CHARLES UNIVERSITY

Faculty of Science

Department of Biochemistry

Mgr. Růžena Filandrová

Structural characterization of interaction between transcription factors and DNA

Strukturní charakterizace interakce transkripčních faktorů s DNA

DOCTORAL THESIS

Supervisor: RNDr. Petr Novák, Ph.D.

Consultant: RNDr. Petr Man, Ph.D.

Prague 2021

Declaration

I declare that I have worked on this thesis under the guidance of my supervisor and that all sources of the previous knowledge are properly cited. No part of this work was used and will ever be used to obtain any other academic degree than Ph.D. from Charles University.

Prague .....

.....

Mgr. Růžena Filandrová

Declaration of authorship

I declare that Mgr. Růžena Filandrová contributed significantly to the experiments and to all scientific publications contained in this Ph.D. thesis. She performed most of the experiments, substantially contributed to their planning, and took a significant part in the primary data interpretation and their preparation for publication.

Prague .....

.....

RNDr. Petr Novák, Ph.D.

Acknowledgements

At this place, I would like to express my gratitude to all the people who supported me and guided me during all the years of my PhD studies.

First and foremost, I would like to thank my supervisor, Petr Novák, for his guidance, patience and for always being open for discussion. I also appreciated very much that he allowed me to taste other aspect of the scientific work such as taking care of funding, teaching younger students, or traveling to conferences.

My thanks belong also to other members of the Laboratory of Structural Biology and Cell Signalling for creating such a friendly place to work at and for helping me whenever it was needed. Namely, I would like to thank Petr Man for all his advice concerning HDX-MS, Petr Pompach for his help with operating the mass spectrometer, Karel Vališ for performing the genome searches for us, Daniel Kavan and Zdeněk Kukačka for cheering me up whenever things did not go exactly the way I wanted and Pavla Vaňková for her friendship and for keeping me company during a number of long evenings in the lab.

I would also like to thank to my bachelor students – Jana Řeháková and Veronika Kůdelová for their contribution to my work and for their patience with me testing my teaching skills on them.

Next, I would like to express my gratitude to our collaborators from other labs: Danielle Fabris for teaching me how to operate the nano electrospray and Tomáš Chum and Marek Cebecauer for introducing me to smFRET.

I would also like to acknowledge all our sources of financial support including the Charles University Grant Agency (1618218), the Czech Science Foundation (grants 16-24309S and 19-16084S), the Ministry of Education of the Czech Republic (project LH15010 and program ‘‘NPU II’’ project LQ1604) and European Commission H2020 (EU_FT-ICR_MS and EPIC-XS grant agreement IDs: 731077 and 823839, respectively)

Finally, my thanks belong to my parents, Pavla Lišková and Pavel Liška, for giving me all the support they possibly could, to my brother, Štěpán Liška, for being there for me anytime and, of course, to my beloved husband, classmate, and colleague in one person, František Filandr, for never stopping to believe in me and for all the evenings he spent taking care of our daughter so I could write this thesis.

Abstract (in English)

Transcription factors are proteins that mediate gene expression regulation through interactions with DNA and other factors. They allow a cell to respond to various stimuli and play a crucial role in many biological processes such as control of cell cycle progression, differentiation of cells during development or immune response. To understand these processes, the knowledge of the transcription factors 3D structure together with the mechanism of their interaction with DNA is essential. However, some of the typical features of transcription factors, such as is for example the presence of intrinsically unstructured regions, make the 3D structure determination by the commonly used high resolution methods challenging. Therefore, utilization of complementary methods like structural mass spectrometry (MS), which was used in this thesis, might prove to be beneficial to explore the structural basis of the transcription factor-DNA interaction.

In first part of this work, a set of structural mass spectrometry methods with the main focus on hydrogen/deuterium exchange mass spectrometry (HDX-MS) was optimized and tested on two transcription factor-DNA complexes and their DNA binding motifs and proved to be able to provide structural information about regions of transcription factors inaccessible by the classical high-resolution methods as well as about structural dynamics of the transcription factor-DNA complex.

In the other part of the thesis, the structural mass spectrometry methods were, together with other techniques, such as smFRET, gel shift or fluorescence anisotropy, used to investigate whether and how the sequential context of the M-CAT motif and its orientation affects its interaction with the DNA binding domain of transcription factor TEAD1 (TEAD1-DBD). The obtained results have shown that the sequences of the DNA regions flanking the M-CAT motif affect its binding affinity to TEAD1-DBD and moreover, the transcription factor was found to be able to bind also to the inverted 5'-CCTTA-3' M-CAT motif, albeit with lower affinity. This low affinity interaction was then structurally characterized and might present a potential way of regulation of the transcription factor activity. Finally, the binding cooperativity of FOXO4 and TEAD1 transcription factors were studied utilizing oligonucleotides with adjacent response motifs.

Abstrakt (in Czech)

Transkripční faktory jsou proteiny, které regulují expresi genů skrze svou interakci s DNA a dalšími faktory. Tím buňce umožňují reagovat na různé vnitřní i vnější podněty a hrají proto důležitou roli v mnoha buněčných dějích jako je například regulace buněčného cyklu, diferenciace buněk během vývoje organismu nebo imunitní reakce. K pochopení těchto dějů je nezbytná nejen znalost 3D struktury samotných transkripčních faktorů, ale i mechanismů jejich vazby na DNA. Nicméně, některé typické vlastnosti transkripčních faktorů, jako je například přítomnost nestrukturovaných oblastí, způsobují, že je velmi obtížné určovat jejich 3D strukturu klasickými metodami s vysokým rozlišením. Z těchto důvodů mohou být pro popis struktury komplexů transkripčních faktorů s DNA s výhodou využity metody s nižším rozlišením, jako je například strukturní hmotnostní spektrometrie, která byla použita v této práci.

V první části této práce byl nejprve optimalizován soubor metod strukturní hmotnostní spektrometrie se zaměřením hlavně na optimalizaci podmínek vodík/deuteriové výměny (HDX-MS) pro jejich využití k analýze komplexů transkripčních faktorů s DNA. Následně pak byly pomocí těchto metod charakterizovány dva komplexy transkripčních faktorů s jejich DNA vazebnými motivy, čímž byla potvrzena schopnost testovaných metod poskytnout informace nejen o oblastech proteinu nedostupných obvykle používanými metodami ale také o strukturní dynamice celého komplexu.

Ve druhé části disertační práce pak byly metody strukturní hmotnostní spektrometrie společně s dalšími technikami jako je smFRET, nativní gelová elektroforéza nebo fluorescenční anisotropie využity v rámci studie zabývající se vlivem sekvence v okolí M-CAT vazebného motivu a jeho orientace na interakci tohoto motivu s DNA vazebnou doménou transkripčního faktoru TEAD1 (TEAD1-DBD). Bylo zjištěno, že sekvence DNA v okolí vazebného motivu má vliv na jeho afinitu k TEAD1-DBD proteinu, a navíc je tento protein schopen, i když s nižší afinitou, se vázat i na invertovanou verzi svého vazebného M-CAT motivu (5'-CCTTA-3'). Schopnost tohoto transkripčního faktoru tvořit nízkoafinní interakce s jiným vazebným motivem může poukazovat na potenciální další způsob regulace jeho aktivity, a proto byla následně také popsána strukturní podstata této interakce. V závěru práce byla také zkoumána možnost kooperativní vazby transkripčních faktorů FOXO4 a TEAD1 za využití oligonukleotidů obsahujících sousedící DNA vazebné motivy obou proteinů.

Table of contents:

Abbreviations

CRM	cis-regulatory module
CryoEM	cryo-electron microscopy
CTD	C-terminal domain of RNA Polymerase II
DAF-16	DAF-16 family member-binding element
DBD	DNA binding domain
DNA	deoxyribonucleic acid
dsDNA	double-stranded DNA
FRET	Förster resonance energy transfer
FTICR	Fourier transform ion cyclotron resonance mass spectrometer
HDX	hydrogen/deuterium exchange
HPLC	high-pressure liquid chromatography
HT-SELEX	high-throughput systematic evolution of ligands by exponential enrichment
ChIP	chromatin immunoprecipitation
ChIP-Seq	chromatin immunoprecipitation combined with DNA sequencing
K_D	dissociation constant
LC	liquid chromatography
M-CAT	muscle CAT motif
MD	molecular dynamics
mRNA	messenger RNA
MS	mass spectrometry
MS/MS	tandem mass spectrometry
Nep1	nepenthesin 1
Nep2	nepenthesin 2
nESI	nanoelectrospray ionization
NMR	nuclear magnetic resonance
Pep	pepsin
P-TEFb	positive transcription elongation factor b
qPCR	quantitative polymerase chain reaction
RNA	ribonucleic acid
Rpn	rhizopuspepsin
SAXS	small-angle X-ray scattering
SDS-PAGE	polyacrylamide gel electrophoresis in the presence of sodium dodecylsulphate
smFRET	single-molecule Förster resonance energy transfer
snRNA	small nuclear RNA

ssDNA	single-stranded DNA
TCEP	tris(2-carboxyethyl)phosphine
TF	transcription factor
TFBS	transcription factor binding site
tPt	trans-dichlorodiamineplatinum(II)
TROSY	transverse relaxation optimized spectroscopy
TSS	transcription start site
UHPLC	ultra high-pressure liquid chromatography
UV	ultraviolet radiation
YBD	YAP binding domain

1. Introduction

1.1. Transcription Factors

The term "transcription factor" is commonly used for proteins that are both capable of binding a specific DNA sequence and able to activate or repress the initiation or elongation phase of synthesis of RNA from DNA template or, in other words, to regulate transcription[1–3]. With these properties, the transcription factors allow a cell to respond to diverse stimuli by promoting expression of specific set of genes. They play a crucial role in many biological processes such as control of cell cycle progression, differentiation of cells during development, immune response or maintenance of intracellular metabolic balance[4–7]. And finally, the general transcription factors are proteins that form the core initiation complex with RNA Polymerase II which is needed for literally every gene expression event to start[3].

In 2018 there were 1639 known human proteins identified as a transcription factor which represent roughly 8 % of all human genes and this number is still expected to grow as new transcription factors are discovered every year[2]. Even though their number seems high, the processes they regulate are so complex, that it could never be sufficient if transcription factors worked on their own – and they in fact almost never do so. Transcription factors can bind to DNA in cooperation with each other and also interact with numerous cofactors to fine-tune the gene expression regulation so the cell will be able to produce the proteins it needs in precisely the right moment[2,8]. Since the transcription regulation is such a complicated process, it is no surprise that mutations in transcription factor genes, their binding sites or errors in their regulation by cofactor presence can lead to various diseases. For example, transcription regulators and nucleic acid binders are significantly over-represented in cancer genes and approximately a third of human developmental disorders is caused by mutation in transcription factor genes[2,9,10]. However, the exact way of how the network of transcription factors and cofactors works and how their combinations affect DNA binding and transcription output is not yet properly understood. That is why the interactions of transcription factors with DNA, as is the case in this work, or other proteins are currently a frequently studied topic.

1.1.1. Transcription-control Regions

Transcription factors bind DNA in a sequence specific manner. Typically, they bind to 4-8 base pairs long protein-binding DNA sequences[8], which are situated in DNA regions that are generically referred to as control elements and could be located either in the genomic region directly preceding the transcription start site (promoter-proximal elements) or up to several megabases away from the promoter in a 50-200 bp long region called an enhancer. Their typical distribution is shown in Figure 1, however, the exact boundaries between promoter-proximal elements and enhancers eventually became blurred with increasing number of discovered control elements. Transcription factor binding sites can also appear as a cluster that forms a cis-regulatory module[11,12].

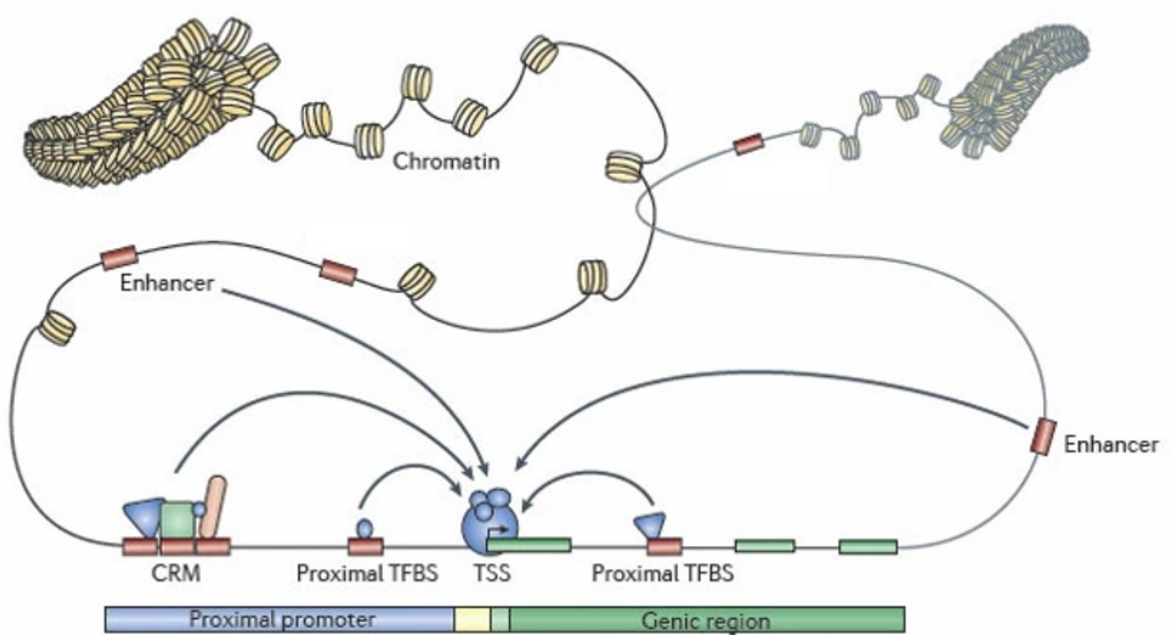


Figure 1: A summary of transcription control elements. *In active genes, chromatin structure must be accessible for proteins. The region around transcription start site (TSS) is called promoter and transcription factor binding sites (TFBS) located near there promoter-proximal elements in contrast with further located enhancers. TFBSs appearing in clusters can form a cis-regulatory module (CRM) [12].*

The very first enhancer (non-coding genomic region with the ability to enhance transcription) was identified as a 72-bp long sequence originating from SV-40 virus that was able to increase transcription of β-globin gene in HeLa cells[13]. Later on, enhancer activity has been shown to correlate with several properties of chromatin such as histone H3 K4 methylation or K27 acetylation. Active enhancers must be accessible for proteins and

therefore free from nucleosomes that are chromatin's base structural unit, which is in agreement with mentioned histone post-translational modifications and DNase I accessibility being well-known markers of active chromatin[14].

Another typical property of enhancers is their ability to act almost independently of the distance and orientation to their target genes. There were several theories trying to explain how this is achieved but the most widely accepted one finally became the "looping" model, according to which the transcription factors bound to enhancers form direct contact with promoters and the proteins forming the preinitiation complex while the DNA between them loops to make the contact possible. The whole structure is then stabilized by cohesin and other proteins[15–17]. However, recent studies suggests that even this model might not be entirely correct and the contacts between enhancer and promoter bound proteins can be more dynamic and flexible than what was believed[8,18]. Finally, like transcription factors, the enhancers can function in cooperation with each other as well, or even be partially redundant resulting in additive effect to gene expression regulation[14].

1.1.2. Mechanisms of Transcription Regulation by Transcription Factors

1.1.2.1. Chromatin Structure Modulation

For a transcription factor to enforce its role in gene expression regulation it must first bind to its DNA response motif. Nevertheless, chromatin at inactive enhancers is usually closed and the presence of nucleosomes at the binding sites is preventing transcription factors from reaching their DNA response motifs[19].

There are several possible mechanisms of how transcription factors overcome this inactivation. One of them relies on the cooperativity between transcription factors. Multiple transcription factors that recognize binding sites located close to each other within an enhancer can compete with proteins of the nucleosome for DNA binding even without forming direct protein-protein interactions. This mechanism is called "collaborative binding" and can be further enhanced by direct protein-protein interactions when two or more transcription factors bind DNA together as a dimer, multimer or, in the extreme case, an enhanceosome[8,20]. Another possible mechanism utilizes a special kind of transcription factors called "pioneer factors" such as is for example FoxA or Sox2[21]. These transcription factors have the ability to bind to nucleosomal DNA. Once bound there, they can interact

with other factors, mainly chromatin remodelers and histone-modifying complexes, to either promote or repress chromatin opening and thus binding of other transcription factors[8,22]. Chromatin remodelers are multi-protein complexes containing a catalytic ATPase subunit similar to DNA translocases that alter the structure, composition or positioning of nucleosomes[23]. Histone-modifying complexes, on the other hand, alter chromatin structure by covalently modifying specific amino acid residues in the histone tails which may directly affect chromatin compaction or create docking sites for chromatin remodelers[24]. Finally, there is also evidence, that the pioneer factors can interact directly with RNA polymerase II[25].

1.1.2.2. Control of RNA Polymerase II Function

Out of the three RNA polymerases that were identified in eukaryotic cells, RNA polymerase II is the enzyme responsible for synthesis of messenger RNA (and various non-coding RNAs) whereas the other two transcribe mostly ribosomal and transfer RNA. The control of RNA Polymerase II function is therefore crucial for gene expression regulation[26]. Alike to what was mentioned in the previous chapter for transcription factor binding, even RNA polymerase II first needs to gain access to promoter region. Although there is a class of promoters containing CpG islands that is mostly open for RNA polymerase II[27], other classes depend on the function of transcription factors and chromatin modulating enzymes to guide the RNA polymerase II to its target promoter[12,26].

It has been known for some time, that transcription factors use their transactivation regions to recruit RNA polymerase and other proteins needed for transcription initiation to promoters[28]. However, only recently a model of how this regulation works was proposed (Fig.2) and proofs started to show up, that the disordered transactivation regions of transcription factors form phase-separated condensates with similarly disordered parts of mediator complex (transcriptional co-activator that stabilizes the preinitiation complex of RNA Polymerase II and general transcription initiation factors)[29] and C-terminal domain of RNA Polymerase II (CTD)[30]. This results in concentration of these factors at promoters and facilitates the preinitiation complex formation[26]. Apart from this, transcription factors and co-factors may also affect the initiation phase by forming direct protein-protein interactions with the mediator complex and through it facilitate or repress the preinitiation complex formation or phosphorylation of CTD by the TFIIH's CDK7 kinase which is needed for the

elongation phase to start[31,32].

Shortly after transition to elongation phase, another event that is highly regulated by transcription factors occurs– the promoter proximal pausing. This happens typically between 20-120 bp downstream from the transcription start site[32] and involves creation of local RNA/DNA hybrids of the newly synthetized RNA and the template DNA strand which leads to RNA Polymerase II pausing and backtracking[33]. The paused state is further stabilized by factors DSIF and NELF that can be later released by SPT6 and PAF complex factors after being, along with CTD, phosphorylated by the positive transcription elongation factor b (P-TEFb)[26]. The release factors can be recruited to their place by transcription factors such as for example BRD4 or C-MYC or inhibited by the non-coding snRNA 7SK and proteins interacting with it[34,35]. When RNA Polymerase II is released from pausing, more transcription factors assist with recruitment of positive elongation factors, RNA processing factors and chromatin modifiers that allows unwinding of nucleosomes and continuation of mRNA synthesis[26,32].

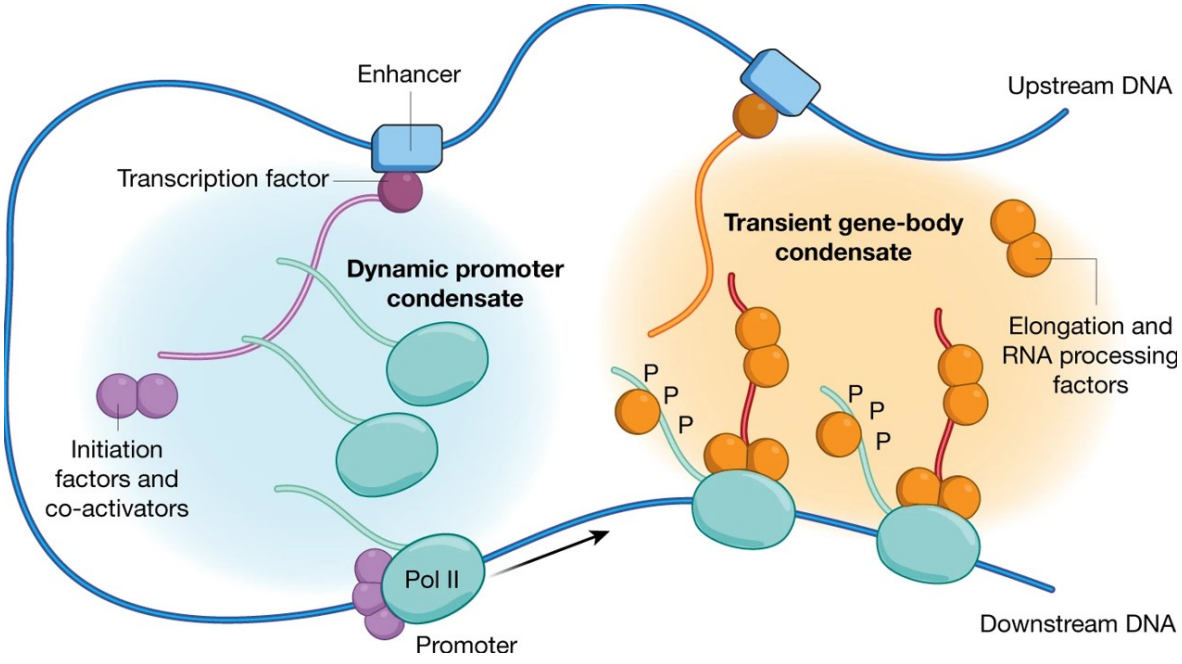


Figure 2: Condensate based model of transcription. *The promoter condensate is formed by transcription factors bound to enhancers which recruit RNA Polymerase II with unphosphorylated C-terminal domain and its co-activating factors. Upon phosphorylation of C-terminal domain RNA polymerase II transfers to elongation phase and associates with elongation and RNA processing factors. It is predicted to form another condensate with assistance of other transcription factors in this phase.*[26]

1.1.3. Structure of Transcription Factors

Transcription factors are typically modular in structure, which means they consist of several structural domains with different functions. A typical transcription factor comprise of one or more DNA binding domains (DBD) that binds a specific DNA sequence and one or more effector domains which might serve either to mediate their transcription regulating function through various mechanisms or to regulate the activity of the transcription factor itself (Fig. 3)[2].

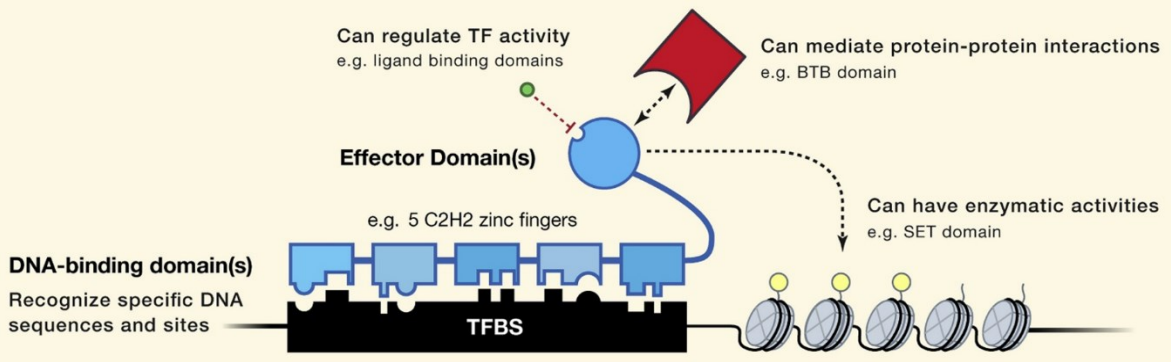


Figure 3: Structure of a typical transcription factor*. A DNA binding domain is used to recognize and bind to a specific DNA sequence whereas an effector domain may use various mechanisms to affect transcription*[2].

1.1.3.1. DNA Binding Domain

As the name suggests, this domain serves mainly to make a sequence specific contact with DNA, but some of these domains are also able to interact with other proteins and thus bind to DNA as a homo or a heteromultimer[36]. Nevertheless, the majority of transcription factors bind to DNA as a monomer or homomultimer[2]. Structure and sequence specificity of DNA binding domains is highly conserved throughout evolution – for example transcription factor orthologs between human and *Drosophila melanogaster* have practically the same sequence specificity and their function also tends to be similar[37].

The structure of DNA binding domain is also a parameter according to which can transcription factors be sorted to families of members using the same DNA binding mechanism. Currently, there is around 100 known DBD families but this number is probably not final since it has been growing for some time and there are still some proteins that have been identified as a transcription factor, but their DNA binding domain is not known. Most of the known human transcription factors carry one of two DBD types – C2H2 zinc fingers or a homeodomain, which are in quantity of members followed by transcription factors

featuring helix-loop-helix, leucine zipper and forkhead DBD types[2].

As was mentioned before, each of the transcription factor families has its own structural motif which they use to bind DNA. The binding is usually a result of a combination of noncovalent interactions between specific amino acids on the interaction interface (which in many cases includes an α-helix inserted to DNA's major grove) and bases of the DNA response motif or the DNA's sugar-phosphate backbone. Nevertheless, sequence independent noncovalent interactions may also contribute to the binding[3]. For example, the second most common DBD type after the C2H2 zinc-finger (which would not be a very good typical example due to the fact that one transcription factor usually contains multiple C2H2 zinc-finger domains and binds to long (20-40 bp) DNA sequences[38]), the homeodomain, consist of three α-helices of which the third (C-terminal) is inserted into DNA's major grove and tree amino acids residues from this helix interact with bases of the TAAT core binding motif by forming one hydrogen bond and several van der Waals contacts. The whole complex is further stabilized by several water-mediated contacts together with ionic interaction between amino acid side chains and the sugar-phosphate backbone[39]. Similar mechanism where some (usually highly conserved) residues make a sequence specific contact with DNA while other residues strengthen the bond by sequence independent interactions with DNA exist for most transcription factors with known structure of the protein-DNA complex[2].

1.1.3.2. Effector Domain

In contrast with the DNA binding domains that have only one single purpose, the effector domains utilize several mechanisms that are very different from each other to achieve its final goal – to activate or repress transcription. As was explained in chapter 1.1.2., the transcription factors can affect transcription in various points including opening of chromatin, recruitment of the basal transcription machinery or release of RNA Polymerase II from pausing and each effector domain can affect one or more of these processes.

The effector domain that interacts with components of the preinitiation complex are called transactivation domains. Many transactivation domains make direct protein-protein contact with the general transcription factors or components of the mediator complex and can be classified according to their amino acid composition. These domains can be rich in either acidic residues, glutamine residues or proline residues which can, to a certain degree, be

used to predict their binding preferences[40]. From the point of view of the 3D structure, the transactivation domains often contain unstructured regions than only become structured upon binding of their interaction partner (a general transcription factor or another protein, DNA, or a small molecular ligand) which serves as a template for its shaping. By this mechanism, the unstructured region may even fold differently depending on the type of the ligand or upon being post-translationally modified[41,42]. Apart from the basal transcription machinery, some effector domains can also interact through specific interfaces with histone modifying enzymes to facilitate or repress access to transcription start sites[40,43].

Another equally important type of effector domain is one that allows transcription factors to cooperatively bind DNA. The dimer (or multimer) can be, depending on the type of effector domain, formed either in solution ahead of DNA binding or following the binding of one of the factors to DNA. Such cooperative binding vastly increases the pool of sequences recognized by transcription factors as it introduces various motif combination and even allows a transcription factor to bind to a sequence not quite corresponding to its preferred response motif. This cooperative binding is also a key part of combinatorial regulation which enables the cell to integrate signals from different pathways to adequately react to external conditions[40].

1.1.4. Low Affinity Binding Sites

In the previous chapter, the possibility of a transcription factor binding to a suboptimal DNA sequence was mentioned. These suboptimal binding sequences have lower affinity to the transcription factor than the optimal consensus binding motif and thus are called low affinity binding sites. As a low affinity binding site is usually considered a sequence that is bound up to 1000-fold more weakly than the optimal DNA response motif but still more strongly than a random DNA sequence[44]. For a long time it was believed, that presence of low affinity binding sites in genome does not have any functional relevance, however, in the last decade a number of studies emerged showing that they might play a very important role in explaining the so called transcription factor specificity paradox[44–47]. This paradox concerns the fact, that eukaryotic transcription factor families usually contain paralogs with very similar DNA binding preferences and yet they affect transcription of different set of genes. Partially it can be explained by cell type specific expression of each paralog but sometimes different paralogs are expressed in the same cell and still they preserve the ability to

distinguish between the affected genes[44,45]. It was also shown, that low affinity sites are involved in distinguishing whether a single transcription factor will behave as an activator or a repressor[44,46].

There are several possible ways of how a transcription factor can bind to a low affinity site or even prefer it to the optimal one. For instance, binding to DNA in complex with an interaction partner can change transcription factor's structure and thus its binding specificity can be shifted to previously not preferred DNA sequence[45,48]. Moreover, spacing between the interaction partners binding sites can in some cases compensate for poor binding affinity[49]. Another possible affinity modifier are epigenetic modifications of DNA, especially the CpG methylation which can alter the binding affinity even in the paralog specific manner[50]. Furthermore, the intrinsic DNA shape of the binding site may also be the source of different paralog specific binding affinity in spite of the fact, that the final complex structure is the same[51]. And finally, there is the factor of local transcription factor concentration. The binding site occupancy is a result of combination of two factors – the binding site affinity (described by the dissociation constant of protein-DNA complex) and transcription factor concentration. Therefore, depending on the concentration and K_D the transcription factor can be either unbound (when concentration of free transcription factor is much lower than K_D), partially bound (when both parameters are approximately equal) or the binding site might become saturated in case that the transcription factor concentration is much higher than K_D[44]. This provides a handy mechanism for concentration dependant transcription regulation where at low transcription factor concentrations only high affinity sites are occupied and when the concertation rises, the transcription factor starts to bind the low affinity sites as well. Such mechanism was already described for example for the MYC transcription factor[52,53]. The binding to a low affinity site can be further enhanced by presence of so-called transcriptional hubs in the nucleus (Figure 4). The hubs are clusters of regulatory elements and the basal transcription machinery components (probably created as a result of phase separation promoted by the disordered regions of transcription factors which was described above) that form a compartment of roughly 100 nm in diameter inside which the presence of a single transcription factor molecule can lead to a concentration high enough to bind even to low affinity sites[44,54,55].

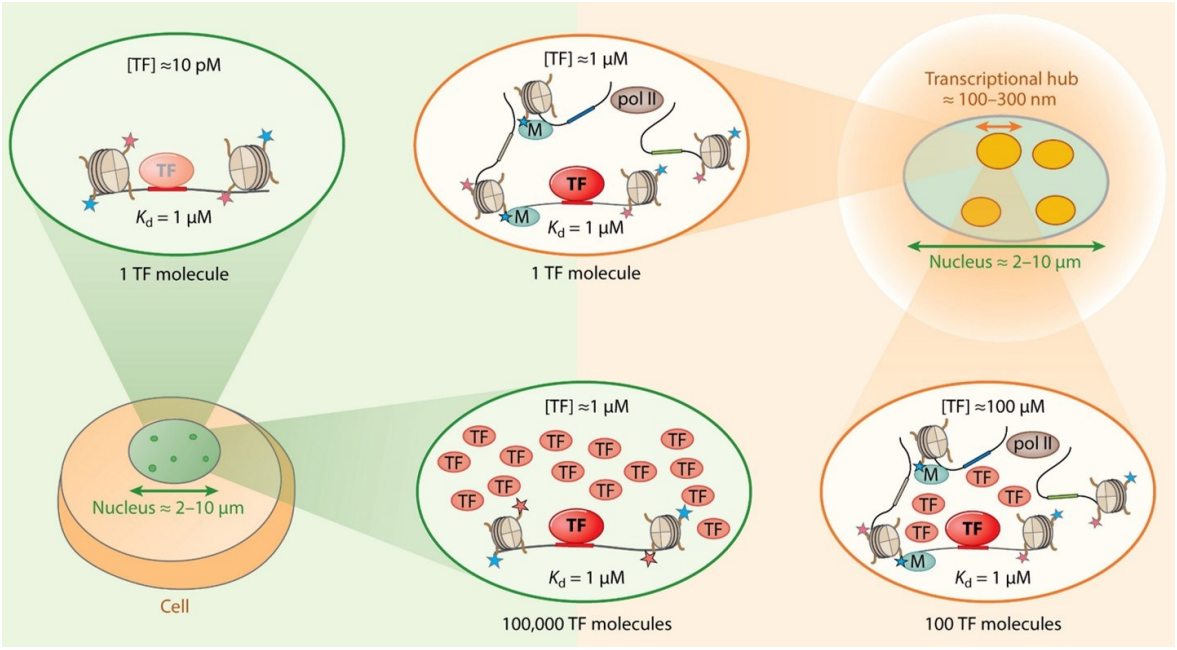


Figure 4: The effect of transcriptional hubs presence on low affinity sites occupancy. *In the green half of the picture, situation inside a nucleus with no compartments is shown. At low transcription factor (TF) concentration (corresponding to presence of only one TF molecule) the low affinity (K_D 1 μM) site would be unbound and much more TF molecules would be needed for this site to become occupied. However, as is shown in the orange part of the picture, restricting one or several TF molecules to a small compartment leads to high enough local concentration to respectively bind or saturate even the low affinity site*[44].

1.2.TEAD Family of Transcription Factors

Transcriptional enhancer associated (or activator for both possibilities are used in the literature) domain (TEAD) is a family of transcription factors that share a highly evolutionary conserved DNA binding domain – the TEA domain[56]. The first member of this family, the transcriptional enhancer factor 1 (TEF1, later named TEAD1), was identified as a small nuclear protein which was able to bind to the GT-IIC motif (5'-ACATTCCAC-3') of the SV40 enhancer in HeLa cells with the ability to upregulate transcription[57]. Later on, other proteins containing the TEA domain were identified throughout different eucaryotic species and the domain is therefore sometimes also called ATTS after the first four known (AbaA and Tec1 in yeast, Tead1 in vertebrates and Scalloped in drosophila)[58]. Comparison of the identified TEAD family proteins showed a high degree of similarity in the DNA binding region revealing the evolutionary conservation of the TEA domain and suggesting its possible importance for development of eucaryotic organisms[56,59].

1.2.1. Function of TEAD proteins in mammals

1.2.1.1. Isoforms and their physiological function

In mammals, the TEAD family consists of four members and each of them is known by several names: Tead1 (Tef-1/Ntef), Tead2 (Tef-4/Etf), Tead3 (Tef-5/Etfr-1), and Tead4 (Tef-3/Etfr-2/Fr-19)[60]. All four members of the family share the same domain structure, which will be discussed in more detail in chapter 1.1.2. They consist of N-terminal DNA binding domain and C-terminal transactivation (or YAP binding) domain, connected with an unstructured linker. All mammal TEADs also express a high degree of homology, especially in the DNA binding domain and YAP binding domain regions (Figure 5)[61]. However, the TEAD isoforms differ in their tissue and stage of development expression patterns, which were most thoroughly studied in mice. Every tested mouse tissue was found to express at least one TEAD protein in a certain point in development, while others were shown capable of expressing all of them[62,63]. Gene inactivation studies in mice also provided an insight into function of each TEAD isoform.

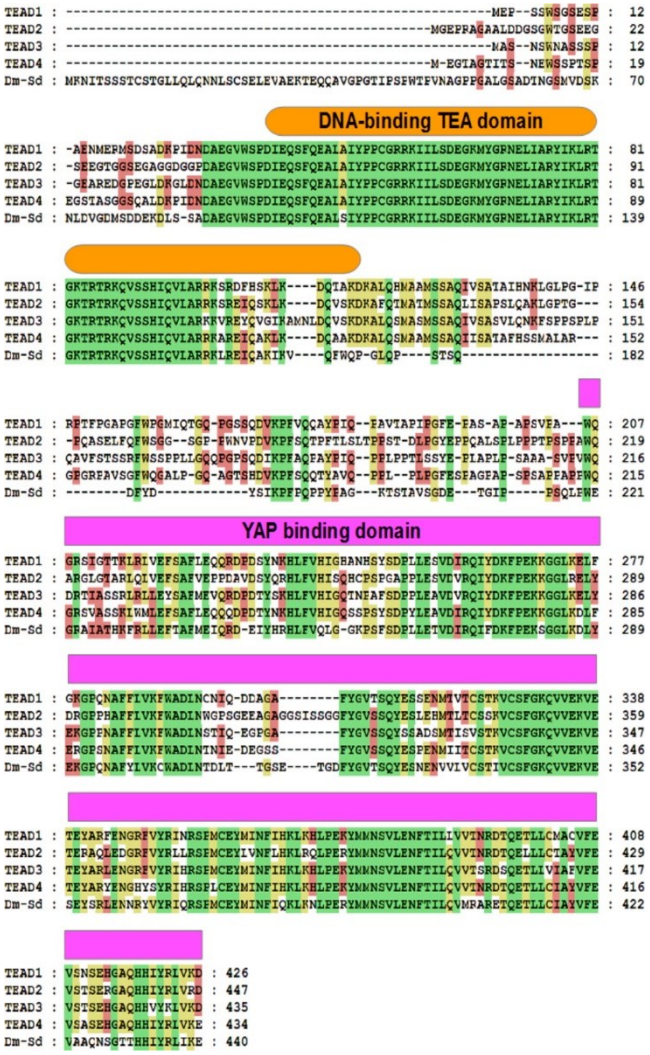


Figure 5: Sequence alignment of the four isoforms of human TEAD and drosophila Scalloped. *Residues highlighted in red are identical among 3 out of 5 sequences, yellow are identical among 4 out of 5 sequences, and green are identical among all the 5 sequences[61].*

Tead1 is in adult mice expressed in many organs and tissues including lungs, heart, kidney, liver, brain or skeletal muscle[64,65]. During embryonic development it has a very important function in regulation of cardiac muscle growth and myocardium differentiation

which manifested in the mice embryos lacking the *Tead1* in the form of lethal heart defects[58,66]. In adult heart, Tead1 is needed for maintaining its normal function[67]. Apart from the cardiac muscle, Tead1 is also involved in other processes such as skeletal and smooth muscle formation and function since it regulates α-actin and α and β myosin heavy chain genes[58] or in development of the neural system through regulation of *Foxa2* gene[68]. In humans, an inactivating missense mutation of *TEAD1* gene is a cause of a genetic disorder called Sveinsson's chorioretinal atrophy. In some cases, other Tead proteins might fill in for the missing Tead1, nevertheless, this is not possible always[68].

Tead2 is the first to be expressed and most abundant TEAD protein during the first seven days of embryonic development[69]. It is expressed in almost every tissue and its most important function is probably regulation of neural crest cells differentiation through affection of *Pax3* gene[58]. Later on, its concentration in organism decreases and its expression starts to be limited to only several tissues such as brain or lungs[69]. Interestingly, it was found, that Tead1 can compensate for Tead2 absence during the first stages of embryonic development. While mutant mouse embryos lacking only the *Tead2* gene appeared normal, the ones that lacked both *Tead1* and *Tead2* genes showed severe developmental defects, especially in the notochord, which appeared sooner than the heart defects caused by missing *Tead1* gene[58,68].

Tead3 remains the least explored of the TEAD family proteins. In both mice and humans, it is expressed mainly in extraembryonic tissue which later forms placenta. In the mouse embryos it could only be detected in later stages of development, mostly in neural and muscle tissues[65]. However, no gene inactivation study has been yet reported for this protein.

Tead4 inactivation in mice disrupted specification of trophoectoderm (a precursor of placenta) and caused the embryos to fail to implant. Nevertheless, if Tead4 was inactivated after implantation, the mice developed normally[62]. Initially, it was thought that Tead4 regulates expression of several trophoectoderm specific genes needed for trophoectoderm cells differentiation[70]. However, a later study provided a proof, that Tead4 is the only one of TEAD proteins present not only in nucleus but also in the mitochondria and during the differentiation of trophoectoderm and that its main role lies in maintaining energy homeostasis and preventing excess accumulation of reactive oxygen species[71].

1.2.1.2. Role of TEAD proteins in cancer and their regulation

Since the deregulation of cell proliferation, growth, differentiation or apoptosis are well known properties of tumorigenesis and given the involvement of TEAD proteins in all of these processes described in the previous chapter, it is no surprise, that the TEAD proteins have been heavily studied in this context[60,62]. In addition to the previously mentioned target genes, TEAD proteins were found to upregulate expression of numerous other genes connected with cell proliferation such as are *CYR61* and *CTGF* which both affect cell migration and adhesion[72,73], anti-apoptotic genes *AXL, Livin or Survivin*[58,74], genes known to be oncogenes or even tumor markers (*MYC, Mesothelin*)[75,76] or genes encoding glucose transporters *GLUT1* and *GLUT3* which are needed by quickly growing cells to satisfy their energy needs[77,78] and many others[60]. On top of that, increased TEAD activity was observed in multiple types of solid tumors including prostate cancers[79], colorectal cancers[80], breast cancers[81] or gastric cancers[82] and in some of them it was also identified as a marker of poor prognosis[60,79]. Therefore, given the involvement of TEADs in cancer development, it is clear, that their activity in organism must be strictly regulated to prevent unchecked cell proliferation.

Soon after the initial discovery of the first TEAD protein, an unusual property typical for this family of transcription factors was found. Although they possess the ability to bind DNA, they are not able to activate transcription on their own and can only do that through interaction with other proteins – their coactivators[83,84]. This provides a handy mechanism through which can the activity of TEAD proteins be regulated in reaction to different signals and conditions. Among the identified coactivators, YES-associated protein (YAP) and its paralog TAZ (transcriptional coactivator with PDZ-binding motif) are the two most well-established[85,86]. They both form nuclear complexes with all TEAD proteins and together they are the main effectors of the Hippo signalling pathway which plays a major role in organ size control, cell proliferation and tumorigenesis[73,87]. This pathway regulates the nuclear localization of YAP/TAZ and therefore its availability for interaction with the DNA-bound nuclear TEAD proteins. Various upstream signals including mechanical signals, cellular stress, extracellular stimuli or cell-cell contact are transferred trough a cytoplasmatic cascade of kinases to large tumor suppressor kinase 1/2 (LATS1/2) which, when activated, phosphorylates YAP or TAZ[88]. Phosphorylation of YAP/TAZ by activated LATS1/2 results, depending on the phosphorylation site, either in cytoplasmic sequestration due to binding to

14-3-3 protein or ubiquitylation and consecutive degradation[84]. Therefore, when the Hippo pathway is active, YAP/TAZ is phosphorylated and not available for TEAD binding whereas in case of deactivation of the Hippo signalling (either due to a defect in some of its components or as a result of physiological signals), YAP/TAZ is unphosphorylated and able to enter nucleus, bind a TEAD protein and increase expression of its target genes[88]. Apart from the Hippo pathway, other signalling pathways connected with cancer cells proliferation were recently found to regulate TEAD activity by using their interaction with YAP/TAZ coactivators both independently or in a crosstalk with the Hippo pathway[60]. This includes the Wnt/β-catenin pathway[89] and alternative Wnt pathway[90], the TGFβ pathway[91] or the LKB1-AMPK signalling[92] (for complete summary of TEAD regulation see Figure 6).

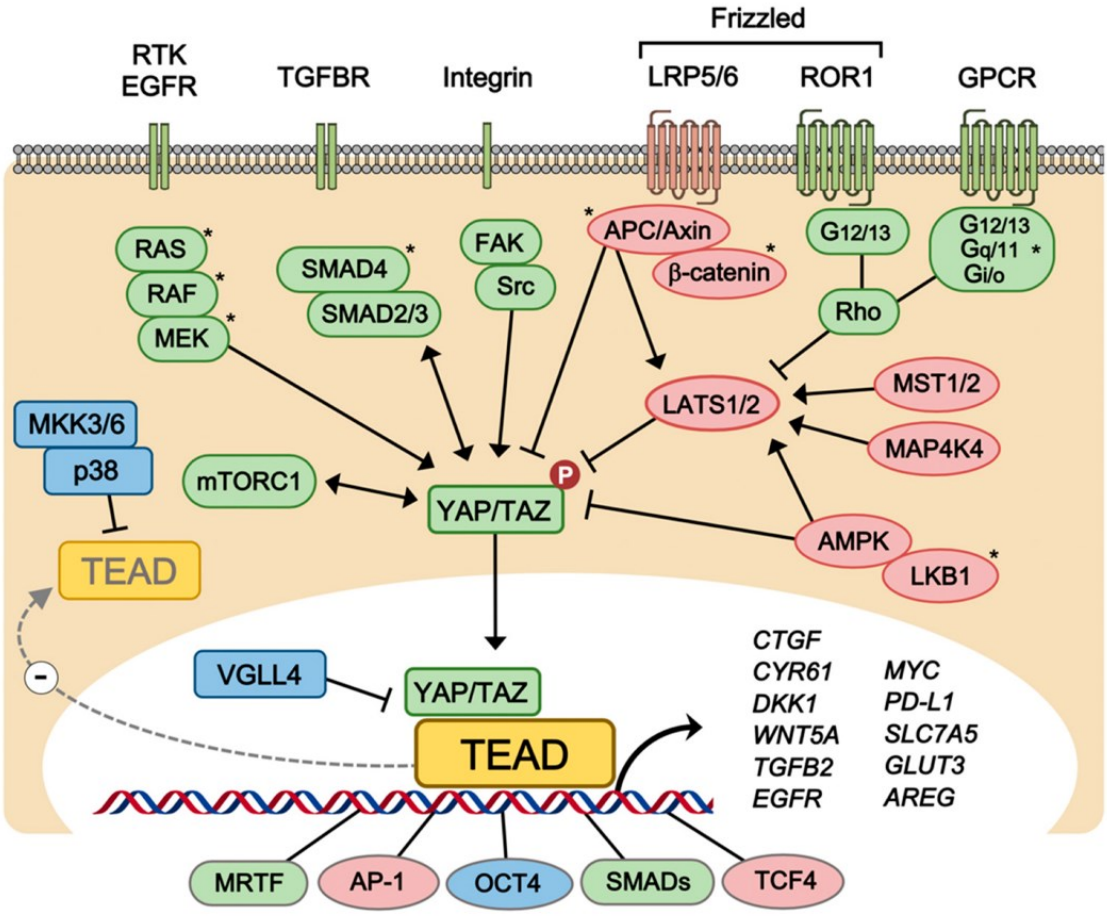


Figure 6: Overview of the regulatory mechanisms of TEAD proteins. *An interplay of signalling pathways and nuclear coactivators is needed for TEAD to either activate or repress transcription of various pro-proliferation genes. Proteins connected with Hippo pathway dependent inhibition of TEAD are highlighted in red, proteins known to activate TEAD are green and those known to inhibit its activity independently of the Hippo signalling are blue. The asterisk labels known oncoproteins*[60].

In addition to YAP/TAZ, other coactivators that act independently on the Hippo pathway were identified. Out of them, the most well-studied is the Vestigial-like protein (VGLL) family whose member VGLL4 was found not only to be able to bind TEAD in the nucleus, but also to do so on a similar interface as YAP/TAZ and thus compete for its binding site on TEAD with it. Increased concentration of VGLL4 was even found to inhibit the effect of deregulated Hippo pathway causing increased nuclear translocation of YAP/TAZ by blocking its binding site on TEAD[93]. Other identified coactivators of TEAD protein then include: p160 family of steroid receptors, serum response factor (SRF), poly-ADP ribose-polymerase (PARP), activator protein-1 (AP-1), myocyte enhancer factor 2 (MEF2), or C-MYC interaction partner MAX[84].

Another way how the activity of TEAD proteins might be regulated is by posttranslational modifications. To date, three such modifications appearing in vivo were described. TEAD proteins might be phosphorylated on serine and threonine residues of the third helix of the DNA binding region by protein kinases A or C and the phosphorylation in both cases results in disruption of DNA binding[94,95]. On the other hand, palmitoylation of cysteine residues in a hydrophobic core of YAP binding domain was found to be crucial for proper folding and stability of TEAD proteins and therefore essential for its physiological function[96].

Finally, a cytoplasmatic translocation of TEAD proteins can be induced as a reaction to cellular stress. It is facilitated by the p38 MAPK pathway where TEAD forms a complex with p38 and is subsequently translocated to cytoplasm. As a result, YAP/TAZ are unable to activate their target genes even when they are not phosphorylated and therefore present in the nucleus[97].

1.2.2. Structure of TEAD proteins

As was already mentioned, all the four mammalian TEAD proteins share the same domain architecture. They consist of two main structural domains (DNA and YAP binding) which are both highly conserved plus two, more variable, unstructured regions. One of the variable regions is present on the very N-terminus and is followed by a DNA binding domain which is then connected by a hydrophobic region rich in proline to a C-terminal transactivation (or YAP/TAZ binding) domain (Figure 7). Although no high-resolution structure is yet available for the full-length protein, structures of the individual domains have been separately solved for some members of the family[58,98].

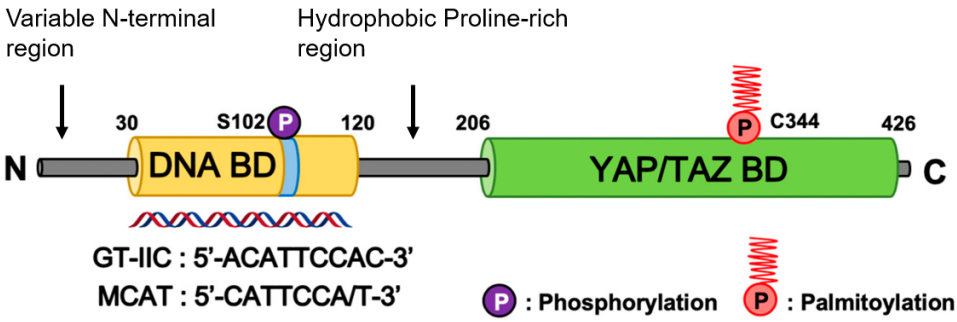


Figure 7: Domain architecture of human TEAD1 protein. *The two highly conserved DNA binding and YAP binding domains are highlighted in yellow and green, respectively. Unstructured variable regions are shown in grey. Adapted from*[58,60].

1.2.2.1. DNA binding (TEA) domain

The first high-resolution structure of a DNA-free TEA domain (TEAD-DBD) in solution was solved for residues 28-104 of human TEAD1 by using NMR. It was found to be a folded globular protein consisting of three α-helices (H1, H2, and H3) connected by two loops (long L1 and a shorter L2). H1 and H2 are nearly antiparallel and each of them folded on the opposite side of the N-terminal end of H3 (Figure 8A). This study also determined the affinity of an isolated TEAD-DBD to DNA to be in nanomolar range, consistent with what was previously found for the full-length protein. This demonstrated, that the TEA domain on its own is the source of DNA binding ability of TEAD proteins[59].

The NMR study provided the first insight to the position of the protein-DNA interaction interface as well. Helix H3 and the L2 loop immediately preceding it were identified as the DNA recognition region[59]. This knowledge was later expanded by two studies that used X-ray crystallography to solve the structure of TEAD-DBD. First, a crystal structure of TEAD1-DBD missing the longer L1 loop was published. The TEAD1-DBD missing the L1 loop quite surprisingly formed a helix swapped homodimer while the H1 helix was swapped between the monomers. But more importantly, the L1 loop was found to be involved in DNA binding (particularly in cooperative binding to tandemly duplicated elements) as well[99]. Finally, a crystallographic structure was published for the whole human TEAD4-DBD (residues 36-139) in complex with DNA which confirmed, that both previously identified regions (helix H3 and L1 loop) form the DNA recognition interface. In the TEAD4-DBD·DNA complex, helix H3 is inserted into DNA's major grove where it is held by specific non-covalent interactions (mainly hydrogen and salt bridges) between H3 residues and bases of the DNA recognition motif while the L1 loop makes a sequence independent

contact with the minor grove where it stabilizes the complex mostly by hydrophobic packing. The main structural difference between the free and bound forms of TEAD-DBD is found in the H3 helix, which is in the bound state prolonged and rotated 30° relative to helices H1 and H2 to better fit in the DNA's major grove (Figure 8B)[100].

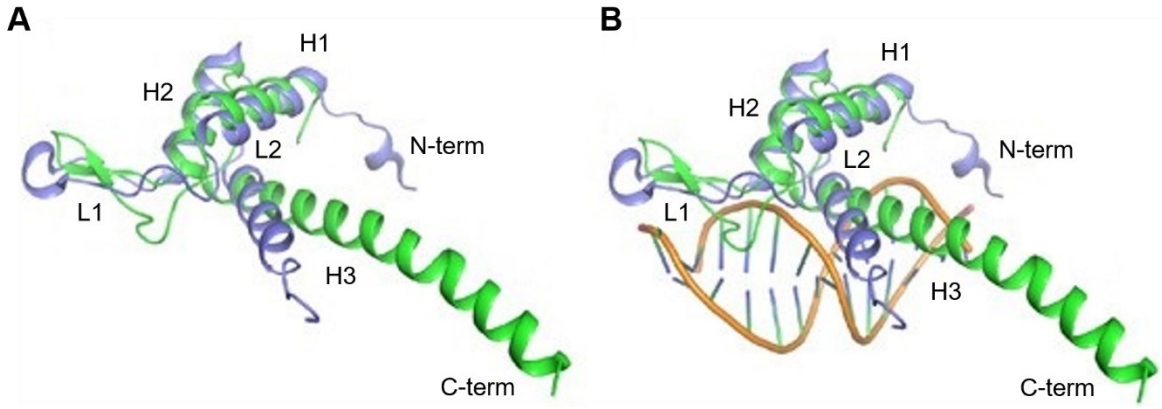


Figure 8: Structural superposition of TEAD-DBD in apo state (A) and in complex with DNA (B) *Comparison of apo state TEAD1-DBD solved by NMR (blue) and crystallographic structure of DNA-complexed TEAD4-DBD (green). Helix H3 and loop L1 were identified as the DNA recognition regions. Helix H3 is prolonged and 30° rotated in the bound form. Adapted from*[100].

1.2.2.1.1. Binding motif

The TEAD1 protein was initially identified bound to the GT-IIC motif of the SV40 enhancer in HeLa cells whose sequence is 5'-ACATTCCAC-3'[57]. Subsequently, similar TEAD binding motives were identified in many muscle specific genes (first example was cardiac troponin T) and thus a consensus sequence of 5'-CATTCCT-3' called M-CAT (muscle CAT) was established[101–103]. Its relationship with TEAD1 was further confirmed by using a protein-binding chip derivatized with randomized DNA duplexes designed to identify the sequence with the highest affinity for TEAD. The results lead to a sequence corresponding to ANATVCZN, in which V can be A, T, or G; Z can be A, T, or C; and N can be any base[59]. Finally, with the introduction of the high throughput sequencing techniques (ChIP-Seq, HT-SELEX) for identification of transcription factor binding sites, the consensus binding motif of TEAD proteins was shortened to 5'-ATTCC-3' (Figure 9)[104].

Shortly after their discovery it was also found, that TEAD proteins bind only double stranded DNA and not the single stranded one, a property that was later explained by X-ray crystallography that identified several hydrogen and salt bridges being formed between H3 helix residues and bases of both DNA strands[100,105].

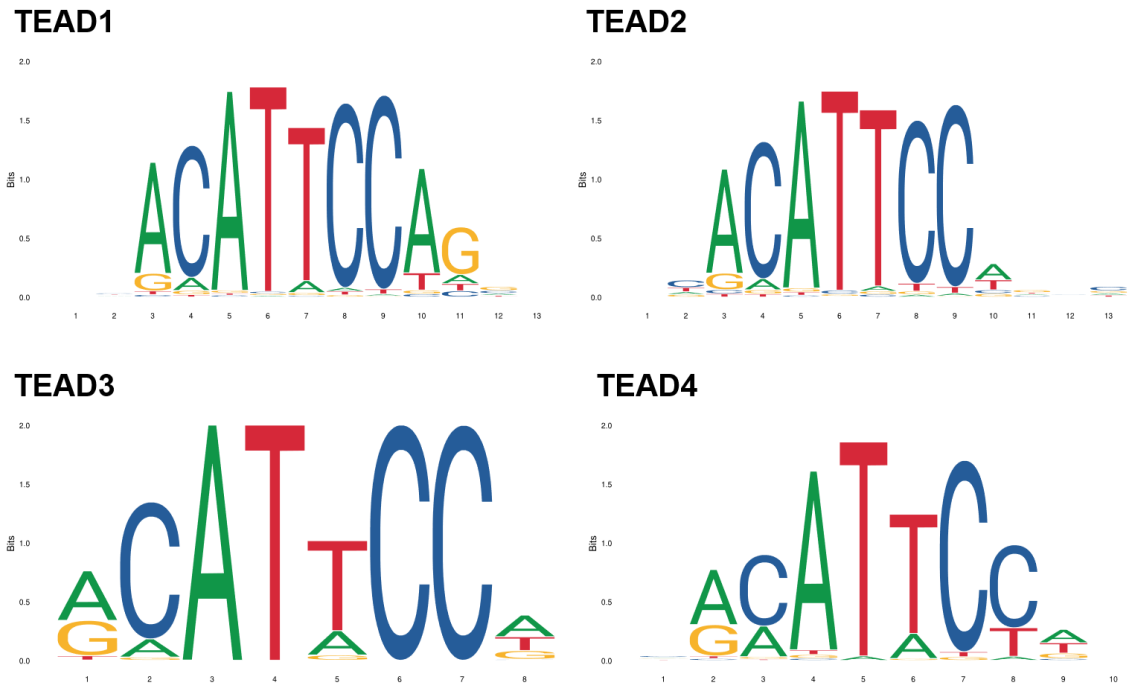


Figure 9: Comparison of binding motifs of the four human TEAD isoforms. *The sequence logos depict the relative frequency of each base on a given position in either HT-SELEX or ChIP-Seq results dataset as are collected in the Jaspar database*[104].

What is also typical for TEAD binding sites is the ability of these transcription factors to bind cooperatively to tandemly repeated M-CAT motifs but non- cooperatively to spaced or inverted repeats[56,106]. Unfortunately, structural basis of this cooperation has not yet been completely explained, although the L1 loop is expected to play a crucial role in it[99].

1.2.2.2. YAP binding domain

As its name suggests, the importance of YAP binding domain (TEAD-YBD) lies in binding of coactivators (out of which is YAP the most well described) and, as was mentioned earlier, TEADs can affect transcription only after forming a complex with these proteins. Therefore, it is also sometimes called "transactivation" domain. The first high resolution structure of this domain was published for residues 217-447 of human TEAD2 and revealed that it adopts an immunoglobulin-like fold with the core of two β-sheets (consisting of five and seven strands) packing against each other to form a β-sandwich plus two helix-turn-helix motifs capping the openings on each end of the β-sandwich (Figure 10)[107].

This study also suggested that its interaction with YAP is formed through a short natively

unfolded segment of YAP which adopts an ordered conformation after binding to TEAD-YBD surface. This finding was subsequently confirmed by two studies that solved the crystallographic structures of human TEAD1-YBD and mouse tead4-YBD in complex with the TEAD-interacting N-terminal domain of YAP[107–109]. It was shown that YAP wraps around the globular structure of TEAD1 and forms extensive interactions via three interfaces, all of which are highly conserved in both YAP and TEAD proteins. Out of them the interface 3, containing the PXXΦP motif (where P is proline, X is any amino acid, and Φ is a hydrophobic residue) was identified as the most critical for complex formation (Figure 10)[108,109].

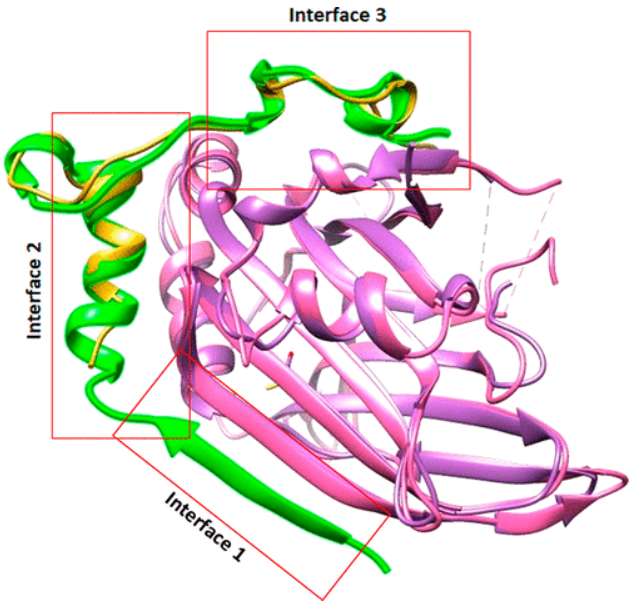


Figure 10: Overlay of the two published TEAD-YBD and YAP complex structures *with highlighted interaction interface. TEAD-YBD is depicted in pink whereas YAP is in green/yellow[98].*

As for other coactivators, the 3D structures of their complex with TEAD-YBD were so far published for mouse tead4-taz, mouse tead4-vgll1 and human TEAD4-VGLL4 which, unlike other VGLL proteins, contains two Vg domains able to bind TEAD[98]. Two different binding modes were found for the Taz coactivator. One of them is very similar to YAP-TEAD complex, while in the other a 2:2 complex is formed by two molecules of Taz jointly forming a bridge to bind two molecules of tead4-YBD. However, in both cases the same binding interfaces as in the TEAD-YAP complex are used[110]. In spite of the significant differences in the primary structures of VGLL and YAP/TAZ, it was shown that Vgll1 adapts a similar fold to YAP/TAZ upon Tead binding with the main difference being the fact, that it interacts with Tead only through interfaces 1 and 2[111]. VGLL4, on the other hand, uses its two Vg domains to form a complex with two TEAD4 molecules at the same time and, more importantly, it was shown to compete with YAP for TEAD binding and therefore to reduce its oncogenic effect[93].

Recently, crystallographic structures of human TEAD2-YBD and TEAD3-YBD were solved and surprisingly, both isoforms were found to be palmitoylated at a conserved

cysteine residue inside a central hydrophobic pocket of the protein. Upon revision of the previously published TEAD-YBD structures, some level of palmitoylation was revealed in all of them and this modification was shown to be responsible for the overall protein stability[96]. The ability of TEAD proteins to autopalmitoylate even at physiological concentrations of palmitoyl-CoA was at the same time discovered by another group which has further shown, that the palmitoylation may, apart from the protein stability, affect YAP/TAZ binding but not VGLL4 binding[112]. Finally, TEAD proteins were shown to be able to incorporate other fatty acids as well, but in this study, the acylation did not have any effect on the coactivator binding. It was, however, confirmed that it highly increases the protein stability[113].

1.3. FOXO transcription factors

The Forkhead box (FOX) family of transcription factors share a conserved DNA binding domain of the same name consisting of a three helix bundle folded into a variant of helix-turn-helix motif and two long β-sheet-bordered loops resembling wings which gave the domain its alternative name "winged helix domain" (Figure 11)[114,115]. So far, 19 subfamilies were identified in organisms ranging from fungi to humans. These subfamilies were classified according to their sequence homology within the Forkhead box and other functional domains (concretely FOXOs do have four of them – the Forkhead DBD, a nuclear localisation sequence, a nuclear export sequence and a transactivation domain[116]) and designated by a letter (A-S)[115,116]. Despite the FOX proteins sharing highly conserved DNA binding domains, they differ in their tissue expression patterns and regulatory mechanisms which allows each of them to have a unique function[117].

First *FOXO* subfamily genes were identified in studies of chromosomal translocations found in human tumors as a part of a fusion gene with *MLL*[118]. Since then, four FOXO members were discovered in humans – namely FOXO1, FOXO3a, FOXO4 and FOXO6. These isoforms are, similarly to TEADs, expressed in almost every tissue and while each of them has its own tissue expression pattern and target gene specificity, their localization and function can sometimes overlap[116,119]. FOXO transcription factors regulate expression of wide range of genes involved in various cellular processes such as cell cycle regulation (*Cyclin D*), apoptosis (*TRAIL*), autophagy (*LC3*), stem cell maintenance and differentiation (MSTN), DNA repair (Gadd45), glucose and lipid metabolism (*G6PC*), stress resistance

(*Catalase*), pluripotency (*OCT4*) or immune response (*IL7R*)[120]. FOXOs themselves are regulated mainly by posttranslational modifications (namely phosphorylation, acetylation and ubiquitination) whose specific combination creates a molecular code to affect FOXO protein's stability, nuclear localization or transcriptional activity in response to external stimuli[119]. Out of these stimuli, the ability to respond to signals transmitted through the insulin or growth factor dependent activation of the PI3K/AKT signalling pathway is a typical property of FOXO family transcription factors and their most thoroughly studied regulatory input[121,122]. If the pathway is active (in presence of insulin or growth factors and importantly in cancer cells), the AKT kinase phosphorylates FOXOs in the nucleus on specific residues resulting in creation of 14-3-3 binding site, subsequent translocation to cytoplasm and thus in inhibition of activity which may result in the suppression of transcriptional programs that control cell proliferation. Therefore, the deregulation of FOXOs is connected with poor prognosis of cancer patients or with insulin resistance[116,119].

