genes

Review

Proteins Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors

Dhanusha Yesudhas, Maria Batool, Muhammad Ayaz Anwar, Suresh Panneerselvam and Sangdun Choi *

Department of Molecular Science and Technology, Ajou University, Suwon 443-749, Korea;
dhanusha2504@gmail.com (D.Y.); mariabatool.28@gmail.com (M.B.);
ayaz@ajou.ac.kr (M.A.A.); sureshcbt@gmail.com (S.P.)
* Correspondence: sangdunchoi@ajou.ac.kr; Tel.: +82-31-219-2600

Received: 10 May 2017; Accepted: 25 July 2017; Published: 1 August 2017

Abstract: Proteins in the form of transcription factors (TFs) bind to specific DNA sites that regulate cell growth, differentiation, and cell development. The interactions between proteins and DNA are important toward maintaining and expressing genetic information. Without knowing TFs structures and DNA-binding properties, it is difficult to completely understand the mechanisms by which genetic information is transferred between DNA and proteins. The increasing availability of structural data on protein-DNA complexes and recognition mechanisms provides deeper insights into the nature of protein-DNA interactions and therefore, allows their manipulation. TFs utilize different mechanisms to recognize their cognate DNA (direct and indirect readouts). In this review, we focus on these recognition mechanisms as well as on the analysis of the DNA-binding domains of stem cell TFs, discussing the relative role of various amino acids toward facilitating such interactions. Unveiling such mechanisms will improve our understanding of the molecular pathways through which TFs are involved in repressing and activating gene expression.

Keywords: base and shape readouts; protein-DNA interaction; protein-DNA recognition; TF domain family

1. Introduction

Most biological activities are governed by multiple protein-DNA interactions. The fundamental phenomenon underlying these interactions is the process by which proteins search and recognize their specific sites on the DNA, thereby enabling the transmission of genetic information to initiate various biological processes. Over the years, theoretical and experimental advances have allowed to improve our understanding of the mechanisms by which transcription factors (TFs) search for, and recognize these binding sites. In addition, researchers have explored how TFs interact with each other and with their binding partners. Although significant progress has been achieved toward understanding the TF search process, the details of this mechanism remain controversial [1,2].

One of the most puzzling phenomena involved in protein search over DNA is the effect of multiple targets, which is particularly important in eukaryotic genomes. Eukaryotic genomes harbor multiple target sites between tightly bound nucleosome core particles on accessible DNA fragments [1]. Recent studies showed that single nucleotide changes can alter TF selectivity, and also influence the sequence of events culminating in the TF binding with its true recognition site. Additionally, other major hurdles faced by TFs regarding their selectivity include the existence of cellular networks, dynamic protein-DNA conformational changes, and tight packing of multiple TFs at the regulator sites of a single DNA section [3]. These factors affect the complexity of protein-DNA recognition processes at both sequence and structural levels, meaning nucleotide sequences and their resulting 3D structures. Furthermore, other factors such as TFs' flexibility for their binding sites, the influence of cofactors, cooperative binding of other TFs, DNA methylation, and other epigenetic modifications add to the complexity of this process. The effect of nucleosomes and their binding with TFs, chromatin accessibility, and nucleosome occupancy will also have an impact on TF-DNA readouts [4]. Along with the nucleosomes, the distribution of sequence-specific TFs (cell-specific and tissue-specific, but also ubiquitous) also greatly affects TF binding. Recent studies show that realistic observations about the readout mechanism vary across the various protein families [4–6]. Most of these readout mechanisms are discussed in this review.

Proteins use a wide range of DNA-binding structural motifs, such as homeodomain (HD), helix-turn-helix (HTH), and high-mobility group box (HMG) to recognize DNA. HTH is the most common binding motif and can be found in several repressor and activator proteins. Despite their structural diversity, these domains participate in a variety of functions that include acting as substrate interaction mediators, enzymes to operate DNA, and transcriptional regulators [7]. Several proteins also contain flexible segments outside the DNA-binding domain to facilitate specific and non-specific interactions. The phage Φ29 transcriptional regulator p4 uses its N-terminal beta-turn substructure for specific contact with DNA [8]. Likewise, HD proteins use N-terminal arms and a linker region to interact with DNA; for example, λ repressor uses its N-terminal arm to make contact with the major groove [9]. The Encyclopedia of DNA Elements (ENCODE) data suggest that about 99.8% of putative binding motifs of TFs are not bound by their respective TFs in the genome [10,11]. It is, therefore, clear that the presence of a single binding motif per TF is not adequate for TF binding.

Over the past decades, developments in computational and structural biology have offered an immense potential toward studying the protein-DNA recognition code. Crystal structures of protein-DNA complexes were first solved in the 1980s [12], and more than 1600 protein–DNA structures have since been deposited in the Protein Data Bank (PDB) [4]. This plethora of information has helped us to conclude that preferential binding of a TF to its cognate site is purely based on its physical interactions, for instance, the physical interaction between the amino acid side chain of the TF and the atoms of DNA base pairs [13]. Most of these physical interactions rely on hydrogen bonds, as well as on hydrophobic and water-mediated contacts. Other mechanisms driving protein-DNA interaction involve the recognition of DNA structural features by proteins; these structural features include the DNA major and minor grooves, backbone features, intrinsic curvature, hydration shells, as well as flexibility of DNA bending [14] and unwinding [15]. The dynamic behavior of DNA structure mostly governs the binding properties, and that can be understood through computational techniques [4,16]. Theoretical studies, such as molecular dynamics (MD) simulations, can provide additional information toward understanding protein-DNA complexes. The monitored dynamic movements of atoms reflect the functional and structural phenomena undergone by proteins or DNA during the initial phase of complex formation.

We have divided this review into two sections. The first section briefly discusses the DNA-recognition mechanisms, including historical mechanisms. The second section summarizes the major DNA-binding protein domains with reference to stem cell factors and their families. This section includes the structural properties of stem cell factor DNA-binding mechanisms and the cooperative binding phenomena driving target gene expression. Since stem cell factors are promising targets in the growing regenerative medicine field, researchers will benefit from the structural aspects of these factors provided in this review.

2. Binding Site Recognition and TFs

Several mechanisms have been proposed to describe how TFs find their target sites on DNA. One of the main scenarios involves a 'sliding' mechanism, in which the protein moves from its initial non-specific site to its actual target site by sliding along the DNA (also known as 1-dimensional (1D) sliding) (Figure 1). The binding of the lactose (*lac*) repressor to non-operator sequences is an ideal example of sliding, since its DNA-binding entirely relies on electrostatic interactions, and consequently, diffusion occurs on an isopotential surface [17–19]. When the TF starts to move and shift counterions from the phosphate backbone, the same number of counterions binds to the site left free by the protein. The detailed sliding mechanism is explained later in this section. The sliding rate is also dependent on the hydrodynamic radius of the protein; the required rotational movement over the DNA backbone is greater for larger proteins, that tend to slide slowly [20,21]. Only a few DNA-binding proteins using the sliding mechanism from non-specific to specific binding have been structurally solved, among these, are BamHI [22], λ-repressor [23], and the lactose repressor [19,24].

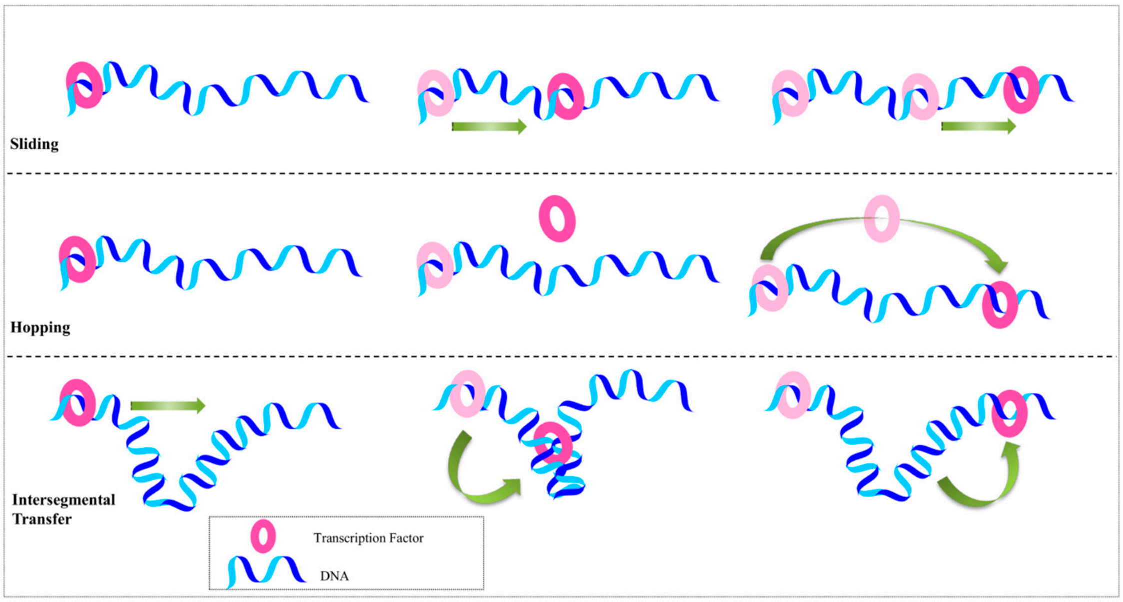
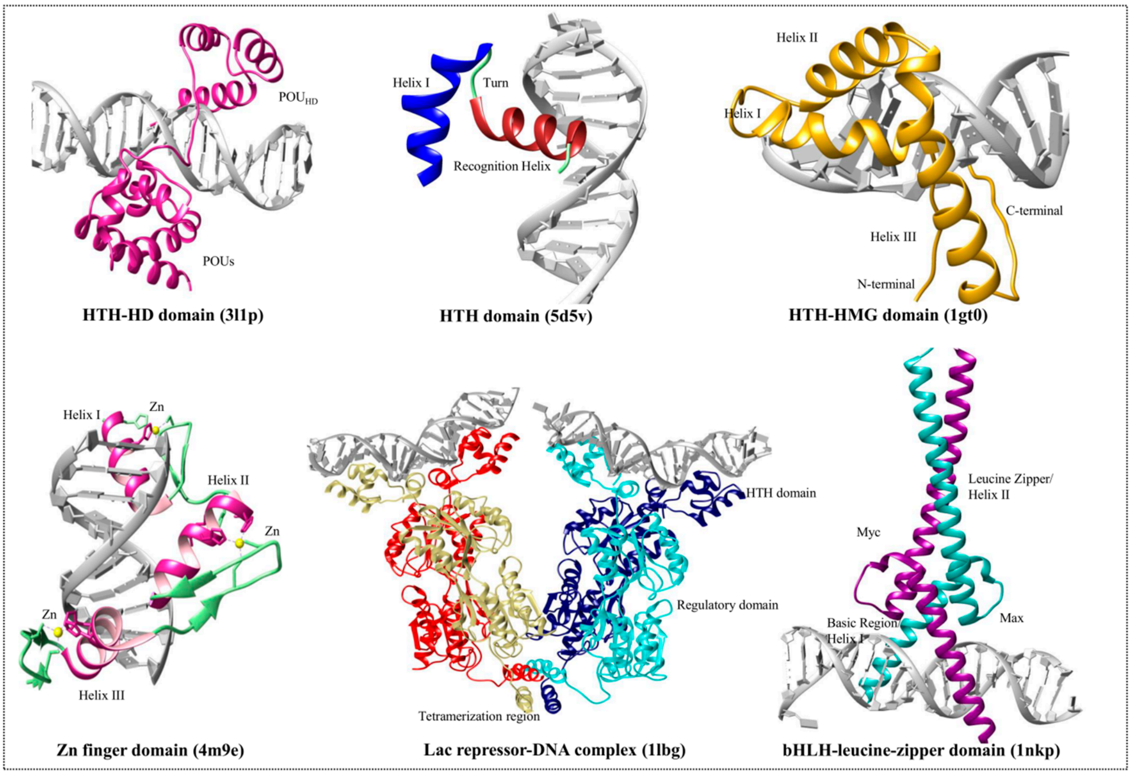


Figure 1. Protein-DNA recognition mechanisms. The main three protein-DNA recognition mechanisms are shown. When the transcription factor (pink ring) moves from one site to another by means of sliding along the DNA and is transferred from one base pair to another without dissociating from the DNA, this mechanism is called sliding (top). Hopping occurs when the transcription factor moves on the DNA by dissociating from one site and re-associating with another site (center). Intersegmental transfer describes the mechanism by which the transcription factor gets transferred through DNA bending or the formation of a DNA loop, resulting in the protein being bound transiently to both sides and subsequently moving from on site to the other (bottom).

