UNIVERSITY OF ICELAND

B.S. thesis

in Biochemistry and Molecular Biology

The expression, purification and labelling of HMGA1a

Elísa Sól Oddgeirsdóttir

May 2023

FACULTY OF LIFE AND ENVIRONMENTAL SCIENCES

The expression, purification and labelling of HMGA1a

Elísa Sól Oddgeirsdóttir

Dissertation submitted in partial fulfillment of a
Baccalaureus Scientiarum degree in Biochemistry and Molecular Biology

Advisor
Pétur Orri Heiðarsson

Supervisor
Mahtab Hafizi Yazdabadi

Co-supervisor
Kateryna Nitsenko

Faculty of Life and Environmental Sciences
School of Engineering and Natural Sciences
University of Iceland
Reykjavík, May 2023

The expression, purification and labelling of HMGA1a
Dissertation submitted in partial fulfillment of a *B.S.* degree in Biochemistry and Molecular Biology

Faculty of Life and Environmental Sciences
School of Engineering and Natural Sciences
University of Iceland
Dunhagi 5
107 Reykjavík
Iceland

Telephone: 525 4000

Bibliographic information:
Elísa Sól Oddgeirsdóttir, 2023, The expression, purification and labelling of HMGA1a, B.S. thesis, Faculty of Life and Environmental Sciences, University of Iceland, 57 pp.

Printing: Háskólaprent
Reykjavík, May 2023

Abstract

The fundamental process of converting our DNA to mRNA, which can then be translated into a protein, is carried out by a special set of proteins called transcription factors. They govern the process of gene transcription by recognizing a particular DNA sequence, which then enables them to recruit cofactors that can either activate or repress transcription. The Eukaryotic transcription factors all have a global architecture that makes them able to carry out their important roles in gene expression. All transcription factors possess some regions that have no well-defined three-dimensional structure, and some are even completely disordered. These proteins are called intrinsically disordered proteins, but it is estimated that they constitute to at least a third of the human proteome. High-Mobility group (HMG) is a group of proteins that belong to a family of disordered architectural transcription factors. They are involved in many different nuclear processes and are known to interact with nucleosomes. Together with linker histone H1, they were the first nuclear proteins known to affect the structure of chromatin. The compaction of chromatin is closely linked to the presence of H1, and as a result, the cell has developed regulatory mechanisms to actively remove H1 from nucleosomes. HMGA1a is one of those, but it's an intrinsically disordered protein known to be able to compete with linker histone H1 on binding the nucleosome. During the time course of this project, we were able to successfully express and purify the HMGA1a transcription factor. In addition, we attempted to label the protein with both donor and acceptor fluorescent dyes, by attaching them at two cysteine residues. Since the native protein contains no such residues in its sequence, we had to use a double-cysteine mutant version of the protein. The goal was to get a double-labelled protein so it could be used for smFRET measurements, ideally in complexes with reconstituted nucleosomes or DNA, to monitor shifts in intra-molecular distances. That would help us understand the molecular mechanism, underlying the HMGA1a nucleosome remodelling ability, and how the competition with H1 takes place.

Útdráttur

Umritun DNA yfir í mRNA er eitt af grundvallarferlum frumunnar en umritunarferlinu sjálfu er stjórnað af sérstökum prótínum sem kallast umritunarþættir. Þeir stjórna umritun með því að bindast á ákveðnar DNA-raðir, og geta í framhaldinu kallað á mismunandi hjálparþætti sem annað hvort virkja eða bæla umritun. Umritunarþættir sem finnast í heilkjörnungum hafa allir svipaða byggingu sem gerir þeim m.a. kleift að halda úti mikilvægum störfum í genatjáningu. Allir umritunarþættir hafa svæði sem innihalda einhverskonar ómótaða byggingu en sumir þeirra eru jafn vel fullkomlega ómótaðir, þ.e. þeir hafa enga þrívíða byggingu í sínu náttúrulega ástandi. Þessi prótín eru oft kölluð óreiðukennd prótín en talið er að þau séu a.m.k. þriðjungur af prótínmengi mannsins. „High-Mobility group“ (HMG) er hópur prótína sem tilheyra fjölskyldu óreiðukenndra umritunarþátta. Þau eru þátttakendur í mörgum mismunandi ferlum innan kjarnans og hafa mikla tengingu við litnisagnirnar. Þessi prótínfjölskylda, ásamt H1 histónprótíninu, eru fyrstu prótínin sem vitað er um að hafi áhrif á byggingu litnis. Myndun þéttlitnis er einnig mjög sterklega tengd við návist H1 og vegna þessa, hefur fruman þróað ýmsa stjórnmekanisma sem gerir prótínum kleyft að keppa við H1 um að bindingu litnis. HMGA1a er eitt þessara prótína en þetta er ómótað DNA-bindiprótín sem tekur þátt í umritun, og hefur einnig þann eiginleika að geta keppt við H1. Á þeim tíma er rannsóknin stóð yfir tókst okkur að tjá og hreinsa HMGA1a prótínið með mjög góðum árangri. Þar að auki reyndum við að merkja það, með því að festa flúrljómandi efni á tvær systein aminósýrur prótínsins. Þar sem upprunalega prótínið inniheldur engar slíkar aminósýrur í röð sinni, þurftum við að nota stökkbreytta gerð þess sem innihélt tvær systein aminósýrur. Markmið rannsóknarinnar var að mynda tvímerkt prótín sem gæti verið notað til smFRET mælinga, t.d. í flóka með litnisögnum, til þess að mæla breytingar á innansameindafjarlægðum þeirra. Það myndi hjálpa okkur að skilja eiginleika HMGA1a sem gerir því kleift að endurskipulegga byggingu litnisagna, og hvernig samkeppni þess við H1 prótínsins er háttað.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First, I would like to say a massive thank you to my family, especially my parents, for their continuous support throughout my academic journey. They have always been there for me and I can't thank them enough for that. I also want to express my gratitude towards my supervisor, Mahtab Hafizi Yazdabadi, who helped me immensely with this project by teaching me everything in the lab. Thank you for answering all my questions and for always being incredibly patient with me. I also want to give many thanks to my advisor, Pétur Orri Heiðarsson, who also helped me a lot throughout his project by giving me great advice and guidance, both regarding my thesis and lab work. I am very thankful for the experience of working in his lab and meeting all the amazing scientists working there. I would like to extend my sincere thanks to everybody in the lab for teaching me and helping me during the time course of this project. Finally, I want to thank all my fellow Biochemistry and Molecular Biology friends for their help and moral support over the past three years.

1 Introduction

1.1 The roles of DNA-binding proteins

DNA-binding proteins carry out central roles in the most fundamental processes that take place in all organisms, such as transcription, replication, packaging, rearrangement and repair[1]. These proteins contain DNA-binding domains, but they search and recognize their specific sites on the DNA and bind to it[2,3]. They either have a specific or general affinity for the DNA but they can bind to either single- or double-stranded DNA. DNA-binding proteins have common folding patterns known as DNA-binding motifs[3], but they use a wide range of different motifs that vary in structure, e.g. helix-turn-helix, the homeodomain (HD) and high-mobility group box (HMG), to recognize DNA[2]. DNA-binding proteins are among the most widespread cellular proteins[3], but their structure, function, and DNA-binding motifs exhibit significant diversity[4]. By binding to the DNA, they enable the transmission of genetic information which initiates many different biological processes. Protein-DNA interactions are therefore extremely important since most biological activities are controlled by them[2]. The importance of examining the nature of the complexes between proteins and DNA is also vital since they form the fundamental basis for understanding the mechanisms behind these processes[1].

1.2 Transcription factors

Transcription factors (TFs) are proteins that govern the process of gene transcription, which is the process of converting our DNA into mRNA, which can then be translated into a protein. The role of TFs is to identify a particular sequence of DNA, and subsequently enlist cofactors that either activate or repress transcription[5]. To do so, they bind to DNA-regulatory sequences that are typically located in the 5'-upstream region of target genes[6]. Eukaryotic TFs all have a global architecture that is made up of DNA-binding domains (DBDs) and effector domains (EDs). The EDs are the primary functional units for controlling transcription but they are really flexible. The amino acid sequence of the DBD is known to be highly conserved within families. Moreover, the majority of DBD structures have been effectively characterized by classical structural biology[5].

A computational analysis of more than 400 human TFs revealed that on average, 49% of their sequence is made up of intrinsically disordered regions (IDRs)[7], but they do not adopt a folded conformation in their functional, native state. These regions are ideal for TFs since they can facilitate precise recognition of interaction partners and coordinate regulatory events in space and time[8]. The disorder can also provide many other advantages to the function and regulation of TFs, such as conformational plasticity, promiscuity, and regulation through various post-translational modifications[9]. In recent years, an increasing number of proteins that do not contain a well-defined secondary or tertiary structure, have been discovered[10], both *in vitro* and by computational predictions[11,12]. Their structure is very flexible and disordered under native conditions. These proteins have been found to be involved in interactions with other biomolecules, such as other proteins, RNA and DNA[10].

1.3 The significance of intrinsically disordered proteins

For a long time, it was thought that all proteins had to have some kind of structure to them in order to be functional[13]. Intrinsically disordered proteins (IDPs) are proteins that contain no well-defined three-dimensional (3D) structure or have long IDRs[14]. Despite that these proteins are indeed functional and carry out many different roles[15]. They can adapt to different structural states and fluctuate among many of them ***(Figure 1)***. It is estimated that IDPs constitute to at least a third of the human proteome, but they are especially important and abundant in the nucleus[14]. IDPs are generally defined by a distinct amino acid composition, which is marked by a relatively low proportion of hydrophobic and aromatic residues, and a relatively high portion of charged and polar residues[16].

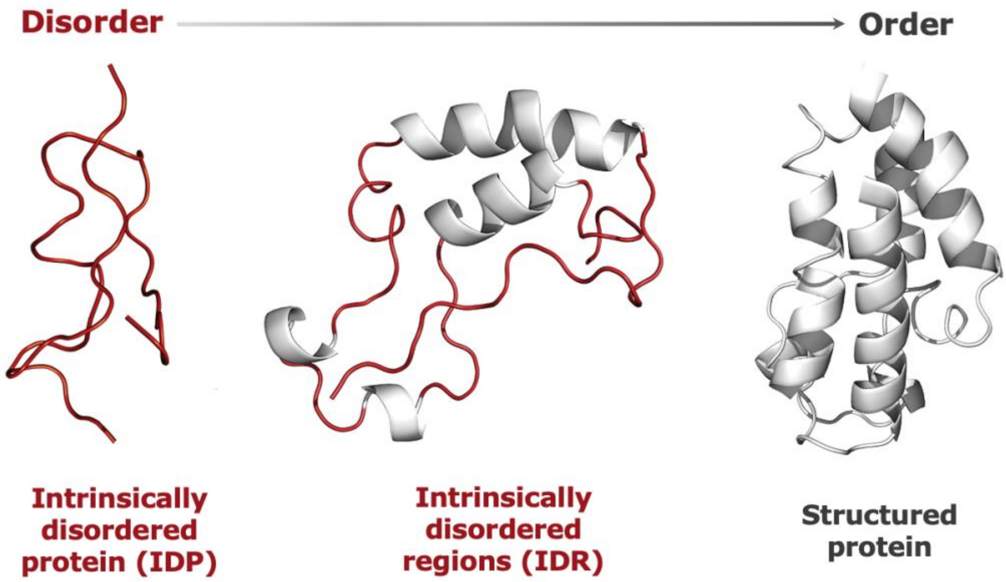

Figure 1: Architecture of disordered and structured proteins*. The protein to the left is completely disordered, the protein in the middle is partially folded, containing IDRs, while the protein to the right has a structure to it (adapted figure[17]).*

IDPs and IDRs play critical roles in the regulation of many signaling pathways and cellular processes, that is regulation of transcription, translation and the cell cycle[18]. They are found in proteomes of all three superkingdoms (Archaea, Bacteria and Eukarya), as well as in their viruses, which is an indication of their presence across various taxa[19]. Notably, eukaryotes contain a higher predicted fraction of disordered residues compared to prokaryotes, suggesting their importance in evolution[20].

1.3.1 The advantages of being intrinsically disordered

The conformational flexibility of IDPs/IDRs can be seen as an advantage in many different aspects. The intrinsic disorder of these proteins may e.g., help to reduce the genome size in organisms, such as humans. Compared to the diverse proteins that are required in higher organisms, the number of genes present in the human genomes is significantly lower. One way to avoid over-large genomes is alternative splicing, where numerous proteins can be produced from a single gene. These regions where the splicing occurs tend to be disordered, which helps to avoid structural disruption in the spliced proteins. If the regions were structured and inflexible, the splicing could disrupt the normal structure of the protein, which could have a negative effect on its function. Another way to control genome size is when a single protein can carry out more than one function[20]. This phenomenon has been referred to as "Moonlighting"[20,21], but moonlighting proteins are multifunctional in which one polypeptide chain exhibits more than one physiologically relevant or biochemical function[21]. IDPs/IDRs are ideal for this since they can use the same or overlapping regions to execute different functions by adopting different conformations upon binding[20].