1.3.1. FOXO4 and its structure

FOXO4 is one of the first two *FOXO* genes discovered in the studies of acute leukemia cells chromosomal translocations as a part of a fusion gene and was therefore initially called *AFX* (acute leukemia fusion gene from chromosome X)[118]. Mammalian FOXO4 was found to be expressed in almost every tissue, however it is most abundant in the skeletal muscle[123]. Despite the fact, that the initial gene knockout studies of *FoxO4* on mice, in contrast to other *FoxO* isoforms, did not reveal any abnormalities, it was later shown, that FOXO4 plays an important role in several cell processes[124,125]. It is for example involved in a cellular response to oxidative stress, where it can both down- and upregulate the cellular antioxidative defence systems[125]. Furthermore, FOXO4 can induce apoptosis[126], downregulate muscle cell proliferation and differentiation[127] and serve as a tumor suppressor by induction of cell cycle arrest[125,128]. Like other FOXO proteins, FOXO4 is regulated by posttranslational modifications but apart from it, it can interestingly also be regulated on the level of protein synthesis by microRNAs[125]. Thanks to its tumor suppressor activity, FOXO4 was subject to many studies with the ultimate goal to find its clinical application which led, among other discoveries, to structural characterisation of its DNA binding domain. This structure, together with its DNA response element served as a model system in this thesis.

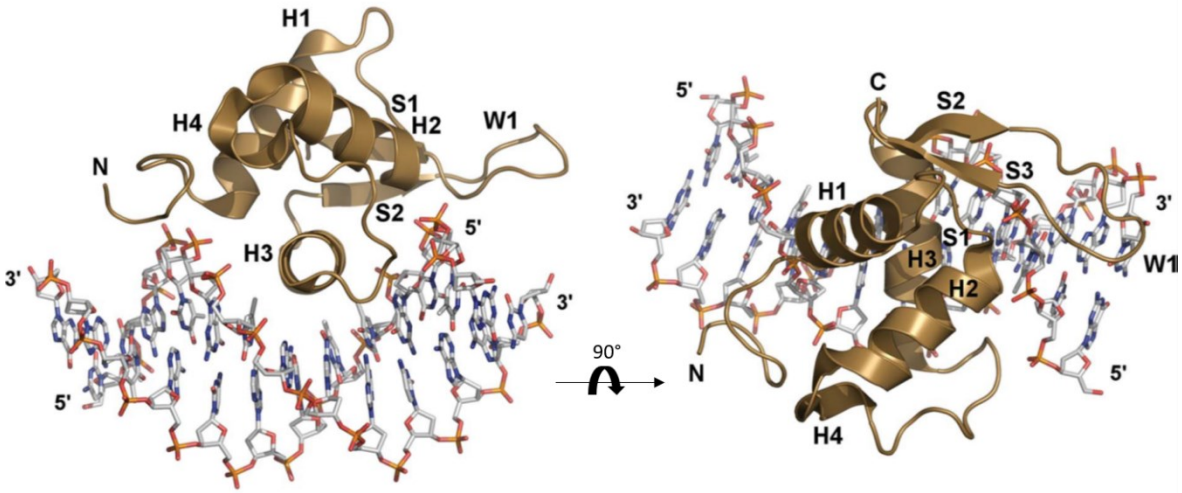


Figure 11: Structure of the FOXO4-DBD·DNA complex. *The complex structure is formed by insertion of the H3 helix to major groove and further stabilized by sequence independent contacts of the N-terminus and wing W1 with the phosphate groups of DNA. Adapted from*[129].

FOXO4 is 505 amino acids long protein consisting of the four main structural domains typical for all FOXOs. The N-terminal Forkhead DNA binding domain is responsible for binding to the consensus sequence 5'-GTAAACAA-3', known as the DAF-16 family member-binding element[130] and it is the only part of the protein whose 3D structure has been solved as the rest of the protein was predicted to be highly disordered[120,131]. The other three regions conserved within the FOXO subfamily (nuclear localisation sequence located on the C-terminal end of the Forkhead domain, nuclear export sequence and C-terminal transactivation domain) all contain numerous posttranslational modification or protein-protein interaction sites and are important for regulation of FOXO4 nuclear localization, stability and transcriptional activity[120]. The first high resolution structure of FOXO4-DBD was obtained by using NMR and confirmed the typical Forkhead domain fold with disordered and highly flexible N- and C-terminal parts[132]. Subsequently, another study has shown, that these two regions (the N-terminal unstructured part preceding the first helix and the C-terminal W2 wing loop) are both involved in stabilization of the protein DNA complex which is, as is typical for all FOX proteins, formed by inserting the third helix into DNA's major grove[133,134]. Finally, a crystallographic structure of the complex of FOXO4-DBD missing the W2 wing with its DAF-16 family member-binding response element was published (Figure 11). It confirmed all the previous discoveries and has shown that FOXO4-DBD uses its helix H3 to dock into the major groove through base-specific contacts, while the N-terminus and wing W1 make additional contacts with the phosphate groups of DNA.

Moreover, in contrast to other FOXO-DBD·DNA structures, the loop between helices H2 and H3 was found to participate in the DNA binding as well[129].

1.4. Methods for Structural Characterization of Transcription Factor-DNA Complexes

To structurally characterize a complex formed between a transcription factor and its DNA response motif a number of biophysical methods can be used. The most commonly utilized are the well-established high-resolution techniques of X-ray crystallography and nuclear magnetic resonance (NMR) with the more recent addition of cryo-electron microscopy (CryoEM). While the obvious advantage of these techniques lies in the ability to solve a protein structure at atomic resolution, they all have their limitations in the experimental conditions that need to be met which will be described further in this chapter. To overcome these limitations, the results received by usage of different high-resolution techniques can be combined and further expanded by lower resolution methods such as small-angle X-ray scattering (SAXS), fluorescence resonance energy transfer (FRET) or structural mass spectrometry (structural MS) which is utilized in this thesis, forming together an approach called "integrative structural biology"[135].

1.4.1. X-ray Crystallography

The X-ray crystallography is the most widely used method for solving 3D structures of proteins or nucleic acids as well as their complexes. The first protein-DNA complex structures were obtained by this methods almost 40 years ago[136,137]. It utilizes the scattering of X-rays of wavelength in the same range as distances between atoms in molecules on electrons in the studied biomolecule. The periodic repeating of the same structural motif in the crystal grid than allows the scattered X-ray radiation to be enhanced or extinguished in specific directions when the waves interfere creating a diffraction pattern characteristic for the given geometry of the crystal unit cell. The effectiveness of this interference then depends on the position of all atoms in the crystal structure. Therefore, after measuring enough reflection intensities while rotating the crystal and after finding phases of individual signals by either experimental or computational means (for both amplitude and phase are needed for deconvolution by Fourier transform but only amplitudes can be obtained from the measured intensities), the measured data can be used to deduce a map of electron density within the crystal

and finally a 3D protein structure in atomic resolution[138].

While the indisputable advantage of this method lies in the high resolution, its use for protein-DNA complexes might be challenging because of the need to crystalize the complex. In case of transcription factors, the main problem lies in the presence of larger unstructured regions that might prevent the crystallization and usually lead to deliberate shortening of the protein and subsequent solving of the structures of only selected domains with stable conformation[139]. In addition to that, the complex needs to be very stable and the conditions used for crystallization need to be carefully selected to prevent creation of artefacts or shielding of recognition sites on the protein by high salt concentrations. Therefore, this technique is not suited for low-affinity complexes[140].

1.4.2. Nuclear magnetic resonance (NMR) spectroscopy

Another frequently utilized method for structural characterization of protein-DNA complexes is NMR spectroscopy. In an NMR experiment, a strong magnetic field is utilized to separate the energy levels associated with nuclei of isotopes such as ^{1}H, ^{13}C, ^{15}N and ^{31}P that carry magnetic dipoles. Upon irradiation of the sample with electromagnetic radiation in the range of radio wave frequency, the nuclei can absorb the radiation and transfer to a higher energy state if the frequency of the radiation precisely correspond to a specific resonant frequency of the nuclei. The nuclei then emit radiation, that can be measured, when returning to the equilibrium state. Most importantly, the specific resonance frequency of each nucleus depends on its chemical environment. Thus, each nucleus in a protein will resonate at different frequency depending on its position in the structure. These frequencies are typically reported as "chemical shifts" in units of ppm to correct for the effect of the size of the used magnet and plotted against intensity of the emitted signal which results in a 1D NMR spectrum. Through a combination of 2D correlation NMR experiments which combines spectra measured for different isotopes it is finally possible to map the distances between pairs of atoms by specifying which pairs are close together in space an therefore to determine a 3D structure of the protein[139,141].

In contrast to X-ray crystallography, the NMR experiments are run in solution and therefore the environment is much closer to native conditions and allows analysis of unstructured parts of the protein and their dynamics. In addition to that, comparison of chemical shifts of the free protein and ligand-bound might be used to identify the regions responsible for binding which is very important for structural characterization of

transcription factor-DNA complexes. However, the main disadvantage of this method has long been the molecular weight upper limit of the analyzed samples of approximately 40 kDa[139]. Most transcription factors, not to mention their complexes with DNA, are bigger than that, so only parts of them could by solved by NMR. Newer techniques, like TROSY, are available to overcome this limitation, however analysis of bigger molecules is still challenging and far from routine[142].

1.4.3. Cryo-electron microscopy

During the last decade, a great improvement in instrumentation finally allowed cryo-electron microscopy (CryoEM) to achieve resolution under 3Å and thus to become a useful tool for determination of 3D structures of proteins. This method is a type of transmission electron microscopy, where electrons emitted by the microscope's source are scattered after passing through the specimen and finally hit a detector to form a 2D projection of the structure in the direction of the electron path[143]. The 3D image is then computationally reconstructed by combining numerous 2D images of the studied object captured from different angles[144]. The cryogenic approach utilizes a process called vitrification, where samples of purified molecules are applied in a very thin layer on a carbon film and subsequently rapidly frozen in liquid ethane cooled by liquid nitrogen. As a result the studied molecules are protected from damage caused by the electron radiation and the same molecule is naturally present in different orientations in the frozen sample to allow the reconstruction of the 3D structure[143,145].

The main advantage of CryoEM lies in the possibility to analyze structures of large protein-DNA complexes such as the nucleosome[146] or the transcription initiation complex[147], that was not accessible by other methods, in decently high resolution. On the other hand, it is still challenging to utilize this technique for analysis of molecules smaller than 100 kDa and particularly of those containing flexible parts or forming more transient interactions which, sadly, includes many transcription factors and their interaction partners[148,149].

1.4.4. Structural Mass Spectrometry

Since transcription factors tend to contain intrinsically disordered regions and their function is often allowed thanks to (or regulated by) interactions with other proteins and, of course, with DNA, they pose a rather difficult target to be studied by the high-resolution methods

described above. The structural mass spectrometry (MS) refers to a set of techniques, that utilize the mass spectrometry analytical method able to measure the exact mass of ionized analyte transferred into the gas phase in the form of mass to charge (m/z) ratio for characterization of structure of macromolecules[150]. These techniques can provide information about the nature and exact position of posttranslational modifications (MS^n), quaternary structure and interactions stoichiometry (native MS), analyte shapes via calculation of collision cross sections (ion mobility MS), both inter- and intramolecular distance restrains (Chemical cross-linking MS) or about solvent accessibility (hydrogen/deuterium exchange MS and stable covalent labeling approaches).

Together, the structural mass spectrometry makes a useful tool to study structures of proteins and their complexes, particularly those whose analysis by other methods might be problematic, as well as their dynamics and interaction interfaces with various ligands[151]. Apart from that, its other advantages are low sample consumption and short analysis times which makes structural MS a high-throughput method. On the other hand, structural MS is not able to solve a full 3D protein structure from scratch since it does not provide atomic coordinates and relies on homology modelling utilizing distance restrains and structures obtained by other methods[152].

In this thesis a combination of hydrogen/deuterium exchange and chemical cross-linking MS[153] was used to structurally characterize transcription factor-DNA complexes and therefore these two approaches will be briefly described in the rest of this chapter.

1.4.4.1.1. Chemical Cross-linking

In a chemical cross-linking experiment a reagent with two reactive groups separated by a spacer of defined length (the cross-linker) is added to a protein or a protein complex solution. The cross-linker then connects two functional groups of amino acid side chains whose maximum distance from each other is defined by the spacer length. In a typical setup, the cross-linked protein is subsequently enzymatically cleaved to peptides which are separated by reverse phase liquid chromatography and analyzed by tandem mass spectrometry (LC-MS/MS) to identify cross-linked residues. Obtained distance restrains are finally used in a computational modelling study to construct a 3D structural model (Figure 12)[154].

Over the years, cross-linkers of various spacer lengths (ranging from zero length to several tens of Ångströms) and reactivity towards different amino acids were developed[151,155].

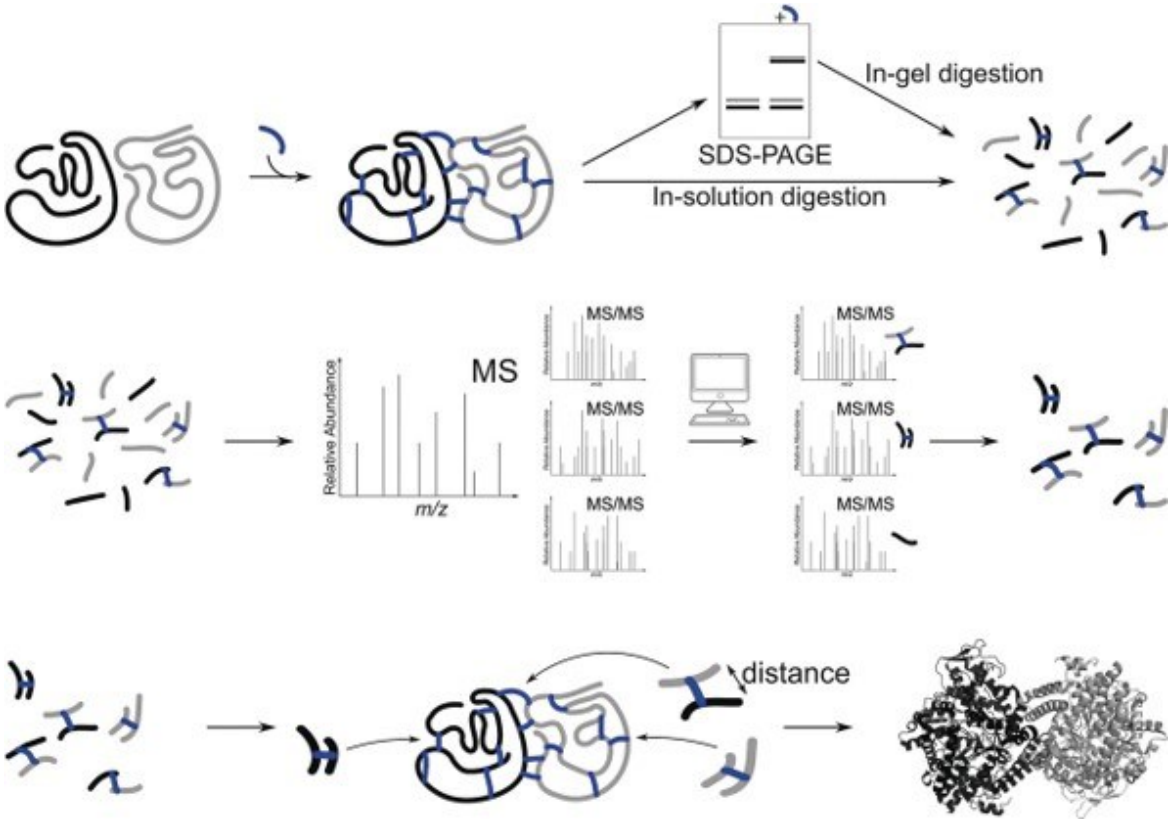


Figure 12: Workflow of a typical chemical cross-linking/MS experiment[154]

However, the most widely used cross-linkers are homobifunctional N-hydroxysuccinimide esters such as is disuccinimidylsuberate which react mainly with primary amine groups (lysine side chains or N-termini) and with limited reactivity also with hydroxy groups (serines, threonines and tyrosines)[154,156]. To facilitate the identification of a cross-link out of many unmodified peptides several modifications of the cross-linking reagent might be used. The cross-linker might be isotopically labeled and using a mixture of labeled and unlabeled cross-linker therefore results in an easily identifiable "double peak" pattern and the labeled reagents can even be used for quantification when utilized on proteins in different conformational states under different conditions[157,158]. Some cross-linkers might also be cleavable inside the mass spectrometer while producing a specific peak pattern as well[159] and an affinity tag can be introduced to enrich the cross-linking products[160].

The cross-linking approach can also be utilized to cross-link proteins with nucleic acids to reveal their mutual placement in the targeted assembly. The UV-irradiation of nucleic

acids can be used to produce highly reactive intermediates that form zero-length covalent cross-links to protein molecules in the vicinity[161,162] and apart from that, several bifunctional reagents able to connect proteins to nucleic acids such as platinum complexes (cis- and transplatin) or nitrogen mustards are also available[163,164]. Nevertheless, none of the currently known protein-nucleic acid cross-linkers can be used universally to any system and conditions and therefore careful selection and optimalization is needed[164].

1.4.4.1.2. Hydrogen/Deuterium Exchange

The hydrogen/deuterium exchange mass spectrometry (HDX-MS) method is based on the effect of hydrogen exchange that occurs between labile hydrogens in -OH, -NH or -SH groups in proteins and surrounding water. When the protein solution is diluted to D_2O based buffer instead of water, the deuterium atoms are over time incorporated into the protein while increasing its mass. The increased mass is then detectable by mass spectrometry[165,166]. Nevertheless, only the amide hydrogens of the protein backbone are typically detected in HDX experiments, because unlike the side chain groups, where the exchange is very fast, they exchange hydrogens in rates ranging from milliseconds to months which is convenient for the measurement[151,167]. The rate of HDX of a given amide hydrogen is dependent, apart from pH and temperature, also on the conformational properties of the protein. Hydrogens participating in the hydrogen bonding network as well as those that are buried inside the protein structure exchange far more slowly than those located in an unstructured region or on the protein solvent accessible surface. Thus, information about protein structure as well as about interaction interfaces of protein ligand complexes can be derived from HDX rates of amide hydrogens that are conveniently uniformly distributed along the peptide chain (with the exception of proline that does not contain an amide hydrogen)[165,167].

In a common workflow (Figure 13) the proteins (or protein complexes) are incubated in deuterated buffer and samples are taken at specific time points. The deuteration reaction is immediately quenched by lowering pH to 2.5 where the exchange rate is minimal and rapid freezing in liquid nitrogen. Denaturants and reducing agents can be added to the quench solution to enhance protein unfolding and facilitate the subsequent protease digestion. Proteases, that are active under the acidic conditions, are then utilized to digest the protein to peptides[168]. While a standard protease for this purpose is pepsin, it may not always produce enough short overlapping peptides to achieve a reasonable spatial resolution and

therefore several other alternative proteases have been developed that might be used instead or in combination with pepsin to increase the sequence coverage and thus to more precisely localize the sites of decreased/increased deuteration[169–171]. The peptides are then desalted, separated by a chilled reversed-phase HPLC system, ionized by electrospray and finally analyzed in a mass spectrometer to determine the increase in mass resulting from deuterium uptake for each peptide. Because the samples were taken in several time points during the deuteration reaction, deuterium uptake plots of individual peptides can be constructed and compared between different conditions of the protein (e.g., bound and unbound to a ligand, phosphorylated or dephosphorylated, etc.) to quickly localize differences in HDX rates[168].

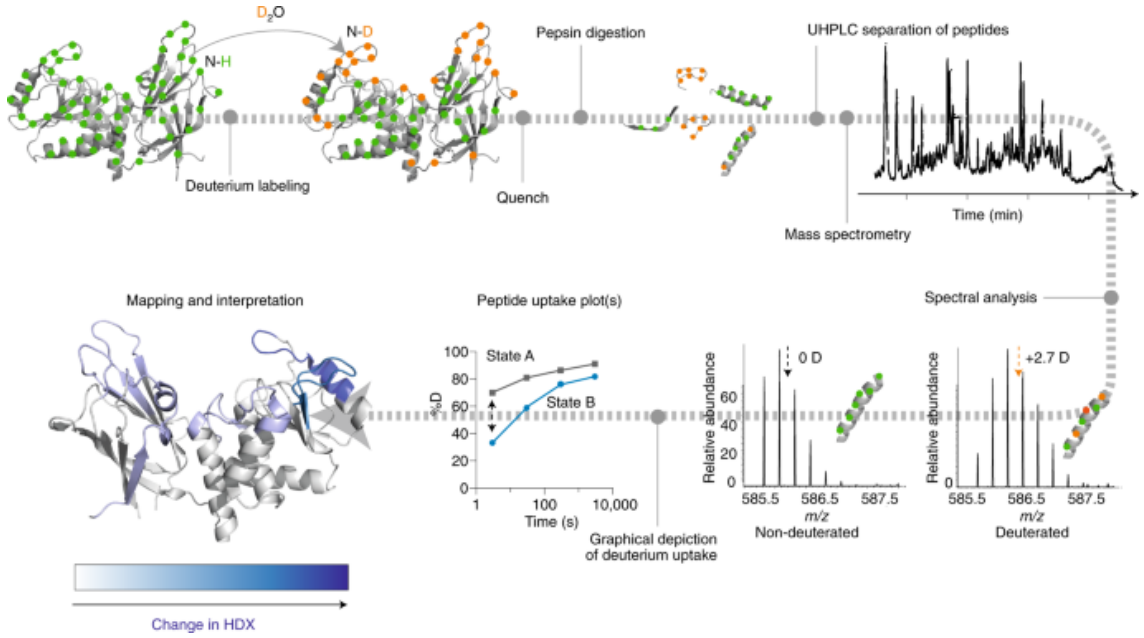


Figure 13: The common workflow of a hydrogen/deuterium exchange MS experiment[168]

In case of transcription factor complexes with DNA, the advantage of HDX-MS is theoretically unlimited protein size and the ability to study proteins in their truly native-like environment together with the possibility to analyze the unstructured regions of the protein and therefore to find, whether they are involved in the complex formation. However, the presence of DNA in the samples might complicate the analysis mainly due to its poor solubility and tendency to precipitate in low pH condition such as those used in the quench solution[172]. Apart from that, the presence of DNA can also have adverse effects on chromatographic separation. To prevent these problems, several approaches were developed and successfully used[173]. It was found, that adding small basic molecules or denaturing agents such as protamine sulfate, guanidine hydrochloride or urea to the quench buffer all

helps prevent the precipitation in low pH[174–176]. Furthermore, addition of strong anion exchange column prior to the reverse-phase chromatographic column might help solving the latter problem[177].

2. Aims of the Thesis

The aims of this thesis were to contribute to the development of a set of structural mass spectrometry methods for characterization of transcription factor complexes with their cognate response motifs and to apply these methods to structurally characterize the interaction between the DNA binding domains of FOXO4 and TEAD1 proteins and their DNA response motifs.

The specific goals were:

- To prepare and characterize a panel of HDX-MS compatible proteases to be used to improve sequence coverage of complex protein systems – namely integral membrane proteins and protein-DNA complexes
- To explore the potential of native MS for quick evaluation of transcription factor-DNA complex formation and their sorting according to binding affinity
- To find out whether and how the sequential context of the M-CAT motif affects its interaction with TEAD1-DBD
- To explain the different affinities of M-CAT motif and its inverted version to TEAD1-DBD

3. Methods

The publications included in this thesis provide a full description of all methods and experimental procedures used which include all details necessary for their reproduction. Therefore, this chapter serves only to list experimental techniques used in the thesis.

List of used research methods:

- Recombinant expression and purification of proteins
- Immobilization of proteolytic enzymes
- Protein and protein-DNA complex sample preparation for MS analysis
- Gel shift analysis of protein-DNA complexes
- Hydrogen/deuterium exchange mass spectrometry
- Fluorescence anisotropy based K_D assay
- Chemical cross-linking in combination with mass spectrometry
- Native mass spectrometry (nESI-FTICR)
- Single-molecule Förster resonance energy transfer

4. Results and Discussion

The aim of the first part of this thesis was to optimize a set of structural mass spectrometry methods to be used for characterization of transcription factor-DNA complexes and subsequently to test them on a model system consisting of FOXO4 transcription factor and its DNA response motif. The second, and main, part was then focused on structural characterization of TEAD1 interaction with its M-CAT response motif in different sequential contexts using not only the previously optimized methods, but also other techniques such as smFRET or native MS.

First, a set of HDX-MS compatible proteases needed to improve sequence coverage and resolution was prepared, characterized, and used to digest two complex protein systems (in addition to transcription factors also integral membrane proteins). The results of this part were included in **Publications I, II and III**. In addition to HDX-MS, protein-protein and protein-DNA chemical cross-linking in combination with mass spectrometry and homology modeling were utilized together to structurally characterize a model system of FOXO4-DBD-DAF16 transcription factor-DNA complex and the results were included in **Publication III**.

In the second part of the thesis, the previously tested MS methods were, together with smFRET, utilized to explain the different affinities of M-CAT motif and its inverted version to TEAD1-DBD with the results being summarized in **Publication IV.** Subsequently, a native MS based method for quick evaluation of protein-DNA complex formation was optimized to be used in an effort to find out how the sequential context of the M-CAT motif affects its interaction with TEAD1-DBD. The results of the last part were partially included in **Publication IV** while the publication including the rest of them is still in preparation.

4.1. Evaluation of MS-based Approaches for Structural Characterization of Transcription Factor-DNA complexes

The information about structure and dynamics of transcription factors and their interacting partners is crucial for understanding their specific mechanisms of DNA recognition and action. However, the size and complexity of these systems, where only the DNA binding domain is typically highly structured, while other regions responsible for either modulating transcription activity, or mediating ligand interactions remains rather flexible, complicates their structural characterization by the commonly used high resolution techniques[41]. Due to

these reasons, most solved transcription factor structures in the Protein Data Bank do not include the entire protein sequence, but mostly consist only of the DNA binding domain[178].

A combination of mass spectrometry-based techniques including hydrogen-deuterium exchange and chemical cross-linking (which were used in this theses) have proved to be able to identify the interaction interfaces between bound biomolecules (including proteins and nucleic acids) and reveal their mutual spatial position as well as to refine the previously known crystallographic structures[153,163,179]. And more importantly, these techniques are not limited by protein size or presence of unstructured regions[151]. Therefore, the application of complementary structural MS techniques has been evaluated to overcome the challenges posed by complexes between transcription factors and their DNA response elements. To do that, the utilized techniques first needed to be adjusted for the protein-DNA complexes. In this thesis, I have focused mostly on optimization of hydrogen/deuterium exchange conditions with the main emphasis on improving the spatial resolution by using alternative proteases for protein digestion.

4.1.1. Improving Sequence Coverage and Resolution in HDX-MS Experiments by Using Alternative Proteases

To achieve a good spatial resolution in HDX-MS, the proteolytic digestion is the key factor, that needs to be optimized. An ideal result would be a full sequence coverage with large numbers of overlapping peptides that are neither too long, nor too short (8-12 amino acids was suggested to be the best option). In that case, the overlaps might be used to localize the differences in deuteration rates nearly to individual amino acids[180–182]. As was mentioned in chapter concerning HDX-MS (1.4.4.1.2.), the standardly used protease, porcine pepsin, does not always provide optimal digestion profile needed to achieve the best possible resolution and therefore other proteases were developed to be used instead or in combination with it[183].

Regarding the proteolytic digestion, both integral membrane proteins and transcription factor-DNA complexes might prove to be particularly troublesome. In case of integral membrane proteins, the main difficulty lies in providing access for the protease to transmembrane regions since the detergents, which are usually used to stabilize their fold for in vitro studies, can disrupt the protease access. Moreover, the peptides arising from the transmembrane parts often tend to precipitate in water and on top of all that are rich in hydrophobic residues that pepsin favors as a cleavage sites. As a result, too short peptides

unable to be retained by the desalting reversed phase precolumn are produced. As for the transcription factors, the problem is quite the opposite. Digestion by pepsin often results in creation of too long peptides that do not provide sufficient spatial resolution due to the fact, that they contain a lot of basic residues after which pepsin does not preferentially cleave. Therefore, alternative proteases with different cleavage specificities might help increase spatial resolution in HDX-MS in both systems.

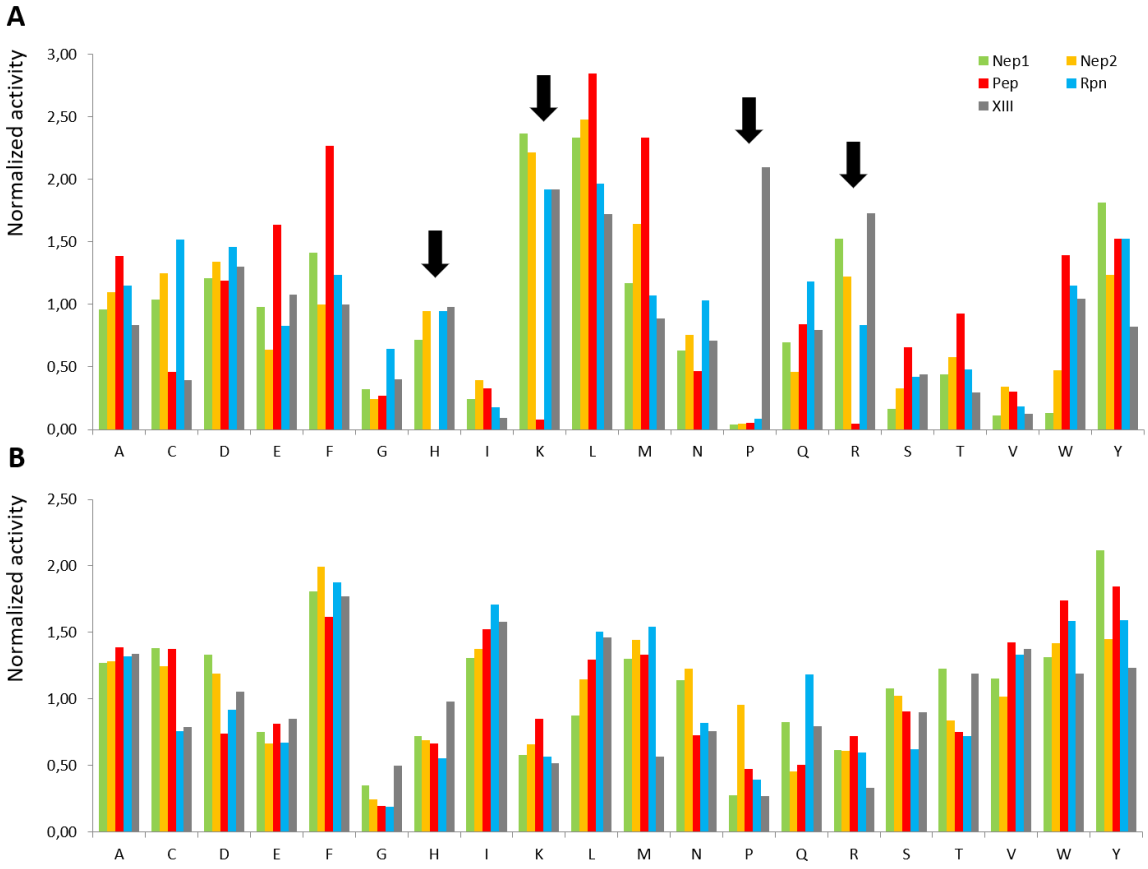


Figure 14: Cleavage preferences of individual immobilized proteases. *Plot (A) shows cleavage at the P1 amino acid position (C-terminal of the residue) and (B) shows cleavage at the P1' amino acid position (N-terminal). Digestion was done using Nepenthesin I (Nep1 - green), Nepenthesin II (Nep2- yellow), Pepsin (Pep – red), Rhizopuspepsin (Rpn – blue) and Aspergillopepsin protease type XIII (XIII – gray). Black arrows indicate amino acids after which pepsin does not cleave but some of the alternative proteases do. Data obtained on several model proteins were normalized according to reference*[169].

For these reasons, a set of alternative proteases compatible with the HDX-MS conditions was either recombinantly produced (Nepenthesins I and II and Rhizopuspepsin) or bought (Aspergillopepsin – protease type XIII) and immobilized on POROS-20 AL resin. To test the columns performance as well as to extract cleavage specificities under HDX-MS conditions for all the immobilized proteases and to finally compare it with pepsin, four

proteins were digested on each protease column and analyzed by LC-MS/MS according to previously published protocol[184]. As is shown in figure 14, in agreement with previously published research[169,171,184], all proteases except pepsin were able to cleave after basic residues (lysine, arginine and histidine). In addition to that, Aspergillopepsin was, in contrast to the other proteases, also able to cleave after proline. The tested protease columns were then used to improve sequence coverage and spatial resolution in HDX-MS analysis of integral membrane proteins and transcription factors and the results were included in **Publication I** and **Publication II and III**, respectively.

4.1.1.1. Publication I

In this study, the digestion of four integral membrane proteins (Cl−/H+ exchange transporter, leucine transporter, dopamine transporter, and serotonin transporter) by porcine pepsin and three alternative aspartic proteases either in-solution or immobilized on-column in HDX-MS compatible conditions was compared. In addition to that, the possibilities of optimizing the quench and chromatography conditions to improve sequence coverage were also screened.

First, the quench conditions were optimized for each membrane protein. The results have shown that addition of the denaturant urea to the quench buffer resulted in improved sequence coverage of all the tested membrane proteins. On the other hand, using guanidine-HCl resulted in decrease of sequence coverage in three out of four tested membrane proteins. Nevertheless, it was previously shown that the optimal choice of additives is highly dependent on the target protein and therefore the charged nature of guanidine, which is probably the cause of decreased deuteration efficiency here, might prove to be beneficial elsewhere[185,186]. Apart from the addition of urea, switching from using C18 analytical column to C8 allowed for a slower increase in acetonitrile concentration in the mobile phase while keeping the overall length of the chromatographic separation constant. This resulted in improved chromatographic separation and ultimately in increased sequence coverage for all four membrane proteins.

The performance of each protease, be it in solution or immobilized on column, was tested by digesting Phosphorylase B under the optimized quench conditions for each membrane protein (thus in presence of the same additives). The results have shown that neither the presence of lipids nor detergents used in the study did impact the performance of any of the

proteases. However, concentrations of urea above 3M in presence of 0.3M TCEP slightly decreased performance of Nepenthesin I and even more so for Nepenthesin II. Luckily, this effect was largely diminished by immobilization of these proteases on POROS beads, which is all in agreement with previously published studies of the Nepenthesins[171,184].

As was already mentioned, all the proteases were studied in both free and immobilized forms. Digestion with immobilized proteases resulted in higher sequence coverage in almost all protease-membrane protein combinations. This is probably due to elevated pressure (~2000 psi) and temperature (20 °C), which can destabilize the detergent micelles and make the transmembrane parts of the proteins accessible while the higher local protease:protein ratio that can be achieved on column can also play its role[187]. Out of the proteases in solution, only Nepenthesin II provided sequence coverages at least close to those of immobilized proteases. Therefore, Nepenthesin II might be an interesting option in cases where the micelle is obstructing digestion by an immobilized protease, as it provided the highest number of identified peptides for in-solution digestions of three out of the four tested membrane proteins.

Interestingly, the choice of the best protease for digestion of membrane proteins was found to be highly protein specific (Figure 15 shows an example comparison of sequence coverage maps of serotonin transporter after digestion with different proteases). The highest sequence coverages were achieved by either immobilized pepsin for the Cl−/H+ exchange transporter and leucine transporter (82.2 % and 33.2 %, respectively) or by immobilized Rhizopuspepsin for dopamine transporter, and serotonin transporter (38.9 % and 58.3 %, respectively). The number of identified peptides also differed considerably. Twice as many peptides were identified after the digestion of leucine transporter with immobilized pepsin compared with its digestion by immobilized Rhizopuspepsin. In contrast to that, compared with pepsin, the digestion of dopamine and serotonin transporters with immobilized Rhizopuspepsin resulted in a 1.6- and 2.8-fold increase in peptide identifications. This is particularly interesting considering the fact, that the leucine, dopamine and serotonin transporters are structurally related and share a similar fold and hydrophobic properties of the transmembrane helices. Thus, the results have shown, that screening of a set of available proteases, immobilized on column if possible, is always beneficial even for closely

structurally related integral membrane proteins.

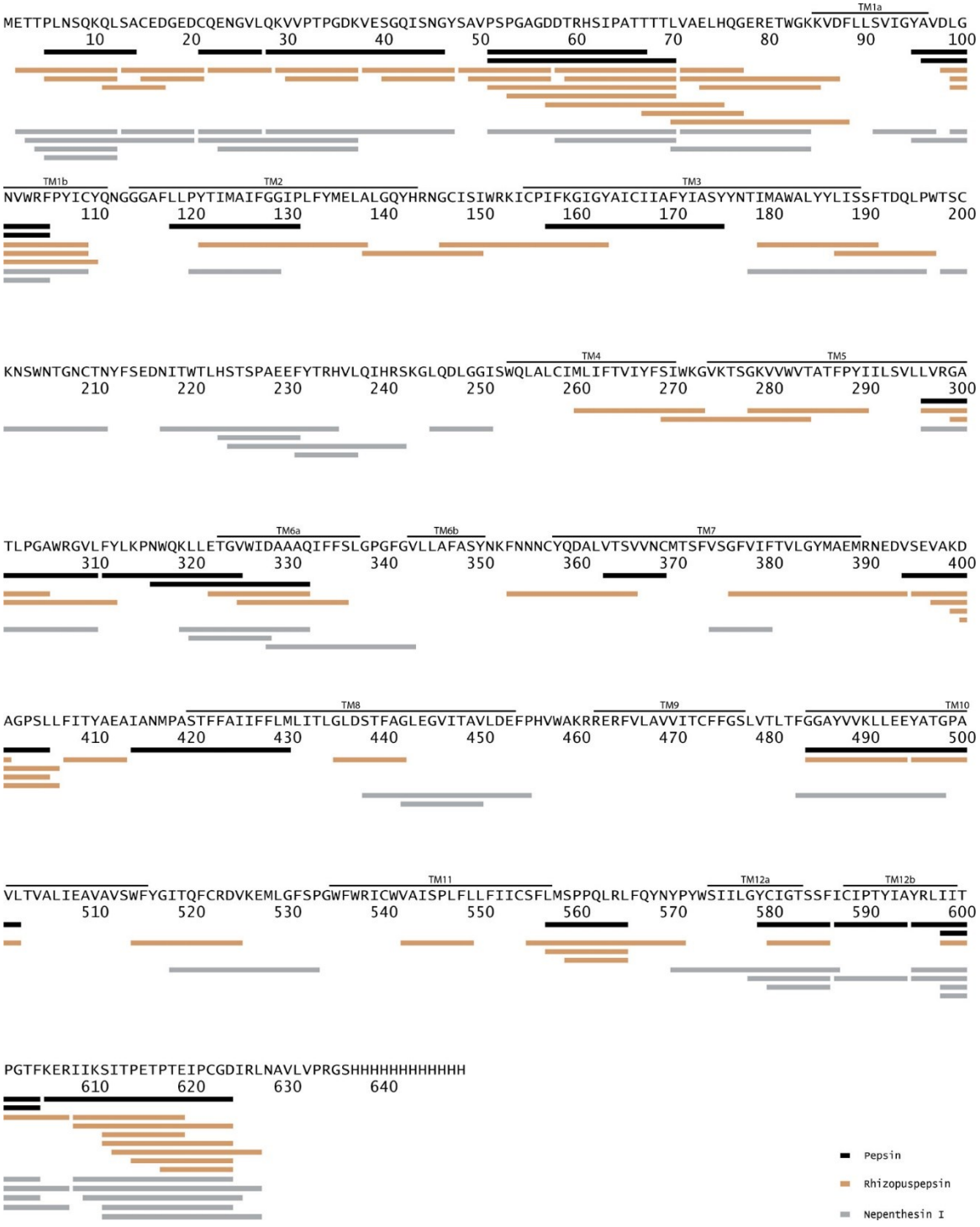


Figure 15: Sequence coverage of serotonin transporter after digestion by different immobilized proteases. *The digestion by immobilized pepsin (black) resulted in 22 identified peptides covering 38.1% of the sequence, 61 identified peptides covering 58.3% of the sequence were obtained through digestion by immobilized rhizopuspepsin (light brown), and immobilized nepenthesin I (gray) provided 46 identified peptides covering 50.9% of the sequence.*

4.1.1.2. Publication II

In this publication, an optimized protocol for performing HDX-MS experiments on transcription factor-DNA complexes was presented. Similar to the integral membrane proteins, it was observed, that protein:DNA systems can behave surprisingly differently even when the proteins as well as DNAs used are similar in size. Out of the two system for which the protocol was optimized, the FOXO4-DBD/DAF16 serves as an example where short dsDNA (13 bp) did not interfere significantly with peptide recovery, and therefore only digestion conditions were optimized to reach good HDX spatial resolution. On the other hand, the 15 bp long dsDNA used in the TEAD1-DBD/M-CAT system had a strong impact on peptide recovery. Therefore, apart from the digestion conditions, the quench and chromatography conditions needed to be adjusted as well.

Before initializing the HDX-MS experiment, the formation of duplex and ability to bind to its cognate transcription factor needed to be verified for all of the used oligonucleotides. For this purpose, a gel shift assay was performed. The results confirmed the formation of all dsDNA duplexes as well as their complexes with transcription factors. Additionally, the bound fraction of the protein could be estimated from the thickness of the band corresponding to dsDNA (Figure 16). This is most beneficial for the subsequent labeling step, where the saturation of the transcription factor by DNA should be as high as possible.

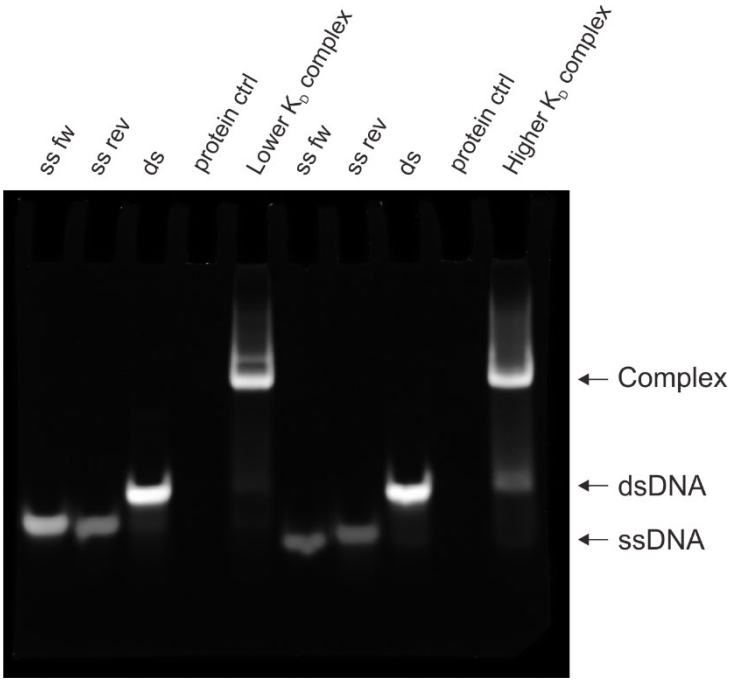


Figure 16: An example of gel shift assay. *Two dsDNAs that form complex with TEAD1-DBD with different affinities are shown. Shifts in electrophoretic mobility in the double strand DNA (ds) and complex lanes indicate duplex DNA and protein–DNA complex formation, respectively. Protein control lane stayed empty since no DNA was present. In both the complex lanes, a band of the same mobility as dsDNA can be observed, though in the "higher K_D complex" lane, this band is much thicker, indicating lower bound fraction of the protein.*

As was mentioned at the beginning of this chapter, several obstacles had to be overcome in case of the TEAD1-DBD/M-CAT system.

First, the 15bp long dsDNA precipitated under the quench conditions while taking the protein with it. According to the literature[175,176], addition of denaturants to the quench buffer was tested to solve this problem. While guanidine hydrochloride did not improve the peptide recovery at all, after addition of urea (finally optimized to 2M in the resulting sample) a complete sequence coverage was achieved. The second issue manifested in widening of the chromatographic peaks after several subsequent injections of samples containing DNA. It was probably caused by accumulation of the DNA on some of the used columns and since washing the desalting trap column or analytical column did not help, the problem was traced to the column with immobilized protease. To get rid of the DNA, incorporation of strong anion exchange trap column prior to the chromatography system[173,177] was tested but resulted in huge decrease in sequence coverage. On the other hand, washing the system after each sample with slightly basic ammonium formate (pH 7.5) slightly improved the chromatography performance, however frequent running of blank samples was still needed even after that.

As for the digestion optimalization, all the protease columns mentioned in chapter 4.1. were tested both alone and in pairs in serial configuration. The sequence coverage maps of the best conditions compared with the standardly used pepsin are shown in figure 17. While pepsin provided a full sequence coverage in both systems, the number of identified peptides and their redundancy significantly improved upon the use of alternative proteases. In case of DNA binding domain of FOXO4 the best results (good spatial resolution and highest redundancy) were obtained by combined digestion with pepsin followed by nepenthesin-1. Aspergillopepsin (protease type XIII), on the other hand, provided the best results for the TEAD1-DBD/M-CAT system. In both examples shown here, the advantage of alternative proteases for the digestion of DNA binding proteins (which are rich in basic residues) lied in their ability to cleave after the Lys and Arg residues giving rise to additional peptides and finally resulting in better spatial resolution. Nevertheless, the choice of the optimal protease for the digestion step out of those able to cleave after basic residues, again, proved to be highly protein dependent.

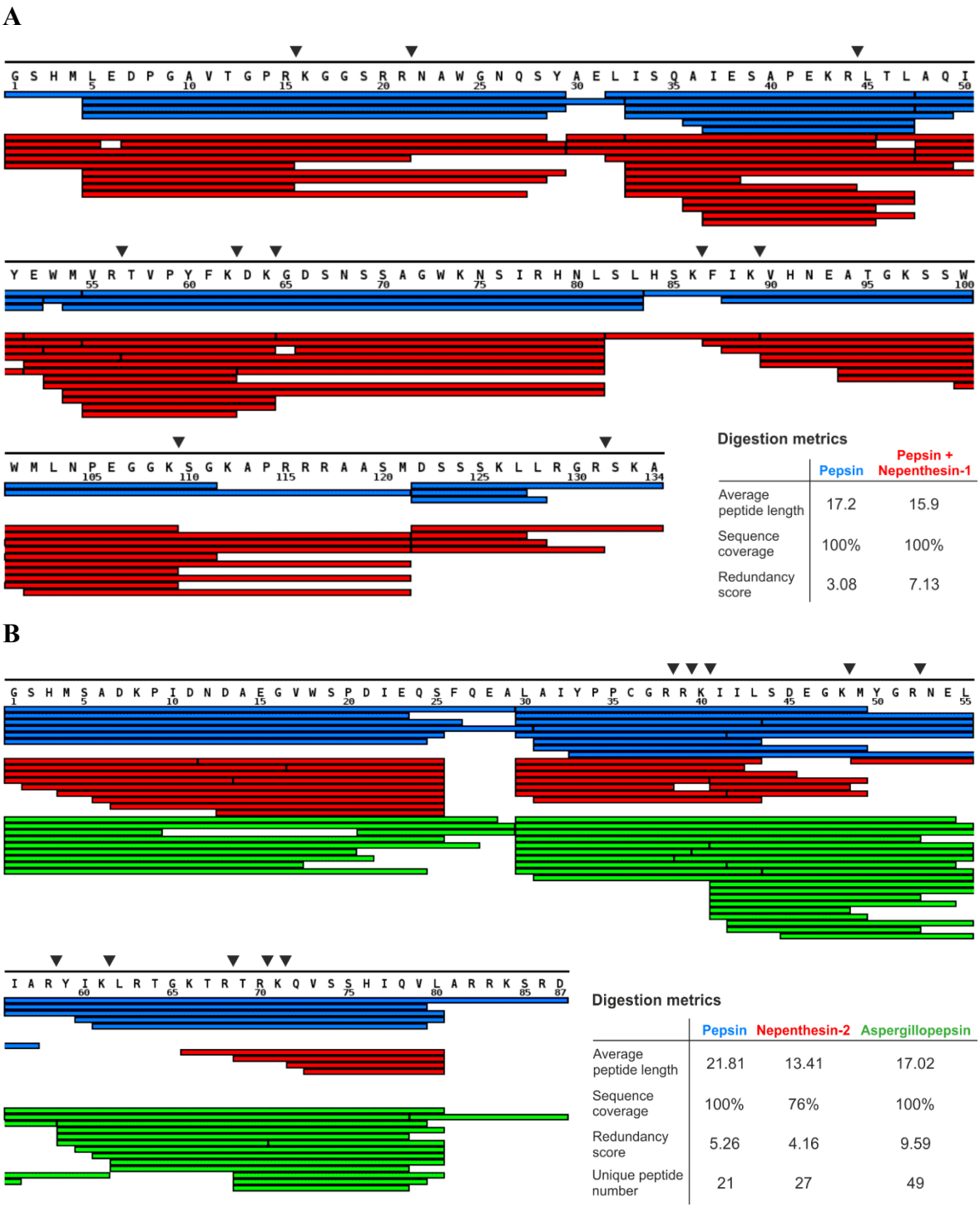


Figure 17: Sequence coverage maps of the two studied transcription factors. *(A) FOXO4-DBD where initial proteolysis by pepsin (blue) was replaced by combined digestion with pepsin followed by nepenthesin-1 (red) after optimization (b) TEAD1-DBD. Pepsin (blue) provided full sequence coverage but lower spatial resolution (longer average peptide length). Nepenthesin- 2 (red) led to over-digestion, which is indicated by gaps in the sequence and the short length of detected peptides. The green bars show the final conditions where aspergillopepsin (protease type XIII) was utilized. Arrowheads above the sequence indicate the cleavage sites after basic residues that were introduced by the alternative proteases.*

4.1.2. Structural Characterization of the FOXO4-DBD/DAF16 Model System

The optimized protocol for performing hydrogen/deuterium exchange MS experiments on transcription factor-DNA complexes was, together with chemical cross-linking and homology modelling, utilized for structural characterization of a model complex consisting of the DNA binding domain of human transcription factor FOXO4 (FOXO4-DBD) and its cognate Daf-16 family member-binding element (DAF16). This model system was selected for two main reasons. The first one was the availability of a crystallographic structure[129] needed for validation of the new experimental constraints and evaluation of the whole method set. The second one was the fact, that the said crystallographic structure revealed some discrepancies when compared to the binding modes of other members of the FOXO family (which could possibly be attributed to crystal-packing issues) and it was also missing the C-terminal region to facilitate crystal formation. Since this region was previously identified to be involved in DNA binding, it presented an interesting opportunity to test the capability of structural MS to study the transcription factor-DNA interactions[129,133].

4.1.2.1. Publication III

To structurally characterize the FOXO4-DBD/DAF16 complex, results obtained by using three structural MS methods were combined to guide model-building operations and create a complete 3D structure of the complex which was finally compared to the available high-resolution structures. Namely, HDX-MS was used to identify the regions of the protein affected by DNA binding, quantitative chemical cross-linking was applied to identify possible variations between free and DNA-bound protein structures and transplatin (trans-dichlorodiamineplatinum(II)) was used to generate protein-DNA cross-links that would help locate the mutual positions of the transcription factor and DNA.

The protocol optimized in **Publication II** was employed to perform the HDX-MS analysis of both DNA-free and bound forms of FOXO4-DBD. As was already mentioned, in case of FOXO4/DAF16 no protease on its own provided a sufficient number of overlapping peptides of optimal length and therefore a combination of two proteases (namely pepsin and nepenthesin1) in serial setting was used. The results, which are summarized in figure 18, revealed three regions to be affected by DNA binding. First region (highlighted in orange in the figure 18) where the deuteration rate significantly differed between the bound

and free forms included helix H1, part of helix H2, strand S2 and the intervening loops. Interestingly, no part of this region was supposed to make direct contact with the duplex DNA according to the crystallographic structure[129]. Therefore, the decrease in deuteration could not be attributed simply to a protection effect but more likely to indirect conformational effects induced by binding.

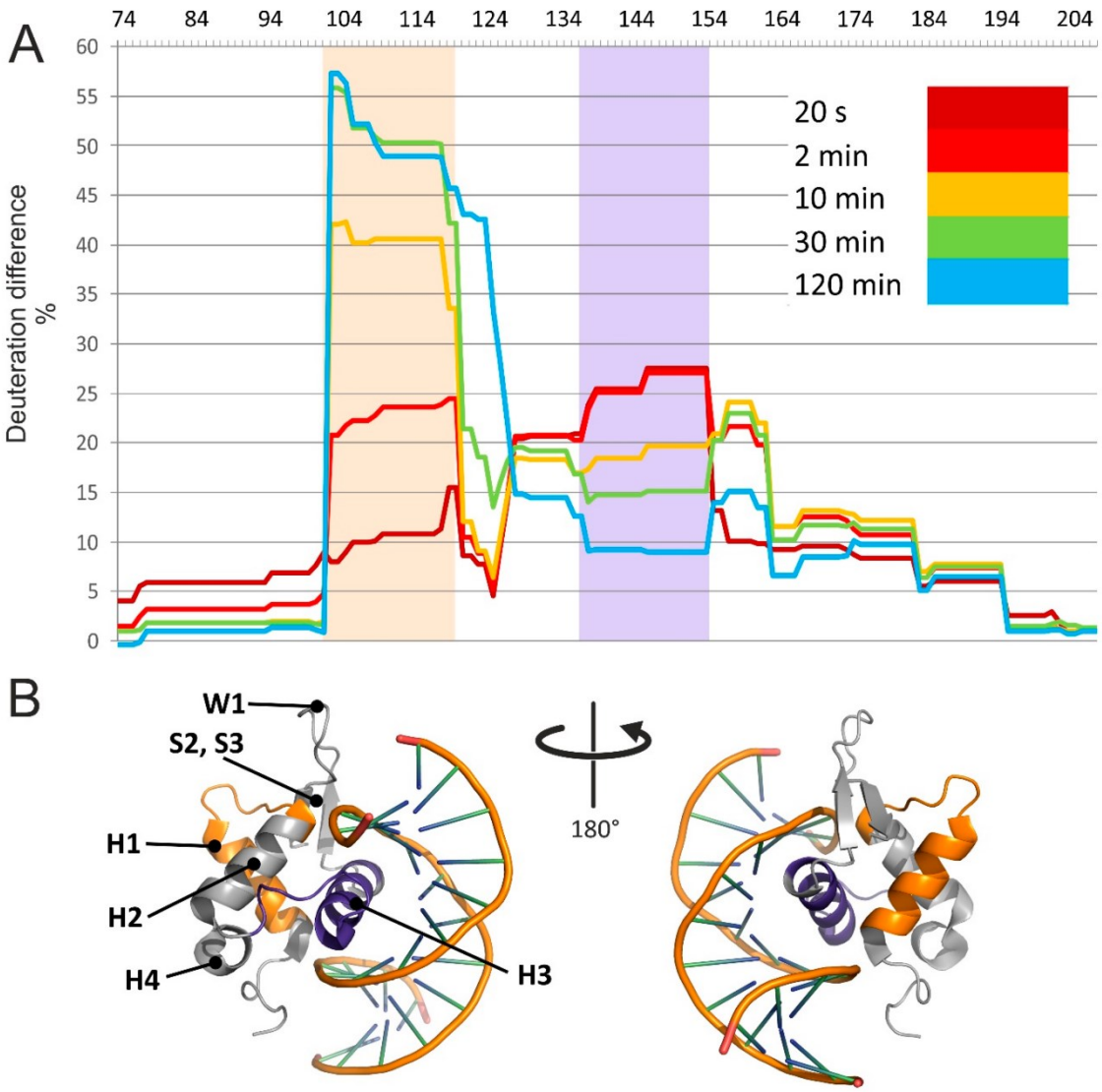


Figure 18: Regions of FOXO4-DBD affected by DNA binding revealed by HDX-MS (A) Relative deuteration differences *between free and bound forms plotted along the FOXO4-DBD sequence and their evolution in time. Highlighted areas show two regions with large differences in deuteration levels caused either by long-distance stabilization of the protein structure (orange) or direct interaction with DNA (blue).* ***(B) FOXO4-DBD•DAF16 structure*** *with the same regions showing significant differences in deuteration highlighted by the appropriate colours.*

The most plausible explanation thus would be, that in the bound form, the dynamics of the local structure may be diminished by interactions with neighbouring structures that, in turn,

make direct contact with the DNA ligand. On the other hand, the second region with significant decrease in deuteration in the bound form (highlighted in blue in figure 18) was helix H3, which the crystal structure places directly in the major groove of the DNA duplex[129] making direct steric protection induced by bound DNA the best explanation for the observed exchange reduction. The last section that manifested smaller, but still significant, differences in deuteration rates was located on the C-terminal side of helix H3 and included strands S2 and S3, as well as the W1 and W2 wings. This region did not make a direct contact with DNA in the crystal structure, so an overall decrease of structural flexibility upon binding could explain the reduced deuterium uptake in this case as well.

To describe the relative position of FOXO4-DBD and its response element, transplatin was utilized to cross-link the protein with its DNA ligand. The treatment of the FOXO4-DBD with transplatin resulted in identification of several peptide-oligonucleotide conjugates. Consistently with the HDX-MS results, all the cross-linked peptides originated either from regions identified by HDX-MS to be affected by DNA binding (helix H3 and the S2-W1-W2 region) or from unstructured regions whose flexibility allowed the susceptible groups to get close enough to each other to allow the cross-link formation (N-terminal loop).

To provide yet another point of view on the conformation changes happening when FOXO4-DBD binds to DNA, quantitative chemical cross-linking was utilized as well[158]. A total of 39 conjugates were identified providing distance restrains needed for subsequent homology modelling. The quantification results have shown that helix H3 located residues are indeed highly affected by DNA binding. Furthermore, many cross-links with residues located in W2 region formed more frequently if FOXO4-DBD was bound to DNA. This is consistent with the stabilization of this region observed by HDX-MS and potential movement of this region closer to DNA suggested by protein-DNA crosslinking.

The information provided by structural MS experiments was finally used to guide the molecular modelling of FOXO4-DBD and its complex with DNA. The obtained 3D structures closely matched the previously published NMR and X-ray structures[129,133], thus supporting the validity of the HDX and cross-linking constraints. The resemblance can also serve as a proof, that the differences in interaction with DNA compared to other FOXO family members were not caused by crystal packing issues[129]. Apart from confirming helix H3 as the main interaction interface, significant conformational changes upon DNA binding have been observed not only in the unstructured C- and N- terminal regions but also in loop

H2–H4–H3 and the S2 and S3 strands of wing W1. These structural changes might suggest an adaptive binding mechanism, where conformational changes may be necessary to establish specific substrate-ligand interactions. Furthermore, the fact that the observed conformational changes also involved regions located further away from the interaction interface may support the mechanism in which binding events trigger structural changes necessary to mediate subsequent interaction with additional factors. In conclusion, the performed experiments have confirmed that the utilized combination of structural mass spectrometry methods could effectively guide model-building operations to obtain information about regions inaccessible by the classical high-resolution methods as well as about structural dynamics of the transcription factor-DNA complex.

4.2. Structural Characterization of TEAD1 Recognition of Genomic DNA

All human TEAD transcription factors (whose importance for development of organisms and involvement in cancer was described thoroughly in chapter 1.2.) bind the core 5'-ATTCC-3' consensus binding motif, which has been broadly referred to as "M-CAT". However, searching the JASPAR database has shown, that the flanking sequences around this motif can vary depending on the TEAD family member[104]. Furthermore, we have noticed, that the regulatory regions of human genes that were previously identified to be regulated by TEAD transcription factors are abundant not only in the M-CAT motifs but also in their inverted variant (5'-CCTTA-3'). While there are two high resolution structures of the complexes between the DNA binding domain of TEAD and DNA containing the M-CAT motif available, they were both created by using an artificial DNA which readily crystallized but did not reflect the flanking sequences around the core motif that are actually present in the human genome. Thus, they can not be used to explain how the sequential context of the M-CAT motif affects the interaction[99,100]. Nevertheless, the existence of the crystallographic structure is beneficial for the structural mass spectrometry approach, that was optimized in **Publication III** for transcription factor-DNA complexes.

4.2.1. Publication IV

All the experience gained throughout the preparation of the previous three publications were finally utilized to investigate the structural basis of the interaction of TEAD1 with its specific target DNA binding site. The focus of this study was to investigate whether and how the

sequential context of the M-CAT motif affects its binding properties towards the isolated TEAD1-DBD as well as to explore the possibility of TEAD1-DBD binding to the inverted 5'-CCTTA-3' motif. After the initial determination of the K_D of each complex, structural MS methods (HDX-MS and quantitative chemical cross-linking), molecular docking, and smFRET were utilized for structural characterization. Finally, ChIP-qPCR was employed to correlate the results with a cell line model.

A series of double stranded DNA constructs that placed either the 5'-ATTCC-3' core of the consensus M-CAT motif or its inverted 5'-CCTTA-3' version in different sequence contexts (further referred to as M-CATs) was prepared and the binding properties of their complexes with TEAD1-DBD were initially compared by using native nanoelectrospray ionization MS (nESI-MS). As a result, the complex formation was confirmed for all the tested M-CATs including those containing the inverted motif. Nevertheless, the ratio of signal intensities of the free protein and the complex differed depending on the binding motif orientation (Figure 19A, B). The M-CAT constructs mimicking the CTGF promoter, SRF promoter, and C-MYC first exon which all contained the classical M-CAT were almost completely bound to TEAD1-DBD, while the inverted M-CATs mimicking the C-MYC enhancer, GLUT1 enhancer, and GLUT1 first exon displayed lower percentages of bound form and thus suggesting, that the inverted M-CAT binds to TEAD1-DBD with lower affinity. This observation was subsequently verified by performing a fluorescence anisotropy-based binding assay, which provided the dissociation constant (K_D) of each selected complex (Figure 19C). Its results have shown that the M-CAT motif orientation indeed strongly affects the binding affinity of the dsDNA constructs to TEAD1-DBD, whereas the sequence of the strand surrounding the M-CAT motif has much lower, albeit still significant, influence.

As a next step, quantitative chemical cross-linking with MS detection was utilized to reveal the spatial arrangement of free TEAD1-DBD versus one bound to the differently oriented M-CATs in solution by using the protocol optimized in **Publication III**. The cross-linking experiment resulted in identification of 16 peptide conjugates that yielded 14 unique distance constraints in the control DNA-free sample. These restrains were subsequently employed to guide the homology modeling of DNA-free TEAD1-DBD. The obtained model differed from the template structure only in the N-terminal region (prolonged in our construct by six amino acids), which bent closer to helix H3 than in the template showing that our

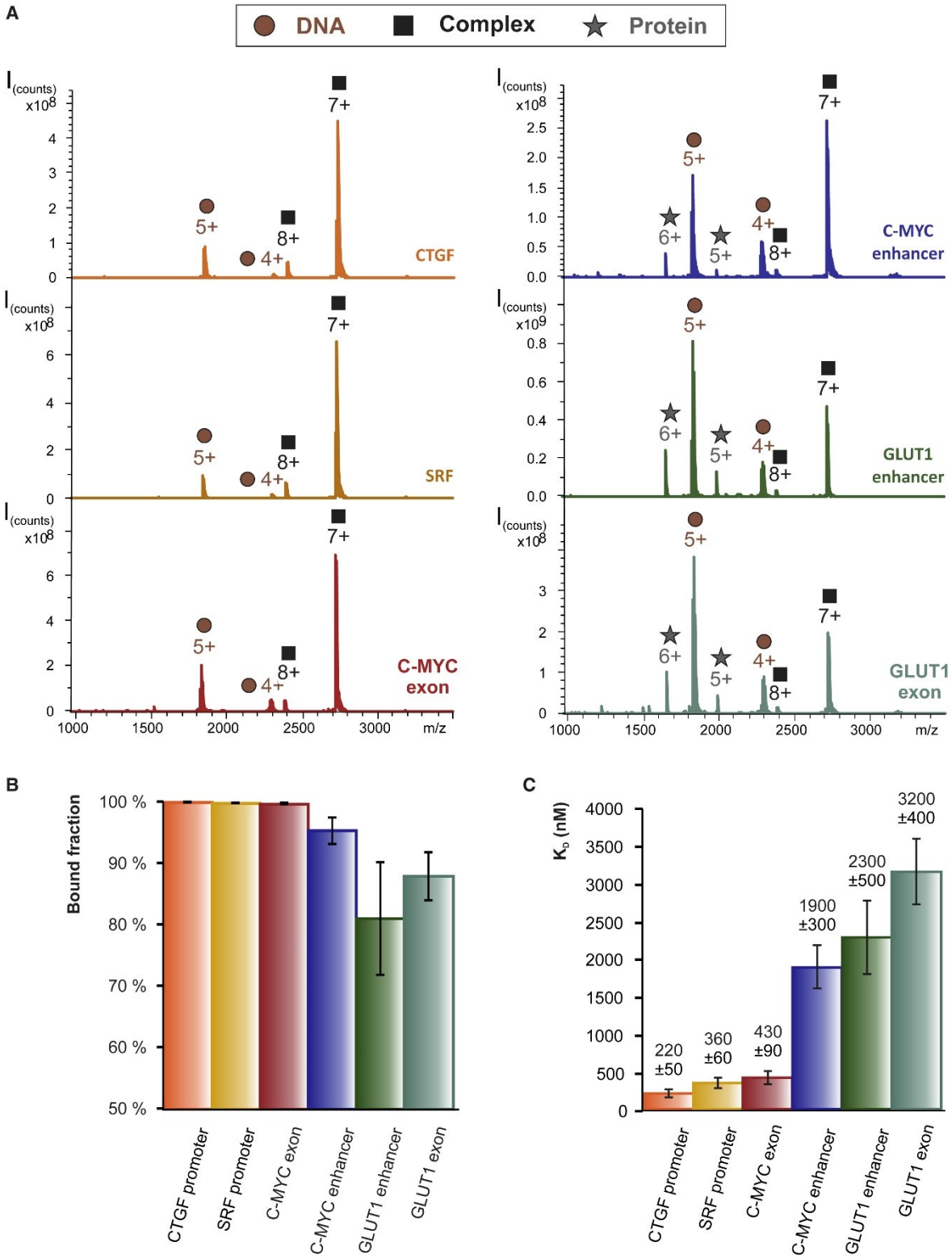


Figure 19: (A) Native ESI-MS spectra of complexes of TEAD1-DBD with each M-CAT *in the study, which revealed the ratios of free versus bound components. The most intense charge states of free protein (gray star), free DNA (brown circle), and the complex (black square) are highlighted* ***(B) Percentage of bound protein*** *derived from the signal intensities in the nESI-MS spectra.* ***(C) Comparison of dissociation constants (K_D) of selected TEAD1-DBD/M-CAT complexes determined by fluorescence anisotropy binding assay.*** *Complexes containing M-CATs with binding motifs in the 5' to 3' orientation (i.e., SRF promoter, CTGF promoter, and C-MYC exon) had approximately 10 times higher K_D than those with the inverted motif.*

construct adopted the same fold as in the previously published high-resolution structures[59]. Similarly to the FOXO4-DBD/DAF16 complex, all 16 cross-linked peptides observed in the free TEAD1-DBD sample were also found in all six TEAD1-DBD/M-CAT complex samples while the abundances of some of them were enhanced or reduced in presence of DNA (Figure 20). Interestingly, the DNA binding affected not only residues located on the H3 helix and L1 loop that were previously identified as the binding interface[59,100] (for example cross-links with K88 located on H3 almost completely vanished in the bound state), but also residues located in the C- and N- terminal regions (probability of formation of cross-links between N-terminal amino group and K65 as well as between K101 and K57 increased in presence of DNA). This effect might be attributed to the fact, that the affected residues are all located on flexible regions that might lose their flexibility upon DNA binding which is in agreement with what was already suggested for the C- terminal helix[100]. The quantitative cross-linking results of the six M-CATs differing in motif orientation were, however, very similar. The only discrepancies induced by the sequential contexts were likely caused by the different affinities of TEAD1-DBD to each M-CAT, since the quantification results correlated well with the measured dissociation constants. Therefore, the results indicate that the overall complex structure may not be significantly affected by the orientation of the consensus motif.