The second scenario is a 'hopping' mechanism, in which a TF might hop from one site to another in 3D space by dissociating from its original site and subsequently binding to the new site. This may happen within the same chain and re-association occurs adjacent to the former dissociated site. For example, in the case of Herpes Simplex Virus Type 1 UL42, the binding between protein and DNA is mostly governed by electrostatic interactions, and its affinity is higher than that of the *lac* repressor [17]. Analyses indicated that non-electrostatic interactions play a key role toward allowing the binding of UL42, as well as facilitating its dissociation from the initial site. When condensation of counterions occurs at high salt concentrations, UL42 becomes more mobile and diffuses faster than at low salt concentrations [17]. Winter et al. [18] suggested that hopping should be slower than sliding, as its mechanism relies on dissociation followed by re-association to a different site. The hopping mechanism describes the search for DNA by proteins through the loss of electrostatic and non-electrostatic interactions, involving a microscopic dissociation constant (Figure 1).

A third search mechanism, proposed by Berg and von Hippel, is described as 'intersegmental transfer' [25]. In this scenario, the protein moves between two sites via an intermediate 'loop' formed by the DNA and subsequently bind at two different DNA sites (Figure 1). This mechanism is applicable to TFs with two DNA-binding sites (e.g., *Lac* repressor or SfiI endonuclease) [26]. Proteins with two DNA-binding sites can occasionally bind non-specifically to two locations situated far apart within the DNA strand, that are brought into close contact through the formation of these loops. Such TFs transfer across a point of close contact without dissociating from the DNA [27]. The EcoRV restriction enzyme binds to two specific or non-specific DNA sites in a deep cleft between two protein subunits, where the cleft is moderately narrow, in order to hold both duplexes simultaneously [28]. The search for a target site by a protein is accelerated through diffusion along the DNA strand [29]. This 3D diffusion can be assisted via the formation of a DNA loop allowing the protein to bind to two DNA segments simultaneously and thereby, enabling its transfer from one location of the segment to the other [30]. This intersegmental transfer accelerates the search for the DNA target site because this mechanism is based on constantly changing random configurations [25,31]. The impact of DNA conformation on target site searching is still unknown, but in the case of EcoRV restriction enzyme, the target finding rate is almost doubled when the DNA changes its conformation from a fully extended structure to a coiled structure [32]. Most of the searching mechanism studies are limited to naked DNA-protein complexes, which do not reflect the actual crowded environment of a cell. Studies have shown that many DNA-binding proteins travel a long distance by 1D diffusion. The search process for eukaryotes must occur in the presence of chromatin, which has the ability to hinder protein mobility. In this case, the protein must dissociate from the DNA, enter a 3D mode of diffusion state, and continue the target site searching process [33].

The sliding and intersegmental transfer mechanisms can be explained through the example of the *lac* repressor. The *lac* repressor contains 4 identical monomers (a dimer of dimers) for its DNA-binding. The binding sequence of these dimers is symmetric or pseudo-symmetric, and each half is identified by these identical monomers [34–36]. The HTH domain of the *lac* repressor is the DNA-binding domain that facilitates the interaction with its target site on DNA (Figure 2). As a result of a rapid search (sliding) along the DNA molecule and intersegmental transfer between distant DNA sequences, the lactose repressor finds its target sites faster than the diffusion limit. The section comprised between residues 1–46 of the HTH domain, characterized by three α-helices, maintains its secondary structure through specific and non-specific binding. When the repressor binds to a non-specific site, the HTH domain interacts with the DNA backbone and maintains the interaction with its helix2 region in the major groove juxtaposition. This arrangement facilitates the interaction of helix2, the recognition helix, with the edges of the DNA bases, enabling the repressor to walk or search for its specific site on the DNA. The C-terminal residues of the DNA-binding domain, residues 47–62, form the hinge region, and are normally disordered during non-specific recognition; however, during specific site recognition, residues 50–58 acquire an α-helix configuration (hinge helix) [34]. The disordered hinge region and the flexibility of the HTH domain allow the protein to move freely along the DNA to search for its target site. In specific binding complexes, the hinge helix of each monomer is located at the symmetrical center of the binding site, thereby causing the hinge helices to interact with each other (intersegmental transfer) to allow better stability. Moreover, DNA bends at the symmetrical center of the specific binding site (37° angle), thereby supporting monomer-monomer interactions [24,35]. Experimental reports suggest that engineering a disulfide bond between the hinge helices of these monomers (Val52Cys mutation) will restrict the HTH domain movement of these monomers and thereby yield higher-affinity complexes [37].



HTH-HD domain (3l1p) **HTH domain (5d5v)** **HTH-HMG domain (1gt0)**

Zn finger domain (4m9e) **Lac repressor-DNA complex (1lbg)** **bHLH-leucine-zipper domain (1nkp)**

Figure 2. Representative figures of the transcription factor binding domains. The figure shows the crystal structures of different types of TF domains (3l1p, 4m9e, 5d5v, 1lbg, 1gt0, and 1nkp). The structures were obtained from the Protein Data Bank (PDB) and redrawn using chimera. The respective domains and important regions have been labeled. HTH stands for helix-turn-helix domain. bHLH stands for basic helix-loop-helix motif. HD and HMG stand for homeodomain and high-mobility group box domain, respectively.

2.1. Historical Mechanism: Base Readout vs. Shape Readout

The direct (physical interaction) and the indirect (alteration of the DNA shape) readout mechanisms are known as the historical mechanisms that drive protein-DNA interactions. Direct recognition occurs when the amino acid side chains of a protein interact with specific DNA bases [6]. Most protein-DNA interactions are mediated by direct physical interaction (hydrogen bonding or hydrophobic interactions) between the protein and the DNA base pairs. Specific binding is mainly obtained via hydrogen bonding between the protein and the major groove base pairs, because each of the four base pairs has a unique pattern of hydrogen bond donors and acceptors in the major groove [38]. The specificity of DNA-binding also depends on the number of hydrogen bonds existing between the protein and the major groove base pairs. Bidentate bonds (two hydrogen bonds with different donor and acceptor atoms) have a higher degree of specificity than bifurcated hydrogen bonds (two hydrogen bonds sharing the same donor). A normal single hydrogen bond does not contribute to specificity, whereas bidentate bonds do [13,39]. Hydrogen bonding between proteins and DNA is also facilitated by water molecules. Highly ordered water molecules mediate the specific base pair readouts in the major groove because they reflect the position of hydrogen bond donors and acceptors at the base edges. For example, the RXR/retinoic acid receptor (RAR)-DNA complex utilizes several lysine and arginine residues to mediate the specific readouts [40]. In the case of the *lac* repressor, the protein-DNA interface is enriched with water molecules when it binds non-specifically; however, this interface is devoid of water when it binds to specific sites [13,19].

Although the base readout exists in all protein-DNA complexes, the structure of bound DNA frequently deviates from its standard one. Often, these deviations also contribute to specific

DNA-binding. For instance, papillomavirus E2 protein and the TATA box binding protein (TBP) both induce some degree of deformation in their cognate DNA to facilitate hydrogen bonding and non-polar interactions with the protein [41,42]. Such indirect methods for protein-DNA recognition are also known as shape readout (indirect readout) mechanisms, in which the binding relies on the base pairs that do not directly contact the protein, but instead, create structural changes within the DNA to facilitate recognition. The elusive conformational shift occurring in biomolecules is the transition among B, A, and Z conformations of DNA [43]. B-form DNA is the most favored conformation for a protein-DNA complex under physiological conditions, whereas A-form DNA is induced locally in some complexes under low water activity. The important factors driving this transition are hydration and electrostatics; however, solvent conditions, counterions condensation, and free energy contribution from phosphate-phosphate repulsion also contribute to this transition [44,45].

DNA bending and kinks are parameters contributing to the shape readout mechanism. DNA kinks result from the complete or partial loss of stacking energies at a single base pair step. Since kinks can occur at any individual base pair step, the adjacent region maintains its B-form conformation. Kinks and bending are stabilized upon protein binding, which compensates for the lack of base pair stacking energies. This energy compensation, resulting from the interaction of protein side-chain hydrophobic residues, makes the protein-DNA complex more stable [13]. Major/minor groove width also plays an important role in protein-DNA-binding. The differences in the hydrogen-bonding pattern of each base pair are due to different stacking energies and therefore, the stacking energies in each dinucleotide step affect the minor groove width [6]. Roll, helical twist, and propeller twist are the rotational parameters that determine the contraction of the minor groove. ApT base pair steps have negative roll angles, causing compression at the minor groove and favoring protein binding [46]. In A-DNA, ApT and ApA exhibit negative roll values and have bifurcated hydrogen bonds (A:T base pair) that lead to propeller twisting and enhance minor groove narrowing [47]. The minor groove of B-DNA shows more electronegative potential than the major groove because perfect B-DNA has a wide, shallow major groove and a narrow, deep minor groove. Likewise, the AT-rich DNA sequence in the minor groove exhibits greater electronegative potential than GC-rich sequences. Thereby, the AT-rich sequences at the minor groove attract the protein and create a bend/kink to enhance interactions. Therefore, high-affinity binding site sequences are more bendable at the minor groove than low-affinity sites, which are straight (sometimes bent into the major groove) and rigid [48]. Hence, DNA curvature and flexibility are also main parameters to be considered when determining the affinity of protein-DNA interactions.

2.2. Beyond the Recognition Mechanism

Other than the defined mechanisms, there are other factors that contribute to protein-DNA recognition. The majority of TFs possess an intrinsically disordered (ID) region, which is known to promote proteins' recognition of specific binding sites. ID regions play important roles in the transition from non-specific to specific binding, and facilitate protein diffusion along the DNA [49]. These regions also control selectivity by forming specific interactions and folding into functional forms, similarly to globular proteins [50]. The tails and linear regions of IDs enhance specific binding affinity, as the charged tails provide constant electrostatic interactions with the available base pair atoms, thereby helping the protein to find and attach to its binding site [50,51]. As a result, the availability of free proteins and the rate of protein diffusion are reduced. The linker region in IDs bridges between multidomain proteins and increases the rate of intersegmental transfer [49,52]. Another interesting fact about ID regions is that without their 3D structural information, their functional sites can be predicted from their primary sequences [53,54]. Moreover, protein-protein interactions use ID regions by targeting small motifs [55]. Intramolecular contacts of ID segments often regulate DNA-binding affinity via a competitive binding mechanism. For example, the transcriptional activation of p53 depends on its interaction with a 70-kDa subunit of human replication protein A (hRPA70), which contains weak- and high-affinity DNA-binding domains [56]. These domains are connected by the intrinsically unstructured linker domain (IULD). The ID region increases the concentration of weak-affinity DNA-binding domains, which in turn, causes a barrier between p53 and hRPA70. The hRPA70 affinity for single-stranded DNA (ssDNA) is greater than that of p53; therefore, depending on the ID region length, flexibility, and orientation, the hRPA70 IULD regulates the binding of p53 to DNA [50]. Alternative splicing modulates DNA-binding by altering ID regions. Alternative splicing affects the number of disordered regions in the ID segments and inserts or deletes post-translational modification sites to improve protein-protein interactions. Finally, ID regions have also been proposed to amplify signals and mediate allosteric responses [50,57].