Because of their conformational flexibility, IDPs/IDRS also facilitate exposure of their modification sites and are easily accessible to modifying enzymes. Therefore, they are predominant sites for posttranslational modifications (PTMs)[20,22]. PTMs are chemical changes that affect proteins after biosynthesis[23] and play a crucial role in regulating protein functions. In addition, they lead to a significant increase in the functional versatility of IDPs[20,22]. By altering the energetics of their conformational landscape, and modulating their interactions with other cellular components, PMTs elicit diverse effects on the biological functions of IDPs[22]. One of the most important and well-studied PMT in eukaryotes is phosphorylation. It is a reversible PTM that results in the addition of a phosphate group to the amino acid chain, but it is frequently found in IDPs. The phosphorylation of IDPs/IDRs is an essential control mechanism in signaling and regulation[20,24]. In addition, there are many other PTMs associated with IDPs/IDRs, i.e., amidation, ubiquitination, glycosylation, sulfation, methylation and more.

Finally, because of their flexible nature, IDPs/IDRs possess some impressive functional advantages over structured proteins. They can e.g., (partially) fold (or undergo disorder-to-order transitions) upon interaction with specific partners. They are also known to participate in highly specific but weak interactions, but this property is useful for reversible signaling transduction since it enables rapid association/dissociation with the partner molecule, without the excessive binding strength[20,23]. With that being said, nothing or no one is perfect, and that includes IDPs. Their lack of folded structure can also result in trouble. Human diseases such as cancer, cardiovascular disease, neurodegenerative diseases, and diabetes are known to be associated with numerous IDPs[25,26]. On top of that, when a structure of a protein is ordered and well understood, it helps scientist to identify and locate the active or binding site of the protein with accuracy. This information is important for conventional rational design, but the goal of this approach is to modify the active or binding site to enhance or alter the function of the protein[20]. This can lead to the development of new drugs[27].

1.4 The HMG protein family

High-Mobility group (HMG) is a group of proteins that belong to a family of disordered architectural transcription factors. They are involved in many different nuclear processes and are known to interact with nucleosomes. Together with histone H1, they were the first nuclear proteins recognized to affect chromatin structure[14]. HMG proteins play significant roles in regulating gene transcription and in remodeling the assembly of chromatin in higher eukaryotic cells. They do so by bending or modifying the DNA structure which forms complexes with histone proteins and other transcription factors[14,28]. They were first discovered in chromatin that had been isolated, along with histones[14], and are the most abundant non-histone chromatin-associated proteins[2]. Through their specific functional motifs, HMG proteins can bind to specific structures in DNA or chromatin in a sequence-independent fashion, meaning they recognize structure rather than a certain nucleotide sequence[29].

1.4.1 The three subgroups of HMG proteins

The HMG superfamily includes three subfamilies, that is HMGA, HMGB and HMGN, formerly known as HMGI/Y, HMG1/2 and HMG14/17. Each family has a unique protein signature and a characteristic sequence motif[29]. The three subgroups are categorized based on their DNA binding domains, that is HMGAs which contain AT-hooks, HMGBs which contain HMG-boxes and HMGNs which have nucleosomal binding domains[30,31]

The HMGB proteins (HMGB1 and HMGB2) contain two DNA binding domains, that is Box A and B. They are structurally conserved and both fold into three helices when in complex with DNA. The HMG-boxes have a low binding affinity towards single-stranded, linear duplex and supercoiled DNA, but prefer DNA that is bent or distorted, which is further intensified by acetylation. Furthermore, they contain a disordered C-terminal region that is enriched with acidic residues and is approximately 30 amino acids in length[14].

The HMGN proteins can be divided into five subgroups: HMGN1 and HMGN2, which were first discovered, and HMGN3-5 which were recognized later on. HMGNs are characterized by a positively charged and conserved nucleosome binding domain (NBD), a nuclear localization signal (NLS) and an acidic C-terminal chromatin regulatory domain (CHUD), that is involved in modulating acetylation of histones. These proteins are completely disordered[14] ***(Figure 2)***.

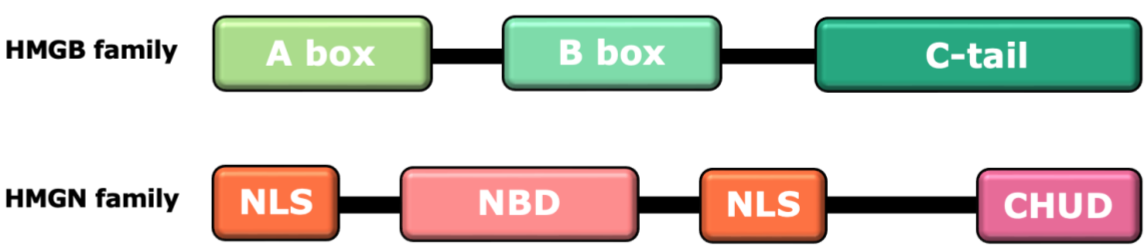

Figure 2: A schematic showing the different domains in the HMGB and HMGN protein families. *Abbreviations: A box, HMG box A; B box, HMG box B; C-tail, C-terminal acidic tail; NLS, nuclear localization sequence; NBD, nucleosome binding domain and CHUD, C-terminal chromatin regulatory domain.*

1.5 HMGA proteins

As mentioned above, the HMGA protein family consists of four members, that is HGMA1a, b, c and HMGA2. Available evidence increasingly suggests that HMGA proteins do not directly regulate the activity of transcription, but instead, they modulate gene expression by changing chromatin structure and organizing multiple TFs on a so-called “enhanceosome”[32]. In differentiated adult cells, the expression of HMGA proteins is low, but in embryonic and cancer cells they are expressed at a very high level. This high expression has been linked to cell growth and to the malignant/metastatic phenotype. When cells are induced to proliferate, the HMGA expression increases[33].

Like other HMG proteins, HMGA undergoes various PTMs. The HMGA1 proteins are e.g., one of the most phosphorylated proteins in the nucleus, but their phosphorylation leads to considerably lower affinity towards the DNA. The reason is that the main sites of phosphorylation are close to the positively charged AT-hooks, which disrupts the binding to the negatively charged DNA[14]. Other PTMs include methylation and acetylation. Furthermore, HMGA1 proteins regulate gene expression, embryogenesis, cell differentiation and adaptive immune responses, by binding to DNA and other TFs. HMGA1 has been shown to be highly over-expressed in multiple human cancers and is viewed as a valuable cancer biomarker. In addition, elevated expression of HMGA1 makes cancer cells resistant to chemotherapy[34]. The HMGA1 protein has three different isoforms, but the HMGA1a isoform is referred to as the canonical sequence by Uniprot. The site defines canonical isoform as the most prevalent, most conserved, longest or the one with a clear description of domains or post-translational modifications. HMGA1a was used for this project and will therefore be the one for discussion hereafter[35] ***(Figure 3)***.

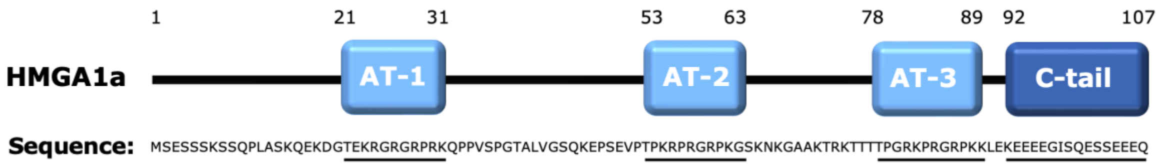

Figure 3: A schematic showing the different domains found in the HMGA1 protein*. Their positions on the sequence are also shown. Abbreviations: AT, AT-hook and C-tail, C-terminal acidic tail.*

1.5.1 The structural features of HMGA1a

The HMGA1a protein itself, obtained from the organism Homo sapiens (human), contains 107 amino acids (a.a.) and has a molecular mass of ~11.7 kDa. The protein has three DNA-binding regions called AT-hooks that can bind to DNA. The first one is positioned from a.a. 21-31, the second one from 53-63 and the last one from 78-89[36]. DNA-proteins derived from nature are designed in such a way that they contact negatively charged DNA phosphates mainly by the positively charged amino acids (arginine and lysine), thereby placing negatively charged amino acids (aspartic acid and glutamic acid) further away from contact with DNA[4]. This is exactly the case for HMGA1a, but these DBD regions contain a lot of positively charged amino acids, such as arginine (R) and lysine (K). The AT-hook has a core motif, RGRP (arginine-glycine-arginine-proline), which has a high binding affinity to the DNA AT-rich region. The protein also has an acidic tail at the C-terminus, enriched in glutamic acid (E), which causes it to be highly negatively charged. The function of the acidic tail is still unknown[37]. In addition, the protein contains no aromatic residues (W, F or Y)[36]. See ***Figure 4***.

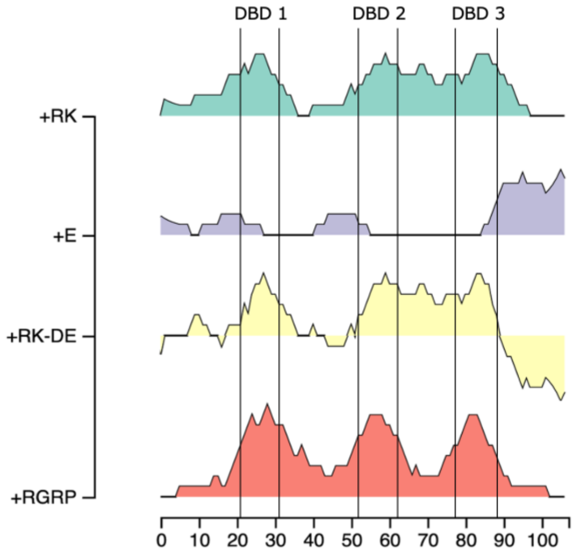

Figure 4: Graphical illustration of amino acid enrichment in HMGA1a*. The graph shows the counts of specific amino acids in the protein. "+RK" shows the count of lysine and arginine residues. "+E" shows the count of aspartic residues. "+RK-DE" shows the count of lysine and arginine minus the count of glutamic and aspartic acid. "+RGRP" shows the count of this repeated sequence of arginine, glycine, arginine, and proline, enriched in the DBD regions. The sequence is obtained from UniProt[36] and the graph is made with http://www2.bio.ku.dk/IDR/ with a 15 a.a. window size (adapted figure).*

Parts of the HMGA1a protein structure have been solved by NMR. As of now, Uniprot provides four different structures of the protein solved by NMR, but they only range from amino acids 51-71, 51-75 or 80-89. These amino acids are all within the DBD of the protein, which is logical because the AT-hook BDBs are the only regions that form some kind of structure upon binding to the DNA[14] ***(Figure 5)***. On the other hand, the site also provides a predicted structure of the protein made by the artificial intelligence system AlphaFold[17] ***(Figure 6)***.