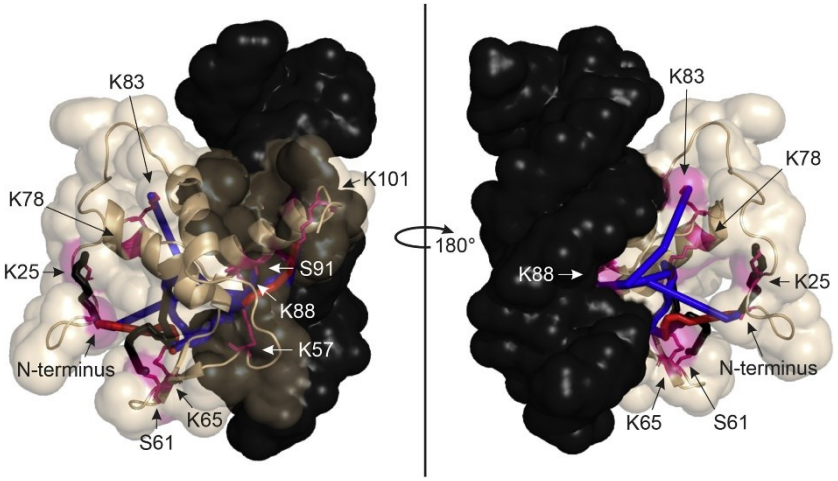


Figure 20: Identified cross-links displayed on a TEAD1-DBD/M-CAT model. *Cross-links favored in the complex state are colored red, cross-links hampered by DNA binding are colored blue, and cross-links that formed independently on DNA are colored black. Residues susceptible to the used cross-linking reagent are highlighted by pink color.*

To identify the binding interface, HDX-MS was performed according to the protocol optimized in **Publication II**. The results have shown that for all M-CATs at short deuteration times, the largest difference in deuterium uptake between free and bound states was observed in helix H3 and the adjacent L2 loop while a slightly above the significance limit protection

from deuteration was observed also in the L1 loop. Since H3 and L1 were previously identified as being directly responsible for DNA binding, this effect was to be expected. On the other hand, L2 was suggested to be involved in DNA binding by an NMR study performed on TEAD1-DBD/M-CAT[59] but the more recently published crystal structure of the TEAD4-DBD/M-CAT complex did not show any direct contact between this loop and DNA[100] which may indicate the two homologs to have a slightly different binding mechanism. In longer deuteration times, similar effect as with the FOXO4-DBD/DAF16 complex discussed in **Publication III** occurred since a significant protection of deuteration in the bound form was displayed by helices H1 and H2. Even here, this protection was probably caused by stabilization of the TEAD1-DBD structure in a more fixed conformation upon DNA binding. No significant differences apart from those that could be attributed to different dissociation constants were observed between the two motif orientations however, as the same regions were affected by DNA binding and the only difference was the degree of the protection effect.

While the information obtained so far did not reveal any significant difference in the mechanisms used by TEAD1-DBD to recognize the different M-CAT orientations, a possibility remained that the protein may bind the inverted M-CATs by using the same interacting region, but in an actual orientation of the entire protein rotated by 180°. This hypothesis was tested by using molecular docking experiments with a series of models of TEAD1-DBD bound to dsDNA constructs which placed the sequences of the C-MYC exon and C-MYC enhancer in different structural contexts. Figure 21A then displays the structures of the two complexes that manifested the most stable interactions which corresponded to the initial TEAD1-DBD/C-MYC exon complex containing the classical M-CAT followed by the C-MYC enhancer sequence, with the M-CAT motif modeled in the complementary strand which is equivalent to a 180° rotation of the entire TEAD1-DBD (Fig. 21B). The simulations also revealed that the shapes and widths of the interacting grooves were very similar in both motif orientations which provides an explanation for how it was possible for helix H3 to fit in the DNA major groove regardless of the motif orientation. Moreover, according to the ΔG calculations, the rotated orientation could lead to less stable interactions between protein and DNA which explains the observed lower affinity.

To support the computational data experimentally, a single-molecule Förster resonance energy transfer (smFRET) study was performed with a 16 bp long version of the dsDNA

with asymmetrically placed binding motif in both orientations labelled at one end with Alexa 647 and TEAD1-DBD modified with Cy3 dye on C53. The high affinity M-CAT served as a control where the distance between fluorophores (based on the X-ray structure model[100]) could be either 46.6 Å if the 5' end of the forward strand is labeled or 25.4 Å if the reverse strand is labeled on the 5' end. As expected, the former setting resulted in energy transfer efficiency of 0.58 while the latter resulted in optimal FRET with energy transfer efficiency of 1.00. In case of the inverted M-CAT labeled on 5' end of the reverse strand, the distance between fluorophores depended on the mutual protein-DNA orientation. If our assumption was correct and the protein binds this motif in a 180° rotated orientation, the distance would be 46.6 Å (and the energy transfer efficiency therefore close to the 0.58 measured for this distance in the control sample), whereas if the protein is not rotated and its orientation toward the DNA stays similar to the published structural model, the distance between fluorophores would be only 28.2 Å which would result in energy transfer efficiency close to 1. The observed energy transfer efficiency was

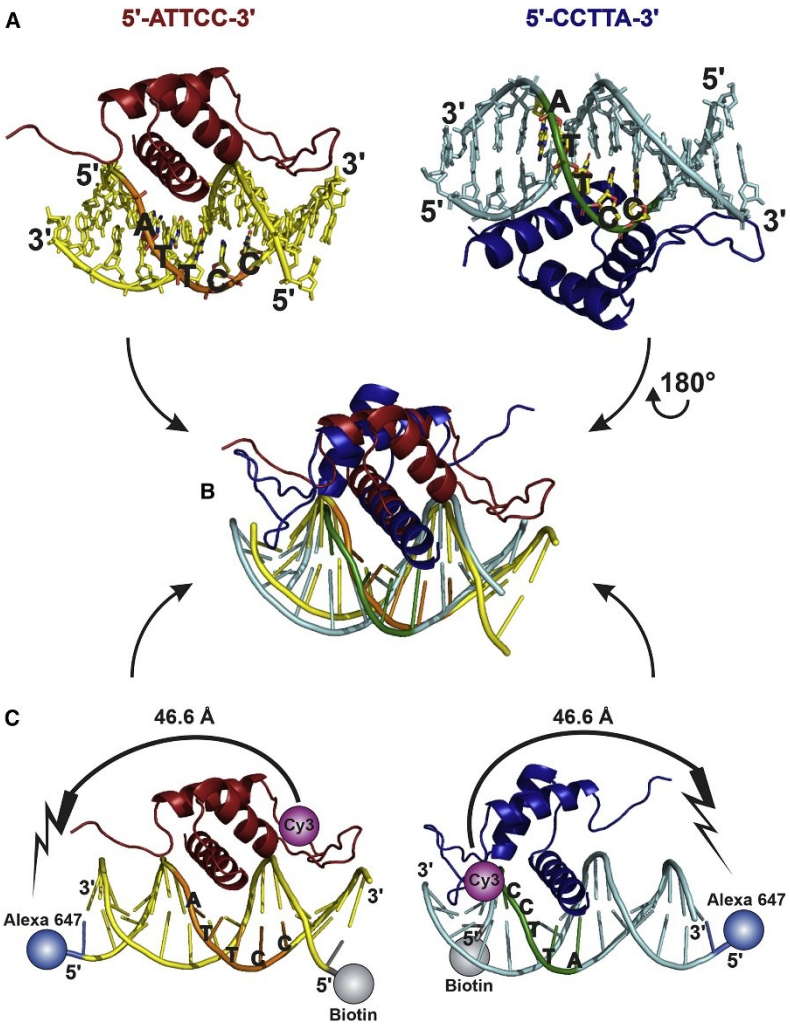


Figure 21: (A) TEAD1-DBD/M-CAT models used for MD simulations *showing the relative position and orientation of the C-MYC enhancer 5'-CCTTA-3' and the C-MYC exon 5'-ATTCC-3' DNA sequences with respect to the TEAD1-DBD.* ***(B) Structure superposition of DNA constructs*** *corresponding to the most stable interactions according to ΔG calculations* ***(C) smFRET study****. DNA and protein were labelled with donor and acceptor fluorophores whose distance (and thus FRET effectivity) depended on the respective orientations of the binding partners.*

0.51 which confirmed the results obtained by molecular docking simulations together with the idea of 180° rotated binding orientation (Fig. 21C).

To assess the possible biological relevance of the presence of the low affinity binding motifs in the human genome in vivo, the relative TEAD1 occupancy of both orientation binding sites was determined by using chromatin immunoprecipitation (ChIP) analysis followed by qPCR quantification. In agreement with the binding affinities observed in in vitro experiments, the resulting data showed high-level occupancy of the 5'-3' oriented C-MYC exon, significantly lower occupancy of the C-MYC enhancer containing the inverted M-CAT, and non-significant occupancy of the control region which did not contain the M-CAT motif in any orientation.

In conclusion, the inverted 5'-CCTTA-3' motif was found to be able to bind TEAD1-DBD with lower affinity than the classical M-CAT both in vivo and in vitro while the surrounding sequence of the core motif also have an influence, although not so significant. The structural MS experiments confirmed the previously identified regions L1, L2 and H3 as the binding interface, revealed a considerable loss of flexibility occurring upon DNA binding but failed to provide an explanation for the low affinity binding of the inverted M-CAT. MD simulations then revealed (and smFRET experiment subsequently confirmed) that TEAD1-DBD can bind to the inverted motif in 180° rotated orientation while suggesting, that TEAD1-DBD may at first recognize the overall shape of the major groove, which is similar in both orientations, and then the specific amino acid-nucleotide interactions, whose number and strength differ depending on the motif orientation, stabilize the complex. Taken together, the presence of M-CAT sites with widely different affinities in the human genome may provide the basis for possible regulatory mechanisms relying on the actual concentration of a certain transcription factor in the proximity to a gene regulatory region as was described in chapter 1.1.4. and already reported for some other transcription factors[52,188].

4.2.2. Influence of the flanking sequences around the core M-CAT motif on its interaction with TEAD1-DBD

In **Publication IV** we have found that the DNA sequence surrounding the M-CAT motif has a significant influence on the affinity of the dsDNA construct to TEAD1-DBD (Fig. 19). To examine this effect more closely, and to find out how exactly the bases flanking the core motif affect the interaction with TEAD1-DBD, a series of dsDNA constructs, which all

originated from regulatory regions of human genes and contained the core ATTCC binding motif either in the classical 5'-3' or in the inverted 3'-5' orientation, was prepared. What these constructs differed in were the identity of the two bases on the 5' side of the M-CAT core motif and one base on its 3' side, which were selected to correspond to the two most frequent bases in each position as is deposited in the JASPAR database and summarized in Figure 9. Two M-CATs were selected for each combination of flanking bases.

First, the TEAD1-DBD's ability to form complexes with all the selected dsDNA constructs had to be checked and the dissociation constant of each complex had to be estimated. Since methods usually used for protein-DNA K_D determination (such as thermophoresis, fluorescence anisotropy or gel shifts) are tedious and sample consuming or need one of the complex components to be labelled, they are not suitable for evaluation of multiple complexes in short time. Thus, we have tested the potential of native nESI-MS for K_D determination using the same six M-CATs with known dissociation constants from **Publication IV**. For all the six M-CATs the bound fraction was calculated and compared with the expected bound fraction calculated from the known K_D. The results (summarised in Figure 22) have shown that, unfortunately, the method could not be used to determine the actual K_D since the bound fractions measured by native MS were higher than expected. This could probably be attributed to the fact, that DNA was in the nESI ion source charged more easily (lower voltage was needed for it to be visible in the spectrum) than the protein because similar effect was, interestingly, already observed in a previous study of ssDNA/dsDNA equilibria[189]. On the other hand, the nESI-MS results followed the

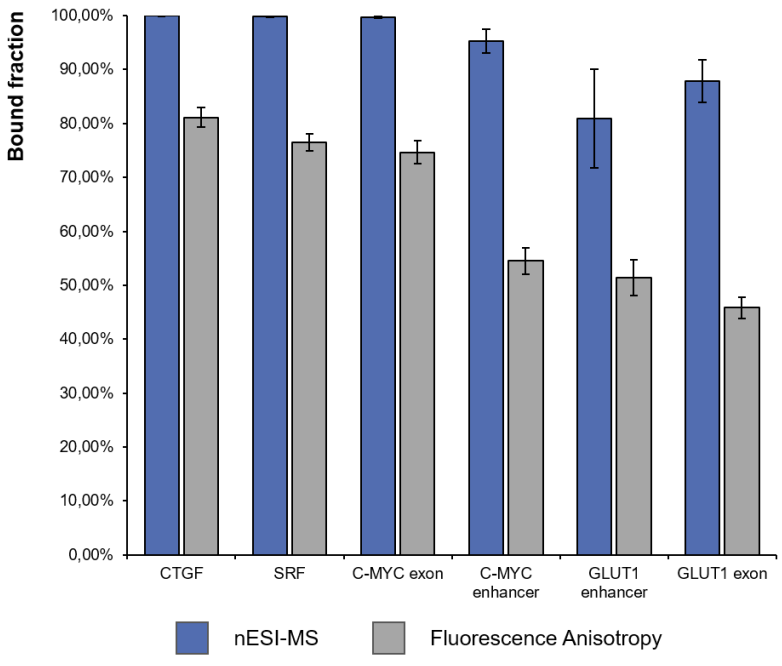


Figure 22: Comparison of calculated bound fractions *from nESI-MS experiments and those expected according to the known K_D of each complex which was previously determined by fluorescence anisotropy assay.*

same trend as those obtained by fluorescence anisotropy assay, and it was clearly possible to distinguish between the weak binders and strong binders. Therefore, the nESI-MS was proved to be a useful tool for fast differentiation between high affinity and low affinity oligonucleotides with low sample consumption.

The native MS was subsequently utilized to evaluate the complex formation and to sort the series of M-CATs with different bases flanking the core binding motif according to their affinity to TEAD1-DBD. The results, which are summarized in Figure 23, have shown that all the tested M-CATs were able to form complex with TEAD1-DBD and that all the M-CATs containing the inverted motif belonged to the weak binders, which is in perfect agreement with the results obtained in **Publication IV**. Moreover, the bound fractions of the two M-CATs with the same flanking base combination were, in most cases, very similar, showing that the identity of the bases flanking the core binding motif might, indeed, be the main reason for the differences in the affinity of these dsDNA constructs to the TEAD1-DBD.

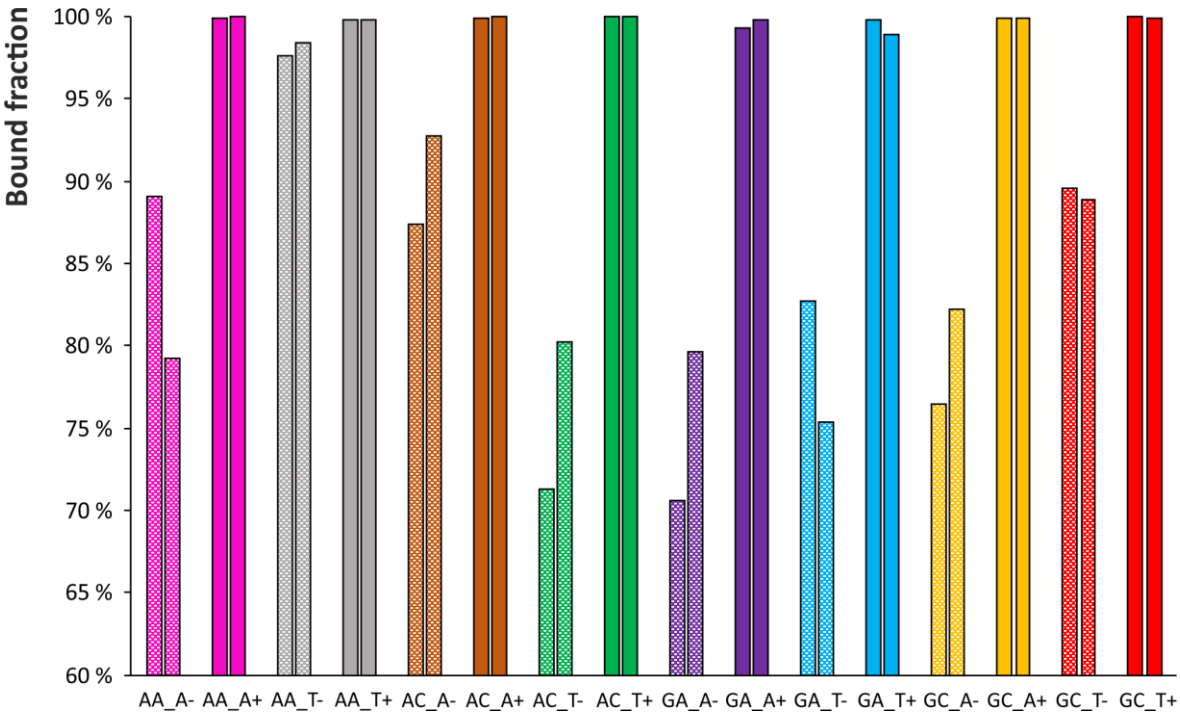


Figure 23: Bound fractions *calculated from native MS spectra of complexes of TEAD1-DBD with oligonucleotides with varying bases around the binding motif. Position of the core binding motif is in the labels shown as an underscore while the motif orientation is shown by + (the classical M-CAT) or – (the inverted one). It is clearly possible to differentiate strong binders (uniform fill) from weak binders (squared).*

Following the initial screening, the HDX-MS was employed to further explore the binding interface. The results (Fig. 24) have shown that, similarily to the six M-CATs examined in **Publication IV**, the highest differences in deuteration rates could be observed in helix H3 and the L2 loop preceding it, thus suggesting, that all M-CATs bind to the same

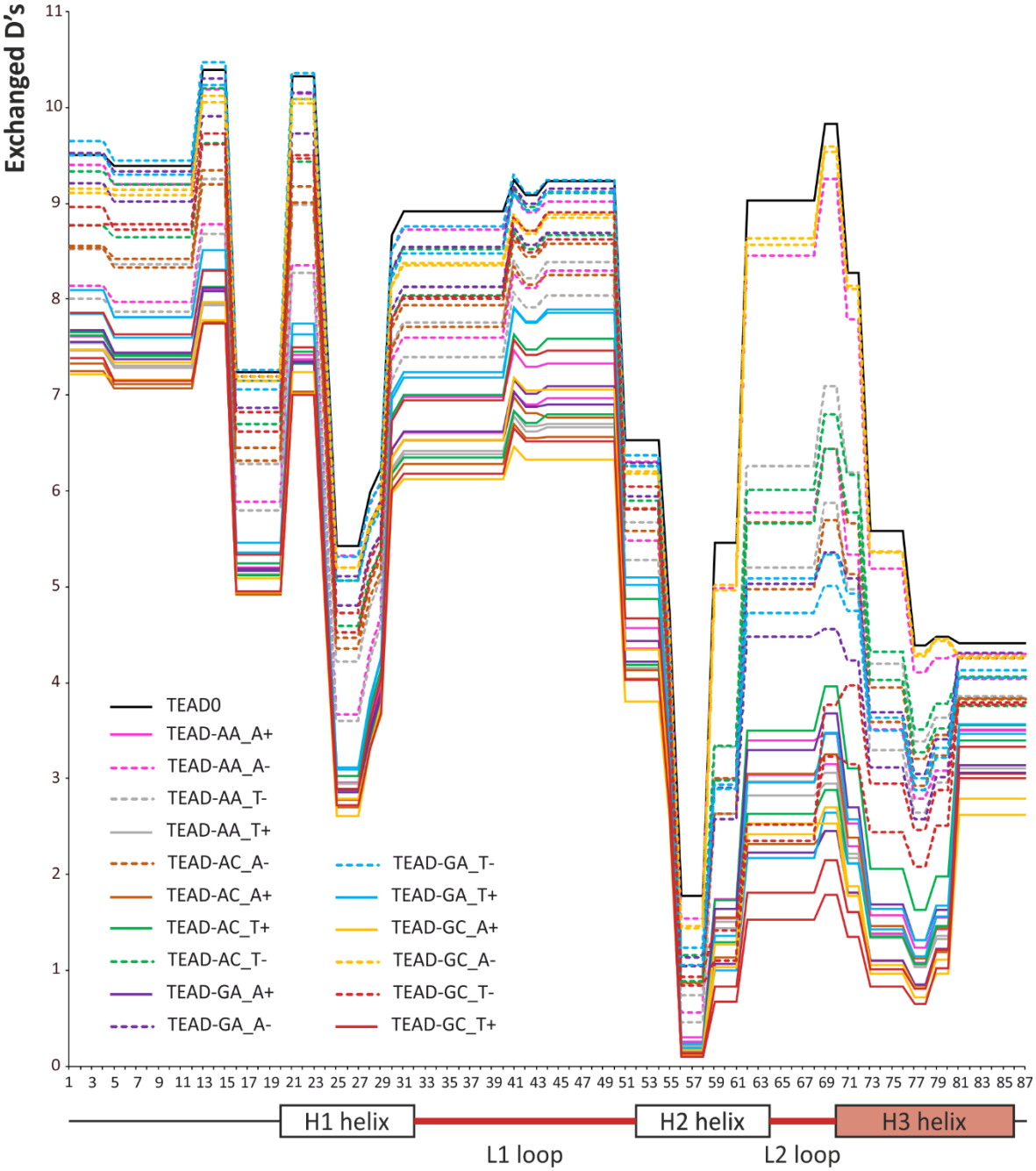


Figure 24: Comparison of HDX-MS results *depicted as number of exchanged deuterium atoms along the sequence for the series of M-CATs differing in the identity of bases flanking the core binding motif (shown as XX_X in the labels where the underscore signifies the position of the core motif). Regions previously identified as responsible for DNA binding are highlighted in red in the structure scheme under the picture. Inverted binding motives are shown as dotted lines and labelled by a minus symbol. The strongest binders exchange less deuterium atoms and therefore are positioned in the bottom of the chart while the weakest binders show a similar pattern as the sample where no DNA was present (TEAD0 - black line).*

region of TEAD1-DBD. However, as was shown in the data published in **Publication IV**, the degree of protection from deuteraion strongly depends on the affinity of each M-CAT to the protein. Therefore, the results allowed identification of the strongest binders (5'-GCATTCC(T/A)-3') as well as the weakest binder (3'-GCATTCCA-5'). Interestingly, the sequence identified as the strongest binder differs from the most abundant sequence of a TEAD binding site deposited in the JASPAR database in which an A was at the 5'-terminal position instead of a G suggesting that the higher affinity might potentially compensate for the lower abundance of this motif in the human genome[104].

Currently, the actual dissociation constants of each complex are being measured to complement the obtained information and a publication concerning the influence of the flanking sequences around the core M-CAT motif on its interaction with TEAD1-DBD is being prepared.

4.2.3. Possible interaction of transcription factors TEAD1 and FOXO4

While searching the regulatory regions of human genes for transcription factor binding sites, one of our colleagues has noticed that the TEAD and FOXO4 binding motifs often colocalize in close proximity to each other. Therefore, several dsDNA constructs mimicking these genomic regions were prepared and their ability to bind the DNA binding domains of both transcription factors at the same time was evaluated by a gel shift assay and native MS. Out of the five tested DNA constructs, which differed in relative orientation of the two binding sites and in the number of base pairs separating them, a complex composed of all three components (TEAD1-DBD, FOXO4-DBD and the dsDNA) was identified in four of them. Results of two example experiments are shown in Figure 25, which, in addition to that, illustrates how native MS allowed identification of the triple complex even when its low abundance prevented its manifestation in the gel shift assay.

Interestingly, the highest amounts of the ternary complex were observed in samples containing dsDNA construct where the two binding motifs overlapped in one base pair. The ternary complex was also found in both samples where the motifs were separated by one base pair, however the mutual orientation of the two motifs significantly affected the observed signal intensities which are proportional to the amount of the formed complex. However, the presence of the ternary complex was not confirmed for the dsDNA construct where the two motifs were separated by two base pairs. The increase in complex formation

in samples where the two motifs are closer to each other may suggest that the two transcription factors interact together and thus present an interesting topic for future research.

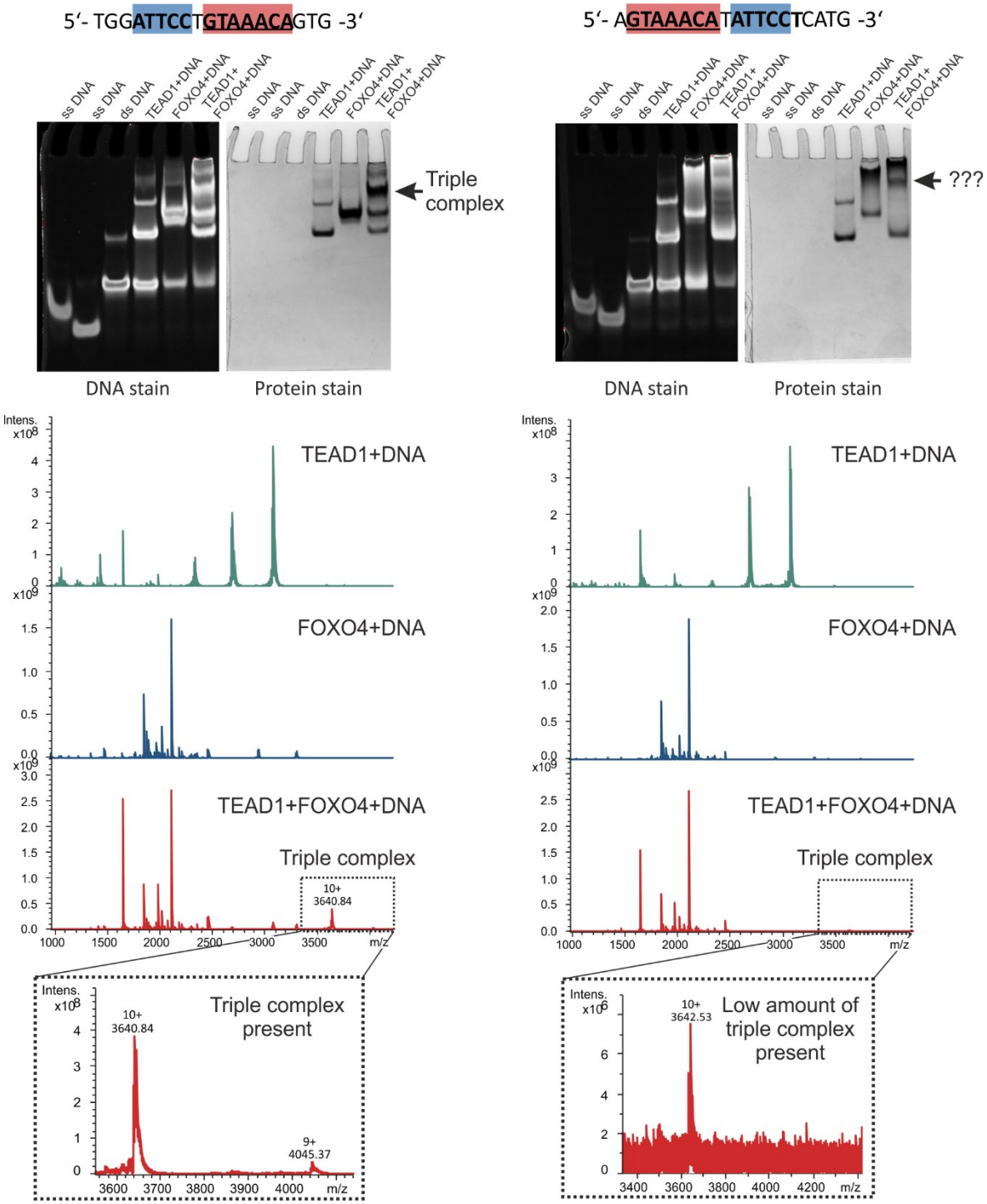


Figure 25: Confirmation of presence of a complex of TEAD1-DBD, FOXO4-DBD and one dsDNA construct. *While in one case (left column), the complex could be observed directly from the gel shift assay and the native MS only confirmed its identity, in the other (right column) due to low amount of the complex being formed, the sensitive native MS analysis was necessary to even observe its presence. The position of the TEAD1 binding motif is highlighted in blue and that of FOXO4 in red in the oligonucleotide sequence in the top part of the picture. Their mutual position is probably the reason for the different amount of the complex formed.*

5. Summary

The aims of this thesis were to contribute to development of a set of structural mass spectrometry methods for characterization of transcription factor complexes with their cognate response motifs and to apply these methods to structurally characterize the interaction between the DNA binding domains of FOXO4 and TEAD1 proteins and their DNA response motifs. The following results were obtained and included in the four attached publications:

- A set of alternative HDX-MS compatible immobilized proteases needed to improve sequence coverage and resolution was prepared, characterized and their usability was tested on two protein systems
- The choice of the best protease for digestion of a givens system in HDX-MS is highly protein specific even in case of proteins sharing a similar fold
- Addition of denaturant urea to quench buffer as well as including washing steps between analyzed samples and screening for the best option from a set of alternative proteases is beneficial to obtain best possible resolution from HDX-MS analysis of transcription factor-DNA complexes
- The combination of structural mass spectrometry methods can effectively guide model-building operations to obtain information about regions inaccessible by the classical high-resolution methods as well as about structural dynamics of the transcription factor-DNA complex
- A significant loss of flexibility upon DNA binding was observed in both FOXO4/DAF16 and TEAD1/M-CAT complexes
- TEAD1-DBD is able to bind to the inverted 5'-CCTTA-3' M-CAT motif with lower affinity both in vitro and in vivo
- TEAD1-DBD binds to the inverted motif in a 180° rotated orientation
- Native nESI-MS was found to be a quick method for confirmation of transcription factor-DNA complex formation with the ability to differentiate between strong and week binders with very low sample consumption
- Sequences of the regions flanking the M-CAT motif affect its binding affinity to TEAD1-DBD as well with 5'-GCATTCC(T/A)-3' being the strongest binder
- TEAD1-DBD and FOXO4-DBD are able to bind together to one oligonuceotide containing both response motifs

List of publications:

1) Möller, I. R., Slivacka, M., Hausner, J., Nielsen, A. K., Pospíšilová, E., Merkle, P. S., **Lišková, R**., Polák, M., Loland, C. J., Kádek, A., Man, P., & Rand, K. D. (2019). Improving the Sequence Coverage of Integral Membrane Proteins during Hydrogen/Deuterium Exchange Mass Spectrometry Experiments. *Analytical Chemistry*, *91*(17), 10970–10978.

2) **Filandrova, R**., Kavan, D., Kadek, A., Novak, P., & Man, P. (2021). Studying Protein–DNA Interactions by Hydrogen/Deuterium Exchange Mass Spectrometry. In *Multiprotein Complexes: Methods and Protocols (Methods in Molecular Biology (2247*)), 193–219, Humana Press Inc.

3) Slavata, L., Chmelík, J., Kavan, D., **Filandrová, R**., Fiala, J., Rosůlek, M., Mrázek, H., Kukačka, Z., Vališ, K., Man, P., Miller, M., McIntyre, W., Fabris, D., & Novák, P. (2019). Ms-based approaches enable the structural characterization of transcription factor/DNA response element complex. *Biomolecules*, *9*(10), 1–21

4) **Filandrová, R**., Vališ, K., Černý, J., Chmelík, J., Slavata, L., Fiala, J., Rosůlek, M., Kavan, D., Man, P., Chum, T., Cebecauer, M., Fabris, D., & Novák, P. (2021). Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA. *Structure*, 29, 1–12

References:

1. Fulton, D. L., Sundararajan, S., Badis, G., Hughes, T. R., Wasserman, W. W., Roach, J. C. & Sladek, R. TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.* **10,** R29 (2009).
2. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. & Weirauch, M. T. The Human Transcription Factors. *Cell* **172,** 650–665 (2018).
3. Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Bretscher, A., Ploegh, H., Amon, A. & Scott, M. P. *Molecular Cell Biology*. (W.H. Freeman and Company, 2013).
4. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10,** 252–63 (2009).
5. Accili, D. & Arden, K. C. FoxOs at the crossroads of cellular metabolism, differentiation, and transformation. *Cell* **117,** 421–426 (2004).
6. De Pooter, R. F. & Kee, B. L. E proteins and the regulation of early lymphocyte development. *Immunol. Rev.* **238,** 93–109 (2010).
7. Lania, L., Majello, B. & Napolitano, G. Transcriptional control by cell-cycle regulators: A review. *J. Cell. Physiol.* **179,** 134–141 (1999).
8. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.* **43,** 73–81 (2017).
9. Boyadjiev, S. & Jabs, E. Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clin. Genet.* **57,** 253–266 (2000).
10. Furney, S. J., Higgins, D. G., Ouzounis, C. A. & López-Bigas, N. Structural and functional properties of genes involved in human cancer. *BMC Genomics* **7,** 1–11 (2006).
11. Andersson, R., Sandelin, A. & Danko, C. G. A unified architecture of transcriptional regulatory elements. *Trends Genet.* **31,** 426–33 (2015).
12. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13,** 233–45 (2012).
13. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).
14. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–286 (2014).
15. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144,** 327–339 (2011).
16. Plank, J. L. & Dean, A. Enhancer function: mechanistic and genome-wide insights come together. *Mol. Cell* **55,** 5–14 (2014).
17. Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H. & Shiroishi, T. Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Dev. Cell* **16,** 47–57 (2009).
18. Fukaya, T., Lim, B. & Levine, M. Enhancer Control of Transcriptional Bursting. *Cell* **166,** 358–368 (2016).
19. Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S. & Natoli, G. Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA

Features of Mammalian Enhancers. *Mol. Cell* **54,** 844–857 (2014).

20. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166,** 538–554 (2016).

21. Zaret, K. S. & Mango, S. E. Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr. Opin. Genet. Dev.* **37,** 76–81 (2016).

22. Li, Z., Gadue, P., Chen, K., Jiao, Y., Tuteja, G., Schug, J., Li, W. & Kaestner, K. H. Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* **151,** 1608–1616 (2012).

23. Saha, A., Wittmeyer, J. & Cairns, B. R. Chromatin remodelling: The industrial revolution of DNA around histones. *Nat. Rev. Mol. Cell Biol.* **7,** 437–447 (2006).

24. Carrera, I. & Treisman, J. E. Message in a nucleus: signaling to the transcriptional machinery. *Curr. Opin. Genet. Dev.* **18,** 397–403 (2008).

25. Hsu, H. T., Chen, H. M., Yang, Z., Wang, J., Lee, N. K., Burger, A., Zaret, K., Liu, T., Levine, E. & Mango, S. E. Erratum for the Report 'Recruitment of RNA polymerase II by the pioneer transcription factor PHA-4' by H.-T. Hsu, H.-M. Chen, Z. Yang, J. Wang, N. K. Lee, A. Burger, K. Zaret, T. Liu, E. Levine, S. E. Mango. *Science (80-.).* **350,** aad5928–aad5928 (2015).

26. Cramer, P. Organization and regulation of gene transcription. *Nature* **573,** 45–54 (2019).

27. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).

28. Kadonaga, J. T., Courey, A. J., Ladika, J. & Tjian, R. Distinct regions of Sp1 modulate DNA binding and transcriptional activation. *Science (80-.).* **242,** 1566–1570 (1988).

29. Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V. & Cisse, I. I. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science (80-.).* **361,** 412–415 (2018).

30. Guo, Y. E., Manteiga, J. C., Henninger, J. E., Sabari, B. R., Dall'Agnese, A., Hannett, N. M., Spille, J., Afeyan, L. K., Zamudio, A. V., Shrinivas, K., Abraham, B. J., Boija, A., Decker, T., Rimel, J. K., Fant, C. B., Lee, T. I., Cisse, I. I., Sharp, P. A., Taatjes, D. J. & Young, R. A. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* **572,** 543–548 (2019).

31. Robinson, P. J., Trnka, M. J., Bushnell, D. A., Davis, R. E., Mattei, P.-J., Burlingame, A. L. & Kornberg, R. D. Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. *Cell* **166,** 1411-1422.e16 (2016).

32. Chen, F. X., Smith, E. R. & Shilatifard, A. Born to run: control of transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **19,** 464–478 (2018).

33. Core, L. & Adelman, K. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes Dev.* **33,** 960–982 (2019).

34. Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., Sharp, P. A. & Young, R. A. C-Myc regulates transcriptional pause release. *Cell* **141,** 432–445 (2010).

35. Li, Y., Liu, M., Chen, L.-F. & Chen, R. P-TEFb: Finding its ways to release promoter-proximally paused RNA polymerase II. *Transcription* **9,** 88–94 (2018).

36. Eguchi, A., Lee, G. O., Wan, F., Erwin, G. S. & Ansari, A. Z. Controlling gene networks and cell fate with precision-targeted DNA-binding proteins and small-molecule-based genome readers. *Biochem. J.* **462,** 397–413 (2014).

37. Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E. E. M. & Taipale, J. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **2015,** 1–20 (2015).

38. Fedotova, A. A., Bonchuk, A. N., Mogila, V. A. & Georgiev, P. G. C2H2 Zinc Finger Proteins: The Largest but Poorly Explored Family of Higher Eukaryotic Transcription Factors. *Acta Naturae* **9,** 47–58 (2017).

39. Fraenkel, E., Rould, M. A., Chambers, K. A. & Pabo, C. O. Engrailed homeodomain-DNA complex at 2.2 Å resolution: A detailed view of the interface and comparison with other engrailed structures. *J. Mol. Biol.* **284,** 351–361 (1998).

40. Frietze, S. & Farnham, P. J. Transcription factor effector domains. *Subcell. Biochem.* **52,** 261–77 (2011).

41. Garza, A. S., Ahmad, N. & Kumar, R. Role of intrinsically disordered protein regions/domains in transcriptional regulation. *Life Sci.* **84,** 189–193 (2009).

42. Minezaki, Y., Homma, K., Kinjo, A. R. & Nishikawa, K. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *J. Mol. Biol.* **359,** 1137–49 (2006).

43. Fry, C. J., Pearson, A., Malinowski, E., Bartley, S. M., Greenblatt, J. & Farnhamt, P. J. Activation of the murine dihydrofolate reductase promoter by E2F1: A requirement for CBP recruitment. *J. Biol. Chem.* **274,** 15883–15891 (1999).

44. Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu. Rev. Cell Dev. Biol.* **35,** 357–379 (2019).

45. Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H. J. & Mann, R. S. Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. *Cell* **147,** 1270–1282 (2011).

46. Ramos, A. I. & Barolo, S. Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philos. Trans. R. Soc. B Biol. Sci.* **368,** (2013).

47. Wang, J., Malecka, A., Trøen, G. & Delabie, J. Comprehensive genome-wide transcription factor analysis reveals that a combination of high affinity and low affinity DNA binding is needed for human gene regulation. *BMC Genomics* **16 Suppl 7,** S12 (2015).

48. Nagy, G. & Nagy, L. Motif grammar: The basis of the language of gene expression. *Comput. Struct. Biotechnol. J.* **18,** 2026–2032 (2020).

49. Farley, E. K., Olson, K. M., Zhang, W., Rokhsar, D. S. & Levine, M. S. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl. Acad. Sci.* **113,** 6508–6513 (2016).

50. Kribelbauer, J. F., Laptenko, O., Chen, S., Martini, G. D., Freed-Pastor, W. A., Prives, C., Mann, R. S. & Bussemaker, H. J. Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. *Cell Rep.* **19,** 2383–2395 (2017).

51. Zeiske, T., Baburajendran, N., Kaczynska, A., Brasch, J., Palmer, A. G., Shapiro, L., Honig, B. & Mann, R. S. Intrinsic DNA Shape Accounts for Affinity Differences between Hox-Cofactor Binding Sites. *Cell Rep.* **24,** 2221–2230 (2018).

52. Lorenzin, F., Benary, U., Baluapuri, A., Walz, S., Jung, L. A., von Eyss, B., Kisker, C., Wolf, J., Eilers, M. & Wolf, E. Different promoter affinities account for specificity in MYC-dependent gene regulation. *Elife* **5,** 1–35 (2016).

53. Zheng, Y. & Levens, D. Tuning the MYC response. *Elife* **5,** 2015–2017 (2016).

54. Tsai, A., Muthusamy, A. K., Alves, M. R. P., Lavis, L. D., Singer, R. H., Stern, D. L. & Crocker, J. Nuclear microenvironments modulate transcription from low-affinity enhancers. *Elife* **6,** 1–18 (2017).

55. Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R. S. & Stern, D. L. Low affinity binding site clusters confer HOX specificity and regulatory robustness. *Cell* **160,** 191–203 (2015).

56. Jacquemin, P., Hwang, J. J., Martial, J. A., Dollé, P. & Davidson, I. A novel family of developmentally regulated mammalian transcription factors containing the TEA/ATTS DNA binding domain. *J. Biol. Chem.* **271,** 21775–85 (1996).

57. Xiao, J. H., Davidson, I., Ferrandon, D., Rosales, R., Vigneron, M., Macchi, M., Ruffenach, F. & Chambon, P. One cell-specific and three ubiquitous nuclear proteins bind in vitro to overlapping motifs in the domain B1 of the SV40 enhancer. *EMBO J.* **6,** 3005–13 (1987).

58. Landin-Malt, A., Benhaddou, A., Zider, A. & Flagiello, D. An evolutionary, structural and functional overview of the mammalian TEAD1 and TEAD2 transcription factors. *Gene* **591,** 292–303 (2016).

59. Anbanandam, A., Albarado, D. C., Nguyen, C. T., Halder, G., Gao, X. & Veeraraghavan, S. Insights into transcription enhancer factor 1 (TEF-1) activity from the solution structure of the TEA domain. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 17225–30 (2006).

60. Huh, H., Kim, D., Jeong, H.-S. & Park, H. Regulation of TEAD Transcription Factors in Cancer Biology. *Cells* **8,** 600 (2019).

61. Santucci, M., Vignudelli, T., Ferrari, S., Mor, M., Scalvini, L., Bolognesi, M. L., Uliassi, E. & Costi, M. P. The Hippo Pathway and YAP/TAZ–TEAD Protein–Protein Interaction as Targets for Regenerative Medicine and Cancer Treatment. *J. Med. Chem.* **58,** 4857–4873 (2015).

62. Pobbati, A. V. & Hong, W. Emerging roles of TEAD transcription factors and its coactivators in cancers. *Cancer Biol. Ther.* **14,** 390–8 (2013).

63. Kaneko, K. J. & DePamphilis, M. L. Regulation of gene expression at the beginning of mammalian development and the TEAD family of transcription factors. *Dev. Genet.* **22,** 43–55 (1998).

64. Kaneko, K. J., Cullinan, E. B., Latham, K. E. & DePamphilis, M. L. Transcription factor mTEAD-2 is selectively expressed at the beginning of zygotic gene expression in the mouse. *Development* **124,** 1963–1973 (1997).

65. Jacquemin, P., Sapin, V., Alsat, E., Evain-Brion, D., Dollé, P. & Davidson, I. Differential expression of the TEF family of transcription factors in the murine placenta and during differentiation of primary human trophoblasts in vitro. *Dev. Dyn.* **212,** 423–36 (1998).

66. Chen, Z., Friedrich, G. A. & Soriano, P. Transcriptional enhancer factor 1 disruption by a retroviral gene trap leads to heart defects and embryonic lethality in mice. *Genes Dev.* **8,** 2293–2301 (1994).

67. Liu, R., Lee, J., Kim, B. S., Wang, Q., Buxton, S. K., Balasubramanyam, N., Kim, J. J., Dong, J., Zhang, A., Li, S., Gupte, A. A., Hamilton, D. J., Martin, J. F., Rodney, G. G., Coarfa, C., Wehrens, X. H., Yechoor, V. K. & Moulik, M. Tead1 is required for maintaining adult cardiomyocyte function, and its loss results in lethal dilated cardiomyopathy. *JCI insight* **2,** (2017).

68. Sawada, A., Kiyonari, H., Ukita, K., Nishioka, N., Imuta, Y. & Sasaki, H. Redundant roles of Tead1 and Tead2 in notochord development and the regulation of cell proliferation and survival. *Mol. Cell. Biol.* **28,** 3177–89 (2008).

69. Kaneko, K. J., Kohn, M. J., Liu, C. & DePamphilis, M. L. Transcription factor TEAD2 is involved in neural tube closure. *Genesis* **45,** 577–87 (2007).

70. Yagi, R., Kohn, M. J., Karavanova, I., Kaneko, K. J., Vullhorst, D., DePamphilis, M. L. & Buonanno, A. Transcription factor TEAD4 specifies the trophectoderm lineage at the beginning of mammalian development. *Development* **134,** 3827–36 (2007).

71. Kaneko, K. J. & De Pamphilis, M. L. TEAD4 establishes the energy homeostasis essential for blastocoel formation. *Dev.* **140,** 3680–3690 (2013).

72. Zhang, H., Pasolli, H. A. & Fuchs, E. Yes-associated protein (YAP) transcriptional coactivator functions in balancing growth and differentiation in skin. *Proc Natl Acad Sci U S A* **108,** 2270–2275 (2011).

73. Zhao, B., Ye, X., Yu, J., Li, L., Li, W., Li, S., Yu, J., Lin, J. D., Wang, C.-Y., Chinnaiyan, A. M., Lai, Z.-C. & Guan, K.-L. TEAD mediates YAP-dependent gene induction and growth control. *Genes Dev.* **22,** 1962–71 (2008).

74. Landin Malt, A., Cagliero, J., Legent, K., Silber, J., Zider, A. & Flagiello, D. Alteration of TEAD1 expression levels confers apoptotic resistance through the transcriptional up-regulation of Livin. *PLoS One* **7,** publikováno online (2012).

75. Hucl, T., Brody, J. R., Gallmeier, E., Iacobuzio-Donahue, C. A., Farrance, I. K. & Kern, S. E. High cancer-specific expression of mesothelin (MSLN) is attributable to an upstream enhancer containing a transcription enhancer factor dependent MCAT motif. *Cancer Res.* **67,** 9055–65 (2007).

76. Schütte, U., Bisht, S., Heukamp, L. C., Kebschull, M., Florin, A., Haarmann, J., Hoffmann, P., Bendas, G., Buettner, R., Brossart, P. & Feldmann, G. Hippo signaling mediates proliferation, invasiveness, and metastatic potential of clear cell renal cell carcinoma. *Transl. Oncol.* **7,** 309–21 (2014).

77. Wang, W., Xiao, Z.-D., Li, X., Aziz, K. E., Gan, B., Johnson, R. L. & Chen, J. AMPK modulates Hippo pathway activity to regulate energy homeostasis. *Nat. Cell Biol.* **17,** 490–9 (2015).

78. Vališ, K., Talacko, P., Grobárová, V., Černý, J. & Novák, P. Shikonin regulates C-MYC and GLUT1 expression through the MST1-YAP1-TEAD1 axis. *Exp. Cell Res.* **349,** 273–281 (2016).

79. Knight, J. F., Shepherd, C. J., Rizzo, S., Brewer, D., Jhavar, S., Dodson, A. R., Cooper, C. S., Eeles, R., Falconer, A., Kovacs, G., Garrett, M. D., Norman, A. R., Shipley, J. & Hudson, D. L. TEAD1 and c-Cbl are novel prostate basal cell markers that correlate with poor clinical outcome in prostate cancer. *Br. J. Cancer* **99,** 1849–58 (2008).

80. Liu, Y., Wang, G., Yang, Y., Mei, Z., Liang, Z., Cui, A., Wu, T., Liu, C.-Y. & Cui, L. Increased TEAD4 expression and nuclear localization in colorectal cancer promote epithelial–mesenchymal transition and metastasis in a YAP-independent manner. *Oncogene* **35,** 2789–2800 (2016).

81. Wang, C., Nie, Z., Zhou, Z., Zhang, H., Liu, R., Wu, J., Qin, J., Ma, Y., Chen, L., Li, S., Chen, W., Li, F., Shi, P., Wu, Y., Shen, J. & Chen, C. The interplay between TEAD4 and KLF5 promotes breast cancer partially through inhibiting the transcription of p27 Kip1. *Oncotarget* **6,** 17685–17697 (2015).

82. Zhou, Y., Huang, T., Zhang, J., Wong, C. C., Zhang, B., Dong, Y., Wu, F., Tong, J. H. M., Wu, W. K. K., Cheng, A. S. L., Yu, J., Kang, W. & To, K. F. TEAD1/4 exerts oncogenic role and is negatively regulated by miR-4269 in gastric tumorigenesis. *Oncogene* **36,** 6518–6530 (2017).

83. Xiao, J. H., Davidson, I., Matthes, H., Garnier, J. M. & Chambon, P. Cloning, expression, and transcriptional properties of the human enhancer factor TEF-1. *Cell* **65,** 551–568 (1991).

84. Lin, K. C., Park, H. W. & Guan, K. L. Regulation of the Hippo Pathway Transcription Factor TEAD. *Trends Biochem. Sci.* **42,** 862–872 (2017).

85. Vassilev, A., Kaneko, K. J., Shu, H., Zhao, Y. & DePamphilis, M. L. TEAD/TEF transcription factors utilize the activation domain of YAP65, a Src/Yes-associated protein localized in the cytoplasm. *Genes Dev.* **15,** 1229–41 (2001).

86. Mahoney, W. M., Hong, J.-H., Yaffe, M. B. & Farrance, I. K. G. The transcriptional co-activator TAZ interacts differentially with transcriptional enhancer factor-1 (TEF-1) family members. *Biochem. J.* **388,** 217–25 (2005).

87. Zhao, B., Wei, X., Li, W., Udan, R. S., Yang, Q., Kim, J., Xie, J., Ikenoue, T., Yu, J., Li, L., Zheng, P., Ye, K., Chinnaiyan, A., Halder, G., Lai, Z. & Guan, K. Inactivation of YAP oncoprotein by the Hippo pathway is involved in cell contact inhibition and tissue growth control. *Genes Dev.* **21,** 2747–61 (2007).

88. Hansen, C. G., Moroishi, T. & Guan, K. L. YAP and TAZ: A nexus for Hippo signaling and beyond. *Trends Cell Biol.* **25,** 499–513 (2015).

89. Azzolin, L., Zanconato, F., Bresolin, S., Forcato, M., Basso, G., Bicciato, S., Cordenonsi, M. & Piccolo, S. Role of TAZ as mediator of wnt signaling. *Cell* **151,** 1443–1456 (2012).

90. Park, H. W., Kim, Y. C., Yu, B., Moroishi, T., Mo, J. S., Plouffe, S. W., Meng, Z., Lin, K. C., Yu, F. X., Alexander, C. M., Wang, C. Y. & Guan, K. L. Alternative Wnt Signaling Activates YAP/TAZ. *Cell* **162,** 780–794 (2015).

91. Hiemer, S. E., Szymaniak, A. D. & Varelas, X. The transcriptional regulators TAZ and YAP direct transforming growth factor β-induced tumorigenic phenotypes in breast cancer cells. *J. Biol. Chem.* **289,** 13461–13474 (2014).

92. Mohseni, M., Sun, J., Lau, A., Curtis, S., Goldsmith, J., Fox, V. L., Wei, C., Frazier, M., Samson, O., Wong, K.-K., Kim, C. & Camargo, F. D. A genetic screen identifies an LKB1–MARK signalling axis controlling the Hippo–YAP pathway. *Nat. Cell Biol.* **16,** 108–117 (2014).

93. Jiao, S., Wang, H., Shi, Z., Dong, A., Zhang, W., Song, X., He, F., Wang, Y., Zhang, Z., Wang, W., Wang, X., Guo, T., Li, P., Zhao, Y., Ji, H., Zhang, L. & Zhou, Z. A peptide mimicking VGLL4 function acts as a YAP antagonist therapy against gastric cancer. *Cancer Cell* **25,** 166–80 (2014).

94. Jiang, S. W., Dong, M., Trujillo, M. A., Miller, L. J. & Eberhardt, N. L. DNA Binding of TEA/ATTS Domain Factors Is Regulated by Protein Kinase C Phosphorylation in Human Choriocarcinoma Cells. *J. Biol. Chem.* **276,** 23464–23470 (2001).

95. Gupta, M. P., Kogut, P. & Gupta, M. Protein kinase-A dependent phosphorylation of transcription enhancer factor-1 represses its DNA-binding activity but enhances its gene activation ability. *Nucleic Acids Res.* **28,** 3168–3177 (2000).

96. Noland, C. L., Gierke, S., Schnier, P. D., Murray, J., Sandoval, W. N., Sagolla, M., Dey, A., Hannoush, R. N., Fairbrother, W. J. & Cunningham, C. N. Palmitoylation of TEAD Transcription Factors Is Required for Their Stability and Function in Hippo Pathway Signaling. *Structure* **24,** 179–186 (2016).

97. Lin, K. C., Moroishi, T., Meng, Z., Jeong, H.-S., Plouffe, S. W., Sekido, Y., Han, J., Park, H. W. & Guan, K.-L. Regulation of Hippo pathway transcription factor TEAD by p38 MAPK-induced cytoplasmic translocation. *Nat. Cell Biol.* **19,** 996–1002 (2017).

98. Gibault, F., Sturbaut, M., Bailly, F., Melnyk, P. & Cotelle, P. Targeting Transcriptional Enhanced Associate Domains (TEADs). *J. Med. Chem.* **61,** 5057–5072 (2018).

99. Lee, D.-S., Vonrhein, C., Albarado, D., Raman, C. S. & Veeraraghavan, S. A Potential Structural Switch for Regulating DNA-Binding by TEAD Transcription Factors. *J. Mol. Biol.* 1–12 (2016). doi:10.1016/j.jmb.2016.03.008

100. Shi, Z., He, F., Chen, M., Hua, L., Wang, W., Jiao, S. & Zhou, Z. DNA-binding mechanism of the Hippo pathway transcription factor TEAD4. *Oncogene* **36,** 4362–4369 (2017).

101. Mar, J. H. & Ordahl, C. P. M-CAT binding factor, a novel trans-acting factor governing muscle-specific transcription. *Mol. Cell. Biol.* **10,** 4271–4283 (1990).

102. Larkin, S. B., Farrance, I. K. & Ordahl, C. P. Flanking sequences modulate the cell specificity of M-CAT elements. *Mol. Cell. Biol.* **16,** 3742–3755 (1996).

103. Farrance, I. K. G., Mar, J. H. & Ordahl, C. P. M-CAT binding factor is related to the SV40 enhancer binding factor, TEF-1. *J. Biol. Chem.* **267,** 17234–17240 (1992).

104. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Van Der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F. & Mathelier, A. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46,** D260–D266 (2018).

105. Carlini, L. E., Getz, M. J., Strauch, A. R. & Kelm, R. J. Cryptic MCAT enhancer regulation in fibroblasts and smooth muscle cells: Suppression of TEF-1 mediated activation by the single-stranded DNA-binding proteins, Purα, Purβ, and MSY1. *J. Biol. Chem.* **277,** 8682–8692 (2002).

106. Davidson, I., Xiao, J. H., Rosales, R., Staub, A. & Chambon, P. The HeLa cell protein TEF-1 binds specifically and cooperatively to two SV40 enhancer motifs of unrelated sequence. *Cell* **54,** 931–942 (1988).

107. Tian, W., Yu, J., Tomchick, D. R., Pan, D. & Luo, X. Structural and functional analysis of the YAP-binding domain of human TEAD2. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 7293–8 (2010).

108. Li, Z., Zhao, B., Wang, P., Chen, F., Dong, Z., Yang, H., Guan, K.-L. & Xu, Y. Structural insights into the YAP and TEAD complex. *Genes Dev.* **24,** 235–40 (2010).

109. Chen, L., Chan, S. W., Zhang, X., Walsh, M., Lim, C. J., Hong, W. & Song, H. Structural basis of YAP recognition by TEAD4 in the hippo pathway. *Genes Dev.* **24,** 290–300 (2010).

110. Kaan, H. Y. K., Chan, S. W., Tan, S. K. J., Guo, F., Lim, C. J., Hong, W. & Song, H. Crystal structure of TAZ-TEAD complex reveals a distinct interaction mode from that of YAP-TEAD complex. *Sci. Rep.* **7,** 2035 (2017).

111. Pobbati, A. V., Chan, S. W., Lee, I., Song, H. & Hong, W. Structural and functional similarity between the Vgll1-TEAD and the YAP-TEAD complexes. *Structure* **20,** 1135–40 (2012).

112. Chan, P., Han, X., Zheng, B., DeRan, M., Yu, J., Jarugumilli, G. K., Deng, H., Pan, D., Luo, X. & Wu, X. Autopalmitoylation of TEAD proteins regulates transcriptional output of the Hippo pathway. *Nat. Chem. Biol.* **12,** 282–9 (2016).

113. Mesrouze, Y., Meyerhofer, M., Bokhovchuk, F., Fontana, P., Zimmermann, C., Martin, T., Delaunay, C., Izaac, A., Kallen, J., Schmelzle, T., Erdmann, D. & Chène, P. Effect of the acylation of TEAD4 on its interaction with co-activators YAP and TAZ. *Protein Sci.* **22,** 1–50 (2017).

114. Weigel, D. & Jäckle, H. The fork head domain: A novel DNA binding motif of eukaryotic transcription factors? *Cell* **63,** 455–456 (1990).

115. Kaestner, K. H., Knöchel, W. & Martínez, D. E. Unified nomenclature for the winged helix/forkhead transcription factors. *Genes Dev.* **14,** 142–146 (2000).

116. Jiramongkol, Y. & Lam, E. W. F. FOXO transcription factor family in cancer and metastasis. *Cancer Metastasis Rev.* **39,** 681–709 (2020).

117. Benayoun, B. A., Caburet, S. & Veitia, R. A. Forkhead transcription factors: Key players in health and disease. *Trends Genet.* **27,** 224–232 (2011).

118. Borkhardt, A., Repp, R., Haas, O. A., Leis, T., Harbott, J., Kreuder, J., Hammermann, J., Henn, T. & Lampert, F. Cloning and characterization of AFX, the gene that fuses to MLL in acute leukemias with a t(X;11)(q13;q23). *Oncogene* **14,** 195–202 (1997).

119. Link, W. Introduction to FOXO biology. *Methods Mol. Biol.* **1890,** 1–9 (2019).

120. Link, W. & Fernandez-Marcos, P. J. FOXO transcription factors at the interface of metabolism and cancer. *Int. J. Cancer* **141,** 2379–2391 (2017).

121. Kops, G. J. P. L., De Ruiter, N. D., De Vries-Smits, A. M. M., Powell, D. R., Bos, J. L. & Burgering, B. M. T. Direct control of the forkhead transcription faotor AFX by protein kinase B. *Nature* **398,** 630–634 (1999).

122. Zhang, X., Tang, N., Hadden, T. J. & Rishi, A. K. Akt, FoxO and regulation of apoptosis. *Biochim. Biophys. Acta - Mol. Cell Res.* **1813,** 1978–1986 (2011).

123. Biggs, W. H., Cavenee, W. K. & Arden, K. C. Identification and characterization of members of the FKHR (FOX O) subclass of winged-helix transcription factors in the mouse. *Mamm. Genome* **12,** 416–425 (2001).

124. Hosaka, T., Biggs, W. H., Tieu, D., Boyer, A. D., Varki, N. M., Cavenee, W. K. & Arden, K. C. Disruption of forkhead transcription factor (FOXO) family members in mice reveals their functional diversification. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 2975–2980 (2004).

125. Liu, W., Li, Y. & Luo, B. Current perspective on the regulation of FOXO4 and its role in disease progression. *Cell. Mol. Life Sci.* **77,** 651–663 (2020).

126. Wang, W., Zhou, P. H. & Hu, W. Overexpression of FOXO4 induces apoptosis of clear-cell renal carcinoma cells through downregulation of Bim. *Mol. Med. Rep.* **13,** 2229–2234 (2016).

127. Shi, X., Wallis, A. M., Gerard, R. D., Voelker, K. A., Grange, R. W., DePinho, R. A., Garry, M. G. & Garry, D. J. Foxk1 promotes cell proliferation and represses myogenic differentiation by regulating Foxo4 and Mef2. *J. Cell Sci.* **125,** 5329–5337 (2012).

128. Medema, R. H., Kops, G. J. P. L., Bos, J. L. & Burgering, B. M. T. AFX-like Forkhead transcription factors mediate cell-cycle regulation by Ras and PKB through p27(kip1). *Nature* **404,** 782–787 (2000).

129. Boura, E., Rezabkova, L., Brynda, J., Obsilova, V. & Obsil, T. Structure of the human FOXO4-DBD-DNA complex at 1.9 Å resolution reveals new details of FOXO binding to the DNA. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66,** 1351–1357 (2010).

130. Furuyama, T., Nakazawa, T., Nakano, I. & Mori, N. Identification of the differential distribution patterns of mRNAs and consensus binding sequences for mouse DAF-16 homologues. *Biochem. J.* **349,** 629–634 (2000).

131. Wang, F., Marshall, C. B. & Ikura, M. Forkhead followed by disordered tail: The intrinsically disordered regions of FOXO3a. *Intrinsically Disord. Proteins* **3,** 32–36 (2015).

132. Weigelt, J., Climent, I., Dahlman-Wright, K. & Wikström, M. Letter to the Editor: 1H, 13C and 15N resonance assignments of the DNA binding domain of the human Forkhead transcription factor AFX [4]. *J. Biomol. NMR* **17,** 181–182 (2000).

133. Boura, E., Silhan, J., Herman, P., Vecer, J., Sulc, M., Teisinger, J., Obsilova, V. & Obsil, T. Both the N-terminal loop and wing W2 of the forkhead domain of transcription factor Foxo4 are important for DNA binding. *J. Biol. Chem.* **282,** 8265–8275 (2007).

134. Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364,** 412–420 (1993).

135. Dimitrova-Paternoga, L., Jagtap, P. K. A., Chen, P. C. & Hennig, J. Integrative Structural Biology of Protein-RNA Complexes. *Structure* **28,** 6–28 (2020).

136. Anderson, W. F., Ohlendorf, D. H., Takeda, Y. & Matthews, B. W. Structure of the cro repressor from bacteriophage λ and its interaction with DNA. *Nature* **290,** 754–758 (1981).

137. McKay, D. B. & Steitz, T. A. Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA. *Nature* **290,** 744–749 (1981).

138. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* **275,** 1–21 (2008).

139. Egli, M. Diffraction Techniques in Structural Biology. *Curr. Protoc. Nucleic Acid Chem.* **41,** 1–35 (2010).

140. Jones, S. Protein–RNA interactions: structural biology and computational modeling techniques. *Biophys. Rev.* **8,** 359–367 (2016).

141. Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F. & MacKay, J. P. Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS J.* **278,** 687–703 (2011).

