Co-crystallization experiments with the NFIX-2/DNA complex were set up by incubating the macromolecules in a 2:1 protein:DNA molar ratio and by decreasing salt concentration of the mixture to 50 mM to promote complex assembly. NFIX binding to 21bp DNA in solution was assessed by SEC. The sample eluted as a single peak, at the predicted molecular weight of 52.34 kDa for the protein dimer/DNA complex (Figure 10b and Figure A6 in IV.6 section). Sitting drop vapour diffusion and under-oil microbatch methods were used for the high-throughput crystallization screening of the NFIX-2 and NFIX-2/DNA complex. We used commercial screens to explore a wide range in search of conditions for preliminary crystallization ‘hits’ (see section IV.7, table A2).

No hits were found for NFIX-2 alone screenings, but a crystal of NFIX-2/DNA preparation grew in sitting drops in condition B2 (20% PEG 3350, 0.2 M sodium

thiocyanate) of the JCSG screen. The crystal was flash-frozen in liquid nitrogen and tested for X-ray diffraction at Diamond Light Source (DLS, UK) synchrotron. This crystal diffracted at 6 Å resolution (Figure 11a). The precipitant composition was optimized by increasing the concentration of PEG 3350 to 22% and adding a buffer (0.1 M HEPES pH 7.0), whilst maintaining 0.2 M sodium thiocyanate. Further optimization of the complex concentration was carried out, increasing it to 15 mg/mL, which increased both crystal quality and reproducibility. Overall, we improved crystals number *per* plate and the diffraction resolution limit to 3.5 Å (Figure 11b). Altering the DNA oligo length to a longer 23bp DNA, which exhibited better DNA binding, also improved crystallization. We tested a 23bp DNA oligo with blunt ends and sticky ends. The latter exhibited improved crystals quality in terms of shape, number of crystals *per* drop, reproducibility, and diffraction resolution limit using the microbatch method (Figure 11c).

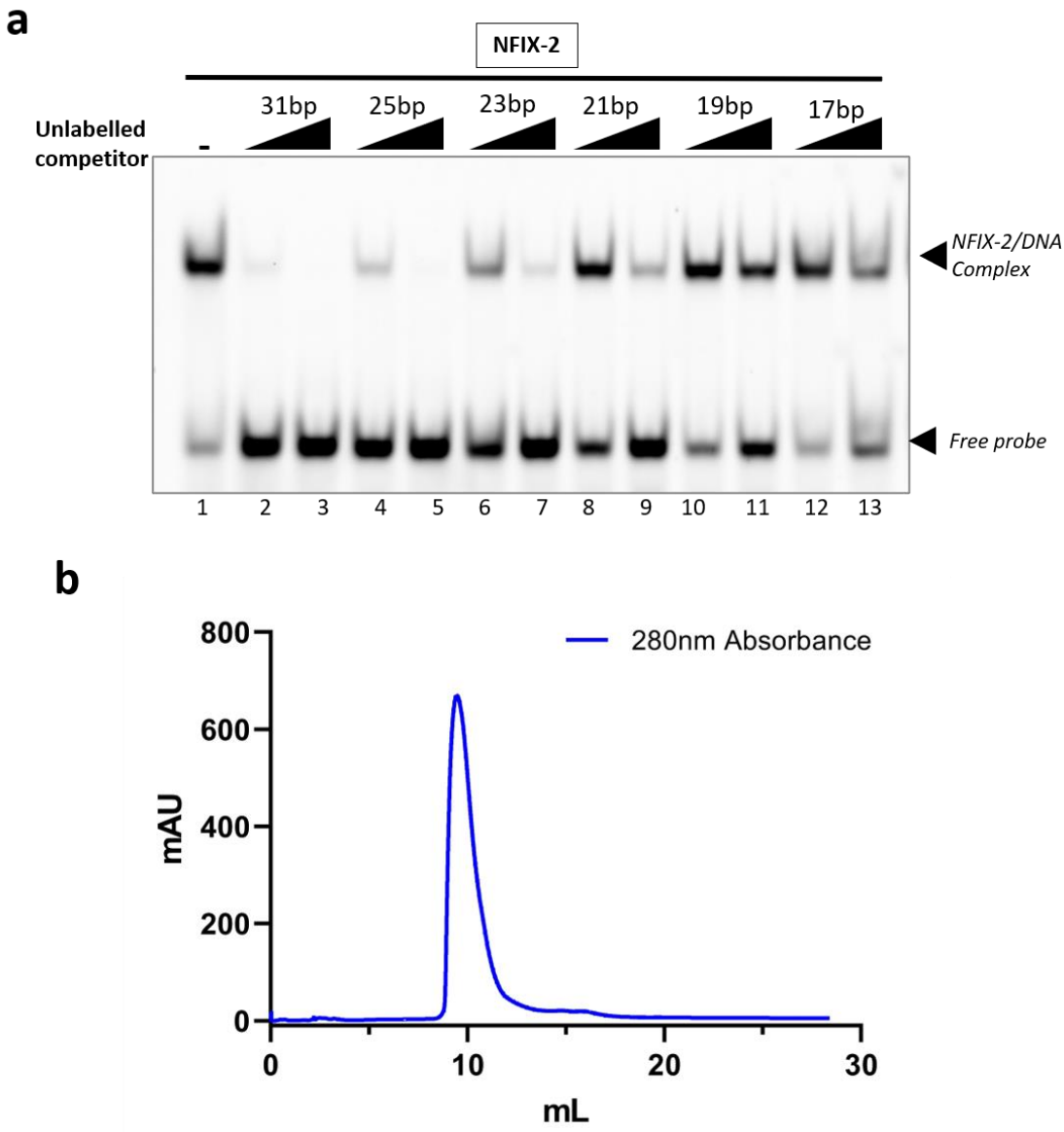


Figure 10. ***Oligo length optimization for co-crystallization.*** **a**) EMSA competition screening different oligo DNA lengths. Unlabelled DNA competitors were used in 10x and 50x excess respect to the Cy5-probe concentration (20 nM). As a negative control, an experiment without competitor was included (lane 1). **b**) SEC chromatogram of the NFIX-2/DNA complex using a 21bp dsDNA.

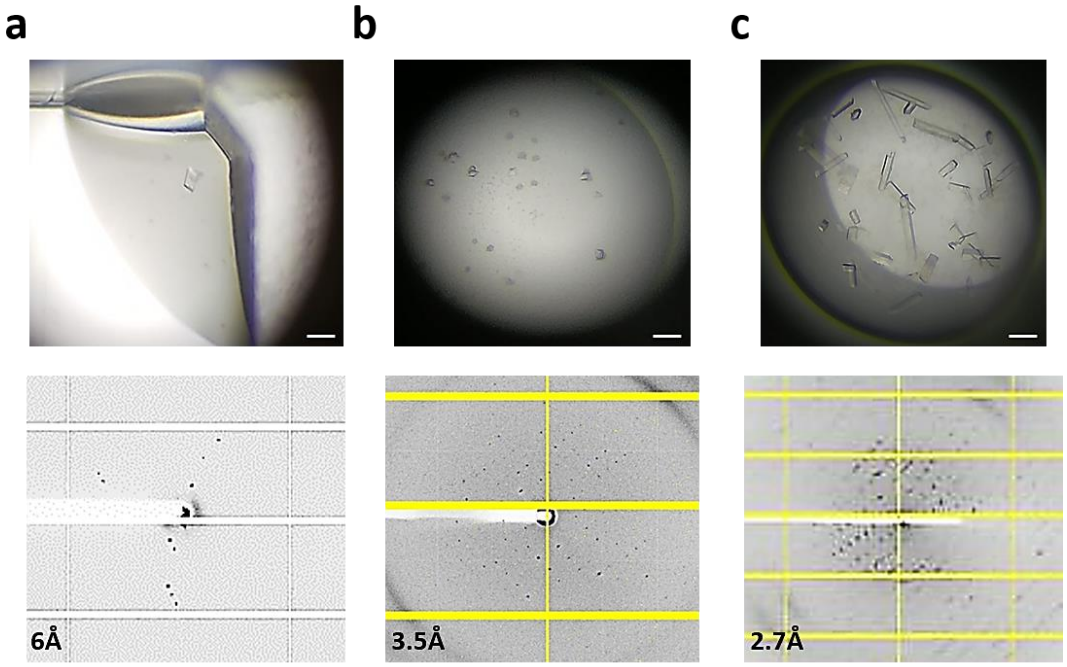


Figure 11. ***NFIX-DNA complex crystals.*** Crystals and their corresponding diffraction patterns are shown above and below, respectively. **a**) Preliminary NFIX-2/DNA (21bp) crystal from JCSG screen that diffracted at 6 Å resolution. **b**) Optimized NFIX-2/DNA (23bp blunt ends) crystals that diffracted at 3.5 Å resolution. **c**) NFIX-2/DNA (23bp sticky ends) crystals in microbatch that diffracted at about 2.7 Å resolution. White scale bar represents 0.2 mm.

II.4.7 NFIX Zn-binding

Sequence analysis on NFIX and alignment with Smad proteins highlighted the conservation in NFIX of a "CCCH" motif, which in Smads is involved in Zn^{2+}-coordination. Accordingly, the Zn^{2+} presence was detected in the NFIX-2 solution sample by Flame Atomic Absorption Spectrometry (FAAS) (see Figure 9).

The presence of Zn^{2+} in the NFIX-2 crystals was further investigated through X-ray fluorescence (XRF) spectroscopy. In XRF, the sample get excited by incident X-ray photons (synchrotron beam). When the electrons transit from the higher energy levels to the vacant inner shell of the atom, subsequently emission of secondary X-ray photons occurs. The released photons have a specific energy, which is a fingerprint of the atom from which it has originated. Therefore, we used XRF on NFIX-2 crystals at DLS synchrotron to analyse the presence of zinc. The X-ray absorbance was scanned over the range 9600–9700 eV, corresponding to K-edge of zinc emission spectrum. Absorption at the zinc K-edge (9670 eV) confirms that zinc is bound to crystallized NFIX-2 (Figure 12).

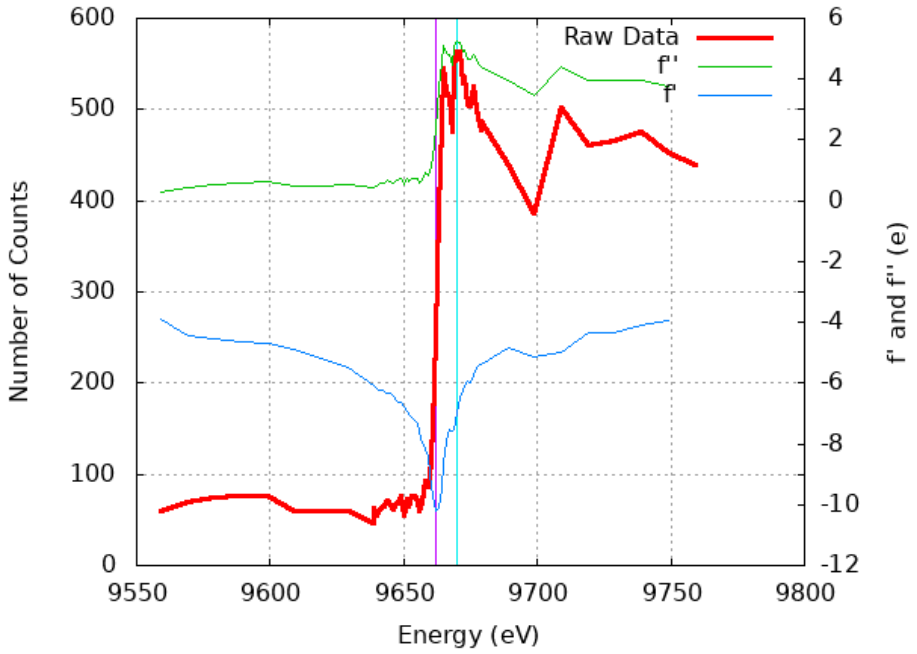


Figure 12. ***Zinc XFR on the NFIX-2 crystal.*** XFR plot of the absorption spectrum of the Zn K-edge. The graph plots the number of photons counts versus their energy (eV). Scattered electrons (e) submitted to derivative operators to calculate the spectra (f' and f''). Absorption at the zinc energy edge (9670.0 eV) indicates that zinc is bound in the NFIX-2 crystal.

II.4.8 Data collection and structure determination

Two different data sets belonging to two different space groups were collected: $P4_12_12$ and $P2_1$.

The $P4_12_12$ crystal was collected at the DLS synchrotron and diffracted up to 3.5 Å resolution. Our initial attempts to solve the structure by molecular replacement using Smad MH1 domain structure (PDB-code 1MHD) as search model failed. The $P2_1$ crystal was further collected at the ELETTRA synchrotron and diffracted at 2.7 Å resolution. This dataset was used for SAD-phasing at the zinc edge (λ = 1.2705 Å). The monoclinic $P2_1$ crystal has unit cell parameters a = 43.50 Å, b = 98.95 Å, c = 61.91 Å, α = 90.00°, β = 92,73° and γ = 90.00°, with two molecules in the asymmetric unit (ASU). The Zn-SAD phasing protocol was carried out using the SHELXC/D/E pipeline as implemented in HKL2MAP (Sheldrick, 2010). Four Zn^{2+} ions were localized in the ASU and used for phasing, with occupancy between 0.75 and 1.0. SAD-phasing statistics are: anomalous completeness 91.0 % (88.0%), R_{anom} 0.072 (2.274), average figure of merit of SAD-phases (after automated fitting of 97 Ala residues) 0.520.

ARP/wARP was used for automated model building (Langer *et al.*, 2008), followed by multiple rounds of manual model building in Coot (Emsley *et al.*,

2010). Surprisingly, only NFIX-2 proteins could be built in the calculated electron density map, with no clear evidence of the presence of the bound DNA. The structure was refined using REFMAC (Murshudov *et al.*, 1997) and Phenix.refine (Adams *et al.*, 2010) to final R_{work}/R_{free} values of 21.0% and 26.9%, respectively (Table 6).

The refined $P2_1$ model was used to solve the structure of the 3.5Å $P4_12_12$ dataset, previously collected at DLS, by molecular replacement with Phaser (McCoy *et al.*, 2007). The tetragonal $P4_12_12$ crystal has unit cell parameters a = 64.98 Å, b = 64.98 Å, c = 127.77 Å, α = 90.00°, β = 90,00° and γ = 90.00°, with one molecule of NFIX-2 in the ASU.

All statistics for the data reduction and the model refinement are summarized in Table 6.

Data set	**NFIX-2 *P*2$_1$**	**NFIX-2 *P*4$_1$2$_1$2**
Space Group	*P* 1 2$_1$ 1	*P* 4$_1$ 2$_1$ 2
a, b, c (Å)	43.17, 98.36, 61.63	65.01 65.01 127.89
α, β, γ (°)	90.00, 92.70, 90.00	90.00, 90.00, 90.00
Wavelength (Å)	1.2705	1.28199
Resolution (upper limit)	2.70 (2.84-2.70)	3.50 (3.85-3.50)
Resolution (lower limit)	49.18	45.97
#Rpim	0.054 (0.559)	0.116 (0.919)
+CC1/2	0.968 (0.817)	0.998 (0.603)
<I/σ(I)>	13.2 (2.0)	7.6 (1.3)
Redundancy	6.6 (6.6)	25.1 (26.6)
Completeness (%)	99.3 (99.3)	100.0 (100.0)
Wilson B-factor (Å^2)	54.55	88.00
Refinement		
Resolution (Å)	49.18-2.70 (2.80-2.70)	45.59 - 3.5 (3.625 - 3.5)
Number of reflections	14029 (1374)	3802 (366)
Rwork/Rfree	0.206/0.262 (0.308/0.407)	0.233/0.252 (0.327/0.525)
Proteins in the ASU	2	1
Total number of protein residues	324	164
Zn^{2+} ions	4	2
HEPES molecules	3	1

Water molecules	39	9
Average B factors (Å^2)	77.88	126.27
Rmsd bond lengths (Å)	0.016	0.005
Rmsd bond angles (°)	1.96	1.00
Ramachandran plot statistics	93.44 % in favoured	96.30% in favoured
	0.94 % outliers	0.62% outliers
Highest-resolution shell is shown in parentheses		
$^{\#}R_{p.i.m.} = \sum_{hkl}\sqrt{1/n-1}\sum_{j=1}^{n}\left\|I_{hkl} - \langle I_{hkl}\rangle\right\| \Big/ \sum_{hkl}\sum_{j} I_{hkl,j}$		
$^{+}$CC1/2 is the correlation coefficient of the mean intensities between two random half-sets of data.		

Table 6. ***Data collection and refinement statistics for the NFIX-2 $P2_1$ and $P4_12_12$ spacegroups.***

II.4.9 Structural analysis

II.4.9.1 NFIX-2 structure

The NFIX-2 structure was solved at 2.7 Å in the $P2_1$ space group by Zn-SAD and refined to final R_{work}/R_{free} values of 21.0% and 26.9%, respectively. The overall structure of the complex is well defined in the electron density map, with the ASU containing two NFIX-2 monomers. The final model accounts for amino acids 13–174 in chains A and B, with good geometry (statistics shown in Table 6). The second structure of the NFIX-2 in the $P4_12_12$ space group was solved at 3.5 Å. Since the structure in $P2_1$ was solved and refined at higher resolution, the results presented here will refer to this structure unless specified otherwise.

The NFIX-2 fold is defined by a "core" domain formed by a four α-helical bundle (α1- α4), two isolated 3_{10} helices, two anti-parallel pairs of short β-strands (β1–β2, β3–β4) each preceded by one-turn of 3_{10} helix (α'5 and α'6), and an "antenna" domain built by two helices (A1 and A2) (Figure 13). The core domain shares the topology of the MH1 domain, the DBD of Smad TFs (Aragon *et al*, 2019; Baburajendran *et al*, 2011; BabuRajendran *et al*, 2010; Chai *et al*, 2015; Martin-Malpartida *et al*, 2017). The antenna domain consists of a helical excursion from the core, localized between the α1 and α4 helices, and it is not present in MH1-containing TFs (Figure 14). Therefore, we can define the NFIX-2 structure as an MH1-Like domain (M-L domain). Considering that sequence alignment of NFIX-2 with other members of the NFI family (A, B, and C from different species) shows a sequence identity >95% (Appendix VI.2 Figure A2),

the M-L domain here reported can be considered as a prototype for all NFI proteins.

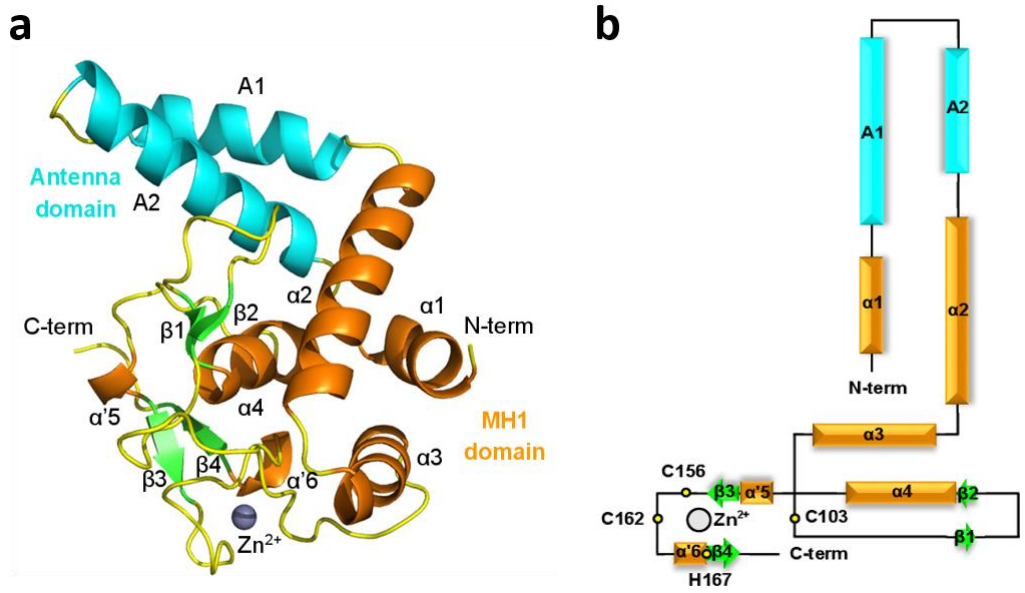


Figure 13. ***Three-dimensional structure of NFIX M-L domain*****. a)** Ribbon diagram showing the 3D structure of NFIX-2. The MH1 core domain is shown in orange (helices) and green (strands), and the helical antenna domain in shown in cyan. The bound Zn^{2+} ion is shown as a grey sphere. **b)** Topology diagram showing the secondary structure organization of the M-L domain. Helices are shown as cylinders, strands as arrows. The bound Zn^{2+} ion is shown in grey and its coordinating residues are shown as yellow circles.

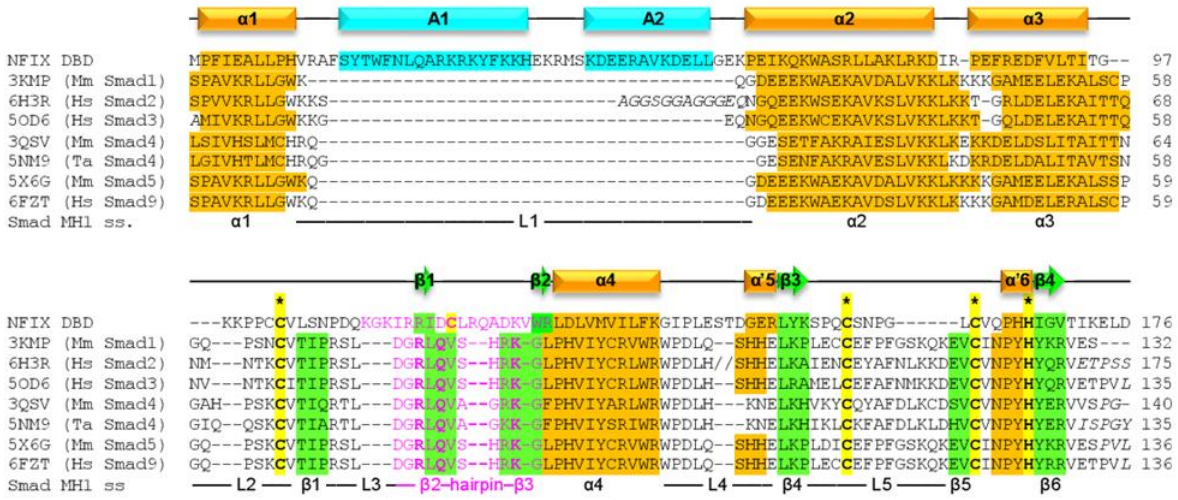


Figure 14. ***Structure-based sequence alignment of the NFIX M-L domain with related MH1-fold proteins***. NFIX M-L domain was aligned with human Smad2 (PDB-code 6H3R), Smad3 (PDB-code 5OD6), and Smad9 (PDB-code 6FZT), mouse Smad1 (PDB-code 3KMP), Smad4 (PDB-code 3QSV), and Smad5 (PDB-code 5X6G), and *Trichoplax adhaerens* Smad4 (PDB-code 5NM9). Sequence alignments were performed using the MUSCLE program (https://www.ebi.ac.uk/Tools/msa/muscle/) and manually corrected based on 3D structure

comparisons. NFIX M-L domain secondary structure elements are indicated above the sequence and shaded (color code as in Figure 13) for all aligned proteins. Residues involved in Zn^{2+}-coordination are highlighted in yellow by an asterisk, while Smad residues forming the DNA-recognition motif are in magenta, with residues that provide sequence-specificity in bold letters. The SMAD MH1 secondary structure is indicated below the alignment as a reference. A 30-residue insertion in Smad2 (6H3R) is indicated by //.

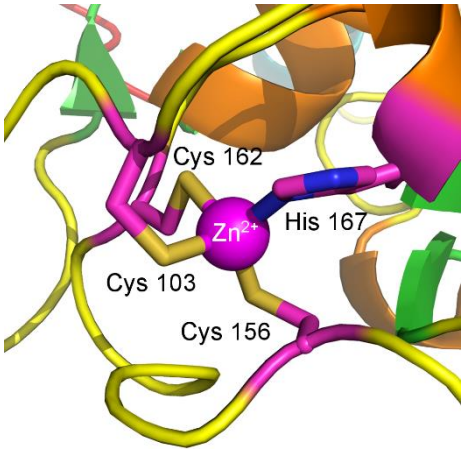


Figure 15. ***The NFIX zinc-coordination site*****.** NFIX residues, coordinating the Zn^{2+}, ion are shown in stick representation and labeled.

A zinc ion was modeled into the MH1 core domain, coordinated by three cysteine (Cys103, Cys156, and Cys162) and one histidine (His167) residues, as indicated by a strong electron density and by the anomalous peak calculated from the Zn-SAD (Figure 15). This is a very interesting structural finding since binding of Zn^{2+} was never reported before either for NFIX or for members of the NFI family, despite the coordinating Cys and His residues align very well with residues forming the Zn^{2+}-binding site typically found in MH1 proteins (Figure 14). Furthermore, NFIX Cys103, Cys156, Cys162, and His167 are invariant in NFI proteins, thus suggesting the conserved nature of zinc-binding within the NFI family (Appendix IV.3 Figure A3). The bound zinc ion is deeply buried in the MH1 core structure, with the four coordinating residues highly or completely solvent inaccessible (solvent-accessible surface area of 19.0 $Å^2$, 14.0 $Å^2$, 0.0 $Å^2$, 7.8 $Å^2$, for Cys103, Cys156, Cys162, and His167, respectively), and it likely plays a structural role by stabilizing the C-terminal region of the DBD (Figure 13b and 15).

The two NFIX M-L domain monomers (A and B chains) present in the crystal ASU assemble in a loose homodimer with a two-fold symmetry. The same crystallographic NFIX-2 dimer is found in the $P4_12_12$ structure. The protein-protein interface involves residues belonging to the helix α1 and the following

loop (21-25), the L1 loop (97-100), the L4 loop (138-141), the L5 loop (153-155), and the C-terminal region (162-172) (Figure 16). The dimeric assembly of NFIX M-L domain buries about 1.510 $Å^2$ solvent-accessible surface area (755.2 $Å^2$ for each monomer, about 8% of the total surface area) with a dissociation free energy of -2.1 kcal/mol, and it is predicted to be unstable in solution (jsPISA) (Krissinel, 2015). This result agrees with our SEC and DLS results (see II.4.2 and II.4.3 paragraph), in which the apparent molecular weight of the NFIX-2 fits with a monomer. At the dimeric interface, two ligands are bound: a Zn^{2+} ion, coordinated by His 166 (from both monomers) and two water molecules; and a HEPES molecule, derived from the crystallization buffer, which fits a pocket lined by residues Arg24, Ala25, Phe26, Gly139, Ile140 and Pro141 from both protein chains. The position of the Zn^{2+} ion at the monomer-monomer interface has been confirmed by the analysis of the SAD data, and it could be ascribed to the presence of a small amount of zinc in the protein crystallization sample most likely due to a marginal zinc dissociation from the protein zinc-binding site (His166 is adjacent to His167 in sequence and in space) and/or from contamination of the crystallization solutions. Indeed, a fluorescence scan (at the Zn^{2+} absorption edge) on the solution present in the cryo-loop after crystal removal revealed the presence of residual zinc in solution (data not shown).

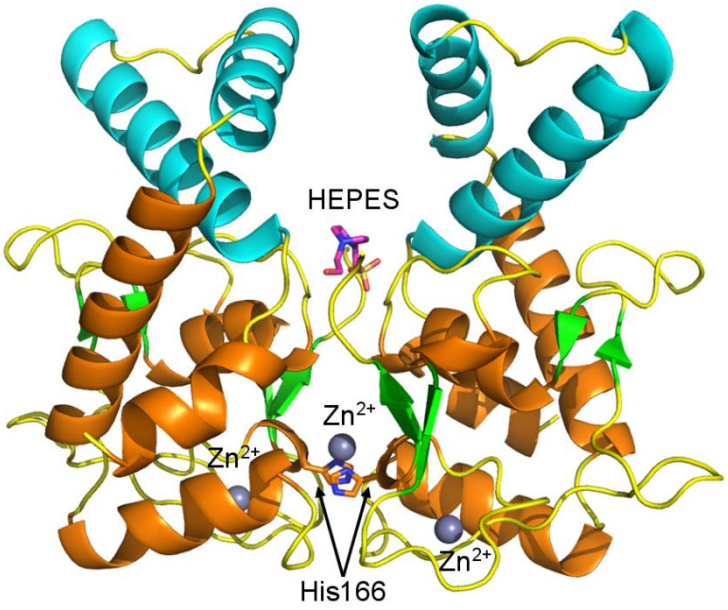


Fig. 16. ***NFIX M-L domain molecules in the crystal ASU.*** Ribbon representation of the NFIX M-L domain packing in the ASU (space group $P2_1$). Colour scheme as in Figure 13. Zn^{2+} ions are shown as grey balls, The HEPES molecule bound at the protein-protein interface and the His166 side chains involved in Zn^{2+} coordination at the same interface are shown as magenta and orange sticks, respectively.

It should be mentioned that the crystallographic packing shows gaps between layers formed by NFIX-2 proteins, with traces of electron density whose shape and connectivity is not clear enough to be modelled. It can be hypothesized that DNA or PEG molecules (both present in the crystallization solution) fill in the protein-protein crystal layers with an unspecific binding that still contribute to stabilizing the crystal packing.

II.4.9.2 Structural relatives

A structural analysis using DALI (Holm & Rosenstrom, 2010) indicates that the NFIX MH1 core domain resembles the MH1 domain of Smad TFs, such as Smad3 (PDB-code 5OD6; DALI Z-score of 12.0, residue identity of 16%), Smad2 (PDB-code 6H3R; DALI Z-score of 11.7, residue identity of 19%), Smad4 (PDB-code 5NM9); DALI Z-score of 11.2, residue identity of 17%) and Smad5 (PDB-code 5X6G; DALI Z-score of 10.3, residue identity of 19%). It is remarkable to note that, despite the low sequence identity (< 20%) the rmsd of the superimposed proteins is relatively low (between 2.6 and 2.7 Å) (Figure 17a). Smad9 (PDB-code 6FZT; DALI Z-score of 9.8, residue identity of 15%), and Smad1 (PDB-code 3KMP; DALI Z-score of 9.7, residue identity of 17%) show a higher rmsd (6.5 Å and 4.6 Å, respectively) due to the simple displacement of the α1 helix from the core of the MH1 domain due to crystal contacts (Figure 17b). If this α1 helix is removed, the rmsd of the superimposition with NFIX MH1 core domain becomes similar to that of other Smad MH1 domains. The structural match between NFIX MH1 core domain and Smad MH1 further improves (rmsd between 1.8 and 2.0 Å) if the loops connecting secondary structure elements are removed (Table 7). The perfect sequence and structural matching of the region (3 Cys-His) involved in Zn^{2+} coordination (Figure 14 and 18) is particularly evident.

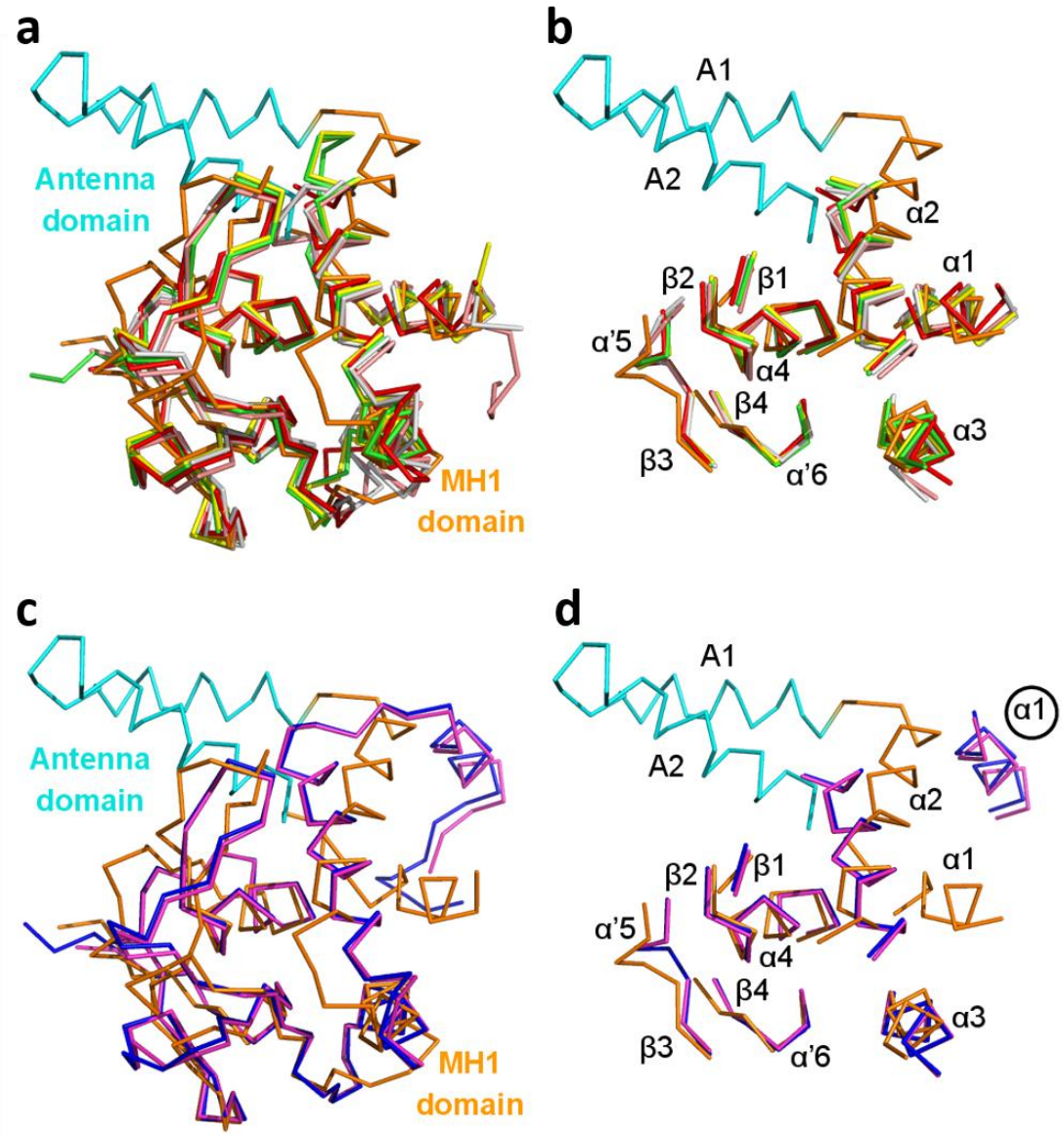


Figure 17. ***Structural superposition of NFIX M-L domain with Smad MH1 domains.*** The NFIX MH1 core domain (orange) is superimposed to the corresponding MH1 domain of **a**) human Smad2 (PDB-code 6H3R, yellow), human Smad3 (PDB-code 5OD6, green), mouse Smad4 (PDB-code 3QSV, pink), mouse Smad5 (PDB-code 5X6G, red), and *Trichoplax adhaerens* Smad4 (PDB-code 5NM9, gray), and **b)** human Smad9 (PDB-code 6FZT, blue), and mouse Smad1 (PDB-code 3KMP, magenta). For clarity, on the right panels only the superimposed secondary structure elements are shown. The NFIX MH1 core and the antenna domains are labeled (left panels), together with the secondary structure elements (right panels). In the b) panels, the different position of the N-terminal α1 helix of 6FZT and 3KMP is highlighted with a circle.

Table 7. ***Structural homologs to NFIX M-L domain*****.**

rmsd (Å)	DALI	Superpose*-CCP4 (all)	Superpose*-CCP4 (no loops)
5OD6 (Smad3)	2.6 (112)**	2.1 (110)	1.8 (90)
6HR3 (Smad2)	2.6 (108)	2.1 (106)	1.8 (87)
5NM9 (Smad4)	2.7 (111)	2.1 (103)	1.8 (85)
5X6G (Smad5)	2.7 (104)	2.1 (101)	1.9 (83)
6FZT (Smad 9)	6.5 (111)	2.2 (95: no α1)	1.9 (76: no α1)
3KMP (Smad 1)	4.6 (106)	2.2 (94: no α1)	2.0 (75: no α1)

*The option Gesamt (General Efficient Structural Alignment of Macromolecular Targets) (Krissinel, 2015) has been used to align structures by using the program Superpose of the CCP4 package (Winn *et al.*, 2011). Gesamt is an algorithm of efficient clustering of short fragments, where the fragments are made from adjacent protein backbone C-alpha atoms, followed by an iterative three-dimensional refinement based on a dynamic programming procedure.
**Aligned residues

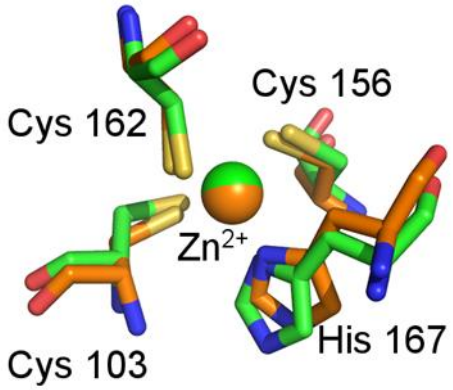


Figure 18. ***Structural superposition of the Zn^{2+} binding site in NFIX and Smad MH1*****.** Zn^{2+} ion coordinating residues are shown in stick representation and labeled. NFIX residues are in orange, and human Smad3 (PDB-code 5OD6), taken as representative of Smad proteins, in green.

II.4.9.3 DNA-binding mode

In the absence of the 3D structure of NFIX in complex with its cognate DNA, the analysis of the DNA-binding mechanism of Smad proteins may provide important pieces of information that can be transferred to NFIX. Several crystal structures of the Smad MH1 domain in complex with DNA are available (Aragon *et al.*, 2019; BabuRajendran *et al.*, 2010; Chai *et al.*, 2015; Martin-Malpartida *et al.*, 2017; Shi *et al*, 1998). In general, Smad proteins recognize a 4bp palindromic GTCT-AGAC defined as Smad-Binding Element (SBE) (Zawel *et al*, 1998). Consequently, two bound Smad MH1 domains are located

approximately on the opposite sides of the DNA duplex, making no physical interaction with each other. They bind identically to the major groove of the SBE, making hydrogen bond interactions to the bases and to the phosphodiester backbone of the DNA. The DNA contacts are provided by an 11-residue DNA-binding motif which build a β hairpin, formed by strands β2 and β3 (Figure 14), both formed by three residues connected by a sharp two-residue turn, which protrudes outward from the globular MH1 core (Figure 19a) (Shi *et al.*, 1998). The β hairpin binds asymmetrically at the DNA major groove, with strand β2 contributing on most of the DNA contacts (Figure 19a, right panel). Specifically, the conserved residues Arg74, Lys81 and Gln76 provide the sequence-specific interactions with the G of the SBE and with the G and A of the complementary strand, respectively. These interactions are conserved in different Smad proteins (Figure 14), while some variations are present in the residues that contact the DNA phosphodiester backbone. Some Smad proteins, such Smad3 and Smad4, have been reported to specifically recognize and bind also a palindromic CG-rich pentanucleotide site (Martin-Malpartida *et al.*, 2017). This implies the possible use of the Smad/DNA complex as a template to model the interaction of the NFIX M-L domain with its cognate pentanucleotide DNA, with the assumption the Smad proteins and NFIX bind to DNA in a similar fashion.
We can generate a plausible (NFIX M-L)$_2$/DNA model by substituting Smad3 with the NFIX M-L domain in a double Smad3/GGCGC complex (PDB-code 5OD6), separated by a 5bp linker, and modifying the DNA to match the palindromic NFI-binding sequence: TTGGC-5bp-GCCAA.