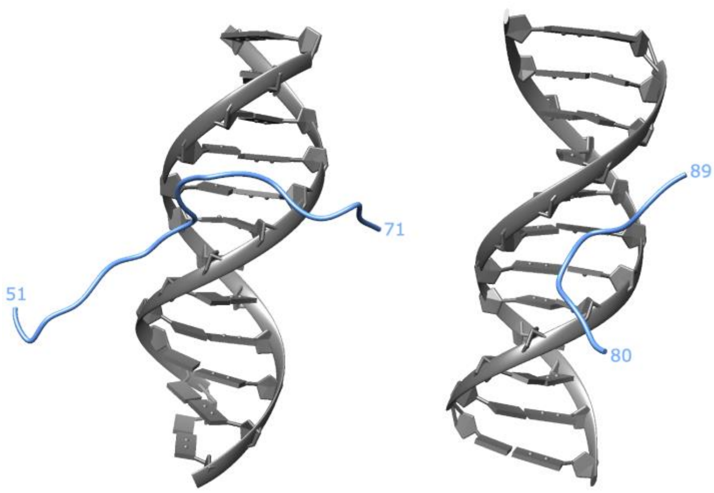

Figure 5: Solved structures of HMGA1a (Homo sapiens) by NMR in parts, bound to DNA*. Structure to the left contains a.a. 51-71 and the structure to the right contains a.a. 80-89. The protein is shown in blue and the DNA helix in grey. Figures are obtained from UniProt[36] (PDB codes: 2EZD and 2EZF) and edited in Chimera.*

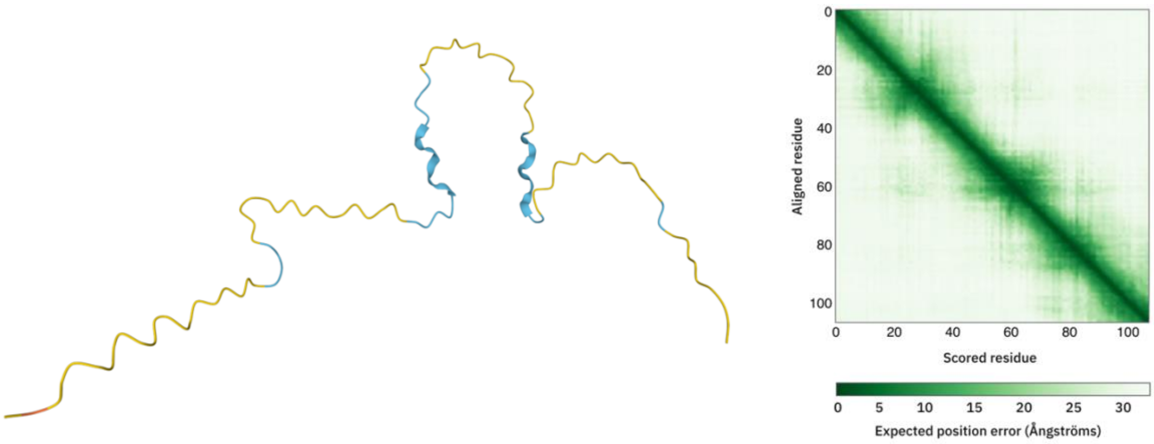

Figure 6: The predicted structure of HMGA1a from Homo sapiens by AlphaFold. *The predicted aligned error is shown to the right. The colors on the protein represent the Model Confidence determined by AlphaFold system. Dark blue = very high confidence, light blue = confident, yellow = low confidence, and orange = very low confidence. The regions colored in light blue are all near or within the DBD domains of the protein (except for a small region near the C-terminal of the protein)*[17].

As previously stated, HMG proteins belong to a family of disordered architectural transcription factors known to interact with nucleosomes, and together with linker histone H1, were the first nuclear proteins known to affect the structure of chromatin. It is also known that the compaction of chromatin is closely linked to the presence of H1. As a result, the cell has developed multiple regulatory mechanisms to actively remove H1 from nucleosomes. One of these mechanisms involves proteins that can compete with H1 for nucleosome binding or in other cases lead to its eviction. Outside of its globular domain, H1 is highly disordered, but that seems to be a common feature shared among the diverse H1 competitors[14].

1.6 The nucleosome and the role of histone H1

Genomic DNA in eukaryotic cells needs to be densely packed but still accessible for essential processes, including DNA repair, transcription, and replication[38]. All eukaryotic cells need to pack an enormous amount of DNA into a very small nucleolus. The solution to this problem is tightly compacted assemblies called chromatins. Chromatins are a highly organized complex of DNA and proteins and a key component of the cell nucleus. They are made up of structural units named nucleosomes, each contains approximately 147 base pairs of DNA which is wound around an octamer of four core histones called H2A, H2B, H3 and H4[39] **(Figure 7)**. Those core particles are connected by a short stretch of linker DNA, which forms a structure that resembles beads on the string. The linker histone H1 binds to the entry/exit sites of DNA on the nucleosomal core particle's surface, thereby completing the nucleosome[38]. H1 can help stabilize both the octamer assembly and higher-order chromatin architecture[38,39]. H1 also plays a role in chromatin condensation through the stabilization of the compact chromatin structures, and by doing so it generally functions as a transcriptional repressor[14]. The polypeptide sequence of H1 is highly positively charged and has two long disordered regions flanking a small folded globular domain, which is known to bind to the dyad axis of the nucleosome[14]. The core and linker histones possess a large amount of disorder regulating nucleosome structure and dynamics, which eventually impacts the global chromatin structure[18].

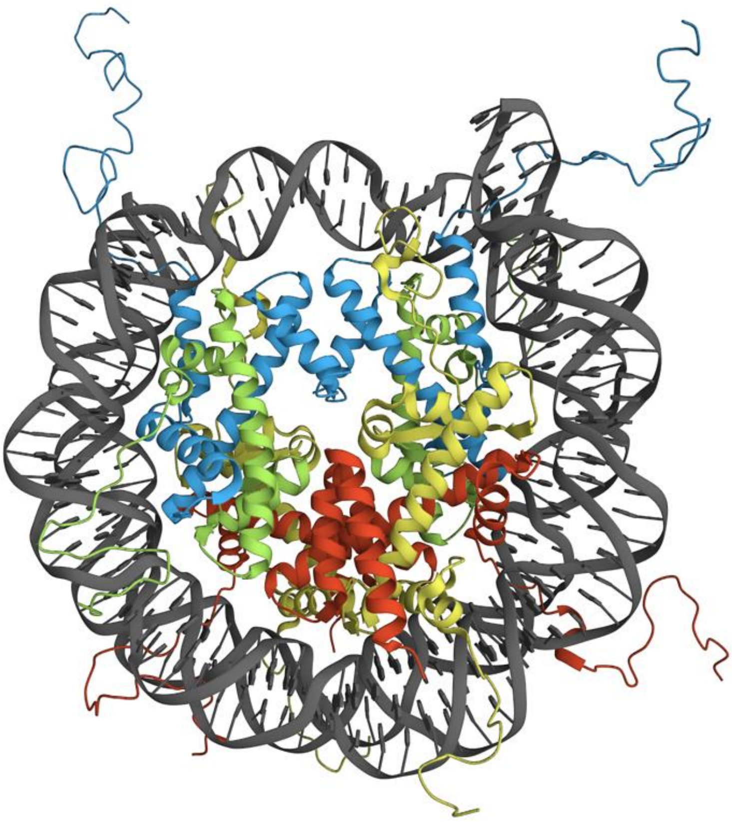

Figure 7: X-ray structure of a nucleosome core particle*. The histones are shown in color: H2A in yellow, H2B in red, H3 in blue and H4 in green. DNA wrapped around octamer is shown in gray. Figure obtained from PDB, code: 1KX5*[40]*.*

1.6.1 The competition of H1 and HMGA1 proteins for binding the nucleosome

As mentioned in the previous chapter, there are some proteins known to be able to compete with linker histone H1 on binding the nucleosome. The fully intrinsically disordered HMGA protein family is one of those competitors[30,41] ***(Figure 8)***. However, the molecular mechanism underlying HMGA nucleosome remodeling ability is still unclear. In order to analyze this phenomenon, there is a need for a good method that can observe the structural changes within the protein upon binding. The problem is that IDPs, such as the HMGA proteins, display a random, dynamic behavior and fluctuate between states[42]. Their lack of a stable secondary structure and tertiary interactions, combined with their sensitivity to measurement conditions, hinders many conventional biochemical and biophysical methods to analyze proteins[43], such as X-ray crystallography[44] or cryo-electron microscopy studies. Nonetheless, integrative approaches have been proven to be powerful tools to study dynamic proteins by combining experimental techniques, (e.g. NMR and small-angle X-ray scattering) with computer simulations, predictions and single molecule approaches[5].

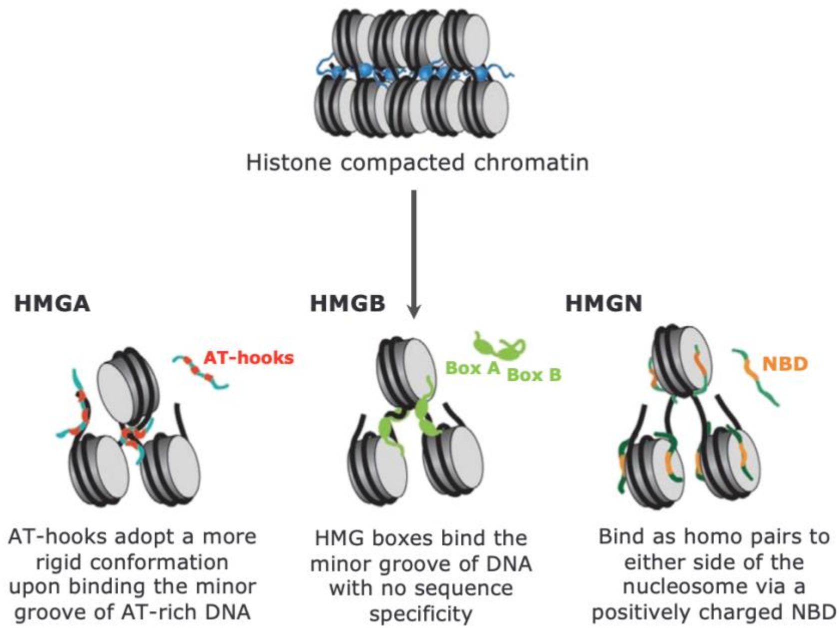

Figure 8: Schematic illustration of the binding modes of H1 competitors in a nucleosomal context. *The tightly packed structure of chromatin with bound H1 is remodeled by the disordered competitors. HMGA, HMGB and HMGN all have distinct binding modes (adapted figure[45]).*

1.7 smFRET: A powerful technique to study IDPs

Single-molecule Förster resonance energy transfer, also known as smFRET, is a powerful technique to study biomolecules. Its application has increased a lot in the past years due to its versatility, but it can be used to determine distances, structures, and the dynamics of biological molecules *in vivo* and *in vitro*[46]. Multiple conformations within and interactions between the molecules themselves can be detected and their kinetics determined as well[47]. The approach is now widely used to characterize the physicochemical properties of intrinsically disordered proteins in isolation and is being increasingly applied to more complex assemblies and experimental environments[43]. The method is dated back to 1996, but since then smFRET has rapidly developed and answered some fundamental questions about replication, transcription, translation, protein folding, and conformational changes, just to name a few[48]. The approach is currently being used by many high-profile labs, e.g. to quantify the strength of interaction between H1 and ProTα[49], to investigate the β-arrestin1 activation mechanism[50], and to investigate receptor dimers in living cells, as well as agonist-induced structural dynamics between promoters[51].

FRET is a non-radiative energy transfer process from a donor to an acceptor, but the occurrence of FRET is determined by the absorption and emission spectra, as well as the distance between neighboring fluorophores. smFRET is an application of FRET[52], where a donor fluorophore is excited by a specific wavelength of light. The donor fluorophore transfers energy to an acceptor fluorophore through a dipole-dipole interaction, between the electronic states of the donor and acceptor[43,52]. The energy transfer takes place when the oscillations of an optically induced electronic coherence of the donor are resonant with the energy gap of the acceptor. FRET is often specified as a “spectroscopic ruler” since it´s not necessarily accurate in measuring absolute distances between donor and acceptor fluorophores. Despite that, smFRET is sensitive enough to probe the temporal fluctuation dynamics of distance, such as the conformational changes of proteins[52]. This provides a conformational landscape of the protein with respect to this specific distance[53]. The distance between a donor and an acceptor molecule determines the efficiency of energy transfer, but the distance between them needs to be within a 1-10 nm range. Then the FRET is considered favorable[52]. The most common and simple smFRET setup is the confocal observation of freely diffusing single molecules. Other setups also exist, e.g., TIRF-FRET and FRET-FCS. ***Figure 9*** shows an example schematic diagram of TIRF-FRET and confocal FRET[43].

Recording high-quality single-molecule data requires samples of high quality. Today, there is a wide range of strategies for labelling proteins and other biomolecules available. The most common approach is the derivatization of two reactive cysteine residues with maleimide-functionalized dyes. By using this method, the proteins are site-specifically labelled with a donor and an acceptor dye for smFRET experiments[54]. The aim of this project was to characterize the potential interactions between HMGA and DNA or histone proteins, and the subsequent modulation of nucleosome structure, using smFRET. The effects of binding HMGA to the nucleosome, and the conformational changes of a labelled HMGA protein, upon binding to DNA, will shed light on the molecular principles of HMGA activity.