142. Wang, G., Zhang, Z. T., Jiang, B., Zhang, X., Li, C. & Liu, M. Recent advances in protein NMR spectroscopy and their implications in protein therapeutics research. *Anal. Bioanal. Chem.* **406,** 2279–2288 (2014).

143. Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., Tran, E. E. H., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A. & Subramaniam, S. Cryo-electron microscopy - A primer for the non-microscopist. *FEBS J.* **280,** 28–45 (2013).

144. Nogales, E., Scheres, S. H. W., Division, L. S., Berkeley, L., Avenue, F. C., Campus, C. B. & Kingdom, U. Cryo-EM: A Unique Tool for the visualization. **58,** 677–689 (2016).

145. Cheng, Y. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **161,** 450–457 (2015).

146. Chua, E. Y. D., Vogirala, V. K., Inian, O., Wong, A. S. W., Nordenskiöld, L., Plitzko, J. M., Danev, R. & Sandin, S. 3.9 Å structure of the nucleosome core particle determined by phase-plate cryo-EM. *Nucleic Acids Res.* **44,** 8013–8019 (2016).

147. Plaschka, C., Hantsche, M., Dienemann, C., Burzinski, C., Plitzko, J. & Cramer, P. Transcription initiation complex structures elucidate DNA opening. *Nature* **533,** 353–358 (2016).

148. Mishyna, M., Volokh, O., Danilova, Y., Gerasimova, N., Pechnikova, E. & Sokolova, O. S. Effects of radiation damage in studies of protein-DNA complexes by cryo-EM. *Micron* **96,** 57–64 (2017).

149. Herzik, M. A., Wu, M. & Lander, G. C. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat. Commun.* **10,** 1–9 (2019).

150. Nordhoff, E. & Lehrach, H. Identification and characterization of DNA-binding proteins by mass spectrometry. *Adv. Biochem. Eng. Biotechnol.* **104,** 111–195 (2006).

151. Konermann, L., Vahidi, S. & Sowole, M. A. Mass spectrometry methods for studying structure and dynamics of biological macromolecules. *Anal. Chem.* **86,** 213–232 (2014).

152. Mihăşan, M., Wormwood, K. L., Sokolowska, I., Roy, U., Woods, A. G. & Darie, C. C. Mass Spectrometry- and Computational Structural Biology-Based Investigation of Proteins and Peptides. *Adv. Exp. Med. Biol.* **1140,** 265–287 (2019).

153. Rozbesky, D., Man, P., Kavan, D., Chmelik, J., Cerny, J., Bezouska, K. & Novak, P. Chemical cross-linking and H/D exchange for fast refinement of protein crystal structure. *Anal. Chem.* **84,** 867–870 (2012).

154. Sinz, A. Cross-Linking / Mass Spectrometry for Studying Protein Structures and Protein – Protein Interactions : Where Are We Now and Where Should We Go from Here ? Minireviews. 6390–6396 (2018). doi:10.1002/anie.201709559

155. Novák, P. & Kruppa, G. H. Intra-Molecular Cross-Linking of Acidic Residues for Protein Structure Studies. *Eur. J. Mass Spectrom.* **14,** 355–365 (2008).

156. Leavell, M. D., Novak, P., Behrens, C. R., Schoeniger, J. S. & Kruppa, G. H. Strategy for selective chemical cross-linking of tyrosine and lysine residues. *J. Am. Soc. Mass Spectrom.* **15,** 1604–1611 (2004).

157. Ihling, C., Schmidt, A., Kalkhof, S., Schulz, D. M., Stingl, C., Mechtler, K., Haack, M., Beck-Sickinger, A. G., Cooper, D. M. F. & Sinz, A. Isotope-Labeled Cross-Linkers and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for Structural Analysis of a Protein/Peptide Complex. *J. Am. Soc. Mass Spectrom.* **17,** 1100–1113 (2006).

158. Kukacka, Z., Rosulek, M., Strohalm, M., Kavan, D. & Novak, P. Mapping protein structural changes by quantitative cross-linking. *Methods* **89,** 112–120 (2015).

159. Müller, M. Q., Dreiocker, F., Ihling, C. H., Schäfer, M. & Sinz, A. Cleavable cross-linker for protein structure analysis: Reliable identification of cross-linking products by tandem MS. *Anal. Chem.* **82,** 6958–6968 (2010).

160. Kang, S., Mou, L., Lanman, J., Velu, S., Brouillette, W. J. & Prevelige, P. E. Synthesis of biotin-tagged chemical cross-linkers and their applications for mass spectrometry. *Rapid Commun. Mass Spectrom.* **23,** 1719–1726 (2009).

161. Steen, H., Petersen, J., Mann, M. & Jensen, O. N. Mass spectrometric analysis of a UV-cross-linked protein-DNA complex: Tryptophans 54 and 88 of E. coli SSB cross-link to DNA . *Protein Sci.* **10,** 1989–2001 (2001).

162. Sharma, K., Hrle, A., Kramer, K., Sachsenberg, T., Staals, R. H. J., Randau, L., Marchfelder, A., van der Oost, J., Kohlbacher, O., Conti, E. & Urlaub, H. Analysis of protein-RNA interactions in CRISPR proteins and effector complexes by UV-induced cross-linking and mass spectrometry. *Methods* **89,** 138–148 (2015).

163. Yu, E. T., Zhang, Q. & Fabris, D. Untying the FIV frameshifting pseudoknot structure by MS3D. *J. Mol. Biol.* **345,** 69–80 (2005).

164. Scalabrin, M., Dixit, S. M., Makshood, M. M., Krzemien, C. E. & Fabris, D. Bifunctional cross-linking approaches for mass spectrometry-based investigation of nucleic acids and protein-nucleic acid assemblies. *Methods* **144,** 64–78 (2018).

165. Konermann, L., Pan, J. & Liu, Y. H. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* **40,** 1224–1234 (2011).

166. Katta V and Chait B. Conformational Changes in Proteins Probed by Hydrogen-exchange

Electrospray-ionization. *Rapid Commun. Mass Spectrom.* **5,** 214–217 (1991).

167. Oganesyan, I., Lento, C. & Wilson, D. J. Contemporary hydrogen deuterium exchange mass spectrometry. *Methods* **144,** 27–42 (2018).

168. Masson, G. R., Burke, J. E., Ahn, N. G., Anand, G. S., Borchers, C., Brier, S., Bou-Assaf, G. M., Engen, J. R., Englander, S. W., Faber, J., Garlish, R., Griffin, P. R., Gross, M. L., Guttman, M., Hamuro, Y., Heck, A. J. R., Houde, D., Iacob, R. E., Jørgensen, T. J. D., Kaltashov, I. A., Klinman, J. P., Konermann, L., Man, P., Mayne, L., Pascal, B. D., Reichmann, D., Skehel, M., Snijder, J., Strutzenberg, T. S., Underbakke, E. S., Wagner, C., Wales, T. E., Walters, B. T., Weis, D. D., Wilson, D. J., Wintrode, P. L., Zhang, Z., Zheng, J., Schriemer, D. C. & Rand, K. D. Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat. Methods* **16,** 595–602 (2019).

169. Yang, M., Hoeppner, M., Rey, M., Kadek, A., Man, P. & Schriemer, D. C. Recombinant Nepenthesin II for Hydrogen/Deuterium Exchange Mass Spectrometry. *Anal. Chem.* **87,** 6681–6687 (2015).

170. Kadek, A., Mrazek, H., Halada, P., Rey, M., Schriemer, D. C. & Man, P. Aspartic protease nepenthesin-1 as a tool for digestion in hydrogen/deuterium exchange mass spectrometry. *Anal. Chem.* **86,** 4287–4294 (2014).

171. Rey, M., Man, P., Brandolin, G., Forest, E. & Pelosi, L. Recombinant immobilized rhizopuspepsin as a new tool for protein digestion in hydrogen/deuterium exchange mass spectrometry. *Rapid Commun. Mass Spectrom.* **23,** 3431–3438 (2009).

172. Sorokin, V., Gladchenko, G. & Valeev, V. DNA protonation at low ionic strength of solution. *Die Makromol. Chemie* **187,** 1053–1063 (1986).

173. Sperry, J. B., Shi, X., Rempel, D. L., Nishimura, Y., Akashi, S. & Gross, M. L. A mass spectrometric approach to the study of DNA-binding proteins: Interaction of human TRF2 with telomeric DNA. *Biochemistry* **47,** 1797–1807 (2008).

174. Poliakov, A., Jardine, P. & Prevelige, P. E. Hydrogen/deuterium exchange on protein solutions containing nucleic acids: utility of protamine sulfate. *Rapid Commun. Mass Spectrom.* **22,** 2423–2428 (2008).

175. Roberts, V. A., Pique, M. E., Hsu, S., Li, S., Slupphaug, G., Rambo, R. P., Jamison, J. W., Liu, T., Lee, J. H., Tainer, J. A., Ten Eyck, L. F. & Woods, V. L. Combining HD exchange mass spectroscopy and computational docking reveals extended DNA-binding surface on uracil-DNA glycosylase. *Nucleic Acids Res.* **40,** 6070–6081 (2012).

176. Graham, B. W., Tao, Y., Dodge, K. L., Thaxton, C. T., Olaso, D., Young, N. L., Marshall, A. G. & Trakselis, M. A. DNA Interactions Probed by Hydrogen-Deuterium Exchange (HDX) Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Confirm External Binding Sites on the Minichromosomal Maintenance (MCM) Helicase. *J. Biol. Chem.* **291,** 12467–12480 (2016).

177. Sperry, J. B., Wilcox, J. M. & Gross, M. L. Strong Anion Exchange for Studying Protein-DNA Interactions by H/D Exchange Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **19,** 887–890 (2008).

178. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–42 (2000).

179. Fabris, D. & Yu, E. T. Elucidating the higher-order structure of biopolymers by structural probing and mass spectrometry: MS3D. *J. Mass Spectrom.* **45,** 841–860 (2010).

180. Mayne, L., Kan, Z.-Y., Sevugan Chetty, P., Ricciuti, A., Walters, B. T. & Englander, S. W.

Many Overlapping Peptides for Protein Hydrogen Exchange Experiments by the Fragment Separation-Mass Spectrometry Method. *J. Am. Soc. Mass Spectrom.* **22,** 1898–1905 (2011).

181. Althaus, E., Canzar, S., Ehrler, C., Emmett, M. R., Karrenbauer, A., Marshall, A. G., Meyer-Bäse, A., Tipton, J. D. & Zhang, H. Computing H/D-Exchange rates of single residues from data of proteolytic fragments. *BMC Bioinformatics* **11,** 424 (2010).

182. Kan, Z. Y., Walters, B. T., Mayne, L. & Englander, S. W. Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 16438–16443 (2013).

183. Cravello, L., Lascoux, D. & Forest, E. Use of different proteases working in acidic conditions to improve sequence coverage and resolution in hydrogen/deuterium exchange of large proteins. *Rapid Commun. Mass Spectrom.* **17,** 2387–2393 (2003).

184. Kadek, A., Mrazek, H., Halada, P., Rey, M., Schriemer, D. C. & Man, P. Aspartic protease nepenthesin-1 as a tool for digestion in hydrogen/deuterium exchange mass spectrometry. *Anal. Chem.* **86,** 4287–4294 (2014).

185. Zhang, X., Chien, E. Y. T., Chalmers, M. J., Pascal, B. D., Gatchalian, J., Stevens, R. C. & Griffin, P. R. Dynamics of the β 2 -Adrenergic G-Protein Coupled Receptor Revealed by Hydrogen−Deuterium Exchange. *Anal. Chem.* **82,** 1100–1108 (2010).

186. Li, S., Lee, S. Y. & Chung, K. Y. in *Methods Enzymol.* **557,** 261–278 (Elsevier Inc., 2015).

187. Ahn, J., Jung, M. C., Wyndham, K., Yu, Y. Q. & Engen, J. R. Pepsin Immobilized on High-Strength Hybrid Particles for Continuous Flow Online Digestion at 10 000 psi. *Anal. Chem.* **84,** 7256–7262 (2012).

188. Ridinger-Saison, M., Boeva, V., Rimmelé, P., Kulakovskiy, I., Gallais, I., Levavasseur, B., Paccard, C., Legoix-Né, P., Morlé, F., Nicolas, A., Hupé, P., Barillot, E., Moreau-Gachelin, F. & Guillouf, C. Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. *Nucleic Acids Res.* **40,** 8927–8941 (2012).

189. Barylyuk, K., Gülbakan, B., Xie, X. & Zenobi, R. DNA Oligonucleotides: A Model System with Tunable Binding Strength to Study Monomer–Dimer Equilibria with Electrospray Ionization-Mass Spectrometry. *Anal. Chem.* **85,** 11902–11912 (2013).

Attached Publications:

Publication I

Möller, I. R., Slivacka, M., Hausner, J., Nielsen, A. K., Pospíšilová, E., Merkle, P. S., **Lišková, R**., Polák, M., Loland, C. J., Kádek, A., Man, P., & Rand, K. D.

Improving the Sequence Coverage of Integral Membrane Proteins during Hydrogen/Deuterium Exchange Mass Spectrometry Experiments

Analytical Chemistry, *91*(17), 10970–10978 (2019)

My contribution: *recombinant production and immobilization of proteases, testing of the cleavage performance of the prepared columns under HDX-MS conditions*

analytical chemistry

Article

Cite This: *Anal. Chem.* 2019, 91, 10970−10978

pubs.acs.org/ac

Improving the Sequence Coverage of Integral Membrane Proteins during Hydrogen/Deuterium Exchange Mass Spectrometry Experiments

Ingvar R. Möller,[†] Marika Slivacka,[†] Jiří Hausner,[‡,⊥] Anne Kathrine Nielsen,[§] Eliška Pospíšilová,[‡,⊥] Patrick S. Merkle,[†] Růžena Lišková,[‡,⊥] Marek Polák,[‡,⊥] Claus J. Loland,[§] Alan Kádek,[‡,⊥] Petr Man,[‡,⊥] and Kasper D. Rand*[,†]

[†]Department of Pharmacy, University of Copenhagen, Universitetsparken 2, Copenhagen E DK-2100, Denmark

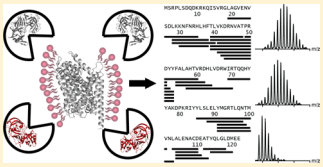
[‡]BioCeV - Institute of Microbiology of the CAS, Prumyslova 595, CZ-252 50 Vestec, Czech Republic

[⊥]Faculty of Science, Charles University, Hlavova 8, CZ-128 20 Prague, Czech Republic

[§]Laboratory for Membrane Protein Dynamics, Department of Neuroscience, University of Copenhagen, Blegdamsvej 3, Copenhagen N DK-2200, Denmark

Supporting Information

ABSTRACT: Insight into the structure−function relationship of membrane proteins is important to understand basic cell function and inform drug development, as these are common targets for drugs. Hydrogen/deuterium exchange mass spectrometry (HDX-MS) is an established technique for the study of protein conformational dynamics and has shown compatibility with membrane proteins. However, the digestion and mass analysis of peptides from membrane proteins can be challenging, severely limiting the HDX-MS experiment. Here we compare the digestion of four integral membrane proteins—Cl^-/H^+ exchange transporter (ClC-ec1), leucine transporter (LeuT), dopamine transporter (DAT), and serotonin transporter (SERT)—by the use of porcine pepsin and three alternative aspartic proteases either in-solution or immobilized on-column in an optimized HDX-MS-compatible workflow. Pepsin was the most favorable for the digestion of ClC-ec1 and LeuT, providing coverage of 82.2 and 33.2% of the respective protein sequence; however, the alternative proteases surpassed pepsin for the digestion of DAT and SERT. By also screening quench solution additives, we observe that the denaturant urea was beneficial, resulting in improved sequence coverage of all membrane proteins, in contrast to guanidine hydrochloride. Furthermore, significant improvements in sequence coverage were achieved by tailoring the chromatography to handle hydrophobic peptides. Overall, we demonstrate that the susceptibility of membrane proteins to proteolytic digestion during HDX-MS is highly protein-specific. Our results highlight the importance of having multiple proteases and different quench buffer additives in the HDX-MS toolbox and the need to carefully screen a range of digestion conditions to successfully optimize the HDX-MS analysis of integral membrane proteins.



Hydrogen/deuterium exchange mass spectrometry (HDX-MS) is a well-established tool for the study of protein conformations and dynamics.[1,2] In a classical continuous labeling HDX-MS experiment, a target protein is diluted into deuterated solution, where the protein exchanges its heteroatom-bound hydrogens for deuterium at a rate primarily determined by the status and strength of the hydrogen bonds in the protein. Because protein higher order structure is defined by such networks of hydrogen bonds, the method is highly sensitive to changes in the conformational dynamics between distinct protein states. Most commonly, upon diluting the target protein in deuterated buffer, the exchange reaction is sampled at various time intervals and quenched.[2] A fundamental aspect for the use of HDX-MS is the ability to efficiently quench the HDX in a protein by lowering the pH and temperature to 2.5 and 0 °C, respectively.[3] The quenched protein sample is then subjected to rapid digestion, and the resulting peptides are separated by a high-performance liquid chromatography/ultra-performance liquid chromatography (HPLC/UPLC) system and mass-analyzed by an electrospray ionization (ESI) mass spectrometer. To achieve the digestion at low pH, an acid-stable protease must be employed. The most regularly used protease in HDX-MS is an aspartic protease from hog stomach, porcine pepsin (EC 3.4.23.1). Although pepsin is a protease with broad specificity and produces, in most cases, a large number of peptides, the digestion of some proteins by this protease may sometimes be suboptimal.[4−7] Therefore, other aspartic proteases, such as rhizopuspepsin (Rpn, also known as protease type XVIII), aspergillopepsin (also known as protease type XIII), nepenthesin I (Nep I), and nepenthesin II (Nep II), have been applied in HDX-MS experimental setups to improve the

Received: February 22, 2019
Accepted: June 20, 2019
Published: June 20, 2019

DOI: 10.1021/acs.analchem.9b00973
Anal. Chem. 2019, 91, 10970−10978

digestion of proteins where pepsin digestion results in low sequence coverage and spatial resolution.[4,8−11]

The sensitivity and protein size/buffer system tolerance of HDX-MS inherently lend themselves to tackling complex protein systems, even integral membrane proteins, not routinely accessible by other methods of structural analysis.[12] Despite the fact that membrane proteins represent around two-thirds of potential drug targets, they are underrepresented in the Protein Data Bank (PDB).[13−15] The main challenges with regards to the study of membrane proteins are their inherent instabilities in solution, low production yields, and the common need for detergents to sufficiently stabilize their fold for in vitro studies. The short sample preparation time needed for HDX-MS makes it applicable to relatively unstable membrane proteins, and the use of highly sensitive mass spectrometers reduces the need for large amounts of protein. The detrimental effect of detergents (and other nonvolatile buffer components) on MS detection can be minimized by optimizing sample preparation and chromatographic separations.[16−18] However, the generation of peptides covering the entire protein sequence is commonly very problematic when applying HDX-MS to large integral membrane proteins. The efficient digestion of these during HDX-MS requires unhindered access by the protease to the membrane protein, which can be hampered by the presence of a detergent micelle surrounding transmembrane (TM) segments. Furthermore, peptides exclusively comprising the TM regions of a membrane protein may be prone to precipitation under aqueous conditions. Additionally, if retained in solution, such peptides can aggregate on chromatographic columns or otherwise bind irreversibly to the hydrophobic stationary phase during the LC separation step of the HDX-MS workflow. Finally, pepsin favors cleavage at most of the hydrophobic residues that are numerous in TM segments, in particular, Phe, Met, and Leu, so the digestion of these protein regions with pepsin may occasionally lead to very short peptides that are not suitable for HDX or are not even retained by the reversed-phase precolumn. Other acid-stable proteases, however, exhibit different cleavage preferences potentially generating TM-spanning peptides with physicochemical properties better suited for the analysis. To our knowledge, the use of these alternative aspartic proteases during the HDX-MS analysis of integral membrane proteins has not been investigated in detail so far.

In this study, we attempt to systematically identify factors that determine the sequence coverage during HDX-MS of three integral membrane proteins from the family of neurotransmitter:sodium symporters (NSSs) with similar functions, topologies, and numbers of TM helices: leucine transporter (LeuT, 59.3 kDa) from *Aquifex aeolicus*, *Drosophila* dopamine transporter (DAT, 72.0 kDa), and human serotonin transporter (SERT, 72.6 kDa). Additionally, we included the Cl^-/H^+ exchange transporter from *E. coli*, also known as chloride channel (ClC-ec1, 50.3 kDa). All of the transporters (Figure 1) comprise more than 10 TM helices, with relatively short extracellular and cytoplasmic loops. We first show that the sequence coverage for HDX-MS can be significantly improved by optimizing quench conditions, specifically with the use of suitable denaturizing quench buffer additives. Furthermore, the choice of chromatographic stationary phase and method used for the peptide separation can be tuned to improve the signal/noise (s/n) ratio of peptides from challenging regions (e.g., due to high hydrophobicity).

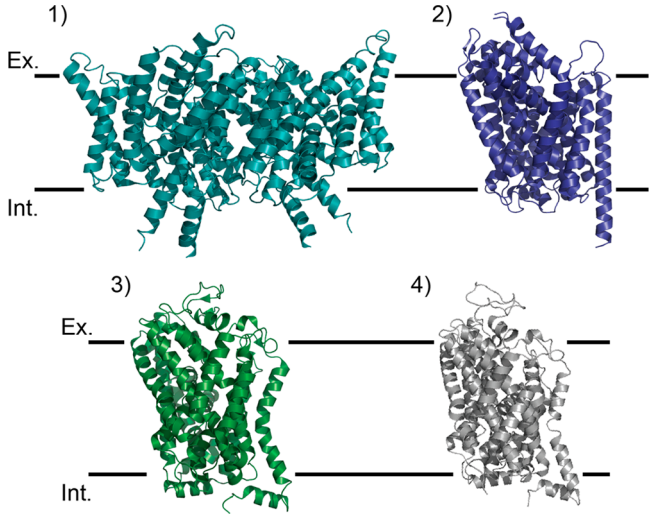


Figure 1. Reported crystal structures of (1) dimeric Cl^-/H^+ exchange transporter (deep teal, PDB: 1KPK), (2) leucine transporter (blue, PDB: 2A65), (3) dopamine transporter (forest green, PDB: 4M48), and (4) serotonin transporter (silver, PDB: 5I75). Black lines show the approximate positions of the extracellular (Ex.) and intracellular (Int.) boundaries of the cell membrane.

Following these optimizations, we investigate the protein digestion step itself and assess the performance of pepsin and three alternative acid-stable aspartic proteases (rhizopuspepsin and nepenthesins I and II) for digestion under HDX-compatible conditions of the four membrane proteins as well as phosphorylase B (Phos B, 97.2 kDa), a commonly used soluble reference protein of considerable size. Our results show that the sequence coverage of the four integral membrane proteins can be further enhanced through the use of other proteases, either in solution or immobilized. Interestingly, the optimal choice of protease is highly protein-specific, even for such closely related proteins, underlining a complex interplay between optimal protease, the sequence of the target membrane protein, and its detergent-solubilized structure.

EXPERIMENTAL PROCEDURES

Proteases. Porcine pepsin-A (no. P7012) in powder form was bought from Sigma-Aldrich. Immobilized porcine pepsin on agarose beads was bought from Thermo Scientific (no. 20343). Recombinant rhizopuspepsin, nepenthesin I, and nepenthesin II were produced as previously described.[9−11] Protease immobilization on POROS-20 AL resin was done as described by Wang and coworkers in 2002.[19]

Target Proteins. Recombinant LeuT and ClC-ec1 were produced similarly to what was described by Yamashita et al. (2005)[20] and Maduke (1999),[21] respectively, and DAT and SERT were produced similarly to what was described by Goehring et al. (2014).[22] In short, LeuT and ClC-ec1 were expressed in *E. coli*, whereas DAT and SERT were expressed in Expi293f cells with recombinantly attached polyhistidine tags. Following cell lysis and membrane preparation, the proteins were purified using single-step nickel affinity chromatography. ClC-ec1 was further subjected to limited LysC digestion for HisTag removal, and the protein was finally purified by size exclusion chromatography. Phos B from rabbit muscle was bought from Sigma-Aldrich (no. P6635).

Digestion Preparations. Twenty-five pmol of the target proteins was diluted with their respective buffer: Phos B, 10

DOI: 10.1021/acs.analchem.9b00973
Anal. Chem. 2019, 91, 10970−10978

mM sodium phosphate (pH 7.0); ClC-ec1, 30 mM sodium citrate-phosphate (pH 7.5), 300 mM KCl, 3.75 mM *N*-decyl-*β*-D-maltoside (DM); LeuT, 20 mM Tris-Cl (pH 8.0), 200 mM KCl, 1 mM *N*-dodecyl-*β*-D-maltoside (DDM); DAT, 20 mM Tris-Cl (pH 8.0), 300 mM NaCl, 14 mM cholesteryl hemisuccinate, 14 μM lipid (POPC/POPG/POPE 1:1:1), 5% glycerol, 1 mM DDM; SERT, 20 mM Tris-Cl (pH 8.0), 300 mM NaCl, 25 mM cholesteryl hemisuccinate (CHS), 24 μM lipid (POPC/POPG/POPE 1:1:1), 5% glycerol, 1 mM DDM. To simulate quench conditions for HDX-MS, the samples were diluted 1:1 with 300 mM potassium phosphate (pH 2.5) quench buffer without additives (Phos B) or with 6 M urea (ClC-ec1, LeuT), 6 M urea and 0.6 M TCEP (DAT), and 2 M urea (SERT), which had been, respectively, identified to be best suited for the on-column digestion of each protein. In-solution digestions were started by the addition of protease in 1:1 ratio to protein (w/w) and incubated for 2 min on ice. The digestion was stopped by freezing the sample at −80 °C. Samples for on-column digestions with immobilized proteases were prepared in the same manner, with the exception of the addition of protease.

Control experiments were performed to evaluate the effect of the sample and quench buffers on the proteases (immobilized pepsin, Rpn, and Nep I; Nep II in solution). Here Phos B was diluted in phosphate buffered saline (PBS, pH 7.4) with or without lipids (24 μM, POPC/POPE/POPG in 1:1:1 ratio; also contained 0.5 mM CHS, 0.05% DDM, 5% glycerol, 300 mM KCl, 6 mM Tris) and mixed 1:1 with quench buffers. The digestion of Phos B in PBS was tested following mixing with potassium phosphate (pH 2.5) with or without 6 M urea, 2 M urea, or 6 M urea and 0.6 M TCEP and in PBS with lipids following mixing with potassium phosphate (pH 2.5). The digestions were performed as described for the integral membrane proteins.

LC-MS. Samples were quickly thawed and injected onto a NanoAcquity UPLC system with an HDX manager from Waters, which allows the chromatographic separation to be performed at 0 °C. The peptides were trapped on a Vanguard column (130 Å, 1.7 μm, 2.1 mm × 5 mm; Waters) and desalted for 4 min with solvent A (0.23% formic acid (pH 2.5)) at 150 μL/min flow rate. The peptides were subsequently separated using an Acquity UPLC column (130 Å, 1.7 μm, 1 mm × 100 mm; Waters) applying a 10 min gradient with a rising concentration of solvent B (0.23% formic acid in acetonitrile) for elution into the mass spectrometer. For the separation of peptides generated from DAT, LeuT, and Phos B, a gradient from 8 to 50% of solvent B was applied using BEH C18 columns, whereas for the separation of peptides from ClC-ec1 and SERT, BEH C8 columns were used with a 10 min gradient from 8 to 30% solvent B. This was chosen to avoid the elution of lipids and detergent from the chromatographic columns during the analysis. The setup was identical for in-solution and on-column digestions, with the exception of a protease column (dimensions: 2 mm ID × 2 cm) being fitted before the trap column in a separate chamber kept at 20 °C. Pressure during the online digestion step was ∼2000 psi. Care was also taken not to cause a false increase in the sequence coverage due to carry-over, and the UPLC columns were rigorously washed using a repetition of quick gradients ranging from 8 to 92% solvent B. To facilitate the removal of peptides from the protease columns, 1 M urea in 300 mM phosphate (pH 2.5) was injected for the wash methods. The peptides were analyzed using a Synapt G2-Si (hybrid Q-TOF) mass spectrometer from Waters in positive ion mode. The peptides were fragmented by collision-induced dissociation (CID) with argon as the collision gas, and the data were collected with the mass spectrometer set in data-independent acquisition (DIA, MS^E). To obtain a meaningful comparison, all experiments on each target protein were performed within the same day and at least in duplicate.

Data Processing. Peptides from MS/MS runs were identified using ProteinLynx Global Server (PLGS), version 3.0 (Waters). Peptides identified by PLGS were filtered with DynamX, version 3.0 (Waters), which was set to consider only peptides having a minimum of 0.2 fragmentations per amino acid in the peptide and 10 ppm maximum allowed error in the precursor mass. Peptides had to be identified in two out of two MS/MS files when comparing proteases. When comparing C8 and C18 columns, DAT peptides had to be identified in two out of three MS/MS files for the comparison between C8 and C18 columns, whereas peptides from SERT were identified using longer chromatographic gradients, and the data were acquired in repeated data-dependent acquisition (DDA) manner and with the use of exclusion lists for the identification of higher number of peptides. Peptides identified by DDA had to be identified only once due to the higher reliability of the method. Peptides had to fulfill the criteria of a PLGS ladder score higher than 1.0, a mass error of the precursor ion lower than 10 ppm, and a manual evaluation of their fragmentation. The quality of the peptide signal in the mass spectra was manually evaluated (s/n values typically >10, although in a few cases s/n values between 3 and 10 were tolerated if the given peptide signal was clearly separated from other ions of similar m/z) for all proteins in DynamX 3.0.

RESULTS AND DISCUSSION

Optimizing Quench Conditions for HDX-MS Analysis of the Four Related Model Membrane Proteins. To allow a quantitative comparison of the ability of each protease to generate peptides suitable for HDX-MS analysis, we first identified quench conditions for each membrane protein, which yielded optimal conditions for digestion using the conventional protease pepsin.

The systematic optimization of quench conditions, which utilized immobilized pepsin due to its common use, showed that the addition of the denaturant urea to the quench buffer resulted in the improved sequence coverage of all of the membrane proteins. The inclusion of 3 M urea in the final quench conditions for ClC-ec1 and LeuT resulted in an 8.4 and 11.3% increase in sequence coverage, respectively, compared with a quench buffer without any additives. Interestingly, using Gnd-HCl as a denaturant resulted in at least an 8% decrease in the sequence coverage of LeuT when compared with quench conditions without any additives. The screening of quench conditions for DAT and SERT showed a similar trend, which could indicate that the mechanism of denaturation by urea is more favorable compared with Gnd-HCl (probably due to guanidine's charged nature), at least toward this particular family of membrane proteins. Even though in our hands urea seems to be the most useful for the denaturation of membrane proteins (possibly owing to its uncharged nature), Gnd as well as tuning the detergent concentration[23] should not be completely disregarded for other membrane protein systems.

Comparing Protease Performance for HDX-MS of Integral Membrane Proteins. *Phos B.* The applicability of

DOI: 10.1021/acs.analchem.9b00973
Anal. Chem. 2019, 91, 10970−10978

all digestion setups was verified by digesting Phos B. The digestions resulted in sequence coverages around 80% for most of the tested setups. All digestions resulted in more than 150 peptides being identified and were considered sufficient to prove that the digestion setups were applicable, at least for the soluble reference protein. The alternative proteases either performed comparably to the standard immobilized pepsin or outperformed it (Table 1, Figures S-1 and S-2).

In a series of separate control experiments (Table S-2), we also performed a comparison of the performance of each protease when digesting PhosB in the presence of additives present in the optimized quench conditions of the membrane proteins studied here (including DDM detergent, lipids, CHS, urea, and TCEP). The results show that the presence of lipids (POPC, POPE, POPG), CHS, and detergent (DDM) at the concentrations used here does not impact the performance of any of the proteases. In contrast, high concentrations of urea (3 M) in combination with TCEP (0.3 M) slightly reduced the performance of Nep I (lower sequence coverage, higher average peptide length), even more so for Nep II. This is in agreement with previously published results[8,9,11] that showed that nepenthesins in solution are susceptible to reduction but that this effect is largely diminished upon protease immobilization and its use in a flow-through setup (e.g., protease columns). Not surprisingly, the most affected was Nep II, which was used in solution. On the contrary, Rpn digestion lead to even better results under highly denaturing and reducing conditions. Rpn used under nondenaturing conditions led to overdigested Phos B, as can be deduced from the digestion metrics. When the protease activity is slowed-down by harsh conditions, larger peptides are created and higher sequence coverage is obtained. However, we can conclude that except for solution digestion by Nep II, the other proteases retain high activity, even under strongly denaturing conditions. Thus apart from the quench conditions used for dDAT (3 M urea and 0.3 M TCEP), none of the other sample/quench conditions used for the membrane proteins tested here had a detrimental impact on the performance of the proteases. We note that the digestion experiments in Table S-1 and the control experiments in Table S-2 were performed separately and under somewhat different conditions. In particular, the immobilized Rpn column used in the control experiments had substantially higher activity (e.g., the higher activity of Rpn, leading to overdigestion of Phos B and a reduction in sequence coverage), and the sample buffer of Phos B consisted of PBS rather than 10 mM phosphate. Thus no direct comparison between the two data sets (i.e., Tables S-1 and S-2) should be performed.

ClC-ec1. The digestion of ClC-ec1 resulted in the highest sequence coverage of all of the membrane proteins. Digestion with immobilized pepsin and immobilized Rpn resulted in high sequence coverage (82.2 and 79.1%, respectively). All other digestions resulted in <30% peptide identifications compared with immobilized Rpn and sequence coverages below 50% (Table 1, Figures S-3 and S-4).

LeuT. The digestion of LeuT with immobilized pepsin resulted in the highest sequence coverage (33.2%) for this transporter. The digestion of the transporter with immobilized Rpn and immobilized Nep I resulted in less than half of the peptide identifications compared with digestion with immobilized pepsin. However, the achieved sequence coverages by Rpn and Nep I were only 5.5 and 12.1% lower, respectively, than the sequence coverage from immobilized pepsin, but the

Table 1. Overview of the Digestion Results Obtained by the Different Proteases[a]

	Phos B			ClC-ec1			LeuT			DAT			SERT		
	sequence coverage (%)	redundancy	no. peptides	sequence coverage (%)	redundancy	no. peptides	sequence coverage (%)	redundancy	no. peptides	sequence coverage (%)	redundancy	no. peptides	sequence coverage (%)	redundancy	no. peptides
iP	79.9	3.3	178	82.2	3.8	124	33.2	2.5	39	29.3	1.8	28	38.1	1.2	22
iRpn	67.9	2.9	164	79.1	3.3	111	27.7	1.6	19	38.9	2.0	45	58.3	1.9	61
iNep I	79.0	4.5	260	48.6	1.7	32	21.1	1.2	11	–	–	–	50.9	1.8	46
sP	77.2	3.0	150	31.9	2.1	22	9.8	1.5	5	13.3	2.3	14	43.5	1.5	31
sRpn	79.2	3.8	214	13.7	1.5	8	3.0	1.0	<5	<5	N/A	<5	39.4	1.4	26
sNep I	61.0	3.9	150	7.8	1.3	<5	3.0	1.0	<5	–	–	–	12.0	1.3	7
sNep II	84.4	4.5	269	27.1	3.5	30	9.6	2.0	6	<5	N/A	<5	40.0	2.1	37

[a]Proteases used for the digestion: P, porcine pepsin; Rpn, rhizopuspepsin; Nep I, nepenthesin I; Nep II, nepenthesin II; i, digestion by protease immobilized on column; s, in-solution digestion; Phos B, phosporylase B; ClC-ec1, Cl^-/H^+ exchange transporter; LeuT, leucine transporter; DAT, dopamine transporter, SERT, serotonin transporter. Digestion of DAT by nepenthesin I was not performed in solution due to the sensitivity of this enzyme to reduction by TCEP. Because Nep I performed poorly in the immobilized form on the other proteins, this condition was not used during the analysis of DAT. The highest achieved sequence coverage is underlined. Note, that the sequence coverages reported here are the result of a fixed non-exhaustive number of MS/MS acquisitions per condition (i.e. two). Redundancy: Description of how often an amino acid is covered by peptides calculated by DynamX version 3.0 for residues making up the sequence coverage. No. peptides: identified peptides consisting of between 4 and 20 amino acids. The peptides had a mass error below 10 ppm with at least 0.2 identified fragments per amino acid in two out of two MS/MS files. Additional details can be found in Table S-1.

DOI: 10.1021/acs.analchem.9b00973
Anal. Chem. 2019, 91, 10970−10978

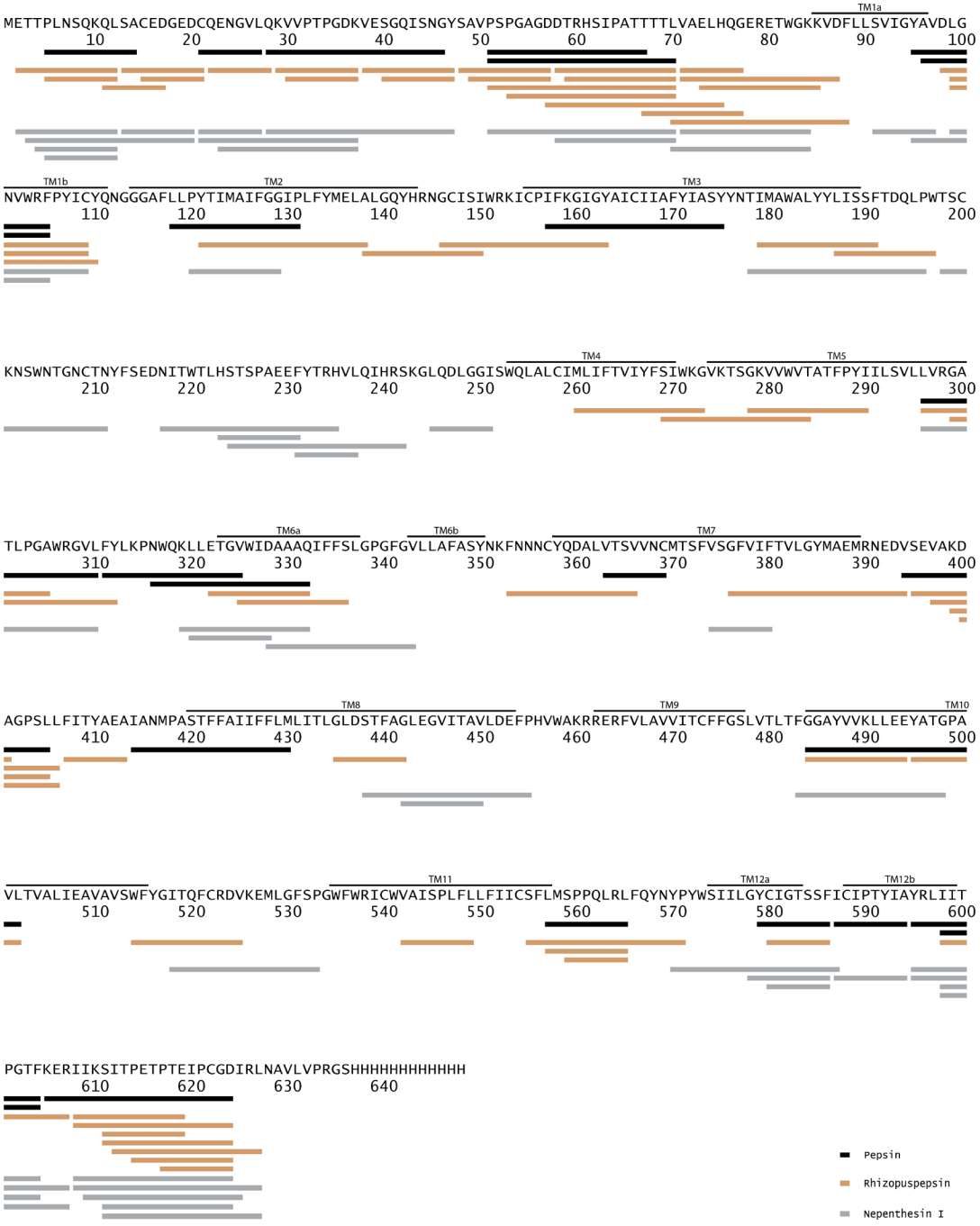


Figure 2. HDX-MS sequence coverage of SERT following digestion by different immobilized proteases. The digestion of SERT resulted in 22 identified peptides covering 38.1% of the sequence following digestion by immobilized pepsin (black), 61 identified peptides covering 58.3% of the sequence following digestion by immobilized rhizopuspepsin (light brown), and 46 identified peptides covering 50.9% of the sequence following digestion by immobilized nepenthesin I (gray).

amino acid redundancy was also impacted (Table 1, Figures S-5 and S-6).

DAT. The digestion of DAT with immobilized Rpn resulted in 38.9% sequence coverage, which was considerably higher than the coverage obtained by immobilized pepsin (29.3%). The in-solution digestion of DAT by pepsin resulted in only 14 identified peptides, covering 13.3% of the sequence, whereas fewer than five peptides were identified when digested by Rpn and Nep II (Table 1, Figures S-7 and S-8).

SERT. The digestion of SERT with immobilized Rpn resulted in around a three-fold increase in peptide identifications compared with immobilized pepsin. The peptides covered 58.3% of the sequence, which was 7.4% higher than when SERT was digested with immobilized Nep I, where the second highest number of peptides was identified (Figure 2). Comparing the in-solution digestions of the four membrane proteins, digestions of SERT resulted in the highest number of identified peptides and the highest sequence coverage (Figure S-9); however, in-solution digestions of SERT did not improve the sequence coverage when compared with digestion with immobilized proteases, with the exception of immobilized pepsin. Digestion with pepsin, Rpn, and Nep II

DOI: 10.1021/acs.analchem.9b00973
Anal. Chem. 2019, 91, 10970−10978

in solution resulted in an increased number of identified SERT peptides, improving the sequence coverage compared with immobilized pepsin (Table 1).

The digestion of SERT is shown in Figure 2 as an example of the on-column digestion of one of the four tested membrane proteins. Other sequence coverage maps can be found in the Supporting Information.

Digestion Using Free versus Immobilized Protease. The accessibility of a protease to substrate sites in a membrane protein may be impacted sterically by the presence of a detergent or lipid micelles. In turn, this could result in different activities of the protease when it is free in solution or immobilized on a stationary phase. Additionally, the effect of concentration may be vastly different between the two digestion formats, as immobilized protease, packed into a column, offers much higher local concentration than when it is free in solution. When comparing the performances of all proteases on Phos B, we observed only minor differences in sequence coverage between the two digestion formats (Table 1, Figures S-1 and S-2). With the exception of digestions with immobilized Rpn and Nep I in solution, which resulted in 67.9 and 61.0% sequence coverages, respectively, all digestion setups resulted in ∼80% sequence coverage of Phos B.

Interestingly, a much larger variation was seen in the digestion of the membrane proteins, primarily when comparing in-solution digestions to digestions performed with immobilized proteases (Table 1). Digestion with immobilized proteases outperformed in-solution digestions on almost all occasions. For ClC-ec1, even the most promising in-solution protease, porcine pepsin, resulted in 16.7% lower sequence coverage than the digestion by immobilized Nep I, which resulted in the lowest sequence coverage from digesting ClC-ec1 with immobilized protease (Table 1). Although lower than sequence coverage from digestion by immobilized proteases, the in-solution digestion by pepsin and Nep II showed promise. Following in-solution digestion by pepsin, we identified 22 peptides covering 31.9% of the protein sequence, whereas 30 peptides were identified covering 27.1% following digestion with Nep II in solution. The success of online digestion can perhaps also be ascribed to elevated pressure (∼2000 psi) and temperature (20 °C), which can destabilize the detergent micelles (Figures S-3 and S-4) and render the TM parts of the protein accessible for protease, a phenomenon similar to the observations previously described where increased pressure was shown to enhance digestion efficiency.[24]

The evaluation of in-solution digestions of the three structurally related transporters (LeuT, DAT, and SERT) further highlights the importance of screening digestion conditions, even for proteins within the same protein families. The in-solution digestions of the membrane proteins were in large part unsuccessful. Only the in-solution digestions of SERT resulted in a higher number of identified peptides and better sequence coverage when compared with the digestion by immobilized pepsin. The digestion of SERT with Nep II showed promise, as 38 peptides were identified from the digestion, covering 40% of the sequence. Similarly, the in-solution digestion of SERT with pepsin resulted in 43.5% sequence coverage, but fewer peptides were identified when compared with the digestion with Nep II (Table 1 and Figure S-9). In cases where in-solution digestion may be required, for example, when a substrate used for HDX-MS of SERT is incompatible with immobilized protease columns, these two proteases would make for an interesting starting point for further optimization of digestion conditions. The in-solution digestion of LeuT and DAT resulted in fewer than 10 peptides being identified from each digestion, with the exception of pepsin digestion of DAT, which resulted in 14 peptides being identified, covering only 13.3% of its sequence (Figures S-6 and S-8).

In general, in-solution digestion yields a number of issues during an HDX-MS experiment, in particular, an increase in sample handling time, resulting in increased back-exchange as well as increased signal interference from peptides originating from autoproteolysis. However, in cases where the micelle is hindering optimal digestion by an immobilized protease, our results indicate that it may be worthwhile to explore in-solution digestion as an alternate route to improve the sequence coverage, or one can try to raise the pressure during on-column digestion, which may destabilize the micelle.

We note that the low amount of protein used for each digestion here may have introduced a bias toward the on-column digestion format. Also, whereas higher protease/protein ratios could be used in the in-solution setup to optimize digestion, a large increase in that ratio can have an undesirable effect during the ensuing chromatographic separation of the HDX-MS experiment and was therefore not tested.

Optimized Chromatographic Separation during HDX-MS of Integral Membrane Proteins. The selection of a suitable chromatographic column for peptide separation can provide a desirable increase in sequence coverage often needed for the HDX-MS analysis of membrane proteins. A back-to-back comparison of peptide identification from the LC-MS/MS analysis of DAT showed an 11.5% increase in sequence coverage when using C8 columns for the peptide separations instead of C18 columns. Under HDX-MS chromatographic conditions, the sequence coverage was increased from 56.4% using C18 resin to 67.9% with C8. This was shown to be true for SERT, as well, but while applying longer chromatographic gradients than is routinely used with HDX-MS. Although the increase in sequence coverage was only ∼10% for DAT, manual evaluation of the mass spectra for individual peptides showed an improvement in signal quality by switching to the shorter chain length columns. Out of 83 DAT peptide identifications shared by both column setups, 60 peptides had higher s/n ratio when using C8 columns for the chromatographic separation. On detailed inspection, the increased s/n ratio was found to be caused by a reduction in the noise level rather than by an increase in the intensity of the individual signals. The difference in the s/n ratio and the increase in the sequence coverage could not be explained by a difference in the hydrophobicity (GRAVY index) of the peptides[25] or by the length of the peptides. The lower concentration of organic solvent required to elute peptides is a possible explanation for the improved s/n ratio. It would allow for a slower increase in organic solvent concentration in the mobile phase per minute while keeping the overall length of the chromatographic separation constant. The increase in organic solvent was 4.2% per minute when using C18 columns compared with 2.2% with C8 columns. This may result in improved chromatographic separation and more stable electrospraying caused by the slower increase in acetonitrile concentration.

Optimal Choice of Aspartic Protease for HDX-MS is Highly Protein-Specific. A striking difference in the

DOI: 10.1021/acs.analchem.9b00973
Anal. Chem. 2019, 91, 10970−10978

digestion of LeuT, DAT, and SERT was observed (Table 1). On the one hand, the digestion of LeuT with immobilized pepsin resulted in double the number of identified peptides compared with its digestion by immobilized Rpn. On the other hand, the digestion of DAT and SERT with immobilized Rpn resulted in a 1.6- and 2.8-fold increase in peptide identifications, respectively, in comparison with pepsin. The difference in the number of identified peptides from in-solution digestions between the three transporters was even more remarkable. Whereas in-solution digestion of SERT resulted in 43.5% sequence coverage using pepsin, the in-solution digestions of DAT and LeuT were largely unsuccessful, with only the digestion of DAT resulting in more than 10 identified peptides. All other in-solution digestions of LeuT and DAT resulted in fewer than 10 peptide identifications. Considering that the three transporters are structurally related and share a common fold, these results highlight the importance of screening a panel of available proteases compatible with HDX-MS to optimize the sequence coverage of individual membrane proteins.

Comparison of Properties of TM Helices. The four integral membrane proteins are largely composed of membrane-spanning helices, which comprise around and over 50% of their respective protein sequence: ClC-ec1: 70.8%;[26] DAT: 49.3%;[27] LeuT: 63.2%;[20] SERT: 52.2%[28]. When their hydropathy was inspected, all four proteins followed similar trends, although, as expected, a larger portion of ClC-ec1 comprises hydrophobic residues compared with the other three transporters.

Although the hydrophobic properties of the proteins are similar, the identification of peptides covering the TM helices differs to a large extent. The highest sequence coverage of the helices was achieved for ClC-ec1, where peptides covered 83.6% of the helices when ClC-ec1 was digested using immobilized pepsin. Conversely, immobilized pepsin resulted in 20.4, 24.2, and 36.2% coverage of the TM helices for DAT, LeuT, and SERT, respectively. Interestingly, digestion using immobilized Rpn resulted in the improved coverage of the membrane-spanning helices in the three transporters. The highest coverage of the helices was achieved for SERT (54.7%), whereas 25.5 and 30.7% of the helices were covered in DAT and LeuT, respectively (Table 2).

The comparison of the overall sequence composition of the membrane-spanning helices (Table 3) does not provide a clear explanation for the differences observed in the sequence coverage of the membrane-spanning helices. As an example, although the membrane-spanning helices in ClC-ec1 contain a higher ratio of leucine and methionine, the helices of one or more of the transporters contain a higher ratio of phenylalanine, tyrosine, and tryptophan, all residues that are favored for pepsin cleavage. Additionally, although a higher ratio of lysine and arginine in these regions can be found in DAT and LeuT, respectively, the ratio of proline and histidine is higher in ClC-ec1, all residues that have been shown to have a negative effect on pepsin cleavage.[29]

Table 2. Sequence Coverage of Membrane Spanning Helices[a]

	ClC-ec1 (%)	DAT (%)	LeuT (%)	SERT (%)
iP	83.6	20.4	24.2	36.2
iRpn	79.1	25.5	30.7	54.7
iNep I	46.3	–	17.6	39.8
sP	31.9	2.5	7.2	33.1
sRpn	12.8	–	4.5	31.0
sNep I	2.7	–	1.8	2.1
sNep II	21.2	–	9.6	17.9

[a]i, immobilized; s, in solution; P, pepsin; Rpn, rhizopuspepsin; Nep I, nepenthesin I; Nep II, nepenthesin II; ClC-ec1, Cl^-/H^+ exchange transporter; DAT, dopamine transporter; LeuT, leucine transporter; SERT, serotonin transporter. The highest sequence coverage achieved for membrane-spanning helices for the respective integral membrane protein is underlined. –, Fewer than five peptides were identified for the entire protein. The sequence coverage was calculated only using the number of identified residues within a transmembrane helix and the total number of residues making up the transmembrane region of each membrane protein.

Table 3. Overall Sequence Composition of Membrane-Spanning Helices in Four Integral Membrane Proteins[a]

residue	ClC-ec1 (%)	DAT (%)	LeuT (%)	SERT (%)
I	9.9	8.2	11.9	12.2
V	7.5	10.1	8.7	9.4
L	16.1	14.8	13.1	11.6
F	6.3	11.3	11.9	9.4
C	0.9	1.6	0.0	3.0
M	4.5	1.9	3.6	2.7
A	12.8	10.7	11.3	9.4
G	12.2	7.5	7.8	7.9
T	6.3	3.5	4.5	5.8
W	1.8	2.8	3.0	2.7
S	2.4	6.6	3.3	6.1
Y	1.8	6.9	3.9	6.7
P	3.6	2.8	3.0	2.7
H	1.2	0.6	0.9	0.3
E	3.0	1.9	3.9	2.4
Q	2.1	0.6	0.9	1.5
D	1.2	2.8	0.9	1.8
N	1.5	1.3	1.8	0.9
K	2.1	2.2	1.8	1.5
R	3.0	1.9	3.9	1.8

[a]Residue: single letter identity of the amino acids. ClC-ec1, Cl^-/H^+ exchange transporter; DAT, dopamine transporter; LeuT, leucine transporter; SERT, serotonin transporter. The composition was calculated only using residues within transmembrane helical regions.

CONCLUSIONS

We performed a systematic optimization of the HDX-MS analysis of digestion-resistant membrane proteins through the screening of denaturing additives, digestion conditions, and chromatography. Importantly, the sequence coverages reported here were the result of a fixed non-exhaustive number of MS/MS acquisitions (i.e. two acquisitions) per condition to allow for a screening solely of the impact of individual sample, digestion, and chromatography conditions. We have recently published the HDX-MS analysis of LeuT,[31] SERT,[32] and DAT[33] and in these works a significant further increase in sequence coverage for the three proteins was obtained by not only combining advantageous sample and digestion conditions for each protein revealed by the current work, but also by increasing the number and types of MS/MS acquisitions to identify peptides for each protein.

As previously shown, the choice of additives such as denaturants is highly dependent on the target protein.[23,30] In this work, the use of urea as a denaturant successfully improved the sequence coverage of all four membrane proteins

DOI: 10.1021/acs.analchem.9b00973
Anal. Chem. 2019, 91, 10970−10978

compared with quenching without the presence of additives. The addition of Gnd-HCl to the quench buffers of LeuT, DAT, and SERT, on the contrary, resulted in reduced sequence coverage. Additionally, through the use of a chromatographic column with a shorter chain length, we observed improved sequence coverage believed to be due to better peptide separation and improved s/n ratios in MS analyses.

The comparison of digestions of LeuT, DAT, and SERT highlighted how protein dependent the choice of protease is. Even though the choice of quench buffer additives had been optimized using immobilized pepsin, immobilized Rpn resulted in the highest sequence coverage for DAT and SERT (38.9 and 58.3%, respectively), proving it to be an interesting option for both target proteins. The lower sequence coverage observed following in-solution digestions relative to on-column digestions of the four integral membrane proteins shows the gain of performing digestions on-column at higher temperature (20 °C) rather than in a cooled solution. Additionally, higher pressure during the on-column digestion may help destabilize the micelle around membrane proteins, allowing us access to cleavage sites otherwise blocked by detergent or lipid molecules. Despite the fact that in-solution digestions were outperformed by immobilized proteases in almost all of our cases, they should definitely not be excluded from the digestion optimization process for other systems. Here Nep II proved to be an interesting option for in-solution digestion, resulting in the highest number of identified peptides for in-solution digestions of ClC-ec1, LeuT, and SERT.

Our findings should also be relevant to HDX-MS experiments on integral membrane proteins embedded in liposomes or nanodiscs. The digestion step in an HDX-MS experiment of a membrane protein in a lipid mimetic system is always performed under denaturing conditions (low pH) and commonly in the presence of chaotropic agents like urea and possibly the reducing agent TCEP. Thus it is unlikely that a significant higher-order protein−lipid structure remains under these conditions to block protease access and thus differentiate digestion conditions in such samples from the digestion conditions of the detergent-solubilized membrane proteins studied here, especially if the enzyme is not restrained by immobilization (i.e., in-solution digestion). Furthermore, as shown in Table S-2, we find that the presence of lipids (POPC, POPE, POPG), detergent (DDM), and CHS does not have a significant impact on the performance of any of the proteases under the conditions used.

Though tempting from the ease-of-use point of view, our results show that a single best HDX-MS protocol cannot be recommended for even closely related membrane proteins of one family, as it can result in suboptimal sequence coverage and results. In particular, with the availability of different aspartic proteases complementary to pepsin, these should be tested when attempting to enhance the sequence coverage and spatial resolution during the HDX-MS analysis of integral membrane proteins.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b00973.

Figure S-1. Sequence coverage of Phos B following digestion by immobilized proteases. Figure S-2. Sequence coverage of Phos B following in-solution digestion. Figure S-3. Sequence coverage of ClC-ec1 following digestion by immobilized proteases. Figure S-4. Sequence coverage of ClC-ec1 following in-solution digestion. Figure S-5. Sequence coverage of LeuT following digestion by immobilized proteases. Figure S-6. Sequence coverage of LeuT following in-solution digestion. Figure S-7. Sequence coverage of DAT following digestion by immobilized proteases. Figure S-8. Sequence coverage of DAT following in-solution digestion. Figure S-9. Sequence coverage of SERT following in-solution digestion. Table S-1. Extended overview of the digestion results obtained by the different proteases. Table S-2. Comparison of the effect of different quench conditions on the digestion of phosphorylase B (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kasper.rand@sund.ku.dk.

ORCID

Ingvar R. Möller: 0000-0003-4497-2955

Kasper D. Rand: 0000-0002-6337-5489

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was made possible through support from the Danish Council for Independent Research (0602-02740B to K.D.R.; 0602-02100B and 4183-00581 to C.J.L.), the UCPH bioSYNergy center of excellence, and the Novo Nordic Foundation (NNF17OC0028582 to C.J.L.). J.H. acknowledges support from GAUK (389115). P.E.M. acknowledges financial support from EU/MEYS NPU II (LQ1604), CZ.1.05/1.1.00/02.0109, and COST-CZ (LD15089). Access to the MS facility in BioCeV was enabled through the LM2015043 CIISB project.

■ REFERENCES

(1) Hvidt, A.; Linderstrøm Lang, K. *Biochim. Biophys. Acta* **1954**, *14*, 574−575.

(2) Zhang, Z.; Smith, D. L. *Protein Sci.* **1993**, *2* (4), 522−531.

(3) Englander, S. V.; Poulsen, A. *Biopolymers* **1969**, *7*, 379−393.

(4) Cravello, L.; Lascoux, D.; Forest, E. *Rapid Commun. Mass Spectrom.* **2003**, *17* (21), 2387−2393.

(5) Man, P.; Montagner, C.; Vernier, G.; Dublet, B.; Chenal, A.; Forest, E.; Forge, V. *J. Mol. Biol.* **2007**, *368* (2), 464−472.

(6) Man, P.; Montagner, C.; Vitrac, H.; Kavan, D.; Pichard, S.; Gillet, D.; Forest, E.; Forge, V. *FEBS J.* **2010**, *277* (3), 653−662.

(7) Marcoux, J.; Man, P.; Petit-Haertlein, I.; Vivès, C.; Forest, E.; Fieschi, F. *J. Biol. Chem.* **2010**, *285* (37), 28980−28990.

(8) Kadek, A.; Mrazek, H.; Halada, P.; Rey, M.; Schriemer, D. C.; Man, P. *Anal. Chem.* **2014**, *86* (9), 4287−4294.

(9) Kadek, A.; Tretyachenko, V.; Mrazek, H.; Ivanova, L.; Halada, P.; Rey, M.; Schriemer, D. C.; Man, P. *Protein Expression Purif.* **2014**, *95*, 121−128.

(10) Rey, M.; Man, P.; Brandolin, G.; Forest, E.; Pelosi, L. *Rapid Commun. Mass Spectrom.* **2009**, *23* (21), 3431−3438.

(11) Yang, M.; Hoeppner, M.; Rey, M.; Kadek, A.; Man, P.; Schriemer, D. C. *Anal. Chem.* **2015**, *87* (13), 6681−6687.

(12) Trabjerg, E.; Nazari, E. Z.; Rand, K. D. *TrAC, Trends Anal. Chem.* **2018**, *106*, 125−138.

DOI: 10.1021/acs.analchem.9b00973
Anal. Chem. 2019, 91, 10970−10978

(13) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. *J. Mol. Biol.* **2001**, *305* (3), 567−580.
(14) Yıldırım, M. A.; Goh, K.-I.; Cusick, M. E.; Barabási, A.-L.; Vidal, M. *Nat. Biotechnol.* **2007**, *25* (10), 1119−1126.
(15) Doerr, A. *Nat. Methods* **2009**, *6* (1), 35−35.
(16) Morgan, C. R.; Hebling, C. M.; Rand, K. D.; Stafford, D. W.; Jorgenson, J. W.; Engen, J. R. *Mol. Cell. Proteomics* **2011**, *10* (9), M111.010876.
(17) Rey, M.; Man, P.; Clémençon, B.; Trézéguet, V.; Brandolin, G.; Forest, E.; Pelosi, L. *J. Biol. Chem.* **2010**, *285* (45), 34981−34990.
(18) Rey, M.; Mrázek, H.; Pompach, P.; Novák, P.; Pelosi, L.; Brandolin, G.; Forest, E.; Havlíček, V.; Man, P. *Anal. Chem.* **2010**, *82* (12), 5107−5116.
(19) Wang, L.; Pan, H.; Smith, D. L. *Mol. Cell. Proteomics* **2002**, *1* (2), 132−138.
(20) Yamashita, A.; Singh, S. K.; Kawate, T.; Jin, Y.; Gouaux, E. *Nature* **2005**, *437* (7056), 215−223.
(21) Maduke, M.; Pheasant, D. J.; Miller, C. *J. Gen. Physiol.* **1999**, *114* (5), 713−722.
(22) Goehring, A.; Lee, C.-H.; Wang, K. H.; Michel, J. C.; Claxton, D. P.; Baconguis, I.; Althoff, T.; Fischer, S.; Garcia, K. C.; Gouaux, E. *Nat. Protoc.* **2014**, *9* (11), 2574−2585.
(23) Zhang, X.; Chien, E. Y. T.; Chalmers, M. J.; Pascal, B. D.; Gatchalian, J.; Stevens, R. C.; Griffin, P. R. *Anal. Chem.* **2010**, *82* (3), 1100−1108.
(24) Ahn, J.; Jung, M. C.; Wyndham, K.; Yu, Y. Q.; Engen, J. R. *Anal. Chem.* **2012**, *84* (16), 7256−7262.
(25) Kyte, J.; Doolittle, R. F. *J. Mol. Biol.* **1982**, *157* (1), 105−132.
(26) Dutzler, R.; Campbell, E. B.; Cadene, M.; Chait, B. T.; MacKinnon, R. *Nature* **2002**, *415* (6869), 287−294.
(27) Penmatsa, A.; Wang, K. H.; Gouaux, E. *Nature* **2013**, *503* (7474), 85−90.
(28) Coleman, J. A.; Green, E. M.; Gouaux, E. *Nature* **2016**, *532* (7599), 334−339.
(29) Hamuro, Y.; Coales, S. J.; Molnar, K. S.; Tuske, S. J.; Morrow, J. A. *Rapid Commun. Mass Spectrom.* **2008**, *22* (7), 1041−1046.
(30) Li, S.; Lee, S. Y.; Chung, K. Y. *Methods Enzymol.* **2015**, *557*, 261−278.
(31) Merkle, P. S.; Gotfryd, K.; Cuendet, M. A.; Leth-Espensen, K. Z.; Gether, U.; Loland, C. J.; Rand, K. D. *Sci. Adv.* **2018**, *4*, eaar6179.
(32) Möller, I. R.; Slivacka, M.; Nielsen, A. K.; Rasmussen, S. G. F.; Gether, U.; Loland, C. J.; Rand, K. D. *Nat. Commun.* **2019**, *10*, 1687.
(33) Nielsen, A. K.; Möller, I. R.; Wang, Y.; Rasmussen, S. G. F.; Lindorff-Larsen, K.; Rand, K. D.; Loland, C. J. *Nat. Commun.* **2019**, *10*, 2714.

DOI: 10.1021/acs.analchem.9b00973
Anal. Chem. 2019, 91, 10970−10978

Publication II

Filandrova, R., Kavan, D., Kadek, A., Novak, P., & Man, P

Studying Protein–DNA Interactions by Hydrogen/Deuterium Exchange Mass Spectrometry

In Multiprotein Complexes: Methods and Protocols (Methods in Molecular Biology (2247)), 193–219, Humana Press Inc. (2021)

My contribution: *recombinant production of proteins and immobilization of proteases, performing experiments (LC-MS, LC-MS/MS and gel shifts), data analysis and interpretation, writing part of the manuscript, figure design*

Chapter 11

Studying Protein–DNA Interactions by Hydrogen/Deuterium Exchange Mass Spectrometry

Ruzena Filandrova, Daniel Kavan, Alan Kadek, Petr Novak, and Petr Man

Abstract

Protein hydrogen/deuterium exchange (HDX) coupled to mass spectrometry (MS) can be used to study interactions of proteins with various ligands, to describe the effects of mutations, or to reveal structural responses of proteins to different experimental conditions. It is often described as a method with virtually no limitations in terms of protein size or sample composition. While this is generally true, there are, however, ligands or buffer components that can significantly complicate the analysis. One such compound, that can make HDX-MS troublesome, is DNA. In this chapter, we will focus on the analysis of protein–DNA interactions, describe the detailed protocol, and point out ways to overcome the complications arising from the presence of DNA.