Cooperative binding is yet another facet of DNA recognition. For instance, HMG box 1 protein (HMGB1) participates in several processes by binding to DNA through two HMG boxes [58]. Cooperative binding prevails when the protein-protein interaction between adjacently bound TFs stabilizes the complex and enhances the transcription activity of both proteins. However, cooperative binding might alter the specificity by extending TF binding [59]. Yeast TFs (MATa2, MATa1, and Mcm1) regulate the genes responsible for mating. MATa2 functions to recruit other cofactors and repress gene expression; nonetheless, its DNA-binding specificity is driven by the cooperative binding of cell type-specific cofactors [60,61]. The haploid-specific genes are repressed by MATa2:MATa1 heterodimers, whereas MATa2:Mcm1 heterotetramers repress the mating-type a-specific genes. The recruitment of co-repressor proteins Tup1 and Ssn6 by MATa2 represses the gene expression of both complexes. In most cases, the cooperative interactions are perpetuated through direct physical contact, but the allosteric effects provoked by DNA structural deformation also contribute to the cooperativity [59,62].

During the search process, only the physical aspect of protein-DNA interaction is addressed, whereas the biological aspects (assisted diffusion, involvement of other TFs, etc.) are more complex and mostly unknown. Although the non-specific protein-DNA binding implies a weaker interaction, their binding mainly mediates the target protein activities. Usually, when the protein encounters DNA, it reaches a random non-specific location on the DNA by diffusion, and then, uses the intramolecular translation process (sliding, hopping, or intersegmental transfer) to search its specific position [63]. The final stage is governed by the formation of specific hydrogen and electrostatic interactions at the protein-DNA interface, resulting in a precise geometrical fit between the protein and its consensus DNA [64]. Therefore, binding to the target site may establish a free energy barrier between non-specific and specific binding conformations [65].

Hydrophobic interaction is the major energetic factor contributing to the protein-DNA complex formation, which is the outcome of bound water molecule release from non-polar surfaces. Ion pairs are also contributing to the protein-DNA complex thermostability through allosteric effects. To fully understand the kinetic and thermodynamic natures of protein-DNA complexes, the salt-independent part of the total electrostatic free energy should be physico-chemically interpreted [66]. Furthermore, to characterize the thermodynamic properties of a biomolecular complex, the changes in enthalpy and entropy have to be calculated among the equilibrium states of unbound and bound conformations. Moreover, single-base variation in the DNA sequence greatly impacts the equilibrium of protein-DNA complexes and consequently, alters thermodynamic properties. Therefore, this phenomenon should also be taken into account when conducting thermodynamic calculations [67].

3. DNA-Binding Domain Families

In this section, we discuss the major DNA-binding protein domains of stem cell factors (Oct4, Sox2, Nanog, c-Myc, and Klf4) and their related families. The major DNA-binding domains, homeodomains, are described using Oct4 and Nanog as examples, whereas high mobility group domains are explained with Sox2, a stem cell transcription factor. The helix-loop-helix (HLH) domain is explained taking c-Myc as an example and the zinc finger (ZF) domain is described with Klf4 protein. Apart from these stem cell TFs, several other proteins belong to main domain superfamilies', and their DNA domain architectures (based on overall secondary structure contents) are summarized in Table 1.

Table 1. Different domain family proteins and their domain architectures.

No	Superfamily Proteins	Domain Motifs	Architecture of DNA-Binding Domains	Representative PROTEIN
1	Winged HTH proteins	Helix-turn-helix	mainly α	hRFX1
2	GCM domain	β-sheet	mixed α/β	WRKY transcription factor
3	Zinc-coordinating proteins	Zinc finger	mixed α/β	SIP1, FOG, Msn2p, A20, Klf4
4	$\beta\beta\alpha$ Zinc-finger family	Zinc finger	mixed α/β	Egr-1
5	Loop-sheet-helix family	Helix-turn-helix	mainly α	p53
6	Leucine zipper family	Helix-loop-helix	mainly α	Jun, Fos
7	POU domain	Helix-turn-helix	mainly α	Oct1, Oct2, Oct4
8	Copper-fist	Zinc finger	mixed α/β	Mac1
9	Histone-fold	NA	mainly α	TBP, TAF proteins, HuCHRAC
10	ETS domain	Helix-turn-helix	mainly α	pointed-P2
11	Bet v1-like	NA	mixed α/β	VASt
12	P-loop domain	NA	multidomain, mixed α/β	ARTS
13	TEA domain	NA	NA	Simian virus 40 (SV40), enhancer factor TEF-1
14	LytTR domain	NA	NA	AlgR/AgrA/LytR family of transcription factors
15	Steroid receptor	Zinc finger	mixed α/β	NA
16	p53-like transcription factors, E-set domains, and Runt domain proteins	Immunoglobulin-like β-sandwich motif	mainly β	NF-κB and Rel
17	TATA-box binding protein-like	TBP (TATA-binding protein) β-sheet	mainly β	HMGB1, HMGB2
18	DNA/RNA polymerases	NA	multidomain, mixed α/β	RNA polymerase I, II, III, IV and V
19	Ribbon-helix-helix	Ribbon-helix-helix	mixed α/β	CopG, NikR, ParG
20	HMG-box	Helix-turn-helix	mainly α	TCF-1, SRY
21	IHF-like DNA-binding proteins	NA	mixed α/β	HBsu
22	RNase A-like	NA	mixed α/β	Train A
23	TrpR-like	Helix-turn-helix	mainly α	TrpR like proteins
24	T4 endonuclease V	Helix-turn-helix	mainly α	RuvC protein
25	ARID-like	Helix-turn-helix	mainly α	SWI-SNF complex protein p270

The superfamily proteins were taken from the structural classification of proteins (SCOP) database and the information was retrieved and updated from Rohs's work [13]. NA: Not available.

3.1. Helix-Turn-Helix Motif

Several motifs combine to form a compact globular structure called a domain; therefore, a motif is believed to be incapable of folding and forming a stable structure, whereas a domain can. HTH is the simplest motif with two α-helices connected by a turn. The dimeric Arc repressor [68] and 3-helix bundle homeodomains [7] are examples of the simplest proteins with HTH motifs [69]. Generally, the HTH proteins dimerize, and each monomer identifies one side of a symmetric DNA sequence [70]. The HD and the HMG domain proteins fall under the same structural arrangement of DNA-binding motif and belong to the HTH superfamily. Both families have HTH in their DNA-binding motifs and exhibit similar behaviors. The recognition helix (mainly the second helix) of HTH motifs binds to DNA bases (especially at major groove) through hydrogen and hydrophobic interactions, whereas the other helices are involved in maintaining protein-DNA stability. Even though the HTH motifs are conserved, their orientation relative to DNA-binding and their structural context differ among different protein families [13]. Outside of the HTH motif, the remaining structures are distinct among all protein families. The representation of the domain families with the referenced crystal structures are shown in Figure 2.

3.2. High-Mobility Group Protein Families

The HMG box domain was initially recognized as a domain that mediates DNA-binding in chromatin-associated proteins of the HMGB type. Later, this domain was observed in other TFs and subunits of chromatin remodeling complexes. This domain can mediate non-sequence-specific (HMGB-type proteins) and sequence-specific (TF like Sox2) DNA-binding [71].

HMG proteins are abundant and ubiquitous nuclear proteins that bind to nucleosomes and induce chromatin structural changes. These non-histone proteins play important roles in transcription, replication, recombination, and DNA repair processes. Structural changes in chromatin are maintained by HMG superfamily proteins. Members of these families are highly expressed in eukaryotic cells; they acquire different structures and unique motifs to bind and affect chromatin fibers by transiently interacting with nucleosomes. HMG protein binding is highly dynamic, and is not confined to a particular site; mostly, these proteins associate in a 'hit and run' fashion [72]. HMG proteins regroup three superfamilies, HMGA, HMGB, and HMGN, characterized by the presence of acidic amino acid-rich C-termini. The domain structures of these HMG family proteins include either an AT hook (HMGA), a HMG-box domain (HMGB), or a HMG-nucleosomal binding domain (HMGN). However, each protein family has an exclusive functional motif that brings specific changes in its DNA-binding site and participates in distinct cellular functions [73]. Several reviews offer wide information on the structure and architectural functions of HMG family proteins [74–77].

HMGB proteins (HMG-Box1 and HMG-Box2) share over 82% sequence identity, and are most abundant in the nucleus and highly conserved [78]. The HMGB superfamily has a unique DNA-binding domain (HMG-box) composed of about 75 residues, which can bind to DNA with high affinity and cause structural deformations, such as bending, kinking, looping, and unwinding [79]. Every HMGB family protein contains two functional HMG box motifs and a highly acidic C-terminal end. The functional motif of HMGB is built by three α-helices folded into an L-shaped structure, and has the capacity to penetrate the DNA minor groove and sharply bend it. Minor differences between the HMG boxes confer specificity to different HMGB proteins, whereas the acidic tails modulate their affinity and contribute to nuclear localization. The protein-protein interactions exhibited by these HMGB domains are mediated by the C-termini (including the helix3 and tail region), but this region does not take part in DNA contact.

Sox2 is one of the HMGB (79 amino acid residues) TFs that binds to DNA in a sequence-specific manner [80]. The HMGB domain-containing proteins bind at the major groove. In contrast, the unique HMG domain of Sox2 binds at the minor groove of DNA and produces a bend (approximately 90°) in the DNA to improve affinity (Figure 2). Usually, if a protein carries more than one HMG domain (e.g., HMGB1, HMGB2, and upstream binding factor) it has a decreased DNA-binding specificity (non-specific binding), whereas the proteins with single HMG domain will bind in a sequence-specific manner (e.g., Sox proteins and T-cell factor/lymphoid enhancer factor family proteins) [81,82]. The sequence-specific Sox2-DNA interaction is mediated by numerous base pair-specific hydrogen bonds and is closely related to Sry/DNA interactions. Asn8, Ser31, Ser34, and Tyr72 are responsible for important protein-DNA interactions. Upon Sox2 binding, the side chain residues of helix1 and helix2 are inserted at the consensus sequence (CTTTGTT) and forces the DNA unwind and open. The C-terminal tail is unstructured in Sox proteins and is specific to these proteins, promoting interaction with DNA [83].

3.3. Homeodomain Proteins

The homeodomains are small, conserved domains with three helices containing approximately 60 amino acid residues, often found with additional flanking domains and cofactors. They have a common fold arrangement; helix1 and helix2 are antiparallel to each other while helix3 lies across both. In HD proteins, the third helix is considered to be the recognition helix and is responsible for interacting with DNA at the major groove. Based on 103 *Drosophila* homeobox genes, 16 and 11 major classes have been identified to date in animals and plants, respectively [84]. The sequence specificity of this domain is governed by the major groove interactions, and the minor groove contacts contribute to binding strength. Arginine, particularly at the 5th position of HD, is found in 99 of the 106 *Drosophila* HDs, and is preferentially found in narrow minor grooves [6]. Different HD domain families differ in their N-terminal arm regions [84], are rich in basic amino acids, and have a hydrophobic residue at the 8th position (start of helix1). In the helix1 region, the 16th and 20th positions are often occupied by hydrophobic residues. The amino acid residues Trp48, Phe49, Asn51, and Arg53 of helix3 are highly conserved within the HD region. The conservation of these residues promotes DNA-binding and overall stability [85,86].

HD domains bind to the genome in a context-dependent and cell/tissue-specific manner. They can drive lineage-specific transcription by recruiting the ubiquitous and tissue-specific TFs. For example, cone-rod homeobox (Crx), a retina-specific TF, recruits ubiquitous myocyte enhancer factor 2D (MEF2D) away from the canonical MEF2D binding site and redirects it to retina-specific enhancers [87]. The reprogramming activity of HD is observed in Oct4 [88] and Pax7 [89] proteins. HD proteins have the capacity to bind closed chromatin structures and enable co-activator binding to promote gene expression.

Although the HDs are conserved class-specific domains [84], they also possess 5–10 residues-long short linear motifs (SLiMs) that are involved in weak and transient interactions [90]. SLiMs are located within the long, disordered regions of the TFs, thereby playing an important role in interactions with a wide range of cofactors in a context-dependent manner [91]. However, the long and disordered regions pose a challenge for the specificity of HD proteins; the increasing number of contacts can displace the HD proteins from their binding sites. However, TFs with HD deletions and SLiM mutations show aberrant and allelic activities, explaining the importance of HD domains and their disordered region motifs.