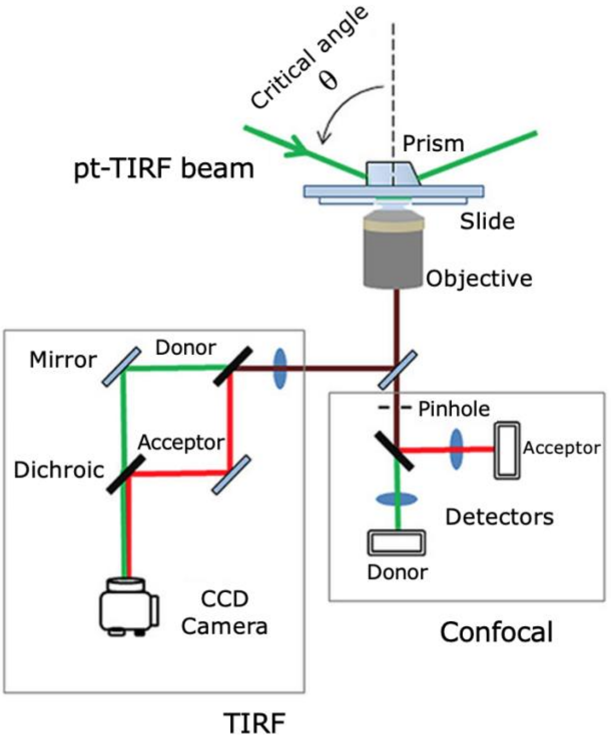

Figure 9: A schematic showing setups of TIRF-FRET and confocal FRET*. The collection of fluorescence emission on a TIRF microscope involves utilizing a CCD camera, whereas confocal techniques employ avalanche photodiodes to achieve faster acquisition times (adapted figure[45]).*

2 Objectives

The original aim of this project was to characterize the potential interactions between HMGA1a and DNA, or histone proteins. As well as the subsequent modulation of the nucleosome structure upon HMGA1a binding to it, using the smFRET approach. To observe such things, we needed to get a pure, double-labelled HMGA1a protein to use for the measurements. To attach the dye molecules to the protein, the original sequence of HMGA1a had to be mutated at two positions, S4C and A69C.

The first step to achieve this was to express the double cysteine mutant HMGA1a protein in *E. coli* cells. Next, the protein had to be purified with a variety of different methods, and finally, it had to be labelled with a single donor and acceptor dye at the two cysteine residue positions.

If all these steps were to be accomplished, the results would help us achieve a better understanding of HMGA1a and its characteristics and structural features. As well as the molecular mechanisms underlying the activity of the protein. It would also shed light on its ability to compete with histone protein H1 on binding the nucleosome. On top of that, the groundwork for smFRET measurements on HMGA1a would be laid out and a successful protocol made for its expression, purification, and labelling.

3 Materials and methods

3.1 Buffers and gels

The following chapter contains the recipes for all the buffers and gels needed for the expression, purification and labelling of the double cysteine mutant HMGA1a protein.

Table 1: *Luria Broth (LB) media for the expression of the protein in E. coli cells. Instructions for 1 L.*

LB media
10 g peptone casein
5 g yeast extract
10 g NaCl

Table 2: *HisTrap binding buffer (also used as a lysis buffer) and HisTrap elution buffer for nickel column purification.*

HisTrap binding buffer	HisTrap elution buffer
Tris 50 mM	Tris 50 mM
NaCl 150 mM	NaCl 150 mM
Urea 6M	Urea 6M
Imidazole 10 mM	Imidazole 300 mM
pH = 8.0	pH = 8.0

Table 3: *Instructions for making 12% and 15% SDS-PAGE separating gels. Protocols for 1 gel.*

Protocol for 1 gel	12% Separating gel	15% Separating gel
Chemical	**Volume**	**Volume**
Acrylamide	1.125 mL	1.408 mL
H_2O	1.633 mL	1.350 mL
Tris 1.5 M	0.975 mL	0.975 mL
10% APS	18.75 µL	18.75 µL
TEMED	1.875 µL	1.875 µL
Total volume	3.753 mL	3.753 mL

Table 4: *Recipe for 4% SDS-PAGE stacking gel. Protocol for 1 gel.*

Protocol for 1 gel	**4% Stacking gel**
Chemical	**Volume**
Acrylamide	1.107 mL
H_2O	0.677 mL
Tris 0.5 M	0.281 mL
10% APS	5.357 µL
TEMED	1.072 µL
Total volume	1.072 mL

Table 5: *Phosphate binding buffer (also used for dialysis) and phosphate elution buffer for cation exchange purification.*

Phosphate binding buffer	**Phosphate elution buffer**
Na_2HPO_4 50 mM	Na_2HPO_4 50 mM
NaCl 50 mM	NaCl 1.2 M
4 M Urea	4 M Urea
H_2O	H_2O
pH = 7.45	pH = 7.45

Table 6: *Ulp1 buffer for digestion of the SUMO tag.*

Ulp1 buffer
25 mM Tris
50 mM NaCl
10 mM DTT
pH = 7-8

Table 7: *The labelling buffer.*

Labelling buffer
0.307 mL K_2HPO_4
0.1925 mL KH_2PO_4
2 M Urea
H_2O
pH = 7-7.4
Final volume = 5 mL

Table 8: *Buffers for smFRET sample preparation.*

TEK165+ buffer	**50x++**	**10x TE buffer**	**TEK165 buffer**
TEK165 buffer	0.05 % Tween	100 mM Tris	10x TE buffer
0.01% Tween added	7.15 M β-mercaptoethanol	10 mM EDTA	1.65 M KCl
	Nuclease free H_2O	pH = 7.6	Diluted w/nuclease free H_2O

3.2 Protocols

3.2.1 Protocol 1: Expression of SUMO-tagged HMGA1a

The following protocol was used to induce the expression of the SUMO-tagged double cysteine mutant HMGA1a (S4C_A69C_HMGA). The SUMO tag makes the expression and purification process easier. The plasmid containing the HMGA gene with the desired mutations was already produced and transformed into expression hosts. For the expression, BL21 pLysS cells were used as ideal expression hosts.

1. Transformed BL21 pLysS cells with SUMO-tagged S4C_A69C_HMGA plasmid were used for overnight culture.
2. For the expression, 10 mL of the overnight culture was added to 1 L LB media **(*Table 1)***, which contained 1000 μL of both kanamycin and chloramphenicol. This was all put in a 5 L flask.
3. The cells were allowed to grow at 37 °C and 150 rpm until the optical density (OD) at 600 nm reached 0.6. Innovia 44 incubator shaker was used for the cultivation.
4. When the $OD_{600\ nm}$ reached 0.6, IPTG (Isopropyl-β-D-Thiogalactopyranoside) was added to the flask with a final concentration of 0.4 mM. IPTG induces the expression of the protein.
5. The cells were allowed to express the protein for about 3 hours before they were harvested.
6. The contents of the flask were then centrifuged at 4 °C for 7 min. at 7000 rpm.
7. Bacteria cells were harvested as pellets and the pellet was frozen immediately in liquid N_2 prior to storage at -80 °C.

3.2.2 Protocol 2: Cell lysis

The following protocol was used to lyse the B21 pLysS cells to extract the SUMO-tagged HMGA1a protein from them.

1. The cell pellet was resuspended in 15 mL nickel (Ni^{2+}) column binding buffer, called HisTrap binding buffer **(*Table 2)***.
2. The solution was then sonicated (power, amp. 50%) for 5 seconds and it was given 15 seconds of rest in between, for a total of 2 minutes on.
3. To collect the supernatant, lysate was centrifuged for 50 minutes 40.000 x g at 4°C.

3.2.3 Protocol 3: Nickel column purification on SUMO-tagged HMGA1a

The following protocol was the first step in purifying the SUMO-tagged double cysteine mutant HMGA1a, using a Nickel column on the “Äkta Pure” liquid chromatography system. The protein purification was carried out using a HisTrap 5 mL HP column. The HisTrap column has highly cross-linked agarose beads with an immobilized chelating group, which is Ni^{2+} in this case. The 6 His-residues that are present in the N-terminal of the SUMO-tag have an affinity for Ni^{2+} ions, allowing separation of proteins based on their affinity for the matrix of the column. Proteins with no affinity for the matrix will go through the column without binding to it, while those that interact with the column matrix will stick to the column until a competitor molecule disrupts their interaction. The proper flow rate for this column was 5 mL/min and the column volume (CV) was 5mL. The wavelengths for monitoring proteins were 220 nm and 280 nm, representing the peptide bonds and aromatic residues, respectively. HMGA1a does not have any aromatic residues, however the SUMO-tag has a small absorption at 280 nm.

1. First, the instrument was equilibrated with 10 CV HisTrap binding buffer **(*Table 2)***.
    - o It is important for the HisTrap binding buffer to contain some amount of imidazole so that the binding of proteins to the column becomes more specific.
2. The next step was sample application where the protein was injected into the column.
3. Following that, the column was washed with 10 CV HisTrap binding buffer.
4. The HisTrap elution buffer **(*Table 2)***, was then added to the flow in a linear gradient from 10 mM imidazole concentration to 300 mM imidazole concentration, to release HMGA1a protein from the column and let it elute.
    - o The high concentration of imidazole in the elution buffer is important because imidazole competes with the His residues on the SUMO-tag found on our protein for binding the column.
5. Every 2 mL of the samples from the elution step was collected in separate 5 mL Eppendorf tubes. They were then analyzed on a SDS-PAGE gel.

3.2.4 Protocol 4: Hand-casted SDS-PAGE gel

Throughout the purification process, SDS-PAGE gels were used to enable us detecting the protein of interest. Here is the protocol for making a home-made SDS-PAGE gel.

1. All the components of a gel-loading cassette needed for casting a gel were cleaned and assembled.
2. The components of the separation gel were mixed together in the right order **(*Table 3)***. The ingredients were added in the following order:
    - o H_2O + Tris
    - o Acrylamide
    - o APS
    - o TEMED
3. The separation gel solution was then transferred quickly but carefully to the cassette, leaving 1 cm empty for the stacking gel.
4. Isopropanol was applied to the top of the gel to get rid of any air bubbles.
5. The gel was allowed to harden for approximately 30 min.
6. The isopropanol was moved from the gel by tilting the cassette over the sink. A paper towel was used to remove most of the isopropanol.
7. The components of a 4% stacking gel were mixed together in the right order **(*Table 4)***.
8. The stacking gel solution was transferred carefully to the cassette on top of the separation gel, filling the cassette all the way to the top.

9. A gel comb was then inserted through the stacking gel while trying to avoid leaving bubbles.
10. The stacking gel was allowed to polymerase for about 20 minutes before removing the comb.

3.2.5 Protocol 5: SDS-PAGE electrophoresis

This section contains protocols for the electrophoresis of home-made gels and pre-made gels, but both were used for the analysis of protein samples throughout the purification process.

SDS-PAGE gel electrophoresis using home-made gels

1. The tank and equipment were cleaned and dried.
2. The assembled cassette was placed in the tank and filled up with used 1X Running Buffer to the correct line, based on how many gels you have.
3. Fresh 1X Running Buffer was added in between the cassettes and filled all the way up so the liquid was covering the whole glass.
4. The samples were then prepared.
    - 14 µL of each sample was mixed with 1 µL of DDT (from 2 M stock) and 5 µL of 4x loading buffer.
    - Then the samples were heated for 10 min. at 90 °C.
5. Then 15 µL of the samples were loaded to the wells of the gel, as well as 2 µL of PageRuler™ Prestained Protein Ladder.
6. The lid was placed on the container and connected to a power supply.
7. The gel was run at 90 V for ~20 min. while the samples passed through the stacking gel.
8. After ~20 min. the gel was run at 120 V for ~100 min or until the front dye eluted.
9. The instrument was turned off and the gel was removed from the cassette.
10. The gel was placed in a plastic box including a staining solution, then heated, and then stained for 15 min while tilting.
11. Then it was washed with heated water for 5 min., three times to destain it, while tilting.
12. Bio-Rad Gel Dock instrument was used for imaging the gels.

SDS-PAGE gel electrophoresis using pre-casted gels

1. The tank and equipment were cleaned and dried.
2. Buffer was added to the tank to fill up to the marked line.
    - Fresh 1X MES Running buffer was used on the side of the cassette where the wells are exposed and used 1X MES running buffer on the other side of it.
3. The cassette was unpacked, the insulating strip removed and washed with a small amount of dH_2O.
4. The samples were then prepared, the same way as for home-made gels.
5. Then 15 µL of the samples were loaded to the wells of the gel, as well as 2 µL of PageRuler™ Prestained Protein Ladder.
6. The lid was placed on the container and connected to a power supply.
7. The gel was run at 200 V for ~20 min., or until the front dye eluted.

8. The instrument was turned off and the gel was removed from the cassette.
9. The gel was stained and destained, the same way as for the home-made gels.
    - Bio-Rad Gel Dock instrument was used for imaging the gels.

3.2.6 Protocol 6: Dialysis

To change the buffer that your protein of interest is in, dialysis is performed.

1. A dialysis membrane with a diameter of 29 cm and a 3.5 kDa cut-off was soaked in de-ionized H_2O for a few minutes.
2. The SUMO-tagged HMGA1a protein sample was then placed in the dialysis membrane and clamps were put on the ends to close them off. The dialysis membrane was put in 3 L of phosphate binding buffer ***(Table 5).***
    - It is important to do this properly to prevent the sample from leaking out of the tube/buffer and getting inside of it.