Key words DNA, Hydrogen/deuterium exchange, Protein–DNA binding, Structural mass spectrometry, Transcription factor

1 Introduction

Structural proteomics is a rapidly developing field focusing specifically on protein structural dynamics and characterization of the architecture of protein macromolecular assemblies. Hydrogen/deuterium exchange (HDX) ranks as one of the most versatile and prominent methods in this area. Its biggest advantages are not being limited by protein size and the ability to study proteins in their truly native-like environment, at virtually any pH, buffer composition, protein concentration, or temperature (*see* **Note 1**). In a typical setup, an HDX experiment starts by the dilution of a protein into a deuterated buffer. At selected time points, aliquots are taken, and the exchange is quenched (nearly stopped) by rapid acidification to pH 2.5 and by lowering the temperature to 0 °C. Subsequently, each sample is digested by acidic, nonspecific proteases, and deuteration of the generated peptides is measured by mass spectrometry (MS) [1–3].

Arnaud Poterszman (ed.), *Multiprotein Complexes: Methods and Protocols*, Methods in Molecular Biology, vol. 2247,
https://doi.org/10.1007/978-1-0716-1126-5_11,

Albeit very powerful and usually straightforward, HDX-MS has also some disadvantages, e.g., rather limited spatial resolution and need for sample dilution as the experiment often starts with ten-fold dilution into a deuterated buffer. However, due to the coupling to sensitive mass spectrometric detection, micro- and submicromolar concentrations are easily managed. Although HDX in combination with mass spectrometry was introduced more than two decades ago and there are hundreds of research articles describing its uses to study protein folding/unfolding, protein–protein and protein–ligand interactions [4, 5], the number of reports utilizing HDX for protein–nucleic acid interaction is rather limited. This is most likely due to the nature of nucleic acids that complicate HDX-MS analyses. During the quenching step (lowering the pH) in an HDX-MS workflow, the DNA backbone becomes protonated, which often results in its poor solubility and precipitation [6]. The protein interacting with DNA then tends to precipitate as well. The presence of DNA can also have adverse effects on chromatographic separation. Such behavior, however, depends on the DNA composition and length. Therefore, reports dealing with larger DNA stretches always utilize technical tricks to deal with the troublesome DNA while those using very short oligonucleotides may sometimes be spared these adverse effects.

The first ever attempt to use HDX for protein–nucleic acid interaction study enabled the determination of a protein–DNA complex dissociation constant by a combination of HDX and matrix-assisted laser desorption/ionization mass spectrometry [7]. However, it took five more years until Sperry et al. in 2008 truly successfully probed structural features of protein–DNA interactions of two biological systems (a protein binding to a telomeric oligonucleotide and the protease thrombin interacting with an aptamer). These studies were successfully accomplished thanks to the incorporation of strong anion exchange trap column prior to the liquid chromatography-mass spectrometric (LC-MS) analysis to selectively trap and remove the oligonucleotides [8, 9]. Besides an ion exchanger, the use of protamine sulfate has also been reported as beneficial in the HDX-MS analysis of oligonucleotide-containing samples. Poliakov and coworkers concluded that the protein–DNA co-precipitation upon the lowering of pH is primarily of electrostatic origin. Thus, inclusion of small, highly basic, and hence, positively charged protamine improved protein recovery [10]. Alternatively, in another study, Roberts et al. counteracted the negative effect of DNA on the formation of peptides, solubility, and chromatographic separation in HDX by tuning the concentration of a positively charged denaturing agent (guanidine hydrochloride) in their quench buffer [11]. Use of another denaturant has recently been exploited by Graham et al. to explain structural and mechanistic aspects of DNA unwinding when guanidine

hydrochloride was replaced by urea, and an alternative aspartic protease was used in the quench buffer [12].

In this study, drawing upon the experience gained through our own work, we present an experimental protocol for the characterization of protein interactions with DNA (in the form of an oligonucleotide) as used in our group. We have repeatedly observed that careful optimization and proper control over the quality of the DNA entering an HDX experiment (amount of dsDNA formed, purity, salt concentration) are crucial for boosting its ability to bind to the interaction partner. Ensuring as strong an interaction as possible in this manner then sometimes largely mitigates the problems stemming from the DNA presence, as the amount of oligonucleotides used can be kept to a minimum (optimally down to 1:1 molar ratio to the protein), while still providing virtually complete saturation of the transcription factor (TF). Sometimes, however, even the best quality of sample is not enough, and the aforementioned techniques for coping with DNA come into play.

In the following set of protocols, we describe ways to perform protein–DNA HDX experiments, together with some of the potential pitfalls, with emphasis on the interaction of transcription factors with their DNA-response elements. As we have observed that protein:DNA systems can behave surprisingly differently even when the proteins as well as DNAs used are similar in size, we will be using two example systems. Forkhead box protein O4 (FOXO4)/DAF16 [13, 14] serves as an example where short dsDNA (13 bp) does not interfere significantly with peptide recovery, and only digestion conditions are optimized to reach good HDX spatial resolution. On the other hand, the TEA domain family member 1 (TEAD1)/M-CAT [15] represents the opposite situation when a 15-bp long oligonucleotide has a strong impact on peptide recovery. These striking differences in behavior stress the importance of having a vast array of conditions, proteases, and technical tricks at one's disposal to fine-tune the HDX protocol. The procedures we use, as described herein, should provide prospective users with a set of tools to successfully optimize conditions for HDX and perform their own analyses of a range of TF–DNA complexes.

2 Materials

1. Protein to be analyzed (*see* **Note 2**).
2. Forward and reverse DNA strands (*see* **Note 3**).
3. Buffer suitable for the complex (e.g., 20 mM HEPES, pH 7.4, 150 mM NaCl).
4. Gel casting tray and electrophoresis apparatus (Bio-Rad).

5. Non-denaturing acrylamide gel (12%)—for 10 ml mix the following: 6 ml H_2O, 1 ml 10× concentrated TBE buffer, 80 µl 10% ammonium persulfate (APS), 8 µl TEMED (*N*,*N*,*N*′,*N*′-tetramethylethylenediamine), 3 ml 40% acrylamide:*N*,*N*′-methylenebisacrylamide (1:19 mix).
6. DNA loading dye (Thermo Fisher).
7. Protein staining solution (45% methanol, 10% acetic acid, 0.25% Coomassie Brilliant Blue R250).
8. TBE buffer (90 mM Tris–Cl, 90 mM boric acid, 2 mM EDTA, pH 8.3).
9. Fluorescent DNA stain GelRed (Thermo Fisher).
10. Deuterium oxide (D_2O).
11. Quench buffer (500 mM glycine in LC-MS grade water, pH 2.3 set with concentrated HCl).
12. Urea or Guanidine hydrochloride.
13. Liquid nitrogen.
14. Peltier cooled box or a polystyrene box filled with ice and water mixture (*see* **Note 4**).
15. Two 6-port valves, one of them with injection needle port (Rheodyne/IDEX).
16. PEEK (polyether ether ketone) tubing (OD 1/16″, ID—250 µm and 750 µm) and stainless steel tubing (OD 1/16″, ID—125 µm).
17. UPLC (ultrahigh-performance liquid chromatography) fittings with ferrules (1/16″).
18. Hamilton syringe with blunt end needle (100 µl).
19. Timer.
20. Columns with immobilized acidic protease—pepsin, rhizopuspepsin (protease type XVIII), nepenthesin-1, nepenthesin-2, or aspergillopepsin (protease type XIII) immobilized on POROS™ 20AL (Thermo Fisher Scientific) using previously described [16, 17] protocols and packed into stainless steel guard columns (2 mm × 20 mm or 1 mm × 20 mm, IDEX).
21. Desalting trap column—VanGuard Pre-column (ACQUITY UPLC BEH C18, 130 Å, 1.7 µm, 2.1 mm × 5 mm, Waters).
22. Reversed-phase analytical column (ACQUITY UPLC BEH C18, 130 Å, 1.7 µm, 1 mm × 100 mm, Waters).
23. Loading and desalting pump (1260 Infinity II Quaternary pump (Agilent Technologies)).
24. Gradient pump (1290 series, Agilent Technologies).
25. LC loading solvent: 0.4% formic acid in water.

26. LC solvent A (0.4% formic acid, 2% acetonitrile, 97.6% water).
27. LC solvent B (0.4% formic acid, 95% acetonitrile, 4.6% water).
28. Electrospray ionization (ESI) mass spectrometer (any ESI-equipped MS with reasonable resolution is usable; in our case, we used an electrospray ionization—Fourier transform ion cyclotron resonance 15T solariX XR, Bruker Daltonics).
29. Software: DataAnalysis 4.1 (Bruker Daltonics), ProteinScape 4 (Bruker Daltonics), DeutEx (http://peterslab.org/downloads.php under HDX tools section), mMass (http://mmass.org), PyMol (https://pymol.org/; v 2.1.0), MSTools (http://peterslab.org/MSTools/).

3 Methods

3.1 Verification of Protein–DNA Complex Formation by a Gel Shift Assay

Prior to the HDX-MS experiment, it is advisable to verify whether the DNA duplex is present and that the complex between protein and ligand (here DNA) is indeed formed upon mixing of the components. In this respect, knowledge of the binding constants offers an advantage as the experimental setup can be adjusted to achieve full protein occupancy (complex presence) during the labeling step [18]. Here, we use gel shift assay to check DNA duplex formation and its binding to the studied protein.

1. Dilute both DNA strands in LC-MS purity water and mix them at equimolar ratio.
2. Heat the mixture at 95 °C for 1 min, and let it slowly cool down to room temperature.
3. Dilute/buffer exchange the protein into a buffer suitable for the complex which is later used for H/D exchange (20 mM HEPES, pH 7.4, 150 mM NaCl, *see* **Note 5**) to a concentration at which full occupancy of the protein by DNA is achieved (*see* **Note 6**).
4. Add equimolar amounts of dsDNA to the protein and let the mixture equilibrate for at least 2 min at a temperature where both the dsDNA and protein are stable (*see* **Note 7**).
5. Prepare following samples: 200 ng of both single-stranded oligonucleotides, 200 ng of free dsDNA, 5 μg of free protein, and the equimolar protein–DNA complex equal to 5 μg of the protein (*see* **Note 8**). All components must be prepared in the same buffer.
6. Add loading buffer to all samples.
7. Prepare non-denaturing polyacrylamide gel by mixing all components except APS. Add APS just before pouring the mixture into a gel casting tray (*see* **Note 9**).

8. Load the samples and run the electrophoresis in TBE buffer at a constant voltage of 50 V for one gel until the loading dye reaches half of the gel (*see* **Note 10**).
9. Stain the gel with any kind of fluorescent dye for DNA staining, take a picture, and stain the gel again in protein staining solution to visualize the protein. Only proteins in complex with DNA will migrate into the gel. Free protein will normally not separate as TFs are generally positively charged and interact strongly with negative DNA.
10. Formation of DNA duplex and protein–DNA complex are manifested as shifts in the electrophoretic mobility (Fig. 1). Compare the results from both the staining methods to confirm the complex formation—position of the complex band should be the same for protein and DNA staining.
11. If multiple complex concentrations are tested, select the one where there is the least free dsDNA (of the same gel mobility as dsDNA control) visible.

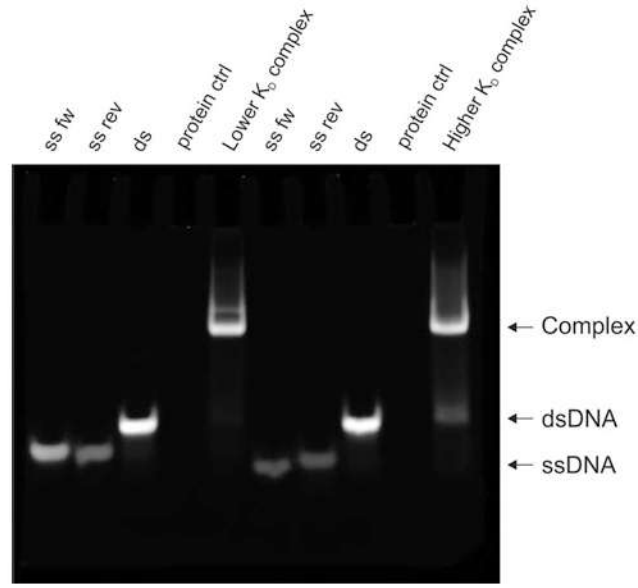


Fig. 1 Gel shift assay. Two oligonucleotides (DNA1 and DNA2) with different binding affinities are shown. Shifts in electrophoretic mobility in the third and fifth lane from the left indicate duplex DNA and protein–DNA complex formation, respectively. Protein control lane stayed empty since no DNA was present. In both the complex lanes, a band of the same mobility as dsDNA can be observed, though in the "higher K_D complex" lane, this band is much thicker, indicating lower bound fraction of the protein

3.2 Optimizing Digestion Conditions

Before starting the H/D exchange, it is crucial to verify how the protein is digested under HDX-MS compatible conditions. Different aspects to be considered are protein sequence coverage, peptide length, and peptide redundancy but also whether the protein remains soluble after quenching and freezing and how the presence of DNA affects the whole procedure. The primary goal is to ideally obtain full sequence coverage with peptides providing good spatial resolution, i.e., not very short or long ones. Recommendations regarding optimal length do vary slightly in the literature, but fragments between 8 and 12 amino acids probably represent the best standard. Besides the spatial resolution obtainable through the generated peptides, one can also target nearly amino acid resolution using suitable fragmentation techniques like electron transfer dissociation [19] or UV photodissociation [20] (*see* **Note 11**). In addition, emphasis should also be put on the generation of large numbers of overlapping peptides as these can further increase the spatial resolution [21, 22] and/or provide higher confidence in the observed changes (*see* **Note 12**). To fulfill all these goals, different proteases, ideally immobilized and packed into a column, should be tested [17, 23–26]. The columns can be used individually or combined in serial or parallel setting [27]. Column size as well as the protease density on the POROS resin, which together dictate the enzyme:protein ratio, can be varied. Other factors affecting the digestion are flow, temperature, pressure, and the use of denaturing agents. Time spent on the proteolytic column is primarily determined by the desalting LC flow. Therefore, higher flow leads to faster digestion and produces longer fragments and higher redundancy (number of overlapping peptides). Slower flow, on the other hand, produces more complete digestion, which however may at times be even detrimental, if the peptides produced are too short. Typical digestion temperature should be close to 0 °C to minimize deuterium-loss (back-exchange), but it can be raised to 15 °C or even 20 °C locally to increase the digestion efficiency. This is usually achieved either in a dedicated separately temperature-controlled chamber for the protease column or by placing the column out of the ice/water bath. However, when tweaking these parameters, one should always keep in mind that optimal balance between efficient digestion and minimal H/D back-exchange conditions must be maintained (hence running the analysis as fast as possible and at the lowest possible temperature). For proteins not offering satisfactory digestion under mild denaturing conditions provided by the acidic environment of the quench buffer alone, addition of denaturing agents is suggested. Urea or guanidine can be used at quite high concentrations with all the immobilized proteases; however, besides rendering the target protein susceptible to digestion, these agents also affect the protease itself as well as peptide binding to the reversed phase resin in the trap column. While guanidine was shown to have mainly a detrimental effect, urea can even enhance

proteolytic activity [17]. In addition, guanidine is difficult to remove completely during the subsequent desalting step, and thus, 4 M urea (final concentration upon quenching) may always be suggested as the first choice.

When analyzing samples containing nucleic acids, specific problems may additionally arise upon the quenching of H/D exchange, as mentioned above. As the pH drops, nucleic acids tend to precipitate, and it is very likely that the protein of interest will coprecipitate as well. Therefore, several strategies described in the Introduction were developed to remove the nucleic acids after quenching or to reduce their detrimental effect on the HDX-MS workflow.

1. Prepare the duplex DNA according to the **steps 2** and **3** of the previous chapter (Subheading 3.1) and mix it with the protein in equimolar ratio and the concentration that will later be used during D_2O labeling. Incubate the mixture for 10–20 min to ensure binding equilibrium. Also, prepare a sample without DNA for the comparison of how DNA affects the digestion pattern. In this step, use normal (H_2O, not D_2O based) buffers.
2. Start the LC-MS/MS system—calibrate mass spectrometer, start analytical gradient pump (Agilent 1290, running at 40 µl/min), loading pump (Agilent 1260, running at 100 µl/min), and pre-cool the LC setup to 0 °C (Fig. 2). If digestion is carried out at higher temperature, make sure that the protease column is well conditioned (*see* **Note 13**).

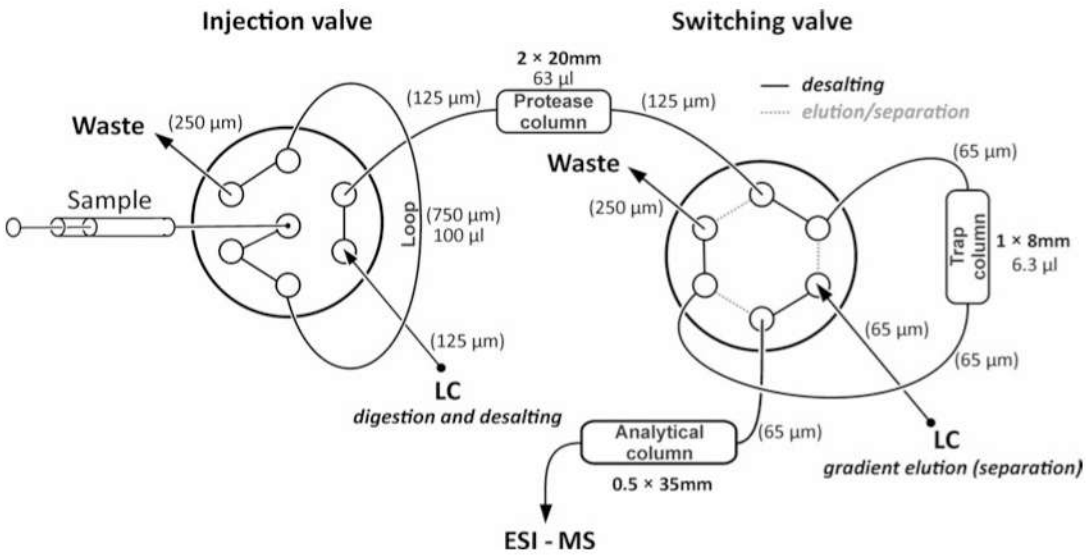


Fig. 2 Schematic representation of HDX LC-MS setup. Dimensions of the columns and tubing are shown. The setup is in position where digestion and desalting are running. Upon switch of the switching valve, the ports are connected by the gray dashed lines and gradient elution of the desalted peptides from trap column onto an analytical column is performed

3. Run a standard protein (e.g., horse heart myoglobin or rabbit muscle phosphorylase B) to verify the LC separation and protease column activity. This standard protein used for system quality control must be different from the protein of interest. Always keep in mind that the protein must be prepared in an acidic buffer as the digestion is done using aspartic proteases. Accidental injection of neutral pH buffer will quickly and irreversibly inhibit pepsin. Other proteases mentioned here are more robust in this regard and can tolerate elevated pH temporarily. Typically prepare 30 pmol of myoglobin in 100 µl of 250 mM glycine-Cl buffer pH 2.3.
4. Run two blank injections (100 µl of pure quench buffer) to clean the system and to ensure proper and stable LC conditions (*see* **Note 14**).
5. Mix the sample with the quench buffer to get the sample volume and concentration as planned for subsequent HDX-MS experiment (*see* **Note 15**). Hence, 50 µl of 500 mM glycine-Cl buffer pH 2.3 with 50 µl of 2 µM protein solution in 20 mM HEPES, pH 7.4, 150 mM NaCl.
6. Inject the sample directly or subject it to a freeze–thaw cycle if you plan to collect aliquots by freezing in liquid nitrogen.
7. Wait for desired time (3 min) until digestion and desalting are finished (*see* **Note 16**) and switch the second LC valve so that the trap column is now in the path of gradient elution and the peptides will continuously elute on the analytical column. Together with the switch, also start the gradient (*see* **Note 17**) and collect the MS data.
8. Process the LC-MS/MS file using the instrument vendor-specific software and generate input for the search engine (*see* **Note 18**).
9. Run a database search using the input file from the previous step. Use a custom-made database containing the sequence of the protein of interest and the protease(s) used and eventually other proteins present in the sample. Do not set any digestion preferences, nor taxonomy filtering. Include possible relevant modifications (fixed or variable) that may occur in the protein. Set the mass tolerance on precursor and fragments according to your instrument performance. Use scoring routines to discard improbable matches. It is also advisable to run the LC-MS/MS analysis 3–4 times and only use peptide identifications that occur in the majority of the analyses.
10. Copy or export the search result (list of identified peptides) to a spreadsheet editor (e.g., Microsoft Excel) and extract the Peptide Start and End columns (peptide limits). Use this in a simple text file to visualize the coverage map using DrawMap

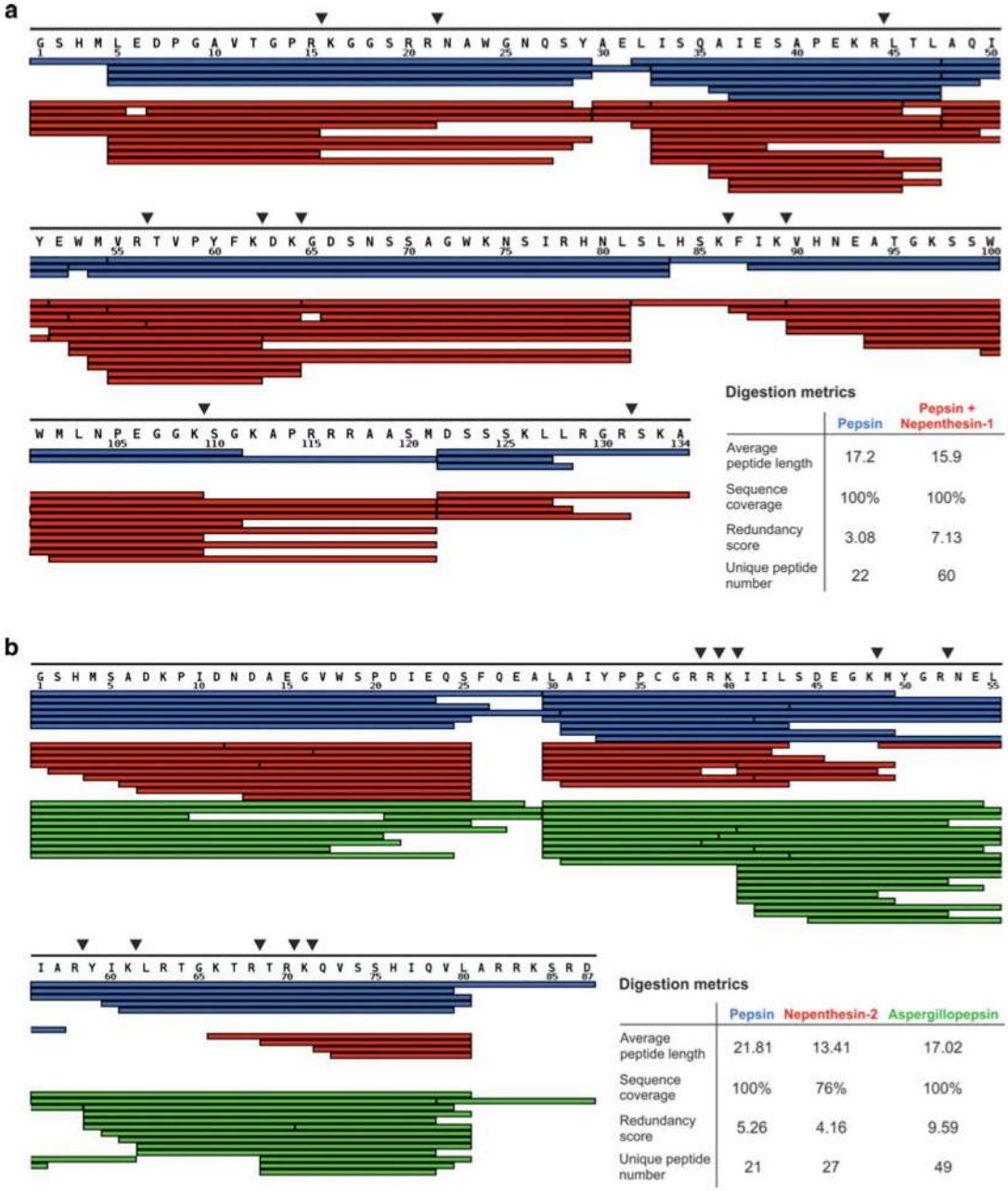


Fig. 3 Sequence coverages of the two studied DNA binding domains. In all instances, immobilized proteases were used. Table at the right lower corner of each panel shows digestion metrics for the individual conditions (coverage, number of peptides, average peptide length, and redundancy—how many peptides are on average covering each residue). (**a**) DNA binding domain of FOXO4 where initial proteolysis by pepsin (blue bars) was after optimization replaced by combined digestion with pepsin followed by nepenthesin-1 (red bars). This led to better spatial resolution and higher redundancy. In case of FOXO4, the presence of oligonucleotides (DAF-16) caused no problems, and thus, only tuning of the digestion itself was necessary. (**b**) Examples from the tuning procedure with TEAD1 DNA binding domain. Pepsin (blue bars) provided full sequence

script at the MSTools website (http://peterslab.org/MSTools/DrawMap/DrawMap.php) and eventually use Excel macro Digestion_metrics.xlsm (http://peterslab.org/downloads.php under HDX tools section) to extract the digestion metrics.

11. Based on the peptide length, redundancy, coverage, and number of gaps in the peptide map, try optimizing the digestion by using alternative proteases or quench buffer additives as described above (example in Fig. 3) (*see* **Note 19**).

3.3 Hydrogen/Deuterium Exchange

The methodology described here represents a manual approach. Even though nowadays there are robotic solutions available that can do the sample preparation and injection automatically, they are quite specific depending on the manufacturer of the robot and the software controlling it. The manual approach represents a straightforward and inexpensive way that can be easily set up in any laboratory possessing an LC system and a mass spectrometer (*see* **Note 20**).

The actual H/D exchange should be started from equilibrated solutions in which the complex is fully formed, and the temperature of all components is stable. The first step in the deuterium labeling reaction is the dilution of the studied protein into a deuterated buffer (*see* **Note 21**) that has otherwise identical composition as the initial buffer used for complex pre-formation. After a predetermined time, the exchange is "stopped" by the addition of acidic quench buffer and by lowering the temperature or eventually by flash freezing in liquid nitrogen.

The experiments should ideally be done (at least) in triplicate to allow for statistical validation. There are several ways how to perform the replication in HDX-MS. The highest validity is achieved by doing a biological replication (in the case of recombinant proteins this means using independent protein production batches). A less significant (though commonly used) level of replication (technical) is reached by separate deuterium labeling, and the lowest tier is represented by measuring the same labeling reaction repeatedly [28]. Replication also allows one to estimate the significance level above which the differences in deuteration between individual states may be considered meaningful [29].

Fig. 3 (continued) coverage but lower spatial resolution—on average it had longest peptides. Nepenthesin-2 (red bars) led to over-digestion, which is indicated by gaps in the sequence, and the identified peptides are mostly short ones. High activity of the protease resulted in the generation of short peptides and incomplete sequence coverage. The green bars show the final conditions where aspergillopepsin (protease type XIII) was utilized in the presence of urea. Addition of denaturing agent was necessary as the protein in the presence of oligonucleotides precipitated. Addition of urea (optimized to final 2 M concentration) led to much better protein recovery. In both examples shown here, the advantage of alternative proteases for the digestion of DNA binding proteins (which are rich in basic residues) is demonstrated through their ability to cleave after the Lys and Arg residues (highlighted by arrowheads above the sequence)

Starting from the moment of quenching and throughout the digestion, desalting, and subsequent analysis, the sample is exposed to protonated solvents. Even though the minimal exchange conditions are maintained through these steps, the sample still undergoes deuterium loss. The level of back-exchange is not uniform and depends on the sequence of individual peptides. However, this loss can be measured by analyzing a fully deuterated sample and then using these data, to correct deuteration levels on partially deuterated peptides [28]. In a typical measurement comparing two protein states, relative deuterium levels are the desired measure, and there is no need for back-exchange correction. Thus, in these cases, this control is often omitted. This is also likely due to the fact that no clear and universal protocol exists and that proteins often tend to precipitate or degrade before their full deuteration is achieved. However, the knowledge of the actual back-exchange is needed when absolute levels of deuteration are sought (e.g., in cases where mutated sequences of the same protein are compared) [30].

1. Prepare the required number of 0.5-ml Eppendorf tubes (for two conditions, three labeling replicates, and six time points, this would mean 36 tubes in total) with quench solution—50 μl of 500 mM glycine-Cl buffer, pH 2.3 (*see* **Note 22**).
2. Prepare (bigger) tubes with the protein alone and with the pre-formed protein–DNA complex. In these, the ten-fold dilution into deuterated buffer and the labeling reaction will be performed. At six preselected time points (*see* **Note 23**), 50 μl aliquots will be removed from the reaction and quenched. As three labeling replicates are to be done, three tubes with protein alone and three with the protein–DNA complex, each containing 40 μl of 2 μM protein solution must be prepared (*see* **Note 24**).
3. Prepare 2 ml of deuterated buffer (20 mM HEPES, 150 mM NaCl, pD 7.4) (*see* **Note 25**).
4. Prepare a time schedule that allows efficient pipetting especially in cases where larger amount of conditions or direct technical (labeling) replicates are performed at once. This can be easily done using MSTools script "Experiment planner" (http://peterslab.org/MSTools/HDExpPlanner/HDExpPlanner.php). For two experimental conditions, each replicated three times, and aliquot collection at 20 s, 2 min, 5 min, 20 min, 1 h, 3 h, this may look as shown in Fig. 4.
5. Prepare automatic pipettes with pre-set volumes for H/D mixing (360 μl) and aliquot collection (50 μl).
6. Use a timer that allows countdown followed by a count-up. Set 20 s count-down and start it. Aspirate 360 μl of the deuterated buffer and when the timer reaches zero (start of the whole

Time	Event
---	Mix TF-r1
20sec	Aliq no. 1 of TF-r1
2min	Aliq no. 2 of TF-r1
3min 30sec	Mix TF-DNA-r2
3min 50sec	Aliq no. 1 of TF-DNA-r2
5min	Aliq no. 3 of TF-r1
5min 30sec	Aliq no. 2 of TF-DNA-r2
7min	Mix TF-r2
7min 20sec	Aliq no. 1 of TF-r2
8min 30sec	Aliq no. 3 of TF-DNA-r2
9min	Aliq no. 2 of TF-r2
10min 30sec	Mix TF-DNA-r2
10min 50sec	Aliq no. 1 of TF-DNA-r2
12min	Aliq no. 3 of TF-r2
12min 30sec	Aliq no. 2 of TF-DNA-r2
14min	Mix TF-r3
14min 20sec	Aliq no. 1 of TF-r3
15min 30sec	Aliq no. 3 of TF-DNA-r2
16min	Aliq no. 2 of TF-r3
17min 30sec	Mix TF-DNA-r3
17min 50sec	Aliq no. 1 of TF-DNA-r3

Time	Event
19min	Aliq no. 3 of TF-r3
19min 30sec	Aliq no. 2 of TF-DNA-r3
20min	Aliq no. 4 of TF-r1
22min 30sec	Aliq no. 3 of TF-DNA-r3
23min 30sec	Aliq no. 4 of TF-DNA-r2
27min	Aliq no. 4 of TF-r2
30min 30sec	Aliq no. 4 of TF-DNA-r2
34min	Aliq no. 4 of TF-r3
37min 30sec	Aliq no. 4 of TF-DNA-r3
1h	Aliq no. 5 of TF-r1
1h 3min 30sec	Aliq no. 5 of TF-DNA-r2
1h 7min	Aliq no. 5 of TF-r2
1h 10min 30sec	Aliq no. 5 of TF-DNA-r2
1h 14min	Aliq no. 5 of TF-r3
1h 17min 30sec	Aliq no. 5 of TF-DNA-r3
3h	Aliq no. 6 of TF-r1
3h 3min 30sec	Aliq no. 6 of TF-DNA-r2
3h 7min	Aliq no. 6 of TF-r2
3h 10min 30sec	Aliq no. 6 of TF-DNA-r2
3h 14min	Aliq no. 6 of TF-r3
3h 17min 30sec	Aliq no. 6 of TF-DNA-r3

Fig. 4 Screenshot of HD Experiment planner script (MSTools: http:/peterslab.org/MSTools/HDExpPlanner/HDExpPlanner.php). Exact planning of aliquot collection for two experimental conditions (TF and TF-DNA) and three labeling replicates. Delays necessary to collect the aliquot or mix the exchange can be varied and are especially useful for larger number of conditions—here 30 s is left for aliquoting and 1 min for mixing

HDX experiment), add the buffer to the protein. Mix thoroughly and prepare the pipette (50 µl) for aliquot collection. Five seconds before the selected time (20 s), aspirate the labeled protein and at the precise time according to the experiment plan, add aspirated solution to the tube with the quench solution. Quickly mix/vortex and flash-freeze the tube in liquid nitrogen. Do this for all the samples according to the time schedule at times indicated by the timer.

7. Transfer the frozen tubes from liquid nitrogen to a deep freezer and store them until analysis (*see* **Note 26**).
8. Also prepare non-labeled reference samples (a non-deuterated control) for each condition. This is the same as preparing the (partially) deuterated samples, except that the buffers used are made with H_2O only.
9. If possible (depends on protein stability), prepare fully deuterated sample (*see* **Note 27**). In order to obtain a fully deuterated control, prepare deuterated buffer containing high concentration (4–6 M final concentration) of urea or guanidine and incubate the protein in it overnight at higher (e.g., 37 °C) temperature. Protein concentration should be as high as feasible, so that it can be lowered before quenching using

non-denaturing deuterated buffer. This will lower the denaturant concentration which can alter the subsequent workflow outcomes. From this perspective, using urea is the better option (*see* Subheading 3.2). Quench the reaction and eventually freeze the sample, if this step is included in the workflow.

3.4 Mass Spectrometric Analysis of Deuterated Samples

1. Use the LC system as in Subheading 3.2 and perform **steps 2–4**. Start all LC pumps, calibrate the mass spectrometer, pre-wash, and pre-condition the protease column. Then run the standard protein followed by two blank injections (*see* **Note 28**).
2. Before the end of previous analysis, take one of the partially deuterated samples from the freezer and start thawing it. Depending on the quench buffer composition, it will melt in 30 s to 1 min. As soon as it is thawed, inject it for analysis.
3. Repeat this for all collected samples (*see* **Note 29**).
4. Analyze fully deuterated control exactly as the partially deuterated samples.
5. Finally run two blanks followed by non-deuterated samples for each experimental condition.
6. Export and/or pre-process all LC-MS acquired data and prepare them for final data processing.

3.5 Interpreting Data from H/D Exchange

Nowadays, there are several different programs that are capable of largely automatic HDX-MS data processing. Waters users rely on their DynamX suite while others can use HDExaminer from Sierra Analytics that supports all native file types as well as open source ones. Another option for Waters and Thermo users is represented by the HDX Workbench [31] from Omics Informatics. Alternatively, there are software tools freely available from different research groups [22, 32–35]. These and other programs and their workflows were recently reviewed in detail by Claesen and Burzykowski and also by Eggertson et al. which in addition offers another view on HDX-MS data processing [36, 37].

The basic principle which is crucial to the understanding of the workflow and its requirements is, however, the same. To make this protocol widely applicable, we will show two possible scenarios. One, relying on a manual interpretation which is nowadays outdated and extremely laborious. However, it demonstrates well the basic principle which the available programs automate. It can also be useful for validation purposes and in specific cases (extraction of EX1/EX2 data). The other workflow presented here employs our own software called DeutEx (*see* **Note 30**).

Data processing in the HDX-MS workflow consists of four major steps. First is the identification of peptides generated during the proteolysis step and their temporal localization within an LC

run. Using the set of coordinates consisting of peptide sequence, its mass-over-charge (m/z) value(s) (*see* **Note 31**), and LC retention time, individual features can be located in the partially deuterated samples. In a subsequent step, information about the number of deuterons carried by individual peptides is extracted. Two major approaches to deuteration readout rely either on centroid readout or on deconvolution and theoretical fitting [36]. Finally, this information is presented in several graphical ways to visualize the effects monitored by the HDX-MS experiment.

3.5.1 Workflow A: (Largely Automated) Data Interpretation Using DeutEx

1. From an MS/MS search engine, e.g., MASCOT (here utilized by Bruker's ProteinScape), export the search result in the form of a csv file (*see* **Note 32**).
2. Prepare a separate simple text file containing the sequence of the studied protein in FASTA format.
3. Open DeutEx and go to *Analysis>Compose digest file from MS*. New window will appear. In this window, upload your protein sequence as a FASTA file, then csv file containing the list of identified peptides from MASCOT and an LC-MS run of a non-deuterated sample exported into simple text files (each MS scan corresponding to one file) containing a list of m/z values and their intensities (*see* **Note 33**).
4. Go to *Digest*, click *Create from report* and then *Extract scan/charge limits*. Each peptide will have its retention limits automatically identified, and extracted ion chromatograms will be drawn for individual charge states (m/z values). User can manually inspect the assignments and either manually correct or completely reject them. Once finished, the result should be exported into a text file that serves as a lookup file for an automated assignment of HDX data for the deuterated samples. Close the "Compose digest from MS window."
5. Go to *Analysis>Open Analysis directory*. It is advisable to pre-compile data for each state into a separate folder (which contains exported LC-MS data in a txt format, with appropriate fasta sequence and digest file—the list of peptides created in **step 4**). However, the data can also be compiled in DeutEx using the "Compose analysis" function.
6. Once all data are opened, check the settings for mass accuracies and filtering under the options menu and run the calculation. Typically, it lasts from seconds to a few minutes, depending on the number of conditions, number of peptides, and size of the exported data on a common laptop/office computer. When calculation is finished, run filtering, which tries to automatically remove incorrect assignments. Then look through the assignments manually, check the data, and correct any misassignments, if needed.

7. The basic output from DeutEx is a set of uptake plots that are displayed either as deuteration percentage or number of exchanged deuterons as a function of time. The data can also be exported in simple text form which allows their further processing in Excel or MSTools (http://peterslab.org/MSTools/) [38].

3.5.2 Workflow B: Manual Data Interpretation

1. In any program that is used for viewing the LC-MS data, open a non-deuterated LC-MS analysis.
2. For each peptide, trace an extracted ion chromatogram based on its m/z (usually included in the search engine report). Check also other charge states that are visible and enter this information into an Excel sheet together with the retention time limits for further use. Sum scans across the LC peak and export the spectrum as a simple text file (typically two columns including m/z values and intensity).
3. Open a partially deuterated LC-MS analysis and locate the deuterated signal corresponding to the peptide information gathered in the previous step. Repeat the step of summing up the scans and exporting the spectra into text files. Repeat iteratively for all peptides. For such a manual approach, it is advisable to select a minimal set of peptides (preferably short ones) covering the entire protein without many overlaps (*see* **Note 34**).
4. Once the data are exported to *.txt files, run mMass program [39] (*see* **Note 35**). Go to *File>Open* and open a file with non-deuterated data. The spectrum is not by default labeled, but it is not necessary at this point. Make sure that the correct way of peak picking is selected. Go to tab *Processing>Deisotoping* and set *"Label envelope tool"* to *Envelope Centroid* option. Now locate the selected peak (m/z) and zoom to the mass range covering the whole isotopic pattern of this peptide. Use *"Label envelope tool"* (select from tab *Tools*). Using Shift + mouse middle wheel scrolling, set the desired charge state—black dots are changing distance as you scroll up and down. Fit it to the desired isotopomer distance and now select the number of dots to cover all visible isotopic peaks from the particular envelope. This is done by Ctrl + Shift + mouse middle button scrolling. This tool helps you to localize all isotopes belonging to the selected peak. Left click and the peak will be labeled. Average m/z value is written into the spectrum, and the deconvoluted value (average mass) appears in the right panel (Fig. 5). Write this number down to an Excel file. Eventually, you can now repeat this step also for other charge states of the same peptide. One by one, all m/z values of each peptide from all deuterated samples are processed using this approach (*see* **Note 36**).

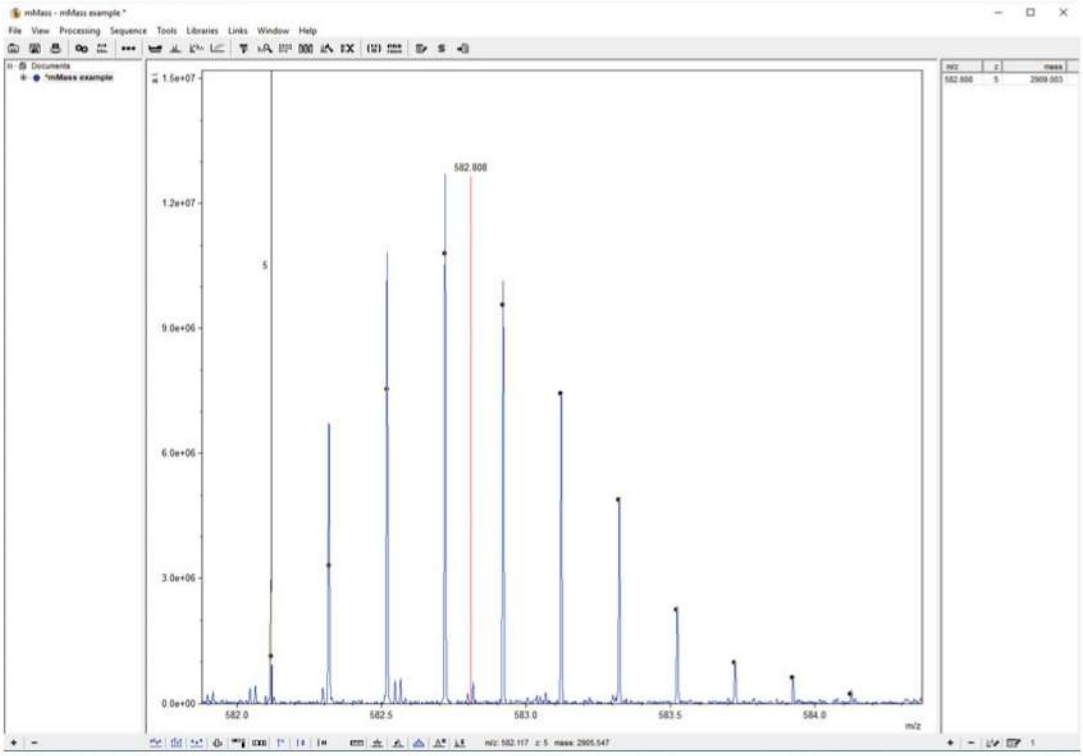


Fig. 5 Screenshot of mMass window with zoom on the isotopic pattern. Black dots are labels showing which isotopes will be used for M_{avg} calculation. The deconvoluted average mass is shown in the right panel

5. When all peptides are processed and the Excel file complete, the percentage of deuteration or the number of deuterons can be easily calculated [2] and used for visualization of the results.

3.6 Data Visualization

There are no set guidelines on how to visualize HDX data, and indeed, a multitude of different plots are used throughout the literature. The first result that is usually obtained is a set of uptake plots in which the percentage of deuteration or the number of deuterons is plotted for each peptide as a function of time. However, from such a display, it is difficult to interpret the data in the structural context and also to efficiently handle the information from overlapping peptides. For a more comprehensive view, heat maps or protection plots [38] are most often utilized. These summarize the information into one picture and present it along the protein sequence. However, their basic form does not allow one to visualize information from the overlapping peptides as the data are plotted using only a subset of selected (usually shortest) peptides continuously covering the entire sequence. To also include the redundancy information, peptide midpoints can be calculated [29] for each peptide, which allows the distinction of closely related ones (e.g., 14–25, 14–27, 16–25, 16–27). Upon this modification,

the data can also efficiently be used in mirror (or so-called butterfly) plots or in bar graphs. The disadvantage of this data display is that the proportions of the peptides versus the protein sequence are lost. Probably one of the most attractive ways of data presentation is thus mapping the observed changes onto an existing high-resolution structure or protein model. In this case, differences between two protein states are calculated (preferably for the number of deuterons not for percentages), and the structure is colored according to the observed changes. Visualization tools are usually included in the HDX specific software (DynamX, HDExaminer, etc.,); however, our web-based MSTools suite makes many visualization tools available for vendor-independent use [38].

Moreover, for structure coloring, PyMol with its scripting capabilities is unparalleled. In this workflow, the b-factors of pdb format's fields in an existing protein structure are hijacked and replaced with the numbers from an HDX experiment. To load the data into PyMol, it is necessary to cut each peptide virtually into individual amino acids and use this as an input for coloring.

1. Prepare a table which contains information about each peptide's start, end, time of deuteration (in seconds), and the difference in deuteration between selected states:

Start	End	Time(s)	(TF-DNA—TF)
1	15	20	0.3
1	15	120	0.5
1	15	300	0.6
1	15	1200	0.8
1	15	3600	1.5
1	15	10,800	2.1
8	17	20	0.2
8	17	120	0.4
8	17	300	0.5

Save this table as tab delimited simple text file into a new folder.

2. Copy Python script HDXPeptideSplitter.py (*see* **Note 37**) into this folder and run it (*see* **Note 38**). Command window will appear, and the script will ask for the name of the input file. Then it requires information about the time to be parsed (selected time or all the times in the file) and finally about the conditions to be extracted (again selected one or all). The script then creates a set of files depending on the number of conditions and time points in the experiment. Names of the files are

created using the name of the input file, time, and number of the column in the input file (e.g., FOXO_20_2.txt means 20-s time point and second column from the input file which was named FOXO.txt).

3. Open PyMol and load the pdb file with the structure. Using "copy" (https://pymolwiki.org/index.php/Copy) command, get the desired number of copies corresponding to the number of states and conditions prepared in the previous step. Also name the objects so that the names correspond to the conditions and times from the dataset.
4. Within the PyMol command line, activate script data2bfactor.py (*see* **Note 39**) by the command "*run path/data2bfactor.py*" (where "path" stands for the absolute path to the folder, where the script is saved on your computer), e.g., *run C:\Python27\Scripts\data2bfactor.py*.
5. Reset the b factors to a number that is outside the dataset (e.g., 99). This will later allow you to color the regions that are not covered by the peptides in the dataset. Enter command "*alter object_name, b=99*" (where "object_name" stands for the name of an individual object/structure chosen by the user).
6. Color the structure(s) to some neutral, uniform color using command "color." E.g., "*color grey70.*"
7. Apply your data onto the individual structures. Use command "*data2b_res object_name, path.*" For example, *data2b_res FOXO_20_WT-MUT, C:\Data\FOXO_20_2.txt* (i.e., second column of the input file for HDXPeptideSplitter.py script contained difference between wild-type and mutant). Do this for all structures/objects in PyMol and all input text files.
8. Now select residues with all negative and positive b values, respectively. Enter commands "*select negative, b<0.0*" and "*select positive, b>0.0.*"
9. Color these selections using commands "*spectrum b, selection=negative, minimum=-X, maximum=0.0, palette=blue_white*" and "*spectrum b, selection=positive, minimum=0.0, maximum=X, palette=white_red*," where X is the highest/lowest number in your dataset. Negative (white-to-blue gradient) differences are regions where the exchange was lowered, and positive (white-to-red gradient) are regions with increased deuteration.
10. Finally, differentiate the regions that were not covered by peptides in the dataset. Run command "*color black, (b=99).*"
11. Optionally show a colorbar with the scale using command "*ramp_new colorbar, none, [-X, 0.0, X], [blue, white, red].*"

4 Notes

1. The influence of pH and temperature on the intrinsic chemical exchange rate must be considered here. If two or more states of a protein are followed at different pH values or temperatures, appropriate correction must be applied. Temperature factor can be calculated from the Arrhenius equation (e.g., comparison of exchange at 10 °C and 37 °C will be corrected by a factor of 13.88). The effect of pH is compensated by 10ΔpH value (e.g., exchange at pH 7 runs 100 times faster than exchange at pH 5) [3].
2. Total amount depends on protein size, the LC system used, and the sensitivity of the mass spectrometer. However, tens (small protein) to hundreds (large protein) of micrograms are enough for the comparison of two states and performing triplicate analysis.
3. Check the oligonucleotides for possible secondary structure formation which may affect their binding properties. When preparing/synthesizing or ordering the oligonucleotides, use HPLC purification as the last step to obtain desalted DNA samples.
4. In case where water/ice mixture is used, the valve body should be properly sealed and isolated from the water (parafilm and heat shrink tubing) as it prolongs the lifetime of the valve.
5. A wide range of buffer components or concentrations is tolerated in H/D exchange workflow as long as they can be removed during desalting on a reversed phase resin that precedes separation and mass spectrometric detection. Nevertheless, the buffer components should be kept at their lowest concentration where both the free protein and the complex are still stable.
6. At least ten times K_D [18]. If K_D is not known, try multiple concentrations. Keep in mind that the complex will be diluted during D_2O labeling. Do not use excess of DNA since high DNA concentration can lead to sample precipitation upon quenching.
7. Depending on the melting point temperature of the used dsDNA, cooling might be needed.
8. Try to keep the volume of samples as low as possible to prevent diffusion from the wells during sample loading—ideally under 5 μl.
9. The concentration of the gel should be adjusted based on oligonucleotide length.

10. If possible, cool the electrophoresis apparatus during the run since the complex or DNA duplex might not be stable at higher temperatures.
11. It should be considered that the fragmentation of a particular peptide does not always lead to sufficient number of fragments. Therefore, the combination of different digestion protocols and dissociation techniques usually represents the best solution.
12. Fragments covering the same region of protein and differing in just few amino acids from the N- or C-terminus should result in fairly similar H/D uptake plots. Hence, the slope, level of deuteration, or eventually the differences between the individual studied states should not be vastly different. If they are, this might be an indication of a possible data processing error. On the other hand, if there are multiple individual peptides spanning one protein region and all show highly similar trend, the confidence in the observed changes is significantly increased.
13. Protease column may suffer from autodigestion upon longer storage. Storing it at conditions that do not affect its activity, while preventing autodigestion (e.g., citrate buffer, pH 4–5, 4 °C), is highly recommended. Also, pre-washing with the loading solvent before connecting the column in front of the trap column should be routinely done during an LC system setup.
14. Specific cleaning solutions containing denaturing agents or higher concentration of acids can be designed to minimize carry-over [40]. Good regeneration of the whole system is likely achieved by combining injection of some other (quality control) protein together with these washing steps.
15. Quench conditions (quench solution composition, concentration, quench ratio) should be tuned prior to the experiment using pure buffers. While protein concentration during the labeling step is given by the requirements on the protein–DNA complex formation, the final concentration and sample volume after quenching depend solely on the requirements of the mass spectrometer and LC setup (injection volume, trap, and protease column size + flow). For the MS analysis, it is worth not considering the true detection limits of the instrument, but staying well above them, as the signal intensity after labeling decreases due to signal spreading to multiple isotopes upon deuteration. Keeping the same protein concentration for LC-MS/MS as for the subsequent LC-MS analyses is crucial as the protein:protease ratio affects the digestion pattern.
16. Sum the volumes of injection loop, protease column, trap column, and the connecting tubing and compare it to the selected flow on loading/desalting pump. Make sure that

the sample will reach the trap column and several volumes of the trap column will pass through to achieve efficient desalting. If guanidine or phosphates are used in sample preparation, make the desalting slightly longer. This step usually lasts from 1 to 4 min, depending on the flow and volumes mentioned above.

17. Duration of the gradient and its slope depend on the column and the LC system used. In general, the gradient should be fast, lasting just a few minutes. However, during this short period, the most efficient elution should be achieved. Therefore, gradient running from 5 or 10% acetonitrile (solvent B) and ending at 25–35% acetonitrile should be used. In specific cases, the upper acetonitrile percentage can be raised to 40 or 50% or even higher.
18. For example, here the data are processed in DataAnalysis using Find>Compounds AutoMS(n) feature and exported into a MASCOT generic file (*.mgf). Two different algorithms are used for peak picking (FTMS and SNAP) and thus two *.mgf files are created for each analysis. The whole procedure can be automated using Visual Basic macros in Data Analysis. These macros can be obtained from the authors.
19. If the coverage map shows many gaps and the covered parts are represented by short peptides, the protein was probably over-digested. Faster flow (even in hundreds of microliters per minute) should be tested. One should, however, keep in mind the trapping efficiency of the trap column at a given flow rate. Alternatively, another protease may be used. These steps are also advisable if higher redundancy is desired. In cases where the peptides are too long, temperature should be increased or addition of a denaturing agent considered. Use of alternative aspartic proteases beside pepsin is advantageous for DNA binding proteins since they tend to be rich in basic residues. While pepsin exerts nearly no cleavage after Lys, His, and Arg residues, the other proteases like nepenthesins, aspergillopepsin, and rhizopuspepsin cleave at these sites with high frequency [17, 24, 26]. These proteases may thus yield good spatial resolution of basic proteins or basic protein regions.
20. It is often stated that the use of robotics leads to higher reproducibility; however, trained user and careful pipetting can lead to very similar outcomes. The main advantage in using automation is the ability to perform the analysis in an unattended manner and thus using the instrument time more efficiently. Most importantly, the manual way utilizes freezing of samples and their subsequent thawing. These steps require additional 0.5–2 min, during which the sample undergoes back-exchange or (in the case of low D_2O dilution by the

quench buffer) artificial, nonnative, in-exchange. In contrast, robotics can set up the mixing of H/D exchange samples consecutively, so that once the exchange is finished by the quench solution addition, the sample is directly injected for analysis. It is also of high importance that some proteins tend to precipitate during freezing/thawing procedures and here the use of robotics provides additional advantage.

21. Dilution is often done ten-fold, but higher (20-fold) or lower (five-fold) ratios are also possible. The key consideration here is the stability of the complex at a given concentration which is influenced by the K_D of the complex.

22. The amount/volume of the quench solution is given by the quench ratio. Optimally, ten-fold dilution is used but other ratios, e.g., 1:1 are also acceptable, if necessary. Quench solution must be acidic and strong enough to override pH of the labeling buffer and should make it 2.3–2.5, instead. Diluted solutions of hydrochloric or phosphoric acid, phosphate buffers, or glycine hydrochloride buffers are typical examples. Quench solution may also contain denaturing agents or other components necessary for the solubility of the protein or its analysis. Only components that are not easily removed by reversed phase desalting and are not compatible with ESI-MS should be avoided (e.g., detergents). However, even for nonionic detergents, specific workflows allowing their nearly complete removal under HDX-MS conditions have been described [41].

    The tubes used for freezing should be PCR tubes as they have to be stable at extreme temperature. Normal Eppendorf microcentrifuge tubes can break or take up some liquid nitrogen, which can lead to their explosion during the subsequent thawing procedure.

23. Preferably run H/D exchange with several time points spanning the region from seconds to a couple of hours. For longer incubation times, a control of protein stability should be performed by incubating the protein for the longest selected time in the normal buffer and then deuterium labeling it for the shortest selected period (e.g., 20 s). If the deuteration is the same as for the sample just deuterated for this time window, the protein is probably stable. If the deuteration profile differs/is indicative of protein unfolding, shorter time windows should be sought or the exchange should be done at lower temperature.

24. Simple setup requires two protein states: protein alone and protein in complex with its DNA binding partner. However, number of conditions to be followed can be much larger, e.g., with different protein:DNA ratios or different ligands.

25. Deuterated buffers must be set to pD (not pH). There is a simple correction factor compensating the difference of measuring deuteron concentration using normal pH electrodes $pD = pH_{read} + 0.41$ [42]. Alternatively, pH reading (clearly marked as such) can be reported instead of the pD value.
26. Samples should be analyzed as soon as possible, e.g., within a few days. Do not store them for more than 1 week, even at very low temperatures.
27. Some proteins are not stable enough to allow the preparation of a fully deuterated control. For the majority of comparative HDX-MS experiments, its inclusion is anyway not crucial, but it can offer valuable insight into the back-exchange levels and provides good quality control of the system used.
28. Tune the mass spectrometer properly to avoid hydrogen scrambling in the ion source [43]. That mainly means lower the desolvation temperature and possible ion activation during ion transfer. These low-scrambling parameters may, however, lead to decrease in signal intensity.
29. It is best not to analyze the samples in any ordered manner, e.g., sequentially by the conditions or replicates. If the analysis is stopped or interrupted, recondition the system by running two or three blanks or a standard protein and two blanks. Cleaning procedures should be included, and potential carryover checked during the analysis. The biggest source of carryover is the protease column, which can, however, be washed during each run [40]. Try to keep the system running continuously so that maximal LC reproducibility is maintained.
30. DeutEx is not yet publicly available but will be offered at our website: http://peterslab.org/downloads.php under HDX tools section.
31. It might be beneficial to search for all detected charge states (m/z values) for each peptide, thus not relying only on those fragmented during the LC-MS/MS run and identified during the subsequent search. This will yield more data points which can be either merged and statistically processed or used separately as redundant information supporting the data validity.
32. DeutEx will accept any text form of the input and has automated detection of the content in the columns. If that fails, the user can correct it and assign the column headers manually.
33. For better clarity, a short video demonstrating the whole workflow is available at http://peterslab.org/downloads/SW/DeutEx.mp4.
34. In Data Analysis, which is delivered together with MS instruments from Bruker Daltonics, this procedure can be partially automated using a Visual Basic macro. This macro can be obtained from the authors upon request.

35. mMass is a freely available software suitable for processing of single spectra. It can be obtained from http://mmass.org/.
36. There are also other ways of data processing. Similar to mMass processing is the use of the MagTran program written by Zhongqi Zhang. Also, HX-Express2 (http://www.hxms.com/HXExpress/) can be used for this purpose [32]. The advantage of HX-Express 2 is its ability to detect and dissect bimodal isotopic envelopes that are typical for mixed EX1/EX2 exchange kinetics.
37. Script can be obtained at http://peterslab.org/downloads.php under HDX tools section.
38. To run Python scripts directly, install python on your PC (https://www.python.org/downloads/) and include it in the PATH variable of the operating system (version used here was 2.7.12).
39. Script was written by Robert L. Campbell and is available here: http://pldserver1.biochem.queensu.ca/~rlc/work/pymol/.

Acknowledgments

Czech Science Foundation projects 16-24309S and 16-20860S are gratefully acknowledged. Additional support was obtained from EU/MEYS projects BioCeV (CZ.1.05/1.1.00/02.0109) and NPU II (LQ1604). R.F. also thanks Charles University Grant Agency (project 1618218) and SVV260427/2019.

References

1. Katta V, Chait BT, Carr S (1991) Conformational changes in proteins probed by hydrogen-exchange electrospray-ionization mass spectrometry. Rapid Commun Mass Spectrom 5:214–217
2. Zhang Z, Smith DL (1993) Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. Protein Sci 2:522–531
3. Bai YW, Milne JS, Mayne L et al (1993) Primary structure effects on peptide group hydrogen-exchange. Proteins 17:75–86
4. Engen JR, Wales TE (2015) Analytical aspects of hydrogen exchange mass spectrometry. Annu Rev Anal Chem 8:127–148
5. Oganesyan I, Lento C, Wilson DJ (2018) Contemporary hydrogen deuterium exchange mass spectrometry. Methods 144:27–42
6. Sorokin VA, Gladchenko GO, Valeev VA (1986) DNA protonation at low ionic strength of solution. Die Makromol Chemie 187:1053–1063
7. Ma L, Fitzgerald MC (2003) A new H/D exchange- and mass spectrometry-based method for thermodynamic analysis of protein-DNA interactions. Chem Biol 10:1205–1213
8. Sperry JB, Wilcox JM, Gross ML (2008) Strong anion exchange for studying protein-DNA interactions by H/D exchange mass spectrometry. J Am Soc Mass Spectrom 19:887–890
9. Sperry JB, Shi X, Rempel DL et al (2008) A mass spectrometric approach to the study of DNA-binding proteins: interaction of human TRF2 with telomeric DNA. Biochemistry 47:1797–1807
10. Poliakov A, Jardine P, Prevelige PE (2008) Hydrogen/deuterium exchange on protein solutions containing nucleic acids: utility of

protamine sulfate. Rapid Commun Mass Spectrom 22:2423–2428

11. Roberts VA, Pique ME, Hsu S et al (2012) Combining HD exchange mass spectroscopy and computational docking reveals extended DNA-binding surface on uracil-DNA glycosylase. Nucleic Acids Res 40:6070–6081
12. Graham BW, Tao Y, Dodge KL et al (2016) DNA interactions probed by hydrogen-deuterium exchange (HDX) Fourier transform ion cyclotron resonance mass spectrometry confirm external binding sites on the minichromosomal maintenance (MCM) helicase. J Biol Chem 291:12467–12480
13. Boura E, Silhan J, Herman P et al (2007) Both the N-terminal loop and wing W2 of the forkhead domain of transcription factor Foxo4 are important for DNA binding. J Biol Chem 282:8265–8275
14. Slavata L, Chmelik J, Kavan D et al (2019) MS-based approaches enable the structural characterization of transcription factor/DNA response element complex. Biomol Ther 9: E535
15. Anbanandam A, Albarado DC, Nguyen CT et al (2006) Insights into transcription enhancer factor 1 (TEF-1) activity from the solution structure of the TEA domain. Proc Natl Acad Sci U S A 103:17225–17230
16. Wang L, Pan H, Smith DL (2002) Hydrogen exchange-mass spectrometry. Mol Cell Proteomics 1:132–138
17. Kadek A, Mrazek H, Halada P et al (2014) Aspartic protease nepenthesin-1 as a tool for digestion in hydrogen/deuterium exchange mass spectrometry. Anal Chem 86:4287–4294
18. Kochert BA, Iacob RE, Wales TE et al (2018) Hydrogen-deuterium exchange mass spectrometry to study protein complexes. In: Methods in molecular biology (Clifton, N.J.). Humana Press, New York, NY, pp 153–171
19. Rand KD, Zehl M, Jensen ON et al (2009) Protein hydrogen exchange measured at single-residue resolution by electron transfer dissociation mass spectrometry. Anal Chem 81:5577–5584
20. Mistarz UH, Bellina B, Jensen PF et al (2018) UV Photodissociation mass spectrometry accurately localize sites of backbone Deuteration in peptides. Anal Chem 90:1077–1080
21. Mayne L, Kan ZY, Sevugan Chetty P et al (2011) Many overlapping peptides for protein hydrogen exchange experiments by the fragment separation-mass spectrometry method. J Am Soc Mass Spectrom 22:1898–1905
22. Kan Z-Y, Walters BT, Mayne L et al (2013) Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. Proc Natl Acad Sci U S A 110:16438–16443
23. Cravello L, Lascoux D, Forest E (2003) Use of different proteases working in acidic conditions to improve sequence coverage and resolution in hydrogen/deuterium exchange of large proteins. Rapid Commun Mass Spectrom 17:2387–2393
24. Rey M, Man P, Brandolin G et al (2009) Recombinant immobilized rhizopuspepsin as a new tool for protein digestion in hydrogen/deuterium exchange mass spectrometry. Rapid Commun Mass Spectrom 23:3431–3438
25. Kadek A, Tretyachenko V, Mrazek H et al (2014) Expression and characterization of plant aspartic protease nepenthesin-1 from *Nepenthes gracilis*. Protein Expr Purif 95:121–128
26. Yang M, Hoeppner M, Rey M et al (2015) Recombinant Nepenthesin II for hydrogen/deuterium exchange mass spectrometry. Anal Chem 87:6681–6687
27. Kadek A, Kavan D, Marcoux J et al (2017) Interdomain electron transfer in cellobiose dehydrogenase is governed by surface electrostatics. Biochim Biophys Acta Gen Subj 1861:157–167
28. Moroco JA, Engen JR (2015) Replication in bioanalytical studies with HDX MS: aim as high as possible. Bioanalysis 7:1065–1067
29. Houde D, Berkowitz SA, Engen JR (2011) The utility of hydrogen/deuterium exchange mass spectrometry in biopharmaceutical comparability studies. J Pharm Sci 100:2071–2086
30. Wales TE, Poe JA, Emert-Sedlak L et al (2016) Hydrogen exchange mass spectrometry of related proteins with divergent sequences: a comparative study of HIV-1 Nef allelic variants. J Am Soc Mass Spectrom 27:1048–1061
31. Pascal BD, Willis S, Lauer JL et al (2012) HDXWorkbench: software for the analysis of H/D exchange MS data. J Am Soc Mass Spectrom 23:1512–1521
32. Guttman M, Weis DD, Engen JR et al (2013) Analysis of overlapped and noisy hydrogen/deuterium exchange mass spectra. J Am Soc Mass Spectrom 24:1906–1912
33. Lindner R, Lou X, Reinstein J et al (2014) Hexicon 2: automated processing of hydrogen-deuterium exchange mass spectrometry data with improved deuteration distribution estimation. J Am Soc Mass Spectrom 25:1018–1028
34. Rey M, Sarpe V, Burns KM et al (2014) Mass spec studio for integrative structural biology. Structure 22:1538–1548

35. Kan ZY, Ye X, Skinner JJ et al (2019) ExMS2: an integrated solution for hydrogen-deuterium exchange mass spectrometry data analysis. Anal Chem 91:7474–7481
36. Claesen J, Burzykowski T (2017) Computational methods and challenges in hydrogen/deuterium exchange mass spectrometry. Mass Spectrom Rev 36:649–667
37. Eggertson MJ, Fadgen K, Engen JR et al (2020) Considerations in the analysis of hydrogen exchange mass spectrometry data. Methods Mol Biol 2051:407–435
38. Kavan D, Man P (2011) MSTools - web based application for visualization and presentation of HXMS data. Int J Mass Spectrom 302:53–58
39. Strohalm M, Kavan D, Novak P et al (2010) mMass 3: a cross-platform software environment for precise analysis of mass spectrometric data. Anal Chem 82:4648–4651
40. Majumdar R, Manikwar P, Hickey JM et al (2012) Minimizing carry-over in an online pepsin digestion system used for the H/D exchange mass spectrometric analysis of an IgG1 monoclonal antibody. J Am Soc Mass Spectrom 23:2140–2148
41. Rey M, Mrazek H, Pompach P et al (2010) Effective removal of nonionic detergents in protein mass spectrometry, hydrogen/deuterium exchange, and proteomics. Anal Chem 82:5107–5116
42. Glasoe PK, Long FA (1960) Use of glass electrodes to measure acidities in deuterium oxide 1,2. J Phys Chem 64:188–190
43. Guttman M, Wales TE, Whittington D et al (2016) Tuning a high transmission ion guide to prevent gas-phase proton exchange during H/D exchange MS analysis. J Am Soc Mass Spectrom 27:662–668

Publication III

Slavata, L., Chmelík, J., Kavan, D., **Filandrová, R.**, Fiala, J., Rosůlek, M., Mrázek, H., Kukačka, Z., Vališ, K., Man, P., Miller, M., McIntyre, W., Fabris, D., & Novák, P

Ms-based approaches enable the structural characterization of transcription factor/DNA response element complex

Biomolecules, 9(10), 1–21 (2019)

My contribution: *optimization of HDX-MS conditions, performing part of the HDX-MS experiments*

biomolecules

MDPI

Article

MS-Based Approaches Enable the Structural Characterization of Transcription Factor/DNA Response Element Complex

Lukáš Slavata [1,2], Josef Chmelík [1,2], Daniel Kavan [1,2], Růžena Filandrová [1,2], Jan Fiala [1,2], Michal Rosůlek [1,2], Hynek Mrázek [1], Zdeněk Kukačka [1], Karel Vališ [1], Petr Man [1,2], Michael Miller [3], William McIntyre [3], Daniele Fabris [3] and Petr Novák [1,2,*]

1 Institute of Microbiology, The Czech Academy of Sciences, 14220 Prague, Czech Republic; lukas.slavata@gmail.com (L.S.); chmelik@biomed.cas.cz (J.C.); kuan@maradan.org (D.K.); ruzena.liskova@gmail.com (R.F.); FiJi.cs@seznam.cz (J.F.); rosulek.michal@gmail.com (M.R.); hynek.mrazek@gmail.com (H.M.); zdenek.kukacka@biomed.cas.cz (Z.K.); karel.valis@biomed.cas.cz (K.V.); pman@biomed.cas.cz (P.M.)