The stem cell TF Nanog is another example of an HD domain protein, whose crystal structure has been published recently [92]. Nanog has a central HD domain composed of 60 amino acid residues, which is highly conserved in *Hox* genes, and preferentially binds at TAAT(G/T)(G/T). Along with the HD domain, Nanog (305 amino acids) possesses a serine-rich N-terminal domain (ND) and a C-terminal domain (CD) including the tryptophan repeats (WR) motif [93] (Figure 3). The HD domain has an unstructured N-terminus and three α-helices, with helix3 determining the binding specificity to the consensus sequence. Nanog's HD domain is distinct from those observed in other HD proteins because it displays variant residues adopting non-canonical conformations during the interactions [93]. Helix1 and helix2 interact with the minor and major grooves, respectively, whereas helix3 interacts with the DNA backbone and helix1. Tyr119, Leu122, Gln124, and Lys137 are important residues for the Nanog-DNA interaction. Mutational studies showed that mutating residues Lys137, Thr141, Asn145, and Arg147 to alanine abolished the HD domain-DNA interaction [92].

Oct4, a POU family protein, is composed of two HTH DNA-binding domains known as POU-specific (POU$_S$) and POU-homeodomain (POU$_{HD}$) [94]. POU$_S$ is only present in POU factors and is more conserved than the POU$_{HD}$ domain, which is distantly related to the classic HD proteins. The POU$_S$ domain has two short α-helices, which bind at the left half of the octamer motif through its HTH. The DNA-POU$_S$ domain interaction is mediated by the third α-helix [95]. These bipartite subdomains (POU$_S$ and POU$_{HD}$) are connected by a linker region of variable length [95]. The unique variable linker region controls the reprogramming efficiency of Oct4, which has an α-helix (α5) structure [96], and helps to recruit other epigenetic factors (such as Sox2, Nanog, and others) to Oct4 for the reprogramming process [97]. The first five residues of the N-terminal arm of the POU$_{HD}$ domain contain either lysine or arginine. This region fits into the minor groove and interacts with the 5$'$ end of its DNA-binding site [98]. The recognition helix of POU$_{HD}$ is highly conserved with a unique cysteine residue at the 50th position, a characteristic that identifies the family members across the diverse phyla [85]. Although POU$_{HD}$ domains are related to the classic HD domains, their DNA recognition mechanism is distinct and exhibits inefficient DNA-binding. Efficient DNA-binding is driven by the cooperative binding of POU$_S$ and POU$_{HD}$ at the major groove [97].

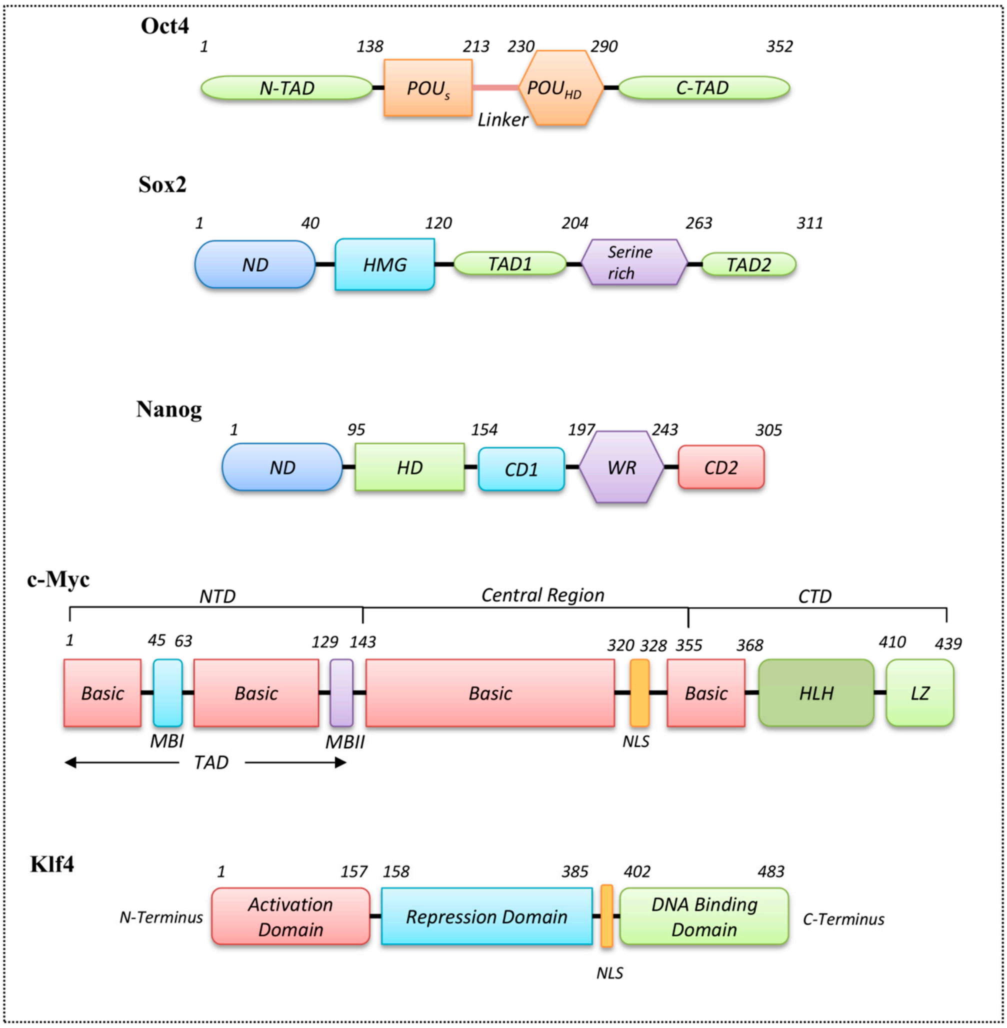


Figure 3. Domain architectures of stem cell transcription factors. A representation of the arrangement of functional domains in stem cell transcription factors Oct4, Sox2, Nanog, c-Myc, and Klf4 are shown. Each domain is marked with the length of its corresponding amino acid sequence. TAD stands for transactivation domain, HMG, HD, WR, HLH, and LZ stand for high-mobility group, homeodomain, tryptophan repeats, helix-loop-helix, and leucine zippers, respectively. NLS stands for nuclear localization sequence. MBI and MBII stand for Myc Boxes I and II, respectively. ND, CD1, CD2, and POU stand for N-terminal domain, C-terminal domain 1, C-terminal domain 2 and POU is derived from the names of three mammalian transcription factors, the pituitary-specific Pit-1, the octamer-binding proteins Oct-1 and Oct-2, and the neural Unc-86 from *Caenorhabditis elegans*.

HD domain proteins prefer AT-rich regions; sites bound by HD TFs generally have very low GC content around the core binding sites [13]. Although a wealth of information is available regarding HD-DNA interactions, the exact mechanism remains elusive. HD-DNA recognition is thought to occur through the action of specific amino acids present in the HD recognition helix, which engage its corresponding nucleotides in the cis-regulatory elements. Other additional recognition mechanisms for HD domain proteins include water-mediated interactions, alterations in the DNA structural parameters, and the presence of cooperative binding factors [93].

3.4. Helix-Loop-Helix Proteins

The HLH family of TF comprises approximately 200 members; each member has a distinct function regarding cell cycle control and differentiation. The HLH domain mediates homo/hetero dimerization, thereby playing an important role in DNA-binding and transcriptional regulation [99]. Most HLH members contain highly basic residues close to the HLH domain that facilitate DNA-binding at the canonical E-box site (CANNTG). The highly conserved HLH region has two α-helices (15–20 residues-long), which are separated by a short loop of variable length [100].

A large number of HLH family proteins have been classified based on distribution, DNA-binding specificity, and dimerization capabilities [101]. Class I HLH proteins (E12, E47, HEB, E2-2, and Daughterless; also known as E proteins) are capable of forming homo- or heterodimers on the E-box binding site. The class II proteins (atonal, MyoD, myogenin, NeuroD/BETA2, and the achaete-scute complex) preferentially form heterodimers with the E proteins. Class III HLH proteins (Myc family of TFs, transcription factor binding to IGHM enhancer 3 (TFE3), sterol regulatory element-binding proteins (SREBP-1), and microphthalmia-associated TF (MITF)) [101,102] have leucine zippers (LZ) adjacent to the HTH domain motif. Class IV proteins (Mad, Myc-associated factor X (Max), and Max interacting protein (Mxi)) [103] dimerize with the Myc proteins. Members of this protein family that lack the DNA-binding region are known as Class V HLH proteins (inhibitor of DNA-binding proteins and emc (Extramachrochaetae) proteins) [104–107], which happen to function separately by dimerizing with other basic-HLH (bHLH) type TF and act as negative regulators toward the binding of bHLH proteins to DNA. Class V members are negative regulators of Class I and II HLH proteins as well [107]. Class VI and VII proteins contain proline in the basic region and bHLH-PER-ARNT-SIM domains, respectively [100,108,109].

The C-terminus of c-Myc harbors bHLH-zipper domain, and the amino-terminal end holds the two highly conserved elements (Myc box 1 and 2 (MBI and MBII)), which are necessary for the transactivation of its target genes. Mutations in the TAD domain or in the bHLH-zipper domain have the potential to abolish c-Myc activity [110]. The heterodimerization of bHLH proteins is mediated through two HLH-zipper interfaces, and for c-Myc, the heterodimerization occurs through a highly specific interaction with the bHLH-zipper protein Max. The heterodimerization of c-Myc with Max enables the association with E-box DNA sequences (CACGTG), thereby stimulating transcription (Figure 2) [111,112]. Max proteins homodimerize weakly, whereas forced Max expression blocks the c-Myc biological activity through competition for E-box sites [113]. Although many biological activities of c-Myc proteins are dependent on Max heterodimers, Max-independent functions of c-Myc proteins are also reported [114,115]. The c-Myc proteins are present with or without the Max protein at the non-E-box binding site with the help of other interacting TFs [115–117]. The c-Myc functional domains involved in transcriptional regulation are highlighted in Figure 3.

3.5. Zinc Finger Domain Proteins

The ZF is a large, widespread domain structure present in 3% of genes composing the human genome. A typical ZF domain contains two histidine and two cysteine residues that pack a zinc ion with coordinate bonds. This motif is composed of an α-helix and antiparallel β-strand (Figure 2); four key amino acid residues, located in specific positions at the tip of the finger, are responsible for DNA recognition by creating hydrogen bonds with the major groove [118]. Each ZF is tandemly linked in a polar fashion to recognize DNA of variable lengths [119]. Even though each finger domain has similar structural arrangements, variations in key amino acid residues drive a large number of combinatorial possibilities in terms of DNA sequence recognition. Thereby, the majority (approximately 80%) of the classical C2H2-ZF proteins have no known DNA-binding motifs [120,121]. Human ZF proteins have approximately 10 C2H2-ZF domains that can bind to approximately 30 base pairs on DNA; however, the entire domain does not necessarily bind simultaneously. Evidence suggests that this domain also binds to other proteins and ligands [122,123].

C2H2-ZF proteins can be divided into three major groups. The first group contains a cluster of three ZF, and is mainly composed of Sp1-like transcription factors [124]. TFs belonging to the Klf family (approximately 17 members in humans) have been identified, displaying three conserved ZF in their C-terminal polypeptide chains [125]. The second group is the smallest, with one or more pairs of ZF. The greater the number of ZF, the further they are located from each other. Tramtrack (TTK-one ZF pair), positive regulatory domain II binding factor (PRDII-BF1- two ZF pairs) [126], and basonudin (three ZF pairs) [127] are some examples of this group. The most abundant group is the third one, which comprises ZF proteins containing clusters made of four or more ZFs. Each protein may have one or more domains containing several closely spaced ZFs. TF CTCF (11 ZFs arranged in one cluster) [128,129], myeloid zinc finger 1 (MZF-1) [130], NEP1 –interacting protein (NIP1) [131], and zinc finger protein 394 (ZNF394) [132] are some examples. Although ZF proteins are classified into these three groups, there are other types of ZFs that contain both paired and clustered groups. For example, zinc finger protein 305 (ZNF305) and paternally-expressed gene 3 (PEG3) TFs contain both pairs and clusters composed of more than four ZFs [118].