3.2.7 Protocol 7: Cation exchange column purification on SUMO-tagged HMGA1a

The following protocol was the second step in purifying the SUMO-tagged HMGA1a protein using a cation exchange column on the "Äkta Pure" liquid chromatography system. The purification was carried out in a HiTrap 5 mL SP HP column. Ion exchange chromatography involves the reversible binding of charged sample molecules to oppositely charged groups that are attached to an insoluble matrix. The pH at which a biomolecule has no net charge is called the isoelectric point (pI). If the pH is below the biomolecule's pI, it becomes positively charged and binds to a cation exchanger[55]. Here the cation exchanger is $-CH_2CH_2CH_2SO_3$. The theoretical pI of the SUMO-tagged HMGA1a is 10.11 and the theoretical pI of the protein without the SUMO is 9.22. The flow rate throughout the process was 5 mL/min and the CV was 5 mL. The wavelengths being monitored were 220 and 280 nm.

1. First, the instrument was equilibrated with a phosphate-binding buffer ***(Table 5)***.
2. The next step was sample application where the protein was injected into the column.
3. Following that, the column was washed with 10 CV of the phosphate-binding buffer.
4. The phosphate elution buffer ***(Table 5)*** was then added to the flow in a step-linear gradient as we used 10 CV of the elution buffer to reach 1.2 M NaCl.
5. Every 2 mL of samples from the elution, were collected and then they were analyzed on a gel.

3.2.8 Protocol 8: Purification of SUMO-tagged HMGA1a with Reversed-phase high-performance liquid chromatography (RP-HPLC)

After the two purification steps using the ÄKTA instrument and the removal of the SUMO-tag, the HMGA1a was further purified with the Agilent 1260 Infinity II RP-HPLC system using a C18 Reprosil column. Before every RP-HPLC step, the protein sample was concentrated down to decrease its volume. For that, we used Amicon Ultra-15 centrifugal filters. The membrane cut-off was 3k and then we centrifuged the solution for 40 min. at 4000 x g and 15 °C. Flow rate throughout the purification was 1 mL/min and the wavelengths being monitored were 220 nm and 280 nm. Two mobile phases were used for the RP-HPLC. Buffer A contains 0.1% trifluoro acetic acid (TFA) in dH_2O while buffer B contains 0.08 % TFA in acetonitrile.

1. After equilibrating the column with 95% buffer A, 5% buffer B, the sample was injected into the column, while 100% of buffer A was running at 1 mL/min.
2. Buffer B is then added to the flow from 5% buffer B to 100% buffer B. See ***Table 9*** for a detailed description of the elution process.

Table 9: *The elution process and the flow of Buffer A and B.*

Time (min)	A (%)	B (%)	Flow (mL/min)
0.00	95.0	5.0	1.000
2.00	95.0	5.0	-
2000	75.0	25.0	-
40.00	20.0	80.0	-
50.00	0.0	100.0	-
55.00	0.0	100.0	-

Samples were collected according to the peaks of different spectra, and were put straight into liquid nitrogen. Then they were either freeze-dried or stored in a -80 °C freezer.

3.2.9 Protocol 9: Ulp1 digestion and the removal of SUMO

For the removal of the SUM tag the Ulp1 protease was used. Freeze-dried samples from RP-HPLC were dissolved in Ulp1 digestion buffer ***(Table 6)***, containing 25 mM Tris and 50 mM NaCl, with a pH between 7-8. DTT (10 mM) was then added, and the digestion reaction was started upon adding the proper amount of the enzyme. Digested samples were then purified again with RP-HPLC, using the same method as in **Protocol 8**. The only difference was that samples were incubated with 10 mM DTT for 20 minutes prior to being injected to the column.

3.2.10 Protocol 10: Fluorophore labelling of HMGA1a

The HMGA1a protein was labelled with Cy3b and CF660R maleimide dyes as fluorescent donor and acceptor, respectively. The dyes are light-sensitive, so they were kept away from light exposure by being wrapped in aluminum. The samples were labelled with a labelling strategy called simultaneous labelling (also known as shotgun labelling), which is when both dyes (Cy3b and CF660R) are added at the same time. The advantage of this technique is that only one chromatography run is needed after the labelling[54]. Two fractions from the last purification step on the RP-HPLC were used for labelling, fractions from min. 14-16 and from min. 26-28.

1. Before labelling, 50 mg/μL aliquotes of the freeze-dried dyes (Cy3B and CF660R) were dissolved in 10 μL of DMSO, to keep the dye concentration similar.
2. The dyes were sonicated for 15 minutes.
3. Freeze-dried protein samples were then dissolved in a de-gassed labelling buffer ***(Table 7)***.
4. The protein concentration was measured in a Nanodrop instrument.
5. Dyes were added to protein aliquots in a ratio of 3:1 (dye:protein).
6. The aliquots were incubated for 2-3 hours at room temperature away from light.
7. Then 10 mM DTT was added to quench the reactions and reduce potentially oxidized cysteines.
8. Then the aliquots were incubated for 20 minutes at room temperature.
9. The labelled protein samples were then flash-freezed in liquid nitrogen, before storing them in a -80 °C freezer.

After the labelling process, the aliquots were purified again with the RP-HPLC system, using the same protocol as before, the only difference being the wavelengths being monitored, that is: 220, 280, 520, 560 and 660 nm to detect the protein and dyes. The samples were collected in 1.5 mL Eppendorf tubes from min every one min since we didn't know exactly when our labelled protein would elute.

3.2.11 Protocol 11: Preparation of the protein sample for smFRET labelling measurements

A few solutions need to be prepared fresh before every measurement session Prior to the measurement of the protein on the fluorescent microscope. First, TEK165+ is used for diluting the protein sample. From that diluted sample, we add 0.4 μL of it to a solution containing 0.8 μL of 50x++ and 38.8 μL of TEK165 ***(Table 8)***. The TEK165 buffer is created from 10x TE buffer.

Before each smFRET measurement, 40 μL of the labelled protein sample was placed on a μ-slide Angiogenesis chambered coverslip. A drop of water was added to the lens before it was located right under the sample to find the focus point of the protein sample solution. All measurements were performed in a dark room and each one lasted for 5 or 15 minutes.

4 Results

4.1 The expression of HMGA1a

Prior to the expression process, the HMGA1 gene was successfully cloned in the proper SUMO vector by my supervisor. This provided us with a SUMO-tagged protein that was ready to be expressed in B21 pLysS cells. SUMO-based fusion tag technology is now often used for the expression of recombinant proteins. *E. coli* recombinant proteins containing a SUMO-tag have e.g., shown enhanced stability and solubility, leading to much better yields over constructs missing the tag. The SUMO-tag can then be easily removed by Ulp1 protease allowing the generation of a native N-terminus of the target protein[56]. The SUMO-tag contains a 6xHis-tag sequence, placed on the N-terminal of the protein. The 6xHis-tag allows us to do an affinity chromatography purification with a nickel column, since the negatively charged His residues bind to the positively charged Ni^{2+} ions in the matrix of the column. The expression of the SUMO-tagged HMGA1a protein was successful, so cell lysis was performed to extract the protein, so we could start its purification.

4.2 The purification of HMGA1a

The purification of HMGA1a was carried out using the Äkta Pure liquid chromatography system and the Agilent 1260 Infinity II RP-HPLC system.

4.2.1 Purification using the ÄKTA chromatography system

The first purification step on the mutant SUMO-tagged HMGA1a protein was done by using a nickel column on the Äkta Pure liquid chromatography system. The isolation of the protein from the cell lysate was carried out in a HisTrap 5 mL HP column containing Ni^{2+} ions in the matrix. For the elution of the protein, the sample was injected while 100% of the phosphate HisTrap binding buffer was running at the speed of 5 mL/min. To release our protein from the column, the HisTrap elution buffer was added to the flow in a linear gradient from 0-100% with a CV of 10. The fractions were collected in 5 mL Eppendorf tubes every 2 minutes, and the wavelengths being monitored were 220 nm, for the peptide bonds of the protein, and 280 nm, for the aromatic residues found in the SUMO-tag. We had one desirable peak that had an absorption at 220 nm and 280 nm, that showed up when 118 mL of the HisTrap elution buffer had passed through the column ***(Figure 10)***. Every fraction from the elution that had an absorption at these two wavelengths was collected and run on an SDS page gel.

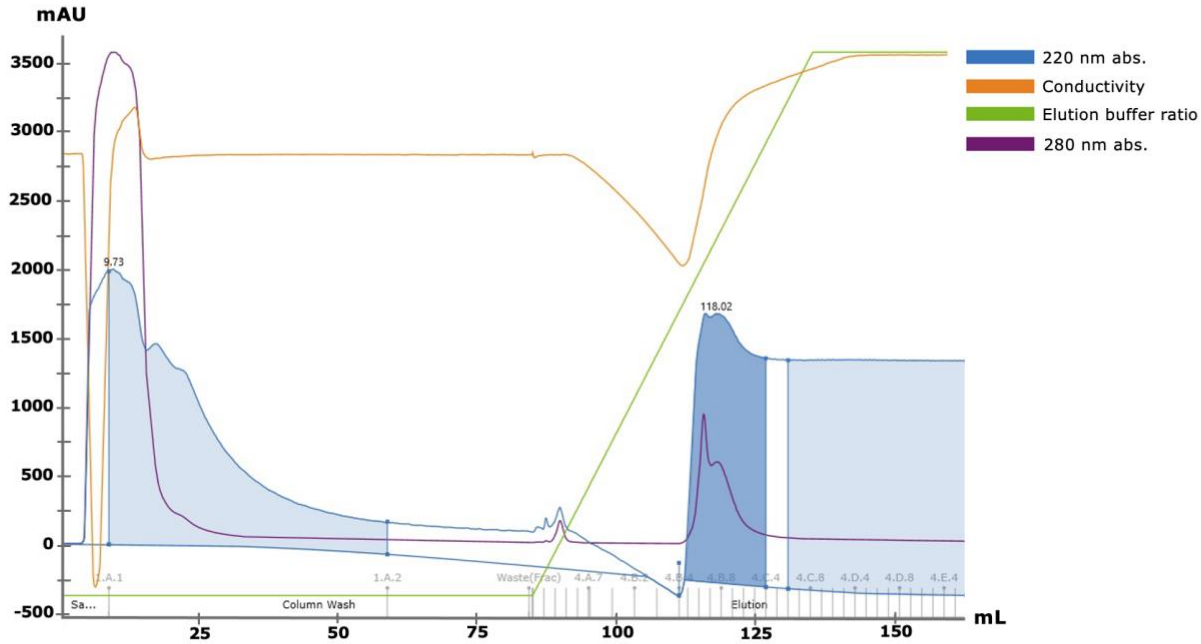

Figure 10: Isolation of SUMO-tagged HMGA1a protein from cell lysate with nickel column purification. *The y-axis shows the signal intensity in absorption units (mAU) and the x-axis shows the retention in mL, that is the amount of liquid that has passed through the column. The blue line shows the signal at 220 nm (for peptide bonds), the purple line shows the signal at 280 nm (for aromatic residues), the green line shows the flow of the His-Trap elution buffer, and the orange line shows the conductivity.*

The samples were analyzed on a SDS-PAGE gel to confirm if our protein of interest was eluted during the process **(*Figure 11)***, and corresponded to the elution peak we obtained. All the wells were loaded with elution samples from the nickel column. The theoretical Mw of the HMGA1a protein with the SUMO tag is around 23.9 kDa. However, the band for the protein shows up on the gel somewhere between 35-40 kDa. The reason is there are many charged residues on the protein and these charges can affect the immigration of the protein on the gel. Wells containing samples 4B1-4B4 contain the supernatant, but the bands for our protein were visible in samples 4B8, 4C1, 4C2, 4C3, 4C4, and 4C5, so they were collected and used for further purification of the protein on the cation exchange column.