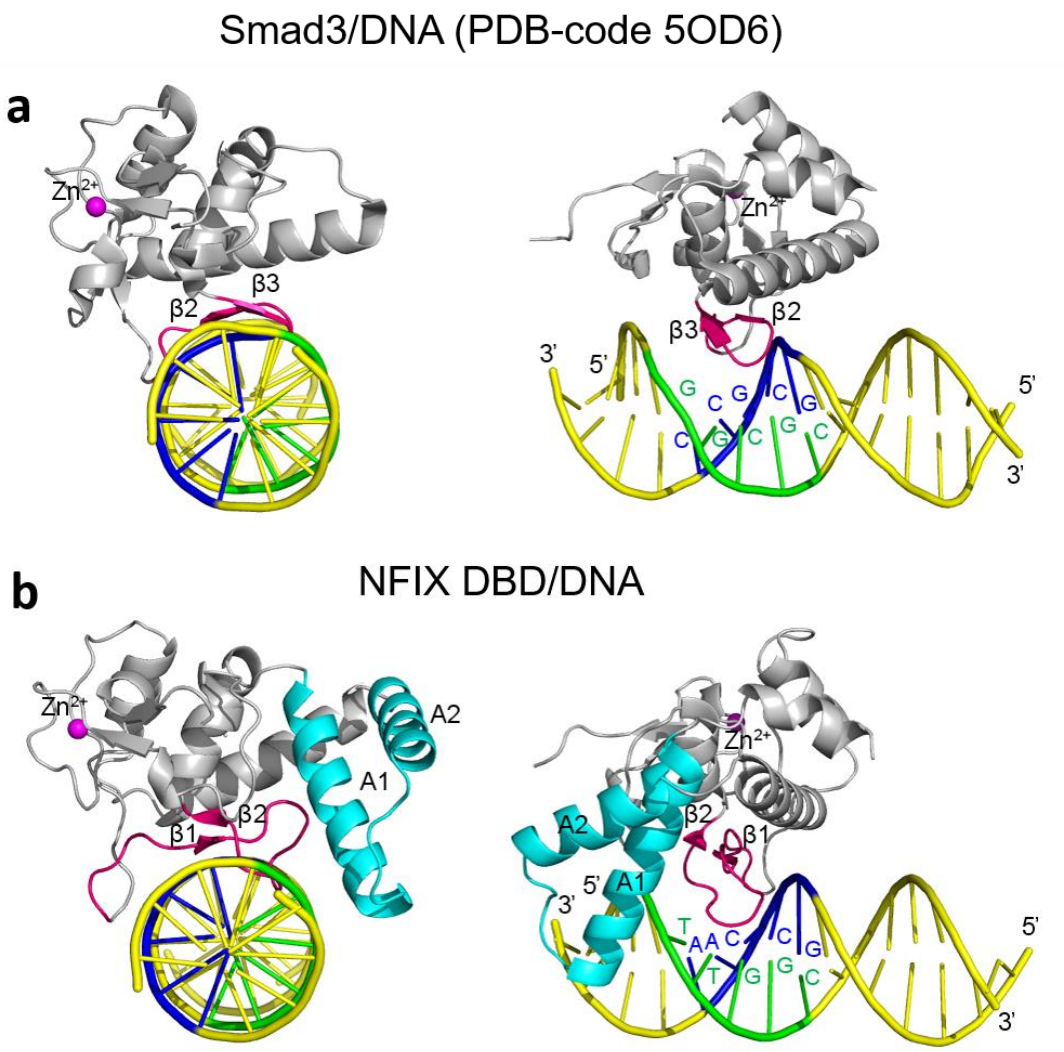


Figure 19. ***Model of the NFIX M-L domain/DNA interaction*****. a**) X-ray structure of the human Smad3/DNA(GGCGC) complex (PDB-code 5OD6) (Martin-Malpartida *et al.*, 2017). **b**) Model of the NFIX M-L domain/DNA interaction based on the Smad3/DNA(GGCGC) complex. The Zn^{2+} ion is shown as a magenta sphere. The NFIX M-L domain and Smad MH1 domain are shown in grey, with the DNA-binding loop, including the β-hairpin, in red, and the NFIX antenna domain in cyan. The NFIX and Smad recognition DNA sites are shown in green (forward strands: TTGGC and GGCGC, respectively) and blue (reverse strands: GCCAA and GCGCC, respectively).

In general, when the NFIX M-L domain is compared/superimposed to Smad, the NFIX β hairpin connecting the α3 and α4 helices matches very well the corresponding Smad region (Figure 19b), which inserts into the DNA major groove and specifically recognize the DNA bases. This is interesting and should not be taking for granted *a priori*, considering that NFIX M-L domain lacks the β1 and β5 strands that in Smads form a small β-sheet able to stabilize and properly orient the β hairpin inside the DNA major groove (Figure 14). On the other hand, it could also be envisaged that such a small β-sheet will form upon

NFIX-DNA binding, since all available Smad MH1 structures are DNA-bound. We can conclude that this small β-sheet is not necessary in the unbound form of NFIX to keep the β hairpin well-exposed on the surface of the M-L domain, ready to insert into the DNA major groove and specifically interact with the target sequence (Figure 19b). Despite a similar 3D location, the NFIX β hairpin differs from that of Smads in several aspects. It is three-residue longer (and also the preceding L1 loop is three-residue longer), the two connected β strands are shorter (only two residues each), and there is no amino acid conservation with Smads (Figure 14), except from the Arg116 and Lys125 that in Smads contact two G bases on opposite strands. Interestingly, the NFIX β hairpin includes three more residues that could potentially interact with DNA bases (Arg115, Arg121, and Gln122), and a Cys residue (Cys119) which has been demonstrated to be oxidation-sensitive and in close proximity of the bound DNA (Bandyopadhyay & Gronostajski, 1994). If we assume no conformational changes for this part of the protein upon DNA-binding, our NFIX model shows that both Arg116 and Lys125 are indeed facing the bases of the NFIX recognition DNA pentanucleotide. Arg115 and Arg121 points out of the DNA groove, being most likely involved in interactions with the DNA phosphodiester backbone. The position of Gln122 would allow it to contact the DNA bases, while Cys119 would not be in direct contact with DNA (Figure 20a). This is in agreement with mutational studies on rat NFIC showing that the majority of this Cys mutants are able to bind DNA, with the exception of Arg or Trp substitutions which abolished DNA-binding activity. From our model, we can propose that these largest side-chain substitutions are likely to have a strong structural impact on the following loop (including Arg121, Gln122, and Lys125), which would be in contact with DNA. A similar structural explanation would justify the inactivation of the TF by the addition of large chemical adducts at the Cys, such as NEM, diamide, and DTNB (Bandyopadhyay & Gronostajski, 1994).

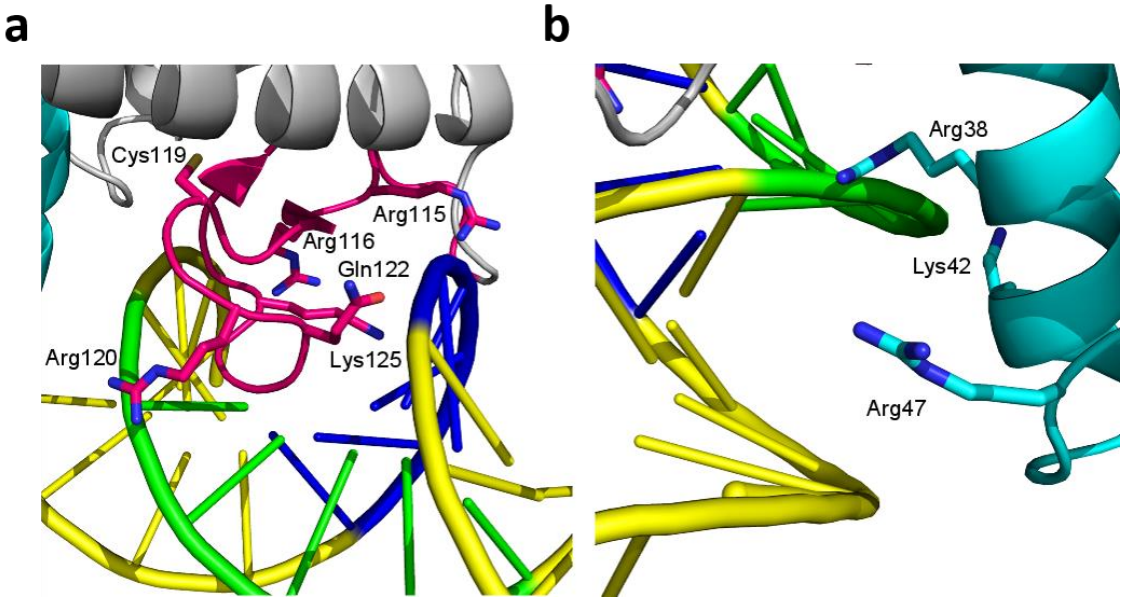


Figure 20. ***Detailed views of the modeled NFIX-2/DNA interface*****. a**) Insertion of the DNA-binding loop, including the β-hairpin into the DNA major groove. **b**) DNA-protein interface at the antenna domain. Relevant residues are shown in stick representation and labeled. Color code as in Figure 19.

The proposed model predicts that several positively charged or polar residues outside the β hairpin may interact with the negatively charged DNA phosphodiester backbone such as Arg38, Lys42, Arg47, Lys78, Lys81, Gln110, Lys111, Ser144, and Thr145 (Figure 21). Interestingly, Arg38, Lys42, and Arg47 belong to the antenna domain (Figure 20b), thus suggesting an important role in DNA-stabilization for this NFI-specific region of the M-L domain. This finding corresponds with previous mutational studies on rat NFIC showing that internal deletions involving regions mapping the A1, and the A2 helices of the NFIX antenna domain, generate proteins that have lost the capacity to bind DNA (Gounari *et al.*, 1990). Apparently, the antenna domain of the unbound NFIX M-L domain locates correctly to allow DNA-binding, without need of large conformational changes (Figure 19b). Finally, the C-terminal region of the NFIX M-L domain points towards the bound DNA, in keeping with the EMSA data that indicate that a NFIX construct with an elongated C-terminus (NFIX-3) binds DNA with higher affinity (see paragraph II.4.4).

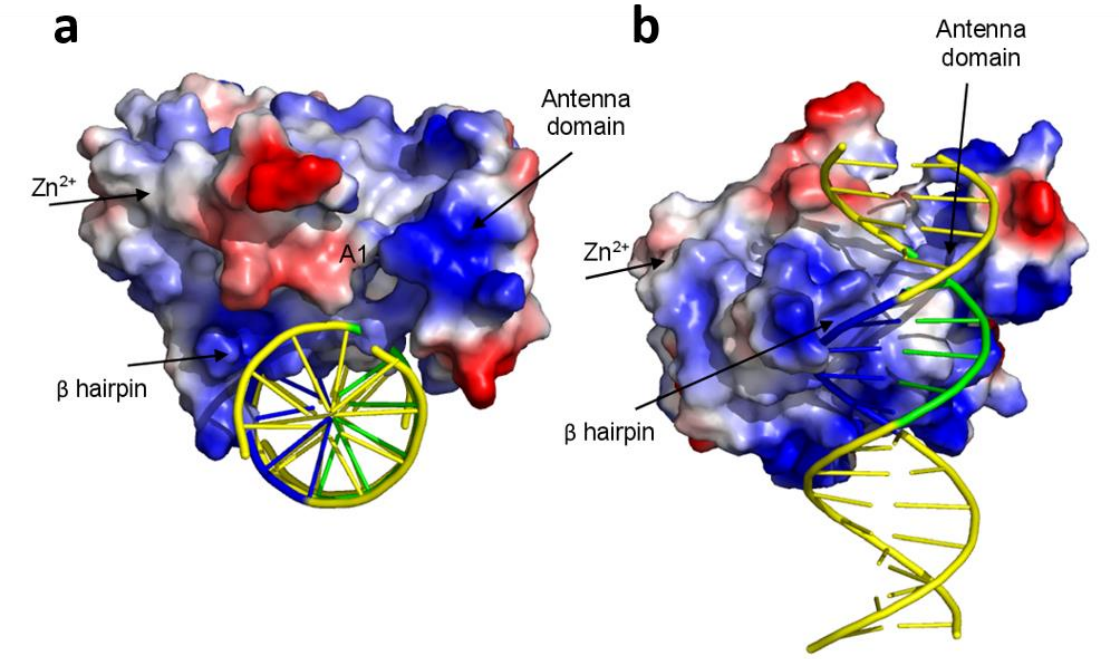


Figure 21. ***Electrostatic surface of the NFIX M-L domain.*** The surface of the NFIX M-L domain in complex with DNA is shown in blue and red for positively and negatively charged side chains, respectively. **a)** The orientation of the protein/DNA complex is identical to Figure 19a, while in panel **b)** the view is rotated of 90°. The location of the Zn^{2+}-binding site, of the antenna domain, and of the β hairpin is highlighted.

II.4.9.4 NFIX dimerization on DNA

Another interesting issue regards NFIX dimerization on DNA. The analysis of the $(NFIX\ M\text{-}L)_2$/DNA model (Figure 22), obtained by using the palindromic NFI-binding sequence TTGGC-5bp-GCCAA, reveals useful indications about the general quaternary architecture of the transcriptional complex. First, the two NFIX M-L domain molecules contact the DNA through two consecutive major groove turns, and second, both protein molecules accommodate at the same side of the DNA helix (Figure 22a), in agreement with previously reported contact-point experiments (de Vries *et al.*, 1987). Furthermore, in the dimeric DNA-bound model, the N-terminal regions of the proteins face each other (Figure 22b). In the case of a shorter DNA linker (4bp), the two facing Met13 would be in very close proximity to one another, and therefore steric clashes could occur between the N-termini of the dimer. This is in keeping with previous data showing that the shortening of the spacer region by one base-pair lowered the apparent affinity to a value similar to that observed with a single half-site. Arguably, a shorter spacer may impact simultaneous interaction of NFI monomers with both half-site sequences (Roulet *et al.*, 2000).

In this model, the two NFIX M-L domains, would be located too distantly on the DNA to allow a direct interaction (Figure 22b), thus suggesting that other regions of the protein might play a role in stabilizing the dimer. This observation is in line with deletion studies on rat NFIC, indicating that a ΔC206 mutant (similar to our 14-176 NFIX M-L domain construct) has an impaired capacity to oligomerize. Thus, the presence of the C-terminal region (or part of it) seems to be necessary to obtain a stable dimer. Interestingly, the mutant missing the antenna domain is still able to dimerize, suggesting that this region is not involved in oligomerization (Gounari *et al.*, 1990), as predicted in our (NFIX M-L)$_2$/DNA model.

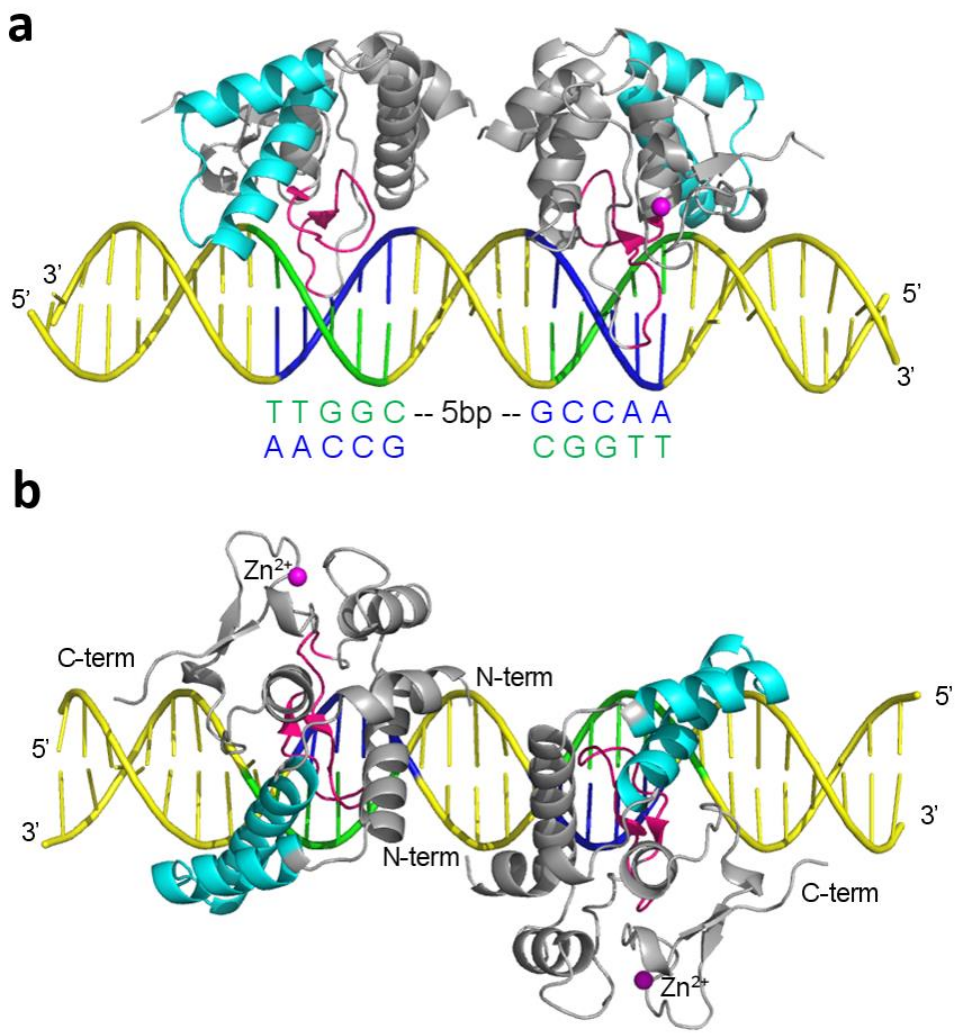


Figure 22. ***Model of the (NFIX M-L)$_2$/DNA interaction*****. a)** side and **b)** top views of the model of the (NFIX ML)$_2$/DNA interaction, based on the duplication of the model shown in Figure 19. The sequence of the DNA corresponds to the 31bp probe used for EMSA experiments (see II.4.4 paragraph) and contains the palindromic NFIX recognition site reported in panel a. Relevant regions of the DNA-bound proteins are labelled. Color code as in Figure 19.

II.5 CONCLUSIONS and FUTURE PERSPECTIVES

NFIX is a member of the NFI family of TFs which also includes NFIA, NFIB, and NFIC. At molecular level, NFIX binds as homo- or hetero-dimer the dyad consensus sequence TTGGC-5bp-GCCAA (Gronostajski, 1986). The N-terminal domain of NFI is sufficient for DNA-binding, whereas the C-terminal domain is implied in modulations by transcription partners (Gounari *et al.*, 1990; Jackson & Tjian, 1988; Mermod *et al.*, 1989; Mukhopadhyay & Rosen, 2007; Singh *et al.*, 2009). NFIX plays an essential role in multiple organ systems that have a large impact on human health. For instance, brain, hematopoietic stem cells, several types of cancer (*e.g*. medulloblastoma, squamous cell carcinoma, prostatic cancer and colorectal cancer), Malan syndrome and Marshall–Smith syndrome (Piper *et al.*, 2019). Nevertheless, NFIX has a prime role in skeletal muscle, leading muscle development and regeneration. In the muscular dystrophy context, the lack of NFIX induces a slower muscle regeneration, thus preserving the tissue and delaying the disease progression (Rossi *et al.*, 2017). With this respect, NFIX has been considered as a promising target for the development of therapeutic strategies for muscular dystrophy treatment. The knowledge of the atomic structure of NFIX and of its interactions with DNA would provide essential information for a rational understanding of its functional DNA-binding properties and, as a future perspective, for developing strategies for its selective inhibition. Up to date, there was no 3D structural information on either NFIX or on other NFIs. We report, for the first time, the successful production of recombinant NFIX and subsequent biochemical, biophysical and 3D structural characterization, describing the first crystal structure of the NFIX DBD.

Based on a preliminary bioinformatics analysis and considering the available information from literature, we focused our work on two NFIX truncated constructs, both including the putative identified DBD, but with a different length at the C-terminus: a shorter construct (14-176) named NFIX-2, and longer construct (14-203) named NFIX-3. In both cases, the C-terminal TAD region, expected to be intrinsically disordered from sequence analysis, was omitted from the constructs. The search for an acceptable expression/purification protocol for large-scale production represented one of the biggest hurdles. After several combinations of affinity chromatography and SEC, we established a successful heparin-based purification protocol for our MBP-His-NFIX recombinant prooteins. The biophysical characterizations of NFIX constructs, including CD, DLS, NMR and Thermofluor measurements, indicated good protein folding and

stability. *In vitro* analyses of DNA-binding by EMSA revealed that NFIX constructs bind DNA in a sequence-specific manner and with great affinity, in the order of nM. Therefore, the NFIX truncated constructs are functional. Moreover, EMSA experiments revealed an important difference between NFIX-2 and NFIX-3. NFIX-3 binds target DNA with 10-fold greater affinity, suggesting that the additional C-terminal residues in NFIX-3, may be involved in mediating DNA-binding/dimerization. Crystallization trials were set up for NFIX-2 alone and in complex with its target DNA. The former trials did not have success, whereas the second one did. The best NFIX-2 crystal diffracted at 2.7 Å resolution at the ELETTRA synchrotron. The structure was solved by SAD method on the absorption edge of zinc, based on the presence of the zinc previously identified by FAAS and confirmed by XRF spectroscopy on the protein crystals.

The DBD of NFIX is characterized by an MH1-Like (M-L) fold which includes a MH1 "core" region and an N-terminal "antenna" helical excursion. The NFIX MH1 core superimposes well with the corresponding domain of Smad proteins, including the Zn^{2+} binding site, coordinated by three cysteine (Cys103, Cys156, and Cys162) and one histidine (His167) residues. This is a very interesting finding, considering that the role of the conserved Cys residues in NFI has been long debated. Previous studies indicated an essential functional role for these cysteine residues in the N-terminal DBD, and their potential involvement in direct interaction with DNA and/or dimerization has been proposed (Novak *et al.*, 1992). Our structure reveals, instead, their role in zinc-coordination, and this result fits data reported in literature showing that NFI binds to the mouse metallothionein (MT)-1 promoter *in vivo* in a zinc-inducible manner (LaRochelle *et al*, 2008). Although two NFIX molecules are present in the ASU of the $P2_1$ crystal, structural analysis, gel-filtration and DLS data indicate that the NFIX M-L domain is a monomer in solution. A similar protein-protein packing was found also in a second crystal form ($P4_12_12$) where the loose dimeric assembly is between two crystallographic related molecules. The dimeric crystal form is stabilized by a HEPES molecule (from the crystallization solution) located at the dimer interface and by a Zn^{2+} ion coordinated by the His166 residues from two opposing monomers. The presence of this additional Zn^{2+} ion, confirmed by the analysis of the SAD data, may be ascribed to a marginal zinc dissociation from the main protein zinc-binding site (His166 is adjacent to His167 in sequence and in space) and/or from contamination of the crystallization solutions.

In the absence of an experimental structure of the NFIX/DNA complex, we can derive useful information on its DNA-binding mode by exploiting its similarity with Smad proteins and computational modelling. It appears that the Smad β hairpin, which recognizes and binds DNA in a sequence specific manner, is structurally conserved in NFIX, although with a different length and sequence. This finding implies a similar DNA-binding mode between Smad and NFIX and indicates residues of the NFIX β hairpin potentially involved in the recognition and binding of the NFI-binding sequence. It is likely that, in the presence of DNA, the NFIX β hairpin adjust its structure to optimize its interactions with DNA. Therefore, our computational modelling cannot provide the atomic details of this interaction but can suggest residues of the β hairpin to be tested for DNA-binding through mutagenesis and EMSA validation. Nevertheless, mutagenesis studies previously reported in literature for Cys119 of the β hairpin concur with our model. Interestingly, in our model the helical antenna domain would provide basic residues for DNA stabilization without requiring structural movement of the domain relative to its unbound position. The involvement of the antenna domain in DNA-binding agrees with previous mutational studies showing that internal deletions in the antenna domain of rat NFIC preclude DNA-binding (Gounari *et al.*, 1990).

The dimerization of NFIX on DNA can be inferred by duplicating the above protein/DNA complex on a palindromic DNA sequence. In this model the two NFIX-2 molecules bind independently the DNA thus suggesting that other regions of the protein, not included in our construct, may play a role in dimer stabilization. This observation is in line with deletion studies on rat NFIC, indicating that the sole DBD construct has an impaired capacity to oligomerize. Of course, this model assumes that the DNA-binding mode of NFIX is analogous to the Smad mechanism and that the formation of the protein/DNA complex does not bend the DNA. In this second scenario, the relative position of the two DNA-bound proteins could be different and not easily predictable.

For the near future, our first aim is the structure determination of NFIX in complex with its target DNA. Our EMSA data suggest using the NFIX-3 construct for this purpose, since it showed higher DNA-binding affinity. Crystallization trials will be attempted together with SAXS measurements. These latter experiments are performed in solution and, therefore, will provide a low resolution 3D envelope where we could fit our structure of the NFIX-2 construct and the DNA, so to observe how the proteins binds their palindromic target sequence, which regions are involved in dimerization, and whether or not protein-binding influences the bending of the DNA. In parallel, we will

challenge our protein/DNA model with a site-directed mutagenesis approach on residues of the β hairpin potentially in contact with the DNA-bases of the NFI sequence-aspecific motif. Furthermore, the role of Zn^{2+} will be also tested, in terms of DNA-binding modulation (EMSA), and structural stability (CD and thermal denaturation).

PART III:

NF-Y

III.1 INTRODUCTION

III.1.1 NF-Y TF and function

The NF-Y TF was identified in the late ‘80s, by the characterization of murine class II major histocompatibility complex (MHC-II) promoter (Dorn *et al*, 1987). This promoter comprises a conserved DNA regulatory module named Y-box, which includes a CCAAT DNA unit. The CCAAT box is a widespread eukaryotic recognition DNA element, found in the forward or reverse orientation (Mantovani, 1998). Bioinformatics studies established that this DNA element exists in every eukaryotic system studied by far (Suzuki *et al*, 2001). Precisely, the CCAAT box is present in about 30% of eukaryotic genes and, regarding humans, in 67% of their promoters (Testa *et al*, 2005).

Owing to the CCAAT box popularity, several DNA-binding proteins that are able to bind this sequence have been identified, and they harbour the word CCAAT in their names (i.e. CTF/NF1, C/EBP, and CDP) (Dorn *et al.*, 1987). Among them, NF-Y is the only one that strictly requires all five nucleotides of this element for binding. Indeed, analysis of CCAAT-binding proteins revealed that their consensus binding sequence occasionally contained an intermediary CCAAT. An example closely related to this thesis work is represented by the so-called CCAAT-Transcription Factor (CTF/NF1), which binds the bipartite consensus TTGGC(n5)GCCAA as a dimer. Clearly, a T following the core consensus would give rise to a CCAAT, but it is not strictly necessary for proper binding. Instead, clear evidence pointed out that NF-Y strictly counts on the intact CCAAT pentanucleotide to bind DNA (Mantovani, 1998). Saturation mutagenesis studies, and most recent ChIP-on-chip and ChIP-seq approaches, defined NF-Y as the prime CCAAT-binding factor (Dolfini *et al*, 2009; Dorn *et al.*, 1987; Mantovani, 1998). Also, NF-Y demands a strong preference for flanking nucleotides to achieve high-affinity interactions (Mantovani, 1998). It was demonstrated a prevalence of two purine bases at -2 and -1 positions, and of CAG trinucleotide just following the core sequence (position +1, +2, and +3, respectively) (Figure 23).

The CCAAT box is generally found at -60/-100bp from the transcriptional start site, in the close vicinity of other promoter elements, hence NF-Y’s major role

is to act synergistically with other TFs (Romier *et al*, 2003). From a functional point of view, NF-Y facilitates TFs binding and promoter architecture organization (Nardini *et al.*, 2013).

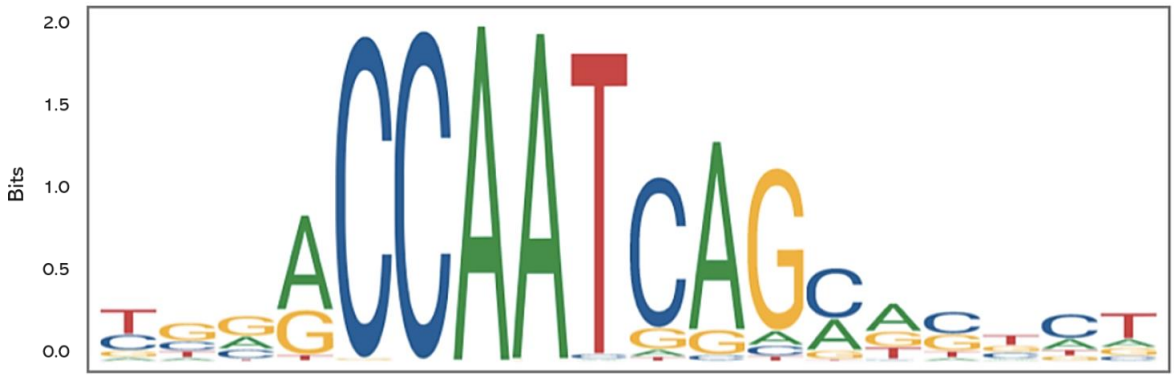


Figure 23. *The CCAAT box*. Sequence logo of the ChIP-seq derived NF-Y binding motif with JASPAR database. Adapted from Dolfini, 2009.

NF-Y works as a heterotrimeric TF, composed of subunits NF-YA, NF-YB, and NF-YC. NF-Y subunit homologs are present in all metazoan, plants, fungi, and protists, pointing out that NF-Y is a column carrier in the eukaryotic domain evolution (Li *et al*, 1992). NF-Y yeast homolog is called HAP complex, composed of HAP2 (NF-YA), HAP3 (NF-YB), HAP5 (NF-YC), and a fourth subunit named HAP4 (Forsburg & Guarente, 1988). The HAP complex binds to the CCAAT in the upstream activation sequence of numerous cytochrome genes, and it is a master regulator of respiratory metabolism. Well-defined roles of NF-Y in developmental and stress response pathways in *Drosophila* and *C. elegans* were established. In zebrafish *D. rerio*, NF-Y expression is crucial for cartilages and notochord development (Dolfini *et al*, 2012).

In mammals, invertebrates, and fungi, there are one or two genes coding for each subunit. Instead, plants have dramatically expanded the number of NF-Y genes, giving rise to multiple tissue-specific and stimulus-specific heterotrimeric combinations (Gnesutta *et al*, 2017). In *Arabidopsis thaliana*, growing pieces of evidence indicate that NF-Y subunits are involved in countless physiological events, such as development, growth, reproduction, adaptation to physiological and adverse environmental conditions (Laloum *et al*, 2013; Petroni *et al*, 2012). Within mammals, the conservation at the protein level is the highest, with more than 90% sequence identity with the human orthologues. NF-Y subunits expression is ubiquitous in most human and mouse cell types. In mammals, NF-Y is essential during early developmental stages; indeed, NF-YA knockout mouse is embryonic lethal due to a block of cell proliferation and induction of apoptosis (Dolfini *et al.*, 2012). NF-Y regulates the expression of many mammalian's cell cycle and house-keeping genes and it is needed for cell proliferation and development. In fact, not only does NF-Y regulate genes with

housekeeping functions through cell type-invariant promoter-proximal binding, but also genes required for cell identity by binding to cell-type-specific enhancers with master TFs (Oldfield *et al*, 2014). It also plays a role in regulating various cell-type-specific genes under different developmental signalling or pathogenic conditions (Maity, 2017). In addition, NF-Y controls biosynthetic pathways of lipids, activates glycolytic genes, and mainly represses mitochondrial respiratory genes (Gurtner *et al*, 2017).

III.1.2 NF-Y structure and DNA-binding mode

In the cytoplasm, a tight NF-YB/NF-YC heterodimer is assembled and then translocated into the nucleus (Goda *et al*, 2005). Here, NF-YA, which is the subunit containing the sequence-specific DBD, binds to NF-YB/NF-YC, forming the functional trimer (Figure 24a). All three NF-Y subunits are required for DNA-binding, allowing the NF-Y complex to recognize and bind CCAAT box sites with an affinity in the 10^{-10} / 10^{-11} M range (Dolfini *et al.*, 2012).

Early sequence alignments revealed that each subunit includes an evolutionarily conserved core domain, positioned at the C-terminal, central and N-terminal regions of NF-YA, NF-YB, and NF-YC subunits, respectively. Both NF-YB and NF-YC contain a histone fold-domain (HFD) that allows them to dimerize (Figure 24b) (Romier *et al.*, 2003). NF-YA and NF-YC comprise also a Glutamine-rich (Q-rich) TAD domain at the N- and C-terminus, respectively (Figure 24b). In NF-YA, the DNA-binding and the trimerization regions are both located at the C-terminus (Figure 24b) (Nardini *et al.*, 2013).

In mammals, the NF-YB gene does not undergo alternative splicing, giving rise to one major protein product of 32 kDa. Instead, NF-YC alternative splicing mainly occurs within the Q-rich TAD, giving rise to three isoforms identified by their molecular weight: 37 kDa, 48 kDa, and 50 kDa (Ceribelli *et al*, 2009). NF-YA can also give rise to two major isoforms: NF-YA long (NF-YA*l*) and NF-YA short (NF-YA*s*), the latter lacking 28 amino acids within the Q-rich TAD (Figure 24b) (Gurtner *et al.*, 2017). Several studies support the idea that NF-YA*l* and NF-YA*s* exert different biological roles. Indeed, NF-YA*s* has been identified as a regulator of stemness and proliferation in mouse and human embryonic cells (m/hESCs) (Dolfini *et al*, 2019). In support of this, experiments in two non-transformed systems (Hematopoietic Stem cells and mESC) indicate that NF-YA*s* is more abundant in “stem”, whereas NF-YA*l* in differentiated cells. Therefore, NF-YA*s* is associated with a proliferative signature (Gurtner *et al.*, 2017).

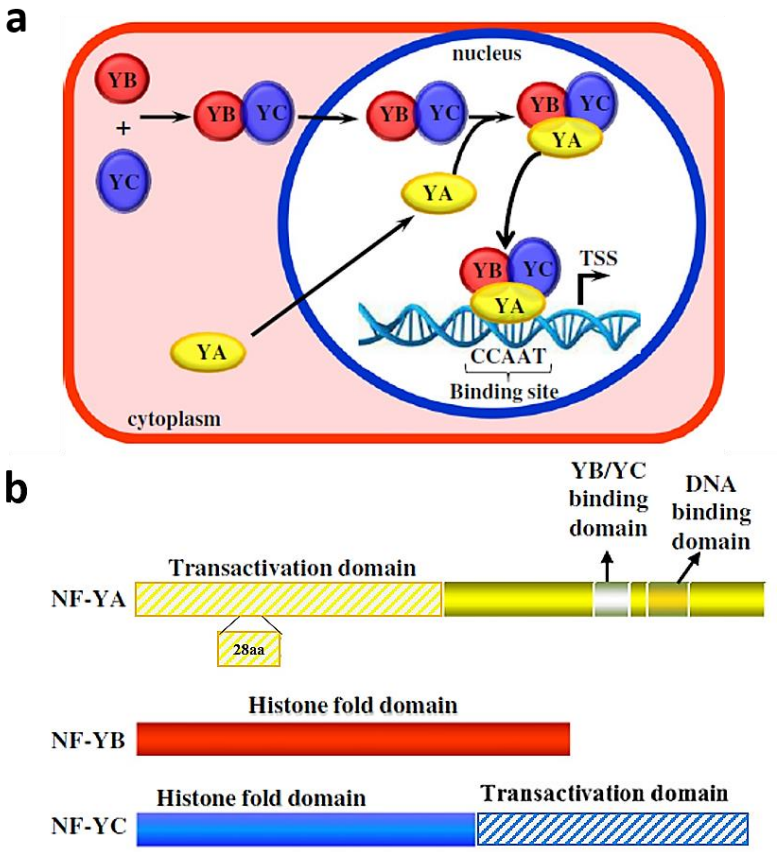


Figure 24. ***NF-Y subunits assembly and composition*****. a)** NF-YB/NF-YC heterodimer is assembled in the cytoplasm, translocated in the nucleus where it binds the NF-YA subunit. The resulting trimer is able to bind the CCAAT sequence on DNA. **b)** Schematic representation of NF-Y subunits, including the differential NF-YA splicing isoforms. Adapted from (Gurtner *et al.*, 2017).

Advancements in the understanding of the details of NF-Y molecular architecture were made through structural biology. Since the earliest characterizations in the second half of the 90s, it was evident that NF-YB and NF-YC subunits were related to core histones H2B and H2A, respectively (Baxevanis *et al*, 1995). In particular, NF-YC and NF-YB display HFD at their N-terminal region (Figure 24b), and these domains serve to form a tight histone-fold heterodimer (Romier *et al.*, 2003). The NF-YB and NF-YC secondary structure is composed of three α-helices connected by three loops. From N- to C-terminal, they are named α1, α2, α1, α3 and αC; the connecting loops are named L1 (between α1 and α2), L2 (between α2 and α3), and LC (between α3 and αC) (Nardini *et al.*, 2013). The structural homology of the NF-YB/NF-YC with H2B/H2A HFD is evident from the superimposition of the respective quaternary structures (Figure 25a) and extends to the additional C-terminal loop-αC element present in both subunits. A characteristic intra-chain Arg-Asp bidentate salt bridge between L2 and α3 in each subunit stabilizes the NF-YB/NF-YC HFD (Romier *et al.*, 2003). The result is a compact heterodimer

stabilised by extensive hydrophobic interactions between the two HFDs arranged in a “hand-shake” assembly (Figure 25b).

The first crystal structure of human NF-YB/NF-YC dimer (PDB-code 1N1J) (Romier *et al.*, 2003) led to the first observation of its histone-like architecture and set the bases for the subsequent crystallization of a mammalian NF-Y heterotrimeric complex bound to CCAAT box DNA (PDB-code 4AWL) (Nardini *et al.*, 2013). Crystals were grown by using the minimal trimerization/DNA-binding core domains of each subunit. The human heterotrimer was complexed with a 25bp double-stranded DNA oligonucleotide harbouring the CCAAT sequence from the HSP70 promoter (Li *et al*, 1998). The crystal structure of NF-Y heterotrimer in complex HSP70 promoter DNA was solved at 3.1 Å resolution. In the NF-Y trimer, NF-YB/NF-YC dimerization provides a stable scaffold, which is a prerequisite for NF-YA association and DNA-binding (Nardini *et al.*, 2013).