The stem cell TF known as Klf4 has been characterized by the C2H2-ZF DNA-binding motif located at its C-terminus (Figure 3). The amino acid residue at position 81 in the ZF domain is highly conserved within the Klf family, and can recognize GC-rich regions (CACCC) [125]. The structural arrangement of the ZF domain of Klf4 comprises two short β-strands followed by an α-helix. The classical ZF arrangement displays a #-X-C-X(1-5)-C-X3-#-X5-#-X2-H-X(3-6)-[H/C] pattern, in which C, H, and X denote cysteine, histidine, and any amino acid, respectively. The # symbol marks an important amino acid, and the associated number defines the number of amino acid residues. Klf proteins have highly conserved linker residues (TGE(R/K)P(Y/F)X) between their ZF domains [125]. The N-termini of Klf proteins have other transcriptional regulatory domains that are specific to each Klf protein. This reflects the functional diversity of Klf proteins, resulting in various interactions with distinct co-activators and repressors [133]. A representation of the domain families with the referenced crystal structures and the domain organization of stem cell TFs are shown in Figures 2 and 3, respectively.

4. Cooperative Binding of Stem Cell TFs

Oct4, Sox2, and Nanog are important TFs, essential toward maintaining the embryonic pluripotent state. Cooperative interactions between these factors drive the pluripotent-specific expression of target genes. Oct4 binds to DNA as a monomer, a homodimer, or a heterodimer with other transcription factors (e.g., Sox2). Even though the Sox family of TFs plays an extensive role in embryonic development, the binding of Sox2 alone does not initiate transcription; it requires a binding partner at an adjacent site on its targeted DNA. Even though Oct4/Sox2 play independent roles in determining other cell types, their cooperative interaction also drives the transcription of a specific set of target genes responsible for cell reprogramming. The known target genes of Oct4/Sox2 heterodimers include fibroblast growth factor 4 (*Fgf4*), undifferentiated embryonic cell transcription factor 1 (*Utf1*), F-box only protein 15 (*Fbxo15*), as well as *Sox2* and *Pou5f1* (the gene encoding Oct4) themselves [134].

Proteins are generally dynamic in nature, and upon binding, protein complexes promote a state that is energetically favorable for efficient transcription. Oct4/Sox2 DNA recognition is driven by the sequence-dependent deformation of DNA, where Oct4, an HTH-containing protein, docks into the major groove of DNA to establish the direct sequence-specific interaction. This causes dynamic and transient contacts between the protein domains and the DNA-binding interface. Combined, the two domains (POU_S and POU_HD) of Oct4 recognize the consensus sequence ATGC(A/T)AAT, where the POU_S recognizes the first half and the POU_HD recognizes the second [94,135]. Furthermore, the Sox2 HMG domain recognizes the CTTTGTT consensus sequence, and its binding introduces a topological deformation (pronounced kink) with respect to B-DNA. The residues of helix1 and helix2 of the Sox2-HMG domain are inserted at the 3-base pair stacks (which are marked) of the consensus sequence (C*T*T*TGTT), and unwind the DNA. Only helix1 and helix2 are needed to enable the DNA interaction, whereas the residues Pro68, Tyr72, and Pro74 from helix3 reorient the C-terminal tail and maintain the interaction with Oct4 [83]. The Oct4 residues Glu82 and Lys85 transiently create salt bridges with the complementary residues in Sox2 [97]. Mutating certain amino acids of the HMG residues of Sox factors correspondingly swaps their ability to dimerize with Oct4 on specific composite motifs and to direct cell fate decisions during the induction of pluripotency or endodermal fate (Figure 4).

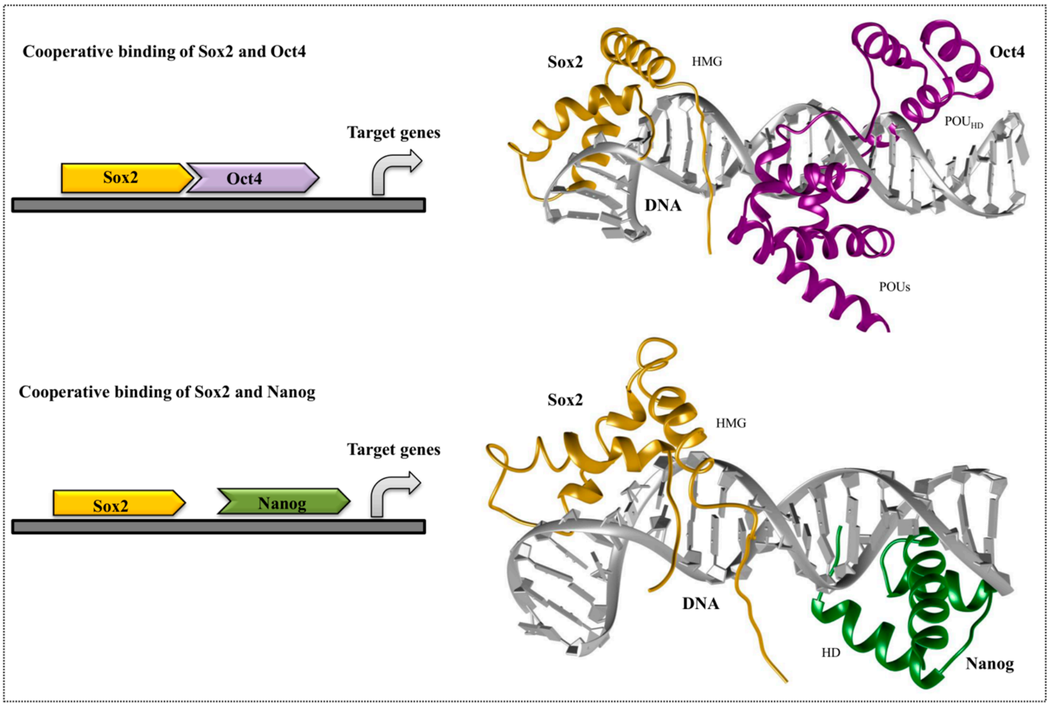


Figure 4. Representative figure of the cooperative binding of stem cell factors. The figure illustrates the cooperative binding of Sox2 and Oct4, as well as Sox2 and Nanog, on their enhancers/promoters of target genes. The Oct4/Sox2 crystal structure is obtained from PDB (1gt0), whereas Sox2/Nanog structure was modeled using chimera.

Even though there is no structural information about the direct physical interaction among Oct4, Sox2, and Nanog, experimental evidence has shown that Nanog-bound promoters are often co-occupied by these two proteins. Within the *Nanog* promoter region, the presence of Sox-Oct-cis regulatory elements confirms the cooperative binding among these factors, which may be necessary for *Nanog* pluripotent transcription [134]. Various experiments have shown that Oct4 and Sox2 regulate their own transcription via Oct4/Sox2 heterodimerization, and positively regulate Nanog transcription to maintain an undifferentiated state in embryonic stem cells (ESCs). Direct physical interactions between Nanog and Sox2 can occur, in which the WR domain of Nanog and the transactivation domain of Sox2 interact (Figure 4). The cooperative binding between Nanog and Sox2 is mediated by the tryptophan residues located in the WR domain of Nanog, and residues 205–263 within the serine-rich region of Sox2. This critical Nanog-interacting region in Sox2 (residues 212–233) is highly enriched with hydroxyamino acids, similarly to the WR domain residues in Nanog [136] (Figure 3). Moreover, careful examination has highlighted three repeats of the sequence S X T/S Y of this 21-amino acid region that may be responsible for mediating the interaction with Nanog.

Studies have proven that Oct4, Sox2, and Klf4 can independently target the nucleosomes using partial or degenerate motifs, and can also recognize the full canonical motifs in the absence of nucleosomes [137]. This differential ability of TFs to recognize their target sites on nucleosomes paves a way to build a hierarchical model for the pioneer factors and identify their targets in silent chromatin structures [138].

The gene regulatory functions are carried out through the molecular interactions between protein and DNA. Uncovering such interactions, as well as identifying precise genome-wide DNA-binding sites for TFs, is performed by chromatin immunoprecipitation, using the DNA sequencing (ChIP-Seq) method [139]. A comprehensive map of Myc binding has been identified within the mouse ESCs genome through ChIP-Seq method. This map revealed 4325 Myc binding sites, of which 2885 were newly identified [140]. Hundreds of target genes have been identified for the Oct4/Sox2 binding, in which Zfp206 is an important TF, playing a role in maintaining stem cell pluripotency [141]. Along with this typical ChIP assay, recent computational methods are also gaining importance toward genome-wide studies on protein-DNA interactions. Position weight matrix (PWM) is the method used for discovering motifs in nucleotide or amino acid sequences, which incorporates the most advanced algorithm in bioinformatics [142]. This method has been extensively used in finding the transcription initiation sites, intron splicing sites, whole-genome screening of regulatory elements, and is also a useful measure for motif strength [143–145].

5. Summary and Perspective

In this review, we summarized the recognition mechanisms that TFs utilize to select their DNA-binding sites. Base readout involves hydrogen bonds and hydrophobic interactions between the DNA bases and the protein, enabling direct readouts. The shape readout mechanism is utilized by many TFs, where the normal or deformed shape of DNA attracts its perfect binding partner. DNA bending, unwinding, and other helical parameters, like major/minor groove width, also influence the shape readout mechanism. The initial search scenarios such as sliding, hopping, and intersegmental jumping are also discussed. To study the structural details and their cooperative mechanisms, we also discussed TF binding domain families with a special emphasis on the stem cell TFs Oct4, Sox2, Nanog, c-Myc, and Klf4. Specifically, the DNA-binding domains of HTH, HLH, and ZF domain structures, as well as their recognition mechanisms, were discussed. Taken together, this review provides basic and advanced information about the recognition mechanisms of TFs, along with a structural study of stem cell factors.

A widely used method to identify the binding site for a protein is PWM [142]. The interdependencies between the nucleotide positions in the binding site can justify quantitative binding, whereas the qualitative binding relies on cofactors, cooperativity, and chromatin accessibility. An accurate model to represent all these highly complex, dependent attributes in protein-DNA binding remains elusive; however, these questions can be resolved at a molecular level by MD simulation methods. These may include the impact of superhelicity on protein-DNA interactions and thorough energy landscape analysis for protein-DNA recognition mechanisms [146]. Experiments have suggested that regulatory proteins should bind firmly to their target sites; however, they also found strong binding to other non-specific sites that act as bait, thereby increasing the time needed for a TF to discover its actual target site [147]. Numerous theoretical models and mechanisms have been proposed to solve this problem, but the predictions of these developing computational models have yet to be tested. Even though the single-molecule technique receives considerable attention regarding the examination of the motion of protein on DNA, it often lacks sufficient resolution to distinguish between sliding and hopping mechanisms [31,148,149]. However, the emerging super-resolution microscopy methods are capable of making direct visualization of individual proteins, which is useful to study DNA repair events [150,151]. Beyond this method, photoactivated localization microscopy (PALM) is used to track the movement of individual molecules in the living cells, which is very useful to study search mechanisms of protein-DNA complexes. In addition to that, PALM provides a direct way of counting the molecules in the cells, and provides a quantitative description of complex reaction networks [150,152]. Advancements in computational processing (graphics processing unit, GPU) and the optimization of MD algorithms have allowed us to analyze conformational ensembles, which represent the real macromolecules in a superior manner. Tools exist in MD simulations that make the setup of a macromolecular system much easier, taking into account macromolecules' flexibility and the dynamic properties (especially thermodynamics). More recent computational methods use machine learning methods to find the characteristic features of protein-DNA interactions. With the availability of recent research contributions and with improved computational techniques, a number of strategies have been utilized toward determining protein-DNA interaction mechanisms. These approaches can be highly successful when combined with experimental data.

Acknowledgments: This work was supported by grants from the Korea Health Industry Development Institute (HI14C1992) and the National Research Foundation of Korea (NRF-2015R1A2A2A09001059, NRF 2012–0006687).