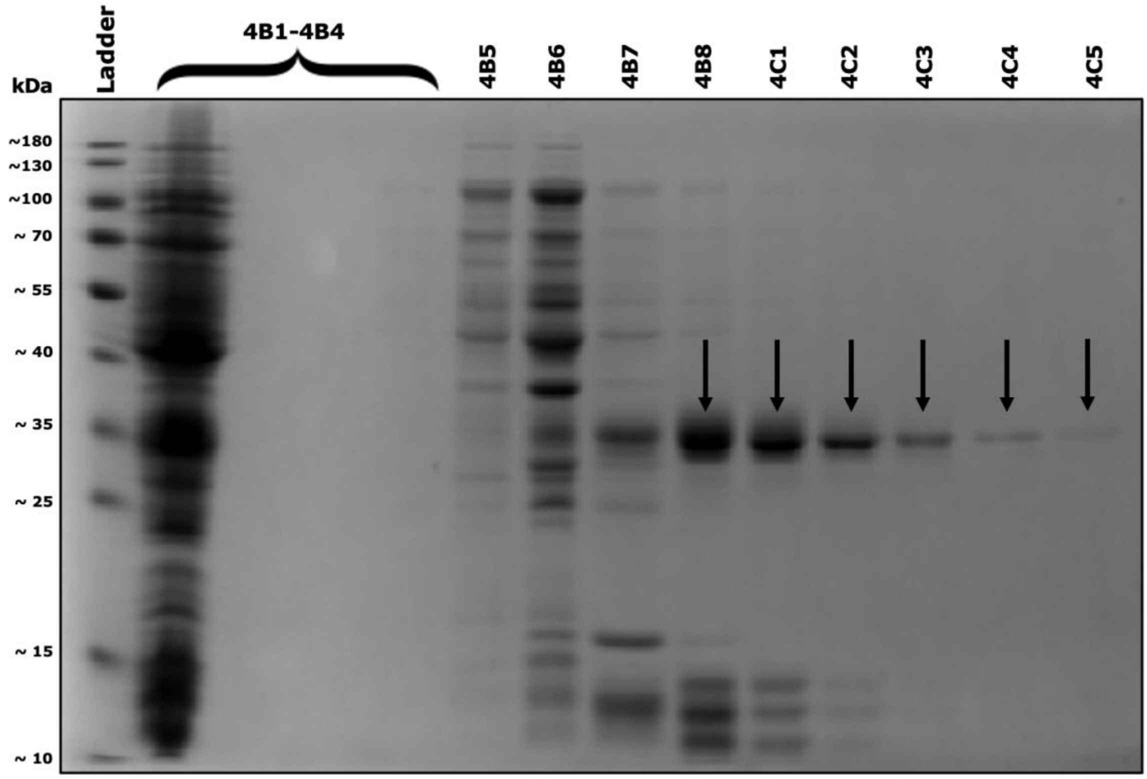

Figure 11: SDS-page gel containing fractions from the nickel column purification on the ÄKTA instrument. *This is a 15% SDS gel containing 13 samples and a 180 kDa ladder (PageRuler™ Prestained Protein Ladder) in the first lane. Samples 4B1-4B4 in lanes 2-5 contain the supernatant, but samples 4B8-4C5 in lanes 9-14 were used for further purification, the arrows point at these samples. The numbering refers to the tubes the samples got eluted into during the nickel purification.*

The second purification step on the SUMO-tagged HMGA1a was also done on the Äkta Pure liquid chromatography system but this time we used a cation exchange column. It was carried out in a HiTrap 5 mL SP HP column which contains cation exchanger -$CH_2CH_2CH_2SO_3$. Before injecting the protein samples that came off the nickel column, on the cation exchange column, they were mixed together and dialyzed to change the buffers. The sample was injected into the column while 100% of the phosphate binding buffer was running. The phosphate elution buffer was then added to the flow, to release our protein, in a step-linear gradient as we used 10 CV of the elution buffer to reach 1.2 M NaCl. The samples were collected every two min. in 5 mL Eppendorf tubes and the wavelengths being monitored were 220 and 280 nm. We obtained one desirable peak that had an absorption at both 220 nm and 280 nm, which showed up when ~65 mL of the phosphate elution buffer had passed through the column ***(Figure 12)***. Every fraction from the elution that had an absorption at these two wavelengths was collected and run on an SDS page gel.

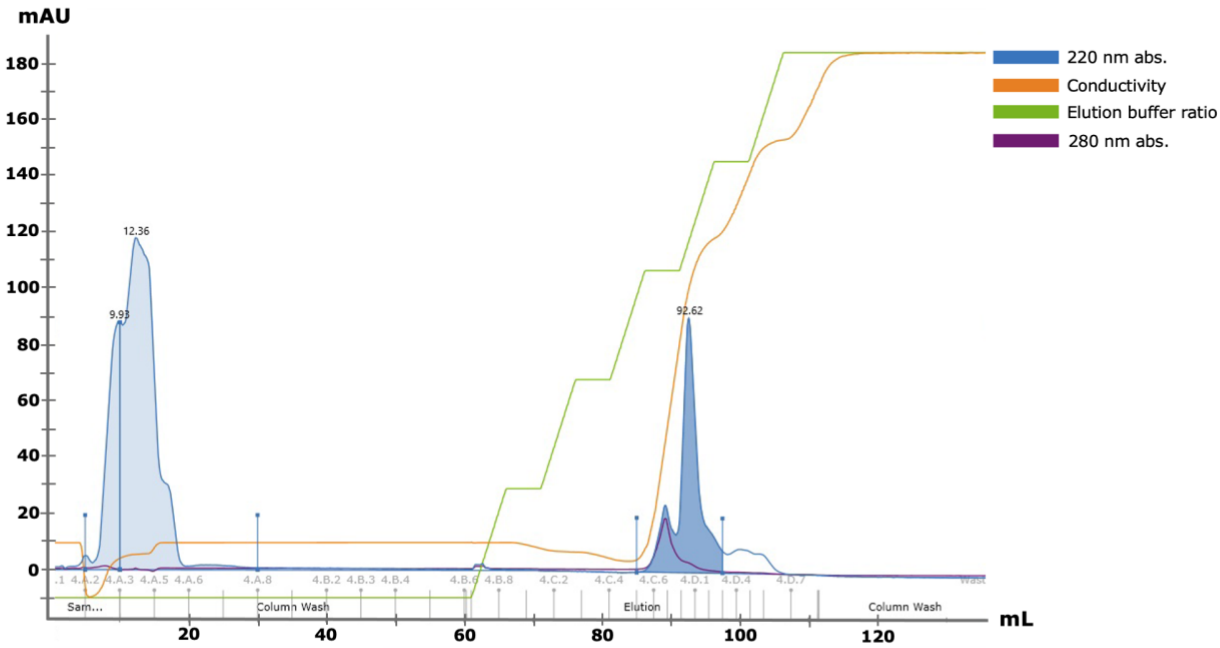

Figure 12: Isolation of SUMO-tagged HMGA1a protein with cation exchange purification. *The y-axis shows the signal intensity in absorption units (mAU) and the x-axis shows the retention in mL, that is the amount of liquid that has passed through the column. The x-axis also shows the collection tubes where the samples got eluted into. The blue line shows the signal at 220 nm (for peptide bonds), the purple line shows the signal at 280 nm (for aromatic residues), the green line shows the flow of the His-Trap elution buffer, and the orange line shows the conductivity.*

The elution samples from the peak were analyzed on an SDS-PAGE gel to see if our protein was eluted during the cation exchange purification ***(Figure 13)***. We expected our protein band to show up between 35 – 40 kDa since it still contained the SUMO tag. The bands for our protein are visible in samples: 5A4 + 5A5 + 5A6 + 5A7, and 5A8 + 5B1, but they were collected and used for further purification of the protein. It is very unlikely that the band from sample 4C8 + 4D1 contained our protein since it eluted at a different time compared to the other samples that had a protein band at ~35 kDa.

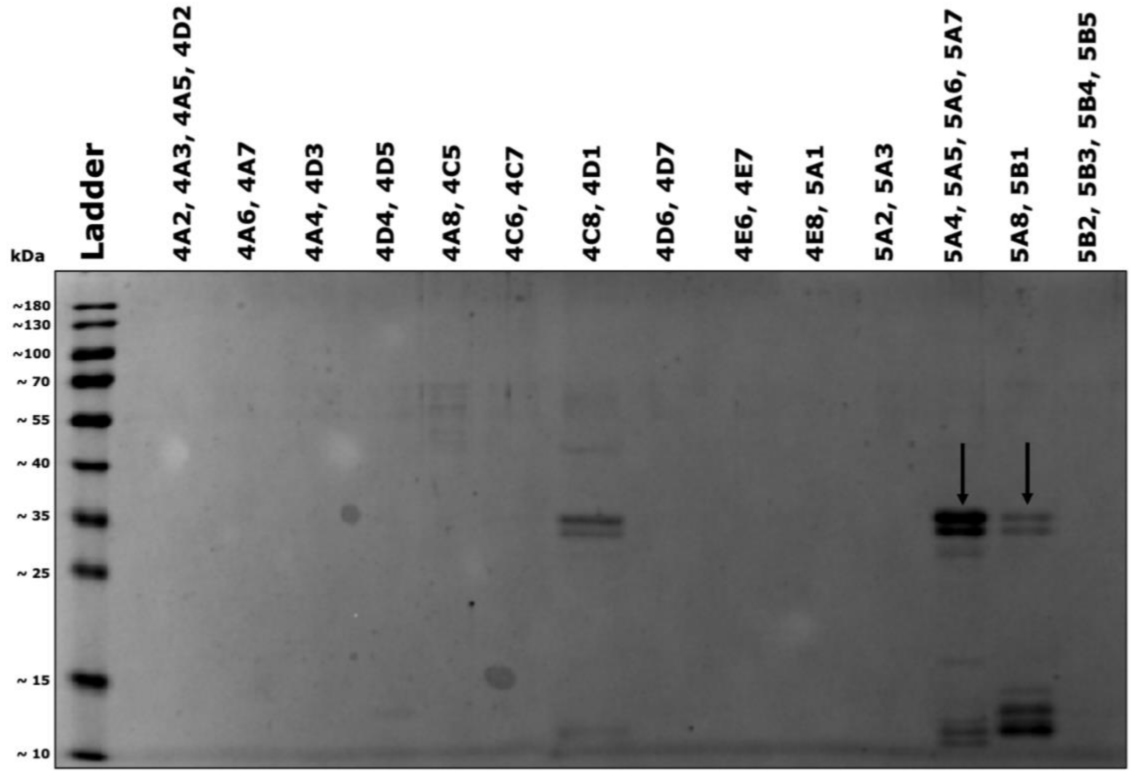

Figure 13: SDS-page gel containing fractions from the cation exchange purification on the ÄKTA instrument. *This is a 12% SDS gel containing 14 samples and a 180 kDa ladder (PageRuler^{TM} Prestained Protein Ladder) in the first lane. Lanes 2-14 contain samples from the elution process that do not contain our protein band. Samples 5A4 + 5A5 + 5A6 + 5A7, and 5A8 + 5B1 (wells 13 and 14) were used for further purification, the arrows point at these samples. The numbering refers to the tubes the samples got eluted into during the cation exchange purification. All wells contain samples that were combined before loading them into the wells.*

4.2.2 Purification using the RP-HPLC system

Further refinements had to be done on the SUMO-tagged HMGA1a protein before labelling. The next steps were carried out using the Agilent 1260 Infinity II RP-HPLC system and a C18 Reprosil column.

The protein sample was injected into the column while 100% of Buffer A was running at 1 mL/min. To release the protein, Buffer B was added to the flow in steps, from 5% buffer B to 100% buffer B, see ***Table 9*** for a detailed description of the elution process. The fractions were collected in 2 mL Eppendorf tubes every 2 minutes, and the wavelengths being monitored were 220 nm and 280 nm, since our protein still contained the SUMO tag. The protein peaks, flow through, and wash solutions were collected. The samples were put straight into liquid nitrogen, freeze-dried, and then stored in a -80 °C freezer. We obtained desirable peaks, with an absorbance of both 220 nm and 280 nm, after 17.5 minutes, from minutes 22.5 – 24.0 and after 28 and 31.5 min ***(Figure 14)***. These fractions from the elution were collected and run on an SDS page gel.