2 Faculty of Science, Charles University, 12843 Prague, Czech Republic

3 RNA Institute, University at Albany, State University of New York, Albany, NY 12222, USA; mrmiller@albany.edu (M.M.); wdmcintyre@albany.edu (W.M.); dfabris@albany.edu (D.F.)

* Correspondence: pnovak@biomed.cas.cz; Tel.: +42-0325-873-610

Received: 19 July 2019; Accepted: 24 September 2019; Published: 26 September 2019

Abstract: The limited information available on the structure of complexes involving transcription factors and cognate DNA response elements represents a major obstacle in the quest to understand their mechanism of action at the molecular level. We implemented a concerted structural proteomics approach, which combined hydrogen-deuterium exchange (HDX), quantitative protein-protein and protein-nucleic acid cross-linking (XL), and homology analysis, to model the structure of the complex between the full-length DNA binding domain (DBD) of Forkhead box protein O4 (FOXO4) and its DNA binding element (DBE). The results confirmed that FOXO4-DBD assumes the characteristic forkhead topology shared by these types of transcription factors, but its binding mode differs significantly from those of other members of the family. The results showed that the binding interaction stabilized regions that were rather flexible and disordered in the unbound form. Surprisingly, the conformational effects were not limited only to the interface between bound components, but extended also to distal regions that may be essential to recruiting additional factors to the transcription machinery. In addition to providing valuable new insights into the binding mechanism, this project provided an excellent evaluation of the merits of structural proteomics approaches in the investigation of systems that are not directly amenable to traditional high-resolution techniques.

Keywords: transcription factor; protein; DNA; protein-nucleic acid cross-linking; cross-linking; transplatin; trans-dichlorodiamineplatinum(II); hydrogen-deuterium exchange; FOXO4; molecular modeling

1. Introduction

A key role in the regulatory of expression machinery is covered by transcription factors (TFs), proteins that recognize target DNA sequences called response elements and establish specific interactions with additional factors to activate or inhibit the transcription process [1–4]. The species involved in the process have been unambiguously identified [5,6], but significant information is still lacking on the effects of structure/dynamics on specific recognition and mechanism of action.

The Protein Data Bank contains the high-resolution structures of at least 483 TFs from different species [7], which include less than 10% of all predicted human TFs [6]. Of such structures, only one third also include the cognate DNA response element, and only one fifth are available in both bound and unbound states (Figure1). Due to the size and complexity of such systems, most solved structures do not include the entire TF sequence, but consist almost exclusively of the DNA binding domain (DBD). This fact reflects the modular organization of TFs, which includes discrete domains acting in rather independent manner [8,9]. While DBDs tend to be highly structured, other regions responsible for either modulating transcription activity, or supporting facultative ligand interactions, are rather flexible and assume well-defined conformations only upon binding to the intended factor [10]. The unstructured nature of these regions poses many challenges to conventional high-resolution approaches, which require adequate conformational stability and homogeneity. In most cases, the natural interactions established in vivo, which are responsible for stabilizing well-defined functional conformations, cannot be properly replicated in vitro. These challenges explain the chronic lack of comprehensive information on full-fledged TF structures, which still hampers the elucidation of their mechanism of action at the molecular level.

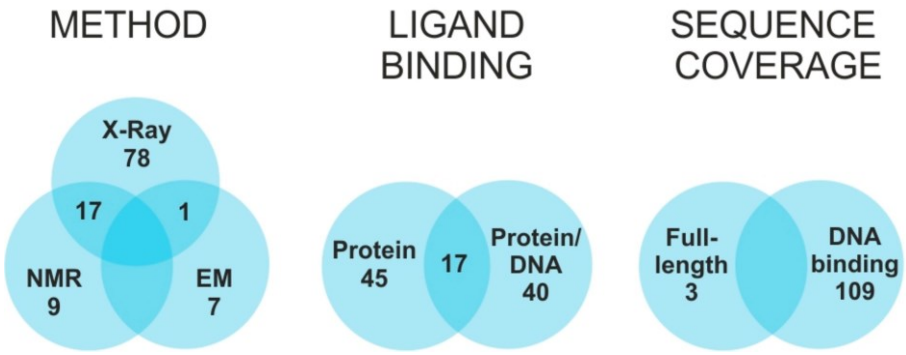


Figure 1. Transcription factors structures—current state: Statistics on high-resolution structures deposited in the Protein Data Bank (80), with consideration of selected methodology, presence of interaction partner, and sequence coverage. To date, the structures of only 112 human TFs have been solved out of a total predicted to be in the 1300–1900 range [6].

Powered by the development of new experimental strategies and mass spectrometric (MS) instrumentation, structural proteomics has rapidly become an essential approach for gathering valuable structural information for species that are not directly amenable to conventional high-resolution techniques [11–14]. In addition to hydrogen-deuterium exchange (HDX) [15], a combination of chemical and biochemical techniques broadly known as MS3D [16–18] have been effectively utilized to identify the regions of contact between bound biomolecules and reveal their mutual spatial organization. In HDX experiments, the exchange rate of backbone amide hydrogens, which is affected by solvent accessibility and possible involvement in hydrogen bonding, can be directly determined by MS analysis. This technique has been broadly employed to study conformational changes [19,20], protein folding [21] and protein-protein interactions [22,23], as well as nucleic acid-protein complexes [24–29]. Among the MS3D techniques, chemical and photo-activated cross-linking (XL) are employed to generate stable covalent bridges between contiguous functional groups, which can reveal their mutual placement in the targeted assembly [13,16]. A variety of bifunctional reagents with different spacing between reactive groups have been developed to determine the distance between susceptible residues. In this way, the sequence position of cross-linked residues and the length of the respective cross-linker provide valid constraints for building accurate molecular models through established computational methods [17,18,30–32]. The excellent versatility of these approaches has prompted the development of reagents capable of targeting functional groups present on protein, as well as nucleic acid substrates [33–39]. Over time, capture tags and isotopic labels have been included in the cross-linker design to facilitate the isolation and analysis of cross-linked products [40,41].

Isotope-labelling, in particular, has provided a valuable tool for highlighting the presence of different conformational states and quantifying their partitioning on the basis of cross-linking probability [13].

In this study, we evaluated the concerted application of complementary structural proteomics techniques to overcome the challenges posed by full-fledged complexes between TFs and respective DNA response elements. We selected a model system consisting of the human transcription factor FOXO4 and its cognate Daf-16 family member-binding element (DBE) [42]. FOXO4 is part of the "O" subfamily of the forkhead box (FOX) class of transcription factors [42–44]. The DBD regions of this class are characterized by winged helix structures, which comprise by approximately 100 amino acids folded into helix-turn-helix motifs and β-sheet-bordered loops that make them resemble butterfly wings [45]. The first high-resolution structure of FOXO4-DBD, which was obtained by NMR spectroscopy, confirmed the presence of a typical forkhead, winged helix fold [46]. At the same time, however, the report indicated that the N- and C- terminal regions of the DBD displayed chemical shifts consistent with highly flexible, disordered structures. The more recent identification of consensus sequences for the FOXO family [47] enabled the crystallization of a complex comprising a selected DBE duplex and a FOXO4-DBD construct that lacked the C-terminal region to facilitate crystal formation [48]. This high-resolution structure provided valuable details on the protein-DNA interaction, but revealed also numerous discrepancies with the binding modes exhibited by other members of the FOXO family [44], which were attributed to possible crystal-packing issues [48]. For these reasons, the FOXO4-DBD•DBE system offered an excellent opportunity for testing the ability of structural proteomics to probe the conformational effects of binding, which would help rectify or corroborate the observed discrepancies. On the other hand, it also afforded sufficient structural information to determine the validity of the new experimental constraints and evaluate the merits of the selected approaches.

The experimental strategies were selected for their ability to provide specific information on a typical protein-DNA complex. For instance, HDX was applied to recognize the regions of the protein affected by DNA binding, either through direct protection of the contact interface, or through allosteric conformational changes involving distal regions of the protein. Quantitative XL was used instead to identify possible variations between free and DNA-bound DBD structures, which would help elucidate the effects of the interaction on overall structure topology. In the case of the DNA component, the fast rate of back-exchange characteristic of nucleic acid hydrogens prevented the application of HDX to recognize the surface of the DBE duplex in direct contact with the DBD. As a possible alternative, we explored the application of transplatin (trans-dichlorodiamineplatinum(II), tPt) to generate protein-DNA cross-links that would help locate the mutual positions of interacting structural features [49]. The spatial constraints afforded by these determinations were combined to guide model-building operations and obtain a full-fledged structure for the complex. The results were compared to the available high-resolution structures to assess possible discrepancies and highlight the new information afforded by the selected techniques. It is necessary to point out, here, that the FOXO4-DBD protein construct used in this study was identical to protein constructs used in previous high-resolution structural studies [46,48]. The outcome clearly demonstrated the benefits of structural proteomics to tackle the elucidation of structure and dynamics in systems that elude established high-resolution techniques.

2. Materials and Methods

2.1. Materials

Non-labelled and isotope-labelled cross-linkers di(N-succinimidyl) glutarate (DSGd0/DSGd4) and di(N-succinimidyl) suberate (DSSd0/DSSd4) were purchased form ProteoChem (Hurricane, UT, USA). Modified protease trypsin (Gold, mass spectrometry grade) was purchased from Promega (Madison, WI, USA). Nuclease Bal-31 was obtained from New England BioLabs (Ipswich, MA, USA). Liquid chromatography solvents of LC/MS grade were purchased from Thermo Fisher Scientific

(Waltham, MA, USA). Other chemicals (highest available purity) were obtained from Sigma-Aldrich. All other chemicals, solvents and buffers for SDS-PAGE were obtained from Bio-Rad Laboratories, Inc. (Hercules, CAUSA). The pET-15b plasmid carrying His-tag, thrombin cleavage site, and FOXO4-DBD (Uniprot ID: P98177-1; residues 86–211) sequences was obtained from Prof. Obsil.

2.2. Design of Oligonucleotides and Duplex DNA Preparation

Both, forward and reverse, complementary oligonucleotide strands (5^0-TTG GGT AAA CAA G-3^0 and 5^0-CTT GTT TAC CCA A-3^0, respectively) were purchased from Integrated DNA Technologies (USA) in standard desalted purity. The reverse oligonucleotide sequence contained DBE (TTG TTT AC) originating from human NOXA promoter localized on chromosome 18 (+ strand) in the position 59,899,552 to 59,899,564 [50]. Both strands were tested for secondary structure presence to avoid interfering structures. Forward and reverse strands were dissolved in water and mixed in an equimolar ratio and then heated up to 95 °C for 1 min. Next, the mixture was let to cool to the room temperature to form the 13 bp duplex DNA.

2.3. FOXO4 Consensus Binding Sequence Determination, Validation and Comparison

The consensus binding sequence of FOXO4 was obtained by searching the HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) v11 [51], which contains the binding models for 680 human transcription factor (TFs). First, we scanned for TF binding models with a Position Count Matrix (PCM) similar to that of FOXO4 by using HOCOMOCO and Matrix CompaRisOn. Approximate P-value Estimation (MACRO-APE) software [52]. We confirmed that the TGTTT consensus sequence was presented in binding models of other FOX factors. Second, we searched for TF binding models with PCM similar to FOXO4 consensus sequence using HOCOMOCO and MACRO-APE to predict TFs with binding mode highly similar to FOXO4.

2.4. Sample Preparation

Full-length DBD (residues 82–207 of the entire FOXO4 sequence) was expressed with an N-terminal His-tag from an appropriate pET-15b plasmid, and then affinity captured on a TALON Superflow Resin (Clontech Laboratories, Mountain View, CA, USA) charged with Co^{2+}. The captured protein was submitted to thrombin digestion to eliminate the tag, followed by gel permeation chromatography. A more detailed description of all experimental procedures is included in the Supplementary Materials. The identity, integrity, and purity of the final sample were verified by MS analysis (vide infra). A duplex DNA construct containing one of the DBE consensus sequences (i.e., TTG TTT AC) [42,53] was obtained by annealing complementary oligonucleotides (i.e., 5^0-TTG GGT AAA CAA G-3^0 forward and 5^0-CTT GTT TAC CCA A-3^0 reverse). Equimolar amounts were mixed and then heated to 95 °C for 1 min. Finally, the sample was let cool to room temperature to form the 13 bp duplex DNA. The desired FOXO4-DBD•DBE complex was obtained by mixing equimolar amounts of protein and DNA samples in 10 mM HEPES buffer with 50 mM NaCl (pH adjusted to 7.4) to a final 33.7 μM concentration, and incubating the mixture at 18 °C for 1 h.

2.5. Product Characterization

Initial stocks of FOXO4-DBD, DBE construct, and FOXO4-DBD•DBE were analyzed to assess sample purity after expression/purification, annealing of the duplex structure, and proper complex formation. The products of XL reaction were also analyzed in the same fashion to assess the incidence of modification. Briefly, each stock was diluted to a final 10-μM concentration by adding a 7.5 mM solution of ammonium acetate (AA) with 50% MeOH (pH 6.85), and then analyzed on a Bruker Daltonics (Billerica, MA, USA) 15T-Solarix XR Fourier transform ion cyclotron resonance (FTICR) mass spectrometer. FOXO4-DBD, as well as FOXO4-DBD•DBE, were analyzed in positive ion mode. Each sample was loaded onto a syringe and introduced into the electrospray ionization (ESI) source at a 2 μL/min flow rate. The FTICR analyzer was calibrated by using a solution of sodium trifluoroacetate

(NaTFA), which afforded a typical 1 ppm accuracy. Mass spectra were acquired over a range of 250–4000 *m*/*z* for 3 min [54].

2.6. Hydrogen-Deuterium Exchange

HDX reactions were performed on 20 μM solutions of either FOXO4-DBD or FOXO4-DBD•DBE complex prepared in an H_2O-based buffer (pH 7.4) containing 10 mM HEPES and 50 mM NaCl. After pre-incubation for an hour at 20 °C, the exchange was initiated by diluting each sample 10-fold into a D_2O-based buffer (pD 7.4) containing 10 mM HEPES and 50 mM NaCl. The reaction was allowed to proceed at 20 °C, while small aliquots containing 100 pmol of protein were taken at predetermined intervals (i.e., 0.33, 2, 5, 10, 30, 60, 180, and 300 min). Quenching was achieved by immediately mixing the aliquot with a 1M glycine/HCl buffer with pH 2.35, and then rapidly freezing the solution in liquid nitrogen. Samples were stored at −80 °C. The analysis was performed by using columns with different immobilized proteases followed by liquid chromatography-mass spectrometry (LC-MS) according to ref. [55]. All experiments were performed in triplicate. A complete description of these procedures is included in the Supplementary Materials (Supplementary Methods) section.

2.7. Quantitative Protein-Protein Cross-Linking

Samples containing 20μM of either FOXO4-DBD or FOXO4-DBD•DBE complex in 10 mM HEPES buffer (pH 7.4) with 50 mM NaCl were pre-incubated for an hour at 20 °C before introducing the cross-linking reagent. Separate samples of FOXO4-DBD were treated with either DSGd0 or DSSd0 in their regular, non-labelled form, whereas FOXO4-DBD•DBE samples were reacted with the deuterium-labelled DSGd4 and DSSd4 versions. The reagents were dissolved in dimethyl-sulfoxide (DMSO) to 6.74 mM concentrations and then added to each substrate to achieve a 10:1 molar ratio. The cross-linking reaction was allowed to proceed undisturbed for 2 h, after which corresponding regular and deuterium-labelled samples (e.g., treated with DSSd0 and DSSd4) were mixed in a 1:1 ratio to enable quantification. In parallel, control samples were also examined, which were treated with pure DMSO lacking cross-linker, or matching cross-linker mixtures with a 1:1 ratio of either DSGd0/DSGd4 or DSSd0/DSSd4. All reactions solutions were analyzed by SDS-PAGE to check whether cross-linking had produced any unwanted non-specific higher-order aggregates, or prevented a sufficient degree of digestion necessary to enable subsequent analysis. Characterization of cross-linked conjugates was achieved according to a bottom-up approach that employed trypsin digestion followed by LC-MS analysis [13]. All experiments were performed as triplicate. A complete description is included in the Supplementary Materials.

2.8. Protein-DNA Cross-Linking

Samples containing 25μM of either FOXO4-DBD or FOXO4-DBD•DBE complex in 150 mM ammonium acetate (pH 6.85) were treated with a 1 mM solution of trans-platinum(II)diammine dichloride (transplatin, tPt), which had been pre-incubated for an hour at 18 °C. Reaction mixtures containing a final 200 μM concentration of transplatin and 20 μM of protein/complex were incubated at 18 °C for 14 h. In parallel, control samples devoid of transplatin were prepared at the same time in the same manner. Reaction and control samples were analyzed by native and denaturing DNA polyacrylamide gel electrophoresis, SDS polyacrylamide gel electrophoresis, and mass spectrometry. Characterization of cross-linked conjugates was achieved according to a bottom up approach that involved treatment with Bal-31 nuclease and trypsin to digest DNA and protein components, respectively, followed by LC-MS/MS analysis with data-independent acquisition with broad isolation window. A complete description is included in the Supplementary Materials.

2.9. Data Processing and Interpretation

The SNAP 2.0 algorithm of the DataAnalysis 4.2 (Bruker Daltonics, Billerica, MA, USA) software package was utilized to generate deconvoluted spectra and lists of monoisotopic masses from the

acquired MS data. The MASCOT 2.2 search engine was used to search MS/MS data and achieve the identification of on-line digest products from a theoretical library of digestion products. Deuteration rate was determined by using the home-built Deutex software (unpublished). The home-built LinX software (available online) and Stavrox (v. 3.6.0.1 by Michael Götze) software was then used to compare the experimental data with a library of theoretical cross-linking products to correctly identify the sought-after conjugates. The proportion of labelled versus un-labelled species was determined by applying mMass 5.4.1 [56] to the signals of such conjugates. In the case of peptide-DNA conjugates, deconvoluted spectra and monoisotopic masses were calculated by using a constant unit to mimic the presence of a certain oligonucleotide cross-linked by a transplatin equivalent, which were subsequently searched by LinX. A complete description of these procedures is included in the Supplementary Materials.

2.10. Molecular Modeling

Initially, six different FOXO4 sequentially related structures (DBD or DBD•DNA complex) were used as templates for homology modeling of protein part (Modeller software [57]). Such templates, however, covered only the 101–176 residues of the DBD sequence, whereas our selected target covered the 82–207 section including the additional flanking sequences that had previously eluded structural elucidation. The HADDOCK [58] program was utilized to perform docking experiments between FOXO4-DBD and DBE substrate, where were meant to rationalize the HDX data. The docking experiments utilized a duplex substrate exhibiting an ideal B-DNA conformation, which was generated by the make-na server (http://structure.usc.edu/make-na/). The WeNMR/WestLife infrastructure [59] was used to carry out the computationally intensive docking calculations. The first structure of the cluster of the FOXO4-DBD•DBE complex, which displayed the best HADDOCK score for each run, was used for the subsequent modeling operations. The missing residues in structures of DBD and DBD•DNA were added using the Modeller incorporating distance restraints derived from XL distances. To search the possible conformational space of flanking residues a restrained simulated annealing protocol in the torsion angle space was accomplished in CNS [60]. An ensemble of 50 structures was calculated for each starting model. During simulations, the coordinates afforded by the initial PDB templates and DBE structure were kept fixed, the XL distances (calculated with consideration of the spacer arms lengths [61–63] and space requirement of side chains) were used as distance restraints. The resulting models were visualized by using Pymol [64].

2.11. Data Availability and Software

- MSTools package—available at http: //peterslab.org/MSTools [65]
- LinX—available at http: //peterslab.org/MSTools
- Stavrox (v. 3.6.0.1 by Michael Götze)—available at http: //www.stavrox.com/
- DeutEx—In-house developed program DeutEx is based on a Tcl macro. It requires protein sequence, list of identified peptides from search engines such as MASCOT or PEAKS. Basic overview of the workflow shown on unrelated example data can be found here: http: //peterslab.org/downloads/SW/DeutEx.mp4
- Mass spectrometry data available at https://www.ebi.ac.uk/pride/archive/ (Project accession: PXD013969)

3. Results and Discussion

The crystal structure available for the FOXO4-DBD•DBE complex does not cover the entire sequence of the DNA binding domain [48], which spans only the 82–207 section of FOXO4 and omits flanking regions that have been hypothesized to promote the recruiting of additional components of the transcription machinery. At the same time, the NMR structure of full-length FOXO4-DBD provides limited information on the G_{138}–A_{144}, E_{166}–K_{170}, and the N- and C-terminal regions, which were described as rather flexible and disordered in solution [46]. Although DNA binding has been

credited with stabilizing at least some of these regions, the structure of the bound form still displayed significant discrepancies with those of homologous members of the FOXO family [44], which were possibly caused by crystal packing [48]. For this reason, we investigated such discrepancies by implementing biochemical approaches to probe the effects of ligand binding directly in solution. The study employed recombinant full-length DBD (i.e., residues G_{82}–A_{207}) and a duplex DNA construct containing the 5^0-TAC CCA A-3^0 consensus sequence, which was obtained by annealing commercial oligo-deoxyribonucleotides. As shown in Figure S1A and Figure S2 of Supplementary Materials, mixing equimolar amounts of protein and duplex DNA provided the expected 1:1 species corresponding to the desired FOXO4-DBD•DBE complex. These samples were submitted to hydrogen-deuterium exchange, quantitative protein-protein cross-linking, protein-DNA cross-linking, and docking experiments to obtain complementary information on their mutual interactions and spatial arrangement. The results were compared to those obtained from the individual FOXO4-DBD protein to investigate the effects induced by specific DNA binding.

3.1. Combined Online Digestion by Pepsin and Nepenthesin I Improves the HDX Resolution

We initially pursued the identification of the regions of FOXO4-DBD, which were making direct contact with the DBE duplex, or were subjected to detectable microenvironment variations upon binding. Hydrogen-deuterium exchange (HDX) was performed on both free and bound forms of full-length FOXO4-DBD. The determinations followed a well-established protocol in which the exchange process was stopped at predetermined intervals to monitor the rate of exchange. Since hydrogen-deuterium exchange has not yet been widely used to study the complex of protein and duplex DNA, we first tuned conditions for on-line digestion in order to obtain the best spatial resolution. It was achieved using the combination of nepenthesin I and pepsin, where rather small and overlapping peptides were observed. The robustness of the setup was approved by multiple injections. The final analysis was carried out at low pH and temperature to minimize back-exchange. The protocol included protein digestion in consecutive on-line columns containing immobilized proteases (i.e., pepsin and nepenthesin-1), followed by LC-MS/MS and LC-MS analysis of digested peptides (see Materials and Methods and Supplementary Materials) [66]. Since HDX has not been widely used to study complexes comprising protein and duplex DNA, the conditions for on-line digestion required fine-tuning to obtain the best possible spatial resolution. This task was accomplished by combining nepenthesin I with pepsin to obtain rather small, overlapping peptides. Multiple analysis were carried out to evaluate the robustness of this approach. The ensuing peptide map demonstrated that the procedure afforded full coverage of the FOXO4-DBD sequence (Figure2).

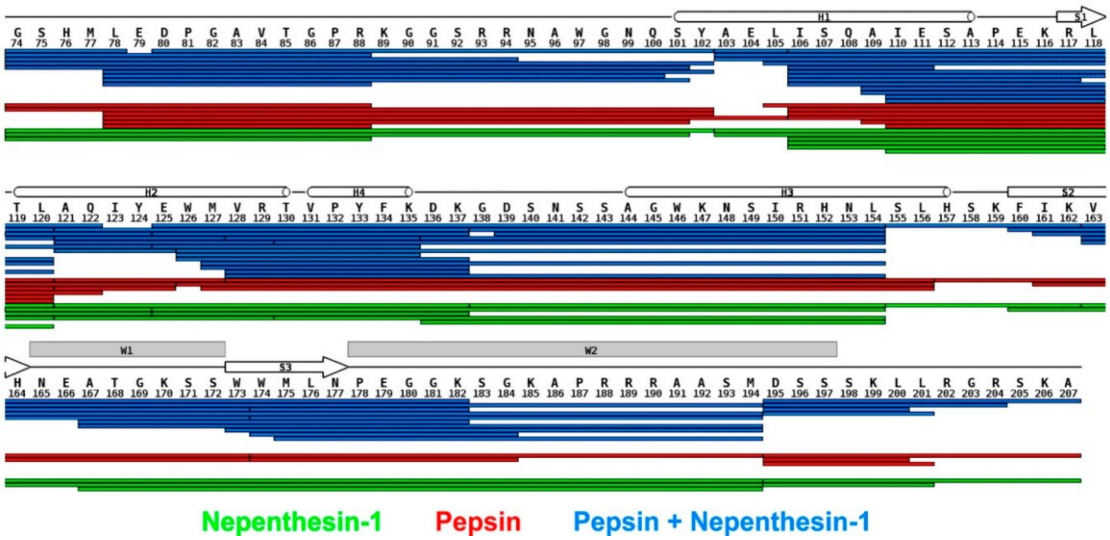


Figure 2. Combined online digestion by pepsin and nepenthesin I improves the HDX resolution. Comparison of peptide maps obtained by on-line proteolysis with nepenthesin-1 (green), pepsin (red), or their combination (blue) is projected on the FOXO4 sequence. In all cases, full sequence coverage was reached but pepsin/nepenthesin-1 digestion provided the highest redundancy and spatial resolution. Secondary structure elements are depicted above the sequence. The N-terminal region G74–P87 originates from the production plasmid and thus is not a part of the wild-type FOXO4 sequence (see also Figure S33 of Supplementary Materials).

3.2. HDX Identified the Interaction Interface and Revealed Long-Distance Structure Stabilization

For each digestion product, a relative deuteration rate was calculated by considering the number of hydrogens exchanged with deuterium atoms against the total number of exchangeable amide hydrogens in the peptide. Relative deuteration rates versus exchange time were calculated at both the peptide and amino acid levels to recognize possible variations between free and bound FOXO4-DBD (see Figures S3–S6 of Supplementary Materials, respectively). This task was facilitated by calculating actual differences for each amino acid in the sequence, which were visualized in a 3D model of FOXO4-DBD•DBE by using an appropriate color palette (Figure S7 of Supplementary Materials).

The results are summarized in an HDX difference plot that provided a comprehensive view of the variations of solvent accessibility induced by the specific interactions between FOXO4-DBD and its cognate DBE duplex (Figure3). Starting from the N-terminus, the G 74–Y102 region displayed relatively high levels of deuteration, regardless of reaction time, with no significant differences between free and bound forms. These observations indicated that this region was rather exposed and capable of exchanging freely with the solvent in both forms. In contrast, the next section spanning the A103–T130 residues displayed the most extensive differences in deuteration rates, which increased significantly as a function of time. This sequence folds helix H1 and H2, strand S1, and intervening loops (see topology annotation in Figure2). According to the crystal structure, none of these distinctive features is supposed to make direct contact with the duplex DNA [48], which would help explain the drop in deuteration by invoking a simple protection effect. In the absence of direct contact, the observed loss of solvent accessibility must be attributed to indirect conformational effects induced by binding. The fact that the difference in deuteration levels increased gradually with time and stabilized after 30 min suggests that, in the free form, this set of secondary structures may undergo slow mutual dynamics that delay the exchange of susceptible hydrogens. In the bound form, such dynamics may be stabilized by interactions with contiguous structures that, in turn, make direct contact with the DNA ligand. Like falling dominoes, a series of relatively minor conformational variations linked together may ultimately induce observable inhibition of the exchange reaction. This long-distance effect is clearly evident, for example, in the relative deuteration plot of peptide A103–L118 (2–3), which shows

increasing uptake in the free FOXO4-DBD as a function of time, but constant low-level deuteration in the bound form (Figure S3 of Supplementary Materials).

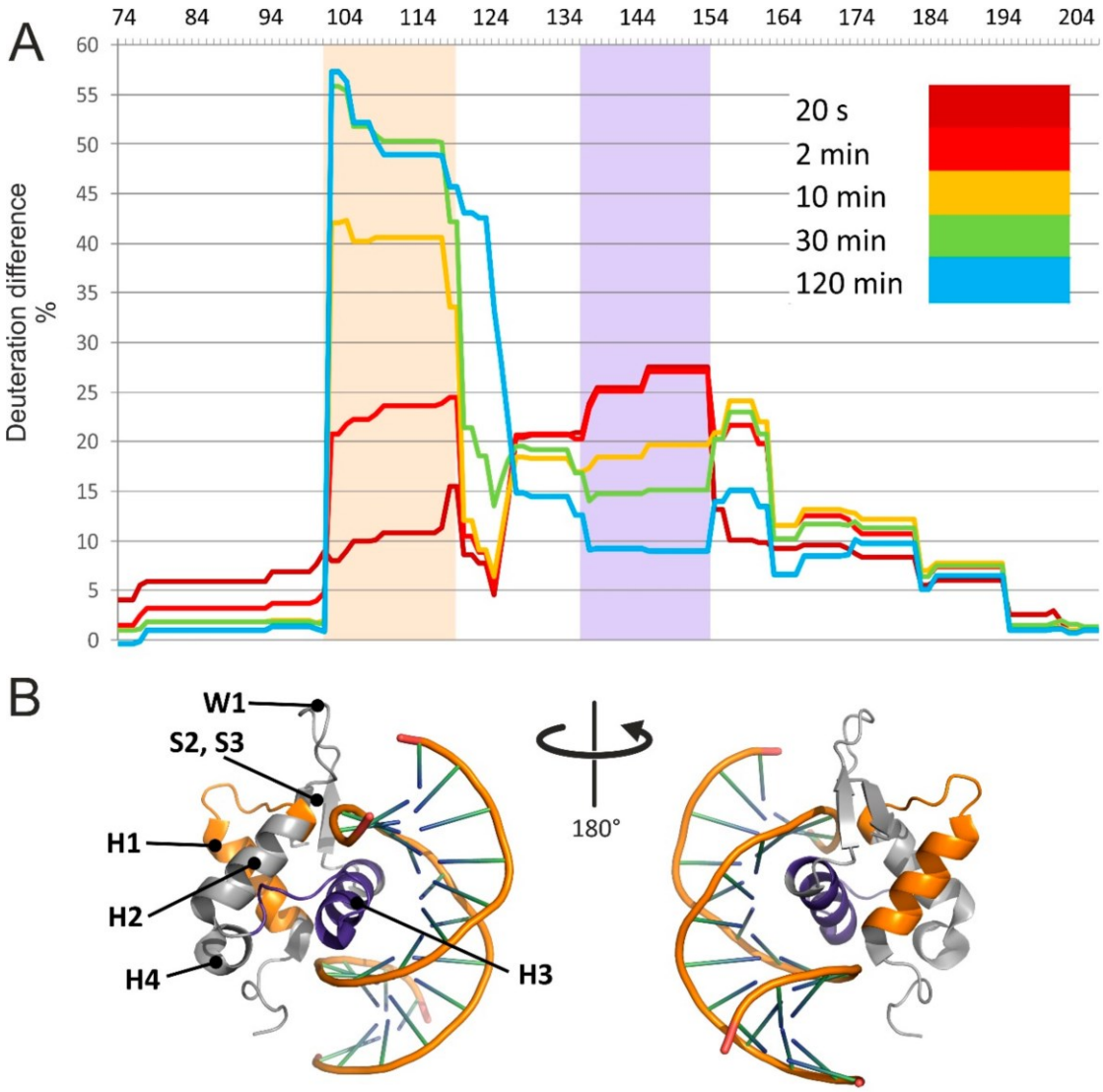


Figure 3. HDX identified the interaction interface and long-distance stabilization of the protein structure. (**A**) Relative deuteration differences [DR(FOXO4-DBD) – DR(FOXO4-DBD•DBE)] plotted along the FOXO4-DBD sequence and their evolution in time. Highlighted areas show two regions with large differences in deuteration levels. (**B**) FOXO4-DBD•DBE structure with highlighted regions showing significant differences in deuteration. Please note that the G_{74}–P_{87} region of the construct was contributed by the recombinant-production vector and, thus, was not part of the wild-type FOXO4 sequence (see also Figures S3–S7 of Supplementary Materials).

The next region, spanning the V_{131}–K_{159} residues, also manifested significant differences between free and bound forms, but their time dependence displayed a rapid increase at shorter intervals, followed by a decline near initial levels at longer reaction times (Figure3A). In particular, the residues forming helix H3, which the crystal structure places directly in the major groove of the DBE duplex [48], experienced the largest differences. For this reason, direct steric protection induced by bound DNA could explain the uptake inhibition observed for such residues. In contrast, the outcome observed for the contiguous helix H4 and intervening loop could indirectly result from the stabilization of the H3 conformation, which could constrain the placement of such residues and restrict their solvent accessibility. The crystal structure also identified a handful of contacts that were mediated by water

molecules trapped in the binding interface. While these interactions have been suggested to further stabilize the dynamics of helix H3 and flanking regions, it is not clear how trapping D_2O versus H_2O present in the solvent may affect the observed deuteration rates. It should be also pointed out that FOXO4 differs from other FOXO homologues by the insertion of five amino acids (K_{137}–N_{141}) in the H4–H3 loop (Figure2). It is not clear whether this insertion may be responsible for the unusual conformation assumed by helix H3 in the bound form, which differs from those assumed in other members of the family [48]. In any case, the observed inhibition pattern supports a role in stabilizing the fold of FOXO4-DBD and reducing its overall flexibility upon complex formation.

The F_{160}–D_{195} section corresponds to sequences located on the C-terminal side of helix H3, which manifested smaller but still perceptible differences of deuteration rates. This region contains strands S2 and S3, as well as the W1 and W2 wings, which the crystal structure placed far removed from the DNA binding interface. Also in this case, an overall decrease of structural flexibility upon binding could explain the reduced deuterium uptake. It should be noted that, in addition to conferring FOXO4-DBD its winged look and fine tuning of the interaction with DNA [47], W1 and W2 could contribute to constitute possible regions of contact for auxiliary components of the transcription complex. The final section covered by HDX determinations consisted of the S_{196}–A_{207} sequence and displayed marginal deuteration differences. This observation indicated the absence of any protection or conformational effects induced by DNA binding.

3.3. Protein-DNA Cross-Linking Revealed the Mutual Placement of Protein and DNA Components

The HDX experiments identified the regions affected directly and indirectly by DNA binding, which experienced clearly detectable variations of solvent accessibility. However, these types of determinations could not identify the structures responsible for limiting the access of solvent to a specific region. In other words, these experiments could not reveal the mutual spatial relationships between such structures, nor assess the effects of binding on such relationships. For this reason, we employed different types of cross-linking strategies to probe the organization of the various structures and recognize their mutual placement in the overall fold. The first approach employed transplatin to generate putative protein-DNA conjugates that may be capable of constraining the position of the DBE ligand onto the FOXO4-DBD substrate. The reactivity of platinum compounds towards specific functional groups of nucleic acids is well documented [67] and involves the preferential attack of the N7 position of guanine base [68]. Although the characteristics of their reactivity towards protein residues are still unclear [49], amino acids with electron-rich S, N and O atoms, such as Cys, Met, His, and Thr, have been described as preferred targets [69]. We treated samples of FOXO4-DBD•DBE complex, as well as free FOXO4-DBD and DBD, with a 10:1 transplatin to substrate molar ratio in 150 mM ammonium acetate (pH 6.85) and incubated at 18 °C for 14 h (see Materials and Methods and Supplementary Materials). The sample mixtures were analyzed by both ESI-MS and SDS-PAGE to assess the distribution of transplatin adducts to estimate the proportion of sought-after intermolecular cross-links. The representative data in Figure4A provides a view of the typical product distributions obtained from these probing reactions, which included monofunctional "dangling" adducts containing a still unreacted chloride function (i.e., marked as tPt-Cl adducts), as well as bifunctional conjugates in which both functions had effectively reacted (i.e., marked as tPt adducts).

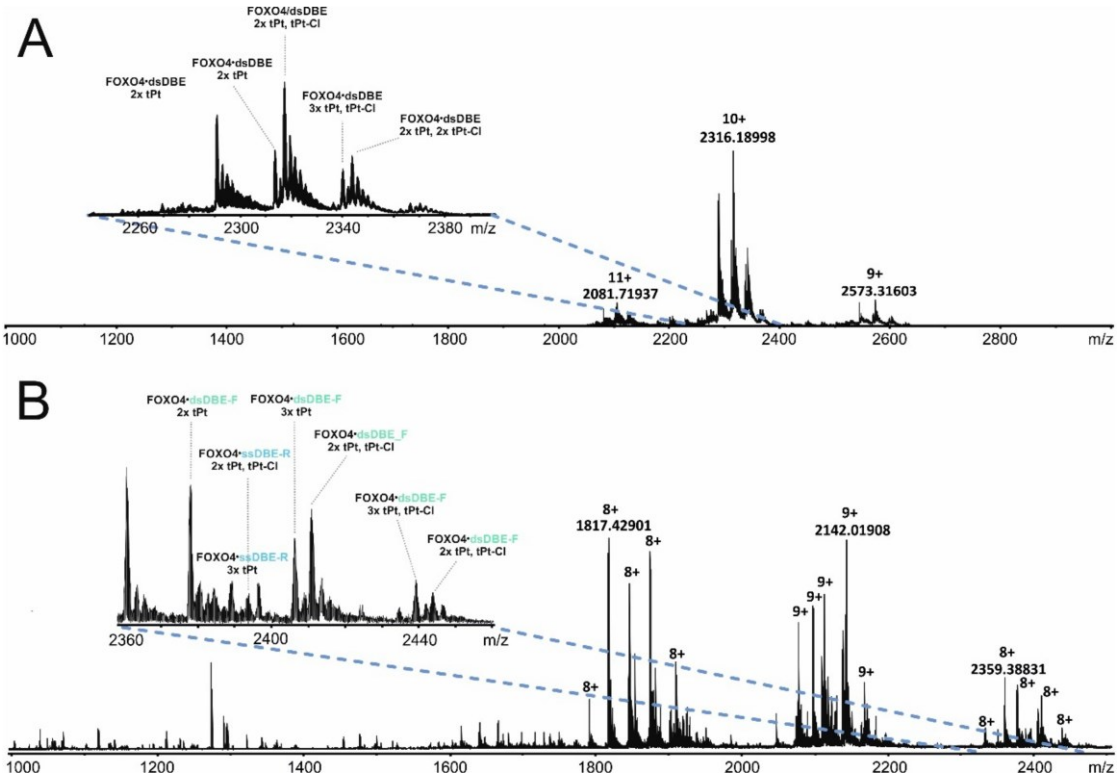


Figure 4. Mass spectra of transplatin-treated FOXO4-DBD•DBE samples, which were obtained under either (**A**) non-denaturing (i.e., 1% acetic acid in 7.5 mM ammonium acetate) or (**B**) denaturing (i.e., 1% acetic acid and 50% methanol in 7.5 mM ammonium acetate solution) conditions. Abbreviations: tPt—bifunctional transplatin conjugate ($Pt(NH2)_2$); tPt-Cl—monofunctional adduct ($Pt(NH2)_2Cl$); dsDBD—duplex DNA; ssDBD-F—forward oligonucleotide strand; ssDBD-R—reverse oligonucleotide strand. The FOXO4-DBD sample employed in these experiments lacked three amino acids at the C-terminus (see also Figure S1, Figure S2, Figure S8, Figure S9 of Supplementary Materials.

Based on mass alone, it is not typically possible to distinguish the desired intermolecular conjugates from intramolecular crosslinks, which share the same elemental composition. For this reason, ESI-MS analysis was repeated in the presence of 50% methanol to achieve mild denaturing conditions. In this way, products stabilized by bridging bifunctional cross-links were still detected intact, such as the conjugates containing FOXO4-DBD and individual DBE-F or DBE-R strands. In contrast, no signal was observed for complexes devoid of any intermolecular conjugation, whereas adducts of their free unbound components were individually detected, such as those of FOXO4-DBD protein, DBE-F, and DBE-R strand (Figure4B). In analogous fashion, the reaction mixtures were also analyzed by PAGE under both native and denaturing conditions (Figure S1 of Supplementary Materials). Direct data comparison enabled the identification of bands that eluded dissociation by elevated concentrations of SDS or urea and, thus, could be attributed to the presence of bridging bifunctional crosslinks. These data enabled us to estimate that the desired intermolecular cross-links amounted to less than 5% of the total material submitted to transplatin reaction.

A classic bottom-up strategy was carried out to complete the characterization of cross-linked products, which included digesting the material with protein- and nucleic acid-specific enzymes to obtain samples amenable to LC-MS and LC-MS/MS analysis. In particular, reaction mixtures were treated with trypsin to map the position of peptides conjugated to DNA strands (see Materials and Methods and Supplementary Materials). In subsequent experiments, the size of the oligonucleotide moieties was reduced by treatment with Bal-31 nuclease to facilitate analysis. The representative data in Figure S8 of Supplementary Materials illustrates the challenges faced by the MS/MS analysis of

these types of hetero-conjugates. Upon gas-phase activation, a precursor ion consisting of G_{74}–R_{88} cross-linked to the DBE-R strand underwent dissociation around the bridging Pt atom, rather than along the backbones of the bridged moieties. The absence of sequence information afforded by this type of fragmentation prevented the identification of the actual residues involved in the cross-linking reaction. Nevertheless, the identity of the conjugated components still represented valuable information on the mutual relationships between contiguous regions (summarized in Figure S9 of Supplementary Materials).

The detected peptide-oligonucleotide conjugates were examined in the context of the results afforded by the HDX determinations and other structural information available for the system. For instance, peptide N_{148}–K_{159} spanning helix H3 was found conjugated to the forward strand of DBE, consistent with the placement of H3 directly into the major groove of the DNA duplex in the crystal structure [48]. This finding agreed also with the prominent protection effects observed in this region during HDX experiments (Figure3A). Peptide F_{160}–K_{170}, covering the end of S2 and beginning of W1, was also cross-linked to the forward strand of DBE, despite the absence of any direct contact in the crystal structure. In this peptide, reactivity and orientation considerations would point towards H164 and T168 as possible conjugation sites, if their distances from susceptible DNA structures were sufficiently favorable. In this direction, the HDX data indicated that this region experienced a detectable decline in deuterium uptake consistent with the adoption of a rather constrained conformation upon DNA binding (Figure3A). The new conformation could place susceptible groups within mutual striking distance, thus promoting the formation of the observed cross-linked product. A similar explanation is applicable also to the S_{171}–K_{182} peptide spanning the end of W1, beginning of W2, and intervening S3 strand, which formed cross-links with both forward and reverse strands of DBE. Also, this region experienced a significant decrease of deuterium exchange upon binding, which was not explainable by direct steric protection, but rather by indirect conformational effects transmitted through contiguous structures. The crystal structure orients S_{171} and S_{172} to face the minor groove of the duplex construct, which would represent prime positions for promoting conjugation with either strand. Also in this case, the respective functional groups could be placed within striking distance by the more constrained conformation revealed by HDX experiments. The remaining products consisted of the G_{74}–R_{88} peptide cross-linked to either the forward or reverse strand. These products address the flexibility of the N-terminal loop, which was supported by the lack of any significant variation of deuteration patterns reported by the HDX experiments.

3.4. DNA Binding Induced Significant Effects on Protein Conformation

The observed protein-DNA cross-links provided valuable information not only on the reciprocal positions of protein and DNA components, but also on the significant changes induced by binding on the initial protein conformation. We employed protein-specific reagents to evaluate the extent of such variations and enable a better appreciation of indirect conformational effects. Our quantitative crosslinking approach involved the concerted application of the homobifunctional reagents DSGd0/d4 and DSSd0/d4 [di(N-succinimidyl) suberate and di(N-succinimidyl) glutarate] to bridge susceptible amino or hydroxy groups that may be respectively placed within 20.5 ± 3.0 or 24.2 ± 3.0 Å of one another (see Supplementary Materials). The utilization of reagents with different bridging spans provided the ability to determine an average distance between residues. At the same time, the isotopic labels facilitated the identification of cross-linked products in complex digestion mixtures from their characteristic 4-Da spacing and enabled the acquisition of unbiased quantitative data on the incidence of cross-linking in the free or bound FOXO4-DBD. Initially, separate samples were treated with 1:1 mixtures of matching unlabeled/labelled reagents of the same length to complete a survey of the regions susceptible to cross-linking (see section Materials and Methods and Supplementary Materials). A total of 39 conjugates were identified, which bridged lysine and serine residues, as well as the N-terminal amino group (Table S1, Figures S10–S26 of Supplementary Materials). The majority of them were detected in matching pairs generated by reagents of either length, and were observed in both

free and bound samples. However, a small portion was unique to just one form and/or cross-linker length. Next, individual aliquots of free FOXO4-DBD were treated with either DSGd0 or DSSd0, whereas those of FOXO4-DBD•DBE complex were separately treated with either DSGd4 or DSSd4 (see Materials and Methods and Supplementary Materials). Corresponding samples treated with the same unlabeled/labelled reagent were mixed in a 1:1 molar ratio prior to protease digestion and analysis to compare the incidence of each conjugate in either free or bound samples. The proportion of each of the 39 conjugates identified earlier was determined for at least one of the cross-linker lengths, as summarized in (Table S2 of Supplementary Materials).

A close examination of the results revealed distinctive cross-linking patterns associated with the presence of bound DNA. In particular, the incidence of some conjugates decreased significantly upon binding, while others increased. Among the former, the DGS conjugates bridging K_{147} to either K_{170} or N-term dropped from 95.9% and 92.8% to 4.1% and 7.2%, respectively (Table S2 of Supplementary Materials). The cross-linking inhibition manifested by K_{147} cannot be merely ascribed to its location on helix H3, in direct contact with the duplex construct, because this residue was still capable of supporting conjugation with both K_{135} and K_{137}. A more plausible explanation could be that DNA binding forced K_{147} out of the reach of either K_{170} on helix H2, or K_{162} on the adjacent S2 region. The limited nature of such conformational changes was revealed by the fact that K_{147} was pushed out of N-term's reach for the shorter DSG reagent, but was still sufficiently close for the longer DSS, with incidence of cross-linking dropping from 92.8 to 7.2% and from 72.7 to 27.3%, respectively (see Table S2 of Supplementary Materials). The limited extent of these changes was also evident in the subtler cross-linking variations between K_{147} and either K_{135} or K_{137} located on the H3-H4 intervening loop.

In other cases, DNA binding increased the incidence of specific conjugates by placing residues within mutual striking distance in the complex, which were marginally susceptible or inert in the free protein. For example, the conjugates bridging residue K_{182} with N-term, K_{89}, K_{116}, K_{159}, K_{162}, or K_{182} were greatly enhanced by the presence of DNA duplex (Table S2 of Supplementary Materials). For the majority of these positions, the levels of cross-linking observed with the longer DSS reagent displayed more significant variations than those with the shorter DSG. Considering that K_{182} is located on the W2 wing region, these observations offered further evidence of the long-range conformational effects of DNA binding suggested by the results of HDX and protein-DNA cross-linking experiments. Another example was provided by the numerous conjugates involving the N-term, which suggested that DNA binding had prominent stabilizing effects on a region that was rather flexible in free FOXO4-DBD. Consistent with the HDX data, the variations of cross-linking patterns confirmed that DNA binding induced significant effects on protein conformation not only within the contact interface, but also in rather distal positions.

3.5. Structural Proteomics Could Effectively Guide Model-Building Operations to Produce Very High-Quality 3D Models

The HDX and XL experiments provided a wealth of information that was used to guide the molecular modelling of a full-fledged FOXO4-DBD and FOXO4-DBD•DBE complex. Our approach took advantage of available high-resolution structures that, although incomplete in their coverage of the protein sequence, still represented excellent templates for homology modelling operations (see Materials and Methods and Supplementary Materials). In particular, the templates were used to obtain the coordinates of what could be defined as the structured core of the complex, a region spanning approximately from R_{93} to N_{177}, which displayed limited discrepancies across the available structures. In contrast, the regions that were either absent from the templates, or displayed significant variations, or had been predicted to possess a high degree of flexibility by PSIPRED [70], were modeled according to the HDX and cross-linking information (see Materials and Methods and Supplementary Materials). These regions corresponded to the G_{74}–Q_{100} and N_{177}–A_{207} sections located respectively at the N- and C-terminal ends of the DBD sequence. These operations were performed in the Modeller suite [57], which was also used to eliminate possible strains and steric

clashes introduced during model building. The program applied the DOPE scoring algorithm to identify the best possible structures that were subsequently employed in docking and simulated annealing procedures. The former was carried out to place the DBE structure, which was created separately by using the make-na server (http://structure.usc.edu/make-na/), onto the putative binding site of the protein. This operation was accomplished in HADDOCK [58] by designating as active those residues that had experienced reduced rates of exchange upon DNA binding (Figure3A and Figure S27 of Supplementary Materials). The mutual positioning between the DBE and FOXO4-DBD components was further refined according to the results of the protein-DNA cross-linking experiments, which were introduced by using Modeller. Finally, the structures of both FOXO4-DBD and FOXO4-DBD•DBE complex were submitted to simulated annealing and energy minimization in CNS [60] to generate the sought-after model ensembles (Figure5).

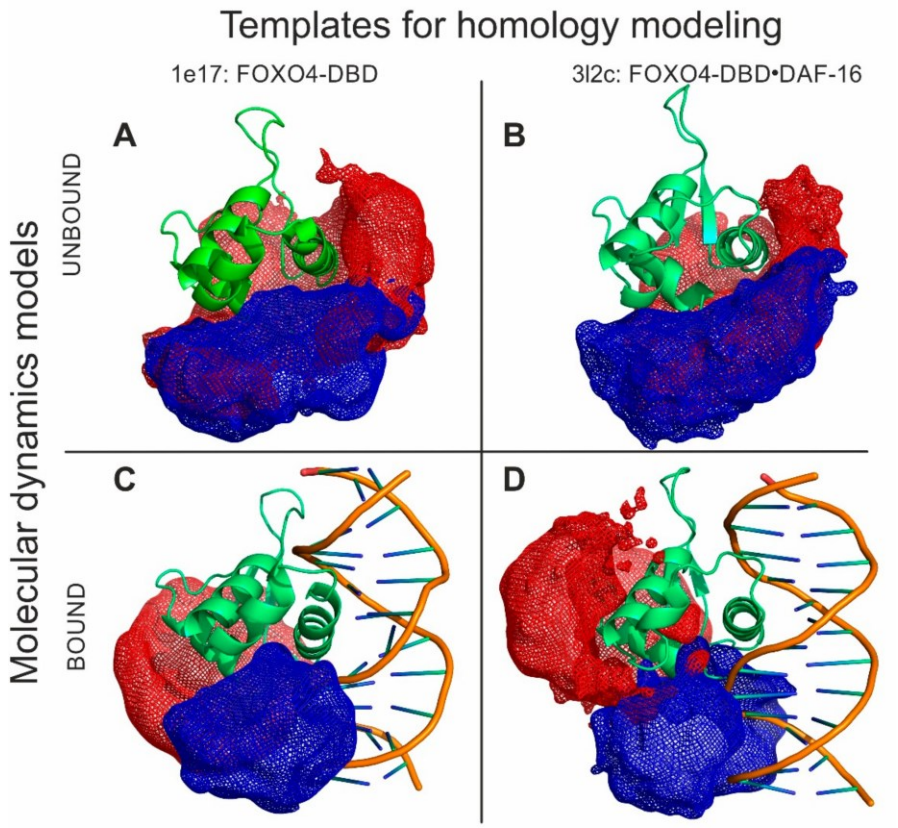


Figure 5. Structural proteomics could effectively guide model-building operations to produce very high-quality 3D models. Models of FOXO4-DBD and FOXO4-DBD•DBE were obtained by combining homology modelling with experimental constraints and molecular dynamics simulations. These models incorporated extensive information from protein-DNA cross-links, quantitative protein-protein cross-links, and hydrogen-deuterium exchange. The green structures show representative models for unbound (**A**,**B**) and bound (**C**,**D**) forms based on corresponding 1E17 (**A**,**C**) or 3L2C (**B**,**D**) high-resolution templates. Mesh areas in blue and red colors represent spaces occupied by all the models in the ensembles, which provided a measure of the flexibility of the N- and C- terminal regions (see also Figures S10–S32 and Tables S1 and S2 of Supplementary Materials).

The structures obtained for FOXO4-DBD and FOXO4-DBD•DBE complex were compared to the corresponding high-resolution structures to assess the robustness of our structural proteomics approach. In the case of individual FOXO4-DBD, the overall topology of the ensemble reflected the typical forkhead structure of the FOXO family and matched very closely that of the NMR structure used

as homology template (see Figure S28A,B of Supplementary Materials), thus supporting the validity of the HDX and cross-linking constraints. A more detailed comparison was obtained by calculating the root mean square deviation (RMSD) between the coordinates of corresponding heavy atoms located in the backbone of each ensemble model and the various templates. The representative plot in Figure S29 of Supplementary Materials, for example, shows that the model obtained from the 1e17 structure deviated very little from the initial template. The fact that the experimental constraints introduced during modelling did not force any significant variation of the initial coordinates indicates that the probing operations did not cause any perturbation of the substrate's 3D structure and corroborated the excellent stability of the structured core of FOXO4-DBD. In contrast, larger RMSD values were obtained when models based on other templates were compared with the initial FOXO4 models used in the study (3L2C, 1E17), as expected from the discrepancies between the various NMR and crystal structures available (see Figure S30 in Supplementary Materials). Although this type of analysis was not possible for the regions that were absent from the templates, the excellent match manifested by the regions present in both model and template warranted a high level of confidence in the entire structures produced by our approach.

The ensemble of the FOXO4-DBD•DBE complex was examined in similar fashion. Also in this case, the overall topology matched very closely that of the corresponding high-resolution template (i.e., 3l2c), with the DBE component oriented in the proper direction and placed in the correct position onto the FOXO4-DBD's binding site (see Figure S28C,D of Supplementary Materials). RMSD comparisons between the model and crystal structure revealed excellent match for the regions present in both, thus ruling out the possibility of inadvertent perturbations introduced by the probing procedures. Additionally, we determined the distances between residues that had been conjugated by the cross-linking reagents, and then compared them with the corresponding distances measured on the crystal structure. The resulting RMSD values revealed excellent agreement across the board, with the sole exception of the distances between the DBE molecule (DBE-F) and specific residues of the H_{164}–M_{175} loop (H_{164}, T_{168}, S_{171}, S_{172} or M_{175}), which were somewhat longer in our model (see Table S1 in Supplementary Materials). These discrepancies; however, were consistent with the high degree of flexibility possessed by the loop, which was manifest also in the higher B-factors displayed by this region in the crystal structure. In agreement with the crystal structure, the models showed that helix H3 represents the main interaction interface, as indicated by both HDX and cross-linking data (see Table S1 in Supplementary Materials). The DBE structure employed here replicated the consensus binding sequences for all related FOX factors, which contain a general TGTTT motif surrounded by more variable sequences (see Figure S31 in Supplementary Materials). Whereas FOXO3 and FOXO6 recognize two nucleotides located after this consensus sequence, FOXP1 only recognize the second nucleotide but not the first one. In contrast, our model indicates that FOXO4 may recognize one nucleotide before and one after the consensus motif, thus affording additional evidence of the uniqueness of the interactions established by this member of the FOXO family. These results were supported also by Position Count Matrix (PCM) values that estimated the binding probability of individual bases at each position in the sequence (see Figure S31 in Supplementary Materials).

A close comparison of the structures of FOXO4-DBD and FOXO4-DBD•DBE complex obtained by our approach allowed us to further explore the effects of binding on protein conformation. The examination confirmed that the N- and C-terminal sequences remained largely unstructured even after DBE binding, as represented by the mesh regions of our models (Figure5). The main interface region consisting of H3 showed limited variations between unbound and bound forms. Similar outcomes were also observed for the contiguous H1–H2 loop and H2 helix. In contrast, loop H2–H4–H3 and the S2 and S3 strands of wing W1 showed rather large variations upon binding. Additionally, also the N- and C- terminal regions manifested extensive variations. These observations were consistent with the HDX data that revealed clearly peculiar time dependencies. For instance, the initially increasing rates in the H1 helix, H1–H2 loop, and H2 helix (A_{103}–T_{130}, in particular) suggested variations of dynamics upon DBE binding, whereas their subsequent decreasing rates

were consistent with the actual structural stabilization resulting from the presence of bound DNA. The RMSD values calculated for corresponding heavy backbone atoms provided an excellent measure of these conformational effects (see Figure S32 of Supplementary Materials). The values obtained from flanking regions near the interface, indicating major changes in loop H4–H3 and W1 wing and minor changes in loops H1–H2 and H3–S1, highlighted the indirect effects of binding, which were consistent with the results of HDX and XL experiments. The conformational changes revealed by this type of treatment were consistent with a classic adaptive binding mechanism by which rather sizeable conformational changes may be necessary to establish specific substrate-ligand interactions. The fact that the observed conformational changes were not limited only to the sequences in direct contact with the ligand DBE, but involved also contiguous regions, supports mechanisms by which binding events may trigger associated activities through allosteric effects, or place bordering regions in positions necessary to mediate the recruiting of additional factors, such as the acetyl transferases that are known to interact with other members of the FOXO family [50,71].

4. Conclusions

The structural investigation of FOXO4-DBD and FOXO4-DBD•DBE provided a thorough assessment of the ability of structural proteomics techniques to obtain valid information on systems that are not directly amenable to classic high-resolution approaches. The outcome showed that the concerted application of HDX and XL could effectively guide model-building operations to produce high-quality 3D models. Our multi-step strategy involved the utilization of high-resolution templates to carry out initial homology modelling. The sections that were not present in the templates were generated from experimental constraints and integrated with the initial structures to cover the entire DNA-binding domain. Although the high-resolution templates did not cover the entire structure folded by our construct, they still provided sufficient overlap to enable an unbiased assessment of the validity of the results afforded by our experimental/computational workflow. In fact, the excellent match between the templates and our structures ruled out the possibility that cross-linking procedures might have introduced unwanted artefacts or structure distortion. Further, an agreement between the templates and the corresponding portions of our models ruled out such possibilities and confirmed the ability of the selected computational strategies to translate these types of experimental constraints into actual 3D structures. Validating the approach on the "known" portions of the structures was essential in supporting the validity of the "unknown" sections that were conspicuously absent from the templates. The fact that the results of HDX and cross-linking experiments were in consistent mutual agreement provided further proof of the robustness of our concerted approach. For these reasons, our models represent comprehensive structures of full-fledged FOXO4-DBD and FOXO4-DBD•DBE.

The pictures painted by the high-resolution templates (identical protein construct was used), which were obtained by NMR and crystallography, are not only incomplete, but in the case of crystallography also static. Our extensive data provide a wealth of new information on the conformational dynamics of the protein in both unbound and bound forms. Our experiments clearly differentiated regions that were conformationally stable from those that underwent significant conformational changes upon DNA binding. The most important finding was that binding affected not only the interface region, but also the conformation of regions that were located away from the interface. This information might be essential to understand the allosteric properties of the complex and their role in recruiting additional transcriptional factors.

Our models confirmed that full-length FOXO4-DBD adopts the classic forkhead topology characteristic of this family of transcription factors, but corroborated also its unusual DNA binding mode that is unique among those manifested by the highly homologous FOXO proteins [44]. The close match between our model and the crystal structure of FOXO4-DBD•DBE ruled out the possibility that crystal packing might be the cause of the significant differences noted between the types of interactions established by FOXO4 versus those involving the other members of the family [48]. Our models confirmed the prominent role of helix H3 in such interactions and highlighted the involvement of

neighboring regions, which may be responsible for fine-tuning the sequence-specific recognition of the correct DNA counterpart.

In conclusion, this study demonstrated the merits of structural proteomics approaches for the elucidation of protein-nucleic acid complexes. The utilization of specific hydrolytic procedures to complete the characterization of HDX and cross-linking products virtually eliminates any limitation pertaining the size of the species of interest. Propelled by continued advances in the computational approaches employed to translate the experimental results into all-atoms models, structural proteomics has rapidly emerged as a valid complement, and often alternative, to the classic high-resolution techniques. By clearly demonstrating the applicability to transcription factor-response element complexes, we hope that this study will lead to a broader utilization of structural proteomics to mitigate the chronic dearth of information on these essential regulatory systems.

Supplementary Materials: The following are available online athttp: //www.mdpi.com/2218-273X/9/10/535/s1, Supplementary methods, Figure S1: Native and denaturing polyacrylamide gel electrophoresis (PAGE) analysis of transplatinated FOXO4-DBD•DBE complex, Figure S2: Native ESI-MS analysis of FOXO4-DBD•DBE complex, Figure S3: Relative deuteration rates of peptides (1/3), Figure S4: Relative deuteration rates of peptides (2/3), Figure S5: Relative deuteration rates of peptides (3/3), Figure S6: HDX butterfly plot, Figure S7: Variations of relative deuteration rates as a function of time mapped onto the protein structure, Figure S8: Collision-induced dissociation (CID) of peptide-oligonucleotide cross-link by transplatin, Figure S9: Representative data obtained from protein-DNA cross-linking by transplatin, Figure S10: Mass spectra of cross-links dissociation products (1/17), Figure S11: Mass spectra of cross-links dissociation products (2/17), Figure S12: Mass spectra of cross-links dissociation products (3/17), Figure S13: Mass spectra of cross-links dissociation products (4/17), Figure S14: Mass spectra of cross-links dissociation products (5/17), Figure S15: Mass spectra of cross-links dissociation products (6/17), Figure S16: Mass spectra of cross-links dissociation products (7/17), Figure S17: Mass spectra of cross-links dissociation products (8/17), Figure S18: Mass spectra of cross-links dissociation products (9/17), Figure S19: Mass spectra of cross-links dissociation products (10/17), Figure S20: Mass spectra of cross-links dissociation products (11/17), Figure S21: Mass spectra of cross-links dissociation products (12/17), Figure S22: Mass spectra of cross-links dissociation products (13/17), Figure S23: Mass spectra of cross-links dissociation products (14/17), Figure S24: Mass spectra of cross-links dissociation products (15/17), Figure S25: Mass spectra of cross-links dissociation products (16/17), Figure S26: Mass spectra of cross-links dissociation products (17/17), Figure S27: Protein-DNA docking, Figure S28: Comparison between models of full-fledged FOXO4-DBD in either free or DBE-bound form with the corresponding high-resolution structures, Figure S29: Average RMSD between the coordinates of backbone heavy atoms in our unbound FOXO4-DBD model and the corresponding 1E17 high-resolution template, Figure S30: Average RMSD between the coordinates of backbone heavy atoms in six created unbound FOXO4-DBD models based on different FOX structures related to FOXO4 and the 3L2C or 1E17 high-resolution template, Figure S31: Putative FOXO4 consensus sequences, Figure S32: Average RMSD between the coordinates of backbone heavy atoms in our unbound FOXO4-DBD and bound FOXO4-DBD•DBE complex, Figure S33: Protein sequence encoded by the express vector pET-15b, Table S1: Deviations between theoretical and back-calculated distances of cross-linked residues in unbound and bound forms obtained from our structural models, Table S2: Quantitative protein-protein cross-linking of unbound FOXO4-DBD and bound FOXO4-DBD•DBE (complex).