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

References

1. Lange, M.; Kochugaeva, M.; Kolomeisky, A.B. Protein search for multiple targets on DNA. *J. Chem. Phys.* **2015**, *143*, 105102. [CrossRef] [PubMed]
2. Kolomeisky, A.B. Physics of protein-DNA interactions: Mechanisms of facilitated target search. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2088–2095. [CrossRef] [PubMed]
3. Pan, Y.; Tsai, C.J.; Ma, B.; Nussinov, R. Mechanisms of transcription factor selectivity. *Trends Genet.* **2010**, *26*, 75–83. [CrossRef] [PubMed]
4. Slattery, M.; Zhou, T.; Yang, L.; Machado, A.C.; Gordan, R.; Rohs, R. Absence of a simple code: How transcription factors read the genome. *Trends Biochem. Sci.* **2014**, *39*, 381–399. [CrossRef] [PubMed]
5. Chen, Y.; Bates, D.L.; Dey, R.; Chen, P.H.; Machado, A.C.; Laird-Offringa, I.A.; Rohs, R.; Chen, L. DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Rep.* **2012**, *2*, 1197–1206. [CrossRef] [PubMed]
6. Rohs, R.; West, S.M.; Sosinsky, A.; Liu, P.; Mann, R.S.; Honig, B. The role of DNA shape in protein-DNA recognition. *Nature* **2009**, *461*, 1248–1253. [CrossRef] [PubMed]
7. Aravind, L.; Anantharaman, V.; Balaji, S.; Babu, M.M.; Iyer, L.M. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiol. Rev.* **2005**, *29*, 231–262. [CrossRef] [PubMed]
8. Badia, D.; Camacho, A.; Perez-Lago, L.; Escandon, C.; Salas, M.; Coll, M. The structure of phage phi29 transcription regulator p4-DNA complex reveals an N-hook motif for DNA. *Mol. Cell* **2006**, *22*, 73–81. [CrossRef] [PubMed]
9. Jordan, S.R.; Pabo, C.O. Structure of the lambda complex at 2.5 A resolution: Details of the repressor-operator interactions. *Science* **1988**, *242*, 893–899. [CrossRef] [PubMed]
10. Wang, J.; Zhuang, J.; Iyer, S.; Lin, X.; Whitfield, T.W.; Greven, M.C.; Pierce, B.G.; Dong, X.; Kundaje, A.; Cheng, Y.; et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **2012**, *22*, 1798–1812. [CrossRef] [PubMed]
11. Dror, I.; Golan, T.; Levy, C.; Rohs, R.; Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.* **2015**, *25*, 1268–1280. [CrossRef] [PubMed]
12. Rohs, R.; West, S.M.; Liu, P.; Honig, B. Nuance in the double-helix and its role in protein-DNA recognition. *Curr. Opin. Struct. Biol.* **2009**, *19*, 171–177. [CrossRef] [PubMed]
13. Rohs, R.; Jin, X.; West, S.M.; Joshi, R.; Honig, B.; Mann, R.S. Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **2010**, *79*, 233–269. [CrossRef] [PubMed]
14. Stella, S.; Cascio, D.; Johnson, R.C. The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes Dev.* **2010**, *24*, 814–826. [CrossRef] [PubMed]
15. Chen, Y.; Zhang, X.; Dantas Machado, A.C.; Ding, Y.; Chen, Z.; Qin, P.Z.; Rohs, R.; Chen, L. Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion. *Nucleic Acids Res.* **2013**, *41*, 8368–8376. [CrossRef] [PubMed]
16. Anwar, M.A.; Yesudhas, D.; Shah, M.; Choi, S. Structural and conformational insights into Sox2/Oct4-bound enhancer DNA: A computational perspective. *RSC Adv.* **2016**, *6*, 90138–90153. [CrossRef]
17. Komazin-Meredith, G.; Mirchev, R.; Golan, D.E.; van Oijen, A.M.; Coen, D.M. Hopping of a processivity factor on DNA revealed by single-molecule assays of diffusion. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 10721–10726. [CrossRef] [PubMed]
18. Winter, R.B.; Berg, O.G.; von Hippel, P.H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli lac* repressor—Operator interaction: Kinetic measurements and conclusions. *Biochemistry* **1981**, *20*, 6961–6977. [CrossRef] [PubMed]

19. Kalodimos, C.G.; Biris, N.; Bonvin, A.M.; Levandoski, M.M.; Guennuegues, M.; Boelens, R.; Kaptein, R. Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **2004**, *305*, 386–389. [CrossRef] [PubMed]
20. Blainey, P.C.; van Oijen, A.M.; Banerjee, A.; Verdine, G.L.; Xie, X.S. A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 5752–5757. [CrossRef] [PubMed]
21. Gorman, J.; Chowdhury, A.; Surtees, J.A.; Shimada, J.; Reichman, D.R.; Alani, E.; Greene, E.C. Dynamic basis for one-dimensional DNA scanning by the mismatch repair complex Msh2-Msh6. *Mol. Cell* **2007**, *28*, 359–370. [CrossRef] [PubMed]
22. Viadiu, H.; Aggarwal, A.K. Structure of BamHI bound to nonspecific DNA: A model for DNA sliding. *Mol. Cell* **2000**, *5*, 889–895. [CrossRef]
23. Albright, R.A.; Mossing, M.C.; Matthews, B.W. Crystal structure of an engineered Cro monomer bound nonspecifically to DNA: Possible implications for nonspecific binding by the wild-type protein. *Protein Sci.* **1998**, *7*, 1485–1494. [CrossRef] [PubMed]
24. Furini, S.; Barbini, P.; Domene, C. DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence. *Nucleic Acids Res.* **2013**, *41*, 3963–3972. [CrossRef] [PubMed]
25. Von Hippel, P.H.; Berg, O.G. Facilitated target location in biological systems. *J. Biol. Chem.* **1989**, *264*, 675–678. [PubMed]
26. Halford, S.E.; Gowers, D.M.; Sessions, R.B. Two are better than one. *Nat. Struct. Biol.* **2000**, *7*, 705–707. [CrossRef] [PubMed]
27. Lomholt, M.A.; van den Broek, B.; Kalisch, S.M.; Wuite, G.J.; Metzler, R. Facilitated diffusion with DNA coiling. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 8204–8208. [CrossRef] [PubMed]
28. Winkler, F.K.; Banner, D.W.; Oefner, C.; Tsernoglou, D.; Brown, R.S.; Heathman, S.P.; Bryan, R.K.; Martin, P.D.; Petratos, K.; Wilson, K.S. The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J.* **1993**, *12*, 1781–1795. [PubMed]
29. Dhavan, G.M.; Crothers, D.M.; Chance, M.R.; Brenowitz, M. Concerted binding and bending of DNA by *Escherichia coli* integration host factor. *J. Mol. Biol.* **2002**, *315*, 1027–1037. [CrossRef] [PubMed]
30. Hu, T.; Grosberg, A.Y.; Shklovskii, B.I. How proteins search for their specific sites on DNA: The role of DNA conformation. *Biophys. J.* **2006**, *90*, 2731–2744. [CrossRef] [PubMed]
31. Halford, S.E.; Marko, J.F. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.* **2004**, *32*, 3040–3052. [CrossRef] [PubMed]
32. Van den Broek, B.; Lomholt, M.A.; Kalisch, S.M.; Metzler, R.; Wuite, G.J. How DNA coiling enhances target localization by proteins. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 15738–15742. [CrossRef] [PubMed]
33. Gorman, J.; Plys, A.J.; Visnapuu, M.L.; Alani, E.; Greene, E.C. Visualizing one-dimensional diffusion of eukaryotic DNA repair factors along a chromatin lattice. *Nat. Struct. Mol. Biol.* **2010**, *17*, 932–938. [CrossRef] [PubMed]
34. Bell, C.E.; Lewis, M. A closer view of the conformation of the *Lac* repressor bound to operator. *Nat. Struct. Biol.* **2000**, *7*, 209–214. [CrossRef] [PubMed]
35. Kalodimos, C.G.; Bonvin, A.M.; Salinas, R.K.; Wechselberger, R.; Boelens, R.; Kaptein, R. Plasticity in protein-DNA recognition: *Lac* repressor interacts with its natural operator 01 through alternative conformations of its DNA-binding domain. *EMBO J.* **2002**, *21*, 2866–2876. [CrossRef] [PubMed]
36. Gilbert, W.; Maxam, A. The nucleotide sequence of the *lac* operator. *Proc. Natl. Acad. Sci. USA* **1973**, *70*, 3581–3584. [CrossRef] [PubMed]
37. Kalodimos, C.G.; Folkers, G.E.; Boelens, R.; Kaptein, R. Strong DNA binding by covalently linked dimeric *Lac* headpiece: Evidence for the crucial role of the hinge helices. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 6039–6044. [CrossRef] [PubMed]
38. Harrison, S.C.; Aggarwal, A.K. DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.* **1990**, *59*, 933–969. [CrossRef] [PubMed]
39. Coulocheri, S.A.; Pigis, D.G.; Papavassiliou, K.A.; Papavassiliou, A.G. Hydrogen bonds in protein-DNA complexes: Where geometry meets plasticity. *Biochimie* **2007**, *89*, 1291–1303. [CrossRef] [PubMed]
40. Rastinejad, F.; Wagner, T.; Zhao, Q.; Khorasanizadeh, S. Structure of the RXR-RAR DNA-binding complex on the retinoic acid response element DR1. *EMBO J.* **2000**, *19*, 1045–1054. [CrossRef] [PubMed]

41. Hegde, R.S.; Grossman, S.R.; Laimins, L.A.; Sigler, P.B. Crystal structure at 1.7 A of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature* **1992**, *359*, 505–512. [CrossRef] [PubMed]
42. Kim, J.L.; Nikolov, D.B.; Burley, S.K. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **1993**, *365*, 520–527. [CrossRef] [PubMed]
43. Lee, O.S.; Cho, V.Y.; Schatz, G.C. A- to B-form transition in DNA between gold surfaces. *J. Phys. Chem. B* **2012**, *116*, 7000–7005. [CrossRef] [PubMed]
44. Waters, J.T.; Lu, X.J.; Galindo-Murillo, R.; Gumbart, J.C.; Kim, H.D.; Cheatham, T.E., 3rd; Harvey, S.C. Transitions of Double-Stranded DNA Between the A- and B-Forms. *J. Phys. Chem. B* **2016**, *120*, 8449–8456. [CrossRef] [PubMed]
45. Jayaram, B.; Sprous, D.; Young, M.A.; Beveridge, D.L. Free Energy Analysis of the Conformational Preferences of A and B Forms of DNA in Solution. *J. Am. Chem. Soc.* **1998**, *120*, 10629–10633. [CrossRef]
46. Leslie, A.G.; Arnott, S.; Chandrasekaran, R.; Ratliff, R.L. Polymorphism of DNA double helices. *J. Mol. Biol.* **1980**, *143*, 49–72. [CrossRef]
47. Haran, T.E.; Mohanty, U. The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.* **2009**, *42*, 41–81. [CrossRef] [PubMed]
48. Zhang, Y.; Xi, Z.; Hegde, R.S.; Shakked, Z.; Crothers, D.M. Predicting indirect readout effects in protein-DNA interactions. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 8337–8341. [CrossRef] [PubMed]
49. Doucleff, M.; Clore, G.M. Global jumping and domain-specific intersegment transfer between DNA cognate sites of the multidomain transcription factor Oct-1. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 13871–13876. [CrossRef] [PubMed]
50. Fuxreiter, M.; Simon, I.; Bondos, S. Dynamic protein-DNA recognition: Beyond what can be seen. *Trends Biochem. Sci.* **2011**, *36*, 415–423. [CrossRef] [PubMed]
51. Garvie, C.W.; Wolberger, C. Recognition of specific DNA sequences. *Mol. Cell* **2001**, *8*, 937–946. [CrossRef]
52. Vuzman, D.; Azia, A.; Levy, Y. Searching DNA via a "Monkey Bar" mechanism: The significance of disordered tails. *J. Mol. Biol.* **2010**, *396*, 674–684. [CrossRef] [PubMed]
53. He, B.; Wang, K.; Liu, Y.; Xue, B.; Uversky, V.N.; Dunker, A.K. Predicting intrinsic disorder in proteins: An overview. *Cell Res.* **2009**, *19*, 929–949. [CrossRef] [PubMed]
54. Meszaros, B.; Simon, I.; Dosztanyi, Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* **2009**, *5*, e1000376. [CrossRef] [PubMed]
55. Diella, F.; Haslam, N.; Chica, C.; Budd, A.; Michael, S.; Brown, N.P.; Trave, G.; Gibson, T.J. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.* **2008**, *13*, 6580–6603. [CrossRef] [PubMed]
56. Vise, P.D.; Baral, B.; Latos, A.J.; Daughdrill, G.W. NMR chemical shift and relaxation measurements provide evidence for the coupled folding and binding of the p53 transactivation domain. *Nucleic Acids Res.* **2005**, *33*, 2061–2077. [CrossRef] [PubMed]
57. Ma, B.; Nussinov, R. Amplification of signaling via cellular allosteric relay and protein disorder. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 6887–6888. [CrossRef] [PubMed]
58. Watson, M.; Stott, K.; Thomas, J.O. Mapping intramolecular interactions between domains in HMGB1 using a tail-truncation approach. *J. Mol. Biol.* **2007**, *374*, 1286–1297. [CrossRef] [PubMed]
59. Siggers, T.; Gordan, R. Protein-DNA binding: Complexities and multi-protein codes. *Nucleic Acids Res.* **2014**, *42*, 2099–2111. [CrossRef] [PubMed]
60. Wolberger, C. Multiprotein-DNA complexes in transcriptional regulation. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 29–56. [CrossRef] [PubMed]
61. Johnson, A.D. Molecular mechanisms of cell-type determination in budding yeast. *Curr. Opin. Genet. Dev.* **1995**, *5*, 552–558. [CrossRef]
62. Kim, S.; Brostromer, E.; Xing, D.; Jin, J.; Chong, S.; Ge, H.; Wang, S.; Gu, C.; Yang, L.; Gao, Y.Q.; et al. Probing allostery through DNA. *Science* **2013**, *339*, 816–819. [CrossRef] [PubMed]
63. Chen, C.; Pettitt, B.M. The binding process of a nonspecific enzyme with DNA. *Biophys. J.* **2011**, *101*, 1139–1147. [CrossRef] [PubMed]
64. Afek, A.; Schipper, J.L.; Horton, J.; Gordan, R.; Lukatsky, D.B. Protein-DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 17140–17145. [CrossRef] [PubMed]