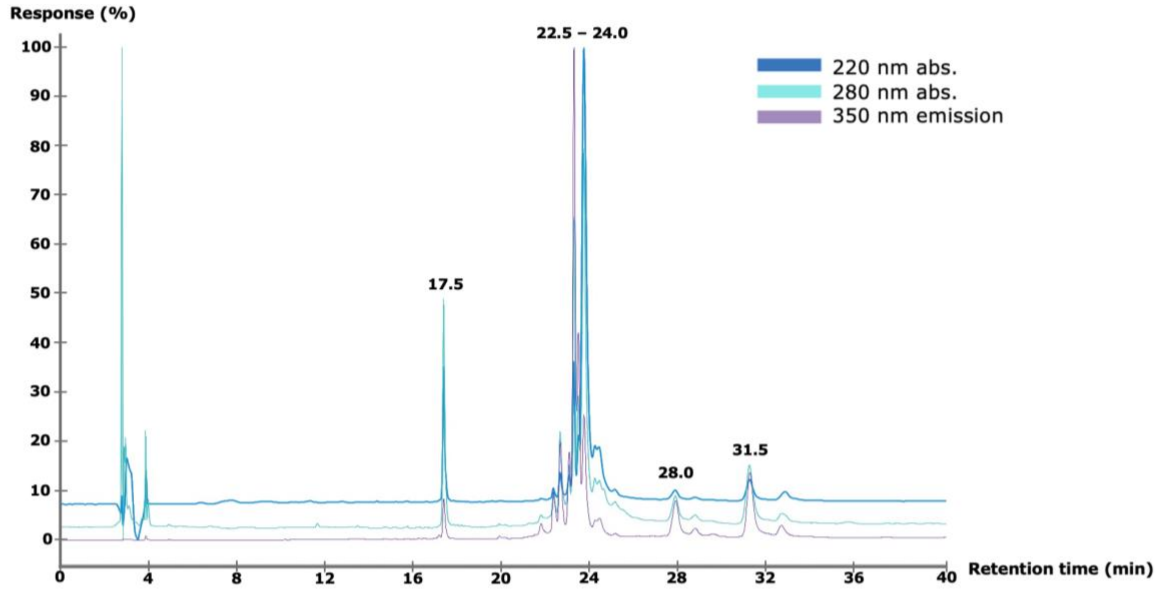

Figure 14: RP-HPLC chromatogram of SUMO-tagged HMGA1a. *The y-axis shows the relative response (%) while the x-axis shows the retention time (min.), that is the time since the sample was injected into the column. The dark blue line shows the signal at 220 nm, the light blue line shows the signal at 280, and the purple line shows the emission at 350 nm if the solution is excited at 280 nm. The peaks with an absorption at 220 nm and 280 nm were collected and run on a SDS-page gel.*

The fractions from the RP-HPLC purification were analyzed on an SDS-PAGE gel to see if they correspond to the size of our protein ***(Figure 15)***. We expected to see our protein band between 35 – 40 kDa so fractions from min. 25, 27, 29 and 31 were used for the Ulp1 digestion, where the SUMO tag was cut off our protein.

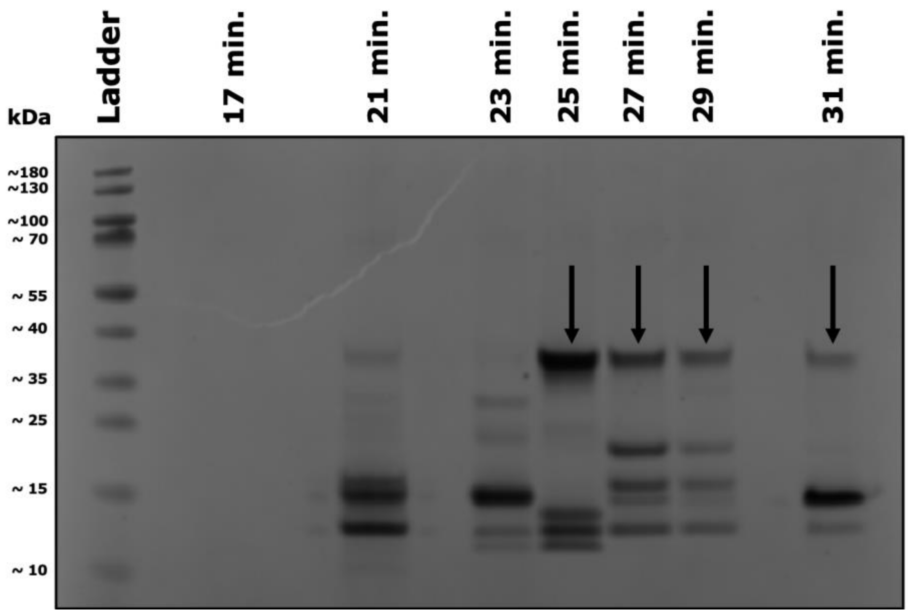

Figure 15: SDS-page gel containing fractions from the first RP-HPLC run on SUMO-tagged HMGA1a. *This is a pre-casted 10% SDS gel containing 7 samples and a 180 kDa protein ladder (PageRuler^{TM} Prestained Protein Ladder) in the first lane. Wells 3, 5, 7-10 and 12 contain elution samples that were each collected for 2 minutes during the RP-HPLC run. Fractions from min. 25, 27, 29 and 31 (wells 8, 9, 10 and 12) were used for the Ulp1 digestion and further purification of the protein, since one of them contained our protein band. The numbering refers to the minute the fraction eluted during the RP-HPLC run.*

After the removal of the SUMO tag with the Ulp1 protease, the same RP-HPLC method as before was used to further purify the HMGA1a protein ***(Figure 16)***. The only difference was that 10-15 min. before the injection, we added 5 mM of DTT (2 M stock) to the sample, to reduce its disulfate bonds. Since now the protein no longer contained the SUMO tag, the protein was expected to elute a bit earlier compared to the previous RP-HPLC runs. The fractions were collected in 2 mL Eppendorf tubes every 2 minutes, and the wavelength being monitored was just 220 nm since we only got absorption at 280 nm from the aromatic residues found in the SUMO tag. Samples were collected from minutes: 7-9, 14-16, 19-21, 21-23, 26-28 and 31-33 and run on a SDS-page gel.

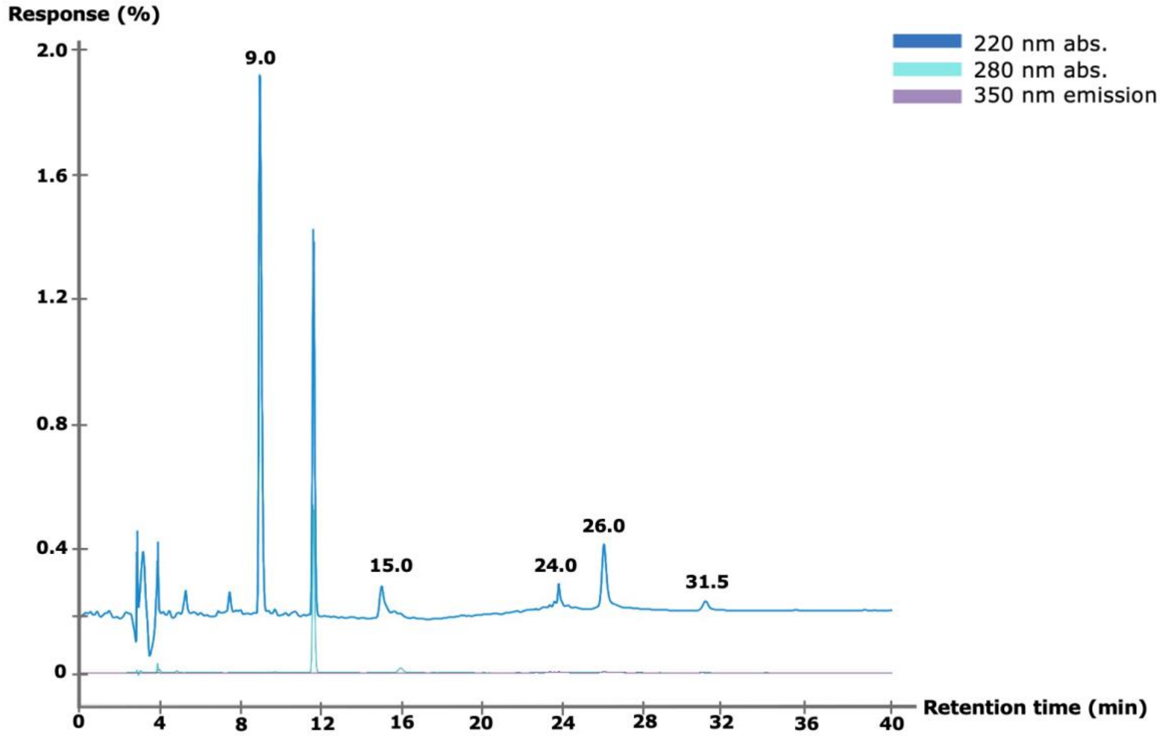

Figure 16: RP-HPLC chromatogram of HMGA1a. *The y-axis shows the relative response (%) while the x-axis shows the retention time (min.), that is the time since the sample was injected into the column. The dark blue line shows the signal at 220 nm, the light blue line shows the signal at 280, and the purple line shows the emission at 350 nm if the solution is excited at 280 nm. The peaks with absorption at 220 nm, starting at min. 7, were collected and run on a SDS-page gel.*

The fractions from the second RP-HPLC purification were analyzed on an SDS-PAGE gel to see if they correspond to the size of our protein ***(Figure 17)***. Now we expected to see our protein band standing a bit closer to its theoretical mW, that is ~11.7 kDa, since it no longer contained the SUMO tag. One of the samples from min. 14-16, 26-28 and 31-33 contained our protein band, so they were collected to be labelled and further purified.

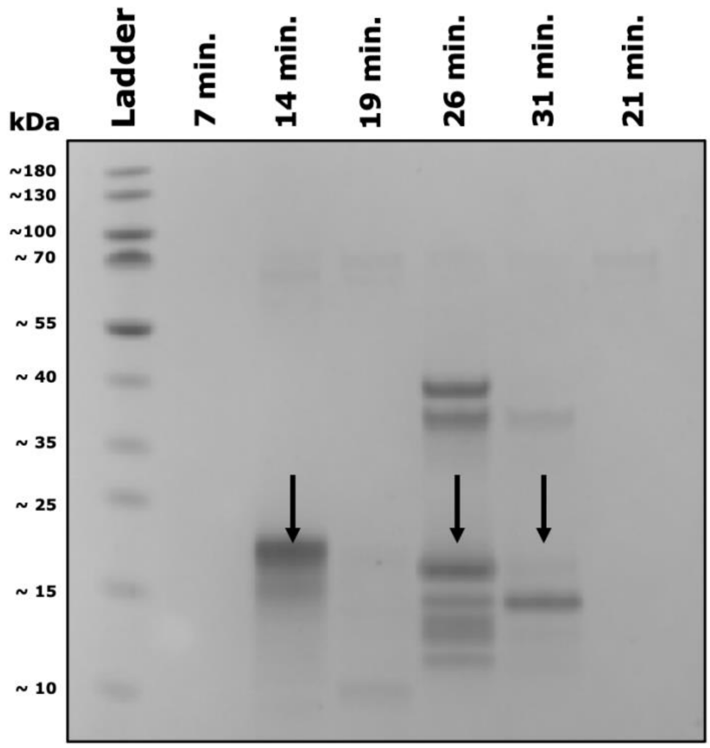

***Figure 17**: **SDS-page gel containing fractions from the second RP-HPLC run on HMGA1a.** This is a pre-casted 10% SDS gel containing 6 samples and a 180 kDa protein ladder (PageRuler™ Prestained Protein Ladder) in the first well. Wells 2-7 contain elution samples that were each collected for 2 minutes during the RP-HPLC run. Fractions from min. 14, 26 and 31 (wells 3, 5 and 6) were used for further purification but only one of them contained our protein band. The numbering refers to the minute the fraction eluted during the RP-HPLC run.*

4.3 The labelling of HMGA1a

For the labelling of HMGA1a, we used the "shotgun" labelling method. We labelled the protein by adding Cy3B and CF660R maleimide dyes to the two cysteine residues found at positions 4 and 69 in the amino acid sequence, which had previously been mutated from the original sequence ***(Figure 18)***. After the labelling process, HMGA1a was purified again on the RP-HPLC to refine it even further, and to isolate as much of the labelled protein without getting a surplus of dye molecules with it. The same RP-HPLC method was used as before, the only difference was the wavelengths being monitored, which were: 220, 280, 520, 560 and 660 nm. Per usual, 220 nm is for the detection of the peptide bonds of the protein. The rest of the wavelengths account for the dyes, but they contain groups in their chemical structure that absorb light at 280 nm. 520 nm is for the detection of Cy3b, but the dye can be excited at 520 nm, and in return it will emit light at 615 nm. 560 nm is the wavelength at which Cy3b has the max absorbance, but for CF660R it is 660 nm. For us to be able to detect the dyes, we must monitor their absorbance at these wavelengths.