Author Contributions: Conceptualization, D.F. and P.N.; Methodology D.F., H.M., P.N. and P.M.; Software D.K. and M.R.; Validation M.M. and W.M.; Formal Analysis J.C. and K.V.; Investigation J.F., J.C., L.S., R.F. and Z.K.; Resources M.M. and W.M.; Data Curation J.F., L.S., M.R. and Z.K.; Writing—Original Draft L.S.; Writing—Review & Editing D.F.; Visualization J.C. and L.S.; Supervision P.M. and P.N.; Project Administration P.N.; Funding Acquisition P.N.

Funding: This research was mainly funded by the Czech Science Foundation (grant number 16-24309S), the Ministry of Education of the Czech Republic (project LH15010, programs "NPU II" project LQ1604 and "NPU I" project LO1509), European Commission H2020 (European Network of Fourier-Transform Ion-Cyclotron-Resonance Mass Spectrometry Centers—project agreement No.731077 and European Proteomics Infrastructure Consortium providing Access—project agreement No.823839), the Charles University Grant Agency (932316) and, in part, by the Czech Academy of Sciences (RVO61388971). This work was also funded by the National Institutes of Health R01 GM121844–01.

Acknowledgments: We acknowledge the Centre of molecular structure Core Facility at BIOCEV, a facility funded by European Regional Development Funds (CZ.1.05/1.1.00/02.0109 BIOCEV) and supported by the Czech Infrastructure for Integrative Structural Biology (LM2015043 CIISB for CMS BIOCEV funded by MEYS CR). The FP7 WeNMR (project# 261572) and H2020 West-Life (project# 675858) European e-Infrastructure projects are acknowledged for the use of their web portals, which make use of the EGI infrastructure and DIRAC4EGI service with the dedicated support of CESNET-MetaCloud.

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

References

1. Lambert, S.A.; Jolma, A.; Campitelli, L.F.; Das, P.K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T.R.; Weirauch, M.T. The Human Transcription Factors. *Cell* **2018**, *172*, 650–665. [CrossRef] [PubMed]
2. Latchman, D.S. Transcription factors: An overview. *Int. J. Biochem. Cell Biol.* **1997**, *29*, 1305–1312. [CrossRef]
3. Latchman, D.S. Transcription factors: Bound to activate or repress. *Trends Biochem. Sci.* **2001**, *26*, 211–213. [CrossRef]
4. Babu, M.M.; Luscombe, N.M.; Aravind, L.; Gerstein, M.; Teichmann, S.A. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **2004**, *14*, 283–291. [CrossRef] [PubMed]
5. Brivanlou, A.H.; Darnell, J.E. Signal transduction and the control of gene expression. *Science* **2002**, *295*, 813–818. [CrossRef] [PubMed]
6. Vaquerizas, J.M.; Kummerfeld, S.K.; Teichmann, S.A.; Luscombe, N.M. A census of human transcription factors: Function, expression and evolution. *Nat. Rev. Genet.* **2009**, *10*, 252–263. [CrossRef]
7. Heck, A.J.R. Native mass spectrometry: A bridge between interactomics and structural biology. *Nat. Methods* **2008**, *5*, 927–933. [CrossRef] [PubMed]
8. Brent, R.; Ptashne, M. A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* **1985**, *43*, 729–736. [CrossRef]
9. Hollenberg, S.M.; Evans, R.M. Multiple and cooperative trans-activation domains of the human glucocorticoid receptor. *Cell* **1988**, *55*, 899–906. [CrossRef]
10. Ma, J.; Ptashne, M. A new class of yeast transcriptional activators. *Cell* **1987**, *51*, 113–119. [CrossRef]
11. Rozbeský, D.; Adámek, D.; Pospíšilová, E.; Novák, P.; Chmelík, J. Solution structure of the lymphocyte receptor Nkrp1a reveals a distinct conformation of the long loop region as compared to in the crystal structure. *Proteins* **2016**, *84*, 1304–1311. [CrossRef] [PubMed]
12. Rozbesky, D.; Man, P.; Kavan, D.; Chmelik, J.; Cerny, J.; Bezouska, K.; Novak, P. Chemical cross-linking and H/D exchange for fast refinement of protein crystal structure. *Anal. Chem.* **2012**, *84*, 867–870. [CrossRef]
13. Kukacka, Z.; Rosulek, M.; Strohalm, M.; Kavan, D.; Novak, P. Mapping protein structural changes by quantitative cross-linking. *Methods* **2015**, *89*, 112–120. [CrossRef] [PubMed]
14. Vandermarliere, E.; Stes, E.; Gevaert, K.; Martens, L. Resolution of protein structure by mass spectrometry. *Mass Spectrom. Rev.* **2016**, *35*, 653–665. [CrossRef]
15. Konermann, L.; Pan, J.; Liu, Y.-H. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* **2011**, *40*, 1224–1234. [CrossRef] [PubMed]
16. Zhang, Q.; Yu, E.T.; Kellersberger, K.A.; Crosland, E.; Fabris, D. Toward building a database of bifunctional probes for the MS3D investigation of nucleic acids structures. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1570–1581. [CrossRef]
17. Fabris, D.; Yu, E.T. Elucidating the higher-order structure of biopolymers by structural probing and mass spectrometry: MS3D. *J. Mass Spectrom.* **2010**, *45*, 841–860. [CrossRef]
18. Sinz, A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom. Rev.* **2006**, *25*, 663–682. [CrossRef]
19. Giladi, M.; van Dijk, L.; Refaeli, B.; Almagor, L.; Hiller, R.; Man, P.; Forest, E.; Khananshvili, D. Dynamic distinctions in the Na+/Ca2+ exchanger adopting the inward- and outward-facing conformational states. *J. Biol. Chem.* **2017**, *292*, 12311–12323. [CrossRef]
20. Kadek, A.; Kavan, D.; Felice, A.K.G.; Ludwig, R.; Halada, P.; Man, P. Structural insight into the calcium ion modulated interdomain electron transfer in cellobiose dehydrogenase. *FEBS Lett.* **2015**, *589*, 1194–1199. [CrossRef]
21. Engen, J.R.; Wales, T.E.; Chen, S.; Marzluff, E.M.; Hassell, K.M.; Weis, D.D.; Smithgall, T.E. Partial cooperative unfolding in proteins as observed by hydrogen exchange mass spectrometry. *Int. Rev. Phys. Chem.* **2013**, *32*, 96–127. [CrossRef] [PubMed]
22. Lennartz, F.; Bengtsson, A.; Olsen, R.W.; Joergensen, L.; Brown, A.; Remy, L.; Man, P.; Forest, E.; Barfod, L.K.; Adams, Y.; et al. Mapping the Binding Site of a Cross-Reactive Plasmodium falciparum PfEMP1 Monoclonal Antibody Inhibitory of ICAM-1 Binding. *J. Immunol.* **2015**, *195*, 3273–3283. [CrossRef] [PubMed]
23. Kacirova, M.; Kosek, D.; Kadek, A.; Man, P.; Vecer, J.; Herman, P.; Obsilova, V.; Obsil, T. Structural Characterization of Phosducin and Its Complex with the 14-3-3 Protein. *J. Biol. Chem.* **2015**, *290*, 16246–16260. [CrossRef] [PubMed]

24. Zhang, J.; Chalmers, M.J.; Stayrook, K.R.; Burris, L.L.; Wang, Y.; Busby, S.A.; Pascal, B.D.; Garcia-Ordonez, R.D.; Bruning, J.B.; Istrate, M.A.; et al. DNA binding alters coactivator interaction surfaces of the intact VDR-RXR complex. *Nat. Struct. Mol. Biol.* **2011**, *18*, 556–563. [CrossRef] [PubMed]
25. Graham, B.W.; Tao, Y.; Dodge, K.L.; Thaxton, C.T.; Olaso, D.; Young, N.L.; Marshall, A.G.; Trakselis, M.A. DNA Interactions Probed by Hydrogen-Deuterium Exchange (HDX) Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Confirm External Binding Sites on the Minichromosomal Maintenance (MCM) Helicase. *J. Biol. Chem.* **2016**, *291*, 12467–12480. [CrossRef]
26. Zhu, M.M.; Chitta, R.; Gross, M.L. PLIMSTEX: A novel mass spectrometric method for the quantification of protein–ligand interactions in solution. *Int. J. Mass Spectrom.* **2005**, *240*, 213–220. [CrossRef]
27. Sperry, J.B.; Shi, X.; Rempel, D.L.; Nishimura, Y.; Akashi, S.; Gross, M.L. A mass spectrometric approach to the study of DNA-binding proteins: Interaction of human TRF2 with telomeric DNA. *Biochemistry* **2008**, *47*, 1797–1807. [CrossRef]
28. Zheng, J.; Yong, H.Y.; Panutdaporn, N.; Liu, C.; Tang, K.; Luo, D. High-resolution HDX-MS reveals distinct mechanisms of RNA recognition and activation by RIG-I and MDA5. *Nucleic Acids Res.* **2015**, *43*, 1216–1230. [CrossRef]
29. Morton, V.L.; Burkitt, W.; O'Connor, G.; Stonehouse, N.J.; Stockley, P.G.; Ashcroft, A.E. RNA-induced conformational changes in a viral coat protein studied by hydrogen/deuterium exchange mass spectrometry. *Phys. Chem. Chem. Phys.* **2010**, *12*, 13468–13475. [CrossRef]
30. Novak, P.; Kruppa, G.H. Intra-molecular cross-linking of acidic residues for protein structure studies. *Eur. J. Mass Spectrom.* **2008**, *14*, 355–365. [CrossRef]
31. Young, M.M.; Tang, N.; Hempel, J.C.; Oshiro, C.M.; Taylor, E.W.; Kuntz, I.D.; Gibson, B.W.; Dollinger, G. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 5802–5806. [CrossRef]
32. Rozbesky, D.; Sovova, Z.; Marcoux, J.; Man, P.; Ettrich, R.; Robinson, C.V.; Novak, P. Structural model of lymphocyte receptor NKR-P1C revealed by mass spectrometry and molecular modeling. *Anal. Chem.* **2013**, *85*, 1597–1604. [CrossRef]
33. Yu, E.T.; Zhang, Q.; Fabris, D. Untying the FIV frameshifting pseudoknot structure by MS3D. *J. Mol. Biol.* **2005**, *345*, 69–80. [CrossRef]
34. Steen, H.; Petersen, J.; Mann, M.; Jensen, O.N. Mass spectrometric analysis of a UV-cross-linked protein-DNA complex: Tryptophans 54 and 88 of E. coli SSB cross-link to DNA. *Protein Sci.* **2001**, *10*, 1989–2001. [CrossRef]
35. Lenz, C.; Kühn-Hölsken, E.; Urlaub, H. Detection of protein-RNA crosslinks by NanoLC-ESI-MS/MS using precursor ion scanning and multiple reaction monitoring (MRM) experiments. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 869–881. [CrossRef]
36. Sharma, K.; Hrle, A.; Kramer, K.; Sachsenberg, T.; Staals, R.H.J.; Randau, L.; Marchfelder, A.; van der Oost, J.; Kohlbacher, O.; Conti, E.; et al. Analysis of protein-RNA interactions in CRISPR proteins and effector complexes by UV-induced cross-linking and mass spectrometry. *Methods* **2015**, *89*, 138–148. [CrossRef]
37. Loeber, R.; Michaelson, E.; Fang, Q.; Campbell, C.; Pegg, A.E.; Tretyakova, N. Cross-Linking of the DNA Repair Protein O 6 -Alkylguanine DNA Alkyltransferase to DNA in the Presence of Antitumor Nitrogen Mustards. *Chem. Res. Toxicol.* **2008**, *21*, 787–795. [CrossRef]
38. Loeber, R.; Rajesh, M.; Fang, Q.; Pegg, A.E.; Tretyakova, N. Cross-Linking of the Human DNA Repair Protein O 6 -Alkylguanine DNA Alkyltransferase to DNA in the Presence of 1,2,3,4-Diepoxybutane. *Chem. Res. Toxicol.* **2006**, *19*, 645–654. [CrossRef]
39. Michaelson-Richie, E.D.; Ming, X.; Codreanu, S.G.; Loeber, R.L.; Liebler, D.C.; Campbell, C.; Tretyakova, N.Y. Mechlorethamine-induced DNA-protein cross-linking in human fibrosarcoma (HT1080) cells. *J. Proteome Res.* **2011**, *10*, 2785–2796. [CrossRef]
40. Müller, D.R.; Schindler, P.; Towbin, H.; Wirth, U.; Voshol, H.; Hoving, S.; Steinmetz, M.O. Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis. *Anal. Chem.* **2001**, *73*, 1927–1934. [CrossRef]
41. Kang, S.; Mou, L.; Lanman, J.; Velu, S.; Brouillette, W.J.; Prevelige, P.E. Synthesis of biotin-tagged chemical cross-linkers and their applications for mass spectrometry. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 1719–1726. [CrossRef]

42. Furuyama, T.; Nakazawa, T.; Nakano, I.; Mori, N. Identification of the differential distribution patterns of mRNAs and consensus binding sequences for mouse DAF-16 homologues. *Biochem. J.* **2000**, *349*, 629–634. [CrossRef]
43. Borkhardt, A.; Repp, R.; Haas, O.A.; Leis, T.; Harbott, J.; Kreuder, J.; Hammermann, J.; Henn, T.; Lampert, F. Cloning and characterization of AFX, the gene that fuses to MLL in acute leukemias with a t(X;11)(q13;q23). *Oncogene* **1997**, *14*, 195–202. [CrossRef]
44. Zhang, X.; Tang, N.; Hadden, T.J.; Rishi, A.K. Akt, FoxO and regulation of apoptosis. *Biochim. Biophys. Acta* **2011**, *1813*, 1978–1986. [CrossRef]
45. Kaestner, K.H.; Knochel, W.; Martinez, D.E. Unified nomenclature for the winged helix/forkhead transcription factors. *Genes Dev.* **2000**, *14*, 142–146.
46. Weigelt, J.; Climent, I.; Dahlman-Wright, K.; Wikström, M. 1H, 13C and 15N resonance assignments of the DNA binding domain of the human forkhead transcription factor AFX. *J. Biomol. NMR* **2000**, *17*, 181–182. [CrossRef]
47. Boura, E.; Silhan, J.; Herman, P.; Vecer, J.; Sulc, M.; Teisinger, J.; Obsilova, V.; Obsil, T. Both the N-terminal loop and wing W2 of the forkhead domain of transcription factor Foxo4 are important for DNA binding. *J. Biol. Chem.* **2007**, *282*, 8265–8275. [CrossRef]
48. Boura, E.; Rezabkova, L.; Brynda, J.; Obsilova, V.; Obsil, T. Structure of the human FOXO4-DBD-DNA complex at 1.9 Å resolution reveals new details of FOXO binding to the DNA. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 1351–1357. [CrossRef]
49. Scalabrin, M.; Dixit, S.M.; Makshood, M.M.; Krzemien, C.E.; Fabris, D. Bifunctional cross-linking approaches for mass spectrometry-based investigation of nucleic acids and protein-nucleic acid assemblies. *Methods* **2018**, *144*, 64–78. [CrossRef]
50. Valis, K.; Prochazka, L.; Boura, E.; Chladova, J.; Obsil, T.; Rohlena, J.; Truksa, J.; Dong, L.F.; Ralph, S.J.; Neuzil, J. Hippo/Mst1 stimulates transcription of the proapoptotic mediator NOXA in a FoxO1-dependent manner. *Cancer Res.* **2011**, *71*, 946–954. [CrossRef]
51. Kulakovskiy, I.V.; Vorontsov, I.E.; Yevshin, I.S.; Sharipov, R.N.; Fedorova, A.D.; Rumynskiy, E.I.; Medvedeva, Y.A.; Magana-Mora, A.; Bajic, V.B.; Papatsenko, D.A.; et al. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **2018**, *46*, D252–D259. [CrossRef]
52. Vorontsov, I.E.; Kulakovskiy, I.V.; Makeev, V.J. Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms Mol. Biol.* **2013**, *8*, 23. [CrossRef]
53. Obsil, T.; Obsilova, V. Structural basis for DNA recognition by FOXO proteins. *Biochim. Biophys. Acta* **2011**, *1813*, 1946–1953. [CrossRef]
54. Rozbeský, D.; Rosůlek, M.; Kukačka, Z.; Chmelík, J.; Man, P.; Novák, P. Impact of Chemical Cross-Linking on Protein Structure and Function. *Anal. Chem.* **2018**, *90*, 1104–1113. [CrossRef]
55. Kadek, A.; Kavan, D.; Marcoux, J.; Stojko, J.; Felice, A.K.G.; Cianférani, S.; Ludwig, R.; Halada, P.; Man, P. Interdomain electron transfer in cellobiose dehydrogenase is governed by surface electrostatics. *Biochim. Biophys. Acta. Gen. Subj.* **2017**, *1861*, 157–167. [CrossRef]
56. Strohalm, M.; Kavan, D.; Novák, P.; Volný, M.; Havlícek, V. mMass 3: A cross-platform software environment for precise analysis of mass spectrometric data. *Anal. Chem.* **2010**, *82*, 4648–4651. [CrossRef]
57. Webb, B.; Sali, A. Comparative Protein Structure Modeling Using Modeller. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2016; pp. 5–6.
58. Dominguez, C.; Boelens, R.; Bonvin, A.M.J.J. Haddock: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125*, 1731–1737. [CrossRef]
59. Wassenaar, T.A.; van Dijk, M.; Loureiro-Ferreira, N.; van der Schot, G.; de Vries, S.J.; Schmitz, C.; van der Zwan, J.; Boelens, R.; Giachetti, A.; Ferella, L.; et al. WeNMR: Structural Biology on the Grid. *J. Grid Comput.* **2012**, *10*, 743–767. [CrossRef]
60. Brünger, A.T.; Adams, P.D.; Clore, G.M.; DeLano, W.L.; Gros, P.; Grosse-Kunstleve, R.W.; Jiang, J.S.; Kuszewski, J.; Nilges, M.; Pannu, N.S.; et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54*, 905–921.
61. Pierce Chemicals. *Double Agents Cross-Linking Reagents Selection Guide*; Pierce Chemicals: Rockford, IL, USA, 1999.

62. Dans, P.D.; Crespo, A.; Estrin, D.A.; Coitiño, E.L. Structural and Energetic Study of Cisplatin and Derivatives: Comparison of the Performance of Density Funtional Theory Implementations. *J. Chem. Theory Comput.* **2008**, *4*, 740–750. [CrossRef]
63. Rosen, M.S.; Spokoyny, A.M.; Machan, C.W.; Stern, C.; Sarjeant, A.; Mirkin, C.A. Chelating Effect as a Driving Force for the Selective Formation of Heteroligated Pt(II) Complexes with Bidentate Phosphino-Chalcoether Ligands. *Inorg. Chem.* **2011**, *50*, 1411–1419. [CrossRef]
64. DeLano, W.L. *The PyMOL Molecular Graphics System*; Schrodinger Inc: Portland, OR, USA, 2002.
65. Kavan, D.; Man, P. MSTools—Web based application for visualization and presentation of HXMS data. *Int. J. Mass Spectrom.* **2011**, *302*, 53–58. [CrossRef]
66. Kadek, A.; Mrazek, H.; Halada, P.; Rey, M.; Schriemer, D.C.; Man, P. Aspartic Protease Nepenthesin-1 as a Tool for Digestion in Hydrogen/Deuterium Exchange Mass Spectrometry. *Anal. Chem.* **2014**, *86*, 4287–4294. [CrossRef]
67. Timerbaev, A.R.; Hartinger, C.G.; Aleksenko, S.S.; Keppler, B.K. Interactions of antitumor metallodrugs with serum proteins: Advances in characterization using modern analytical methodology. *Chem. Rev.* **2006**, *106*, 2224–2248. [CrossRef]
68. Deubel, D.V. Factors governing the kinetic competition of nitrogen and sulfur ligands in cisplatin binding to biological targets. *J. Am. Chem. Soc.* **2004**, *126*, 5999–6004. [CrossRef]
69. Nafisi, S.; Norouzi, Z. A comparative study on the interaction of cis- and trans-platin with DNA and RNA. *DNA Cell Biol.* **2009**, *28*, 469–477. [CrossRef]
70. Buchan, D.W.A.; Minneci, F.; Nugent, T.C.O.; Bryson, K.; Jones, D.T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **2013**, *41*, W349–W357. [CrossRef]
71. Eijkelenboom, A.; Burgering, B.M.T. FOXOs: Signalling integrators for homeostasis maintenance. *Nat. Rev. Mol. Cell Biol.* **2013**, *14*, 83–97. [CrossRef]

Publication IV

Filandrová, R., Vališ, K., Černý, J., Chmelík, J., Slavata, L., Fiala, J., Rosůlek, M., Kavan, D., Man, P., Chum, T., Cebecauer, M., Fabris, D., & Novák, P

Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA

Structure, 29, 1–12 (2021)

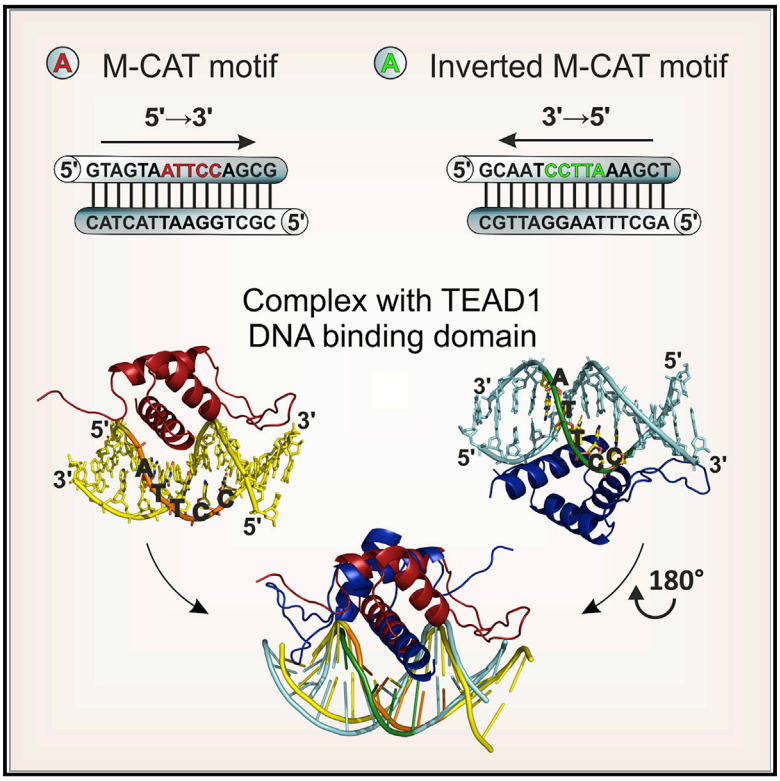
My contribution: *recombinant protein production, planning and conducting experiments (native nESI-MS, fluorescence anisotropy binding assay, HDX-MS, quantitative chemical cross-linking, part of smFRET), data analysis & interpretation, manuscript writing, figure design*

Article

Structure

Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA

Graphical Abstract



Authors

Růžena Filandrová, Karel Vališ, Jiří Černý, ..., Marek Cebecauer, Daniele Fabris, Petr Novák

Correspondence

pnovak@biomed.cas.cz

In Brief

TEAD transcription factors regulate gene expression during the development of eukaryotic organisms. Filandrova et al. investigated how the context of its DNA binding motif affects its affinity with DNA and how TEAD1 can bind to high- and low-affinity inverted motifs using structural mass spectrometry, computational modeling, and smFRET.

Highlights

- Different "best fit" motifs affect M-CAT recognition by TEAD1
- TEAD1 binds the 5′-CCTTA-3′ motif with low affinity both *in vitro* and *in vivo*
- TEAD1 binds the low-affinity inverted M-CAT motif in 180° rotated orientation
- Helices facing the other side from DNA are stabilized upon DNA binding

Filandrová et al., 2021, Structure *29*, 1–12
April 1, 2021
https://doi.org/10.1016/j.str.2020.11.018

CellPress

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

Structure

CellPress

Article

Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA

Růžena Filandrová,[1,2] Karel Vališ,[1] Jiří Černý,[3] Josef Chmelík,[1,2] Lukáš Slavata,[1,2] Jan Fiala,[1,2] Michal Rosůlek,[1,2] Daniel Kavan,[1,2] Petr Man,[1,2] Tomáš Chum,[4] Marek Cebecauer,[4] Daniele Fabris,[5] and Petr Novák[1,2,6,*]

[1]Institute of Microbiology, Czech Academy of Sciences, Prague 142 20, Czech Republic
[2]Faculty of Science, Charles University, Prague 128 43, Czech Republic
[3]Institute of Biotechnology, Czech Academy of Sciences, Vestec 252 50, Czech Republic
[4]J. Heyrovsky Institute of Physical Chemistry, Czech Academy of Sciences, Prague 182 00, Czech Republic
[5]University of Connecticut, Department of Chemistry, 55 N. Eagleville Road, Storrs, CT 06269, USA
[6]Lead Contact
*Correspondence: pnovak@biomed.cas.cz
https://doi.org/10.1016/j.str.2020.11.018

SUMMARY

TEAD transcription factors regulate gene expression through interactions with DNA and other proteins. They are crucial for the development of eukaryotic organisms and to control the expression of genes involved mostly in cell proliferation and differentiation; however, their deregulation can lead to tumorigenesis. To study the interactions of TEAD1 with M-CAT motifs and their inverted versions, the K_D of each complex was determined, and H/D exchange, quantitative chemical cross-linking, molecular docking, and smFRET were utilized for structural characterization. ChIP-qPCR was employed to correlate the results with a cell line model. The results obtained showed that although the inverted motif has 10× higher K_D, the same residues were affected by the presence of M-CAT in both orientations. Molecular docking and smFRET revealed that TEAD1 binds the inverted motif rotated 180°. In addition, the inverted motif was proven to be occupied by TEAD1 in Jurkat cells, suggesting that the low-affinity binding sites present in the human genome may possess biological relevance.

INTRODUCTION

Transcription enhancer activator domain (TEAD) proteins are widely distributed transcription factors that share a highly conserved DNA-binding domain (DBD) (Jacquemin et al., 1996; Kaneko and DePamphilis, 1998). The first member of the family was identified as a small nuclear protein bound to the SV40 enhancer in HeLa cells, in which it regulates transcription of SV40 virus (Xiao et al., 1987). Over time, additional members (i.e., TEAD2 to 4) were recognized in mammals, in which almost every tissue was found to express at least one type of TEAD, while others were shown capable of expressing all of them (Jacquemin et al., 1996, 1998; Kaneko and DePamphilis, 1998; Yasunami et al., 1996). Genetic manipulation in mice exposed Tead1 as a crucial regulator of cardiac muscle differentiation and growth and also revealed its importance for maintaining normal adult heart function (Chen et al., 1994; Liu et al., 2017; Sawada et al., 2008). Tead2 was shown to be responsible for neural tube closure (Kaneko et al., 2007) and, in cooperation with Tead1 and Tead4, for heart development (Liu et al., 2017; Sawada et al., 2008). Silencing Tead4 demonstrated its activity in primary myoblast differentiation (Joshi et al., 2017), whereas no functional information has thus far emerged for Tead3. Perhaps not coincidentally, the essential roles of TEAD transcription factors in developmental processes translate into their possible implication in several types of tumors. Multiple genes known to be connected with tumorigenesis, including connective tissue growth factor (*CTGF*) (Zhao et al., 2008), transcription factor *C-MYC* (Dong et al., 2007), or glucose transporter *GLUT1* (Vališ et al., 2016), were shown to be regulated by TEAD proteins (Lin et al., 2017; Zhou et al., 2016). Thus, TEAD protein activity must be strictly regulated to prevent developmental defects or malignant transformation. Since TEAD proteins on their own are not able to activate transcription and can achieve that only through interactions with other regulatory proteins known as coactivators (Lin et al., 2017; Xiao et al., 1991), most studies focus on TEAD-coactivator interactions in order to understand their regulation. Two of the identified coactivators, YES-associated protein (YAP) (Vassilev et al., 2001) and its paralog TAZ (Mahoney et al., 2005), represent well-known targets of the organ-size-regulating Hippo signaling pathway (Zhao et al., 2007), whereas other TEAD coactivators act independently on Hippo signaling and include VGLL, the p160 family of steroid receptors, serum response factor (SRF), poly-ADP ribose-polymerase, activator protein-1, myocyte enhancer factor 2, and the C-MYC interaction partner MAX, as reviewed by Lin et al. (Lin et al., 2017).

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

TEAD proteins consist of two main structural domains: the DNA-binding (or TEA) and the transactivation (or YAP-binding) domain. Although no high-resolution structure is available for the full-length protein, structures of the individual domains have been separately solved for different members of the family. For instance, the TEA structure of TEAD1 was initially obtained by NMR in the absence of bound DNA, identifying helix H3 as a part of the possible binding interface (Anbanandam et al., 2006). The crystal structure of the DBD of TEAD4 in complex with a model duplex subsequently confirmed the location of the DNA-binding region. This study also brought insight into what the structural determinants of DNA recognition by a TEAD protein are. As is the case for most transcription factors, the DNA recognition and binding are governed by non-covalent interactions between amino acid side chains and bases or the deoxyribose-phosphate backbone of DNA. TEAD4 was shown to have two main interaction interfaces: helix H3 and the L1 loop. Several hydrogen and salt bridges between H3 helix residues and bases of the DNA recognition motif, as well as sequence-independent hydrophobic packing in the L1 loop, were identified as responsible for the complex formation (Shi et al., 2017). At the same time, the crystal structures of transactivation domains of TEAD1 and TEAD2 in complex with the coactivator YAP displayed a high degree of similarity, consistent with the highly conserved nature of these proteins (Li et al., 2010; Tian et al., 2010). Further, the structures of complexes including either TAZ or Vgll coactivator demonstrated the ability of this domain to bind different cognate factors in very similar fashions (Kaan et al., 2017; Pobbati et al., 2012). Finally, palmitoylation has been shown to ensure proper folding and stabilization of the transactivation domain structure, thus providing another possible mechanism for regulating its functions (Noland et al., 2016). Despite the wealth of information provided by these high-resolution structures, how DNA recognition is affected by the binding motif context and the possible interplay between DNA-binding and transactivation domains in stabilizing such interaction are still not well understood.

In this report, we have investigated the structural bases of the recognition between TEAD1 and its specific target DNA. Searching the JASPAR database revealed that all TEAD proteins possessed a putative 5′-ATTCC-3′ consensus binding motif, but flanking sequences varied widely within the family (Khan et al., 2018). This core sequence with an additional cytosine on the 5′ end has been broadly referred to as the M-CAT (muscle–cytosine, adenine, thymine) motif by virtue of its abundance in the regulatory regions of genes that are specifically expressed in muscle tissue (Mar and Ordahl, 1990). Its relationship with TEAD1 was further confirmed by using a protein-binding chip derivatized with randomized DNA duplexes designed to identify the sequence with the highest affinity for TEAD. The results led to a putative consensus sequence corresponding to ANATVCZN, in which V can be A, T, or G; Z can be A, T, or C; and N can be any base (Anbanandam et al., 2006). Based on these premises, we compared the affinity of isolated TEAD1-DBD toward different DNA duplexes that replicated the context of the 5′-ATTCC-3′ core of the M-CAT motif within the regulatory sequences of selected human genes. We also investigated whether TEAD1-DBD is able to bind its response motif in the inverted (5′-CCTTA-3′) orientation since this motif is abundant in the regulatory regions of human genes that were previously identified as regulated by TEAD transcription factors and how the orientation of the motif (i.e., 5′-ATTCC-3′ versus 5′-CCTTA-3′) affects binding affinity and, thus, recognition. Further, we employed a combination of mass spectrometry (MS) techniques, such as H/D exchange (HDX) and quantitative chemical cross-linking (Kukacka et al., 2015; Rozbesky et al., 2012; Slavata et al., 2019), and molecular docking experiments to explore the structural basis of the interactions between TEAD1-DBD and M-CAT constructs. We finally evaluated the significance of these observations obtained *in vitro* by determining the occupancy of selected M-CATs in leukemic cells.

RESULTS AND DISCUSSION

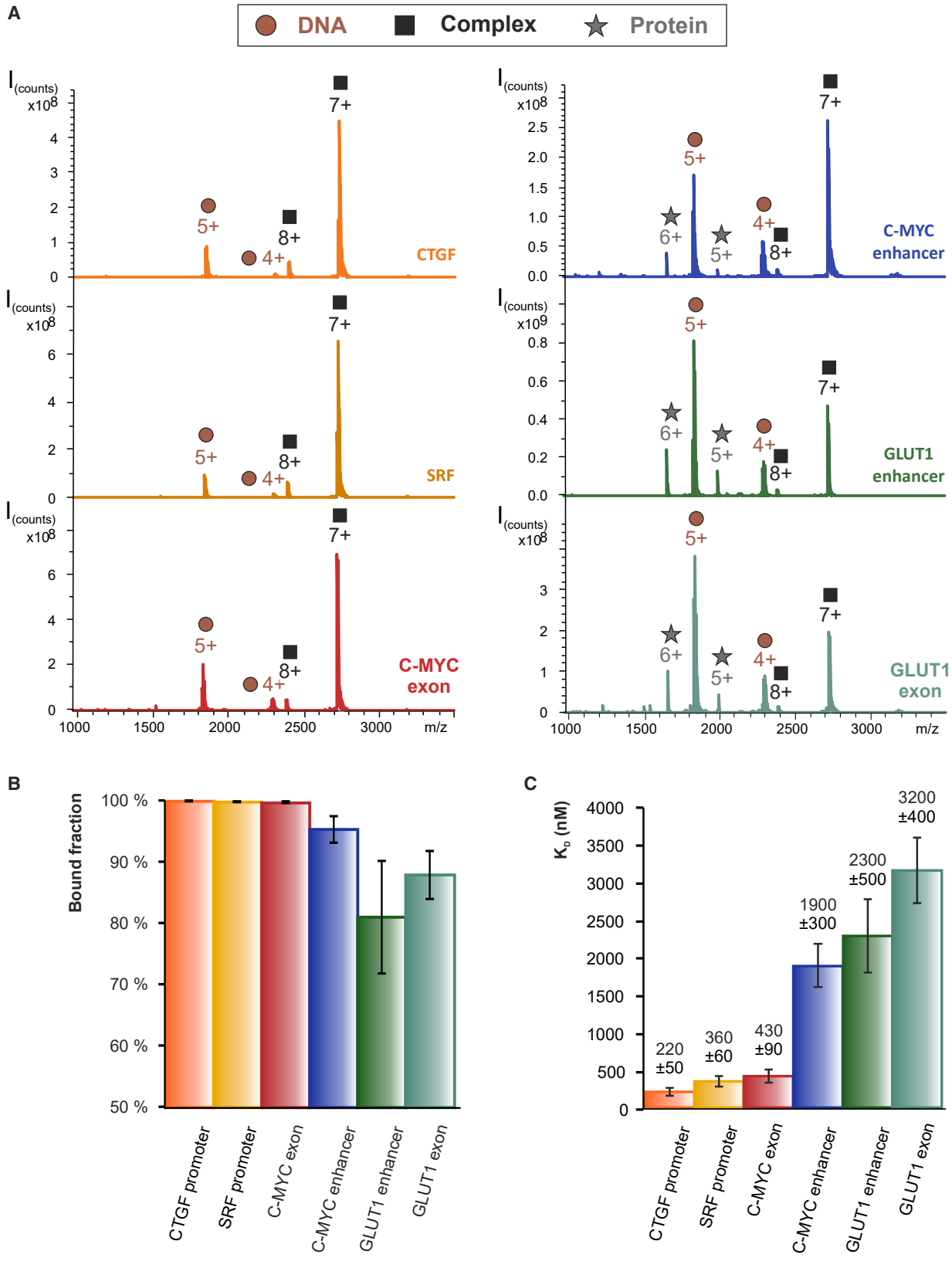
The structural characterization of the interactions between TEAD1-DBD and cognate DNA targets was carried out by using a multipronged approach (Slavata et al., 2019). For the purpose of exploring the effects of DNA sequence on protein recognition, we compared the binding properties of a series of double-stranded DNA (dsDNA) constructs (see Table S1) that placed the 5′-ATTCC-3′ core of the consensus M-CAT motif in different sequence contexts (further referred to as M-CATs). All the sequences of the selected M-CATs originated from regulatory regions of human genes, which in the genome colocalize with markers of active chromatin (see Figure S1). Structural proteomic techniques were then employed to compare the spatial arrangements of free versus bound components of the complex to enable the characterization of the contact interface and the evaluation of possible conformational effects induced by binding of different M-CATs. The dsDNAs were submitted to UV-melting analysis to confirm their proper formation (see Figure S2), whereas the identity, purity, and proper folding of recombinant TEAD1-DBD and complex assemblies were verified by MS analysis (see Figures 1 and S4), as detailed in the STAR Methods section.

M-CAT context affects TEAD1 recognition

Together with verification of the identity and purity of the complex components, the binding properties of the selected M-CATs were initially compared by using native electrospray ionization MS (ESI-MS), as described under STAR Methods. Figure 1 displays representative spectra obtained from samples prepared by mixing equimolar amounts of recombinant TEAD1-DBD with each of the M-CAT dsDNAs. Consistent with the establishment of a binding equilibrium in solution, each sample provided signals corresponding to both free and bound components, which were recognized from their characteristic masses. The respective signal intensities enabled us to determine the partitioning between free and bound species, which is a direct reflection of the binding affinity of the complex. In particular, the fact that M-CAT constructs mimicking the *CTGF* promoter, *SRF* promoter, and *C-MYC* first exon were almost completely bound to TEAD1-DBD, with percentages of bound form ranging from 99.6% to 100%, indicated that these sequences possessed the higher affinities in the series for the protein component. In contrast, the M-CATs mimicking the *C-MYC* enhancer, *GLUT1* enhancer, and *GLUT1* first exon displayed percentages of bound form ranging only from 72% to 97%,

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

CellPress



(legend on next page)

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

consistent with markedly lower affinities (Figures 1A and 1B). More importantly, the higher percentages observed in these experiments were obtained from oligonucleotides in which the orientation of the M-CAT read from the 5′ to 3′ direction relating to the DNA strand, unlike the other three, where the motif was oriented in the 3′ to 5′ direction.

These observations were verified by performing binding assays based on fluorescence anisotropy determinations, which provided the actual dissociation constant (K_D) of each selected complex (see STAR Methods for details). As shown in Figure 1C, two distinctive groups could be readily distinguished with K_D values that differed by at least one order of magnitude. The group with lower values, corresponding to higher binding affinities, included the *CTGF* promoter, *SRF* promoter, and *C-MYC* first exon complexes, in agreement with the results afforded by the native ESI-MS determinations. The higher K_D values (lower affinities) were instead provided by the *C-MYC* enhancer, *GLUT1* enhancer, and *GLUT1* first exon complexes, which display the M-CAT sequence in the 3′ to 5′ orientation. Altogether, these results provided evidence that the M-CAT motif orientation strongly affects the binding affinity of the dsDNA constructs to TEAD1-DBD, whereas the sequence of the strand surrounding the M-CAT motif has much lower, but nevertheless still significant, influence.

Effect of M-CAT orientation on structure of the complex

The application of complementary structural proteomics approaches allowed us to structurally characterize the specific interaction between TEAD1-DBD and differentially oriented M-CAT motifs. Initially, quantitative chemical cross-linking with MS detection was implemented to reveal the spatial arrangement of free TEAD1-DBD versus bound to 3′ → 5′ or 5′ → 3′ oriented M-CATs in solution (Kukacka et al., 2015). The homobifunctional reagent disuccinimidyl adipate (DSA) was employed to conjugate susceptible primary amines (i.e., N-terminal and Lys ε-amino groups) or hydroxyl groups that may be placed within 8.9 ± 3.0 Å of one another in the structural context. The identity of cross-linked residues was then obtained by performing protease digestion and MS analysis, as detailed under STAR Methods. This strategy enabled the identification of 16 peptide conjugates that yielded 14 unique distance constraints in the control DNA-free sample (see Table 1 and Figure 2). This spatial information was subsequently employed to guide the homology modeling of DNA-free TEAD1-DBD that also included residues 22–28, which were not included in the high-resolution structure PDB: 2hzd used as a template (Anbanandam et al., 2006). The resulting model differed from the template structure only in the N-terminal region (prolonged in our construct by six amino acids), which bent closer to helix H3 than in the template. The fact that the distances of all cross-linked residues in this model fit into the maximum allowed distance for DSA confirmed that our TEAD1-DBD construct adopted the same fold exhibited in the previously published high-resolution structures (Anbanandam et al., 2006; Shi et al., 2017) (Figure S4).

A quantitative cross-linking strategy that utilized isotopically labeled versions of the reagent was subsequently implemented to evaluate TEAD1-DBD in complex with differentially oriented M-CATs (CTGF, SRF, and C-MYC exon in 5′ → 3′ and GLUT1 exon, GLUT1 enhancer, and C-MYC enhancer in 3′ → 5′ orientation). As described under STAR Methods, ^{12}C-DSA was used to label the free protein, ^{13}C-DSA to label the complex, and a 1:1 mixture of these samples was proteolytically digested and analyzed by high-performance liquid chromatography (HPLC)-MS. In the resulting spectra, the ratio of ^{12}C-DSA:^{13}C-DSA was calculated not only to detect structural changes induced by DNA binding but also to obtain a view of the actual binding interface (Kukacka et al., 2015). All 16 cross-linked peptides observed in the free TEAD1-DBD sample were also found in all six TEAD1-DBD•M-CAT complex samples. The formation of three such conjugates was not affected by DNA binding, whereas the abundances of the other two and nine of them were enhanced or reduced, respectively (the quantitative data are summarized in Table 1, whereas the conjugates are displayed on the structure in Figure 2). The N-terminal amino group showed an increased ability to form cross-links with K65 in the presence of DNA. This information should be placed in the context of the homology model obtained for free TEAD1-DBD, which showed this region bending closer to the H3 helix that was previously identified as responsible for DNA binding (Anbanandam et al., 2006). Taken together, these observations suggested that the N-terminal region might move from its position near the H3 helix upon DNA binding, placing itself closer to K65 and promoting facile DSA cross-linking. Another cross-link that manifested an increased incidence in the bound state comprised K101 connected to K57. The former is located almost at the C terminus of TEAD1-DBD, in a region that is rather unstructured in the DNA-free form (Anbanandam et al., 2006), but becomes a part of a long α helix in the TEAD4-DBD•DNA complex (Shi et al., 2017). In contrast, K57 is located on the long and flexible L1 loop that interacts with the DNA minor groove in the TEAD4-DBD•DNA complex. The fact that both residues are located in regions that lose their flexibility in the bound state might be the reason the DSA cross-link is formed with higher probability in the complexed state. On the other hand, the lost flexibility and direct interaction with DNA of the L1 loop could be responsible for the decreased abundance of K57 cross-linking to K65 and K78. Finally, all cross-links relying on K88 almost completely vanished in the bound state. This demonstrates the direct effect of DNA binding, since K88 is located precisely inside the DNA major groove in TEAD4-DBD•DNA (Shi et al., 2017) and

Figure 1. M-CAT context affects TEAD1 recognition

(A) Native ESI-MS spectra of complexes of TEAD1-DBD with each M-CAT in the study, which revealed the partitioning of free versus bound components. The most intense charge states of free protein (gray star), free DNA (brown circle), and the complex (black square) are highlighted

(B) Percentage of bound protein observed in the samples containing the various M-CATs. Data are represented as mean ± SD.

(C) Comparison of dissociation constants (K_D) of selected M-CAT·TEAD1-DBD complexes determined by fluorescence anisotropy binding assay. These values were obtained from corresponding experimental data shown in Figure S3. Complexes containing M-CATs with binding motifs in the 5′ to 3′ orientation (i.e., *SRF* promoter, *CTGF* promoter, and *C-MYC* exon) had approximately 10 times higher K_D than those with the 3′ to 5′-oriented motif. Data are represented as mean ± SD.

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

Table 1. Summary of identified cross-links and associated quantitative ratios

AA1	AA2	Sequence	CTGF	SRF	C-MYC exon	C-MYC enhancer	GLUT1 enhancer	GLUT1 exon
N-term	K65	GSHMSA-DEGKMYGR	11.10 ± 0.3	1.65	2.53	1.37	1.65	2.85 ± 0.45
N-term	K88	GSHMSA-KQVSSHIQVLAR	0.15 ± 0.02	0.18	0.25	0.69	0.59	0.58 ± 0.03
N-term	K25	GSHMSADKPIDN	0.86 ± 0.08	1.07	1.11	1.28	1.17	1.00 ± 0.19
K57	K65	KIILSDEGKMYGR	0.40 ± 0.03	0.55	0.47	0.66	0.75	0.84 ± 0.02
K57	K78	KIILSDEGK-YIKLR	0.39 ± 0.04	0.61	0.51	0.83	0.69	0.76 ± 0.02
K57	K88	KIILS-KQVSSHIQVLAR	0.41 ± 0.1	0.25	0.20	0.55	0.60	0.74 ± 0.1
K57	K88	KIILSDEGK-KQVSSHIQVLAR	0.41 ± 0.04	0.67	0.43	0.76	0.68	0.77 ± 0.02
K57	K101	KIILSDEGK-KSRD	0.95 ± 0.04	1.37	1.41	1.90	2.52	1.87 ± 0.07
K57	K101	RKIILSDEGK-KSRD	1.08 ± 0.05	1.60	1.79	2.18	3.10	2.03 ± 0.06
S61	K65	IILSDEGKMYGR	1.01 ± 0.05	0.83	0.88	0.87	0.92	1.03 ± 0.04
K65	K78	DEGKMYGR-YIKLR	0.69 ± 0.06	1.03	1.14	0.82	0.85	0.89 ± 0.01
K65	K88	DEGKMYGR-KQVSSHIQVLAR	0.25 ± 0.04	0.11	0.12	0.47	0.53	0.74 ± 0.02
K78	K88	YIKLR-KQVSSHIQVLAR	0.16 ± 0.02	0.22	0.27	0.56	0.56	0.49 ± 0.05
K83	K88	TGKTR-KQVSSHIQVLAR	0.16 ± 0.02	0.30	0.32	0.59	0.61	0.52 ± 0.03
K88	K101	KQVSSHIQVLAR-KSRD	0.27 ± 0.04	0.31	0.31	0.74	0.83	0.71 ± 0.09
K88	S91	KQVSSHIQVLAR	0.14 ± 0.03	0.24	0.25	0.60	0.68	0.62 ± 0.03

Values in the table correspond to the ratio of ^{13}C-labeled cross-links (complex) over ^{12}C (DNA-free). Values lower/higher than 1 thus indicate how many times the probability of cross-link formation decreased/increased in the presence of each M-CAT.

thus the DNA presence makes this residue inaccessible to solvent or chemical probe. Together, the cross-linking results suggested that DNA binding affected not only the previously identified binding interface (H3 helix and L1 loop), but also C- and N-terminal regions that might contribute to stabilizing the protein in the complex, or the interaction with the transactivation domain or other regulatory proteins. However, further investigation utilizing longer protein constructs will be necessary to properly investigate this effect.

Upon comparison of the quantitative cross-linking results of the six M-CATs differing in motif orientation, the first thing that could be observed is the fact that the same cross-links were identified in all samples, and in each group with the same motif orientation, the quantification results were also very similar. The only discrepancies induced by the various M-CATs' contexts were likely caused by the different affinities of TEAD1-DBD to each M-CAT, since the quantification results correlate well with the measured dissociation constants. When the equilibrium between unbound and bound complex components is shifted toward the former by lower affinities, the greater population of free protein could manifest higher reactivity with the cross-linking reagent and even form cross-links that are not favorable in the bound state. This effect is even stronger when the motifs in different orientations are compared. For instance, the ratio of ^{12}C- to ^{13}C-DSA was 0.49 for the K78-K88 cross-link in the complex with *GLUT1* exon M-CAT oriented in the 3′ → 5′ direction. This value was significantly lower than the ratio of 1 expected from unaffected residues, but also significantly higher than the 0.16 observed in the complex with the 5′ → 3′-oriented *CTGF*.

Another effect that can be observed in the quantification results shown in Table 1 is the fact that, unlike the decrease in cross-link formation, which is highly affected by the dissociation constant of the complex and therefore the motif orientation, the cross-links that are favored in the bound form seem unaffected by the motif orientation and the quantification results are similar for all tested M-CATs. As discussed above, the reason these cross-links are allowed to preferentially form in the bound state is probably the loss of flexibility of the unstructured regions in complex with DNA and the apparent independence of the dissociation constant, and thus it might be caused by the two effects (association/dissociation with DNA and loss of flexibility) happening on different time scales. Nevertheless, the fact that all M-CATs produced similar results indicates that the overall complex structure may not be significantly affected by the orientation of the consensus motif, since the same residues were cross-linked in the presence of all M-CATs tested in the study and the only differences were likely caused by different affinities of TEAD1-DBD to DNA (see Table 1).

Mapping the binding interface

Although chemical cross-linking can reveal the mutual positions of bound structures, it may not be necessarily capable of identifying the residues placed in the contact interface, which must be typically inferred from the models derived from the cross-linking data. For this reason, complementary information is usually sought by performing HDX experiments (Rozbesky et al., 2012), which allow us to observe the changes in solvent accessibility of amide hydrogens or changes in their involvement in secondary structure formation via the hydrogen bonding network upon ligand binding. In our case, we compared the exchange pattern produced by free TEAD1-DBD with those observed for its complexes with the six M-CAT constructs differing in the binding motif orientation.

The residue resolution achieved by these experiments was greatly enhanced by utilizing the aspartic protease aspergillopepsin (protease type XIII) to carry out the bottom-up characterization of the protein components (see STAR Methods). Upon online digestion, the ensuing peptides were mapped by data-dependent liquid chromatography-tandem mass spectrometry

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

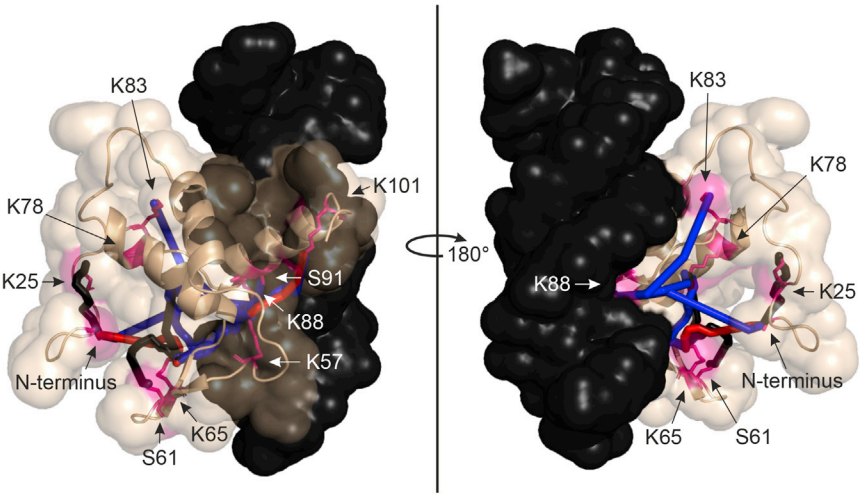


Figure 2. Identified cross-links displayed on a TEAD1-DBD•M-CAT model

Cross-links favored in the complex state are colored red, cross-links hampered by DNA binding are colored blue, and cross-links that formed independently on DNA are colored black. Residues susceptible to the DSA cross-linking reagent are highlighted by pink. For the table of all identified cross-links and validation of the cross-linking experiment see Table 1 and Figure S4.

analysis in positive-ion mode. Figure S5C demonstrates that the selected procedure generated a large variety of overlapping products, which resulted in 100% sequence coverage. The number of exchanged deuterium atoms versus the deuteration time for all identified peptides was calculated and, after recalculation to smaller regions with the use of the information from overlapping peptides, all significant (>2× SD) differences between the free and the in-complex (with each of the six studied DNA duplexes) state of TEAD1-DBD represented by a color palette were mapped on the structure model of TEAD1-DBD/M-CAT.

As is shown in Figure S5A, each of the three complexes with the same motif orientation produced a very similar protection pattern, with the same protein regions protected from deuteration compared with the DNA-free state, and even the intensity of the protection varied only slightly. However, these mild variations were in perfect agreement with the dissociation constants of each complex (see Figure S5B for examples of deuterium uptake plots of selected peptides). Figure 3 thus highlights the differences between the free and the bound states of the two tested motif orientations via color coding of the structures of TEAD1-DBD·*C-MYC* exon (5′ → 3′ oriented) and TEAD1-DBD·*C-MYC* enhancer (3′ → 5′ oriented) complexes.

For all M-CATs at short deuteration times, the largest difference in deuterium uptake between free and bound states was observed in helix H3 and the adjacent L2 loop, while more limited deuteration protection was observed also in the L1 loop. As discussed above, H3 and L1 were previously identified as being directly responsible for DNA binding. Although prior NMR studies suggested that L2 was affected by DNA binding in the TEAD1-DBD•M-CAT complex (Anbanandam et al., 2006), the crystal structure of the TEAD4-DBD•M-CAT complex did not show any direct contact between this loop and DNA (Shi et al., 2017), thus failing to support firm conclusions. At longer deuteration times, other regions that are not in contact with DNA in the crystal structure (i.e., helices H1 and H2) displayed significant protection in the bound form as well. Keeping in mind the fact that HDX is always affected not only by shielding of the structure by bound ligand but also by the changes to protein structure dynamics, especially higher amounts of hydrogen bonding, the protection observed in longer deuteration times could probably be attributed to the stabilization of the TEAD1-DBD structure in a more fixed conformation, resulting in more hydrogen bonding and a lower local exchange rate. On the other hand, the protection at short deuteration times is likely a result of the interplay of the two effects—reduced solvent accessibility caused by shielding by DNA and stabilization of the protein structure with increased hydrogen bonding. Together, the obtained results point to a large loss of structural dynamics happening in the presence of any of the tested DNA duplexes and identify helix H3 and both L1 and L2 loops as the presumable binding interface.

When the results obtained from complexes are compared with those from M-CATs present in different orientations, the most noticeable difference is the intensity of the observed protection (Figure 3). The most affected region at 10 s deuteration time was in both cases the H3 helix together with the L2 loop. With longer deuteration times this protection weakens to the point of not being significant, in contrast to the protection of the H1 and H2 helices, which dissipates more slowly. The only difference between the two motif orientations is the timescale on which this happens, which is in perfect agreement with the effect that was observed in the cross-linking study. As discussed above, decreased protection in the case of the low-affinity inverted motif was likely caused by intrinsic differences in dissociation constants, with higher populations of unbound components decreasing the average deuteration levels. With no other observable difference between the two states, the results suggest that the region of TEAD1-DBD responsible for DNA binding is the same for both motif orientations.

TEAD1 binds inverted M-CAT in a rotated orientation

The complementary cross-linking and HDX information did not reveal any significant variations in the putative mechanisms used by TEAD1-DBD to recognize the different M-CAT orientations. At the same time, however, these experiments could not rule out the possibility that the protein may bind M-CATs with the inverted 5′-CCTTA-3′ consensus sequence by using the same interacting region, but in an actual orientation of the entire protein rotated by 180°. To test this hypothesis, we carried out molecular docking experiments by using Amber14 software (Case et al., 2014) (see STAR Methods). Protein-DNA recognition relies not only on specific interactions supporting a direct readout (Anderson et al., 1987), such as hydrogen bonds and hydrophobic interactions, but also on indirect interactions mediated by water (Biedermannová and Schneider, 2016; Patikoglou

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

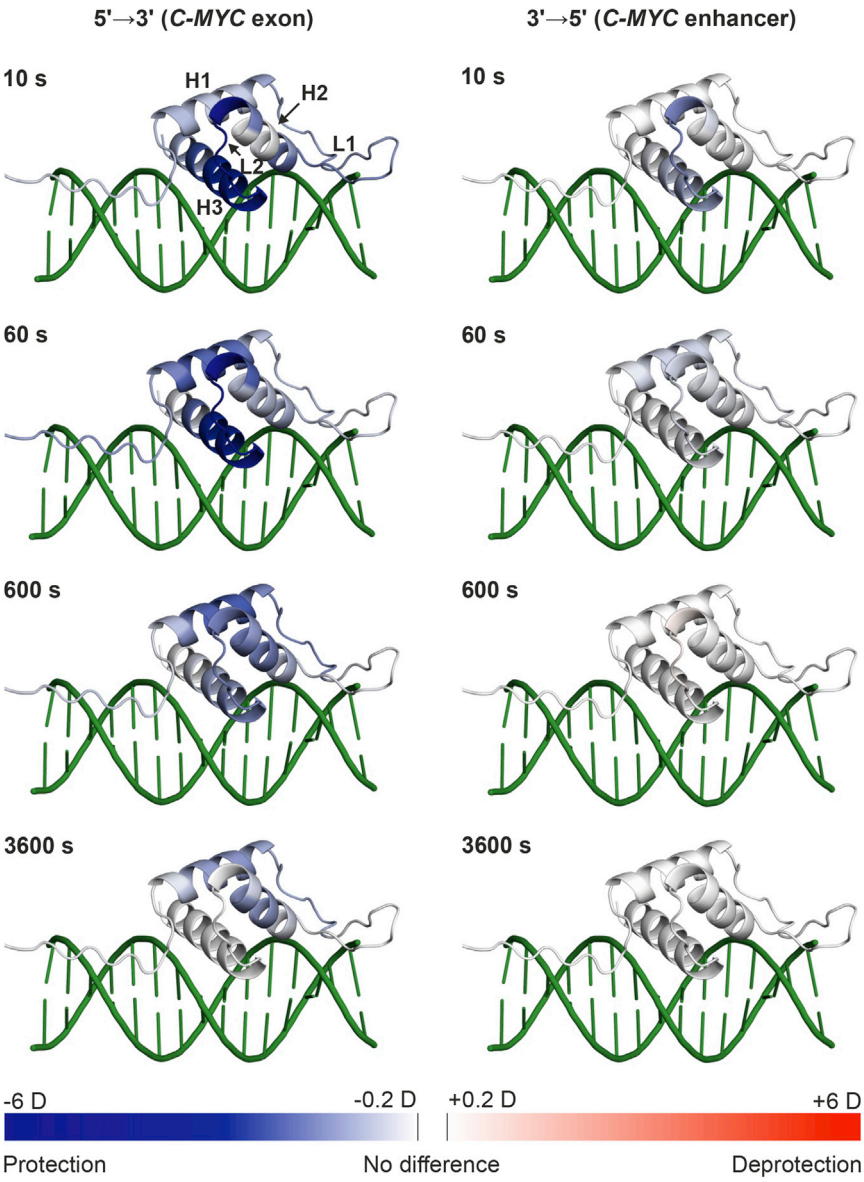


Figure 3. Mapping the binding interface

Structural changes induced by binding of TEAD1-DBD to 5′ → 3′ C-MYC exon (left) and 3′ → 5′ C-MYC enhancer (right) projected onto the structure of DNA-bound TEAD1-DBD. The deuteration level (in number of Ds) of the DNA-bound form was subtracted from the deuteration level of free TEAD1-DBD for the selected time points (10–3600 s, indicated in the top left corner of each structure) and the differences were used to replace B-factor values of each residue and subsequently color the structure using a blue-white-red gradient. Differences in negative values represent lower deuteration in the DNA-bound form compared with the DNA-free one and are colored in blue tones. In contrast, higher deuteration in the DNA-bound state would manifest in red tones. No or insignificant (−0.2 to 0.2 D) differences are in white. A much stronger effect induced by solvent protection, structure compaction, and changes in hydrogen bonding was detected for 5′ → 3′ C-MYC exon (left) than for 3′ → 5′ C-MYC enhancer (right). For more information regarding the H/D exchange experiment, see Figure S5.

and Burley, 1997; Schneider et al., 2014; Woda et al., 1998) and on the balance of electrostatic potentials and mutual shape adaptation (Dror et al., 2014; Lavery, 2005). Docking experiments can account for the overall combination of these disparate factors and enable a direct comparison of the binding abilities of different M-CATs. For this reason, we prepared a series of 15 models of TEAD1-DBD bound to dsDNA constructs based on the homologous TEAD4-DBD•M-CAT structure (PDB: 5gzb) (Shi et al., 2017), which placed the sequences of the *C-MYC* exon and *C-MYC* enhancer in different structural contexts (Figure 4 and Table S2). The models were subjected to 20 ns molecular dynamics (MD) simulations. Regular snapshots along the simulation trajectories were examined to assess the local DNA backbone conformation. In addition, the molecular mechanics Poisson Boltzmann surface area method was employed to calculate the ΔG of binding and predict the putative affinity of the protein-DNA interaction.

Figure 4A displays the structures of the two TEAD1-DBD•M-CAT complexes that manifested the most stable interactions according to the ΔG calculations carried out before and after docking and MD simulations. The most stable interaction corresponded to the TEAD1-DBD•*C-MYC* exon complex, followed by the *C-MYC* enhancer sequence, with the M-CAT motif modeled in the complementary strand while the latter is equivalent to a 180° rotation of the entire TEAD1-DBD (Figure 4B). These two forms displayed a difference of about 0.5 kcal/mol between one another, which would correspond to an ~2.4-fold difference in the respective dissociation constants (K_D). Consistent with the results of structural proteomics analysis, the simulations revealed that the distributions of local conformations assumed by the deoxyribose-phosphate backbone in the interface regions were similar for the two M-CAT orientations. Moreover, a similar range of A-B and B-A DNA conformers as in the template crystal structure was observed for either orientation, which in turn accounted for the similar shapes and widths of the interacting grooves (Černý et al., 2016; Schneider et al., 2018). The similarity of the interacting grooves showed that it was possible for helix H3 to fit in the DNA major groove regardless of the motif orientation. Moreover, the ΔG calculations suggested that in the case of the inverted M-CAT motif, TEAD1-DBD compensates for the inversion by binding in a 180° rotated orientation, which could lead to less stable interactions between protein and DNA. For this reason, the simulations provided a very plausible explanation for the widely different affinities exposed by the binding experiments.

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

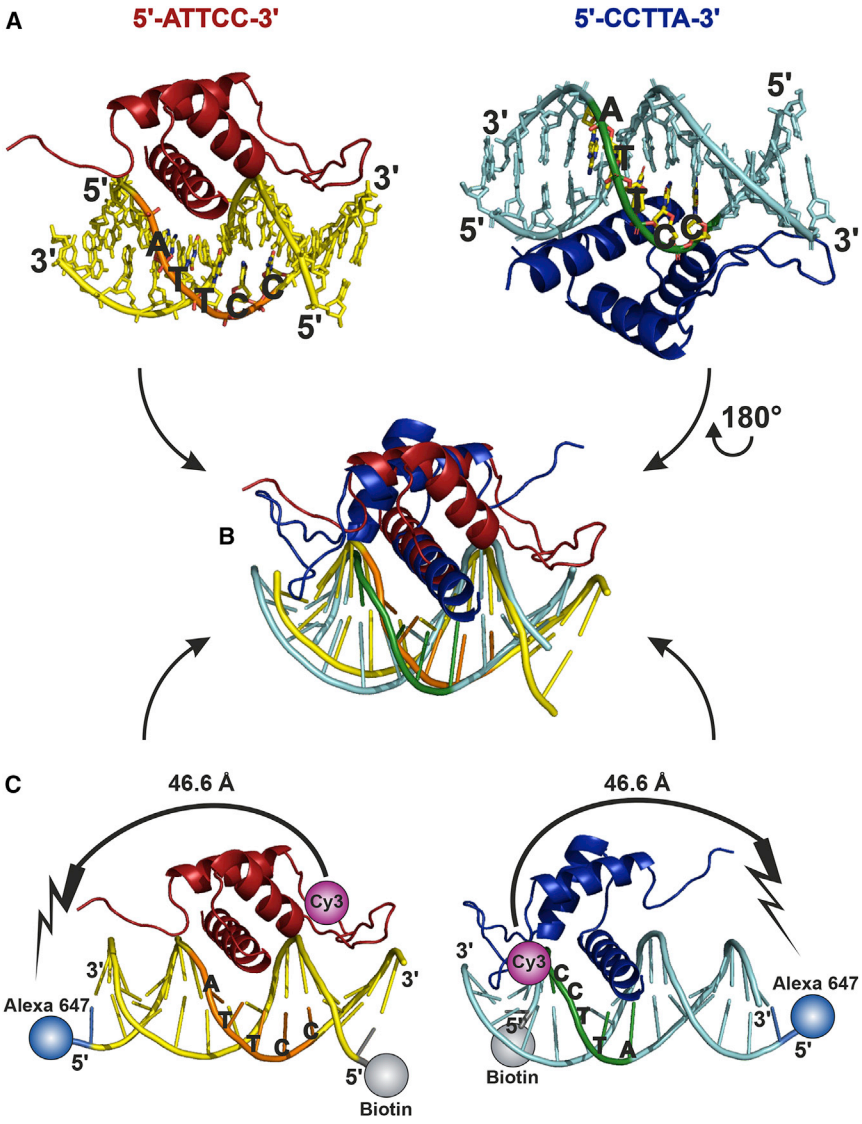


Figure 4. TEAD1 binds inverted M-CAT in a rotated orientation

(A) TEAD1-DBD●M-CAT models used for MD simulations showing the relative position and orientation of the C-MYC enhancer 5′-CCTTA and the C-MYC exon 5′-ATTCC DNA sequences with respect to the TEAD1-DBD.