The upper region of the HFD exposes an extended basic surface, responsible for non-specific contacts with the DNA backbone (Nardini *et al.*, 2013). A genuine characteristic of the NF-YB/NF-YC HFD dimer with respect to other HFD-containing proteins is the presence of an acidic surface groove, built by NF-YC αC, NF-YC α1, and NF-YB α2 residues, responsible for the binding of the NF-YA subunit (Figure 25c) (Nardone *et al*, 2017). While NF-YB/NF-YC heterodimer provides a wide positive surface for the non-specific DNA contacts, NF-YA is devoted to CCAAT recognition, thus accounting for DNA-binding specificity of the protein complex (Figure 25c) (Nardini *et al.*, 2013).

The NF-YA subunit displays an elongated structure that hosts the N-terminal and the C-terminal helices, named A1 and A2, followed by a loop containing a GxGGRF motif (named Gly loop; x = any residue). The two helices are separated by a 15-residue linker loop (A1A2 linker) (Figures 25b and c). The NF-YA A1 helix docks in the extended acidic groove displayed by the HFD thanks to specific positively charged sidechains, thus allowing trimerization (Figure 25c) (Nardini *et al.*, 2013).

The structural hallmark of the NF-Y/DNA complex is the wide insertion of NF-YA A2 helix into the DNA minor groove at the first CCAAT adenine, hence imposing CCAAT-binding specificity. The sequence recognition is achieved thanks to Arg274, His277, Arg281, Arg288, and Phe289, key residues of the A2 and to the kinked backbone of the Gly-loop that directly place functional contacts with the CCAAT nitrogenous bases (Figure 25d) (Nardone *et al.*, 2017). Moreover, the Phe289 residue belonging to the Gly-loop is inserted between the two consecutive C and A base pairs of the CCAAT. This induces a positive roll of 48° between those base pairs, at the centre of the CCAAT box (Figure 25d). Overall, the NF-YB/NF-YC convex DNA-binding surface and minor-groove

insertion by NF-YA result in a global DNA bending angle of ~80° (Figure 25c), similar to nucleosome DNA (Nardini *et al.*, 2013).

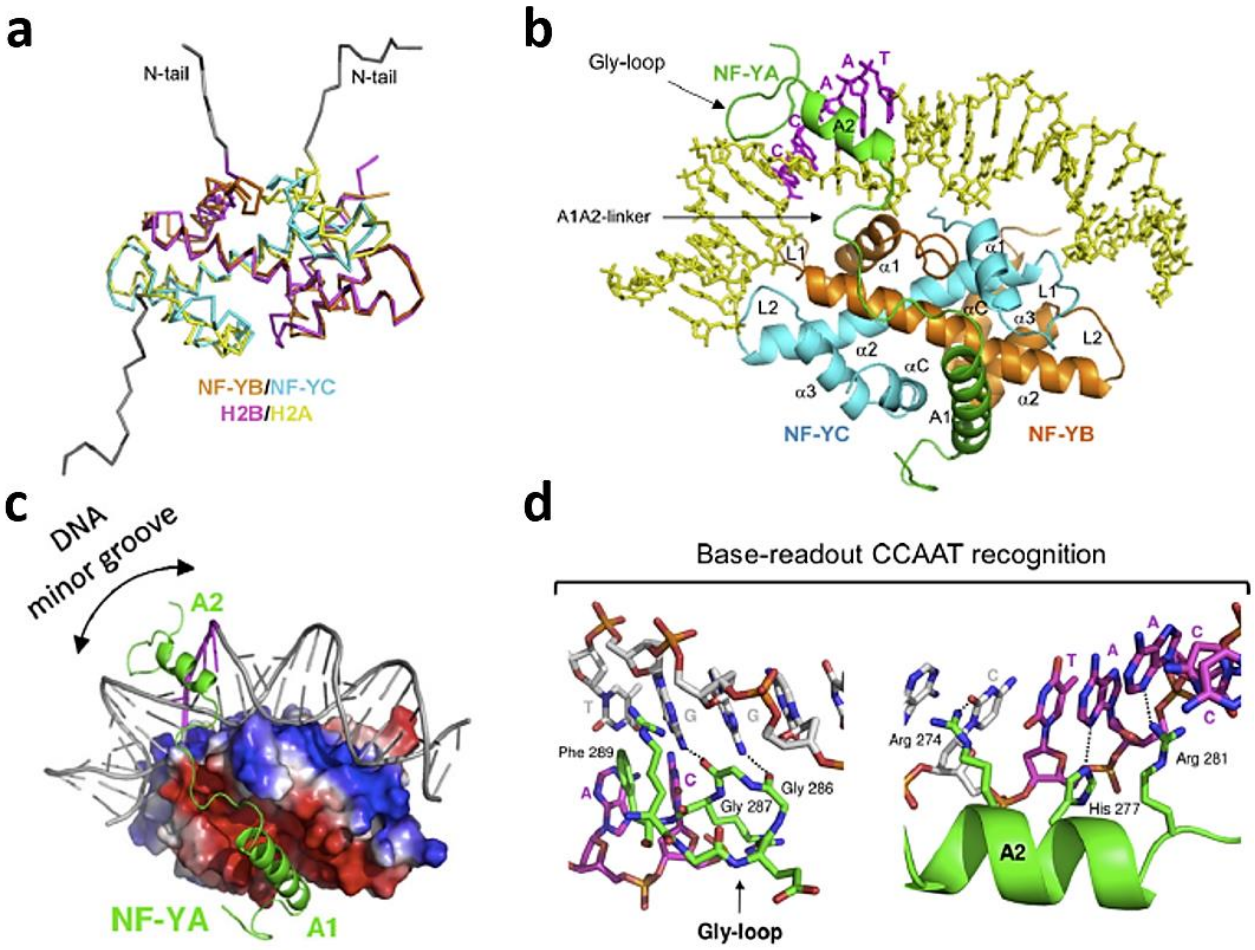


Figure 25. ***NF-Y/DNA structure*****. a)** Structural superimposition of coil diagrams between NF-YB/NF-YC HFD dimer (PDB-code 4CSR) and H2B/H2A dimer (PDB-code 1AOI, chains C and D). **b)** Ribbon representation of the NF-Y core heterotrimer in complex with a 25 bp oligonucleotide from HSP70 promoter CCAAT box (PDB-code 4AWL). **c)** Electrostatic surface of HFD NF-YB/NF-YC dimer in complex with NF-YA and DNA as in b. Red and blue colours indicate negatively and positively charged regions, respectively. NF-YA minor groove insertion is highlighted. **d)** Close-up on the base-readout mechanism employed by NF-YA to specifically recognize CCAAT nucleotides. The CCAAT nucleotides are in magenta, complementary strand bases in grey; hydrogen bonds interactions involving Gly-loop and A2-helix residues are indicated by dashed lines. Adapted from (Nardini *et al.*, 2013).

Besides DNA bending, a striking feature of the complex is the extension of the contacted DNA, spanning at least 25 bp. There are 41 detected protein-DNA contacts, widely distributed on the protein-DNA interface, explaining the high-affinity of the complex for consensus DNA. The combination of HFD non-specific DNA contacts provided by NF-YB/NF-YC dimer and CCAAT

recognition by NF-YA subunit might suggest a scouting model in which the preassembled heterotrimer transiently contacts the DNA backbone through its HFD component. Structurally, the A1A2 linker adopts an extended conformation that provides the flexibility required to direct the NF-YA chain toward DNA (Nardini *et al.*, 2013). The complex could proceed by local sliding on the DNA surface, while the flexible NF-YA A1A2-linker allows the A2-helix residues to constantly search for high-affinity CCAAT nucleotides, thereby locking the complex in place with the Gly-loop and Phe318 side chain insertion. Interestingly, the A2 region is unstructured in the complex prior to DNA binding (Huber *et al*, 2012).

The resulting NF-Y/DNA complex has a compact aspect, with the bent DNA that faithfully follows the HFD upper convex surface. Note that, however, both NF-YA and NF-YC full-length proteins possess long and disordered N- and C-terminal Gln-rich TADs, respectively. These regions make up a large portion of the full-length complex and, although not necessary for DNA-binding, they may play an important functional role in the interaction of NF-Y with other TFs.

III.1.3 The pioneering action of NF-Y and its role in cancer

Most TFs are unable to gain access to repressed or non-modified chromatin domains, even if high-affinity binding sites are present. Exceptions to this paradigm are the so-called 'pioneer' TFs, whose capability to colonize neutral or hostile chromatin environments is cardinal during cell differentiation, but also reprogramming and establishment of altered transcriptional patterns in cancer cells, where they have been used as prognostic biomarkers (Magnani *et al*, 2011). The main function of this subclass of TFs is to determine gene expression, either by promoting the cooperative binding of a second, non-pioneer TF or by recruiting chromatin remodelling/modifying complexes which in turn physically provide DNA access to other TFs (Morris, 2016). The following studies demonstrated that NF-Y belongs to the pioneer TFs class. First, a new unbiased computational method that models the magnitude and shape of genome-wide DNase I profiles to identify TF binding sites. This method was applied to differentiating mESCs and led to the identification and experimentally validation of 'pioneer' TF families that dynamically open chromatin, enabling other TFs to bind to adjacent DNA. Among them, NF-YA was identified as a 'directional-pioneer', since the open-chromatin state was asymmetrical and strongly enriched towards the CCAAT downstream region (Sherwood *et al*, 2014). A second piece of evidence demonstrated that NF-Y's distinct DNA-binding mode facilitates a permissive chromatin conformation and promotes enhanced binding of master ESC TFs at enhancers. NF-Y promotes chromatin accessibility for the key pluripotency factors Oct4 and Sox2 in mouse embryonic stem cells (mESCs);

therefore the binding of these master TFs is dependent on the pioneering action of NF-Y (Oldfield *et al.*, 2014). In addition, genome-wide mapping of NF-Y binding sites from the ENCODE project from separate cell lines revealed that a significant portion of NF-Y locations falls in repressed chromatin regions (Fleming *et al*, 2013). The challenge is to define the molecular mechanism by which NF-Y can gain access to repressed chromatin DNA.

Concerning the few other examples of TFs studied in their pioneer activity, they all share the capability to interact with nucleosomes (Sekiya *et al*, 2009; Soufi *et al*, 2015). All pioneer TFs studied so far bind their target sites on the surface of the nucleosome, to then establish and extend a competent chromatin environment for other TFs.

In nucleosomes, two left-handed super-helical turns of bent DNA are wrapped around an octameric-histone core consisting of two copies each of the core histone proteins H2A, H2B, H3, and H4. At the heart of the octamer, there are two H3/H4 heterodimers with two H2A/H2B heterodimers bound to opposite faces (Luger *et al.*, 1997). NF-Y/DNA complex is remarkably similar to nucleosome components H2A/H2B. Notably, DNA curvature induced by NF-Y is comparable to nucleosome's DNA bending (Drew & Travers, 1985). An appealing model would assume that NF-Y pioneer activity would rely on its HFD subunits and their natural homology with H2A/H2B, as observed in Figure 25a, suggesting an evolutionary relationship of NF-YB/NF-YC to core histone proteins (Nardini, 2013). In fact, NF-YB/NF-YC dimer has been reported to interact with histones H3/H4 tetramers during nucleosome assembly *in vitro* and NF-YA can interact to such preformed NF-YB/NF-YC/$(H3/H4)_2$ complex and impart CCAAT specificity (Caretti *et al*, 1999; Motta *et al*, 1999).

The pioneering role of NF-Y is particularly relevant in cell proliferation and transformation (Gurtner *et al.*, 2017). A plethora of NF-Y targets is upregulated in different types of cancer. Analysis of transcriptome profiles in normal *vs* tumour cells from different tissues (breast, colon, leukaemia, prostate, thyroid) has shown that NF-Y binding sites are enriched in promoters of genes specifically overexpressed in diverse types of cancers (Dolfini *et al.*, 2012). NF-Y subunits have not been found to be altered, mutated, or grossly overexpressed in cancer, yet, a growing number of profiling experiments support the notion that activation of CCAAT-dependent genes is crucial in changing the transcriptome profiles during cell transformation (Dolfini *et al.*, 2009). Consequently, specific cancer-driving nodes are under NF-Y control (Gurtner *et al.*, 2017).

NF-Y, through its binding to the CCAAT motif present on promoters, contributes to the modulation of several cancer-associated genes. Inevitably, the list of proteins deemed to interact with NF-Y is long and rapidly increasing, the most significant being: p53/p63/p73, C/EBPα/ζ, Smad2/3, E box proteins, USF1/2, MYC/FOS, and APC, TFIID. In most cases, the interactions lead to

synergistic activation of transcription, typically through neighbouring binding sites (Dolfini *et al.*, 2012). The interaction between NF-Y and wild type (wt) / mutant (mut)p53 is crucial (Imbriano *et al*, 2012). The NF-Y/(wt)p53 complexes recruit histone-deacetylases (HDACs) on repressed promoters causing cell cycle arrest. On the other end, in cells carrying (mut)p53 the NF-Y/(mut)p53 complexes have the opposite effect causing transactivation of proliferative genes involved in cell transformation (Di Agostino *et al*, 2006).

Interestingly, while the NF-YB and NF-YC subunits are mainly ubiquitously expressed, the NF-YA subunit is downregulated in some postmitotic cells. Thus, from a biochemical perspective, NF-YA is the limiting subunit of the trimer. Loss of NF-YA expression results in loss of a functional NF-Y complex, suggesting that, although NF-Y is a ubiquitous TF, differential expression of the NF-YA subunit can occur during growth and differentiation of different cell types (Gurtner *et al.*, 2017).

Furthermore, NF-YA is present in two alternatively spliced isoforms, NF-YA*l* and NF-YA*s*. In most of the experimental conditions and cell culture in which they have been tested, the two isoforms showed similar activities and hence for this reason were long thought to be functionally equivalent. The first indirect evidence that they have different biological effects came from analysis of proliferating human fibroblasts, where NF-YA*l* was the most expressed isoform. However, upon SV40-dependent oncogenic transformation of the cells, the expression of the NF-YA*s* isoform increased (Gu *et al*, 1999). Indeed, NF-YA*s* enhances cell proliferation, while NF-YA*l* inhibits it (Basile *et al*, 2016). Of note, the NF-YA*l* is the main form expressed in benign tissues, whereas NF-YA*s* is almost absent in these samples (Gurtner *et al.*, 2017).

III.1.4 NF-Y as a target of anti-cancer drugs

Since NF-Y modulates the transcription of multiple genes involved in cell transformation, several efforts have been made to inhibit NF-Y transcription activity as an anti-cancer strategy. Interfering with NF-Y pioneer activity for oncogenic activators could bear key implications for cancer control. So far, such search has been focused on minor-groove binding drugs e.g. pyrrole-imidazole polyamides, which are sequence-specific compounds that bind DNA non covalently, affecting the DNA structure and, potentially, the activity of TFs (Dervan & Edelson, 2003). These polyamide intercalating derivatives compounds had been tested to interfere with Topoisomerase IIα (Topo IIα) promoter, being an essential nuclear enzyme and the primary target for several clinically important anticancer drugs (Chen *et al*, 2011). Polyamide compounds were able to block interactions of NF-Y with the promoter of Topo IIα by binding at the AT-rich sequences corresponding to the preferred binding site of

NF-Y and to displace NF-Y trimer bound to the CCAAT box (Kotecha *et al*, 2008).

A completely different approach reached similar results, by using a miRNA (miR-485-3p) which targets NF-YB. NF-YB causes down-regulation of Topo IIα in drug-resistant cells, whereas miR-485-3p leads to an increase of Topo IIα expression through NF-YB down-regulation (Chen *et al.*, 2011).

Another set of compounds that alter NF-Y activity are HDAC-inhibitors. In most studies, the Class I and II inhibitor TSA, and the current pharmaceutical compounds SAHA and VPA, have been employed. As a result, HDACs inhibition impacts on NF-Y acetylation, as well as non-specifically increase of core histones H3 and H4 acetylation, leading in turn to the "opening" of a large set of chromatin and to increase promoter activity of a plethora of genes (Dolfini *et al.*, 2012).

Various other compounds have been shown to alter the activity of NF-Y. Genistein, which is a phytoestrogen contained in soy and a potent inhibitor of cell proliferation, antagonized the binding of NF-Y to the CCAAT sequences in the HSP70 promoter ER-stress genes (Zhou & Lee, 1998). Quercetin, belonging to the subclass of flavonoid, was shown to inhibit NF-Y binding to the cyclin B1 promoter leading to cell-cycle repression (Kim *et al*, 2008).

The number of studies searching for chemical/biological compounds able to inhibit NF-Y activity is significantly long. Altogether, the proof of principle of altering NF-Y binding *in vitro* and *in vivo* has been obtained, and this line of experiments shows promise to specifically target subclasses of CCAAT boxes, implicated in specific molecular pathways. The findings clearly indicate that in specific cellular contexts, the strategy to impair the activity of a master regulator of gene expression could be a valuable way to improve anticancer therapies (Gurtner *et al.*, 2017).

However, the mechanisms behind the above observations, whether exerting their effect through direct binding to any of the subunits, are completely unknown. So far, the numerous data on compounds and drugs affecting NF-Y activity await a rationalization by *in silico* and *in vitro* experiments with available structures of NF-Y. Thus, limiting NF-Y activity may represent a desirable anti-cancer strategy, which is an ongoing field of research.

Here, we present a detailed study of the molecular bases that underlie the modulation of DNA-binding activity of NF-Y by an already known drug, suramin.

III.2 AIM

The NF-Y project is the result of a long-term collaboration with multiple research partners. Among several ongoing lines of research, one is focussed on the search for compounds able to modulate/inhibit the DNA-binding of NF-Y, which may represent a desirable anti-cancer strategy. A preliminary virtual-screening approach on a library of pharmacologically active compounds, allowed us to identify suramin as a potential NF-Y inhibitor. My contribution in the NF-Y:suramin project includes the biochemical/biophysical characterization of the NF-Y/suramin interaction, focused on ITC experiments, and the refinement and analysis of the NF-Y/suramin crystal complex, whose data were previously collected.

III.3 ARTICLE

cells
an Open Access Journal by MDPI

MDPI

Article

Structural Basis of Inhibition of the Pioneer Transcription Factor NF-Y by Suramin

Valentina Nardone [1,†], Antonio Chaves-Sanjuan [1,†], Michela Lapi [1,†], Cristina Airoldi [2], Andrea Saponaro [1], Sebastiano Pasqualato [3], Diletta Dolfini [1], Carlo Camilloni [1], Andrea Bernardini [1], Nerina Gnesutta [1], Roberto Mantovani [1] and Marco Nardini [1,*]

[1] Department of Biosciences, University of Milano, Via Celoria 26, 20133 Milano, Italy; valentina.nardone83@gmail.com (V.N.); antonio.chaves@unimi.it (A.C.-S.); michela.lapi@unimi.it (M.L.); andrea.saponaro@unimi.it (A.S.); diletta.dolfini@unimi.it (D.D.); carlo.camilloni@unimi.it (C.C.); andrea.bernardini@unimi.it (A.B.); nerina.gnesutta@unimi.it (N.G.); mantor@unimi.it (R.M.)

[2] Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza 2, 20126 Milan, Italy; cristina.airoldi@unimib.it

[3] Department of Experimental Oncology, IEO, European Institute of Oncology IRCCS, Via Adamello 16, 20139 Milan, Italy; sebastiano.pasqualato@ieo.it

* Correspondence: marco.nardini@unimi.it; Tel.: +39 0250314893

† These authors equally contributed to the work.

Received: 28 September 2020; Accepted: 26 October 2020; Published: 29 October 2020.

Abstract: NF-Y is a transcription factor (TF) comprising three subunits (NF-YA, NF-YB, NF-YC) that binds with high specificity to the CCAAT sequence, a widespread regulatory element in gene promoters of prosurvival, cell-cycle-promoting, and metabolic genes. Tumor cells undergo "metabolic rewiring" through overexpression of genes involved in such pathways, many of which are under NF-Y control. In addition, NF-YA appears to be overexpressed in many tumor types. Thus, limiting NF-Y activity may represent a desirable anti-cancer strategy, which is an ongoing field of research. With virtual-screening docking simulations on a library of pharmacologically active compounds, we identified suramin as a potential NF-Y inhibitor. We focused on suramin given its high water-solubility that is an important factor for *in vitro* testing, since NF-Y is sensitive to DMSO. By electrophoretic mobility shift assays (EMSA), isothermal titration calorimetry (ITC), STD NMR, X-ray crystallography, and molecular dynamics (MD) simulations, we showed that suramin binds to the histone fold domains (HFDs) of NF-Y, preventing DNA-binding. Our analyses, provide atomic-level detail on the interaction between suramin and NF-Y and reveal a region of the protein, nearby the suramin-binding site and poorly conserved in other HFD-containing TFs, that may represent a promising starting point for rational design of more specific and potent inhibitors with potential therapeutic applications.

Keywords: transcription factor; histone fold; CCAAT box; NF-Y; suramin; inhibition

1. Introduction

The transcription factor (TF) NF-Y is a nuclear protein that binds the CCAAT sequence in promoters with a very high specificity [1]. The CCAAT box is an important regulatory element, typically located at a conserved distance of −60/−100 bp from the Transcriptional Start Site (TSS) and it is present in 25% of eukaryotic promoters [2]. This occurrence is similar to that of the TATA box, and the CCAAT box is mostly found in TATA-less promoters [2,3]. Genome-wide assays and *in vitro* experiments have demonstrated that NF-Y is the primary CCAAT-binding protein [4].

NF-Y is a heterotrimer formed by evolutionarily conserved subunits: NF-YA, NF-YB, and NF-YC. NF-YB and NF-YC form a heterodimer *via* interacting histone fold domains (HFDs), while NF-YA provides DNA sequence-specificity to the trimer. A multitude of genes have been described to be positively or negatively regulated by NF-Y, including prosurvival and cell-cycle-promoting genes, in addition to genes involved in metabolism [5–10]. Regarding metabolism, the NF-Y yeast-homologue HAP2/3/4/5 was originally identified as the activator of oxygen-fueled metabolism in the presence of non-fermentable carbon sources, by binding the CCAAT box at Upstream Activating Sequences (UAS) of nuclear genes of the mitochondrial respiratory complexes [11]. In mammals, the NF-Y regulome is far more complex, yet functional dissection of individual promoters suggested the importance of NF-Y for high level expression of metabolic genes: available genomic data and gene expression experiments after inactivation of NF-Y subunits confirmed this [12]. Specifically, following NF-Y removal, expression of anabolic or catabolic genes was found to be reduced or increased, respectively; among the formers, rate-limiting steps in amino acid (Ala, Asp, Glu; Ser, Gly; Gln), lipid (cholesterol and fatty acids) and nucleic acid pathways. Furthermore, as for carbohydrate, carbon metabolism (mostly glycolysis) is almost entirely under NF-Y control. All these metabolic pathways are particularly crucial in cancer cells, where "metabolic reprogramming" is a hallmark of tumor development and progression [13–15]. Analysis of expression profiling of large tumor datasets indicate that the NF-Y matrix is enriched in promoters of genes overexpressed in cancer cells [16]. Recently, we, and others, have also reported the overexpression of the NF-YA subunit in different types of tumors and that this correlates with poor disease prognosis [17–21]. It should be noted that while the trimer is found in all growing transformed or immortalized cell lines, cells in specific tissues, particularly post-mitotic ones, lack or contain very low levels of NF-YA. In general, overexpression of oncogenic TFs not only leads to profound and persistent changes in gene expression, but also to the "addiction" of the tumor cell for high TF gene expression level. In such context, a decrease, even if partial, in TF activity, which would normally be marginal in normal cells, could lead to disproportionately higher effects in tumor cells. Based on these premises, NF-Y has been listed among TFs whose targeting could restrict uncontrolled cell growth.

In general, it is known the poor "druggability" of TFs, which lack ligand pockets and act by means of protein–protein and protein–DNA interactions. Yet, attempts to target TFs have been tried with some success [22,23], including employing unbiased screenings without *a priori* knowledge of protein structures [24]. In the absence of the structure of the NF-Y trimer, screening for compounds that inhibit proliferation by targeting the NF-Y/CCAAT complex has mainly focused on minor-groove binding drugs, typically polyamide intercalating derivatives that bind at the preferred NF-Y binding site (i.e., on CCAAT-boxes of the Topoisomerase α promoter) [25–30]. More recently, the molecular mechanism of DNA-recognition and binding by NF-Y was revealed at the atomic level *via* X-ray crystallography [31–33]. All three NF-Y subunits were shown to be necessary for DNA binding, covering different roles. The NF-YB and NF-YC subunits dimerize through their HFDs and bind the DNA non-specifically over a long sequence (about 25–30bp). The NF-YA subunit, once associated to the HFD dimer with its A1 α-helix, recognizes and binds the CCAAT nucleotides *via* its A2 α-helix and the following Gly-loop, inserting in the minor groove of DNA. These notions

represent the basis for more knowledge-based approaches for targeting the subunits and thus NF-Y activity. In fact, recent studies describing the use of peptides that mimic the A1 α-helix of NF-YA, showed their ability to prevent trimer association and therefore CCAAT binding [34].

Here, we present a detailed study of the molecular bases that underlie the possible modulation of DNA binding activity of NF-Y by an already known drug, suramin. Suramin was identified from a large library of pharmacologically active compounds, *via in silico* docking-simulations carried out on the NF-Y structure. The water solubility of the compound allowed us to test its inhibitory potential without using DMSO, which induces precipitation of NF-Y. Using a combination of biochemical and biophysical approaches, we showed that suramin inhibits the NF-Y/DNA interaction by binding to the HFD of the NF-YB/NF-YC subunits. This demonstrates that NF-Y presents at least one ligandable surface, thus creating the starting point for the rational design of new antiproliferative compounds.

2. Materials and Methods

2.1. In Silico Search for NF-Y Inhibitors

The virtual Library of Pharmacologically Active Compounds (LOPAC®1280) employed for the docking analysis was provided by Sigma-Aldrich and included 1280 commercially available compounds (https://www.sigmaaldrich.com). The AutoDock4 software package [35] was used for a docking screen of the LOPAC®1280 library. The Python Molecule Viewer 1.5.6 of the MGL-tools package (https://ccsb.scripps.edu/mgltools/) was used to analyze the data. The atomic coordinates of NF-Y in complex with DNA (PDB ID 4AWL) [31] were used for docking; both the DNA and the NF-YA subunit were removed prior to *in silico* screening. Hydrogen atoms and Kollman charges were added using the program AutoDock4. The protein model was then used to build a discrete grid within a box (58 × 94 × 68 grid points, with a spacing of 0.375Å) as the explored volume for the compound docking search. The grid was centered on the DNA binding site and, alternatively, on the NF-YA binding site. Fifty independent genetic algorithm runs were performed for each LOPAC library compound (with 150 individuals in the population and 27,000 generations). The docking poses produced were ranked based on the predicted binding free-energy values ΔG (kcal/mol).

2.2. Protein Expression and Purification

The recombinant protein constructs for the expression of the minimal functional domains (md) of the NF-Y HFD dimer (YB/YCmd, hereafter NF-Yd), and of the NF-Y trimer (NF-Ymd, hereafter NF-Yt), which constitute the minimal regions for subunit interaction and DNA binding (NF-YB, aa 49–141; NF-YC, aa 27–120; NF-YA, aa 262–332-long subunit numbering-), and of NF-YA C-terminal portion (YA3; aa 239–347), were previously described [31,36]. Proteins were produced in BL21 (DE3) *E. coli* cells, exploiting a subunit coexpression system strategy [37] for NF-Yd and NF-Yt, and purified as previously described [31,36,38]. Briefly, soluble expression of NF-Yt and NF-Yd was achieved upon induction with 0.2 mM IPTG, incubating overnight at 25 °C. Cells were lysed by sonication in Buffer A (10 mM Tris-HCl pH 8, 400 mM NaCl, 2 mM $MgCl_2$, and 2 mM imidazole). The cell lysate was loaded on a His-Select Nickel affinity column (Sigma Aldrich, St Louis, MO, USA), and proteins were purified, exploiting the presence of the 6His-tag at NF-YA C-terminus in NF-Yt, and at NF-YB N-terminus in NF-Yd. The His-tag was removed from the target proteins by incubating pooled, peak fractions (proteins elute in Buffer A + 250 mM imidazole) with thrombin (Thrombin CleanCleave kit, Sigma Aldrich, St Louis, MO, USA), overnight at 20 °C. Cleaved proteins were further purified on a HiLoad®16/60 Superdex®75 prep grade size-exclusion column

(GE Healthcare, Uppsala, Sweden) pre-equilibrated in Buffer B (10 mM Tris-HCl pH 8, 400 mM NaCl, 2 mM DTT) using an Akta chromatography system (GE Healthcare, Uppsala, Sweden). NF-YA, expressed in *E. coli* BL21 (DE3) with a C-terminal His-tag, was purified by Nickel affinity chromatography, as previously described [36]. Analytical size exclusion chromatography studies of NF-Yd in presence of different concentrations of suramin were performed on a Superdex 75 10/300 GL column (GE Healthcare, Uppsala, Sweden) equilibrated in buffer C (10 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM DTT).

2.3. Electrophoretic Mobility Shift Assays (EMSA)

For EMSA experiments, recombinant proteins were added to a Binding Mix in 16-μL reaction volume, containing a 31bp Cy5-labelled DNA probe derived from the human HSP70 promoter [31]. The final composition of the binding reaction was: 40 nM protein, 20 nM HSP70 probe, 20 mM Tris·Cl pH 7.5, 50 mM NaCl, 5 mM $MgCl_2$, 0.5 mM EDTA, 6.5% glycerol, 2.5 mM DTT, 0.1 mg/mL BSA. Suramin (Sigma-Aldrich, St Louis, MO, USA—cat no. S2671) was included at indicated concentrations in the binding mix (containing NF-Yd in the case of reconstituted trimer reactions), before the addition of recombinant proteins (NF-YA, or NF-Yt), or nuclear extracts. Reactions were assembled on ice and incubated at 30 °C for 30 min in the dark. An aliquot of each reaction was loaded on a 6% non-denaturing polyacrylamide gel and run in 0.25× TBE at 80 V at 4 °C. EMSAs with HeLa cell nuclear extracts were performed essentially as previously described [36]: to obtain nuclear extracts for NF-Y overexpression and control samples, HeLa cells were grown at 37 °C in DMEM high glucose with L-glutamine and 10% FBS (EuroClone, Pero, MI, Italy) and seeded in 6-cm plates; the next day, cells were co-transfected with 350 ng of each full-length NF-Y subunit expression vector (pSG5-NF-YA, pSG5-NF-YB, pCMV2-flag-NF-YC), or empty control plasmid, to a total of 2.3 μg DNA. After 24 h, cells were harvested for nuclear extract preparation as described in [39]. Nuclear extracts from transfected and control cells were used in EMSA as described in [36]. For EMSAs shown in Figure 1b,c, suramin inhibition of DNA binding was assayed in three independent experiments, using the same preparation of nuclear extracts for Figure 1c.

2.4. Isothermal Titration Calorimetry (ITC)

Experiments were performed at 25 °C using a VP-ITC MicroCalorimeter (MicroCal, Malvern Instruments Ltd. Malvern UK) following the general procedure, as previously described [40]. Briefly, the volume of the sample cell was 1.4 mL; the reference cell contained water. Suramin (250 μM) was titrated using injection volumes of 8 μL into a solution containing the required protein at 20 μM. Both protein and suramin were diluted with the same buffer to obtain a final solution in Buffer C (10 mM Tris-HClpH 8, and 150 mM NaCl). Calorimetric data were analyzed with the software packages NITPIC, SEDPHAT, and GUSSI [41].

2.5. Saturation-Transfer Difference (STD) NMR

NMR spectra were acquired on a Bruker AVANCE III 600 MHz NMR spectrometer equipped with a QCI (^{1}H, ^{13}C, $^{15}N/^{31}P$, and ^{2}H lock) cryogenic probe. Samples for STD NMR experiments were prepared as follows: the stock solution of protein (NF-YB/NF-YC or trimer in 10 mM Tris-DCl and 0.4 M NaCl, pH 8) was diluted to 25 μM in the NMR sample and the stock solution of suramin (5 mM) in D_2O was diluted to 1 mM in the NMR sample. The final concentration of the buffer was brought to 10 mM Tris-DCl and 150 mM NaCl in the NMR sample, changing gradually the ionic force over 30 min and keeping the sample at 4 °C. Samples containing the small molecule (1 mM) in 10 mM Tris-DCl and 0.4 M NaCl, pH 8, were also prepared to record the corresponding 1H-NMR and STD NMR blank experiments. The total sample volumes were 560 μL. The pH of each sample

was measured with a microelectrode (Mettler Toledo, Columbus, OH, USA) for 5 mm NMR tubes and adjusted to pH 8 with small amounts (few microliters) of NaOD and/or DCl. All pH values were corrected for the isotope effect. The acquisition temperature was 25 °C. ^{1}H NMR spectra were recorded (*zgesgp* pulse sequences in Bruker library) with 64 scans, with a spectral width of 12 ppm, and a relaxation delay of 3 s. 1D STD NMR spectra were recorded (*stddiffesgp*.3 pulse sequences in Bruker library) with 1024 scans, with a spectral width of 12 ppm, and saturation times of 3 s, 2 s, 1 s, 0.6 s, 0.3 s, with on-resonance frequency = −1.0 ppm and off-resonance frequency = 40 ppm. They were processed with a line broadening of 0.2 Hz and corrected for phase and baseline.

2.6. Crystallization, Data Collection, Structure Determination and Refinement

Suramin sodium salt (catalogue No. S2671) was obtained from Sigma–Aldrich (St Louis, MO, USA). After overnight incubation at 4 °C of NF-Yd with a tenfold molar excess of suramin in Buffer B, crystals of the complex NF-Yd/suramin were prepared at 20 °C in 200 mM ammonium citrate tribasic, and 20% *(w/v)* polyethylene glycol 3350, using the sitting-drop technique. Crystals were cryoprotected in the same reservoir solution supplemented with 20% *(w/v)* glycerol before cooling in liquid nitrogen. X-ray diffraction data were collected at the ESRF synchrotron (ID29 beamline, Grenoble, France). The crystal diffracted at 2.7 Å with the $P2_12_12_1$ space group, with two molecules of NF-Yd and one molecule of suramin per asymmetric unit. Raw data were processed with XDS [42] and Scala [43]. The structure was solved by molecular replacement using the software Phaser [44], and the NF-YB/NF-YC dimer as the search model (PDB-code 4CSR). Iterative cycles of model building with Coot [45] and refinement with Refmac5 and Phenix [46,47] were carried out to produce the final model. The stereochemical parameters of the final model were checked with Molprobity [48]. Data processing and refinement statistics are summarized in Table 1. Atomic coordinates and the structure factors have been deposited in the Protein Data Bank (www.rcsb.org) with entry code 7AH8.

Table 1. Data collection and refinement statistics.

Data Collection	
Space group	$P2_12_12_1$
Cell dimensions	
a, *b*, *c* (Å)	45.697, 61.213, 123.533
α, β, γ (°)	90.00, 90.00, 90.00
Resolution (Å)	45.7–2.7 (2.83–2.70) *
Unique reflections	10061 (1299)
*R*merge (%)	0.11 (0.75)
$I/\sigma(I)$	16.2 (3.5)
Multiplicity	12.1 (12.8)
Completeness (%)	100 (100)
Refinement	
R_{work}/R_{free} (%)	22.2/27.4
No. residues/molecules	
NF-YB	88 (A chain); 89 (C chain)
NF-YC	79 (B chain); 80(D chain)
Suramin	1
Glycerol	1
Citrate	1
Water	47
B-factors (Å^2)	55.1
R.m.s. deviations	
Bond lengths (Å)	0.009
Bond angles (°)	1.41
Ramachandran statistics	
allowed region (%)	98.1
favorably allowed region (%)	1.9
outliers	0

* Highest resolution shell is shown in parenthesis.

2.7. Molecular Dynamics (MD) Simulations

NF-Yd was described by the Amber99SB force-field and solvated in ~23,000 TIP3P water molecules [49]. Suramin was parametrized using GAFF2 [50]. Partial charges were derived by a density functional theory B3LYP all-electron calculation on a 6-31G** basis set by RESP. Density functional calculations have been performed using CP2K [51] while MD simulations were performed using GROMACS 2016 [52]. The initial conformation of the system was taken from the symmetrical dimer found in the crystal and solvated in a dodecahedron box of 735 nm^3. Short-range Coulomb and van der Waals interactions were cut-off at 0.9 nm with long-range Coulomb interactions treated using Particle Mesh Ewald. After energy minimization, the temperature and density of the system were equilibrated keeping the NF-Yd/suramin position fixed for 1 ns. A production 1.2 μs simulation was run at 300K and 1 atm using the Bussi's thermostat [53] and the Parrinello-Rahman barostat [54].

3. Results

3.1. Identifying Suramin as a Compound Binding to NF-Y

Virtual screening was initially set up to discover synthetic compounds that interfere with NF-Y activity. Docking simulations were carried out using the NF-YB/NF-YC HFD dimer (NF-Yd: derived from the 4AWL PDB structure) as a receptor. The volume target for binding included the NF-Y trimerization interface, a mostly negatively charged groove of NF-Yd, or the DNA-binding surface, a wide mostly positively charged region (Supplementary Figure S1). These regions were explored using a library of small molecules as described in Materials and Methods. The docking search produced a list of compounds with predicted binding free-energy values (ΔG) up to −12.3 kcal/mol. Among the ten top ranking compounds, we noticed that suramin (ΔG = −12.1 kcal/mol; Figure 1a) was the only compound that is water-soluble, while the other ligands were soluble only in DMSO. Dynamic Light Scattering experiments indicated that DMSO (1–10%) induces high polydispersity (>25%) and non-specific aggregation of NF-Yd (1 mg/mL) (data not shown), thus precluding any reliable biophysical characterization of the NF-Y-binding activity for DMSO-soluble compounds. For this reason, we focused our attention on suramin, which was assayed in its ability to inhibit NF-Y function in EMSA.