65. Marcovitz, A.; Levy, Y. Frustration in protein-DNA binding influences conformational switching and target search kinetics. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 17957–17962. [CrossRef] [PubMed]
66. Norberg, J. Association of protein-DNA recognition complexes: Electrostatic and nonelectrostatic effects. *Arch. Biochem. Biophys.* **2003**, *410*, 48–68. [CrossRef]
67. Tsourkas, A.; Behlke, M.A.; Rose, S.D.; Bao, G. Hybridization kinetics and thermodynamics of molecular beacons. *Nucleic Acids Res.* **2003**, *31*, 1319–1330. [CrossRef] [PubMed]
68. Anderson, T.A.; Cordes, M.H.; Sauer, R.T. Sequence determinants of a conformational switch in a protein structure. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 18344–18349. [CrossRef] [PubMed]
69. Religa, T.L.; Johnson, C.M.; Vu, D.M.; Brewer, S.H.; Dyer, R.B.; Fersht, A.R. The helix-turn-helix motif as an ultrafast independently folding domain: The pathway of folding of Engrailed homeodomain. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 9272–9277. [CrossRef] [PubMed]
70. Rigali, S.; Derouaux, A.; Giannotta, F.; Dusart, J. Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *J. Biol. Chem.* **2002**, *277*, 12507–12515. [CrossRef] [PubMed]
71. Stros, M.; Launholt, D.; Grasser, K.D. The HMG-box: A versatile protein domain occurring in a wide variety of DNA-binding proteins. *Cell. Mol. Life Sci.* **2007**, *64*, 2590–2606. [CrossRef] [PubMed]
72. Gerlitz, G.; Hock, R.; Ueda, T.; Bustin, M. The dynamics of HMG protein-chromatin interactions in living cells. *Biochem. Cell Biol.* **2009**, *87*, 127–137. [CrossRef] [PubMed]
73. Hock, R.; Furusawa, T.; Ueda, T.; Bustin, M. HMG chromosomal proteins in development and disease. *Trends Cell Biol.* **2007**, *17*, 72–79. [CrossRef] [PubMed]
74. Fedele, M.; Battista, S.; Manfioletti, G.; Croce, C.M.; Giancotti, V.; Fusco, A. Role of the high mobility group A proteins in human lipomas. *Carcinogenesis* **2001**, *22*, 1583–1591. [CrossRef] [PubMed]
75. Reeves, R. Molecular biology of HMGA proteins: Hubs of nuclear function. *Gene* **2001**, *277*, 63–81. [CrossRef]
76. Reeves, R.; Adair, J.E. Role of high mobility group (HMG) chromatin proteins in DNA repair. *DNA Repair* **2005**, *4*, 926–938. [CrossRef] [PubMed]
77. Giavara, S.; Kosmidou, E.; Hande, M.P.; Bianchi, M.E.; Morgan, A.; d'Adda di Fagagna, F.; Jackson, S.P. Yeast Nhp6A/B and mammalian HMGB1 facilitate the maintenance of genome stability. *Curr. Biol.* **2005**, *15*, 68–72. [CrossRef] [PubMed]
78. Bustin, M.; Lehn, D.A.; Landsman, D. Structural features of the HMG chromosomal proteins and their genes. *Biochim. Biophys. Acta* **1990**, *1049*, 231–243. [CrossRef]
79. Stros, M. HMGB proteins: Interactions with DNA and chromatin. *Biochim. Biophys. Acta* **2010**, *1799*, 101–113. [CrossRef] [PubMed]
80. Wegner, M. From head to toes: The multiple facets of Sox proteins. *Nucleic Acids Res.* **1999**, *27*, 1409–1420. [CrossRef] [PubMed]
81. Grosschedl, R.; Giese, K.; Pagel, J. HMG domain proteins: Architectural elements in the assembly of nucleoprotein structures. *Trends Genet.* **1994**, *10*, 94–100. [CrossRef]
82. Kamachi, Y.; Uchikawa, M.; Kondoh, H. Pairing SOX off: With partners in the regulation of embryonic development. *Trends Genet.* **2000**, *16*, 182–187. [CrossRef]
83. Remenyi, A.; Lins, K.; Nissen, L.J.; Reinbold, R.; Scholer, H.R.; Wilmanns, M. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.* **2003**, *17*, 2048–2059. [CrossRef] [PubMed]
84. Burglin, T.R.; Affolter, M. Homeodomain proteins: An update. *Chromosoma* **2016**, *125*, 497–521. [CrossRef] [PubMed]
85. Banerjee-Basu, S.; Baxevanis, A.D. Molecular evolution of the homeodomain family of transcription factors. *Nucleic Acids Res.* **2001**, *29*, 3258–3269. [CrossRef] [PubMed]
86. Gehring, W.J.; Qian, Y.Q.; Billeter, M.; Furukubo-Tokunaga, K.; Schier, A.F.; Resendez-Perez, D.; Affolter, M.; Otting, G.; Wüthrich, K. Homeodomain-DNA recognition. *Cell* **1994**, *78*, 211–223. [CrossRef]
87. Andzelm, M.M.; Cherry, T.J.; Harmin, D.A.; Boeke, A.C.; Lee, C.; Hemberg, M.; Pawlyk, B.; Malik, A.N.; Flavell, S.W.; Sandberg, M.A.; et al. MEF2D drives photoreceptor development through a genome-wide competition for tissue-specific enhancers. *Neuron* **2015**, *86*, 247–263. [CrossRef] [PubMed]
88. Soufi, A.; Donahue, G.; Zaret, K.S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **2012**, *151*, 994–1004. [CrossRef] [PubMed]

89. Budry, L.; Balsalobre, A.; Gauthier, Y.; Khetchoumian, K.; L'Honore, A.; Vallette, S.; Brue, T.; Figarella-Branger, D.; Meij, B.; Drouin, J. The selector gene Pax7 dictates alternate pituitary cell fates through its pioneer action on chromatin remodeling. *Genes Dev.* **2012**, *26*, 2299–2310. [CrossRef] [PubMed]
90. Dinkel, H.; Van Roey, K.; Michael, S.; Kumar, M.; Uyar, B.; Altenberg, B.; Milchevskaya, V.; Schneider, M.; Kuhn, H.; Behrendt, A.; et al. ELM 2016—Data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.* **2016**, *44*, D294–D300. [CrossRef] [PubMed]
91. Uversky, V.N. The multifaceted roles of intrinsic disorder in protein complexes. *FEBS Lett.* **2015**, *589*, 2498–2506. [CrossRef] [PubMed]
92. Hayashi, Y.; Caboni, L.; Das, D.; Yumoto, F.; Clayton, T.; Deller, M.C.; Nguyen, P.; Farr, C.L.; Chiu, H.J.; Miller, M.D.; et al. Structure-based discovery of NANOG variant with enhanced properties to promote self-renewal and reprogramming of pluripotent stem cells. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 4666–4671. [CrossRef] [PubMed]
93. Jauch, R.; Ng, C.K.; Saikatendu, K.S.; Stevens, R.C.; Kolatkar, P.R. Crystal structure and DNA binding of the homeodomain of the stem cell transcription factor Nanog. *J. Mol. Biol.* **2008**, *376*, 758–770. [CrossRef] [PubMed]
94. Jerabek, S.; Merino, F.; Scholer, H.R.; Cojocaru, V. OCT4: Dynamic DNA binding pioneers stem cell pluripotency. *Biochim. Biophys. Acta* **2014**, *1839*, 138–154. [CrossRef] [PubMed]
95. Zhao, F.Q. Octamer-binding transcription factors: Genomics and functions. *Front. Biosci.* **2013**, *18*, 1051–1071. [CrossRef]
96. Kong, X.; Liu, J.; Li, L.; Yue, L.; Zhang, L.; Jiang, H.; Xie, X.; Luo, C. Functional interplay between the RK motif and linker segment dictates Oct4-DNA recognition. *Nucleic Acids Res.* **2015**, *43*, 4381–4392. [CrossRef] [PubMed]
97. Esch, D.; Vahokoski, J.; Groves, M.R.; Pogenberg, V.; Cojocaru, V.; Vom Bruch, H.; Han, D.; Drexler, H.C.; Arauzo-Bravo, M.J.; Ng, C.K.; et al. A unique Oct4 interface is crucial for reprogramming to pluripotency. *Nat. Cell Biol.* **2013**, *15*, 295–301. [CrossRef] [PubMed]
98. Klemm, J.D.; Rould, M.A.; Aurora, R.; Herr, W.; Pabo, C.O. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* **1994**, *77*, 21–32. [CrossRef]
99. Norton, J.D. ID helix-loop-helix proteins in cell growth, differentiation and tumorigenesis. *J. Cell Sci.* **2000**, *113 (Pt 22)*, 3897–3905. [PubMed]
100. Massari, M.E.; Murre, C. Helix-loop-helix proteins: Regulators of transcription in eucaryotic organisms. *Mol. Cell. Biol.* **2000**, *20*, 429–440. [CrossRef] [PubMed]
101. Murre, C.; Bain, G.; van Dijk, M.A.; Engel, I.; Furnari, B.A.; Massari, M.E.; Matthews, J.R.; Quong, M.W.; Rivera, R.R.; Stuiver, M.H. Structure and function of helix-loop-helix proteins. *Biochim. Biophys. Acta* **1994**, *1218*, 129–135. [CrossRef]
102. Henthorn, P.S.; Stewart, C.C.; Kadesch, T.; Puck, J.M. The gene encoding human TFE3, a transcription factor that binds the immunoglobulin heavy-chain enhancer, maps to Xp11.22. *Genomics* **1991**, *11*, 374–378. [CrossRef]
103. Ayer, D.E.; Kretzner, L.; Eisenman, R.N. Mad: A heterodimeric partner for Max that antagonizes Myc transcriptional activity. *Cell* **1993**, *72*, 211–222. [CrossRef]
104. Benezra, R.; Davis, R.L.; Lockshon, D.; Turner, D.L.; Weintraub, H. The protein Id: A negative regulator of helix-loop-helix DNA binding proteins. *Cell* **1990**, *61*, 49–59. [CrossRef]
105. Ellis, H.M.; Spann, D.R.; Posakony, J.W. *extramacrochaetae*, a negative regulator of sensory organ development in *Drosophila*, defines a new class of helix-loop-helix proteins. *Cell* **1990**, *61*, 27–38. [CrossRef]
106. Garrell, J.; Modolell, J. The *Drosophila extramacrochaetae* locus, an antagonist of proneural genes that, like these genes, encodes a helix-loop-helix protein. *Cell* **1990**, *61*, 39–48. [CrossRef]
107. Lasorella, A.; Benezra, R.; Iavarone, A. The ID proteins: Master regulators of cancer stem cells and tumour aggressiveness. *Nat. Rev. Cancer* **2014**, *14*, 77–91. [CrossRef] [PubMed]
108. Klambt, C.; Knust, E.; Tietze, K.; Campos-Ortega, J.A. Closely related transcripts encoded by the neurogenic gene complex enhancer of split of *Drosophila melanogaster*. *EMBO J.* **1989**, *8*, 203–210. [PubMed]
109. Crews, S.T. Control of cell lineage-specific development and transcription by bHLH-PAS proteins. *Genes Dev.* **1998**, *12*, 607–620. [CrossRef] [PubMed]
110. Pelengaris, S.; Khan, M.; Evan, G. c-MYC: More than just a matter of life and death. *Nat. Rev. Cancer* **2002**, *2*, 764–776. [CrossRef] [PubMed]