(B and C) (B) Structure superposition of DNA constructs containing the C-MYC exon and enhancer motifs corresponding to the most stable interactions according to ΔG calculations, which were later confirmed by (C) an smFRET study, where DNA and protein were labeled with donor and acceptor fluorophores whose distance (and thus FRET effectivity) depended on the respective orientations of the protein and DNA.

To support the computational data experimentally, a single-molecule Förster resonance energy transfer (smFRET) study was carried out. For one oligonucleotide of each M-CAT orientation a 16 bp long version with asymmetrically placed binding motif and an Alexa 647 label was designed and Cys53 of the TEAD1-DBD protein was modified with Cy3 dye (for sequences and label positions see Figure 4C and Table S3). The oligonucleotides with the higher-affinity 5′-ATTCC-3′ motif served as control, where the acceptor (Alexa 647) was either on the 5′ end of the forward strand or the 5′ end of the reverse strand and the donor (Cy3) was on Cys53. Based on the structural model (Shi et al., 2017), the distance between fluorophores is 46.6 Å (forward strand labeled) or 25.4 Å (reverse strand labeled) and, as expected, energy transfer efficiency of 0.58 and 1.00, respectively, was observed in the smFRET experiment. The lower-affinity 5′-CCTTA-3′ motif had an acceptor on the 5′ end of the reverse strand and the donor was on Cys53 as well. Depending on the relative orientation of the protein with respect to the DNA, the distance between donor and acceptor could be either 46.6 Å (in the case that our assumption is correct and the protein binds this motif in a 180° rotated orientation) or 28.2 Å (if the orientation of TEAD1-DBD toward the DNA is similar to the published structural model). The shorter 28.2 Å distance would result in optimal FRET with energy transfer efficiency close to 1, whereas the longer distance would lead to lower efficiency. In our case, we observed an energy transfer efficiency of 0.51 (similar to the high-affinity motif), which confirms the results obtained by molecular docking simulations together with the idea of 180° rotated binding orientation.

Inverted M-CAT sites are occupied by TEAD1 *in vivo*

The DNA sequences selected for our study were initially subjected to a bioinformatics search of the human genome, which revealed the presence of M-CAT motifs in both orientations near typical signatures of active chromatin (i.e., H3K27Ac, H3K4Me1, and DNase sensitivity markers) (see Figure S1). This observation suggests a potential occupancy of all tested M-CATs by TEAD proteins inside human cells. Furthermore, we have also previously reported that *C-MYC* expression in T cell acute lymphoblastic leukemia cells is regulated by TEAD1 (Vališ et al., 2016) and two of the selected sequences were derived from regulatory regions of the *C-MYC* gene: one from the first exon (5′ → 3′ oriented) and the other from the enhancer (~2,000 bp from the transcription start site, 3′ → 5′ oriented). To assess possible differences in the recognition mechanisms *in vivo*, we determined the relative rates of TEAD1 occupancy by using chromatin immunoprecipitation (ChIP) analysis followed by qPCR quantification (see STAR Methods). First, primers for regions containing the M-CAT sequences and control primers for intervening stretches (i.e., ~1,000 bp from the transcription start site) were designed. All primers were then tested for specificity and effectivity. Melting curve analysis of the resulting PCR products suggested specific amplification of regions of interest. The

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

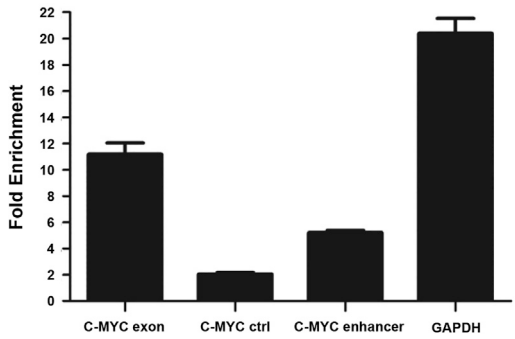


Figure 5. Relative occupancy of selected *C-MYC* regions by TEAD1
ChIP analysis followed by relative quantification of selected regions was used. Individual quantities are shown as fold enrichment of TEAD1 ChIP samples against negative control IgG ChIP samples. C-MYC ctrl consisted of a control region without any M-CAT motif. ChIP using RNA polymerase II antibody and primers to the GAPDH promoter was used as a positive control. Data are represented as mean ± SD.

effectiveness of individual primers was calculated as 100% (±5%) for each set of primers. ChIP analysis used rabbit antibody against TEAD1, rabbit isotype IgG as a negative control, and RNA polymerase II antibody together with primers to glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) as a positive control. The resulting data showed high-level occupancy of the *C-MYC* exon by TEAD1 (11.2× enrichment against negative IgG), significantly lower occupancy of the *C-MYC* enhancer by TEAD1 (5.3× enrichment against negative IgG), and non-significant occupancy of the control region by TEAD1 (2× enrichment against negative IgG). This was in excellent agreement with our *in vitro* observations, since the enhancer region contained an inverted M-CAT motif, whereas the first exon region included the M-CAT motif oriented in the 5′ → 3′ direction (Figure 5).

Additional information was provided by a more detailed analysis of the positions of M-CAT sequences in the context of individual DNA strands and their orientation versus the direction of transcription of regulated genes. The results showed that all high-affinity M-CATs shared a core sequence of 5′-ATTCC-3′ localized on the non-transcribed DNA strand, whereas all low-affinity M-CATs shared an inverted 3′-ATTCC-5′ sequence on the transcribed strand. This fact suggested that distinctive orientations of TEAD transcription factors might have a significant impact on intact chromatin with diverse possible effects on transcriptional outputs.

Conclusions

In this report, we employed complementary techniques to investigate the structural determinants and the possible modes of TEAD1-DBD interactions with its DNA response motif. The results provided valuable information on the binding interface and surrounding regions, some of which were not immediately evident from the available high-resolution structures (Anbanandam et al., 2006; Shi et al., 2017). On one hand, the placement of helix H3 and the L1 loop in direct contact with the DNA was in excellent agreement with recently published data. On the other, the deuteration patterns revealed by HDX experiments suggested that DNA binding might stabilize the conformation of helices H1 and H2 in the complex, even though these structures are not directly involved in actual contacts.

The systematic examination of a series of M-CAT constructs has shown that the structural context of the consensus motif may induce detectable effects on the strength of the binding interaction. The highest affinity was observed for the *CTGF* M-CAT with the sequence 5′-AGTCACATTCCTCCG-3′, which contained the highest-affinity ACATTCCT motif predicted by detailed transcription factor flexible models accessible in JASPAR (Khan et al., 2018) and confirmed by protein-binding chip experiments (Anbanandam et al., 2006). Progressively lower affinities were observed for M-CATs with different flanking sequences surrounding the core ATTCC motif.

Moreover, the inverted 5′-CCTTA-3′ motif was found to be able to bind TEAD1-DBD with lower affinity. The differences between M-CATs containing motifs in either the 5′ → 3′ or the 3′ → 5′ orientation were immediately noticeable in the outcomes of native ESI-MS and fluorescence anisotropy binding assays, but also in those of quantitative cross-linking and HDX experiments. In fact, 5′ → 3′ orientations produced more significant alterations of cross-link formation and deuteration rates upon binding than the inverted 3′ → 5′ counterparts. The results of molecular docking experiments indicated not only that the different orientation might contribute to an average stabilization of 0.5 kcal/mol, but also that TEAD1-DBD might have a marked preference for binding 3′ → 5′ M-CATs in an orientation rotated by 180°, seemingly to compensate for the inverted motif. The results of the smFRET study thereafter proved this assumption to be true and showed that TEAD1-DBD actually binds the inverted motif in 180° rotated orientation. The MD simulations have also shown that the two M-CAT orientations produced similar distributions of local conformations, which resulted in major grooves sharing similar shapes, thus explaining the ability of TEAD1-DBD to establish stable interactions in either arrangement. Taken together, these observations suggest that TEAD1-DBD may at first recognize the overall shape of the major groove and that the specific interactions necessary to stabilize complex association may be established only when the H3 helix is properly seated inside the groove. The fact that the number and strength of these specific amino acid-nucleotide interactions are influenced by M-CAT orientation may explain the different affinities revealed by our experiments.

At the end, the presence in the human genome of M-CAT sites with widely different affinities may provide the basis for possible regulatory mechanisms relying on the actual concentration of a certain transcription factor in the cell to ensure site occupancy. Our experiments revealed that site occupancy was affected by the orientation of the M-CAT motif in living cells, with sites that exhibited a 5′ → 3′ orientation displaying significantly greater occupancy than those with 3′ → 5′ orientation. On one hand, these results together with the fact that the M-CAT motifs examined in this study were located very close to markers of active chromatin support the possible physiological relevance of differentially oriented M-CATs. On the other, the situation in living cells is never isolated to just one protein interacting with DNA, and in the case of TEAD proteins, there are a number of known interaction partners already known, with new ones emerging every year, and there is a chance that some of them may affect the TEAD-DNA

interaction and possibly the M-CAT site occupancy (Lin et al., 2017). Despite that, our results are consistent with recent reports that, depending on the specific transcription factor, the inversion of the binding motif may have negative effects on transcription output in *S. cerevisiae* (Sharon et al., 2012), to the point where total loss of transcriptional output could be observed in *Ciona intestinalis* (Farley et al., 2016). Moreover, other reports show an inhibitory effect of low-affinity sites on the transcription of regulated genes enabling precise tuning of transcriptional outputs in reliance on individual transcription factor levels (Ridinger-Saison et al., 2012). In the case of the C-MYC transcription factor, the presence of both low- and high-affinity sites has been interpreted as the basis for a fine-tuned regulatory mechanism (Lorenzin et al., 2016). According to this mechanism, low C-MYC concentrations may induce exclusive occupancy of high-affinity sites, whereas high concentrations can lead to saturation of high-affinity sites, as well as binding to low-affinity ones. For this reason, the high C-MYC concentrations detected in cancer cells have been implicated as the possible trigger for the distinctive transcription patterns associated with cancer (Lorenzin et al., 2016; Zheng and Levens, 2016). Similar regulation was suggested also for other transcription factors (Wang et al., 2015). Our data strongly suggest that analogous regulatory mechanisms may be supported by the ability of TEAD transcription factors to bind differentially with specific M-CAT sites. This functional hypothesis, together with the way it cooperates with other known regulation mechanisms of TEAD proteins and its significance for the development of possible therapeutic strategies, will be the object of future studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCE TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Chemicals
  - Protein expression and Cy3 labeling
  - Design of oligonucleotides and preparation of dsDNA
  - Melt curves of individual DNA duplexes
  - Preparation of protein-DNA complexes
  - Fluorescence anisotropy binding assay
  - Quantitative chemical cross-linking
  - Hydrogen-deuterium exchange
  - Homology modelling
  - Molecular docking
  - Molecular dynamics
  - Single-molecule förster resonance energy transfer
  - Chromatin immunoprecipitation and qPCR
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Fluorescence anisotropy binding assay
  - Quantitative chemical cross-linking
  - Hydrogen-deuterium exchange

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.str.2020.11.018.

ACKNOWLEDGMENTS

We acknowledge the Czech Science Foundation (grant 16-24309S), the Ministry of Education of the Czech Republic (project LH15010, program "NPU II" project LQ1604, and "NPU I" project LO1509), European Commission H2020 (EPIC-XS grant agreement ID: 823839), the Charles University Grant Agency (1618218), and, in part, the Czech Academy of Sciences (RVO61388971) for financial support. We further acknowledge the Centre for Molecular Structure Core Facility at BIOCEV, a facility funded by European Regional Development Funds (CZ.1.05/1.1.00/02.0109 BIOCEV) and supported by the Czech Infrastructure for Integrative Structural Biology (LM2015043 CIISB for CMS BIOCEV funded by MEYS CR). The FP7 WeNMR (project 261572) and H2020 West-Life (project 675858) European e-Infrastructure projects are acknowledged for the use of their web portals, which make use of the EGI infrastructure and DIRAC4EGI service with the dedicated support of CESNET-MetaCloud. The authors also acknowledge support from the EU (Operational Program Prague–Competitiveness Project, CZ.2.16/3.1.00/24023).

AUTHOR CONTRIBUTIONS

Conceptualization, P.N., K.V., and R.F.; Methodology, R.F., L.S., J.F., J.C., J.Ch., P.M., T.Ch., M.C., D.F., and P.N.; Software, K.V., D.K., and M.R.; Investigation, R.F., K.V., J.C., and T.Ch.; Writing – Original Draft, R.F., K.V., J.C., J.Ch., and M.C.; Writing – Review & Editing, R.F., D.F., and P.N.; Visualization, R.F., J.C., and J.Ch; Supervision, P.N. and M.C.; Funding Acquisition, P.N. and R.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 27, 2020
Revised: October 9, 2020
Accepted: November 24, 2020
Published: December 16, 2020

REFERENCES

Anbanandam, A., Albarado, D.C., Nguyen, C.T., Halder, G., Gao, X., and Veeraraghavan, S. (2006). Insights into transcription enhancer factor 1 (TEF-1) activity from the solution structure of the TEA domain. Proc. Natl. Acad. Sci. U S A *103*, 17225–17230.

Anderson, J.E., Ptashne, M., and Harrison, S.C. (1987). Structure of the represser–operator complex of bacteriophage 434. Nature *326*, 846–852.

Barylyuk, K., Gülbakan, B., Xie, X., and Zenobi, R. (2013). DNA Oligonucleotides: a model system with tunable binding strength to study monomer–dimer equilibria with electrospray ionization-mass spectrometry. Anal. Chem. *85*, 11902–11912.

Biedermannová, L., and Schneider, B. (2016). Hydration of proteins and nucleic acids: advances in experiment and theory. A review. Biochim. Biophys. Acta *1860*, 1821–1835.

Boura, E., Silhan, J., Herman, P., Vecer, J., Sulc, M., Teisinger, J., Obsilova, V., and Obsil, T. (2007). Both the N-terminal loop and wing W2 of the forkhead domain of transcription factor Foxo4 are important for DNA binding. J. Biol. Chem. *282*, 8265–8275.

Brunger, A.T. (2007). Version 1.2 of the crystallography and NMR system. Nat. Protoc. *2*, 2728–2733.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D. Biol. Crystallogr. *54*, 905–921.

Case, D.A., Babin, V., Berryman, J., Betz, R., Cai, Q., Cerutti, D., Cheatham, T.E., Darden, T.A., Duke, R., Gohlke, H., et al. (2014). Amber 14 (Univ. California).

Černý, J., Božíková, P., and Schneider, B. (2016). DNATCO: assignment of DNA conformers at dnatco.org. Nucleic Acids Res. *44*, W284–W287.

Chen, Z., Friedrich, G.A., and Soriano, P. (1994). Transcriptional enhancer factor 1 disruption by a retroviral gene trap leads to heart defects and embryonic lethality in mice. Genes Dev. *8*, 2293–2301.

Dong, J., Feldmann, G., Huang, J., Wu, S., Zhang, N., Comerford, S.A., Gayyed, M.F., Anders, R.A., Maitra, A., and Pan, D. (2007). Elucidation of a universal size-control mechanism in Drosophila and mammals. Cell *130*, 1120–1133.

Dror, I., Zhou, T., Mandel-Gutfreund, Y., and Rohs, R. (2014). Covariation between homeodomain transcription factors and the shape of their DNA binding sites. Nucleic Acids Res. *42*, 430–441.

Edelstein, A.D., Tsuchida, M.A., Amodaj, N., Pinkard, H., Vale, R.D., and Stuurman, N. (2014). Advanced methods of microscope control using μManager software. J. Biol. Methods *1*, 10.

Farley, E.K., Olson, K.M., Zhang, W., Rokhsar, D.S., and Levine, M.S. (2016). Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc. Natl. Acad. Sci. U S A *113*, 6508–6513.

Fiala, J., Kukačka, Z., and Novák, P. (2020). Influence of cross-linker polarity on selectivity towards lysine side chains. J. Proteomics *218*, https://doi.org/10.1016/j.jprot.2020.103716.

Filandrová, R., Kavan, D., Kádek, A., Novák, P., and Man, P. (2021). Studying protein–DNA interactions by hydrogen/deuterium exchange mass spectrometry. In Multiprotein Complexes: Methods and Protocols (Methods in Molecular Biology (2247), A. Poterszman, ed. (Springer US).

Götze, M., Pettelkau, J., Schaks, S., Bosse, K., Ihling, C.H., Krauth, F., Fritzsche, R., Kühn, U., and Sinz, A. (2012). StavroX-A software for analyzing crosslinked products in protein interaction studies. J. Am. Soc. Mass Spectrom. *23*, 76–87.

Götze, M., Pettelkau, J., Fritzsche, R., Ihling, C.H., Schäfer, M., and Sinz, A. (2015). Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis. J. Am. Soc. Mass Spectrom. *26*, 83–97.

Hagan, N., and Fabris, D. (2003). Direct mass spectrometric determination of the stoichiometry and binding affinity of the complexes between nucleocapsid protein and RNA stem-loop hairpins of the HIV-1 ψ-recognition element. Biochemistry *42*, 10736–10745.

Hwang, S., Guo, Z., and Kuznetsov, I.B. (2007). DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics *23*, 634–636.

Jacquemin, P., Hwang, J.J., Martial, J.A., Dollé, P., and Davidson, I. (1996). A novel family of developmentally regulated mammalian transcription factors containing the TEA/ATTS DNA binding domain. J. Biol. Chem. *271*, 21775–21785.

Jacquemin, P., Sapin, V., Alsat, E., Evain-Brion, D., Dollé, P., and Davidson, I. (1998). Differential expression of the TEF family of transcription factors in the murine placenta and during differentiation of primary human trophoblasts in vitro. Dev. Dyn. *212*, 423–436.

Jensen, P.F., Larraillet, V., Schlothauer, T., Kettenberger, H., Hilger, M., and Rand, K.D. (2015). Investigating the interaction between the neonatal Fc receptor and monoclonal antibody variants by hydrogen/deuterium exchange mass spectrometry. Mol. Cell. Proteomics *14*, 148–161.

Jones, D.T., and Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics *31*, 857–863.

Joshi, S., Davidson, G., Le Gras, S., Watanabe, S., Braun, T., Mengus, G., and Davidson, I. (2017). TEAD transcription factors are required for normal primary myoblast differentiation in vitro and muscle regeneration in vivo. PLoS Genet. *13*, e1006600.

Kaan, H.Y.K., Chan, S.W., Tan, S.K.J., Guo, F., Lim, C.J., Hong, W., and Song, H. (2017). Crystal structure of TAZ-TEAD complex reveals a distinct interaction mode from that of YAP-TEAD complex. Sci. Rep. *7*, 2035.

Kahraman, A., Malmström, L., and Aebersold, R. (2011). Xwalk: computing and visualizing distances in cross-linking experiments. Bioinformatics *27*, 2163–2164.

Kaneko, K.J., and DePamphilis, M.L. (1998). Regulation of gene expression at the beginning of mammalian development and the TEAD family of transcription factors. Dev. Genet. *22*, 43–55.

Kaneko, K.J., Kohn, M.J., Liu, C., and DePamphilis, M.L. (2007). Transcription factor TEAD2 is involved in neural tube closure. Genesis *45*, 577–587.

Kapanidis, A.N., Lee, N.K., Laurence, T.A., Doose, S., Margeat, E., and Weiss, S. (2004). Fluorescence-aided molecule sorting: analysis of structure and interactions by alternating-laser excitation of single molecules. Proc. Natl. Acad. Sci. U S A. *101*, 8936–8941.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., Van Der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2018). Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. *46*, D260–D266.

Kukacka, Z., Rosulek, M., Strohalm, M., Kavan, D., and Novak, P. (2015). Mapping protein structural changes by quantitative cross-linking. Methods *89*, 112–120.

Kuznetsov, I.B., Gou, Z., Li, R., and Hwang, S. (2006). Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins *64*, 19–27.

Lavery, R. (2005). Recognizing DNA. Q. Rev. Biophys. *38*, 339–344.

Li, Z., Zhao, B., Wang, P., Chen, F., Dong, Z., Yang, H., Guan, K.-L., and Xu, Y. (2010). Structural insights into the YAP and TEAD complex. Genes Dev. *24*, 235–240.

Lin, K.C., Park, H.W., and Guan, K.L. (2017). Regulation of the Hippo pathway transcription factor TEAD. Trends Biochem. Sci. *42*, 862–872.

Liu, R., Lee, J., Kim, B.S., Wang, Q., Buxton, S.K., Balasubramanyam, N., Kim, J.J., Dong, J., Zhang, A., Li, S., et al. (2017). Tead1 is required for maintaining adult cardiomyocyte function, and its loss results in lethal dilated cardiomyopathy. JCI Insight *2*, e93343.

Lorenzin, F., Benary, U., Baluapuri, A., Walz, S., Jung, L.A., von Eyss, B., Kisker, C., Wolf, J., Eilers, M., and Wolf, E. (2016). Different promoter affinities account for specificity in MYC-dependent gene regulation. Elife *5*, 1–35.

Mahoney, W.M., Hong, J.-H., Yaffe, M.B., and Farrance, I.K.G. (2005). The transcriptional co-activator TAZ interacts differentially with transcriptional enhancer factor-1 (TEF-1) family members. Biochem. J. *388*, 217–225.

Mar, J.H., and Ordahl, C.P. (1990). M-CAT binding factor, a novel trans-acting factor governing muscle-specific transcription. Mol. Cell. Biol. *10*, 4271–4283.

Noland, C.L., Gierke, S., Schnier, P.D., Murray, J., Sandoval, W.N., Sagolla, M., Dey, A., Hannoush, R.N., Fairbrother, W.J., and Cunningham, C.N. (2016). Palmitoylation of TEAD transcription factors is required for their stability and function in Hippo pathway signaling. Structure *24*, 179–186.

Ovesný, M., Křížek, P., Borkovec, J., Švindrych, Z., and Hagen, G.M. (2014). ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. Bioinformatics *30*, 2389–2390.

Patikoglou, G., and Burley, S.K. (1997). Eukaryotic transcription factor-dna complexes. Annu. Rev. Biophys. Biomol. Struct. *26*, 289–325.

Pobbati, A.V., Chan, S.W., Lee, I., Song, H., and Hong, W. (2012). Structural and functional similarity between the Vgll1-TEAD and the YAP-TEAD complexes. Structure *20*, 1135–1140.

Ridinger-Saison, M., Boeva, V., Rimmelé, P., Kulakovskiy, I., Gallais, I., Levavasseur, B., Paccard, C., Legoix-Né, P., Morlé, F., Nicolas, A., et al. (2012). Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. Nucleic Acids Res. *40*, 8927–8941.

Rozbesky, D., Man, P., Kavan, D., Chmelik, J., Cerny, J., Bezouska, K., and Novak, P. (2012). Chemical cross-linking and H/D exchange for fast refinement of protein crystal structure. Anal. Chem. *84*, 867–870.

Rozbeský, D., Rosůlek, M., Kukačka, Z., Chmelík, J., Man, P., and Novák, P. (2018). Impact of chemical cross-linking on protein structure and function. Anal. Chem. *90*, 1104–1113.

Sawada, A., Kiyonari, H., Ukita, K., Nishioka, N., Imuta, Y., and Sasaki, H. (2008). Redundant roles of Tead1 and Tead2 in notochord development and the regulation of cell proliferation and survival. Mol. Cell. Biol. *28*, 3177–3189.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

Schneider, B., Černý, J., Svozil, D., Čech, P., Gelly, J.C., and De Brevern, A.G. (2014). Bioinformatic analysis of the protein/DNA interface. Nucleic Acids Res. *42*, 3381–3394.

Schneider, B., Božíková, P., Nečasová, I., Čech, P., Svozil, D., and Černý, J. (2018). A DNA structural alphabet provides new insight into DNA flexibility. Acta Crystallogr. Sect. D Struct. Biol. *74*, 52–64.

Schrangl, L., Göhring, J., and Schütz, G.J. (2018). Kinetic analysis of single molecule FRET transitions without trajectories. J. Chem. Phys. *148*, 123328.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res. *33*, 382–388.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Shi, Z., He, F., Chen, M., Hua, L., Wang, W., Jiao, S., and Zhou, Z. (2017). DNA-binding mechanism of the Hippo pathway transcription factor TEAD4. Oncogene *36*, 4362–4369.

Slavata, L., Chmelík, J., Kavan, D., Filandrová, R., Fiala, J., Rosůlek, M., Mrázek, H., Kukačka, Z., Vališ, K., Man, P., et al. (2019). Ms-based approaches enable the structural characterization of transcription factor/dna response element complex. Biomolecules *9*, 1–21.

Stephen, A.G., Datta, S.A.K., Worthy, K.M., Bindu, L., Fivash, M.J., Turner, K.B., Fabris, D., Rein, A., and Fisher, R.J. (2007). Measuring the binding stoichiometry of HIV-1 Gag to very-low-density oligonucleotide surfaces using surface plasmon resonance spectroscopy. J. Biomol. Tech. *18*, 259–266.

Strohalm, M., Kavan, D., Novák, P., Volný, M., and Havlíček, V. (2010). MMass 3: a cross-platform software environment for precise analysis of mass spectrometric data. Anal. Chem. *82*, 4648–4651.

Tian, W., Yu, J., Tomchick, D.R., Pan, D., and Luo, X. (2010). Structural and functional analysis of the YAP-binding domain of human TEAD2. Proc. Natl. Acad. Sci. U S A *107*, 7293–7298.

Vališ, K., Talacko, P., Grobárová, V., Černý, J., and Novák, P. (2016). Shikonin regulates C-MYC and GLUT1 expression through the MST1-YAP1-TEAD1 axis. Exp. Cell Res. *349*, 273–281.

Vassilev, A., Kaneko, K.J., Shu, H., Zhao, Y., and DePamphilis, M.L. (2001). TEAD/TEF transcription factors utilize the activation domain of YAP65, a Src/Yes-associated protein localized in the cytoplasm. Genes Dev. *15*, 1229–1241.

Wang, J., Malecka, A., Trøen, G., and Delabie, J. (2015). Comprehensive genome-wide transcription factor analysis reveals that a combination of high affinity and low affinity DNA binding is needed for human gene regulation. BMC Genomics *16* (*Suppl 7*), S12.

Wang, W., Kitova, E.N., and Klassen, J.S. (2003). Influence of solution and gas phase processes on Protein−Carbohydrate binding affinities determined by nanoelectrospray fourier transform ion cyclotron resonance mass spectrometry. Anal. Chem. *75*, 4945–4955.

Wassenaar, T.A., van Dijk, M., Loureiro-Ferreira, N., van der Schot, G., de Vries, S.J., Schmitz, C., van der Zwan, J., Boelens, R., Giachetti, A., Ferella, L., et al. (2012). WeNMR: structural Biology on the grid. J. Grid Comput. *10*, 743–767.

Webb, B., and Sali, A. (2014). Comparative protein structure modeling using MODELLER. Curr Protoc Bioinformatics *54*, 5.6.1–5.6.37.

Weinkam, P., Pons, J., and Sali, A. (2012). Structure-based model of allostery predicts coupling between distant sites. Proc. Natl. Acad. Sci. U S A *109*, 4875–4880.

Woda, J., Schneider, B., Patel, K., Mistry, K., and Berman, H.M. (1998). An analysis of the relationship between hydration and protein-DNA interactions. Biophys. J. *75*, 2170–2177.

Xiao, J.H., Davidson, I., Ferrandon, D., Rosales, R., Vigneron, M., Macchi, M., Ruffenach, F., and Chambon, P. (1987). One cell-specific and three ubiquitous nuclear proteins bind in vitro to overlapping motifs in the domain B1 of the SV40 enhancer. EMBO J. *6*, 3005–3013.

Xiao, J.H., Davidson, I., Matthes, H., Garnier, J.M., and Chambon, P. (1991). Cloning, expression, and transcriptional properties of the human enhancer factor TEF-1. Cell *65*, 551–568.

Yasunami, M., Suzuki, K., and Ohkubo, H. (1996). A novel family of TEA domain-containing transcription factors with distinct spatiotemporal expression patterns. Biochem. Biophys. Res. Commun. *228*, 365–370.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res. *46*, D754–D761.

Zhao, B., Wei, X., Li, W., Udan, R.S., Yang, Q., Kim, J., Xie, J., Ikenoue, T., Yu, J., Li, L., et al. (2007). Inactivation of YAP oncoprotein by the Hippo pathway is involved in cell contact inhibition and tissue growth control. Genes Dev. *21*, 2747–2761.

Zhao, B., Ye, X., Yu, J., Li, L., Li, W., Li, S., Yu, J., Lin, J.D., Wang, C.-Y., Chinnaiyan, A.M., et al. (2008). TEAD mediates YAP-dependent gene induction and growth control. Genes Dev. *22*, 1962–1971.

Zheng, Y., and Levens, D. (2016). Tuning the MYC response. Elife *5*, 2015–2017.

Zhou, Y., Huang, T., Cheng, A.S.L., Yu, J., Kang, W., and To, K.F. (2016). The TEAD family and its oncogenic role in promoting tumorigenesis. Int. J. Mol. Sci. *17*, 1–15.

van Zundert, G.C.P., Rodrigues, J.P.G.L.M., Trellet, M., Schmitz, C., Kastritis, P.L., Karaca, E., Melquiond, A.S.J., van Dijk, M., de Vries, S.J., and Bonvin, A.M.J.J. (2016). The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. J. Mol. Biol. *428*, 720–725.

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

CellPress

STAR★METHODS

KEY RESOURCE TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Polyclonal rabbit IgG anti-TEAD1	GeneTex	Cat No. GTX32918; RRID: AB_2884896
Rabbit IgG isotype negative control	GeneTex	Cat No. GTX35035; RRID: AB_10623175
Mouse IgG anti-RNA polymerase II	Active Motif	Cat No. 39497; RRID: AB_2732926
Bacterial and virus strains		
E. Coli BL21-CodonPlus (DE3)-RIPL	Agilent Technologies	Cat No. 230280
Chemicals, peptides, and recombinant proteins		
DNA binding domain of TEAD1 (22-104 region)	This paper	N/A
Cy3 maleimide	GE Healthcare	Cat No. PA23031
Disuccinimidyl adipate (DSA) ^{12}C and ^{13}C	Creative Molecules	Cat No. 013SC
Trypsin from bovine pancreas	Merck	Cat No. T1426
Endoproteinase Asp-N	Roche	Cat No. ENDOARGS-RO
Albumin, biotin labeled bovine	Merck	Cat No. A8549
Streptavidin from *Streptomyces avidinii*	Merck	Cat No. S4762
Micrococcal Nuclease	NEB	Cat No. M0247S
Proteinase K	Thermo Fisher	Cat No. EO0491
Column with immobilized Aspergillopepsin	AffiPro	Cat No. AP-PC-005
Deposited data		
Ensembl database	(Zerbino et al., 2018)	http://www.ensembl.org/index.html
JASPAR database	(Khan et al., 2018)	http://jaspar.genereg.net/
Cross-linking source data	This paper	ProteomeXchange: PXD012127 (http://www.ebi.ac.uk/pride)
HDX source data	This paper	ProteomeXchange: PXD012127 (http://www.ebi.ac.uk/pride)
Supporting data for the structural mass spectrometry experiments – deuterium uptake plots of all identified peptides and mass spectra of cross-links dissociation products	This paper	Mendeley data: 10.17632/27zkz3v729.1 (https://data.mendeley.com/datasets/27zkz3v729/1)
X-ray structure of the complex of DNA binding domain of protein TEAD4 with DNA	(Shi et al., 2017)	PDB: 5GZB (https://www.rcsb.org/structure/5GZB)
NMR structure of the DNA binding domain of protein TEAD1	(Anbanandam et al., 2006)	PDB: 2HZD (http://www.rcsb.org/structure/2HZD)
Model of DNA binding domain of protein TEAD1 based on cross-linking and HDX data	This paper	PDB-Dev: PDBDEV_00000062 (https://pdb-dev.wwpdb.org/entry.html?PDBDEV_00000062)
Model of complex of DNA binding domain of protein TEAD1 with DNA based on cross-linking and HDX data	This paper	PDB-Dev: PDBDEV_00000063 (https://pdb-dev.wwpdb.org/entry.html?PDBDEV_00000063)
Computational model of DNA binding domain of TEAD1 docked to DNA with 3'-5' oriented M-CAT motif	This paper	Model Archive: 10.5452/ma-3qrtm (https://www.modelarchive.org/doi/10.5452/ma-3qrtm)
Computational model of DNA binding domain of TEAD1 docked to DNA with 5'-3' oriented M-CAT motif	This paper	Model Archive: 10.5452/ma-xpro7 (https://www.modelarchive.org/doi/10.5452/ma-xpro7)

(*Continued on next page*)

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: cell lines		
Jurkat E6.1 cell line	ATCC	RRID:CVCL_0367
Oligonucleotides		
Oligonucleotides used in native MS, KD determination, H/D exchange and cross-linking experiments, see Table S1	This paper	N/A
Oligonucleotides used for smFRET study, see Table S3	This paper	N/A
Primers for ChIP-qPCR, see Table S4	This paper	N/A
Recombinant DNA		
Plasmide pET28a(+) with 22-104 region of human TEAD1 cloned between *EcoRI* and *NdeI* sites.	This paper	N/A
Software and algorithms		
Vinci	ISS	http://www.iss.com/fluorescence/software/vinci.html
DataAnalysis 4.4	Bruker Daltonics	https://www.bruker.com/service/support-upgrades/software-downloads/mass-spectrometry.html
MeroX	(Götze et al., 2012, 2015)	http://www.stavrox.com
LinX	(Slavata et al., 2019)	http://peterslab.org/downloads.php
mMass 5.4.1	(Strohalm et al., 2010)	http://www.mmass.org/download/old.php
ProteinScape 4	Bruker Daltonics	https://www.bruker.com/products/mass-spectrometry-and-separations/ms-software/proteinscape.html
DeutEx	(Filandrová et al., 2021)	http://peterslab.org/downloads.php
Modeller	(Webb and Sali, 2014)	https://salilab.org/modeller/
Xwalk	(Kahraman et al., 2011)	www.xwalk.org
FoldX	(Schymkowitz et al., 2005)	http://foldxsuite.crg.eu/
AmberTools14	(Case et al., 2014)	https://ambermd.org/
μManager 1.4.22	(Edelstein et al., 2014)	https://micro-manager.org/wiki/Download_Micro-Manager_Latest_Release
ImageJ Thunderstorm plugin	(Ovesný et al., 2014)	https://zitmen.github.io/thunderstorm/
Crystallography & NMR System (CNS) 1.3	(Brunger, 2007; Brünger et al., 1998)	http://cns-online.org/v1.3/
HADDOCK 2.2	(van Zundert et al., 2016)	https://www.bonvinlab.org/software/haddock2.2/
DISOPRED	(Jones and Cozzetto, 2015)	http://bioinf.cs.ucl.ac.uk/psipred/
DP-bind	(Hwang et al., 2007)	http://lcg.rit.albany.edu/dp-bind/
DNATCO	(Černý et al., 2016)	http://dnatco.org

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Petr Novák (pnovak@biomed.cas.cz)

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Software used in this study is available under these links: **MS Tools** – web based application for visualization and presentation of HXMS data (http://peterslab.org/MSTools), **LinX** – software for data processing and interpretation of cross-linking data (http://peterslab.org/downloads.php), **DeutEx -** Software for interpretation of HDX-MS data (available upon request http://peterslab.org/downloads.php), **DNATCO** – web based application for assignment of DNA and RNA conformers (http://dnatco.org)
- The accession number for the original/source data for the cross-linking and HDX experiments reported in this paper is ProteomeXchange: PXD012127 (http://www.ebi.ac.uk/pride)
- Summary of all deuterium uptake plots and mass spectra of cross-links dissociation product have been deposited to Mendeley data: 10.17632/27zkz3v729.1 (https://data.mendeley.com/datasets/27zkz3v729/1)
- The accession number for the models based on cross-linking and HDX data is PDB-Dev: PDBDEV_00000062, PDBDEV_00000063 (https://pdb-dev.wwpdb.org/entry.html?PDBDEV_00000062 and https://pdb-dev.wwpdb.org/entry.html?PDBDEV_00000063)
- Computational models have been deposited to Model Archive: 10.5452/ma-3qrtm, 10.5452/ma-xpro7 (https://www.modelarchive.org/doi/10.5452/ma-3qrtm and https://www.modelarchive.org/doi/10.5452/ma-xpro7)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

***E. Coli* BL21-CodonPlus (DE3)-RIPL** (Agilent Technologies, USA) was used for recombinant production of TEAD1-DBD. The cells were stored in a pH 6.8 buffer containing 10mM MOPS, 10mM RbCl, 75mM $CaCl_2$ and 15% glycerol at -80°C and after transformation grown on LB agar and in LB medium at 37°C.

Jurkat E6.1 cell line (RRID:CVCL_0367) was used for ChIP-qPCR study and also as a source of material for preparation of plasmid containing the *TEAD1-DBD* DNA. The cells were cultured in RPMI1640 medium supplemented with 5 mM glutamine (Lonza, Switzerland), and 10% fetal bovine serum (Gibco, USA) at 37°C and 5% CO_2.

METHOD DETAILS

Chemicals

Unless stated otherwise, all chemicals were purchased from Merck (Darmstadt, Germany) and used directly without any further purification.

Protein expression and Cy3 labeling

The 22-104 region of human TEAD1, which covers the entire DNA-binding domain (TEAD1-DBD) obtained by PCR amplification of cDNA from Jurkat cells, was cloned immediately adjacent to an N-terminal His-tag of a pET-28a(+) vector. The protein was over-expressed in BL21-CodonPlus (DE3)-RIPL cells (Agilent Technologies, USA). After initial growth at 37°C the cell culture was cooled to 30°C and protein expression in transformed cells was induced by addition of 0.2M IPTG and performed for 16 hours. Harvested cells were lysed by sonication in phosphate buffer (pH 7.4) containing 0.5M NaCl and 1mM 2-mercaptoethanol. Subsequent affinity purification was performed on column filled with a TALON Superflow Resin (Clontech Laboratories, USA). Cleavage of His-tag was done overnight at 4°C by specific cleavage by human thrombin (1U per mg of recombinant protein). Removal of thrombin, cleaved His-tag and buffer exchange to 20mM HEPES, 150mM NaCl, 1mM TCEP, pH 7.4 (HEPES buffer) was carried out at gel permeation chromatography column (ENrich SEC 70 10 × 300 Column, Bio-Rad Laboratories, USA). Protein intended for FRET experiment was subsequently labeled with Cy3 dye. Cy3 maleimide (Thermo Fisher, USA) was dissolved in dimethyl sulfoxide (DMSO) to 10mM concentration. To perform the reaction a 5-fold molar excess of the dye in DMSO was added to 1 mg/ml protein solution in the HEPES buffer and the mixture was incubated for 16 h in 4°C. The unreacted dye was then removed by using a Zeba Spin desalting column (7000 MWCO) (Thermo Fisher, USA) followed by gel permeation chromatography on the same column as before.

Design of oligonucleotides and preparation of dsDNA

Six oligonucleotides containing the M-CAT binding motif ATTCC were obtained from the human genome sequence by using the *Ensembl* database (Zerbino et al., 2018), namely 5'-GTA GTA ATT CCA GCG-3' from the *C-MYC* first exon; 5'-GCA ATC CTT AAA GCT-3' from the *C-MYC* enhancer; 5'-TTG CCT TAC AGC CGG-3' from the *GLUT1* first exon; 5'-AGC CCG CCT TAC TCA-3' from the *GLUT1* enhancer; 5'-AGG CAG ATT CCA GAT- 3' from the *SRF* promoter; and 5'-AGT CAC ATT CCT CCG- 3' from the *CTGF* promoter (for summary see Table S1). For each of these oligonucleotides, both forward and reverse strands were purchased from Integrated DNA Technologies (USA) in standard desalted purity. The same oligonucleotides bearing a fluorescein amidite (FAM) label at the 5'-end were purchased from Eurofins Genomics (Austria) to enable fluorescence anisotropy determinations. To minimize the incidence of competing folding processes, all samples were tested for possible formation of secondary structures or self-dimerization by performing melting experiments. The desired duplex samples were obtained by dissolving corresponding forward/reverse strands in water, mixing them in equimolar ratio, and heating the solution to 95°C for 1 min. Each mixture was finally allowed to slowly cool down to room temperature to generate a 15-bp dsDNA construct.

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

Melt curves of individual DNA duplexes

Dissolved DNA duplexes were mixed 1:1 with SYBR Green fluorophore (Bio-Rad, USA) to reach final reaction volume of 25 μl. Emission of SYBR Green fluorescence was recorded during temperature increment from 10°C to 90°C (increment 1°C, read after 5 seconds) on CFX-96 Real-Time PCR instrument (Bio-Rad, USA). CFX manager software was used for non-linear regression model fitting and individual melting curves visualization.

Preparation of protein-DNA complexes

Recombinant TEAD1-DBD was transferred into a 150mM solution of ammonium acetate (pH adjusted to 6.8) and diluted to a 5μM concentration. An equimolar amount of the desired dsDNA sample was added, and the resulting mixture was incubated for at least 5 min at room temperature. A triplicate of samples was prepared for each dsDNA. The samples were subsequently analyzed by nano-electrospray ionization MS on a 15T-solariX XR Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (Bruker Daltonics, USA). All data were acquired in positive mode over the m/z 1000 – 5000 range. The fraction of bound protein was calculated according to the equation:

$$F_{\%} = \frac{\sum_{zmax}^{zmin} \frac{I_{F(z_i)}}{z_i}}{\sum_{zmax}^{zmin} \frac{I_{F(z_i)}}{z_i} + \sum_{zmax}^{zmin} \frac{I_{C(z_i)}}{z_i}} \cdot 100\%$$

in which F% is the fraction of bound protein; $I_{F(zi)}$ is the signal intensity of each z_i charge state detected for the free protein in the spectrum; $I_{C(zi)}$ is the signal intensity of each z_i charge state detected for the complex; z_{min} and z_{max} are the minimum and maximum charge states detected for the respective species. This treatment relies on the fact that all the samples in the study displayed the same charge state distribution regardless of the type of dsDNA in the complex, thus enabling for unbiased comparisons between the respective binding properties (Barylyuk et al., 2013; Hagan and Fabris, 2003; Stephen et al., 2007; Wang et al., 2003).

Fluorescence anisotropy binding assay

Fluorescence anisotropy determinations were performed on a PC1 photon count, steady-state spectrofluorometer (ISS, USA) according to reference (Boura et al., 2007). Double-stranded DNA (dsDNA) labelled with FAM at the 5'-terminus of each strand was diluted to 100nM with buffer (20mM HEPES, 150mM NaCl, 1mM TCEP, pH 7.4) and titrated with 100 or 20 μM (depending on K_D of the complex) TEAD1-DBD in the same buffer. After each increment the solution in cuvette was left to equilibrate for 2 minutes and steady-state fluorescence anisotropy was recorded with 494 nm excitation and emission at 520 nm. Each determination was repeated at least ten times to achieve proper sampling statistics. Vinci software (ISS, USA) was used for data processing. Every point was calculated as an average of three independent measurements, each already consisting of ten technical replicates. The initial fluorescence anisotropy observed for protein-free dsDNA was subtracted from that obtained after each addition of TEAD1-DBD. The values were plotted against the corresponding TEAD1-DBD concentrations and fitted according to the equation:

$$FA = \frac{FAmax \cdot c}{K_D + c}$$

in which FA is the increase in fluorescence anisotropy; c is the concentration of TEAD1-DBD; FA_{max} is the anisotropy increase at saturation; and K_D is the dissociation constant of the complex.

Quantitative chemical cross-linking

Cross-linking reaction

A previously described protocol was followed to perform cross-linking reactions (Kukacka et al., 2015; Rozbeský et al., 2018; Slavata et al., 2019). DNA-free TEAD1-DBD or TEAD1-DBD•M-CAT complex (1:1 molar ratio) were prepared in a 20 mM HEPES buffer solution with pH 7.4, which contained 150 mM NaCl and 1 mM tris(2-carboxyethyl)phosphine (TCEP). Each sample was preincubated for at least 1 hr at 20°C before mixing with the cross-linking reagent disuccinimidyl adipate (DSA). DNA-free TEAD1-DBD samples were mixed with non-labelled cross-linker ($DSA^{12}C$), whereas TEAD1-DBD•M-CAT samples were mixed with the ^{13}C-labelled version ($DSA^{13}C$). Both cross-linkers were purchased from Creative Molecules (Canada) and were dissolved in dimethyl sulfoxide (DMSO) to 14.6mM concentration. The amount of cross-linker stock added to the substrate was calculated to provide a 15:1 molar excess over the protein with a final overall protein concentration of 40μM. After 45 min incubation at room temperature, a 5-fold molar excess of ethanolamine over cross-linker was added to quench the reagent and stop the reaction. Corresponding non-labelled and ^{13}C-labelled samples were subsequently mixed in 1:1 ratio to enable quantitative determinations. Negative control samples (DMSO added instead of cross-linker) as well as samples treated with a 1:1 mixture of both labelled/unlabelled cross-linkers, were prepared in parallel to enable proper comparison. The reaction samples were initially analysed by SDS-PAGE to evaluate the presence of any unwanted higher oligomeric structures. To demonstrate the cross-link formation and quantification reproducibility, the experiment was performed in triplicates for one forward (*CTGF*) and one inverted (*GLUT1* exon) M-CAT motif. The samples were subsequently characterised according to an established bottom up approach in which digestion with a combination of proteases (i.e., trypsin and AspN) was carried out before LC-MS analysis (Kukacka et al., 2015).

Proteolytic digestion

Samples were mixed with 100mM ethylmorpholine buffer pH 8.5 containing 20% ACN in 1:1 ratio. AspN dissolved in water was added to reach the final weight ratio 1:100 (AspN:protein). Digestion was performed overnight at 40°C. Following the first digestion, trypsin was added in 1:20 (trypsin:protein) weight ratio and second digestion was performed for 8h at 45°C and quenched by lowering pH with 0,4% FA added in 1:1 volume ratio.

HPLC/ESI-FTICR-MS

Peptide samples were diluted in H_2O to final peptide and FA concentration of 0.05 μg/μL and 0.1%, respectively. Afterwards, peptides were injected (0.15 μg per injection) onto a reversed phase trap column (Zorbax 300SB-C18 5 μm, 0.3 × 5 mm, Agilent Technologies, USA), where the trapped peptides were desalted by 10 μL/min flow of 0.1% FA in water for 5 min (LC-20AD HPLC pump, Shimadzu, JPN). Desalted peptides were eluted from trap column at a constant flow rate of 10 μL/min (HPLC pump 1200, Agilent Technologies, USA) by a two-step linear gradient, first from 5 to 40% of solution B (0.1% FA, 98% ACN in water) in 35 min, followed by fast step from 40 to 99% of solution B in 5 min. Solution A consisted of 0.1% FA and 2% ACN in water. Eluted peptides were then on-line separated on the analytical column (ZORBAX 300SB-C18, 0.3 x 150 mm, 3.5 μm, Agilent, USA). The HPLC system was directly coupled with an electrospray ionization source of Fourier transform ion cyclotron resonance mass spectrometer (15T-SolariX XR FT-ICR, Bruker Daltonics, USA). Mass spectral data were acquired in positive LC-MS mode over the m/z range 250 – 2500, where each MS scan is followed by MS/MS scan with fixed collision-induced fragmentation energy of 15eV and broad isolation window (Fiala et al., 2020)

Data processing and quantification

The MS data obtained from the protease-digested reaction mixtures were processed by using DataAnalysis 4.4 (Bruker Daltonics, USA). The SNAP 2.0 algorithm was employed to generate deconvoluted spectra and corresponding lists of monoisotopic masses. The data provided by negative controls and samples treated with the 1:1 cross-linkers mixture were searched for potential cross-links by using MeroX (Götze et al., 2012, 2015) and a home-built software Links. For each identified cross-link an extracted ion chromatogram was created. The scan with the highest signal intensity was further processed by using mMass 5.4.1. The ratio of incorporation of light (^{12}C) versus heavy (^{13}C) cross-linker was calculated for each conjugated product by using the envelope fit tool, which relies on fitting putative theoretical profiles with corresponding experimental data following both a linear combination and least-square fitting.

Hydrogen-deuterium exchange

HDX reaction

HDX reactions were performed according to an established procedure (Filandrová et al., 2021; Slavata et al., 2019) on either DNA-free TEAD1-DBD (20μM), or TEAD1-DBD•M-CAT complex (20μM, 1:1 molar ratio) in a 20 mM HEPES buffer solution with pH 7.4, which contained 150 mM NaCl and 1 mM TCEP. Samples were preincubated for at least 1 hr at 20°C before deuterium exchange was initiated by 10-fold dilution into a D_2O-based buffer (20mM HEPES, 150mM NaCl, 1mM TCEP, pD 7.4). HDX was allowed to proceed at 20°C, while aliquots containing 300 pmol of protein were taken at pre-determined intervals (i.e., at 10 sec, 1 min, 10 min, and 1.5 h of reaction time.). Each aliquot was immediately quenched by 1:1 mixing with quench buffer (4M Urea, 1M glycine/HCl, pH 2.3) and then rapidly freezing in liquid nitrogen. Samples were stored at -80°C and thawed only immediately before LC-MS analysis. All experiments were performed in triplicates.

HPLC/ESI-FTICR-MS analysis of HDX samples

Each sample was injected into a home-built LC system that was maintained at 0°C to minimize the possible back-exchange of hydrogen. Protein digestion was carried out on a column with immobilized aspergillopepsin (protease type XIII) at a flow rate of 200 μL/min (LC-20AD HPLC pump, Shimadzu, JPN) of 0.4% formic acid (FA) in water. Produced peptides were then trapped and desalted on a peptide microtrap (C8 reversed-phase, Michrom BioResources, USA) for 3 min under the same flow rate and solvent composition. Trapped peptides were eluted at a constant flow rate of 20 μL/min (HPLC pump 1200, Agilent Technologies, USA) first by a linear 5 min gradient from 5-35% of solution B (0.4% FA, 95% Acetonitrile (ACN) in water) followed by quick step to 99% B and separated on the reversed-phase analytical column (ZORBAX 300SB-C18, 0.5 x 35 mm, 3.5 μm, Agilent, USA). Solution A consisted of 0.4% FA, 2% ACN in water. For a peptide mapping of non-deuterated samples, the same protocol was used, except of the H_2O-based buffer was used instead of D_2O-based. Peptide species eluting from the LC system were introduced online into the electrospray ionization source of our 15T-solariX XR FTICR (Bruker Daltonics, USA) mass spectrometer. Peptide mapping experiments were performed in broadband positive ion mode according to a data-dependent MS/MS approach, in which each MS scan was followed by the collision-induced dissociation (CID) analysis of the six most abundant precursor ions.

Data processing

The MS/MS data collected in the 300 – 1800 m/z range were processed by using DataAnalysis 4.1 (Bruker Daltonics, USA) and then searched by ProteinScape 4 (Bruker Daltonics, USA) against a library of possible digestion products generated from the TEAD1-DBD sequence. The corresponding MS data collected also in the 300 – 1800 m/z range were processed by using the home-built DeutEx software. Peptides with overlapping regions were used to calculate the number of exchanged deuterium atoms to a resolution corresponding to smaller regions of the protein sequence, sometimes even to single aminoacids. Deuteration level of DNA-bound form was subtracted from the deuteration level of free TEAD1-DBD and the differences were used to replace B-factors and to colour the structure using blue-white-red gradient. The significance level was defined as two times standard deviation (Jensen et al., 2015).

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

Homology modelling

The MODELLER (Webb and Sali, 2014) package was employed to generate initial homology models of free and bound TEAD1-DBD from templates consisting of the high-resolution structures of TEAD1 (PDB: 2HZD) (Anbanandam et al., 2006) and the complex including TEAD4 and M-CAT DNA (PDB: 5GZB) (Shi et al., 2017). The procedure employed cross-linking information as distance restraints with increased weight to 20 to reflect the actual conformation of the substrates under the selected experimental conditions. The LYS:Cα-LYS:Cα and LYS:Cα-SER:Cα distances identified by the DSA reagent employed in the study were respectively set to 21.7 and 17.7 Å. The Nterm-LYS:Cα distances identified by conjugates involving the N-terminal residue were set to 15.3 Å. All distances included a 3 Å standard deviation to account for the putative flexibility of linker and side chains in the conjugate. To exploit positions of N-terminal and C-terminal residues structures with the best DOPE scores were subsequently used to carry out restrained molecular dynamics simulations with the simulated annealing protocol (rMD-SA) using the CNS 1.3 software suite (Brunger, 2007; Brünger et al., 1998). Atom positions of residues predicted as flexible by DISOPRED (Jones and Cozzetto, 2015) were allowed to move during the rMD-SA, atom positions of the DBD were restrained. (Weinkam et al., 2012) Solvent accessible surface distances were evaluated by using the Xwalk program (Kahraman et al., 2011). For additional information on the cross-linking method validation see Figure S4.

Molecular docking

For the docking of the TEAD1-DBD and the C-MYC exon oligonucleotide the shortened model of TEAD1-DBD (I39–R100]) without flexible residues on C-terminal and N-terminal ends was used. The structure of the C-MYC exon oligonucleotide was generated by the online tool make-na (http://structure.usc.edu/make-na/) in the ideal B-DNA conformation. The docking was performed using the program HADDOCK 2.2. (van Zundert et al., 2016) Residues T81–R85, R87–H93, and Q95 indicating the significant difference of the H/D exchange between apo and DNA-bound TEAD1 and at the same time predicted as DNA-binding residues using DP-Bind web server (Hwang et al., 2007; Kuznetsov et al., 2006) were considered as active residues. The core M-CAT motif and nucleotides on the complementary strand of the C-MYC exon oligonucleotide were considered as active residues. The docking calculations were performed using the WeNMR/WestLife infrastructure (Wassenaar et al., 2012). The missing N-terminal and C-terminal residues for structures obtained from HADDOCK in the cluster with the lowest Z-score were added using the MODELLER and further subjected rMD-SA using CNS software suit. During the simulated-annealing atom positions of the TEAD1-DBD and C-MYC exon oligonucleotide were restrained. Solvent accessible surface distances were evaluated by using the Xwalk program.

Molecular dynamics

In systematic fashion, a series of protein-DNA complexes were generated by docking the structures of selected dsDNA constructs with that of the TEA domain (see Table S2). The automodel function of MODELLER ver. 9.14 (Webb and Sali, 2014) was used to generate initial models from a template consisting of the high-resolution structure of the TEAD4•dsDNA complex (PDB: 5gzb) (Shi et al., 2017). This strategy exploited the high degree of homology between TEAD4 and TEAD1 to obtain a representation of the binding region, while the structure of the bound DNA component was kept frozen. In this way, the homology modelling process was mainly used to add the atoms missing in the template. To obtain the structure of each protein-DNA complex, the BuildModel function of FoldX (Schymkowitz et al., 2005) was subsequently used to "mutate" appropriate bases of the dsDNA component, while keeping the biopolymer's backbone fixed. MD simulations were then carried out in Amber14 (Case et al., 2014) by using the ff14SB parameters for the TEAD1 protein and ff99bsc0 for the DNA component. The complex was placed in a periodic box and solvated by using a TIP3P water box. The total charge was adjusted to neutral by adding an appropriate number of sodium ions. A 20 ns Langevin MD simulation was then carried out at constant pressure with a time step of 2 fs and frame collection every 10 ps. The DNA conformation of each snapshot extracted from the trajectory was examined by using the assignment protocol (Schneider et al., 2018) accessible on the dnatco.org web service (Černý et al., 2016). The theoretical protein-DNA binding affinities (ΔG) were calculated from the MD trajectories using the Molecular Mechanics Poisson Boltzmann Surface Area (MMPBSA) method as implemented in the MMPBSA.py script from AmberTools14 (Case et al., 2014).

Single-molecule förster resonance energy transfer

Imaging surface preparation

High precision, round microscopy coverslips (25 mm; Marienfeld 0117650) in Teflon holder (Wash-N-Dry Coverslip Rack; Diversified Biotech Inc.) were submerged into 2% Hellmanex III solution (Hellma Analytics) diluted in the ultrapure water (Milli-Q®, Millipore) and incubated overnight at 70°C. Afterwards, 30 minutes sonication in a heated sonication bath and several washes in ultrapure water with one additional sonication step were applied. Cleaned coverslips were stored in ultrapure water. To immobilize the DNA-protein complex, coverslips were first dried, placed in glass Petri-dishes and coated with 500 μl of 0.1 mg/ml BSA/BSA-biotin (ratio 1000:1) mixture in PBS for 1 hour at room temperature. Coverslips were then washed 3 times with ultrapure water and incubated with 500 μl of 2 μg/ml streptavidin in PBS for 1 hour. After 3 washing steps with ultrapure water, coverslips were incubated with 500 μl of 2nM biotinylated oligonucleotides (Table S3) in 20 mM HEPES buffer solution (pH 7.4), which contained 150 mM NaCl and 1 mM TCEP (HEPES buffer) for 30 minutes. After a brief wash, 500 μl of 2nM Cy3-labeled TEAD1-DBD protein in HEPES buffer was added for 15 minutes. Before acquisition, the coverslips were placed into ChamLide holder (Live Cell Instruments) and washed 3 times with the HEPES buffer.

Data acquisition

smFRET measurements were performed on a home-build microscope based on IX71 body (Olympus) equipped with 150 mW 561 nm (Sapphire, Coherent) and 100 mW 643 nm lasers (Cube, Coherent), 100x TIRF objective (UApoN, NA 1.49, Olympus) and an acousto-optic tunable filter (AOTFnC-400.650-TN, AA Optoelectronics) for the synchronization of the EM-CCD camera (iXon DU-897, Andor; EM gain: 500) with lasers. The two channels of the fluorescence emission were simultaneously detected on a single camera positioned behind Optosplit II (Cairn research) equipped with emission splitter cube: T635lpxr – UF1 dichroic mirror (Chroma); HQ 500 LP (for the Donor channel) and 635 LP Edge (for the Acceptor channel; Semrock). Total internal reflection fluorescence (TIRF) mode of sample illumination was employed. μManager software (version 1.4.22 (Edelstein et al., 2014)) controlled the acquisition in an Alternating-Laser Excitation (ALEX (Kapanidis et al., 2004)) scheme at the rate 50 ms per frame (20 frames/field of view).

Data analysis

Images were first processed by standard tools of the ImageJ software (version 1.49 (Schindelin et al., 2012)) to separate Donor and Acceptor channels. Photon counts (intensity) for individual complexes were acquired in each channel separately by a single-molecule localization analysis using Thunderstorm plugin of the ImageJ software (Ovesný et al., 2014). The intensity values measured for the Donor (IDD) and Acceptor (IDA) of the same DNA-protein complex, acquired under the donor illumination (561 nm), were found by the nearest neighbor analysis (maximum displacement: 800 nm). FRET efficiency (E_{FRET}) values for the individual DNA-protein complexes were calculated as $E_{FRET} = \frac{IDA}{(IDD+IDA)}$ where IDD and IDA are the intensity values measured for the donor and acceptor, respectively. The average FRET efficiency for a specific DNA-protein pair was calculated from all complexes exhibiting $E_{FRET} > 0.25$. Localizations with $E_{FRET} < 0.25$ represent donors missing the acceptor in the complex (as evaluated by the ALEX illumination scheme using the acceptor illumination (647 nm (Schrangl et al., 2018)). At least 100 complexes were evaluated for each DNA-protein pair to calculate the average E_{FRET}.

Chromatin immunoprecipitation and qPCR

Design of primers

Sequences of individual genomic regions were obtained with *Ensembl* database and primers were designed using GeneRunner software enabling detailed analysis of individual primers. Specificity and effectivity of individual sets of primers were tested using genomic DNA of Jurkat T-ALL cells as matrices (clone E6.1, ATTC Collection, USA). C_t for individual sets of primers across logarithmic dilution range of genomic DNA and slopes of individual curves were determined. Effectivity of sets of primers was calculated according equation: $E = \left(10^{-\frac{1}{Slope}} - 1\right) x\ 100$. Melting curves of individual PCR products were recorded to confirm primers specificity. Primers used during qPCR quantification are stated in Table S4.

Cross-linking, isolation and enzymatic shearing of chromatin

Jurkat cells were grown to typical densities in the 500,000-700,000 cells/mL range in RPMI1640 medium supplemented with 5 mM glutamine (Lonza, Switzerland), and 10% fetal bovine serum (Gibco, USA) at 37°C and 5% CO_2. $5x10^7$ cells was harvested by centrifugation (400g, 4min, RT) and resuspended in 20 ml of minimal RPMI1640 medium. After harvesting, 1% formaldehyde was added to induce protein-DNA cross-linking which took place for 7 minutes at room temperature. The cells were lysed and their nuclei collected. To do that, the cells were centrifuged (400g, 4min, RT), resuspended in 0.3 mol/L glycine solution in PBS and incubated at RT for 5 minutes. After incubation, cells were pelleted, resuspended in PBS, pelleted again and the pellet was resuspended in 500 μl of hypotonic buffer and incubated on ice for 15 minutes. Nonidet NP-40 was then added to final concentration of 1%, cells were resuspended by pipetting and nuclei were pelleted by centrifugation (12 000g, 30s, 4°C). Pelleted nuclei were resuspended in 350 μl of Micrococcal Nuclease Buffer (NEB, UK), supplemented with HaltT Protease Phosphatase Inhibitor Cocktail (Thermo, USA), preincubated at 37°C for 5 minutes and treated with Micrococcal Nuclease (NEB, USA) which was added to final dilution 500x . The released chromatin was vortexed in periods of two minutes and sheared for 12 min at 37°C. Shearing was stopped by addition of ice-cold EDTA to final concentration of 10 mmol/L and samples were incubated on ice for 10 minutes. Samples were centrifuged (15 000g, 10min, 4°C), supernatant was removed and stored at -80°C for ChIP analysis.

Precipitation of TEAD1 cross-linked complexes

50 μL aliquots of the chromatin samples were finally submitted to immunoprecipitation by using Protein G Magnetic Beads (Active Motif, USA). Polyclonal rabbit IgG against TEAD1 (GeneTex, USA), rabbit IgG isotype negative control (GeneTex, USA), and mouse IgG against RNA polymerase II (Active Motif, USA) were used in equimolar concentrations. Each IP reaction consisted of 50 μl of chromatin, 25 μl of Protein G Magnetic Beads, 50 μl of ChIP buffer 1 (Active Motif, USA), 5 μg of individual antibody and HaltT Protease Phosphatase Inhibitor Cocktail. The reaction mixtures were incubated overnight at 4°C. After incubation, the beads were pelleted using Magnetic Separation Rack (NEB, UK) and washed one-time by 800 μl of ChIP buffer 1 (Active Motif, USA) and two-times by ChIP buffer 2 (Active Motif, USA). After final wash, beads were resuspended in 50 μl of AM2 elution buffer (Active Motif, USA) and incubated 15 minutes at RT to release the chromatin. Additional incubation with 50 μl of ammonium bicarbonate buffer (0.2 mol/L, pH 8.3) at 95°C for 15 min was carried out to reverse the cross-links formation and enable protein removal by Proteinase K. Finally, 1 μl of Proteinase K solution (5 μg/μl, Thermo, USA) was added to each sample and samples were incubated for one hour at 37°C. Proteinase K was then inactivated by PMSF at final concentration of 5 mmol/L. Samples were stored for qPCR analysis at -20°C.

Please cite this article in press as: Filandrová et al., Motif orientation matters: structural characterization of TEAD1 recognition of genomic DNA, Structure (2020), https://doi.org/10.1016/j.str.2020.11.018

*Relative quantification of individual regions by qPCR B*efore qPCR analysis, samples were diluted 10-fold and pipetted individually into 96-wells plate (Bio-Rad, USA) in technical triplicates. Each reaction mixture contained 10.5 μL of diluted ChIP sample, 2 μL of a 100 μmol/L solution containing a 1:1 mixture of forward and reverse primers (the sequence of which are provided in Table S4), and 12.5 μL of 2x SsoFast™ EvaGreen Supermix (Bio-Rad, USA). Plates were briefly centrifuged and finally analyzed by using a two-step PCR program (95°C for 30 s and 40x (95°C for 5 s, 60°C for 20s, plate read) followed by melt curve from 65°C to 95°C, increment 0,5°C) on a CFX-96 Real Time PCR System. Fold enrichment was calculated as $2^{-\Delta Ct}$ from C_t values afforded by specific antibody samples versus IgG samples used as negative controls. Two independent ChIP analyses were completed in the study.

QUANTIFICATION AND STATISTICAL ANALYSIS

Fluorescence anisotropy binding assay

As described in the Method Details section, every point was calculated as an average of three independent measurements, each already consisting of ten technical replicates and Vinci software (ISS, USA) was used for data processing. QtiPlot was used for curve fitting and standard deviation calculation.

Quantitative chemical cross-linking

Details of data processing are included in Method Details section. mMass 5.4.1 envelope fit tool was used for quantification and the results shown in Table 1 were calculated as a mean of three measurements and standard deviations were calculated by the excel standard deviation function.

Hydrogen-deuterium exchange

All experiments were performed in triplicates and a lab-made software DeutEx was utilized for data processing, quantification and calculating standard deviations. Differences higher than two times the average standard deviation were considered significant.

Details concerning quantification and statistics of the **Molecular Docking**, **smFRET** and **qPCR** experiments are each described in the appropriate section of the Method Details.