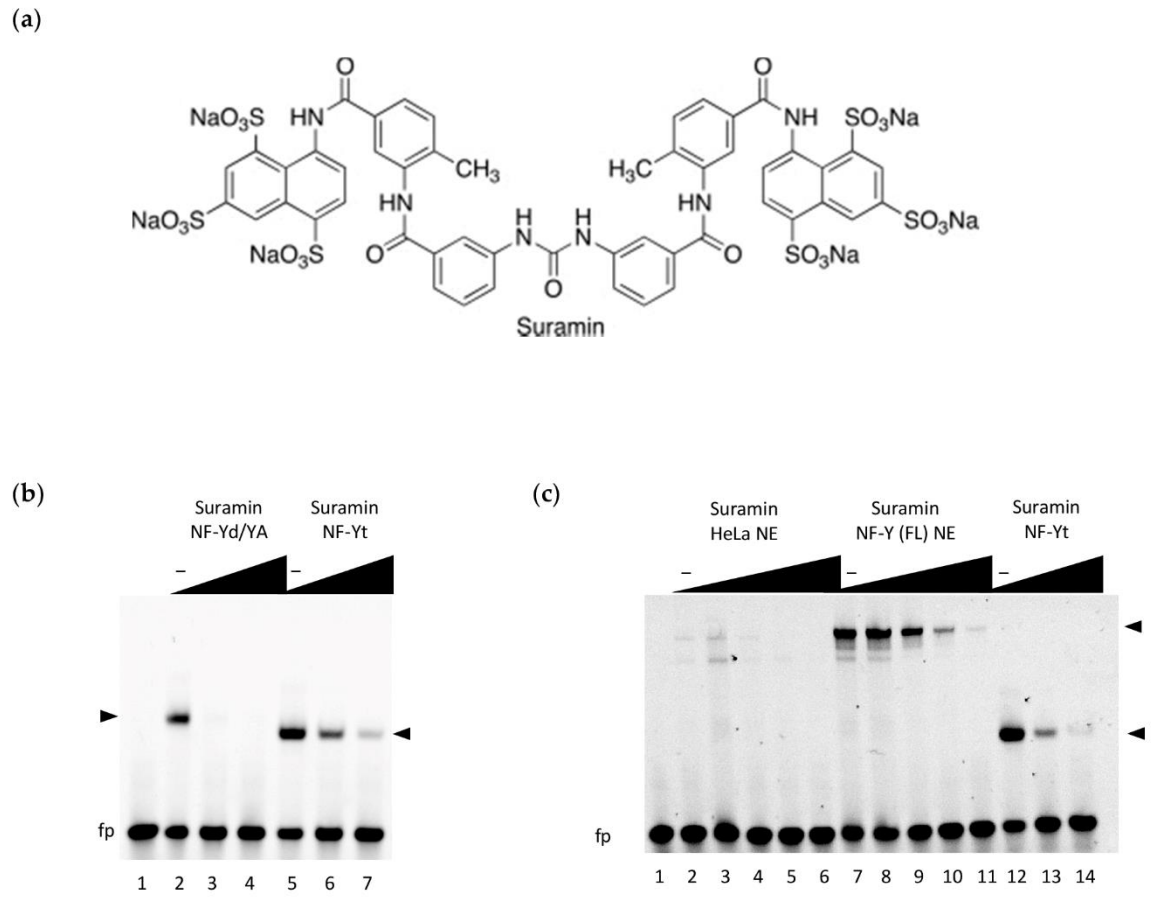


Figure 1. DNA-binding inhibition by suramin; (**a**) chemical structure of suramin; (**b**) DNA binding inhibition of the NF-Y trimer (40 nM) by suramin was assessed by electrophoretic mobility shift assays (EMSA) using a Hsp70 CCAAT box DNA probe (20 nM). Inhibition was tested at increasing doses of suramin (0, 50, 100 μM; lanes 2–4 and 5–7) against the reconstituted trimer, obtained by mixing equimolar ratios of purified NF-YA with NF-Y histone fold domain (HFD) dimer (NF-Yd) (NF-Yd/YA) or on the co-purified trimeric subunit protein (NF-Yt). Lane 1: probe alone DNA binding mix in the absence of NF-Y. NF-Y/DNA complexes are indicated by arrowheads. fp: free probe. Slower migration of NF-Yd/YA/DNA, as compared to NF-Yt/DNA (composed of the minimal DNA-binding domains), reflects the higher molecular weight of the purified NF-YA subunit within the complex; (**c**) EMSA experiments using nuclear extracts (NE) from HeLa cells, obtained from control cells (HeLa NE: lanes 2–6) or from cells overexpressing the full-length NF-Y subunits (including the transactivation domains) (NF-Y(FL) NE: lanes 7–11). NF-Yt recombinant protein was used as a positive control for suramin inhibition of DNA binding (lanes 12–14). In lanes 2–11, the reaction mix includes 2.3 μg of nuclear extract, and an increasing concentration of suramin for NE and NE + NF-Y(FL) (0, 1, 50, 100, and 200 μM), lanes 12-14 reactions include NF-Yt protein (40 nM) and suramin (0, 100, and 200 μM). Lane 1: probe alone DNA binding mix without NF-Y subunits or NE added. The migration of FL and minimal domain NF-Y/DNA complexes are indicated by black arrowheads. "fp": free probe, "–": no suramin.

3.2. Inhibition of NF-Y DNA-Binding by Suramin

To evaluate whether the suramin can inhibit the specific binding of NF-Y to DNA, we firstly performed EMSA experiments with purified NF-Y recombinant proteins, using a fluorescently

labeled high-affinity CCAAT box probe derived from the human HSP70 promoter [31]. EMSA binding reactions (Figure 1b) were assembled in the presence of increasing concentrations of suramin, either using the reconstituted NF-Y trimer (NFYd/YA), obtained by combining equimolar ratios of the individually purified NF-YA and the HFD co-expressed dimer (NF-Yd) proteins, or with the co-expressed minimal NF-Y DNA binding domain trimer (NF-Yt). We observed that in both cases, suramin addition substantially decreased the formation of NF-Y-bound DNA complexes, in a dose-dependent manner. Suramin's action is more evident when the combined NF-YA and NF-Yd proteins are assayed, as compared to the pre-formed NF-Yt, suggesting that the C-terminal region of NF-YA (not involved in trimerization and demonstrated to be flexible in the absence of DNA [32]) might partly interfere with suramin binding before repositioning itself at the CCAAT for DNA binding, thus decreasing the effective concentration of the inhibited protein. Considering that the tested constructs are composed of the minimal DNA-binding homology regions of NF-Y, in order to ascertain whether suramin functional inhibition could also hold true on full-length (FL) native proteins expressed in mammalian cells, we performed further EMSAs in which the recombinant proteins were replaced with nuclear extracts. To obtain a significant (specific) signal of NF-Y-bound DNA complexes, HeLa cells were transfected with NF-Y subunit vectors (NF-Y(FL) NE) to obtain extracts that overexpress the three full-length subunit proteins. As a comparison, the NF-Yt recombinant protein was assayed in parallel, together with control (untransfected cells) nuclear extracts (HeLa NE). The results demonstrate that, in the presence of all nuclear components of the mammalian cell (HeLa) extracts, suramin addition efficiently interferes with DNA binding of the native NF-Y FL protein, similarly to NF-Yt (Figure 1c). Importantly, suramin can modulate the binding of NF-Y to DNA at similar concentration ranges as observed for the recombinant protein.

3.3. Interaction Between NF-Yd and Suramin

In order to quantify the affinity of suramin for NF-Y, we used isothermal titration calorimetry (ITC). Figure 2a reports a representative ITC experiment in which suramin was injected into a sample cell containing NF-Yd. The thermogram exhibited a biphasic behavior, indicating the occurrence of at least two binding events. Since it is known that suramin can induce the dimerization of its receptors [55,56], in order to determine the appropriate binding model for fitting ITC data we analyzed the oligomerization state of NF-Yd as a function of the suramin relative concentration by gel filtration (GF) (Figure 2b). In the presence of equimolar concentration or low excess of suramin (fivefold more than the protein), the GF chromatograms displayed a second elution peak, compatible with the formation of a NF-Yd protein dimer. This peak was not present at high suramin molar excess (100-fold more than the protein). Therefore, it appears that at low (limiting) concentrations suramin promotes dimerization of NF-Yd, while at high molar ratios suramin saturates all NF-Yd molecules, preventing NF-Yd dimer formation.

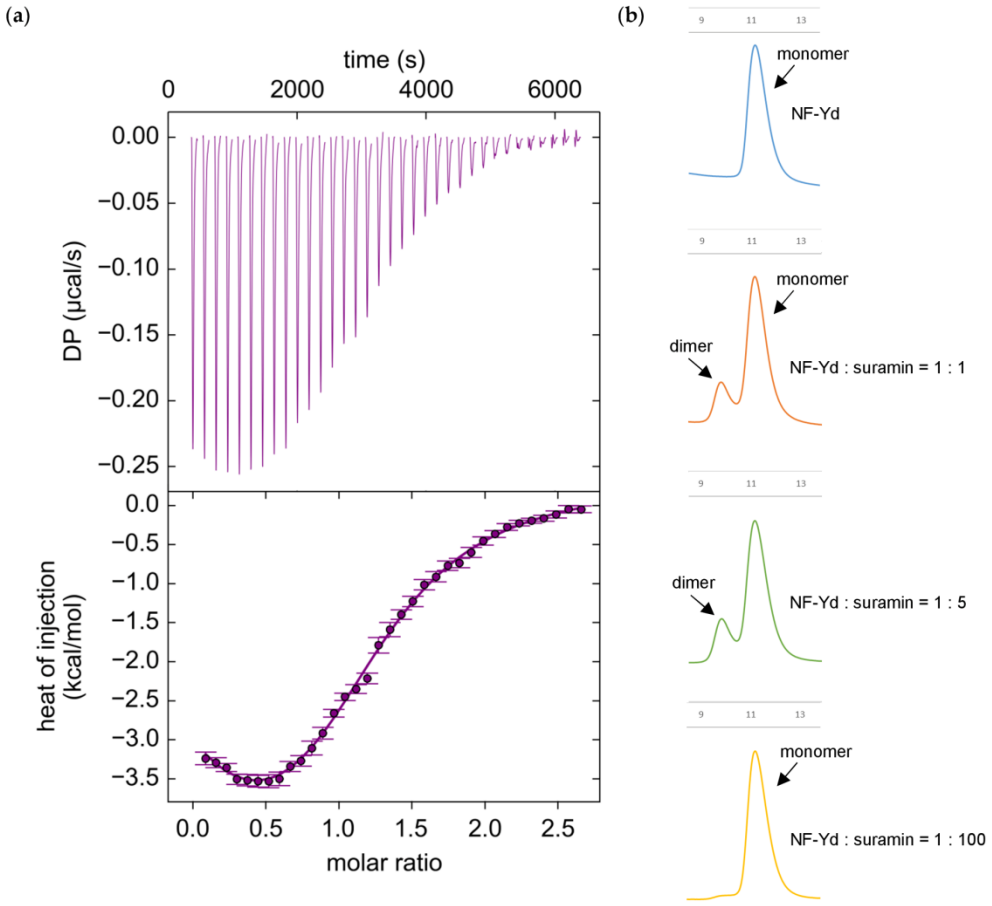


Figure 2. Biophysical analysis of suramin binding to NF-Yd; (**a**) Suramin binding to purified NF-Yd proteins measured by isothermal titration calorimetry (ITC). Upper panel, heat changes (µcal/s) during successive injections of 8 µL of suramin (250 µM) into the chamber containing NF-Yd (20 µM). Lower panel, binding curve obtained from data displayed in the upper panel. The peaks were integrated and normalized per mole of suramin injected. Both panels are plotted against the molar ratio suramin:NF-Yd. The solid line represents a nonlinear least-squares fit to a double dependent sites binding model (see Materials and Methods). The experiment was repeated three times; (**b**) gel-filtration analysis (Superdex 75 10/300 GL column, GE Healthcare, Uppsala, Sweden) of the oligomerization state of NF-Yd induced by different molar ratios of suramin.

Based on these data, we designed a double-dependent binding site model where the binding of suramin to NF-Yd forms a ligand–receptor complex, which in turn generates a binding surface for a second NF-Yd molecule. As suramin concentration increases, the suramin-induced dimerization of NF-Yd is penalized (for details see Material and Methods). The proposed model well fits the ITC data (solid line in Figure 2a) yielding two mean dissociation constants (K_d) values of 2.9 ± 0.7 µM and of 61.5 ± 9.4 µM for the first (NF-Yd + suramin) and the second (NF-Yd-suramin + NF-Yd) binding event, respectively. Both dissociation constants are in the low micromolar range, with suramin–NF-Yd interaction twenty-fold stronger than the suramin-induced NF-Yd dimerization.

Altogether, ITC and GF analyses indicate that suramin binding to NF-Yd has a dimerization side effect on the transcription factor, which is evident only at low suramin concentrations when NF-Yd can bind to both free suramin or to a NF-Yd–suramin complex. In contrast, when each NF-Yd is saturated by suramin, no dimerization between NF-Yd–suramin complexes can occur. This strongly suggests that one suramin can bind to two NF-Yd molecules simultaneously (see also crystallography section).

3.4. STD NMR Binding Experiments

STD NMR spectroscopy [57–59] was employed to characterize the molecular recognition events involving the NF-Yd or NF-Yt and suramin in solution. STD NMR experiments can reveal interactions between a small molecule and a high molecular weight biomolecule, such as a protein or, as in this case, a protein complex, by detecting the transfer of magnetization from the receptor to the ligand that can occur only if both molecular entities bind to each other. Some receptor resonances are selectively saturated and, after binding, the magnetization is transferred from the receptor to the ligand. The detection of ligand NMR signals in the STD spectrum is an unequivocal evidence of its interaction with the receptor. On the contrary, any signal from non-binding compounds is erased in the difference spectrum (STD spectrum). STD NMR spectra (Figure 3b,c) were acquired on samples containing a mixture of the protein (dimer or trimer 25 μM) and suramin (1 mM) dissolved in 10 mM Tris-DCl, pH 8, 150 mM NaCl.

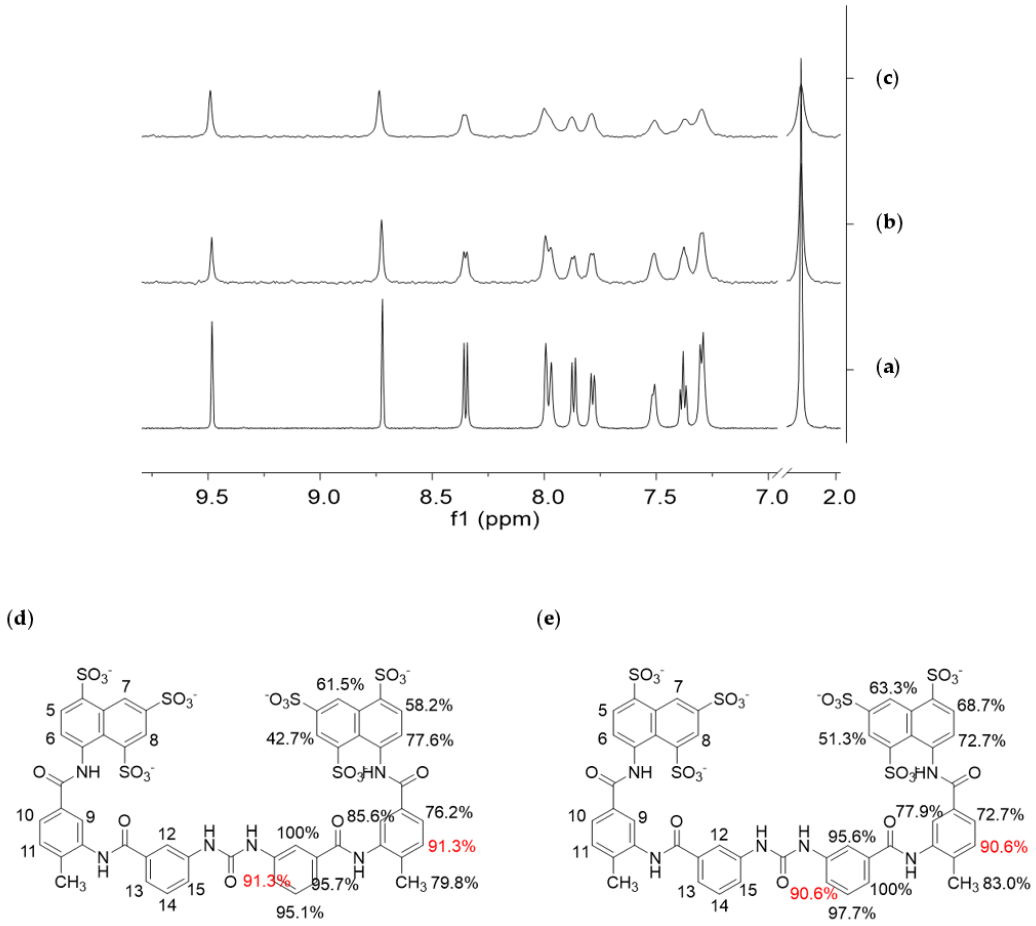


Figure 3. Saturation-Transfer Difference (STD) Nuclear Magnetic Resonance (NMR) experiments on suramin–NF-Y complexes; (**a**) ^{1}H-NMR spectrum of suramin 1 mM; (**b**) STD NMR spectrum of a mixture of suramin (1 mM) and NF-Yd (25 μm); (**c**) STD NMR spectrum of a mixture of suramin (1 mM) and NF-Yt (25 μm). All samples were dissolved in 10 mM Tris-DCl, 150 mM NaCl, pH 8, and analyzed at 25 °C and 600 MHz. The ^{1}H-NMR spectra were recorded with 64 scans and the STD NMR spectra with 1024 scans and 3-s saturation times; (**d**) binding epitopes of suramin to NF-Yd, and (**e**) trimer obtained for 0.6-s saturation time. The largest relative STD intensity was scaled to 100%. Values are referred to half molecule and are intended duplicated symmetrically. Values in red refer to protons whose overlapping prevented the discrimination of their single contribution.

Selective saturation of some aliphatic protons of the protein was achieved by irradiating at −1.00 ppm (on-resonance frequency), a spectral region where no resonances of the putative ligand are present. Figure 3a shows the ^{1}H NMR spectrum of the ligand, used as a reference for its characterization, and Figures 3b and 3c report the STD NMR spectra recorded on the different protein/ligand mixture. The presence of suramin resonances in the STD spectra demonstrates the interaction of the compound with both the dimer and the trimer.

From a qualitative point of view, the stronger intensity of a ligand signal in the STD NMR spectrum indicates shorter inter-proton distances between that ligand proton and the receptor surface in the bound state, allowing the identification of the portion of a ligand most involved in the interaction with a receptor (epitope mapping). Here, STD experiments were acquired with five different saturation times (0.3, 0.6, 1.0, 2.0, and 3.0 s) (data not shown), to obtain the relative STD intensity for each proton of the ligand that gives STD signal. Thus, from these data, the ligand-binding epitopes were characterized, showing to be essentially similar/superimposable for the NF-Yd and NF-Yt (Figure 3b).

3.5. NF-Yd–Suramin Complex

To shed light on the molecular mechanisms of NF-Y inhibition described above, we undertook crystallographic analyses of the TF in complex with suramin. The NF-Yd–suramin co-crystals belonged to space group $P2_12_12_1$ and diffracted to 2.7 Å resolution. The refinement converged to an R_{factor} value of 22.2% (R_{free} = 27.4%), with a final model composed of two NF-Yd protomers (identified with A/B, and C/D chains for NF-YB/NF-YC, respectively), one suramin molecule, one glycerol, one citrate, and 47 water molecules in the asymmetric unit. After initial refinement cycles of the protein structure, residual electron density was visible for the full molecule of the bound suramin that was modeled accordingly. The refinement statistics and other information are provided in Table 1.

The suramin molecule adopts an elongated conformation in the $(NF\text{-}Yd)_2$–suramin complex and binds at the interface between two NF-Yd protomers, in an almost symmetric manner (Figure 4a,b).

Half of the ligand molecule establishes interactions in a cleft lined by residues from the NF-YC B and D chains, and from the NF-YB A chain. In detail, the half suramin-binding site is lined by the NF-YC N-terminus (Gln(D)41), helix α2 (Glu(D)64, and Pro(B)66), helix α1 (Leu(B)45, Lys(B)49, and Lys(B)53), and loop L1 (Lys(B) 59, Met(B)60, Ile(B)61, Ser(B)62, and Ala(D)63), and by the NF-YB C-terminal residues (Phe(C)139, and Arg(C)140) (Figures 5a,c and 6).

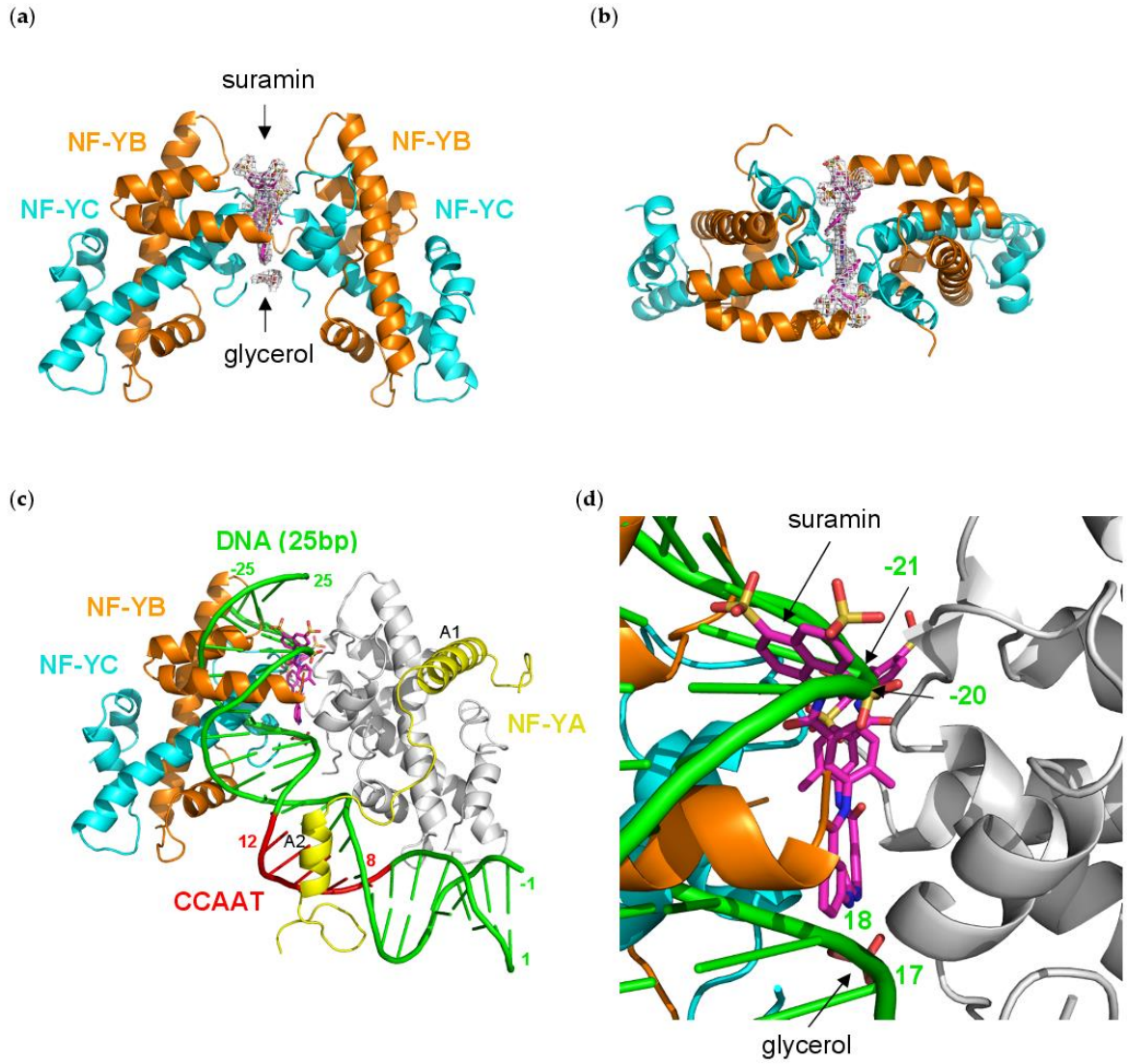


Figure 4. Structure of the $(NF\text{-}Yd)_2$–suramin complex; (**a**) ribbon diagram showing the bound suramin (magenta sticks) at the dimerization interface of two NF-Yd molecules (NF-YB in orange, and NF-YC in cyan), together with a glycerol molecule (from the cryoprotectant solution). A representative electron density map (grey net), contoured at 1.0 σ, is shown around the bound molecules; (**b**) top view of $(NF\text{-}Yd)_2$–suramin complex; (**c**) Structural superposition of one NF-Yd molecule of the $(NF\text{-}Yd)_2$–suramin complex with the DNA/NF-Y complex (NF-YB/NF-YC in grey, NF-YA in yellow, DNA in green) (PDB-code 4AWL). The CCAAT box is shown in red. The NF-YA A1 and A2 α-helices are labeled, and the DNA numbering is indicated. About 10bp of the bound DNA superimpose with the second NF-Yd molecule of the $(NF\text{-}Yd)_2$–suramin complex; (**d**) close-up of panel (**c**) showing that two sulfonic acid groups of suramin and the glycerol molecule approximately match the positions of two phosphate groups in both DNA strands of the DNA/NF-Y complex (−20, −21, and 17, 18, respectively).

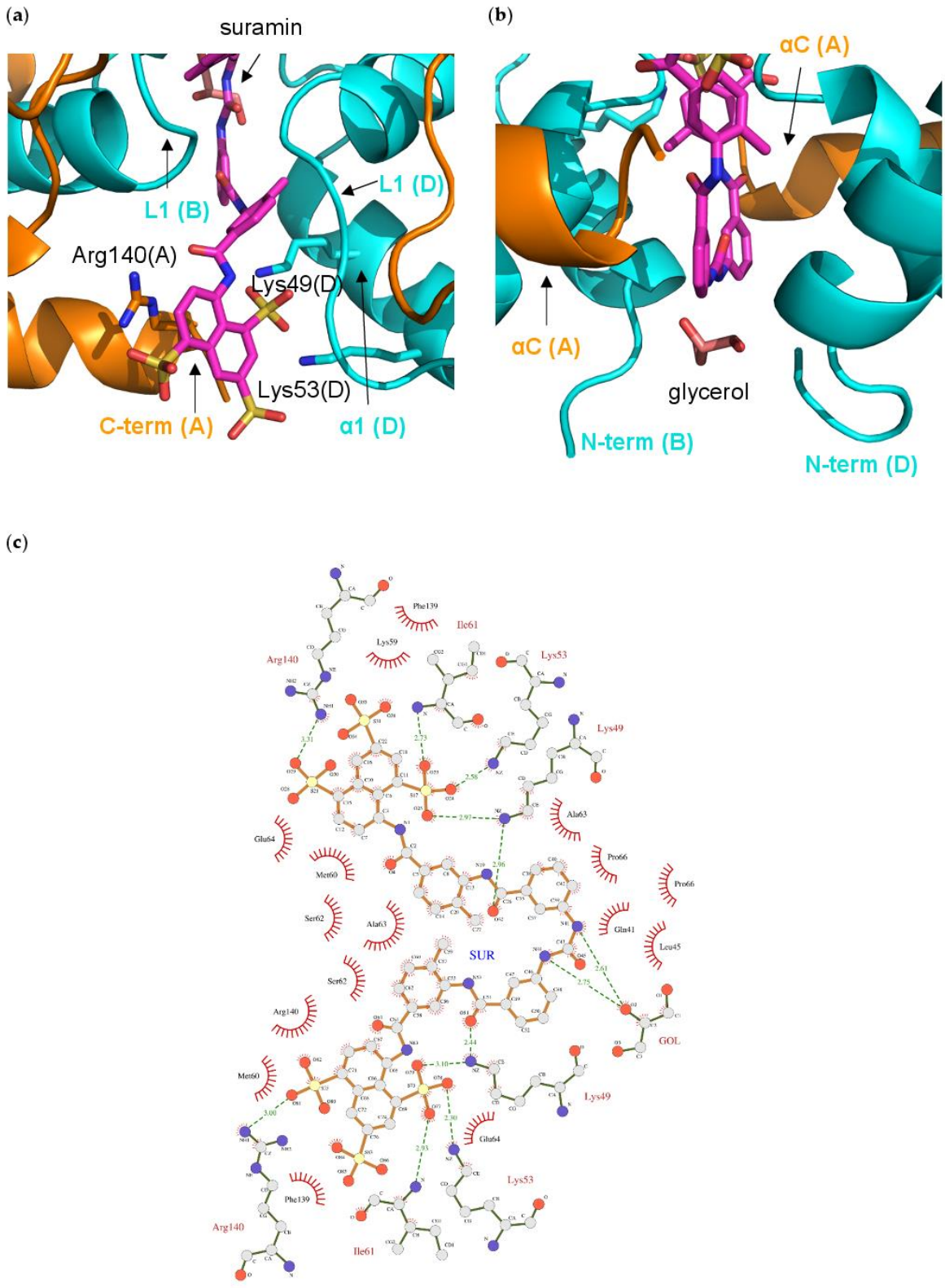


Figure 5. Interactions of suramin in the NF-Yd structure; (**a**) close-up of the binding pocket of half-suramin at the NF-Yd dimerization interface. Secondary structure elements and protein chains are indicated. Color coding as in Figure 4; (**b**) glycerol-binding pocket relative to the suramin binding site. The glycerol molecule is shown in pink sticks; (**c**) schematic representation with LIGPLOT [60]. Polar contacts are depicted as broken lines and hydrophobic contacts are indicated by arcs with radiating spokes.

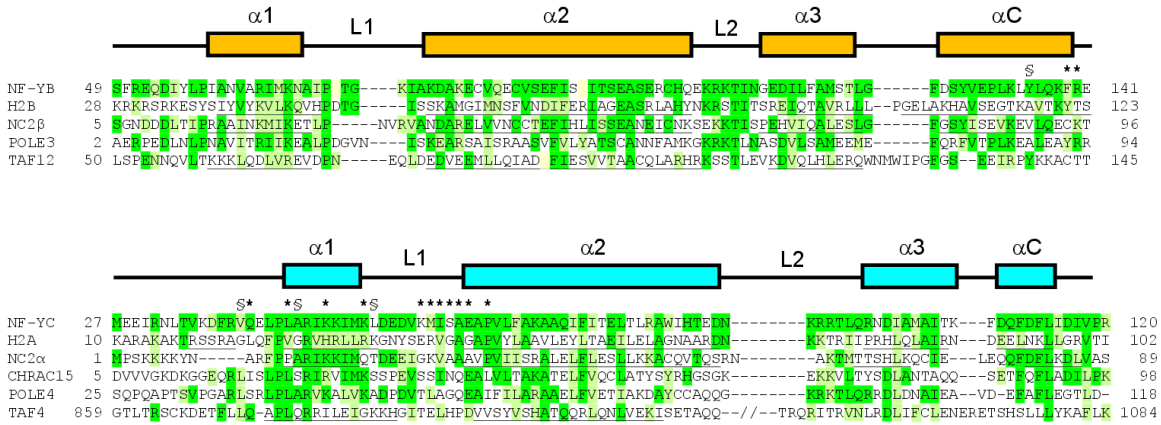


Figure 6. Sequence alignment of NF-YB and NF-YC with HFD proteins; the NF-YB sequence is aligned with human proteins H2B (sp|P06899|; PDB-code 4AFA), NC2β (sp|Q01658|; PDB-code 1JFI), POLE3 (sp|Q9NRF9|), and TAF12 (sp|Q16514|; PDB-code 1H3O). The NF-YC sequence is aligned with human proteins H2A (sp|P04908|; PDB-code 4AFA), NC2α (sp|Q14919|; PDB-code 1JFI), CHRAC15 (sp|Q9NRG0|), POLE4 (sp|Q9NR33|), and TAF4 (sp|O00268|; PDB-code 1H3O). The secondary structure arrangement of the HFD of NF-YB and NF-YC is shown above the sequences. When available, secondary structure information for the other aligned proteins was included (helical residues underlined). In TAF4, the predicted α3 region of the sequence is also aligned, separated by two slashes indicating the ~100 aa loop. Identical residues are highlighted in green, similar residues in light green. NF-Yd residues involved in suramin and glycerol binding are indicated by * and §, respectively.

In particular, the charges of two out of the three sulfonates are compensated by Arg(C)140, and Lys(B)49/Lys(B)53, while the third $-SO^{3-}$ points toward the solvent. Lys(B)49 also interacts with the carbonyl oxygen of the carbonylimino group linking the naphthalene ring and the benzene ring of suramin (Figure 5c). The other half of the suramin molecule binds to the corresponding protein regions of the A, B, and D chains, respectively. Overall, the suramin molecule fits well at the dimeric interface of the two NF-Yd molecules. A further suramin interaction is provided by a glycerol molecule from the cryoprotectant solution, which is symmetrically hydrogen-bonded to both the NH groups of the functional urea moiety (Figures 4a and 5b,c).

Structural comparisons made between the $(NF\text{-}Yd)_2$–suramin complex and the NF-Yd structure (PDB-code 4CSR) show that, besides promoting the dimeric assembly of the protein, binding of suramin is not associated with any significant tertiary structure modifications (rmsd 0.75 Å). However, local residue adaptations to host the ligand are evident in the C-terminal region of NF-YB (residues 135–141), and in the L1-loop of NF-YC (residues 59–64). In particular, the NF-YB Phe139 side chain rotates by about 90° to accommodate the two benzene rings attached to the urea moiety of the ligand (Supplementary Figure S2).

Structural superimposition with the NF-Y/DNA complex (PDB-code 4AWL) provides hints on the mechanism of DNA-binding inhibition induced by suramin binding (Figure 4c). In the $(NF\text{-}Yd)_2$–suramin complex, the ligand is located close to the DNA binding region of NF-Y, contacting NF-YC L1 and α1 and NF-YB αC (Figure 5a,b), which are responsible for more than 40% of the contacts between the HFD NF-Yd and the phosphate backbone of the NF-Y-bound DNA [31]. Furthermore, two sulfonic acid groups that decorate the suramin naphthalene moiety, in particular the one salt-bridged to the NZ atoms of NF-YC Lys49 and Lys53, match well with the position of two DNA phosphate groups, corresponding to position −21 and −20 in the complementary strand of the CCAAT box (Figure 4d) in the DNA/NF-Y complex [31]. Additionally, the suramin-induced

homodimerization of the HFD NF-Yd sterically precludes the binding of about 10bp present in the NF-Y/DNA complex (16–25 in 4AWL) (Figure 4c).

On the contrary, suramin binding seems not to interfere with NF-Y trimerization, since the interaction surface between NF-YA and the NF-Yd (NF-YB α2, NF-YC L2 and αC) is distant from the suramin-binding site (Figure 4c). These results are in line with our STD NMR data that indicate that suramin binds in a similar manner to both NF-Yd and NF-Yt.

3.6. Molecular Dynamics Simulation

Given that the binding of suramin resulted in the formation of a symmetrical dimer of NF-Yd in the crystal, we probed whether the interaction between suramin and one NF-Yd molecule is sufficient to provide a stable complex or if a sandwich between two NF-Yd protomers is compulsory for suramin binding. To this aim, we performed 1.2 μs MD simulations at 300K. The MD results show that the interaction of suramin with a single copy of the NF-Yd is stable for the full duration of the simulations (Figure 7a,c). The overall binding mode is preserved with the central urea functional moiety of suramin well-accommodated in the complementary surface groove formed by NF-YC α1, and α2. Diversely, some variations are found for the position of the two naphthalene rings that tend to readjust their position to optimize the interactions with the single NF-Yd molecule (Figure 7b). In the absence of the second (heterodimeric) NF-Yd protomer, the NF-YC L1 reorients its position towards the NF-YC α1 and creates a pocket (lined by Lys49, Lys53, Leu54, and Glu56) that hosts the naphthalene ring of the orphan half suramin molecule. The other naphthalene ring moves towards the NF-YC α2 and NF-YB L2 and binds in a surface cleft lined by NF-YC Ser62, Ala63, Glu64, and NF-YB Arg108, Thr110, Asn112, Glu114, and Phe139 and Arg140 at the NF-YB C-terminus. In the MD simulation, two of the suramin sulfate groups mimic the DNA phosphate backbone as found in the DNA/NF-Y trimer complex.

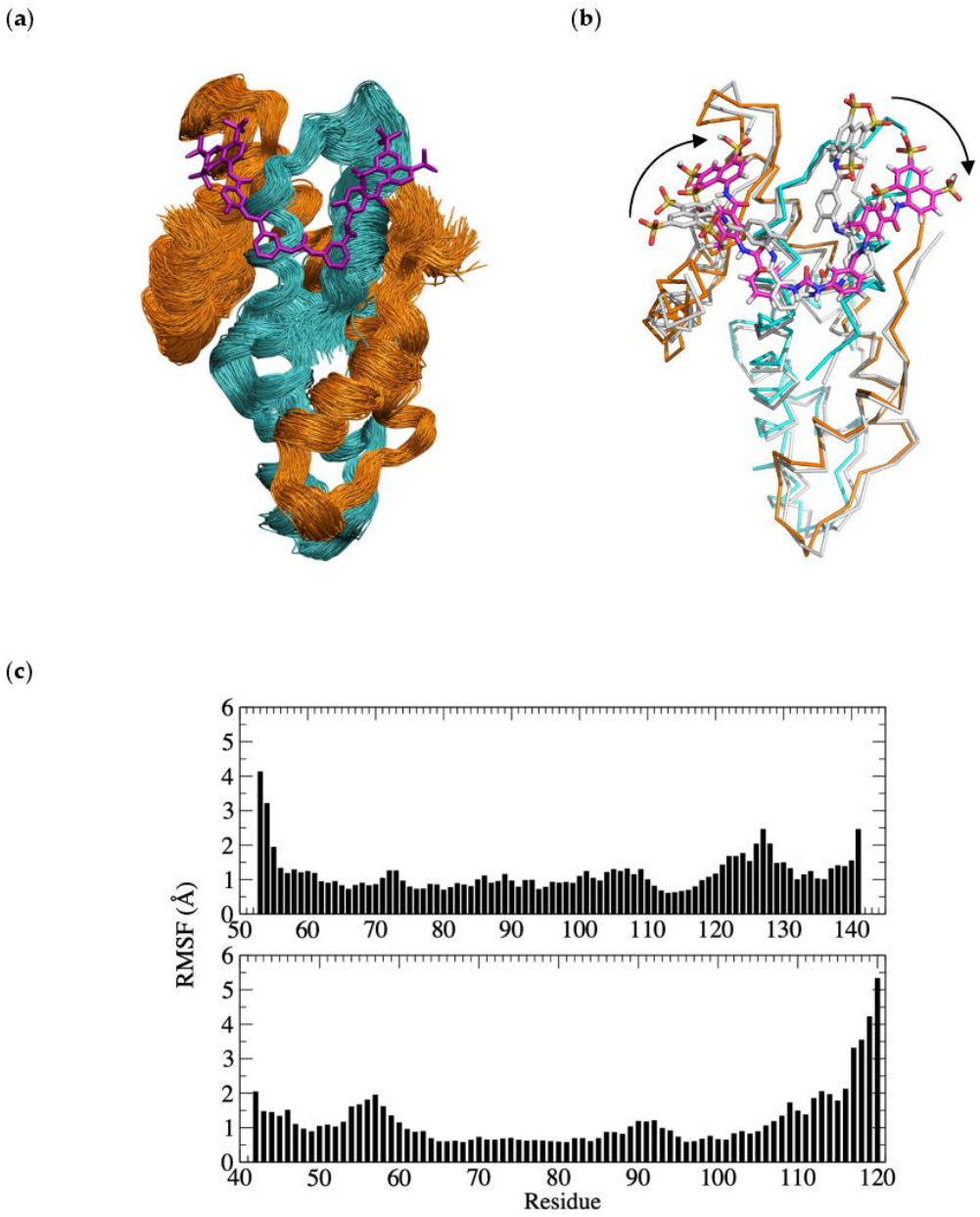


Figure 7. Molecular dynamics (MD) simulations of the NF-Yd–suramin complex; (**a**) superimposed frames (one per ns) for the last μs of simulation (orange NF-YB, cyan NF-YC, and magenta suramin); (**b**) structural superposition of the middle structure of the most populated cluster of the MD simulations (color code as in panel (**a**) with the crystal structure of the complex (grey)). The readjustment of the two naphthalene rings to optimize the interactions with the single NF-Yd molecule are highlighted by arrows; (**c**) root mean square fluctuations (RMSF) per NF-YB and NF-YC residues (upper and lower panels, respectively).