111. Nair, S.K.; Burley, S.K. X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell* **2003**, *112*, 193–205. [CrossRef]
112. Conacci-Sorrell, M.; McFerrin, L.; Eisenman, R.N. An overview of MYC and its interactome. *Cold Spring Harb. Perspect. Med.* **2014**, *4*, a014357. [CrossRef] [PubMed]
113. Canelles, M.; Delgado, M.D.; Hyland, K.M.; Lerga, A.; Richard, C.; Dang, C.V.; Leon, J. Max and inhibitory c-Myc mutants induce erythroid differentiation and resistance to apoptosis in human myeloid leukemia cells. *Oncogene* **1997**, *14*, 1315–1327. [CrossRef] [PubMed]
114. Gallant, P. Myc function in *Drosophila*. *Cold Spring Harb. Perspect. Med.* **2013**, *3*, a014324. [CrossRef] [PubMed]
115. Steiger, D.; Furrer, M.; Schwinkendorf, D.; Gallant, P. Max-independent functions of Myc in *Drosophila melanogaster*. *Nat. Genet.* **2008**, *40*, 1084–1091. [CrossRef] [PubMed]
116. Gomez-Roman, N.; Grandori, C.; Eisenman, R.N.; White, R.J. Direct activation of RNA polymerase III transcription by c-Myc. *Nature* **2003**, *421*, 290–294. [CrossRef] [PubMed]
117. Izumi, H.; Molander, C.; Penn, L.Z.; Ishisaki, A.; Kohno, K.; Funa, K. Mechanism for the transcriptional repression by c-Myc on PDGF beta-receptor. *J. Cell Sci.* **2001**, *114*, 1533–1544. [PubMed]
118. Razin, S.V.; Borunova, V.V.; Maksimenko, O.G.; Kantidze, O.L. Cys2His2 zinc finger protein family: Classification, functions, and major members. *Biochemistry* **2012**, *77*, 217–226. [CrossRef] [PubMed]
119. Klug, A. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu. Rev. Biochem.* **2010**, *79*, 213–231. [CrossRef] [PubMed]
120. Jolma, A.; Yan, J.; Whitington, T.; Toivonen, J.; Nitta, K.R.; Rastas, P.; Morgunova, E.; Enge, M.; Taipale, M.; Wei, G.; et al. DNA-binding specificities of human transcription factors. *Cell* **2013**, *152*, 327–339. [CrossRef] [PubMed]
121. Emerson, R.O.; Thomas, J.H. Adaptive evolution in zinc finger transcription factors. *PLoS Genet.* **2009**, *5*, e1000325. [CrossRef] [PubMed]
122. Najafabadi, H.S.; Mnaimneh, S.; Schmitges, F.W.; Garton, M.; Lam, K.N.; Yang, A.; Albu, M.; Weirauch, M.T.; Radovani, E.; Kim, P.M.; et al. C_2H_2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.* **2015**, *33*, 555–562. [CrossRef] [PubMed]
123. Brayer, K.J.; Segal, D.J. Keep your fingers off my DNA: Protein-protein interactions mediated by C_2H_2 zinc finger domains. *Cell Biochem. Biophys.* **2008**, *50*, 111–131. [CrossRef] [PubMed]
124. Kaczynski, J.; Cook, T.; Urrutia, R. Sp1- and Kruppel-like transcription factors. *Genome Biol.* **2003**, *4*, 206. [CrossRef] [PubMed]
125. Swamynathan, S.K. Kruppel-like factors: Three fingers in control. *Hum. Genom.* **2010**, *4*, 263–270. [CrossRef]
126. Fan, C.M.; Maniatis, T. A DNA-binding protein containing two widely separated zinc finger motifs that recognize the same DNA sequence. *Genes Dev.* **1990**, *4*, 29–42. [CrossRef] [PubMed]
127. Tseng, H.; Green, H. Basonuclin: A keratinocyte protein with multiple paired zinc fingers. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10311–10315. [CrossRef] [PubMed]
128. Ohlsson, R.; Bartkuhn, M.; Renkawitz, R. CTCF shapes chromatin by multiple mechanisms: The impact of 20 years of CTCF research on understanding the workings of chromatin. *Chromosoma* **2010**, *119*, 351–360. [CrossRef] [PubMed]
129. Dunn, K.L.; Davie, J.R. The many roles of the transcriptional regulator CTCF. *Biochem. Cell Biol.* **2003**, *81*, 161–167. [CrossRef] [PubMed]
130. Morris, J.F.; Hromas, R.; Rauscher, F.J., 3rd. Characterization of the DNA-binding properties of the myeloid zinc finger protein MZF1: Two independent DNA-binding domains recognize two DNA consensus sequences with a common G-rich core. *Mol. Cell. Biol.* **1994**, *14*, 1786–1795. [CrossRef] [PubMed]
131. Nielsen, A.L.; Jorgensen, P.; Lerouge, T.; Cervino, M.; Chambon, P.; Losson, R. Nizp1, a novel multitype zinc finger protein that interacts with the NSD1 histone lysine methyltransferase through a unique C2HR motif. *Mol. Cell. Biol.* **2004**, *24*, 5184–5196. [CrossRef] [PubMed]
132. Huang, C.; Wang, Y.; Li, D.; Li, Y.; Luo, J.; Yuan, W.; Ou, Y.; Zhu, C.; Zhang, Y.; Wang, Z.; et al. Inhibition of transcriptional activities of AP-1 and c-Jun by a new zinc finger protein ZNF394. *Biochem. Biophys. Res. Commun.* **2004**, *320*, 1298–1305. [CrossRef] [PubMed]
133. Geiman, D.E.; Ton-That, H.; Johnson, J.M.; Yang, V.W. Transactivation and growth suppression by the gut-enriched Kruppel-like factor (Kruppel-like factor 4) are dependent on acidic amino acid residues and protein-protein interaction. *Nucleic Acids Res.* **2000**, *28*, 1106–1113. [CrossRef] [PubMed]

134. Rodda, D.J.; Chew, J.L.; Lim, L.H.; Loh, Y.H.; Wang, B.; Ng, H.H.; Robson, P. Transcriptional regulation of nanog by OCT4 and SOX2. *J. Biol. Chem.* **2005**, *280*, 24731–24737. [CrossRef] [PubMed]
135. Merino, F.; Bouvier, B.; Cojocaru, V. Cooperative DNA recognition modulated by an interplay between protein-protein interactions and DNA-Mmediated allostery. *PLoS Comput. Biol.* **2015**, *11*, e1004287. [CrossRef] [PubMed]
136. Gagliardi, A.; Mullin, N.P.; Ying Tan, Z.; Colby, D.; Kousa, A.I.; Halbritter, F.; Weiss, J.T.; Felker, A.; Bezstarosti, K.; Favaro, R.; et al. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J.* **2013**, *32*, 2231–2247. [CrossRef] [PubMed]
137. Soufi, A.; Garcia, M.F.; Jaroszewicz, A.; Osman, N.; Pellegrini, M.; Zaret, K.S. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **2015**, *161*, 555–568. [CrossRef] [PubMed]
138. Iwafuchi-Doi, M.; Zaret, K.S. Pioneer transcription factors in cell reprogramming. *Genes Dev.* **2014**, *28*, 2679–2692. [CrossRef] [PubMed]
139. Valouev, A.; Johnson, D.S.; Sundquist, A.; Medina, C.; Anton, E.; Batzoglou, S.; Myers, R.M.; Sidow, A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **2008**, *5*, 829–834. [CrossRef] [PubMed]
140. Krepelova, A.; Neri, F.; Maldotti, M.; Rapelli, S.; Oliviero, S. Myc and max genome-wide binding sites analysis links the Myc regulatory network with the polycomb and the core pluripotency networks in mouse embryonic stem cells. *PLoS ONE* **2014**, *9*, e88933. [CrossRef] [PubMed]
141. Wang, Z.X.; Teh, C.H.; Kueh, J.L.; Lufkin, T.; Robson, P.; Stanton, L.W. Oct4 and Sox2 directly regulate expression of another pluripotency transcription factor, Zfp206, in embryonic stem cells. *J. Biol. Chem.* **2007**, *282*, 12822–12830. [CrossRef] [PubMed]
142. Xia, X. Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica* **2012**, *2012*, 917540. [CrossRef] [PubMed]
143. Li, G.L.; Leong, T.Y. Feature selection for the prediction of translation initiation sites. *Genom. Proteom. Bioinform.* **2005**, *3*, 73–83. [CrossRef]
144. Aerts, S.; van Helden, J.; Sand, O.; Hassan, B.A. Fine-tuning enhancer models to predict transcriptional targets across multiple genomes. *PLoS ONE* **2007**, *2*, e1115. [CrossRef] [PubMed]
145. Hertzberg, L.; Izraeli, S.; Domany, E. STOP: Searching for transcription factor motifs using gene expression. *Bioinformatics* **2007**, *23*, 1737–1743. [CrossRef] [PubMed]
146. Zakrzewska, K.; Lavery, R. Towards a molecular view of transcriptional control. *Curr. Opin. Struct. Biol.* **2012**, *22*, 160–167. [CrossRef] [PubMed]
147. Benichou, O.; Kafri, Y.; Sheinman, M.; Voituriez, R. Searching fast for a target on DNA without falling to traps. *Phys. Rev. Lett.* **2009**, *103*, 138102. [CrossRef] [PubMed]
148. Bustamante, C.; Smith, S.B.; Liphardt, J.; Smith, D. Single-molecule studies of DNA mechanics. *Curr. Opin. Struct. Biol.* **2000**, *10*, 279–285. [CrossRef]
149. Kabata, H.; Kurosawa, O.; Arai, I.; Washizu, M.; Margarson, S.A.; Glass, R.E.; Shimamoto, N. Visualization of single molecules of RNA polymerase sliding along DNA. *Science* **1993**, *262*, 1561–1563. [CrossRef] [PubMed]
150. Uphoff, S.; Kapanidis, A.N. Studying the organization of DNA repair by single-cell and single-molecule imaging. *DNA Repair* **2014**, *20*, 32–40. [CrossRef] [PubMed]
151. Gahlmann, A.; Moerner, W.E. Exploring bacterial cell biology with single-molecule tracking and super-resolution imaging. *Nat. Rev. Microbiol.* **2014**, *12*, 9–22. [CrossRef] [PubMed]
152. Manley, S.; Gillette, J.M.; Patterson, G.H.; Shroff, H.; Hess, H.F.; Betzig, E.; Lippincott-Schwartz, J. High-density mapping of single-molecule trajectories with photoactivated localization microscopy. *Nat. Methods* **2008**, *5*, 155–157. [CrossRef] [PubMed]