4. Discussion

In this study, we demonstrated that suramin, a polysulfonated naphthylamine derivative of urea, inhibits the DNA interaction of NF-Y with the CCAAT box by binding the HFD subunits, either as isolated heterodimers or in the complete trimer.

The first issue concerns whether NF-Y is a reasonable target for pharmacological anti-cancer intervention. The NF-Y subunits are evolutionarily conserved, from yeast to plants to mammals, as

is the NF-Y binding site, the CCAAT box, which is relatively widespread in human promoters (about 25%). The three NF-Y subunits are expressed widely in many (but not all) tissues. This implies that to target cancer cells, NF-Y drugs may be too "unspecific", or too toxic, as they are likely to also heavily affect normal cells. However, a specific role of NF-Y in cancer progression recently emerged. First, the CCAAT box is selectively present in cell-cycle and metabolic genes. Furthermore, essentially all G2/M genes, whose expression is typically altered in cancer, and genes relating to certain metabolic pathways (lipids, glycolysis, nucleotides, SOCG, and glutamine) show altered expression. This is essentially the reason why bioinformatics analyses of promoters of genes overexpressed in cancers often rank CCAAT at the top of the list; in down-regulated genes, instead, CCAAT has never been found. Second, it is becoming clear that NF-YA is robustly and widely overexpressed in epithelial cancers and this is correlated to a worse tumor prognosis [18–21]. Third, with the possible exception of the hematopoietic system, mice models with conditional KO of NF-YA in different cell types (adipocytes, neurons, hepatocytes) showed chronic, but not abrupt, acute effects [61]. Fourth, metabolic rewiring of cancer cells is associated with, and possibly caused by, overexpression of genes at crucial crossroads in the nucleic acid, amino acid, glucose and lipid metabolic pathways, the vast majority of which are under NF-Y control, based on subunits inactivation and genomic location analysis [12]. Explicative of this is our recent dissection of the glutamine biosynthetic pathway: tumor cells that are sensitive to glutamine starvation become more resistant upon NF-YA stable overexpression entailing up-regulation of CCAAT-dependent genes [62]. In summary, in agreement with other active studies in this field, NF-Y may represent an interesting target for anti-cancer therapy. To date, two directions have been pursued: inhibition of selected sites by DNA-binding drugs [25–30] and, following the 3D structure determination [31], interference of trimerization by modified peptides [34]. A third approach, also possible due to structural knowledge, is described here, based on the NF-Y–suramin interaction.

Pharmacologically, suramin has been used for the treatment of early stages of human sleeping sickness and onchocerciasis, an infectious cause of blindness [63]. It is also an inhibitor of reverse transcriptase of retroviruses [64], and, because of the antiproliferative effects, it is currently considered for use as an anticancer agent and chemosensitizer in cancer therapy [65,66]. The antiproliferative mechanism(s) of suramin is far from understood, as its activity has been linked to inhibition of various, and disparate pathways. (i) Suramin can dissociate receptor-bound growth factors, consequently resulting in loss of signal transduction [67]. (ii) Suramin and its analogues have been shown to inhibit Hpa in many human cancer cell lines [68–71]. (iii) Recently, it has been shown to bind HMGA2 and to potently inhibit DNA interactions, providing new insights into its anti-cancer and anti-metastasis functions, since the expression levels of HMGA proteins are associated with metastasis and poor prognosis for many cancer types [72].

In the case of NF-Y, the binding constants of suramin are modest (K_d in the low μM range) but the structural characterization of binding provides room for a rational interpretation of the process and for its improvement. The crystal structure of the complex reveals that suramin binds to the surface of the HFDs, thus promoting homodimerization through the formation of a $(NF\text{-}YB/NF\text{-}YC)_2$–suramin quaternary structure (Figure 4a,b). The presence of this new $(NF\text{-}Yd)_2$ oligomerization state for HFDs is detectable in solution by gel filtration and is in line with the ITC results, indicating two distinct binding events (Figure 2). Based on our structural evidence, these two events may be attributed to the binding of suramin to the surface of one HFD and to the binding of the second to HFD–suramin, to produce the symmetric "sandwiched" $(NF\text{-}Yd)_2$–suramin homodimer. This interpretation is in agreement with the MD simulations, indicating that the complex between suramin and a single HFD is stable, in addition to gel filtration experiments, showing that when the protein sample is saturated by suramin forming the NF-Yd–suramin complex, HFDs homodimerization was not further sustained, because the $(NF\text{-}Yd\text{-}suramin)_2$ complex is not feasible.

The quaternary structure assembly observed for NF-YB/NF-YC in the presence of suramin is not uncommon, and has already been reported in the literature for ecarpholin S, a Ser49-PLA2 from *E. carinatus* venom [55], for MjTX-II, a myotoxic Lys49-PLA2 from *B. moojeni* [56], and for human NAD^+-Dependent Deacetylase SIRT5 [73]. This suramin-induced oligomerization behavior is related to its symmetrical chemical structure (Figure 1a), that favors symmetrical protein–protein interactions when bound to the target protein surface.

NF-Y represents the first example of an HFD-containing TF that interacts with suramin. The suramin-binding site is located at the HFD surface involved in DNA interactions within the NF-Y/DNA complex (Figure 4c) [31]. On the contrary, the NF-YB/NF-YC region involved in trimerization with NF-YA is located far from the suramin-binding site (Figure 4c) and, accordingly, our STD NMR data show that suramin can bind to the NF-YB/NF-YC heterodimer and to the trimer in a similar way: the suramin binding epitopes are coherent with the binding shown in the crystallographic complex (Figure 3d,e). Within the HFD heterodimer, NF-YC is the subunit that provides the majority of the suramin-interacting residues: located at the first turn of helix α1 and in the L1 region, they are mostly apolar and generate the homodimeric cleft where suramin symmetrically fits (Figures 5a,c and 6). Other residues, i.e., Lys49 and Lys53, provide electrostatic interactions to the suramin sulfonic acid group, matching the position of DNA phosphate groups of the CCAAT complementary strand in the NF-Y/DNA complex (−21 and −20 in 4AWL) (Figure 4d). Additionally, the C-terminus of the NF-YB subunit, in particular Arg140, provides an additional electrostatic interaction to stabilize a second suramin sulfonic acid group (Figure 5a,c). Interestingly, a glycerol molecule is bound to the NH groups of the urea functional moiety of suramin (Figure 5c), with two hydroxyl groups almost matching the position of two phosphate groups of the CCAAT strand in the NF-Y/DNA complex (17 and 18 in 4AWL) (Figure 4d). The pocket hosting the glycerol is solvent-exposed and lined by residues belonging to the N-terminal region of NF-YC (Val40, Gln41, Leu45, and Ala46), but also to the C-terminal region of NF-YB (Tyr135, and Phe139) (Figures 5b and 6). Furthermore, the suramin-induced homodimerization of the HFD heterodimers sterically precludes binding of about 10bp present in the NF-Y/DNA complex (Figure 4c). Our tests on nuclear extracts obtained from HeLa cells overexpressing full-length NF-Y provide the first validation that the presence of other protein domains (including the transactivation domains) of the native protein, or other components present in HeLa cells nuclear fractions, do not substantially affect the suramin inhibitory activity, as it can efficiently modulate NF-Y binding to DNA in the same molar concentrations as observed with the recombinant protein (Figure 1c).

Due to its large, flexible, and multifunctional nature, suramin tends to be a nonselective drug. In the case of NF-Y, an issue may be its interaction with other HFD-containing proteins [74]. Sequence alignment of NF-YB and NF-YC with other human HFD-containing proteins indicates that the NF-Y suramin-binding site is lined by residues that are partly conserved (Figure 6). The conservation is located to NF-YC α1 and α2 and NF-YB L2, while differences are present at the NF-YC N-terminus and L1 and at the NF-YB C-terminus; based on our crystallographic data, these latter regions line the glycerol-binding cleft. Such areas with low sequence homology with other HFD proteins could be used to drive chemical modifications on suramin to generate either symmetrical or asymmetrical compounds that can improve ligand-binding affinity, and/or selectivity.

In conclusion, many proof-of-concept experiments have suggested various TFs as promising therapeutic targets. As cancer cells can become addicted to the activity of specific oncogenic TFs for survival, then their inhibition can lead to the selective killing of cancer cells compared with normal cells. Targeting TFs has traditionally been challenging due to disordered structures and the necessity to modulate large protein–protein or protein–DNA interfaces. Screening large compound libraries may generally select hit compounds that inhibit the transcriptional activity or may be designed to home in on a specific mechanism of action. This approach has further hurdles due to limited cell

internalization of the hit, difficult or even impossible improvement depending on its chemical nature, and, most importantly, off-target activity when the hit is tested in cells. Certainly, an important requirement to increase the probability of success is a more precise knowledge of the structure and mechanism of action of a TF interacting with its cognate DNA sequence (or protein partners). In this context, our data provide a clear picture that dissects the molecular mechanism of NF-Y inhibition by suramin, thus setting the rationale grounds for the development of new potential NF-Y inhibitory compound(s), with improved properties for *in vivo* testing.

Supplementary Materials: The following figures are available online at www.mdpi.com/xxx/s1, Figure S1: Electrostatic surface of NF-Yd, Figure S2: Structural changes induced by suramin binding.

(See Appendix IV.8)

Author Contributions: Conceptualization, M.N.; methodology, A.S., C.A., C.C. and S.P.; software, A.C.S., A.S., C.A., and C.C.; validation, D.D., N.G.; investigation, V.N., A.C.-S., M.L., C.A., A.S., C.C. and A.B.; data curation, V.N., A.C.-S., M.L., A.S., C.A. and C.C.; writing—original draft preparation, M.N.; writing—review and editing, V.N., A.S., C.A., C.C., A.C.-S., N.G. and R.M.; funding acquisition, M.N. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Italian Association for Cancer Research (AIRC) grant number IG-15267 to M.N., IG-19050 to R.M., and MIUR Progetto di Rilevante Interesse Nazionale (PRIN) 2017 grant number 2017SBFHLH to R.M.

Acknowledgments: We acknowledge the European Synchrotron Radiation Facility (ESRF, Grenoble, France) for provision of synchrotron radiation facilities, and for support during data collection and processing at beamline ID29.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mantovani, R. The molecular biology of the CCAAT-binding factor NF-Y. *Gene* **1999**, *239*, 15–27, doi:10.1016/s0378-1119(99)00368-6.
2. Dolfini, D.; Zambelli, F.; Pavesi, G.; Mantovani, R. A perspective of promoter architecture from the CCAAT box. *Cell Cycle* **2009**, *8*, 4127–4137, doi:10.4161/cc.8.24.10240.
3. Yang, C.; Bolotin, E.; Jiang, T.; Sladek, F.M.; Martinez, E. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **2007**, *389*, 52–65, doi:10.1016/j.gene.2006.09.029.
4. Dolfini, D.; Mantovani, R. Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y? *Cell Death Differ.* **2013**, *20*, 676–685, doi:10.1038/cdd.2013.13.
5. Bergh, K.T.; Litzka, O.; Brakhage, A.A. Identification of a major cis-acting DNA element controlling the bidirectionally transcribed penicillin biosynthesis genes acvA (pcbAB) and ipnA (pcbC) of *Aspergillus nidulans*. *J. Bacteriol.* **1996**, *178*, 3908–3916, doi:10.1128/jb.178.13.3908-3916.1996.
6. Ceribelli, M.; Dolfini, D.; Merico, D.; Gatta, R.; Viganò, A.M.; Pavesi, G.; Mantovani, R. The histone-like NF-Y is a bifunctional transcription factor. *Mol. Cell. Biol.* **2008**, *28*, 2047–2058, doi:10.1128/MCB.01861-07.
7. Deng, H.; Sun, Y.; Zhang, Y.; Luo, X.; Hou, W.; Yan, L.; Chen, Y.; Tian, E.; Han, J.; Zhang, H. Transcription factor NFY globally represses the expression of the C. elegans Hox gene abdominal-B homolog egl-5. *Dev. Biol.* **2007**, *308*, 583–592, doi:10.1016/j.ydbio.2007.05.021.
8. Littlejohn, T.G.; Hynes, M.J. Analysis of the site of action of the amdR product for regulation of the amdS gene of *Aspergillus nidulans*. *Mol. Gen. Genet.* **1992**, *235*, 81–88, doi:10.1007/BF00286184.
9. Litzka, O.; Bergh, K.T.; Brakhage, A.A. The *Aspergillus nidulans* penicillin biosynthesis gene aat (penDE) is controlled by a CCAAT-containing DNA element. *Eur. J. Biochem.* **1996**, *238*, 675–682, doi:10.1111/j.1432-1033.1996.0675w.x.
10. Steidl, S.; Hynes, M.J.; Brakhage, A.A. The *Aspergillus nidulans* multimeric CCAAT binding complex AnCF is negatively autoregulated via its hapB subunit gene. *J. Mol. Biol.* **2001**, *306*, 643–653, doi:10.1006/jmbi.2001.4412.

11. Bolotin-Fukuhara, M. Thirty years of the HAP2/3/4/5 complex. *Biochim. Biophys. Acta Gene Regul. Mech.* **2017**, *1860*, 543–559, doi:10.1016/j.bbagrm.2016.10.011.
12. Benatti, P.; Chiaramonte, M.L.; Lorenzo, M.; Hartley, J.A.; Hochhauser, D.; Gnesutta, N.; Mantovani, R.; Imbriano, C.; Dolfini, D. NF-Y activates genes of metabolic pathways altered in cancer cells. *Oncotarget* **2016**, *7*, 1633–1650, doi:10.18632/oncotarget.6453.
13. DeBerardinis, R.J.; Lum, J.J.; Hatzivassiliou, G.; Thompson, C.B. The biology of cancer: Metabolic reprogramming fuels cell growth and proliferation. *Cell Metab.* **2008**, *7*, 11–20, doi:10.1016/j.cmet.2007.10.002.
14. Kroemer, G.; Pouyssegur, J. Tumor cell metabolism: cancer's Achilles' heel. *Cancer Cell* **2008**, *13*, 472–482, doi:10.1016/j.ccr.2008.05.005.
15. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674, doi:10.1016/j.cell.2011.02.013.
16. Goodarzi, H.; Elemento, O.; Tavazoie, S. Revealing global regulatory perturbations across human cancers. *Mol. Cell* **2009**, *36*, 900–911, doi:10.1016/j.molcel.2009.11.016.
17. Cicchillitti, L.; Corrado, G.; Carosi, M.; Dabrowska, M.E.; Loria, R.; Falcioni, R.; Cutillo, G.; Piaggio, G.; Vizza, E. Prognostic role of NF-YA splicing isoforms and Lamin A status in low grade endometrial cancer. *Oncotarget* **2017**, *8*, 7935–7945, doi:10.18632/oncotarget.13854.
18. Dolfini, D.; Andrioletti, V.; Mantovani, R. Overexpression and alternative splicing of NF-YA in breast cancer. *Sci. Rep.* **2019**, *9*, 12955, doi:10.1038/s41598-019-49297-5.
19. Bie, L.Y.; Li, D.; Mu, Y.; Wang, S.; Chen, B.B.; Lyu, H.F.; Han, L.L.; Nie, C.Y.; Yang, C.C.; Wang, L.; et al. Analysis of cyclin E co-expression genes reveals nuclear transcription factor Y subunit alpha is an oncogene in gastric cancer. *Chronic Dis. Transl. Med.* **2018**, *5*, 44–52, doi:10.1016/j.cdtm.2018.07.003.
20. Bezzecchi, E.; Ronzio, M.; Dolfini, D.; Mantovani, R. NF-YA Overexpression in Lung Cancer: LUSC. *Genes* **2019**, *10*, 937, doi:10.3390/genes10110937.
21. Bezzecchi, E.; Ronzio, M.; Semeghini, V.; Andrioletti, V.; Mantovani, R.; Dolfini, D. NF-YA Overexpression in Lung Cancer: LUAD. *Genes* **2020**, *11*, 198, doi:10.3390/genes11020198.
22. Tsigelny, I.F.; Mukthavaram, R.; Kouznetsova, V.L.; Chao, Y.; Babic, I.; Nurmemmedov, E.; Pastorino, S.; Jiang, P.; Calligaris, D.; Agar, N.; et al. Multiple spatially related pharmacophores define small molecule inhibitors of OLIG2 in glioblastoma. *Oncotarget* **2017**, *8*, 22370–22384, doi:10.18632/oncotarget.5633.
23. Bykov, V.J.; Wiman, K.G. Mutant p53 reactivation by small molecules makes its way to the clinic. *FEBS Lett.* **2014**, *588*, 2622–2627, doi:10.1016/j.febslet.2014.04.017.
24. Gayvert, K.M.; Dardenne, E.; Cheung, C.; Boland, M.R.; Lorberbaum, T.; Wanjala, J.; Chen, Y.; Rubin, M.A.; Tatonetti, N.P.; Rickman, D.S.; et al. A Computational Drug Repositioning Approach for Targeting Oncogenic Transcription Factors. *Cell Rep.* **2016**, *15*, 2348–2356, doi:10.1016/j.celrep.2016.05.037.
25. Hochhauser, D.; Kotecha, M.; O'hare, C.; Morris, P.J.; Hartley, J.M.; Taherbhai, Z.; Harris, D.; Forni, C.; Mantovani, R.; Lee, M.; et al. Modulation of topoisomerase IIalpha expression by a DNA sequence-specific polyamide. *Mol. Cancer Ther.* **2007**, *6*, 346–354, doi:10.1158/1535-7163.MCT-06-0503.
26. Kotecha, M.; Kluza, J.; Wells, G.; O'Hare, C.C.; Forni, C.; Mantovani, R.; Howard, P.W.; Morris, P.; Thurston, D.E.; Hartley, J.A.; et al. Inhibition of DNA binding of the NF-Y transcription factor by the pyrrolobenzodiazepine-polyamide conjugate GWL-78. *Mol. Cancer Ther.* **2008**, *7*, 1319–1328, doi:10.1158/1535-7163.MCT-07-0475.
27. Mackay, H.; Brown, T.; Sexton, J.S.; Kotecha, M.; Nguyen, B.; Wilson, W.D.; Kluza, J.; Savic, B.; O'Hare, C.; Hochhauser, D.; et al. Targeting the inverted CCAAT Box-2 of the topoisomerase IIalpha gene: DNA sequence selective recognition by a polyamide-intercalator as a staggered dimer. *Bioorganic Med. Chem.* **2008**, *16*, 2093–2102, doi:10.1016/j.bmc.2007.10.059.
28. Franks, A.; Tronrud, C.; Kiakos, K.; Kluza, J.; Munde, M.; Brown, T.; Mackay, H.; Wilson, W.D.; Hochhauser, D.; Hartley, J.A.; et al. Targeting the ICB2 site of the topoisomerase IIalpha promoter with a formamido-pyrrole-imidazole-pyrrole H-pin polyamide. *Bioorganic Med. Chem.* **2010**, *18*, 5553–5561, doi:10.1016/j.bmc.2010.06.041.
29. Brucoli, F.; Hawkins, R.M.; James, C.H.; Jackson, P.J.; Wells, G.; Jenkins, T.C.; Ellis, T.; Kotecha, M.; Hochhauser, D.; Hartley, J.A.; et al. An extended pyrrolobenzodiazepine-polyamide conjugate with

selectivity for a DNA sequence containing the ICB2 transcription factor binding site. *J. Med. Chem.* **2013**, *56*, 6339–6351, doi:10.1021/jm4001852.
30. Pett, L.; Kiakos, K.; Satam, V.; Patil, P.; Laughlin-Toth, S.; Gregory, M.; Bowerman, M.; Olson, K.; Savagian, M.; Lee, M.; et al. Modulation of topoisomerase II*α* expression and chemosensitivity through targeted inhibition of NF-Y: DNA binding by a diamino p-anisyl-benzimidazole (Hx) polyamide. *Biochim. Biophys. Acta Gene Regul. Mech.* **2017**, *1860*, 617–629, doi:10.1016/j.bbagrm.2016.10.005.
31. Nardini, M.; Gnesutta, N.; Donati, G.; Gatta, R.; Forni, C.; Fossati, A.; Vonrhein, C.; Moras, D.; Romier, C.; Bolognesi, M.; et al. Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. *Cell* **2013**, *152*, 132–143, doi:10.1016/j.cell.2012.11.047.
32. Huber, E.M.; Scharf, D.H.; Hortschansky, P.; Groll, M.; Brakhage, A.A. DNA minor groove sensing and widening by the CCAAT-binding complex. *Structure* **2012**, *20*, 1757–1768, doi:10.1016/j.str.2012.07.012.
33. Chaves-Sanjuan, A.; Gnesutta, N.; Gobbini, A.; Martignago, D.; Bernardini, A.; Fornara, F.; Mantovani, R.; Nardini, M. Structural determinants for NF-Y subunit organization and NF-Y/DNA association in plants. *Plant J.* **2020**, Oct 24, doi:10.1111/tpj.15038.
34. Jeganathan, S.; Wendt, M.; Kiehstaller, S.; Brancacci, D.; Kuepper, A.; Pospiech, N.; Carotenuto, A.; Novellino, E.; Hennig, S.; Grossmann, T.N. Constrained peptides with fine-tuned flexibility inhibit NF-Y transcription factor assembly. *Angew. Chem. Int. Ed. Engl.* **2019**, *58*, 17351–17358, doi:10.1002/anie.201907901.
35. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. Autodock4 and AutoDockTools4: Automated docking with selective receptor flexiblity. *J. Comp. Chem.* **2009**, *16*, 2785–2791, doi:10.1002/jcc.21256.
36. Bernardini, A.; Lorenzo, M.; Nardini, M.; Mantovani, R.; Gnesutta, N. The phosphorylatable Ser320 of NF-YA is involved in DNA binding of the NF-Y trimer. *FASEB J.* **2019**, *33*, 4790–4801, doi:10.1096/fj.201801989R.
37. Diebold, M.L.; Fribourg, S.; Koch, M.; Metzger, T.; Romier, C. Deciphering correct strategies for multiprotein complex assembly by co-expression: Application to complexes as large as the histone octamer. *J. Struct. Biol.* **2011**, *175*, 178–188, doi:10.1016/j.jsb.2011.02.001.
38. Romier, C.; Cocchiarella, F.; Mantovani, R.; Moras, D. The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. *J. Biol. Chem.* **2003**, *278*, 1336–1345, doi:10.1074/jbc.M209635200.
39. Ceribelli, M.; Benatti, P.; Imbriano, C.; Mantovani, R. NF-YC complexity is generated by dual promoters and alternative splicing. *J. Biol. Chem.* **2009**, *284*, 34189–34200, doi:10.1074/jbc.M109.008417.
40. Saponaro, A. Isothermal Titration Calorimetry: A Biophysical Method to Characterize the Interaction between Label-free Biomolecules in Solution. *Bio-Protocol* **2018**, *8*, e2957, doi:10.21769/BioProtoc.2957.
41. Brautigam, C.A.; Zhao, H.; Vargas, C.; Keller, S.; Schuck, P. Integration and global analysis of isothermal titration calorimetry data for studying macromolecular interactions. *Nat. Protoc.* **2016**, *11*, 882–894, doi:10.1038/nprot.2016.044.
42. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 125–132, doi:10.1107/S0907444909047337.
43. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62*, 72–82, doi:10.1107/S0907444905036693.
44. Storoni, L.C.; McCoy, A.J.; Read, R.J. Likelihood-enhanced fast rotation functions. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 432–438, doi:10.1107/S0907444903028956.
45. Emsley, P.; Cowtan, K. Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 2126–2132, doi:10.1107/S0907444904019158.
46. Murshudov, G.N.; Vagin, A.A.; Dodson, E.J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **1997**, *53*, 240–255, doi:10.1107/S0907444996012255.
47. Adams, P.D.; Afonine, P.V.; Bunkóczi, G.; Chen, V.B.; Davis, I.W.; Echols, N.; Headd, J.J.; Hung, L.W.; Kapral, G.J.; Grosse-Kunstleve, R.W.; et al. PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 213–221, doi:10.1107/S0907444909052925.

48. Chen, V.B.; Arendall, W.B.; Headd, J.J.; Keedy, D.A.; Immormino, R.M.; Kapral, G.J.; Murray, L.W.; Richardson, J.S.; Richardson, D.C. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr*. **2010**, *66*, 12–21, doi:10.1107/S0907444909042073.
49. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712–725, doi:10.1002/prot.21123.
50. Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174, doi:10.1002/jcc.20035.
51. VandeVondele, J.; Krack, M.; Fawzi, M.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comp. Phys. Commun.* **2005**, *167*, 103–128, doi:10.1016/j.cpc.2004.12.014.
52. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19-25, doi:10.1016/j.softx.2015.06.001.
53. Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101, doi:10.1063/1.2408420.
54. Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190, doi:10.1063/1.328693.
55. Zhou, X.; Tan, T.-C.; Valiyaveettil, S.; Go, M.L.; Manjunatha Kini, R.; Velazquez-Campoy, A.; Sivaraman, J. Structural Characterization of Myotoxic Ecarpholin S from *Echis carinatus* Venom. *Biophys. J.* **2008**, *95*, 3366–3380, doi:10.1529/biophysj.107.117747.
56. Salvador, G.H.; Dreyer, T.R.; Cavalcante, W.L.; Matioli, F.F.; Dos Santos, J.I.; Velazquez-Campoy, A.; Gallacci, M.; Fontes, M.R. Structural and functional evidence for membrane docking and disruption sites on phospholipase A2-like proteins revealed by complexation with the inhibitor suramin. *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71*, 2066–2078, doi:10.1107/S1399004715014443.
57. Mayer, M.; Meyer, B. Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew. Chem. Int. Ed.* **1999**, *38*, 1784–1788, doi:10.1002/(SICI)1521-3773(19990614)38:12<1784::AID-ANIE1784>3.0.CO;2-Q.
58. Peri, F.; Airoldi, C.; Colombo, S.; Mari, S.; Jiménez-Barbero, J.; Martegani, E.; Nicotra, F. Sugar-Derived Ras Inhibitors: Group Epitope Mapping by NMR Spectroscopy and Biological Evaluation. *Eur. J. Org. Chem.* **2006**, *16*, 3707–3720, doi:10.1002/ejoc.200600132.
59. Airoldi, C.; Merlo, S.; Sironi, E. NMR Molecular Recognition Studies for the Elucidation of Protein and Nucleic Acid Structure and Function. In *Applications of NMR Spectroscopy*; Rahman, A.U., Iqbal Choudhary, M., Eds.; Bentham Science Publishers Ltd.: Sharjah, UAE, 2015; Volume 2, pp. 147–219, doi:10.2174/9781608059652115020l.
60. Wallace, A.C.; Laskowski, R.A.; Thornton, J.M. LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **1995**, *8*, 127–134, doi:10.1093/protein/8.2.127.
61. Maity, S.N. NF-Y (CBF) regulation in specific cell types and mouse models. *Biochim. Biophys. Acta Gene Regul. Mech.* **2017**, *1860*, 598–603, doi:10.1016/j.bbagrm.2016.10.014.
62. Dolfini, D.; Minuzzo, M.; Sertic, S.; Mantovani, R. NF-YA overexpression protects from glutamine deprivation. *Biochim. Biophys. Acta Mol. Cell Res.* **2020**, *1867*, 118571, doi:10.1016/j.bbamcr.2019.118571.
63. Voogd, T.E.; Vansterkenburg, E.L.; Wilting, J.; Janssen, L.H. Recent research on the biological activity of suramin. *Pharmacol. Rev.* **1993**, *45*, 177–203.
64. Mitsuya, H.; Popovic, M.; Yarchoan, R.; Matsushita, S.; Gallo, R.C.; Broder, S. Suramin protection of T cells in vitro against infectivity and cytopathic effect of HTLV-III. *Science* **1984**, *226*, 172–174, doi:10.1126/science.6091268.
65. Villalona-Calero, M.A.; Wientjes, M.G.; Otterson, G.A.; Kanter, S.; Young, D.; Murgo, A.J.; Fischer, B.; DeHoff, C.; Chen, D.; Yeh, T.K.; et al. Phase I study of low-dose suramin as a chemosensitizer in patients with advanced non-small cell lung cancer. *Clin. Cancer Res.* **2003**, *9*, 3303–3311.
66. Wiedemar, N.; Hauser, D.A.; Mäser, P. 100 Years of Suramin. *Antimicrob. Agents Chemother.* **2020**, *64*, e01168-19, doi:10.1128/AAC.01168-19.

67. Larsen, A.K. Suramin: An anticancer drug with unique biological effects. *Cancer Chemother. Pharmacol.* **1993**, *32*, 96–98, doi:10.1007/BF00685609.
68. Nakajima, M.; De Chavigny, A.; Johnson, C.E.; Hamada, J.; Stein, C.A.; Nicolson, G.L. Suramin. A potent inhibitor of melanoma heparanase and invasion. *J. Biol. Chem.* **1991**, *266*, 9661–9666.
69. Firsching, A.; Nickel, P.; Mora, P.; Allolio, B. Antiproliferative and angiostatic activity of suramin analogues. *Cancer Res*. **1995**, *55*, 4957–4961.
70. Parish, C.R.; Freeman, C.; Brown, K.J.; Francis, D.J.; Cowden, W.B. Identification of sulfated oligosaccharide-based inhibitors of tumor growth and metastasis using novel in vitro assays for angiogenesis and heparanase activity. *Cancer Res.* **1999**, *59*, 3433–3441.
71. Marchetti, D.; Reiland, J.; Erwin, B.; Roy, M. Inhibition of heparanase activity and heparanase-induced angiogenesis by suramin analogues. *Int. J. Cancer* **2003**, *104*, 167–174, doi:10.1002/ijc.10930.
72. Su, L.; Bryan, N.; Battista, S.; Freitas, J.; Garabedian, A.; D'Alessio, F.; Romano, M.; Falanga, F.; Fusco, A.; Kos, L.; et al. Suramin potently inhibits binding of the mammalian high mobility group protein AT-hook 2 to DNA. *BioRxiv* **2019**, 838656, doi:10.1101/838656.
73. Schuetz, A.; Min, J.; Antoshenko, T.; Wang, C.L.; Allali-Hassani, A.; Dong, A.; Loppnau, P.; Vedadi, M.; Bochkarev, A.; Sternglanz, R.; et al. Structural basis of inhibition of the human NAD^+-dependent deacetylase SIRT5 by suramin. *Structure* **2007**, *15*, 377–389, doi:10.1016/j.str.2007.02.002.
74. Gnesutta, N.; Nardini, M.; Mantovani. R. The H2A/H2B-like histone-fold domain proteins at the crossroad between chromatin and different DNA metabolisms. *Transcription* **2013**, *4*, 114–119, doi:10.4161/trns.25002.

III.4 CONCLUSIONS AND FUTURE PERSPECTIVES

The TF NF-Y is a heterotrimer consisting of the NF-YB/NF-YC histone-like domain (HFD) dimer and the sequence-specific NF-YA subunit. Our research group solved the crystal structure of NF-Y in complex with its CCAAT-containing DNA target, which provided the first detailed molecular picture of a unique sequence-specific DNA recognition mechanism. The HFD dimer binds non-specifically to the DNA sugar-phosphate backbone and provides the binding surface for NF-YA, which in turn contacts the CCAAT-box nucleotides by deeply inserting into the DNA minor groove (Nardini *et al.*, 2013). Genome-wide analysis within ENCODE (Encyclopedia of DNA Elements) indicated NF-Y as a key pioneer TF that binds the CCAAT-box in thousands of genomic sites, including promoters of genes overexpressed in cancer (Fleming *et al.*, 2013).
DNA intercalating derivatives have been long studied as proliferation inhibitors targeting NF-Y/CCAAT complex. Our aim was instead to exploit NF-Y/DNA 3D structure information for targeting the NF-Y protein component, and not its DNA element, with small-molecule ligands for its inhibition. To do that, we applied a discovery pipeline consisting of *in silico* screening of small molecule virtual chemical libraries in search for drug hits and compounds potentially hampering CCAAT-box interaction. Such discovery phase was followed by experimental confirmation of inhibition *in vitro*, by combining site-directed mutagenesis, X-ray crystallography, electrophoretic mobility shift assay, isothermal titration calorimetry, and saturation-transfer difference (STD) NMR. In the published study, we demonstrated that suramin, a charged bis-hexasulfonated napthylurea, inhibits the DNA interaction of NF-Y with the CCAAT box by binding the HFD subunits, either as isolated heterodimers or in the complete trimer. Suramin is a known pharmacological compound, used for its antiproliferative effects. Furthermore, suramin has been used for the treatment of early stages of human sleeping sickness and onchocerciasis, an infectious cause of blindness (Voogd *et al*, 1993). ITC measurements indicate that the binding affinity to the NF-Y dimer for suramin is modest (K_d=2.92 μM). Interestingly, the suramin/NF-Y dimer complex exhibits a biphasic curve in ITC, indicating that a different and more complex event occurs when suramin binds. Gel filtration experiments on NF-Y HFD dimer/suramin at different ligand/protein molar ratios, indicate the presence of two peaks in solution, corresponding to a dimer of NF-Y HFD dimers in presence of equimolar or excess of suramin (5-fold). ITC and gel filtration analysis indicate that suramin binding has a dimerization side effect on NF-Y HFD dimer. This happens only at low suramin concentration, indeed when each NF-Y dimer is saturated by suramin binding, no dimerization of NF-Y dimers occurs. This suggests that one

suramin molecule could bind every two NF-Y HFD dimers. The structural analysis of the NF-Y/suramin complex provides a rational interpretation of this process. suramin binds to the surface of the HFDs and promotes homodimerization through the formation of an $(NF\text{-}YB/NF\text{-}YC)_2$–suramin quaternary structure. Besides, structural superimposition with the NF-Y/DNA complex shed light on the mechanism of DNA-binding inhibition induced by suramin. In the $(NF\text{-}Yd)_2$–suramin complex, the ligand is located close to the DNA binding region of NF-Y, contacting NF-YC L1 and α1 and NF-YB αC, which are responsible for more than 40% of the contacts with the DNA phosphate backbone. Furthermore, two sulfonic acid groups that decorate the suramin naphthalene moiety match well with the position of two DNA phosphate groups, corresponding to positions −21 and −20 in the complementary strand of the CCAAT box in the DNA/NF-Y complex (Nardini *et al.*, 2013). On the contrary, suramin binding seems not to interfere with NF-Y trimerization, since the interaction surface between NF-YA and the HFD is distant from the suramin-binding site.

Due to its large, flexible, and multifunctional nature, suramin tends to be a nonselective drug and it also presents poor membrane permeability. In the case of NF-Y, an issue may be its interaction with other HFD-containing proteins because sequence alignment of NF-YB and NF-YC with other human HFD-containing proteins indicates that the NF-Y suramin-binding site is lined by partly conserved residues. Therefore, suramin itself may not be a good candidate for NF-Y selective inhibition but it could be a good lead compound for further chemical improvement. As plan, such areas with low sequence homology with other HFD proteins could be used to drive chemical modifications on suramin to generate either symmetrical or asymmetrical compounds that can improve ligand-binding affinity, and/or selectivity, thus setting the rationale grounds for the development of new potential NF-Y inhibitory compound(s), with improved properties for *in vivo* testing. In line with this purpose, we started a collaboration with the "High-throughput crystallization laboratory" of the European Synchrotron Radiation Facility (ESRF) to apply a fragment-based drug design (FBDD) approach, taking advantage of the high reproducibility of the NF-Y HFD crystals and of their good diffraction (about 1.5 Å). The aim is to find small fragments hits for driving the rational design of new inhibitors and/or to design suramin hybrids with more specificity and druggability features.

PART IV:

APPENDIX

IV.1 Overview of DNA-binding motifs in TFs.

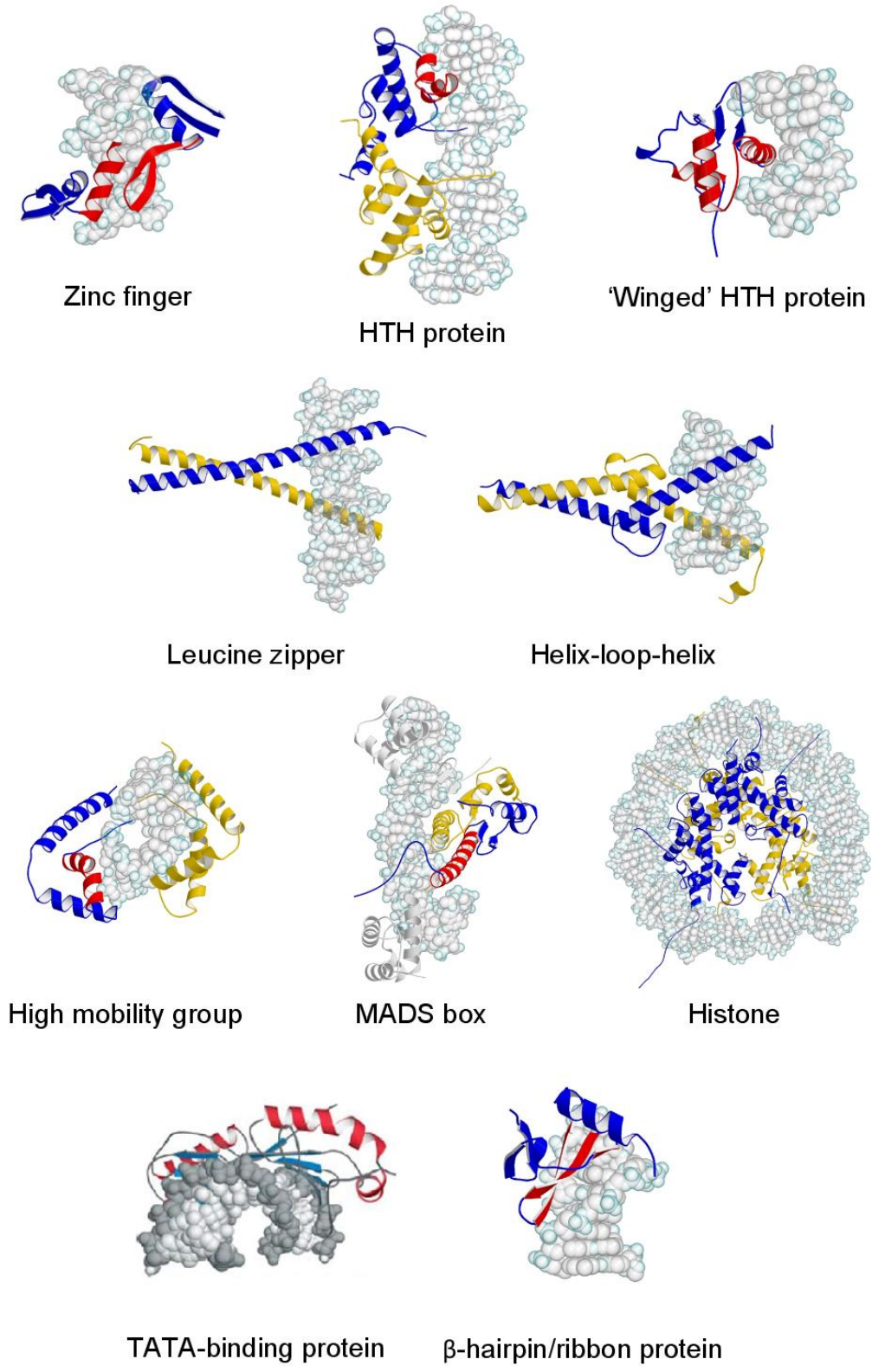


Figure A1. ***Overview of DNA-binding motifs in TFs*.** TFs are shown in ribbon representation, while the DNA is shown as a space-filling model. The PDB-codes of the structures are 1AAY (Zinc finger), 1LMB (Helix-turn-helix), 1BC8 ('Winged' HTH protein), 2DGC (Leucine zipper), 1AM9 (Helix-loop-helix), 1QRV (High mobility group), 1MNM (MADS box), 1AOI (Histone), 1YTB (TATA-binding protein), 1AZP (β-hairpin/ribbon protein). Adapted from Luscombe et al., 2000.

IV.2 Examples of TFs functional inhibition.

Inhibitor name	Company	Mode of action	Clinical trial status
Protein–protein interaction inhibitors			
RG7388	Roche	Inhibits MDM2–p53 binding leading to reduced ubiquitylation of p53, thereby increasing p53 levels leading to increased cell death	NCT02633059 (ref.a) NCT03287245 (ref.b) NCT02670044 (ref.c) NCT03135262 (ref.d) NCT03566485 (ref.e) NCT03850535 (ref.f)
HDM201	Novartis	Inhibits MDM2–p53 binding leading to reduced ubiquitination of p53, thereby increasing p53 levels leading to increased cell death	NCT02890069 (ref.g) NCT02780128 (ref.h) NCT02601378 (ref.i)
KO-539	Kura Oncology	Inhibits menin–MLL binding for treatment of MLL fusion-positive leukaemia; displaces MLL fusion proteins from target genes to reduce MLL fusion- driven gene activation	NCT04067336(ref.j)
SNDX-5613	Syndax	Inhibits menin–MLL binding for treatment of MLL fusion-positive leukaemia; displaces MLL fusion proteins from target genes to reduce MLL fusion- driven gene activation	NCT04065399(ref.k)
LeuSO (AI-10-49)	Systems Oncology	Inhibits CBFβ–SMMHC binding to RUNX1 for treatment of inv(16) AML; restores occupancy of RUNX1 on target genes	IND, enabling studies underway
RITA	NHS Trust	Inhibits p53–HDM-2 interaction	NCT03052036(ref.l)
Nutlin-3	Novartis	Inhibits p53–HDM-2 interactions	NCT02780128(ref.m)
PROTACs			
ARV-110	Arvinas	PROTAC- based degrader of the AR for the treatment of castration- resistant prostate cancer	NCT03888612 (ref.n)
ARV-471	Arvinas	PROTAC- based degrader of the ER for treatment of ER-positive breast cancer	NCT04072952(ref.o)
Modulators of transcription factor gene expression			
SY-1365	Syros	CDK7 inhibitor that alters gene expression, including RUNX1 expression	NCT03134638 (ref.p)
INCB057643	Incyte	Inhibitor of BET protein–acetylated lysine binding	NCT02959437 (ref.q) NCT02711137 (ref.r)
BMS-986158	Bristol-Myers Squibb	Inhibitor of BET protein–acetylated lysine binding	NCT02419417 (ref.s)

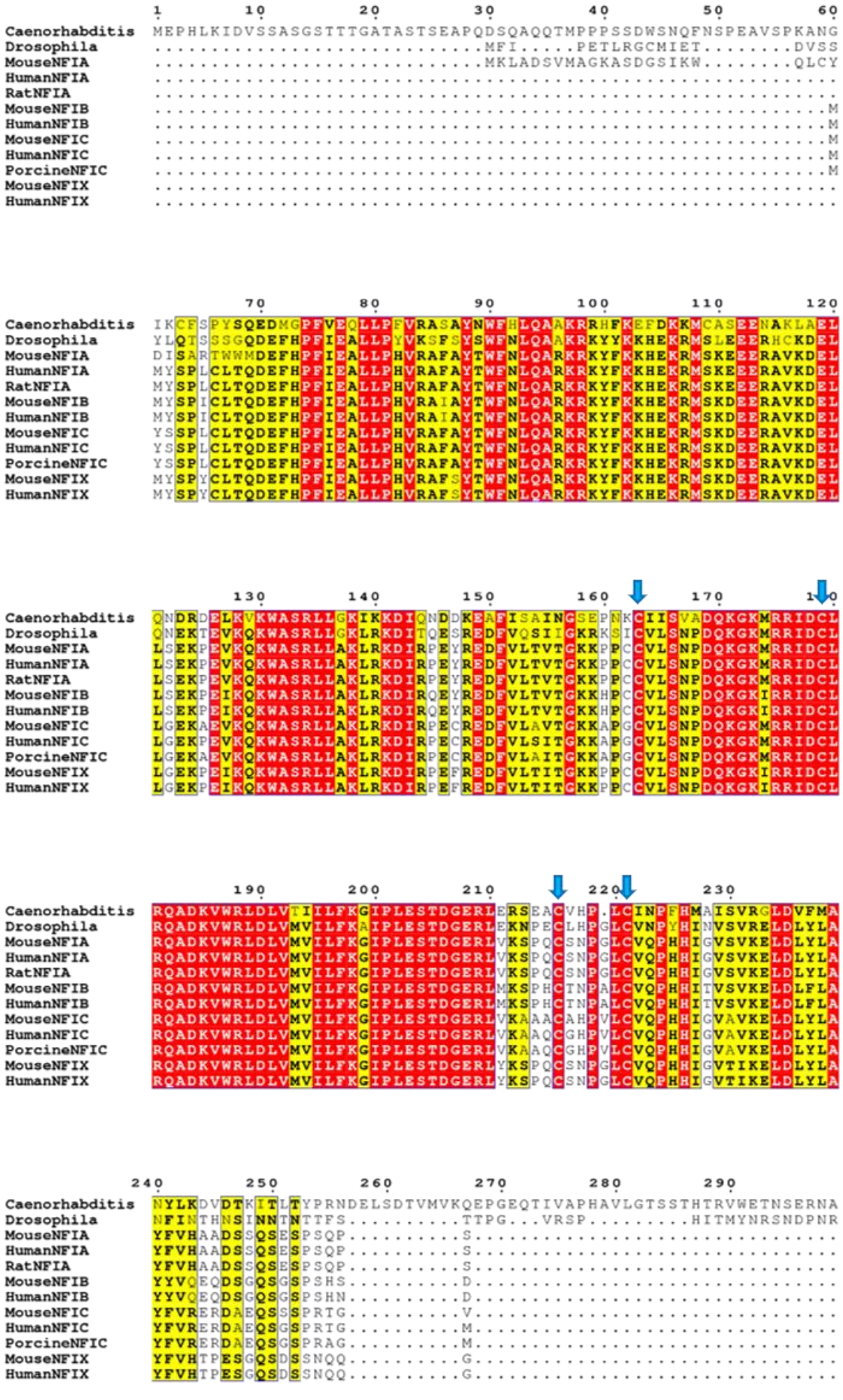
Table A1. ***Examples of TF inhibitors in clinical trials or in clinical development.*** Acronims: AML, acute myeloid leukaemia; AR, androgen receptor; BET, bromodomain and extra- terminal; CBFβ, core binding factor β; CDK7, cyclin- dependent kinase 7; ER, oestrogen receptor; HDM-2, human double minute-2; IND, investigational new drug; MLL, mixed lineage leukaemia; PROTAC, proteolysis targeting chimaera; RUNX1, runt- related transcription factor 1; SMMHC, smooth muscle myosin heavy chain (Bushweller et. al. 2019).

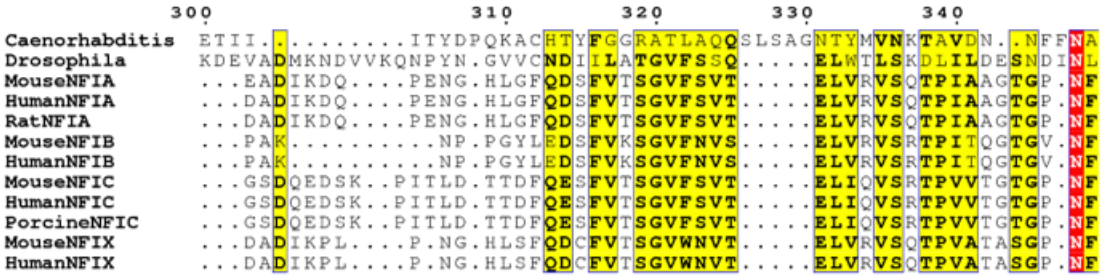
References:

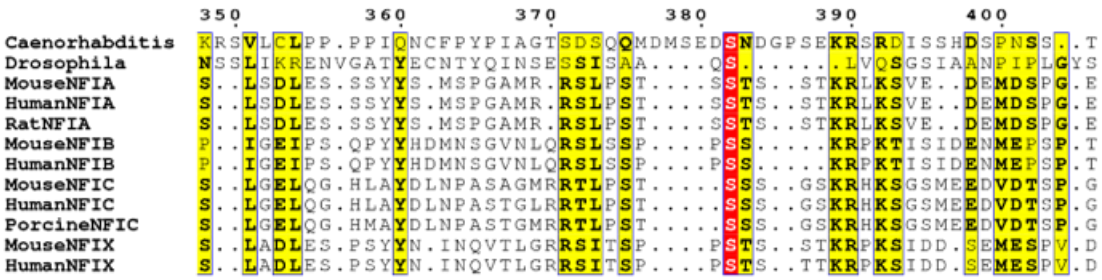
a. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT02633059 (2015).
b. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT03287245 (2017).
c. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT02670044 (2016).
d. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT03135262 (2017).
e. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT03566485 (2018).
f. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT03850535 (2019).

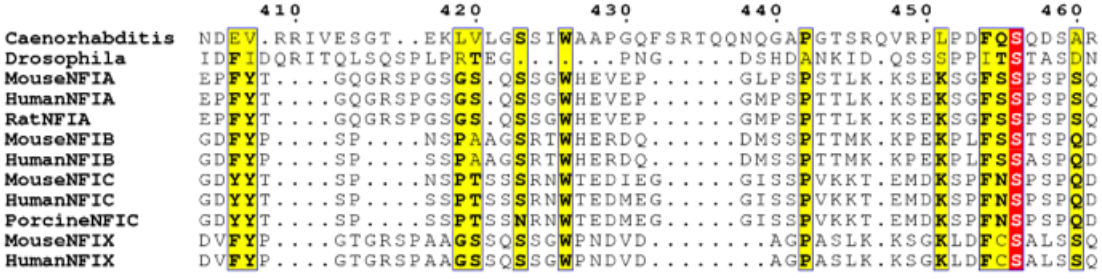
g. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT02890069 (2016).
h. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT02780128 (2016).
i. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT02601378 (2016).
j. US National Library of Medicine. *ClinicalTrials.gov*, https://clinicaltrials.gov/ct2/show/NCT04067336 (2021).
k. US National Library of Medicine. *ClinicalTrials.gov*, https://www.clinicaltrials.gov/ct2/show/NCT04065399 (2020).
l. US National Library of Medicine. *ClinicalTrials.gov*, https://clinicaltrials.gov/ct2/show/NCT03052036 (2020)
m. US National Library of Medicine. *ClinicalTrials.gov*, https://www.clinicaltrials.gov/ct2/show/NCT02780128 (2021)
n. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT03888612 (2019).
o. US National Library of Medicine. *ClinicalTrials.gov*, https://clinicaltrials.gov/ct2/show/NCT04072952 (2021).
p. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT03134638 (2019).
q. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT02959437 (2016).
r. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT02711137 (2016).
s. US National Library of Medicine. *ClinicalTrials.gov*, http://www.clinicaltrials.gov/ct2/show/NCT02419417 (2015).

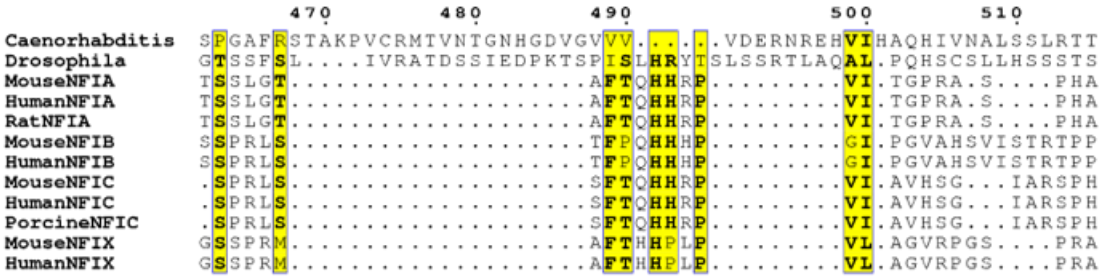
IV.3 Sequence alignment of NFI amino acid sequences

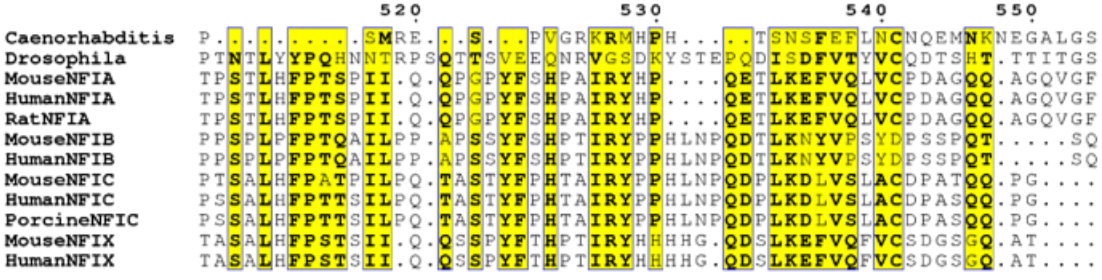

Caenorhabditis
Drosophila
MouseNFIA
HumanNFIA
RatNFIA
MouseNFIB
HumanNFIB
MouseNFIC
HumanNFIC
PorcineNFIC
MouseNFIX
HumanNFIX

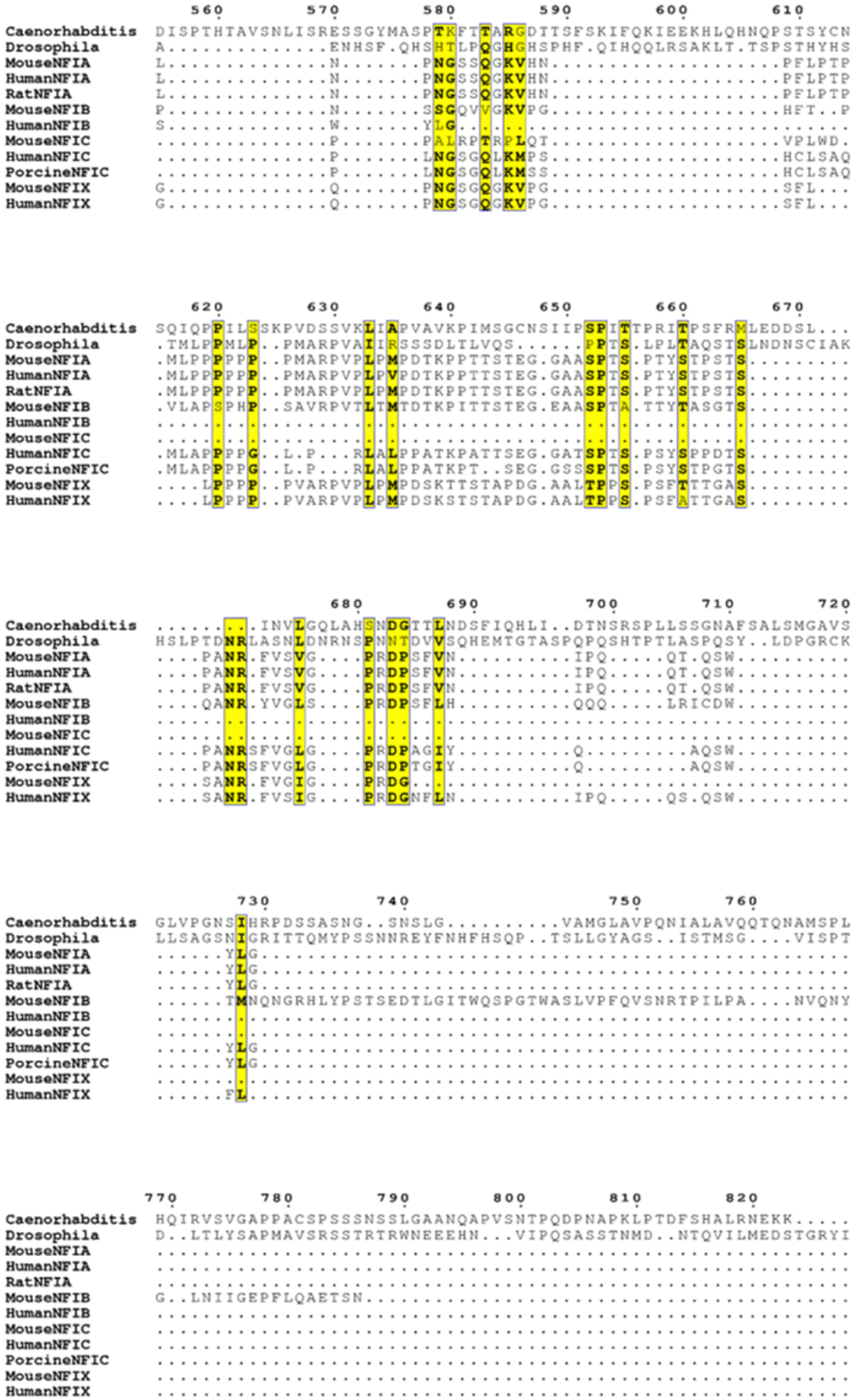

300 310 320 330 340
Caenorhabditis ETII.....ITYDPQKACHTYFGGRATLAQQSLSAGNTYMVNKTAVDN..NFFNA
Drosophila KDEVADMKNDVVKQNPYN.GVVCNDIILATGVFSSQ.....ELWTLSKDLILDESNDINL
MouseNFIA ...EADIKDQ....PENG.HLGFQDSFVTSGVFSVT.....ELVRVSQTPIAAGTGP.NF
HumanNFIA ...DADIKDQ....PENG.HLGFQDSFVTSGVFSVT.....ELVRVSQTPIAAGTGP.NF
RatNFIA ...DADIKDQ....PENG.HLGFQDSFVTSGVFSVT.....ELVRVSQTPIAAGTGP.NF
MouseNFIB ...PAK.........NP.PGYLEDSFVKSGVFNVS.....ELVRVSRTPITQGTGV.NF
HumanNFIB ...PAK.........NP.PGYLEDSFVKSGVFNVS.....ELVRVSRTPITQGTGV.NF
MouseNFIC ...GSDQEDSK..PITLD.TTDFQESFVTSGVFSVT.....ELIQVSRTPVVTGTGP.NF
HumanNFIC ...GSDQEDSK..PITLD.TTDFQESFVTSGVFSVT.....ELIQVSRTPVVTGTGP.NF
PorcineNFIC ...GSDQEDSK..PITLD.TTDFQESFVTSGVFSVT.....ELIQVSRTPVVTGTGP.NF
MouseNFIX ...DADIKPL....P.NG.HLSFQDCFVTSGVWNVT.....ELVRVSQTPVATASGP.NF
HumanNFIX ...DADIKPL....P.NG.HLSFQDCFVTSGVWNVT.....ELVRVSQTPVATASGP.NF


350 360 370 380 390 400
Caenorhabditis KRSVLCLPP.PPIQNCFPYPIAGTSDSQQMDMSEDSNDGPSEKRSRDISSHDSPNSS..T
Drosophila NSSLIKRENVGATYECNTYQINSESSISAA....QS.....LVQSGSIAANPIPLGYS
MouseNFIA S..LSDLES.SSYYS.MSPGAMR.RSLPST....SSTS..STKRLKSVE..DEMDSPG.E
HumanNFIA S..LSDLES.SSYYS.MSPGAMR.RSLPST....SSTS..STKRLKSVE..DEMDSPG.E
RatNFIA S..LSDLES.SSYYS.MSPGAMR.RSLPST....SSTS..STKRLKSVE..DEMDSPG.E
MouseNFIB P..IGEIPS.QPYYHDMNSGVNLQRSLSSP....PSS.....KRPKTISIDENMEPSP.T
HumanNFIB P..IGEIPS.QPYYHDMNSGVNLQRSLSSP....PSS.....KRPKTISIDENMEPSP.T
MouseNFIC S..LGELQG.HLAYDLNPASAGMRRTLPST.....SSS..GSKRHKSGSMEEDVDTSP.G
HumanNFIC S..LGELQG.HLAYDLNPASTGLRRTLPST.....SSS..GSKRHKSGSMEEDVDTSP.G
PorcineNFIC S..LGELQG.HMAYDLNPASTGMRRTLPST.....SSS..GSKRHKSGSMEEDVDTSP.G
MouseNFIX S..LADLES.PSYYN.INQVTLGRRSITSP....PSTS..STKRPKSIDD.SEMESPV.D
HumanNFIX S..LADLES.PSYYN.INQVTLGRRSITSP....PSTS..TTKRPKSIDD.SEMESPV.D


410 420 430 440 450 460
Caenorhabditis NDEV.RRIVESGT..EKLVLGSSIWAAPGQFSRTQQNQGAPGTSRQVRPLPDFQSQDSAR
Drosophila IDFIDQRITQLSQSPLPRTEG.....PNG.....DSHDANKID.QSSSPPITSTASDN
MouseNFIA EPFYT....GQGRSPGSGS.QSSGWHEVEP.....GLPSPSTLK.KSEKSGFSSPSPSQ
HumanNFIA EPFYT....GQGRSPGSGS.QSSGWHEVEP.....GMPSPTTLK.KSEKSGFSSPSPSQ
RatNFIA EPFYT....GQGRSPGSGS.QSSGWHEVEP.....GMPSPTTLK.KSEKSGFSSPSPSQ
MouseNFIB GDFYP....SP....NSPAAGSRTWHERDQ.....DMSSPTTMK.KPEKPLFSSTSPQD
HumanNFIB GDFYP....SP....SSPAAGSRTWHERDQ.....DMSSPTTMK.KPEKPLFSSASPQD
MouseNFIC GDYYT....SP....NSPTSSSRNWTEDIEG.....GISSPVKKT.EMDKSPFNSPSPQD
HumanNFIC GDYYT....SP....SSPTSSSRNWTEDMEG.....GISSPVKKT.EMDKSPFNSPSPQD
PorcineNFIC GDYYT....SP....SSPTSSNRNWTEDMEG.....GISSPVKKT.EMDKSPFNSPSPQD
MouseNFIX DVFYP....GTGRSPAAGSSQSSGWPNDVD.....AGPASLK.KSGKLDFCSALSSQ
HumanNFIX DVFYP....GTGRSPAAGSSQSSGWPNDVD.....AGPASLK.KSGKLDFCSALSSQ


470 480 490 500 510
Caenorhabditis SPGAFRSTAKPVCRMTVNTGNHGDVGVVV.....VDERNREHVIHAQHIVNALSSLRTT
Drosophila GTSSFSL....IVRATDSSIEDPKTSPISLHRYTSLSSRTLAQAL.PQHSCSLLHSSSTS
MouseNFIA TSSLGT.....AFTQHHRP.....VI.TGPRA.S....PHA
HumanNFIA TSSLGT.....AFTQHHRP.....VI.TGPRA.S....PHA
RatNFIA TSSLGT.....AFTQHHRP.....VI.TGPRA.S....PHA
MouseNFIB SSPRLS.....TFPQHHHP.....GI.PGVAHSVISTRTPP
HumanNFIB SSPRLS.....TFPQHHHP.....GI.PGVAHSVISTRTPP
MouseNFIC .SPRLS.....SFTQHHRP.....VI.AVHSG...IARSPH
HumanNFIC .SPRLS.....SFTQHHRP.....VI.AVHSG...IARSPH
PorcineNFIC .SPRLS.....SFTQHHRP.....VI.AVHSG...IARSPH
MouseNFIX GSSPRM.....AFTHHPLP.....VL.AGVRPGS....PRA
HumanNFIX GSSPRM.....AFTHHPLP.....VL.AGVRPGS....PRA


520 530 540 550
Caenorhabditis P.....SMRE..S...PVGRKRMHPH....TSNSFEFLNCNQEMNKNEGALGS
Drosophila PTNTLYYPQHNNTRPSQTTSVEEQNRVGSDKYSTEPQDISDFVTYVCQDTSHT.TTITGS
MouseNFIA TPSTLHFPTSPII.Q.QPGPYFSHPAIRYHP....QETLKEFVQLVCPDAGQQ.AGQVGF
HumanNFIA TPSTLHFPTSPII.Q.QPGPYFSHPAIRYHP....QETLKEFVQLVCPDAGQQ.AGQVGF
RatNFIA TPSTLHFPTSPII.Q.QPGPYFSHPAIRYHP....QETLKEFVQLVCPDAGQQ.AGQVGF
MouseNFIB PPSPLPFPTQAILPP.APSSYFSHPTIRYPPHLNPQDTLKNYVPSYDPSSPQT.....SQ
HumanNFIB PPSPLPFPTQAILPP.APSSYFSHPTIRYPPHLNPQDTLKNYVPSYDPSSPQT.....SQ
MouseNFIC PTSALHFPATPILPQ.TASTYFPHTAIRYPPHLNPQDPLKDLVSLACDPATQQ.PG....
HumanNFIC PSSALHFPTTSILPQ.TASTYFPHTAIRYPPHLNPQDPLKDLVSLACDPASQQ.PG....
PorcineNFIC PSSALHFPTTSILPQ.TASTYFPHTAIRYPPHLNPQDPLKDLVSLACDPASQQ.PG....
MouseNFIX TASALHFPSTSII.Q.QSSPYFTHPTIRYHHHHG.QDSLKEFVQFVCSDGSGQ.AT....
HumanNFIX TASALHFPSTSII.Q.QSSPYFTHPTIRYHHHHG.QDSLKEFVQFVCSDGSGQ.AT....


560 570 580 590 600 610
Caenorhabditis DISPTHTAVSNLISRESSGYMASPTKFTTARGDTTSFSKIFQKIEEKHLQHNQPSTSYCN
Drosophila A.....ENHSF.QHSHTLPQGHGHSPHF.QIHQQLRSAKLT.TSPSTHYHS
MouseNFIA L.....N.....PNGSSQGKVHN.....PFLPTP
HumanNFIA L.....N.....PNGSSQGKVHN.....PFLPTP
RatNFIA L.....N.....PNGSSQGKVHN.....PFLPTP
MouseNFIB P.....N.....SSGQVVGKVPG.....HFT..P
HumanNFIB S.....W.....YLG.....
MouseNFIC .....P.....PALRPTRPLQT.....VPLWD.
HumanNFIC .....P.....LNGSGQLKMPS.....HCLSAQ
PorcineNFIC .....P.....LNGSGQLKMSS.....HCLSAQ
MouseNFIX G.....Q.....PNGSGQGKVPG.....SFL...
HumanNFIX G.....Q.....PNGSGQGKVPG.....SFL...
620 630 640 650 660 670
Caenorhabditis SQIQPPILSSKPVDSSVKLIAPVAVKPIMSGCNSIIPSPITTPRITPSFRMLEDDSL...
Drosophila .TMLPPMLP..PMARPVAIIRSSSDLTLVQS.....PPTS.LPLTAQSTSLNDNSCIAK
MouseNFIA .MLPPPPPP..PMARPVPLPMPDTKPPTTSTEG.GAASPTS.PTYSTPSTS.....
HumanNFIA .MLPPPPPP..PMARPVPLPVPDTKPPTTSTEG.GAASPTS.PTYSTPSTS.....
RatNFIA .MLPPPPPP..PMARPVPLPMPDTKPPTTSTEG.GAASPTS.PTYSTPSTS.....
MouseNFIB .VLAPSPHP..SAVRPVTLTMTDTKPITTSTEG.EAASPTA.TTYTASGTS.....
HumanNFIB .....
MouseNFIC .....
HumanNFIC .MLAPPPPG..L.P...RLALPPATKPATTSEG.GATSPTS.PSYSPPDTS.....
PorcineNFIC .MLAPPPPG..L.P...RLALPPATKPT..SEG.GSSSPTS.PSYSTPGTS.....
MouseNFIX ....LPPPP..PVARPVPLPMPDSKTTSTAPDG.AALTPPS.PSFTTTGAS.....
HumanNFIX ....LPPPP..PVARPVPLPMPDSKSTSTAPDG.AALTPPS.PSFATTGAS.....
680 690 700 710 720
Caenorhabditis .....INVLGQLAHSNDGTTLNDSFIQHLI..DTNSRSPLLSSGNAFSALSMGAVS
Drosophila HSLPTDNRLASNLDNRNSPNNTDVVSQHEMTGTASPQPQSHTPTLASPQSY..LDPGRCK
MouseNFIA ....PANR.FVSVG....PRDPSFVN.....IPQ.....QT.QSW.....
HumanNFIA ....PANR.FVSVG....PRDPSFVN.....IPQ.....QT.QSW.....
RatNFIA ....PANR.FVSVG....PRDPSFVN.....IPQ.....QT.QSW.....
MouseNFIB ....QANR.YVGLS....PRDPSFLH.....QQQ.....LRICDW.....
HumanNFIB .....
MouseNFIC .....
HumanNFIC ....PANRSFVGLG....PRDPAGIY.....Q.....AQSW.....
PorcineNFIC ....PANRSFVGLG....PRDPTGIY.....Q.....AQSW.....
MouseNFIX ....SANR.FVGIG....PRDG..
HumanNFIX ....SANR.FVSIG....PRDGNFLN.....IPQ.....QS.QSW.....
730 740 750 760
Caenorhabditis GLVPGNSIHRPDSSASNG..SNSLG.....VAMGLAVPQNIALAVQQTQNAMSPL
Drosophila LLSAGSNIGRITTQMYPSSNNREYFNHFHSQP..TSLLGYAGS..ISTMSG....VISPT
MouseNFIA .....YLG.....
HumanNFIA .....YLG.....
RatNFIA .....YLG.....
MouseNFIB .....TMNQNGRHLYPSTSEDTLGITWQSPGTWASLVPFQVSNRTPILPA....NVQNY
HumanNFIB .....
MouseNFIC .....
HumanNFIC .....YLG.....
PorcineNFIC .....YLG.....
MouseNFIX .....
HumanNFIX .....FL.....
770 780 790 800 810 820
Caenorhabditis HQIRVSVGAPPACSPSSSNSSLGAANQAPVSNTPQDPNAPKLPTDFSHALRNEKK.....
Drosophila D..LTLYSAPMAVSRSSTRTRWNEEEHN...VIPQSASSTNMD..NTQVILMEDSTGRYI
MouseNFIA .....
HumanNFIA .....
RatNFIA .....
MouseNFIB G..LNIIGEPFLQAETSN.....
HumanNFIB .....
MouseNFIC .....
HumanNFIC .....
PorcineNFIC .....
MouseNFIX .....
HumanNFIX .....

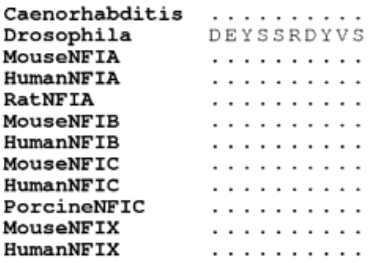
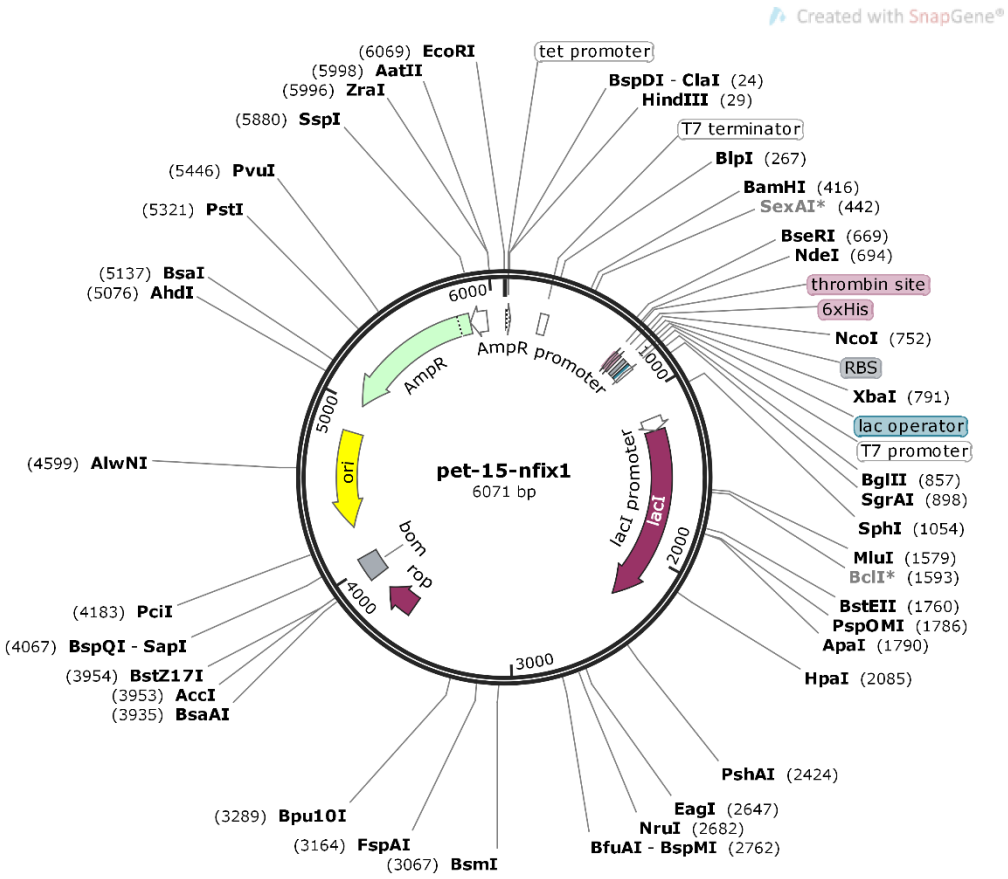
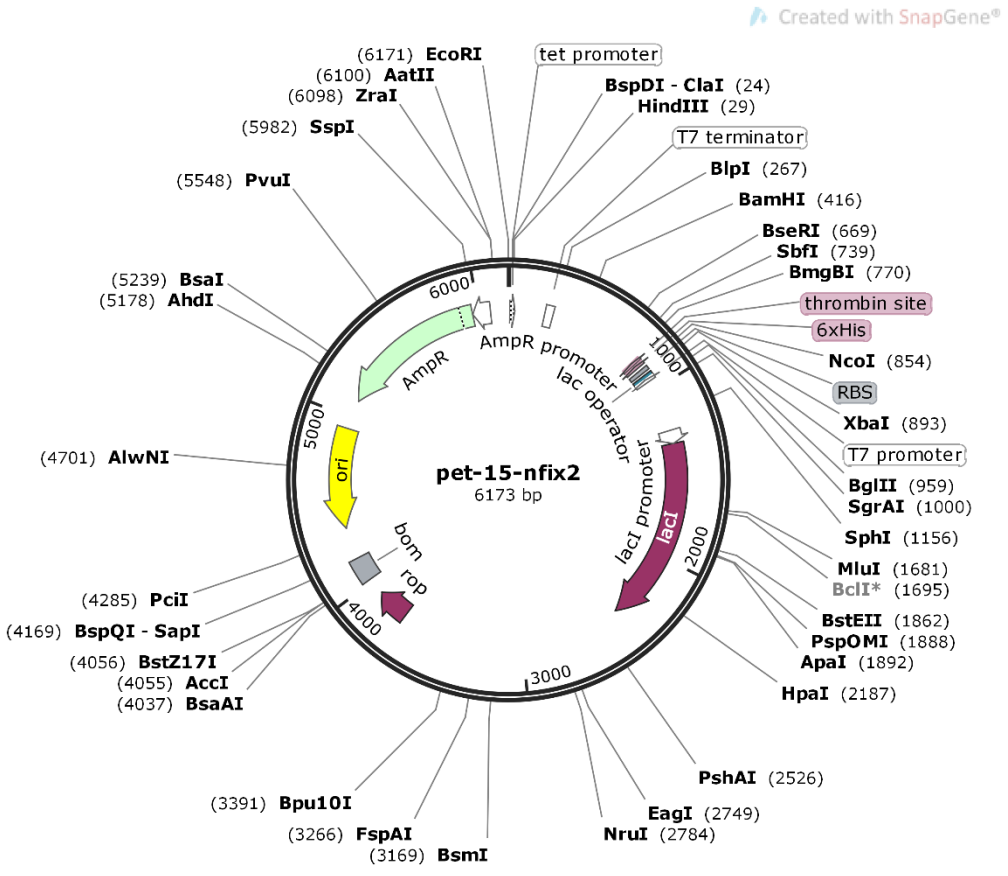
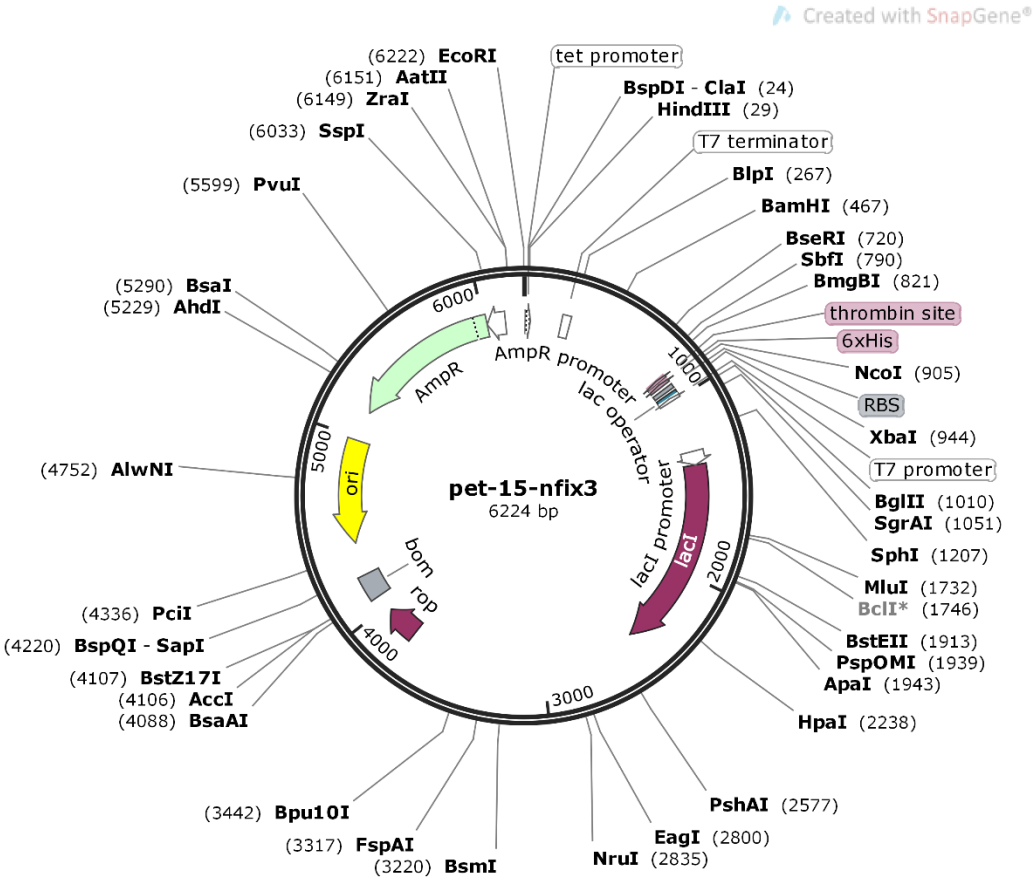


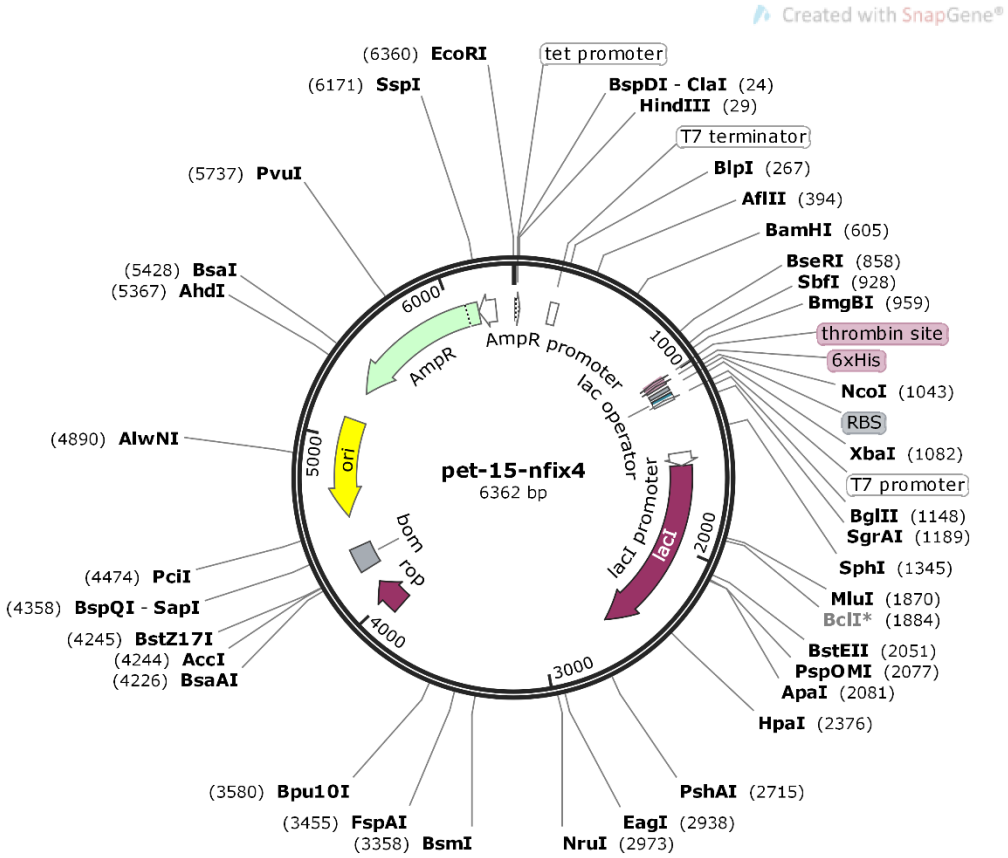
Figure A3. ***Comparison of the NFI amino acid sequences*****.** C. *elegans*, *Drosophila*, mouse, rat and human NFI amino acid sequences are aligned and compared with ESPript 3.x online service (ESPript - http://espript.ibcp.fr) (Robert & Gouet, 2014). NFIC porcine sequence was also included. Multiple sequence alignments of homologous proteins are boxed in colours according to residue conservation. A score is calculated for each column of residues. By default, residue names are written in black if score is below 0.7 (low similarity); they are in black (bold) and framed in yellow if score is in the range 0.7-1 (high similarity); they are in white on a red background in case of strict identity. Conserved cysteine residues are indicated with blue arrows.

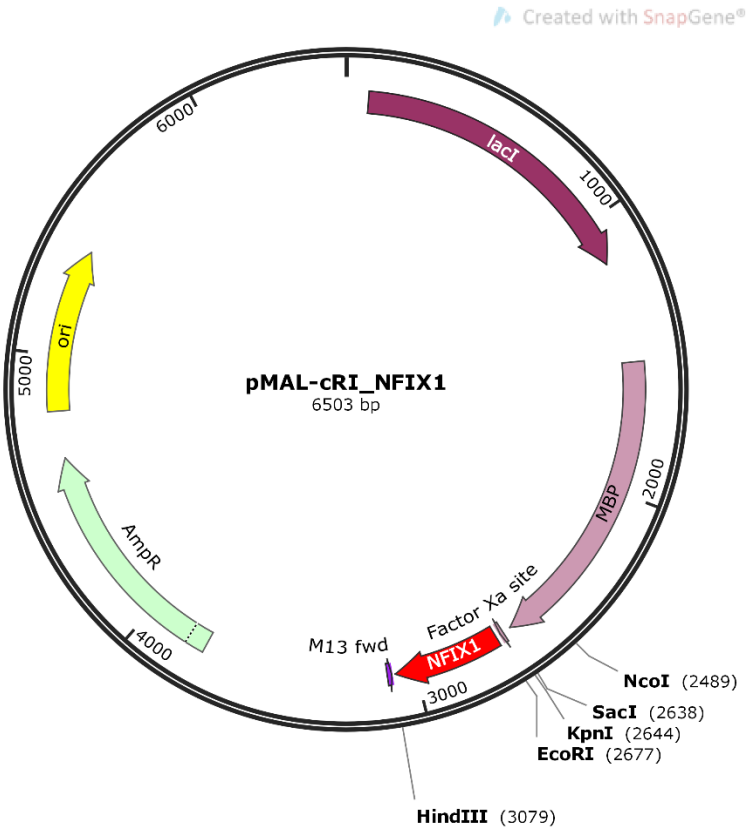
IV.4 Vectors maps used for recombinant NFIX expression

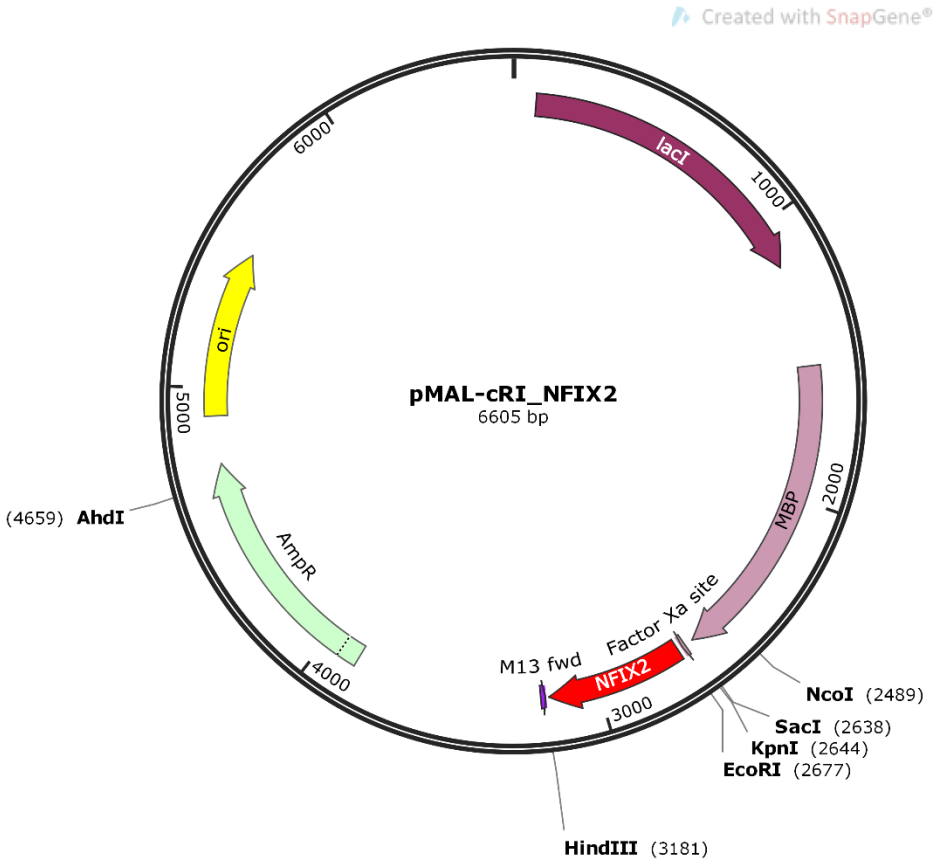


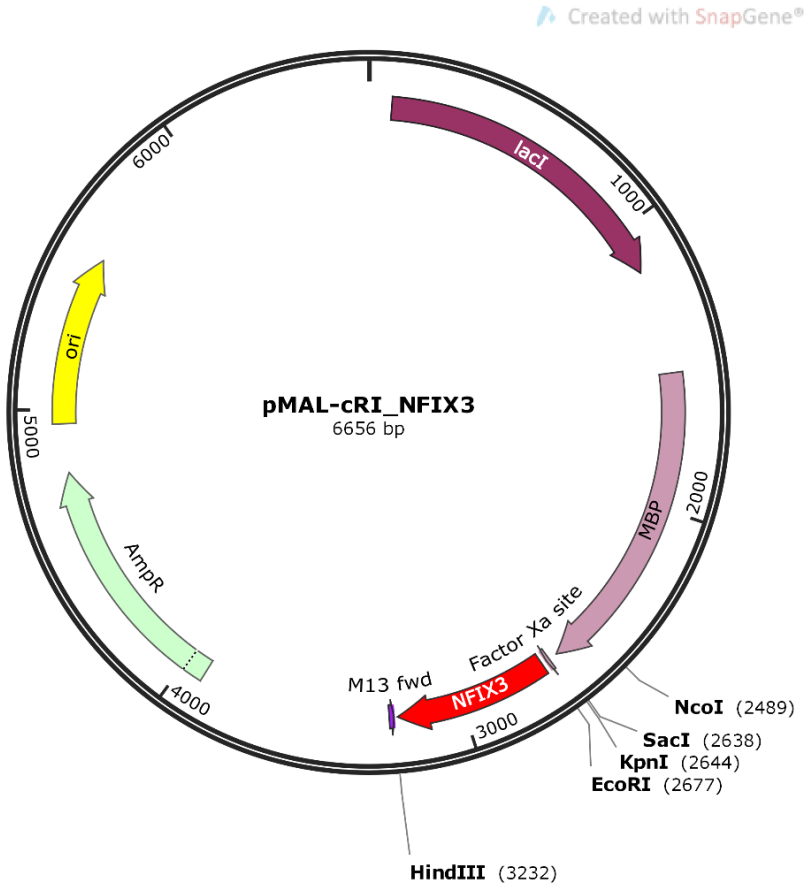


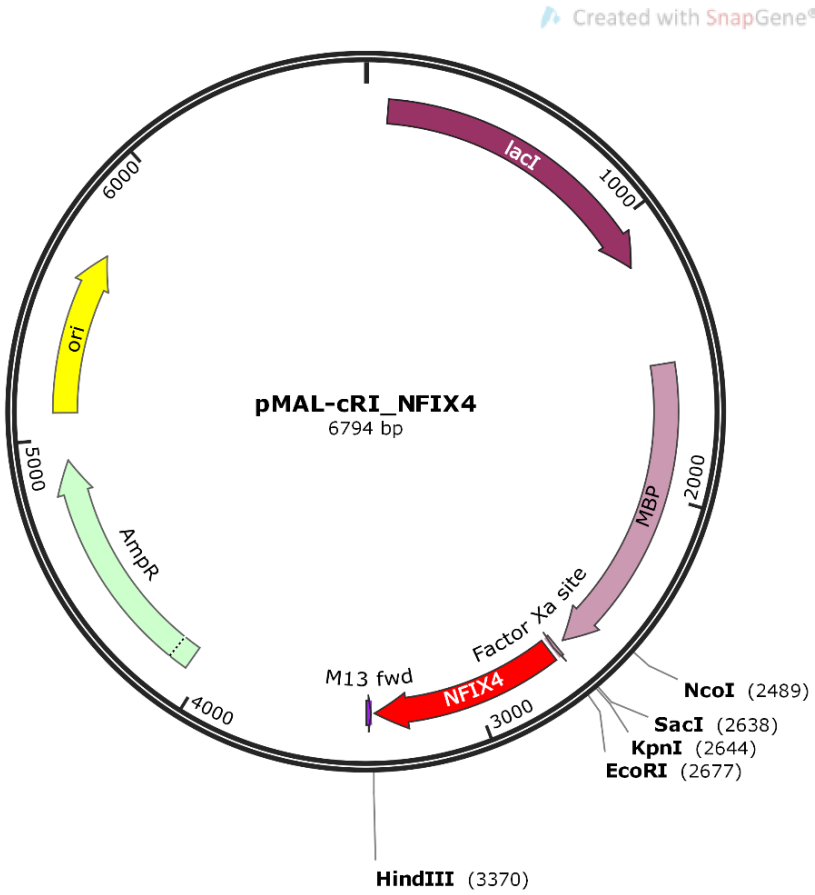

Created with SnapGene®
(6222) EcoRI
(6151) AatII
(6149) ZraI
(6033) SspI
(5599) PvuI
(5290) BsaI
(5229) AhdI
(4752) AlwNI
(4336) PciI
(4220) BspQI - SapI
(4107) BstZ17I
(4106) AccI
(4088) BsaAI
(3442) Bpu10I
(3317) FspAI
(3220) BsmI
tet promoter
BspDI - ClaI (24)
HindIII (29)
T7 terminator
BlpI (267)
BamHI (467)
BseRI (720)
SbfI (790)
BmgBI (821)
thrombin site
6xHis
NcoI (905)
RBS
XbaI (944)
T7 promoter
BglII (1010)
SgrAI (1051)
SphI (1207)
MluI (1732)
BclI* (1746)
BstEII (1913)
PspOMI (1939)
ApaI (1943)
HpaI (2238)
PshAI (2577)
EagI (2800)
NruI (2835)
AmpR promoter
lac operator
lacI promoter
AmpR
lacI
ori
bom
rop
1000
2000
3000
4000
5000
6000
pet-15-nfix3
6224 bp

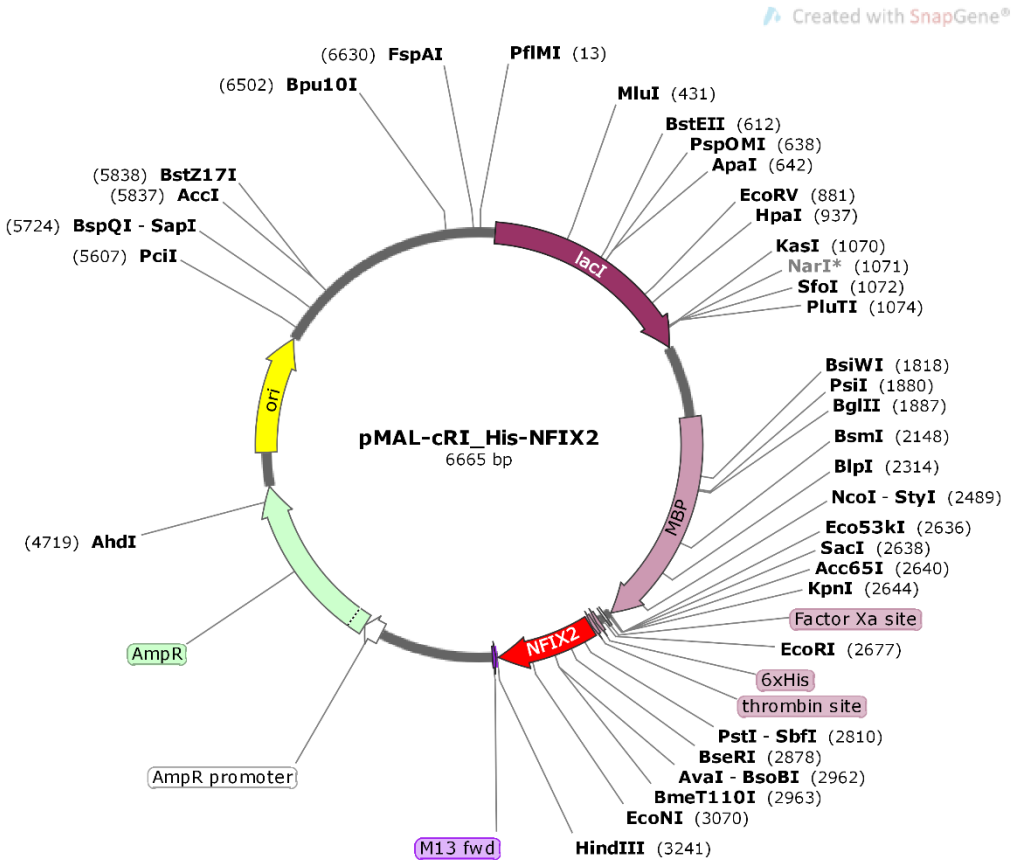

Created with SnapGene®
(6360) EcoRI
(6171) SspI
(5737) PvuI
(5428) BsaI
(5367) AhdI
(4890) AlwNI
(4474) PciI
(4358) BspQI - SapI
(4245) BstZ17I
(4244) AccI
(4226) BsaAI
(3580) Bpu10I
(3455) FspAI
(3358) BsmI
tet promoter
BspDI - ClaI (24)
HindIII (29)
T7 terminator
BlpI (267)
AflII (394)
BamHI (605)
BseRI (858)
SbfI (928)
BmgBI (959)
thrombin site
6xHis
NcoI (1043)
RBS
XbaI (1082)
T7 promoter
BglII (1148)
SgrAI (1189)
SphI (1345)
MluI (1870)
BclI* (1884)
BstEII (2051)
PspOMI (2077)
ApaI (2081)
HpaI (2376)
PshAI (2715)
EagI (2938)
NruI (2973)
AmpR promoter
lac operator
lacI promoter
AmpR
lacI
ori
bom
rop
1000
2000
3000
4000
5000
6000
pet-15-nfix4
6362 bp

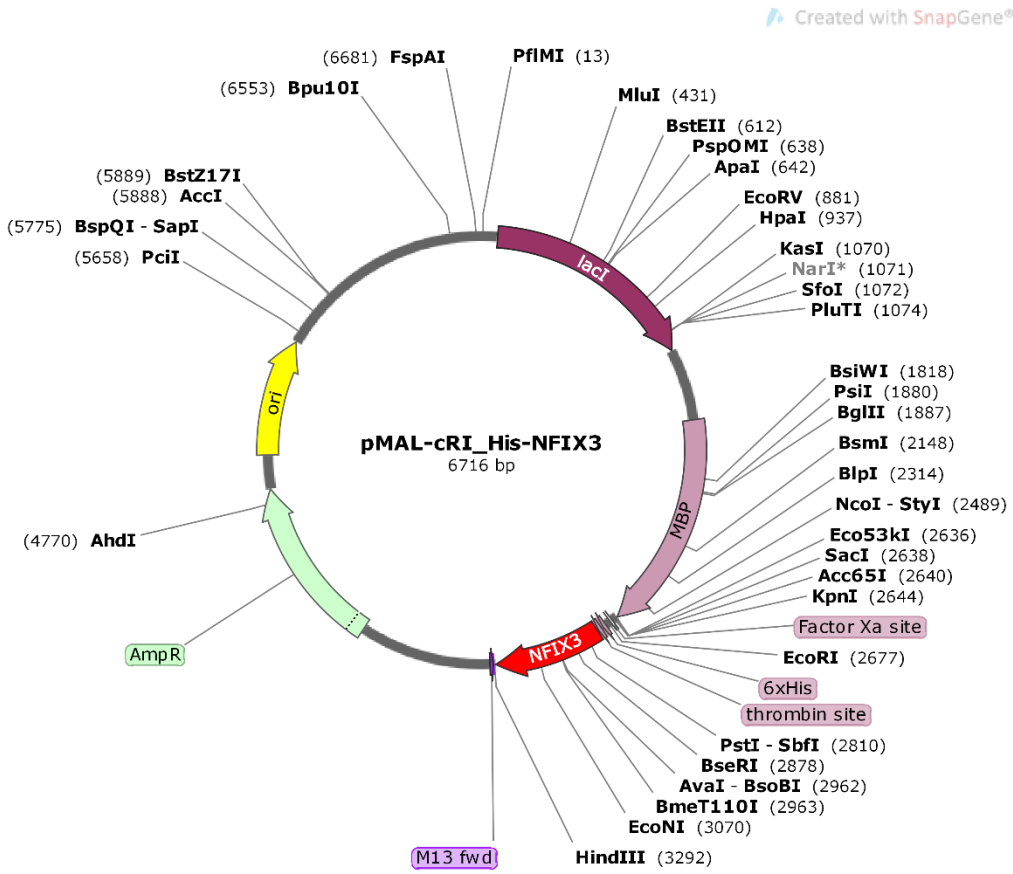

Created with SnapGene®
6000
lacI
1000
ori
5000
pMAL-cRI_NFIX1
6503 bp
MBP
2000
AmpR
Factor Xa site
4000
M13 fwd
NFIX1
3000
NcoI (2489)
SacI (2638)
KpnI (2644)
EcoRI (2677)
HindIII (3079)

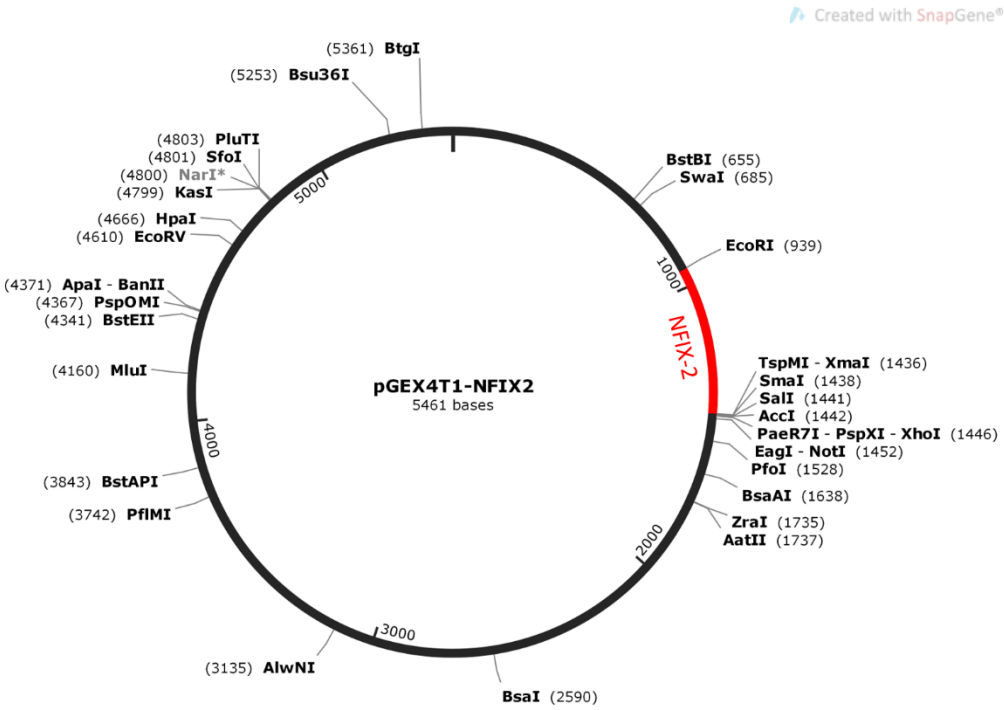

Created with SnapGene®
6000
lacI
1000
ori
5000
pMAL-cRI_NFIX2
6605 bp
(4659) AhdI
MBP
2000
AmpR
Factor Xa site
4000
M13 fwd
NFIX2
3000
NcoI (2489)
SacI (2638)
KpnI (2644)
EcoRI (2677)
HindIII (3181)

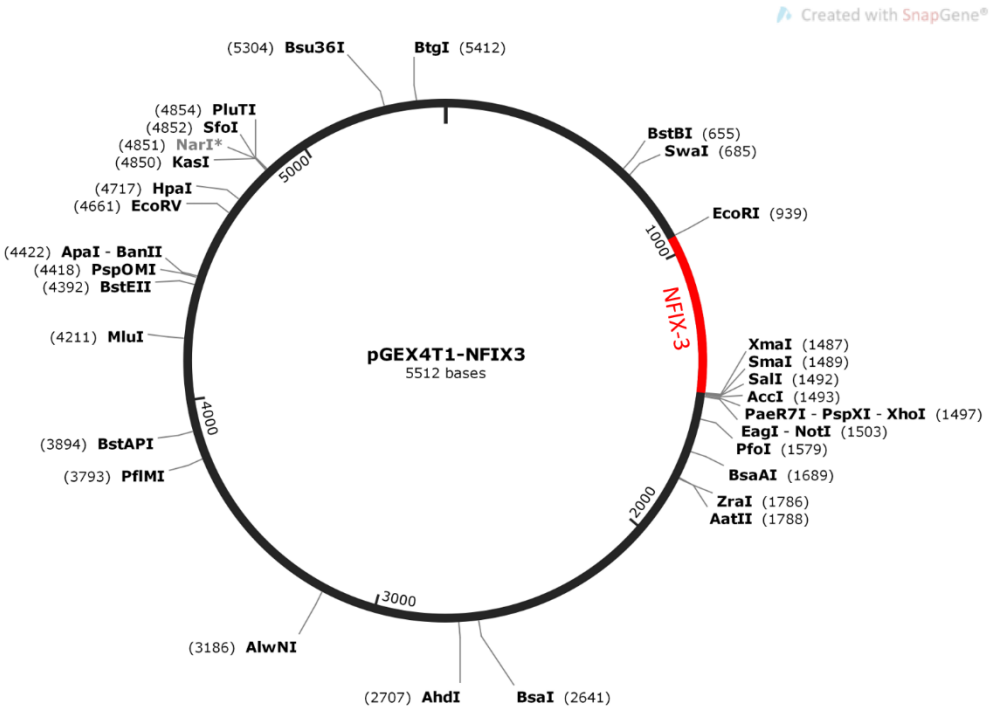

Created with SnapGene®
pMAL-cRI_NFIX3
6656 bp
lacI
1000
MBP
2000
Factor Xa site
NFIX3
M13 fwd
3000
NcoI (2489)
SacI (2638)
KpnI (2644)
EcoRI (2677)
HindIII (3232)
AmpR
4000
ori
5000
6000

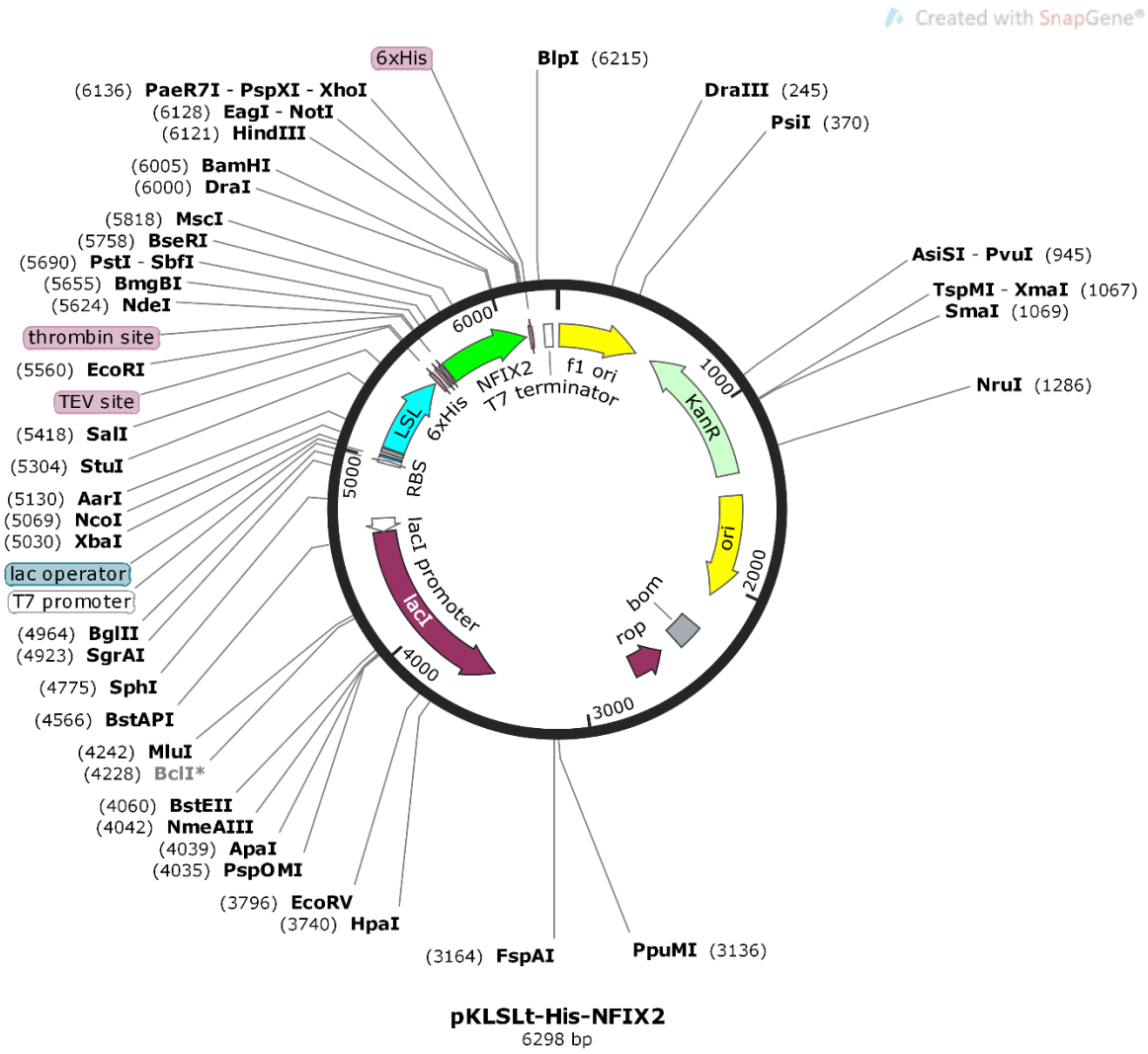

Created with SnapGene®
pMAL-cRI_NFIX4
6794 bp
lacI
1000
MBP
2000
Factor Xa site
NFIX4
M13 fwd
3000
NcoI (2489)
SacI (2638)
KpnI (2644)
EcoRI (2677)
HindIII (3370)
AmpR
4000
ori
5000
6000


Created with SnapGene®
(6630) FspAI
PflMI (13)
(6502) Bpu10I
MluI (431)
BstEII (612)
PspOMI (638)
ApaI (642)
(5838) BstZ17I
(5837) AccI
EcoRV (881)
HpaI (937)
(5724) BspQI - SapI
(5607) PciI
KasI (1070)
NarI* (1071)
SfoI (1072)
PluTI (1074)
lacI
BsiWI (1818)
PsiI (1880)
BglII (1887)
ori
pMAL-cRI_His-NFIX2
6665 bp
BsmI (2148)
BlpI (2314)
MBP
NcoI - StyI (2489)
Eco53kI (2636)
(4719) AhdI
SacI (2638)
Acc65I (2640)
KpnI (2644)
Factor Xa site
NFIX2
AmpR
EcoRI (2677)
6xHis
thrombin site
PstI - SbfI (2810)
BseRI (2878)
AmpR promoter
AvaI - BsoBI (2962)
BmeT110I (2963)
EcoNI (3070)
M13 fwd
HindIII (3241)


Created with SnapGene®
(6681) FspAI
PflMI (13)
(6553) Bpu10I
MluI (431)
BstEII (612)
PspOMI (638)
ApaI (642)
(5889) BstZ17I
(5888) AccI
EcoRV (881)
HpaI (937)
(5775) BspQI - SapI
(5658) PciI
KasI (1070)
NarI* (1071)
SfoI (1072)
PluTI (1074)
lacI
BsiWI (1818)
PsiI (1880)
BglII (1887)
ori
pMAL-cRI_His-NFIX3
6716 bp
BsmI (2148)
BlpI (2314)
MBP
NcoI - StyI (2489)
Eco53kI (2636)
(4770) AhdI
SacI (2638)
Acc65I (2640)
KpnI (2644)
Factor Xa site
NFIX3
AmpR
EcoRI (2677)
6xHis
thrombin site
PstI - SbfI (2810)
BseRI (2878)
AvaI - BsoBI (2962)
BmeT110I (2963)
EcoNI (3070)
M13 fwd
HindIII (3292)


Created with SnapGene®
(5361) BtgI
(5253) Bsu36I
(4803) PluTI
(4801) SfoI
(4800) NarI*
(4799) KasI
(4666) HpaI
(4610) EcoRV
(4371) ApaI - BanII
(4367) PspOMI
(4341) BstEII
(4160) MluI
(3843) BstAPI
(3742) PflMI
(3135) AlwNI
BsaI (2590)
BstBI (655)
SwaI (685)
EcoRI (939)
NFIX-2
TspMI - XmaI (1436)
SmaI (1438)
SalI (1441)
AccI (1442)
PaeR7I - PspXI - XhoI (1446)
EagI - NotI (1452)
PfoI (1528)
BsaAI (1638)
ZraI (1735)
AatII (1737)
pGEX4T1-NFIX2
5461 bases
1000
2000
3000
4000
5000


Created with SnapGene®
(5304) Bsu36I
BtgI (5412)
(4854) PluTI
(4852) SfoI
(4851) NarI*
(4850) KasI
(4717) HpaI
(4661) EcoRV
(4422) ApaI - BanII
(4418) PspOMI
(4392) BstEII
(4211) MluI
(3894) BstAPI
(3793) PflMI
(3186) AlwNI
(2707) AhdI
BsaI (2641)
BstBI (655)
SwaI (685)
EcoRI (939)
NFIX-3
XmaI (1487)
SmaI (1489)
SalI (1492)
AccI (1493)
PaeR7I - PspXI - XhoI (1497)
EagI - NotI (1503)
PfoI (1579)
BsaAI (1689)
ZraI (1786)
AatII (1788)
pGEX4T1-NFIX3
5512 bases
1000
2000
3000
4000
5000



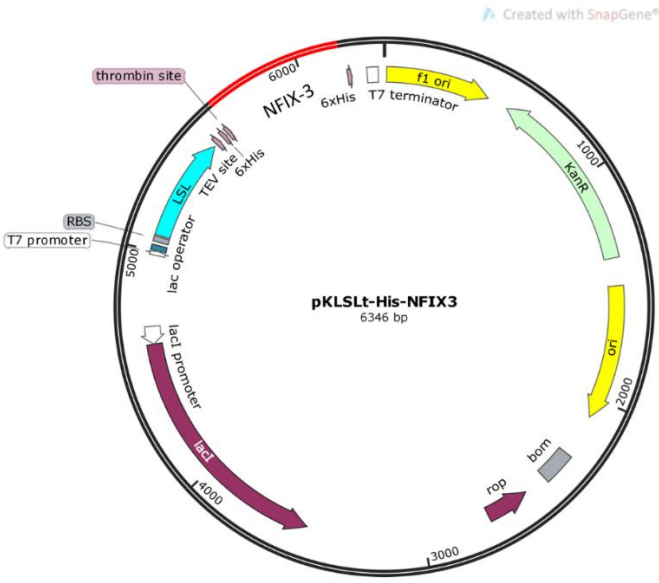


Figure A4. ***NFIX constructs vector maps*****.** Vectors used for expression and purification trials of NFIX constructs -1, -2, -3 and -4. Created using SnapGene.

IV.5 MBP-His-NFIX recombinant proteins purification steps by SDS-PAGE

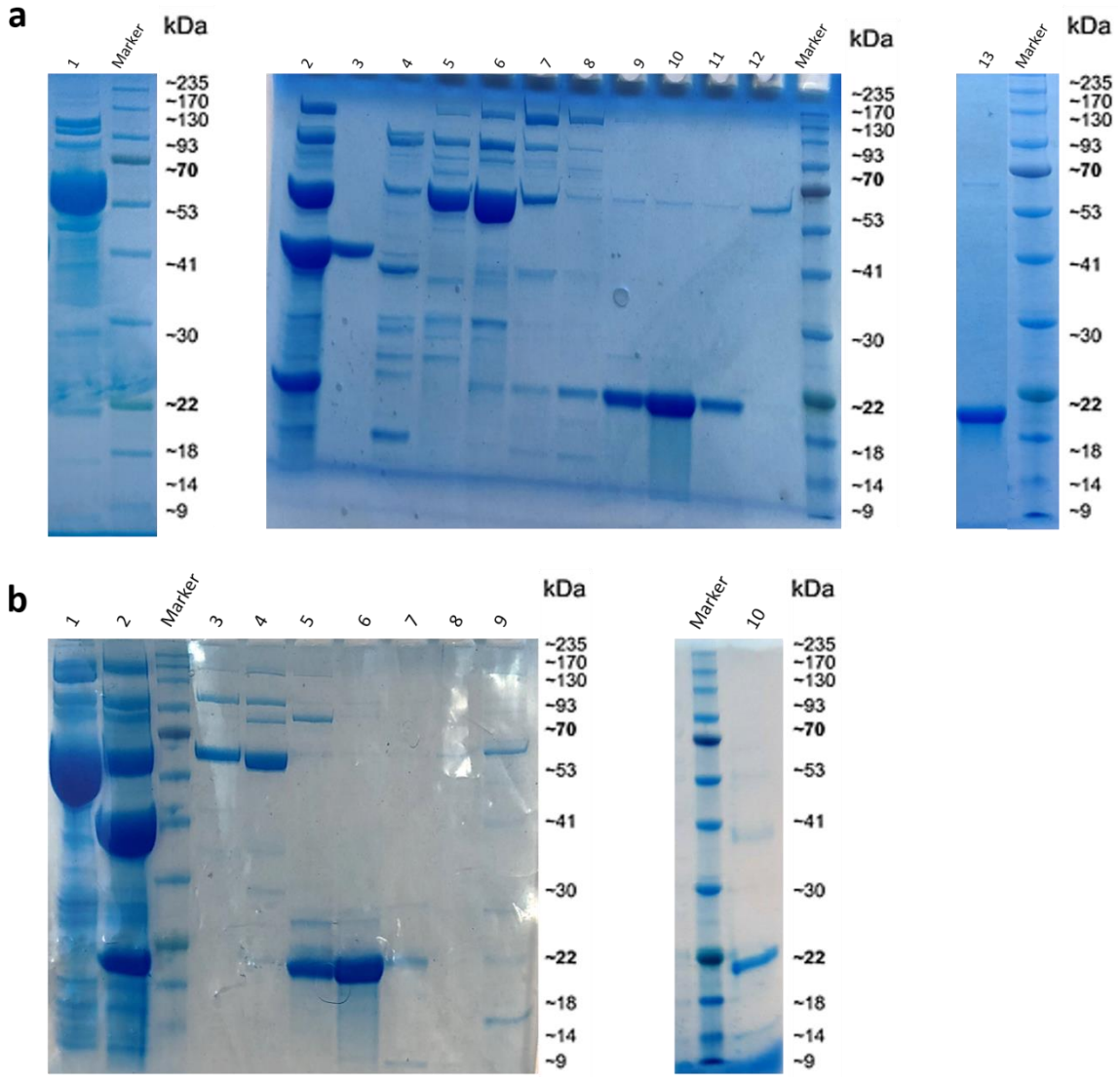


Figure A5. ***MBP-His-NFIX constructs 2 and 3 purification steps illustrated by SDS-PAGE*****. a)** MBP-His-NFIX-2 recombinant protein purification steps. First SDS-PAGE shows first heparin chromatography output; lane 1, MBP-His-NFIX-2 fusion protein (MW=62246 Da); Second SDS-PAGE shows second heparin chromatography output; lane 2, fusion protein after Thrombin cleavage (Fusion protein uncleaved MW=62246 Da, MBP MW=42500 Da, NFIX-2 MW= 19746 Da); lane 3, flow through composed by MBP only, lane 4-7, fractions containing uncleaved fusion protein; lane 8-11 fractions containing NFIX-2; lane 12, protein contaminants. Third SDS-PAGE shows SEC output; lane 13, NFIX-2 pure. **b)** MBP-His-NFIX-3 recombinant protein purification steps. First SDS-PAGE shows first and second heparin chromatography output; lane 1, MBP-His-NFIX-3 fusion protein (MW=64142 Da; lane 2, fusion protein after Thrombin cleavage (Fusion protein uncleaved MW=64142 Da, MBP MW=42500 Da, NFIX-3 MW=21642 Da); lane 3-4, fractions containing uncleaved fusion protein; lane 5-8 fractions containing NFIX-3; lane 9, protein contaminants. Second SDS-PAGE shows SEC output; lane 10, NFIX-3 pure.

IV.6 SEC of the NFIX-2/DNA complex

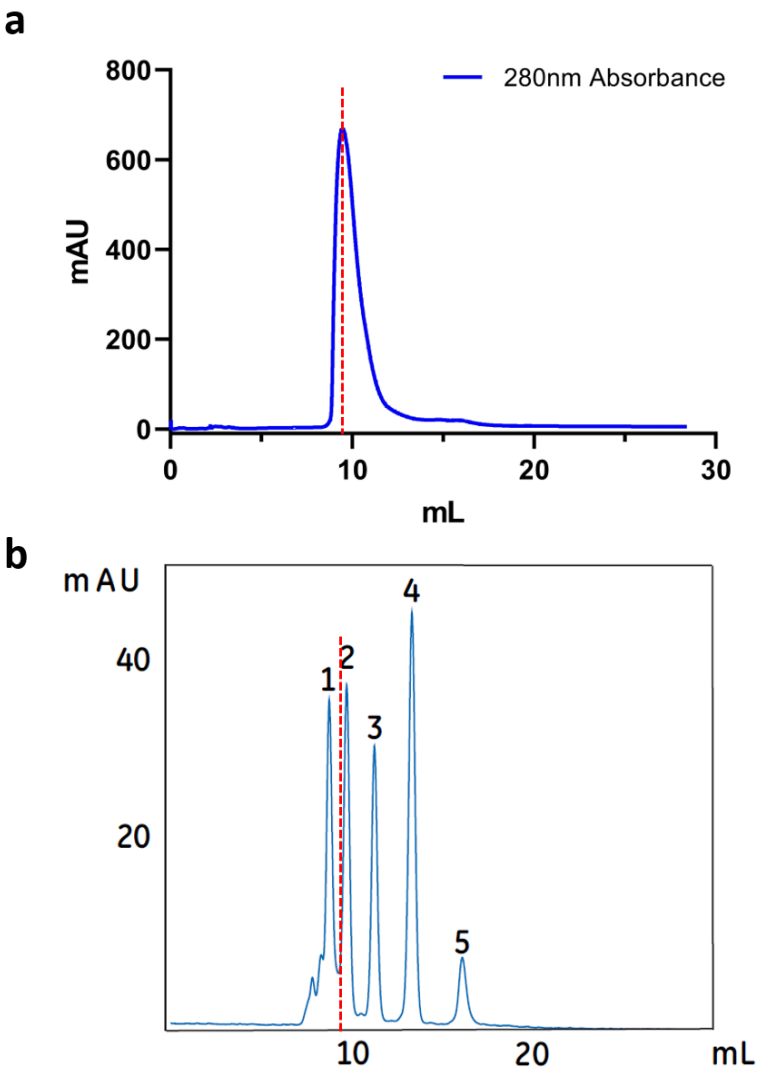


Figure A6. Comparison of NFIX-2/DNA complex SEC elution profile with that of MW markers. **a)** NFIX-2/21bp DNA complex SEC using Superdex 75 Increase 10/300 GL (GE healthcare) and ÄKTA pure 25 system. Sample volume: 100 µL, buffer: 50 mM HEPES pH 7.0, 50 mM NaCl, 2 mM DTT. The theoretical MW of the NFIX-2/DNA complex is 52.34 kDa, considering the MW of a NFIX-2 dimer (19746.13 Da), of the forward DNA strand TCTTTGGCAGGCAGCCAACCA) (6391.2 Da), and of the reverse DNA strand TGGTTGGCTGCCTGCCAAAGA (6462.2 Da). Theorical NFIX-2/DNA complex is 52.34 kDa. NFIX-2/DNA complex elution volume is about 9 mL. **b)** MW markers (GE healthcare) SEC using Superdex 75 Increase 10/300 GL (GE healthcare) and ÄKTA pure 25 system. Sample 1: Conalbumin (MW 75 000 Da) 1.5 mg/mL; sample 2: Ovalbumin (MW 44 000 Da) 4 mg/mL; sample 3: Carbonic anhydrase (MW 29 000 Da) 1.5 mg/mL; sample 4: Ribonuclease A (MW 13 700 Da) 3 mg/mL: sample 5. Aprotinin (MW 6 500 Da) 1 mg/mL A red dash line is used to show the elution volume of NFIX-2/DNA complex, which corresponds to ~50 kDa, in line with theoretical MW of the complex.

IV.7 NFIX crystallization screens

Screen name	Company	Conditions	Methods
Crystal screen I/II	Hampton Research	96-well	Vapor diffusion, microbatch
Index		96-well	
PEG/Ion		48-well	
PACT	Molecular Dimension	48-well	
MacroSol		48-well	
Proplex		96-well	
Morpheus		96-well	
JCSG		96-well	
Wizard		96-well	
JBScreen Classic I	Jena Bioscience	24-well	
JBScreen Classic II		24-well	
JBScreen Classic III		24-well	
JBScreen Classic IV		24-well	

Table A2. ***Commercial screens and methods used in NFIX crystallization trials*****.**

IV.8 Supplementary figures of the article: Structural Basis of Inhibition of the Pioneer Transcription Factor NF-Y by Suramin

Supplementary Figure S1

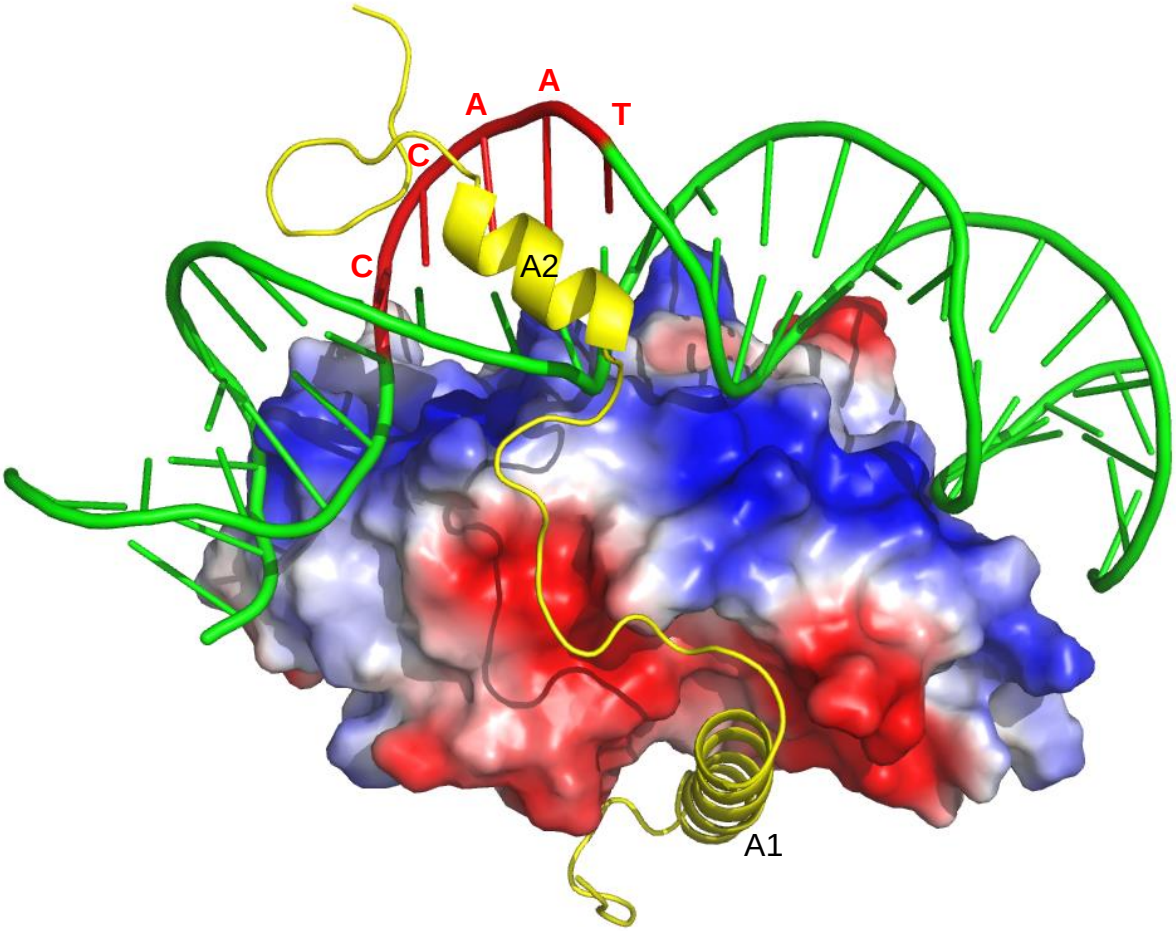


Figure S1. Electrostatic surface of NF-Yd; Blue and red colors indicate positively and negatively charged regions, respectively. NF-YA (yellow) and DNA (green) are represented in ribbon and stick models. NF-YA secondary structure elements (the A1 and A2 α helices) are labeled, and the CCAAT nucleotides highlighted in red.

Supplementary Figure S2

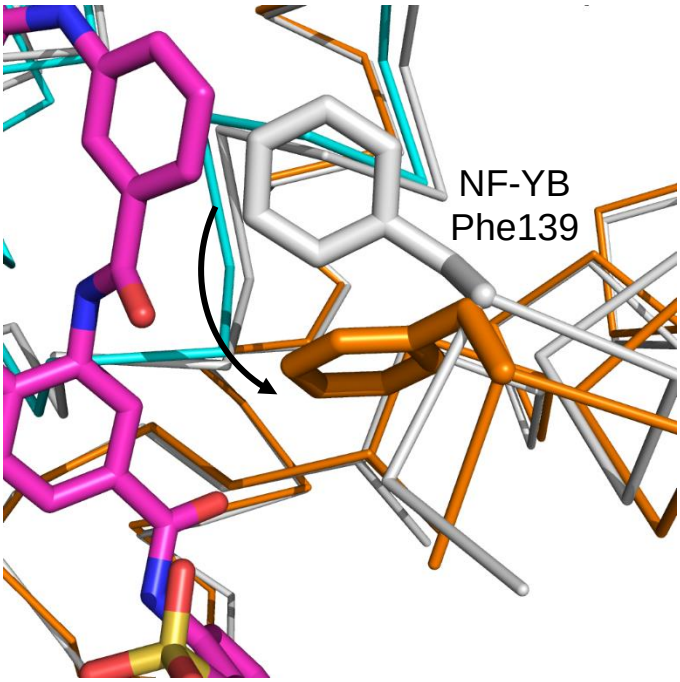


Figure S2. Structural changes induced by suramin binding; rotation (shown with an arrow) of the NF-YB Phe139 side-chain (orange) upon suramin (magenta sticks) binding, relative to its position in the ligand-free NF-Yd (PDB-code 4CSR).

REFERENCES

Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW *et al* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66: 213-221

Alevizopoulos A, Dusserre Y, Tsai-Pflugfelder M, von der Weid T, Wahli W, Mermod N (1995) A proline-rich TGF-beta-responsive transcriptional activator interacts with histone H3. *Genes Dev* 9: 3051-3066

Apt D, Liu Y, Bernard HU (1994) Cloning and functional analysis of spliced isoforms of human nuclear factor I-X: interference with transcriptional activation by NFI/CTF in a cell-type specific manner. *Nucleic Acids Res* 22: 3825-3833

Aragon E, Wang Q, Zou Y, Morgani SM, Ruiz L, Kaczmarska Z, Su J, Torner C, Tian L, Hu J *et al* (2019) Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF-beta signaling. *Genes Dev* 33: 1506-1524

Armentero MT, Horwitz M, Mermod N (1994) Targeting of DNA polymerase to the adenovirus origin of DNA replication by interaction with nuclear factor I. *Proc Natl Acad Sci U S A* 91: 11537-11541

Baburajendran N, Jauch R, Tan CY, Narasimhan K, Kolatkar PR (2011) Structural basis for the cooperative DNA recognition by Smad4 MH1 dimers. *Nucleic Acids Res* 39: 8213-8222

BabuRajendran N, Palasingam P, Narasimhan K, Sun W, Prabhakar S, Jauch R, Kolatkar PR (2010) Structure of Smad1 MH1/DNA complex reveals distinctive rearrangements of BMP and TGF-beta effectors. *Nucleic Acids Res* 38: 3477-3488

Bandyopadhyay S, Gronostajski RM (1994) Identification of a conserved oxidation-sensitive cysteine residue in the NFI family of DNA-binding proteins. *J Biol Chem* 269: 29949-29955

Bandyopadhyay S, Starke DW, Mieyal JJ, Gronostajski RM (1998) Thioltransferase (glutaredoxin) reactivates the DNA-binding activity of oxidation-inactivated nuclear factor I. *J Biol Chem* 273: 392-397

Basile V, Baruffaldi F, Dolfini D, Belluti S, Benatti P, Ricci L, Artusi V, Tagliafico E, Mantovani R, Molinari S *et al* (2016) NF-YA splice variants have different roles on muscle differentiation. *Biochim Biophys Acta* 1859: 627-638

Baxevanis AD, Arents G, Moudrianakis EN, Landsman D (1995) A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic Acids Res* 23: 2685-2691

Bouhlel MA, Lambert M, David-Cordonnier MH (2015) Targeting Transcription Factor Binding to DNA by Competing with DNA Binders as an Approach for Controlling Gene Expression. *Curr Top Med Chem* 15: 1323-1358

Bryksin A, Matsumura I (2013) Overlap extension PCR cloning. *Methods Mol Biol* 1073: 31-42

Buchwald P (2010) Small-molecule protein-protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations. *IUBMB Life* 62: 724-731

Bushweller JH. Targeting transcription factors in cancer - from undruggable to reality. *Nat Rev Cancer*. 2019 Nov;19(11):611-624

Campbell CE, Piper M, Plachez C, Yeh YT, Baizer JS, Osinski JM, Litwack ED, Richards LJ, Gronostajski RM (2008) The transcription factor Nfix is essential for normal brain development. *BMC Dev Biol* 8: 52

Caretti G, Motta MC, Mantovani R (1999) NF-Y associates with H3-H4 tetramers and octamers by multiple mechanisms. *Mol Cell Biol* 19: 8591-8603

Ceribelli M, Benatti P, Imbriano C, Mantovani R (2009) NF-YC complexity is generated by dual promoters and alternative splicing. *J Biol Chem* 284: 34189-34200

Chai N, Li WX, Wang J, Wang ZX, Yang SM, Wu JW (2015) Structural basis for the Smad5 MH1 domain to recognize different DNA sequences. *Nucleic Acids Res* 43: 9051-9064

Chaudhry AZ, Lyons GE, Gronostajski RM (1997) Expression patterns of the four nuclear factor I genes during mouse embryogenesis indicate a potential role in development. *Dev Dyn* 208: 313-325

Chaudhry AZ, Vitullo AD, Gronostajski RM (1998) Nuclear factor I (NFI) isoforms differentially activate simple versus complex NFI-responsive promoters. *J Biol Chem* 273: 18538-18546

Chen CF, He X, Arslan AD, Mo YY, Reinhold WC, Pommier Y, Beck WT (2011) Novel regulation of nuclear factor-YB by miR-485-3p affects the expression of DNA topoisomerase IIalpha and drug responsiveness. *Mol Pharmacol* 79: 735-741

Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66: 12-21

Cleat PH, Hay RT (1989) Co-operative interactions between NFI and the adenovirus DNA binding protein at the adenovirus origin of replication. *EMBO J* 8: 1841-1848
de Vries E, van Driel W, van den Heuvel SJ, van der Vliet PC (1987) Contactpoint analysis of the HeLa nuclear factor I recognition site reveals symmetrical binding at one side of the DNA helix. *EMBO J* 6: 161-168
Dekker J, van Oosterhout JA, van der Vliet PC (1996) Two regions within the DNA binding domain of nuclear factor I interact with DNA and stimulate adenovirus DNA replication independently. *Mol Cell Biol* 16: 4073-4080
Dervan PB, Edelson BS (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Curr Opin Struct Biol* 13: 284-299
Di Agostino S, Strano S, Emiliozzi V, Zerbini V, Mottolese M, Sacchi A, Blandino G, Piaggio G (2006) Gain of function of mutant p53: the mutant p53/NF-Y protein complex reveals an aberrant transcriptional mechanism of cell cycle regulation. *Cancer Cell* 10: 191-202
Dolfini D, Andrioletti V, Mantovani R (2019) Overexpression and alternative splicing of NF-YA in breast cancer. *Sci Rep* 9: 12955
Dolfini D, Gatta R, Mantovani R (2012) NF-Y and the transcriptional activation of CCAAT promoters. *Crit Rev Biochem Mol Biol* 47: 29-49
Dolfini D, Zambelli F, Pavesi G, Mantovani R (2009) A perspective of promoter architecture from the CCAAT box. *Cell Cycle* 8: 4127-4137
Donaldson LW, Petersen JM, Graves BJ, McIntosh LP (1996) Solution structure of the ETS domain from murine Ets-1: a winged helix-turn-helix DNA binding motif. *EMBO J* 15: 125-134
Dorn A, Bollekens J, Staub A, Benoist C, Mathis D (1987) A multiplicity of CCAAT box-binding proteins. *Cell* 50: 863-872
Drew HR, Travers AA (1985) DNA bending and its relation to nucleosome positioning. *J Mol Biol* 186: 773-790
Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* 43: W389-394
El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A *et al* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* 47: D427-D432
Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66: 486-501
Filippakopoulos P, Picaud S, Mangos M, Keates T, Lambert JP, Barsyte-Lovejoy D, Felletar I, Volkmer R, Muller S, Pawson T *et al* (2012) Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell* 149: 214-231
Fleming JD, Pavesi G, Benatti P, Imbriano C, Mantovani R, Struhl K (2013) NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. *Genome Res* 23: 1195-1209
Fletcher CF, Jenkins NA, Copeland NG, Chaudhry AZ, Gronostajski RM (1999) Exon structure of the nuclear factor I DNA-binding domain from C. elegans to mammals. *Mamm Genome* 10: 390-396
Forsburg SL, Guarente L (1988) Mutational analysis of upstream activation sequence 2 of the CYC1 gene of Saccharomyces cerevisiae: a HAP2-HAP3-responsive site. *Mol Cell Biol* 8: 647-654
Gee P, Xu H, Hotta A (2017) Cellular Reprogramming, Genome Editing, and Alternative CRISPR Cas9 Technologies for Precise Gene Therapy of Duchenne Muscular Dystrophy. *Stem Cells Int* 2017: 8765154
Geourjon C, Deleage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci* 11: 681-684
Gnesutta N, Saad D, Chaves-Sanjuan A, Mantovani R, Nardini M (2017) Crystal Structure of the Arabidopsis thaliana L1L/NF-YC3 Histone-fold Dimer Reveals Specificities of the LEC1 Family of NF-Y Subunits in Plants. *Mol Plant* 10: 645-648
Goda H, Nagase T, Tanoue S, Sugiyama J, Steidl S, Tuncher A, Kobayashi T, Tsukagoshi N, Brakhage AA, Kato M (2005) Nuclear translocation of the heterotrimeric CCAAT binding factor of Aspergillus oryzae is dependent on two redundant localising signals in a single subunit. *Arch Microbiol* 184: 93-100
Gonzalez-Sandoval A, Gasser SM (2016) On TADs and LADs: Spatial Control Over Gene Expression. *Trends Genet* 32: 485-495
Gounari F, De Francesco R, Schmitt J, van der Vliet P, Cortese R, Stunnenberg H (1990) Amino-terminal domain of NF1 binds to DNA as a dimer and activates adenovirus DNA replication. *EMBO J* 9: 559-566
Gronostajski RM (1986) Analysis of nuclear factor I binding to DNA using degenerate oligonucleotides. *Nucleic Acids Res* 14: 9117-9132

Gronostajski RM (1987) Site-specific DNA binding of nuclear factor I: effect of the spacer region. *Nucleic Acids Res* 15: 5545-5559
Gronostajski RM (2000) Roles of the NFI/CTF gene family in transcription and development. *Gene* 249: 31-45
Grunder A, Ebel TT, Mallo M, Schwarzkopf G, Shimizu T, Sippel AE, Schrewe H (2002) Nuclear factor I-B (Nfib) deficient mice have severe lung hypoplasia. *Mech Dev* 112: 69-77
Gu Z, Kuntz-Simon G, Rommelaere J, Cornelis J (1999) Oncogenic transformation-dependent expression of a transcription factor NF-Y subunit. *Mol Carcinog* 24: 294-299
Gurtner A, Manni I, Piaggio G (2017) NF-Y in cancer: Impact on cell transformation of a gene essential for proliferation. *Biochim Biophys Acta Gene Regul Mech* 1860: 604-616
Hagenbuchner J, Ausserlechner MJ (2016) Targeting transcription factors by small compounds--Current strategies and future implications. *Biochem Pharmacol* 107: 1-13
Harris L, Genovesi LA, Gronostajski RM, Wainwright BJ, Piper M (2015) Nuclear factor one transcription factors: Divergent functions in developmental versus adult stem cell populations. *Dev Dyn* 244: 227-238
Holm L, Rosenstrom P (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res* 38: W545-549
Holmfeldt P, Pardieck J, Saulsberry AC, Nandakumar SK, Finkelstein D, Gray JT, Persons DA, McKinney-Freeman S (2013) Nfix is a novel regulator of murine hematopoietic stem and progenitor cell survival. *Blood* 122: 2987-2996
Hsu YC, Osinski J, Campbell CE, Litwack ED, Wang D, Liu S, Bachurski CJ, Gronostajski RM (2011) Mesenchymal nuclear factor I B regulates cell proliferation and epithelial differentiation during lung maturation. *Dev Biol* 354: 242-252
Huang W, Dong Z, Chen Y, Wang F, Wang CJ, Peng H, He Y, Hangoc G, Pollok K, Sandusky G *et al* (2016) Small-molecule inhibitors targeting the DNA-binding domain of STAT3 suppress tumor growth, metastasis and STAT3 target gene expression in vivo. *Oncogene* 35: 783-792
Huber EM, Scharf DH, Hortschansky P, Groll M, Brakhage AA (2012) DNA minor groove sensing and widening by the CCAAT-binding complex. *Structure* 20: 1757-1768
Imbriano C, Gnesutta N, Mantovani R (2012) The NF-Y/p53 liaison: well beyond repression. *Biochim Biophys Acta* 1825: 131-139
Issaeva N, Bozko P, Enge M, Protopopova M, Verhoef LG, Masucci M, Pramanik A, Selivanova G (2004) Small molecule RITA binds to p53, blocks p53-HDM-2 interaction and activates p53 function in tumors. *Nat Med* 10: 1321-1328
Jackson SP, Tjian R (1988) O-glycosylation of eukaryotic transcription factors: implications for mechanisms of transcriptional regulation. *Cell* 55: 125-133
Jiang H, Bower KE, Beuscher AEt, Zhou B, Bobkov AA, Olson AJ, Vogt PK (2009) Stabilizers of the Max homodimer identified in virtual ligand screening inhibit Myc function. *Mol Pharmacol* 76: 491-502
Johnson PF, McKnight SL (1989) Eukaryotic transcriptional regulatory proteins. *Annu Rev Biochem* 58: 799-839
Kabsch W (2010) Xds. *Acta Crystallogr D Biol Crystallogr* 66: 125-132
Kim YH, Lee DH, Jeong JH, Guo ZS, Lee YJ (2008) Quercetin augments TRAIL-induced apoptotic death: involvement of the ERK signal transduction pathway. *Biochem Pharmacol* 75: 1946-1958
Kotecha M, Kluza J, Wells G, O'Hare CC, Forni C, Mantovani R, Howard PW, Morris P, Thurston DE, Hartley JA *et al* (2008) Inhibition of DNA binding of the NF-Y transcription factor by the pyrrolobenzodiazepine-polyamide conjugate GWL-78. *Mol Cancer Ther* 7: 1319-1328
Krissinel E (2015) Stock-based detection of protein oligomeric states in jsPISA. *Nucleic Acids Res* 43: W314-319
Kruse U, Qian F, Sippel AE (1991) Identification of a fourth nuclear factor I gene in chicken by cDNA cloning: NFI-X. *Nucleic Acids Res* 19: 6641
Kruse U, Sippel AE (1994) The genes for transcription factor nuclear factor I give rise to corresponding splice variants between vertebrate species. *J Mol Biol* 238: 860-865
Kulkarni S, Gronostajski RM (1996) Altered expression of the developmentally regulated NFI gene family during phorbol ester-induced differentiation of human leukemic cells. *Cell Growth Differ* 7: 501-510
Laloum T, De Mita S, Gamas P, Baudin M, Niebel A (2013) CCAAT-box binding transcription factors in plants: Y so many? *Trends Plant Sci* 18: 157-166
Lambert M, Jambon S, Depauw S, David-Cordonnier MH (2018a) Targeting Transcription Factors for Cancer Treatment. *Molecules* 23

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT (2018b) The Human Transcription Factors. *Cell* 172: 650-665
Langer G, Cohen SX, Lamzin VS, Perrakis A (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* 3: 1171-1179
LaRochelle O, Labbe S, Harrisson JF, Simard C, Tremblay V, St-Gelais G, Govindan MV, Seguin C (2008) Nuclear factor-1 and metal transcription factor-1 synergistically activate the mouse metallothionein-1 gene in response to metal ions. *J Biol Chem* 283: 8190-8201
Latchman DS (1997) Transcription factors: an overview. *Int J Biochem Cell Biol* 29: 1305-1312
Li B, Carey M, Workman JL (2007) The role of chromatin during transcription. *Cell* 128: 707-719
Li Q, Herrler M, Landsberger N, Kaludov N, Ogryzko VV, Nakatani Y, Wolffe AP (1998) Xenopus NF-Y pre-sets chromatin to potentiate p300 and acetylation-responsive transcription from the Xenopus hsp70 promoter in vivo. *EMBO J* 17: 6300-6315
Li XY, Mantovani R, Hooft van Huijsduijnen R, Andre I, Benoist C, Mathis D (1992) Evolutionary variation of the CCAAT-binding transcription factor NF-Y. *Nucleic Acids Res* 20: 1087-1091
Lin K, Simossis VA, Taylor WR, Heringa J (2005) A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* 21: 152-159
Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003a) Protein disorder prediction: implications for structural proteomics. *Structure* 11: 1453-1459
Linding R, Russell RB, Neduva V, Gibson TJ (2003b) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 31: 3701-3708
Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* 389: 251-260
Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. *Genome Biol* 2000;1(1):REVIEWS001.
Magnani L, Eeckhoute J, Lupien M (2011) Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends Genet* 27: 465-474
Maity SN (2017) NF-Y (CBF) regulation in specific cell types and mouse models. *Biochim Biophys Acta Gene Regul Mech* 1860: 598-603
Mancheno JM, Tateno H, Goldstein IJ, Martinez-Ripoll M, Hermoso JA (2005) Structural analysis of the Laetiporus sulphureus hemolytic pore-forming lectin in complex with sugars. *J Biol Chem* 280: 17251-17259
Mantovani R (1998) A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res* 26: 1135-1143
Martin-Malpartida P, Batet M, Kaczmarska Z, Freier R, Gomes T, Aragon E, Zou Y, Wang Q, Xi Q, Ruiz L *et al* (2017) Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors. *Nat Commun* 8: 2070
Martynoga B, Mateo JL, Zhou B, Andersen J, Achimastou A, Urban N, van den Berg D, Georgopoulou D, Hadjur S, Wittbrodt J *et al* (2013) Epigenomic enhancer annotation reveals a key role for NFIX in neural stem cell quiescence. *Genes Dev* 27: 1769-1786
McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. *J Appl Crystallogr* 40: 658-674
Meisterernst M, Rogge L, Foeckler R, Karaghiosoff M, Winnacker EL (1989) Structural and functional organization of a porcine gene coding for nuclear factor I. *Biochemistry* 28: 8191-8200
Mermod N, O'Neill EA, Kelly TJ, Tjian R (1989) The proline-rich transcriptional activator of CTF/NF-I is distinct from the replication and DNA binding domain. *Cell* 58: 741-753
Messina G, Biressi S, Monteverde S, Magli A, Cassano M, Perani L, Roncaglia E, Tagliafico E, Starnes L, Campbell CE *et al* (2010) Nfix regulates fetal-specific transcription in developing skeletal muscle. *Cell* 140: 554-566
Miller J, McLachlan AD, Klug A (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J* 4: 1609-1614
Morris SA (2016) Direct lineage reprogramming via pioneer factors; a detour through developmental gene regulatory networks. *Development* 143: 2696-2705
Motta MC, Caretti G, Badaracco GF, Mantovani R (1999) Interactions of the CCAAT-binding trimer NF-Y with nucleosomes. *J Biol Chem* 274: 1326-1333
Mukhopadhyay SS, Rosen JM (2007) The C-terminal domain of the nuclear factor I-B2 isoform is glycosylated and transactivates the WAP gene in the JEG-3 cells. *Biochem Biophys Res Commun* 358: 770-776
Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53: 240-255

Nagata K, Guggenheimer RA, Enomoto T, Lichy JH, Hurwitz J (1982) Adenovirus DNA replication in vitro: identification of a host factor that stimulates synthesis of the preterminal protein-dCMP complex. *Proc Natl Acad Sci U S A* 79: 6438-6442
Nardini M, Gnesutta N, Donati G, Gatta R, Forni C, Fossati A, Vonrhein C, Moras D, Romier C, Bolognesi M *et al* (2013) Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. *Cell* 152: 132-143
Nardone V, Chaves-Sanjuan A, Nardini M (2017) Structural determinants for NF-Y/DNA interaction at the CCAAT box. *Biochim Biophys Acta Gene Regul Mech* 1860: 571-580
Nilsson J, Bjursell G, Kannius-Janson M (2006) Nuclear Jak2 and transcription factor NF1-C2: a novel mechanism of prolactin signaling in mammary epithelial cells. *Mol Cell Biol* 26: 5663-5674
Novak A, Goyal N, Gronostajski RM (1992) Four conserved cysteine residues are required for the DNA binding activity of nuclear factor I. *J Biol Chem* 267: 12986-12990
O'Connor C, Campos J, Osinski JM, Gronostajski RM, Michie AM, Keeshan K (2015) Nfix expression critically modulates early B lymphopoiesis and myelopoiesis. *PLoS One* 10: e0120102
Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S, Jothi R (2014) Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. *Mol Cell* 55: 708-722
Papavassiliou KA, Papavassiliou AG (2016) Transcription Factor Drug Targets. *J Cell Biochem* 117: 2693-2696
Park JC, Herr Y, Kim HJ, Gronostajski RM, Cho MI (2007) Nfic gene disruption inhibits differentiation of odontoblasts responsible for root formation and results in formation of short and abnormal roots in mice. *J Periodontol* 78: 1795-1802
Petroni K, Kumimoto RW, Gnesutta N, Calvenzani V, Fornari M, Tonelli C, Holt BF, 3rd, Mantovani R (2012) The promiscuous life of plant NUCLEAR FACTOR Y transcription factors. *Plant Cell* 24: 4777-4792
Piper M, Barry G, Hawkins J, Mason S, Lindwall C, Little E, Sarkar A, Smith AG, Moldrich RX, Boyle GM *et al* (2010) NFIA controls telencephalic progenitor cell differentiation through repression of the Notch effector Hes1. *J Neurosci* 30: 9127-9139
Piper M, Gronostajski R, Messina G (2019) Nuclear Factor One X in Development and Disease. *Trends Cell Biol* 29: 20-30
Prado F, Vicent G, Cardalda C, Beato M (2002) Differential role of the proline-rich domain of nuclear factor 1-C splice variants in DNA binding and transactivation. *J Biol Chem* 277: 16383-16390
Priolo M, Grosso E, Mammi C, Labate C, Naretto VG, Vacalebre C, Caridi P, Lagana C (2012) A peculiar mutation in the DNA-binding/dimerization domain of NFIX causes Sotos-like overgrowth syndrome: a new case. *Gene* 511: 103-105
Priolo M, Schanze D, Tatton-Brown K, Mulder PA, Tenorio J, Kooblall K, Acero IH, Alkuraya FS, Arias P, Bernardini L *et al* (2018) Further delineation of Malan syndrome. *Hum Mutat* 39: 1226-1237
Puzianowska-Kuznicka M, Shi YB (1996) Nuclear factor I as a potential regulator during postembryonic organ development. *J Biol Chem* 271: 6273-6282
Rivera CM, Ren B (2013) Mapping human epigenomes. *Cell* 155: 39-55
Robert X, Gouet P (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* 42: W320-324
Romier C, Cocchiarella F, Mantovani R, Moras D (2003) The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. *J Biol Chem* 278: 1336-1345
Rosenfeld PJ, O'Neill EA, Wides RJ, Kelly TJ (1987) Sequence-specific interactions between cellular DNA-binding proteins and the adenovirus origin of DNA replication. *Mol Cell Biol* 7: 875-886
Rossi G, Antonini S, Bonfanti C, Monteverde S, Vezzali C, Tajbakhsh S, Cossu G, Messina G (2016) Nfix Regulates Temporal Progression of Muscle Regeneration through Modulation of Myostatin Expression. *Cell Rep* 14: 2238-2249
Rossi G, Bonfanti C, Antonini S, Bastoni M, Monteverde S, Innocenzi A, Saclier M, Taglietti V, Messina G (2017) Silencing Nfix rescues muscular dystrophy by delaying muscle regeneration. *Nat Commun* 8: 1055
Roulet E, Armentero MT, Krey G, Corthesy B, Dreyer C, Mermod N, Wahli W (1995) Regulation of the DNA-binding and transcriptional activities of Xenopus laevis NFI-X by a novel C-terminal domain. *Mol Cell Biol* 15: 5552-5562
Roulet E, Bucher P, Schneider R, Wingender E, Dusserre Y, Werner T, Mermod N (2000) Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. *J Mol Biol* 297: 833-848

Rupp RA, Kruse U, Multhaup G, Gobel U, Beyreuther K, Sippel AE (1990) Chicken NFI/TGGCA proteins are encoded by at least three independent genes: NFI-A, NFI-B and NFI-C with homologues in mammalian genomes. *Nucleic Acids Res* 18: 2607-2616
Saclier M, Lapi M, Bonfanti C, Rossi G, Antonini S, Messina G (2020) The Transcription Factor Nfix Requires RhoA-ROCK1 Dependent Phagocytosis to Mediate Macrophage Skewing during Skeletal Muscle Regeneration. *Cells* 9
Savojardo C, Fariselli P, Martelli PL, Casadio R (2016) INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics* 32: 2542-2544
Schneider TR, Sheldrick GM (2002) Substructure solution with SHELXD. *Acta Crystallogr D Biol Crystallogr* 58: 1772-1779
Secchiero P, Bosco R, Celeghini C, Zauli G (2011) Recent advances in the therapeutic perspectives of Nutlin-3. *Curr Pharm Des* 17: 569-577
Sekiya T, Muthurajan UM, Luger K, Tulin AV, Zaret KS (2009) Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev* 23: 804-809
Sheldrick GM (2010) Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr D Biol Crystallogr* 66: 479-485
Shen Y, Wei W, Zhou DX (2015) Histone Acetylation Enzymes Coordinate Metabolism and Gene Expression. *Trends Plant Sci* 20: 614-621
Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* 32: 171-178
Shi Y, Wang YF, Jayaraman L, Yang H, Massague J, Pavletich NP (1998) Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signaling. *Cell* 94: 585-594
Singh U, Bongcam-Rudloff E, Westermark B (2009) A DNA sequence directed mutual transcription regulation of HSF1 and NFIX involves novel heat sensitive protein interactions. *PLoS One* 4: e5050
Slabinski L, Jaroszewski L, Rychlewski L, Wilson IA, Lesley SA, Godzik A (2007) XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics* 23: 3403-3405
Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS (2015) Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* 161: 555-568
Stros M, Launholt D, Grasser KD (2007) The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. *Cell Mol Life Sci* 64: 2590-2606
Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y *et al* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res* 11: 677-684
Tanese N, Pugh BF, Tjian R (1991) Coactivators for a proline-rich activator purified from the multisubunit human TFIID complex. *Genes Dev* 5: 2212-2224
Testa A, Donati G, Yan P, Romani F, Huang TH, Vigano MA, Mantovani R (2005) Chromatin immunoprecipitation (ChIP) on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters. *J Biol Chem* 280: 13606-13615
Tsafou K, Tiwari PB, Forman-Kay JD, Metallo SJ, Toretsky JA (2018) Targeting Intrinsically Disordered Transcription Factors: Changing the Paradigm. *J Mol Biol* 430: 2321-2341
Ulasov AV, Rosenkranz AA, Sobolev AS (2018) Transcription factors: Time to deliver. *J Control Release* 269: 24-35
Vinson CR, Sigler PB, McKnight SL (1989) Scissors-grip model for DNA recognition by a family of leucine zipper proteins. *Science* 246: 911-916
Voogd TE, Vansterkenburg EL, Wilting J, Janssen LH (1993) Recent research on the biological activity of suramin. *Pharmacol Rev* 45: 177-203
Wingender E (2013) Criteria for an updated classification of human transcription factor DNA-binding domains. *J Bioinform Comput Biol* 11: 1340007
Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A *et al* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67: 235-242
Wintjens R, Rooman M (1996) Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. *J Mol Biol* 262: 294-313
Xu M, Osada S, Imagawa M, Nishihara T (1997) Genomic organization of the rat nuclear factor I-A gene. *J Biochem* 122: 795-801

Zawel L, Dai JL, Buckhaults P, Zhou S, Kinzler KW, Vogelstein B, Kern SE (1998) Human Smad3 and Smad4 are sequence-specific transcription activators. *Mol Cell* 1: 611-617
Zhou Y, Lee AS (1998) Mechanism for the suppression of the mammalian stress response by genistein, an anticancer phytoestrogen from soy. *J Natl Cancer Inst* 90: 381-388
Zhu H, Wang G, Qian J (2016) Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet* 17: 551-565