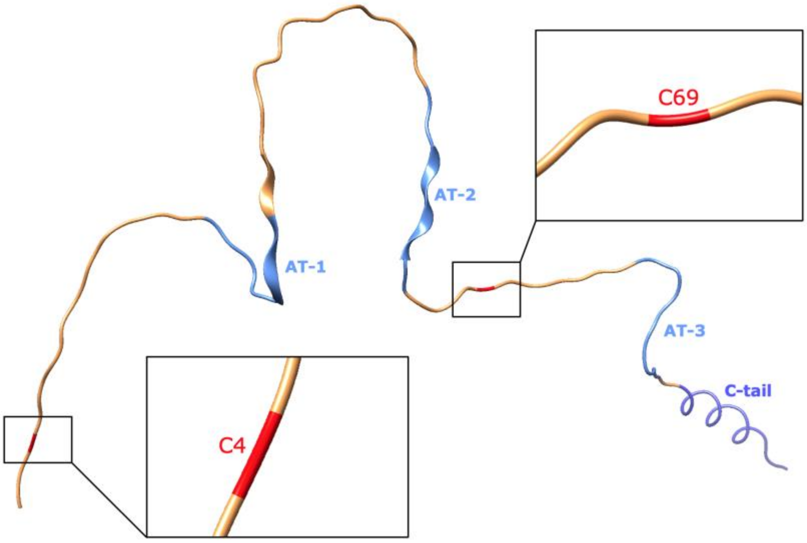

Figure 18: The labelling sites on HMGA1a. *The protein is labelled at positions 4 and 69 on the sequence at two cysteine residues (shown in red). The AT-hooks are also shown in blue and the acidic C-tail in purple. The picture is obtained from AlphaFold[17] and edited in Chimera.*

Figure 19 shows the results from the purification of labelled fraction 14 min. (min. 14-16) from the last RP-HPLC run before labelling, which can be seen on the SDS-page gel ***(Figure 17)***. To prevent missing any of the protein, samples were collected from min. 18-32 every one min. We collected everything because we had no idea which peak corresponded to our hopefully double-labelled protein.

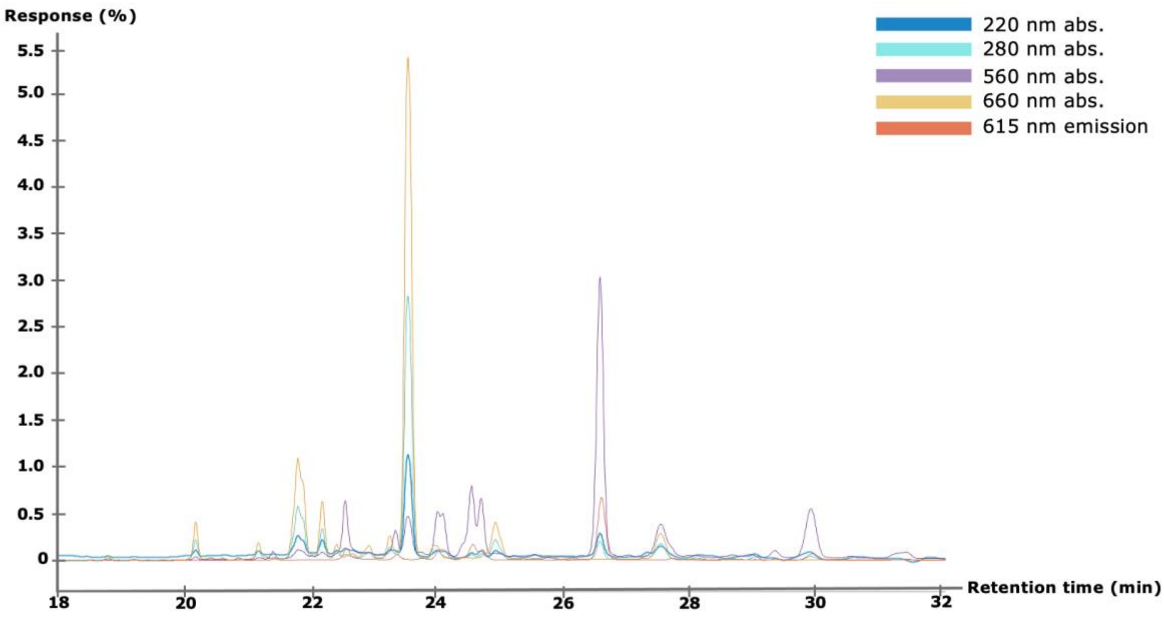

Figure 19: RP-HPLC chromatogram of the fraction "14 min." from the labelled HMGA1a. *The y-axis shows the relative response (%) while the x-axis shows the retention time (min.), that is the time since the sample was injected into the column. The dark blue line shows the signal at 220 nm, the light blue line shows the signal at 280, the purple line shows signal at 560 nm, the yellow line shows signal at 660 nm, and the orange line shows the emission at 615 nm if the solution is excited at 520 nm.*

Figure 20 shows the results from the purification of labelled fraction 26 min. (min. 26-28), also from the last RP-HPLC run before labelling, which can be seen on the SDS-page gel (***Figure 17***). Samples were collected from min. 18-40 every one min, for the same reason as before, so we wouldn't miss any of our protein samples.

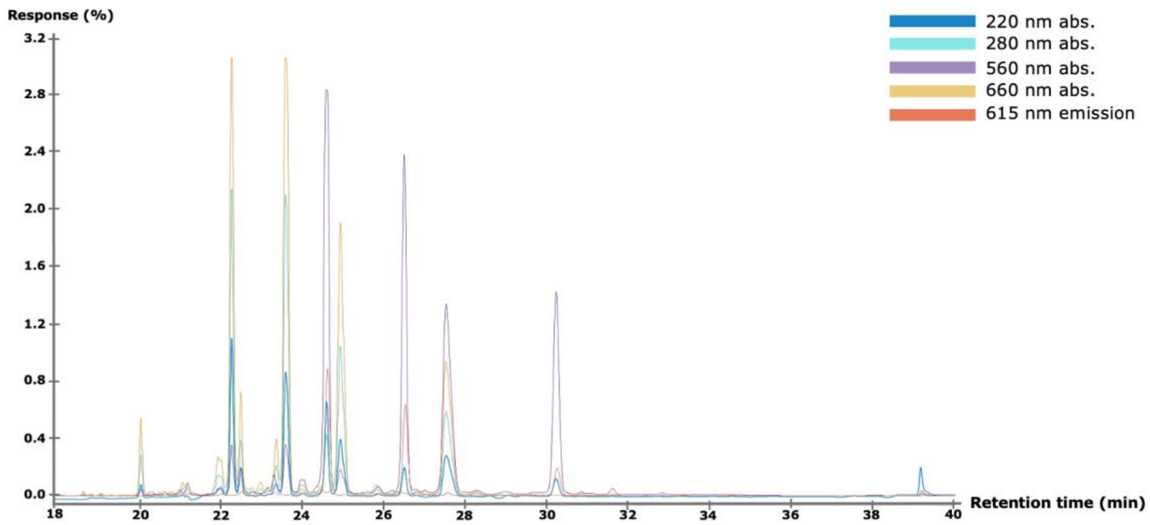

Figure 20: RP-HPLC chromatogram of the fraction "26 min." from the labelled HMGA1a. *The y-axis shows the relative response (%) while the x-axis shows the retention time (min.), that is the time since the sample was injected into the column. The dark blue line shows the signal at 220 nm, the light blue line shows the signal at 280, the purple line shows signal at 560 nm, the yellow line shows signal at 660 nm, and the orange line shows the emission at 615 nm if the solution is excited at 520 nm.*

After this RP-HPLC purification, samples were immediately flash-freezed with liquid nitrogen and then freeze-dried before storing them in a -80 °C freezer. Now they could be prepared in sample buffers for smFRET measurements ***(Table 11)***. To do actual smFRET measurements, including the HMGA1a mutant and e.g., a nucleosome, we needed to have a double-labelled protein. After the labelling process was complete, we remained hopeful that some of our protein was labelled. The only way to see if the labelling was successful was to measure the labelled protein samples on the smFRET instrument. Unfortunately, no double-labelled protein was obtained so further optimization must be done during the labelling process. Below are some examples of our smFRET results that looked the most promising to us in terms of the ratio between the absorbances at each wavelength. ***Figure 21*** shows a zoomed-in version of ***Figure 20***, showcasing the two most promising peaks from the shotgun labelling of fraction 26 min. (see ***Figure 17*** for the SDS-PAGE gel). The elution from these two peaks were used for smFRET test measurements, we called them "Sample 1" and "Sample 2".

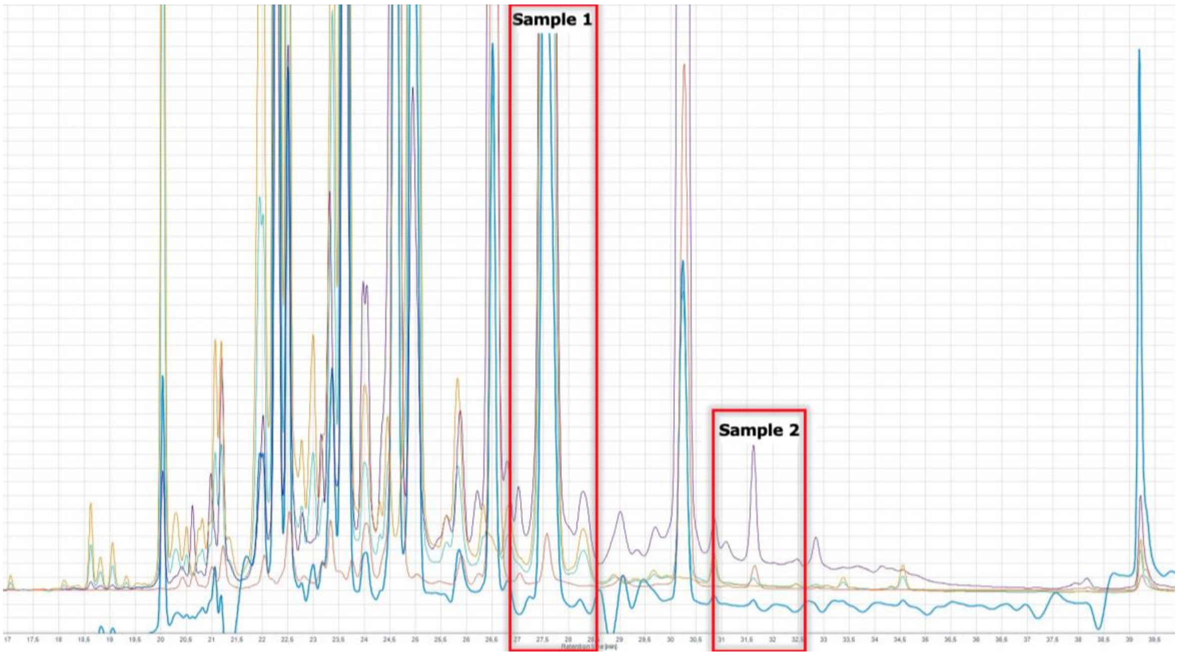

Figure 21: RP-HPLC chromatogram showing Sample 1 and Sample 2. This is a zoomed-in version of ***Figure 20***, *showing the most promising peaks in the red boxes, used for smFRET test measurements on HMGA1a, called Sample 1 and 2. These peaks are from the shotgun labelling of fraction "26 min.".*

4.4 FRET results

Below are the smFRET results of "Sample 1" and "Sample 2" from the labelled fraction 26 min. These results include a smFRET time trace ***(Figure 22)***, a FRET efficiency histogram and a Pulse Interleaved Excitation (PIE) stoichiometry graph ***(Figure 23)***.

A stoichiometry ratio of around 0.5 tells us that acceptor molecules are almost 50% of the population of donor/acceptor molecules, while a FRET efficiency of 1.0 for that ratio can mean that almost all excited donor molecules are exciting almost all of the acceptor molecules. However, the stoichiometry of 1.0 with 0 efficiency might be due to a large population of donor-only labelled samples ***(Figure 23)***. Since the histogram is from a single sample, it is very unlikely that we observe different populations (although not entirely impossible).

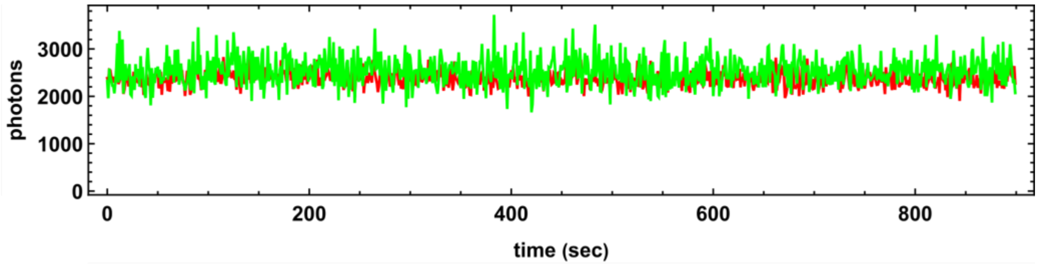

Figure 22: smFRET time trace for "Sample 1". *It shows how many donor (green) photons and acceptor (red) photons the microscope detects during the whole measurement, here 15 min. The y-axis shows the photon count, and the x-axis shows the measuring time in seconds.*

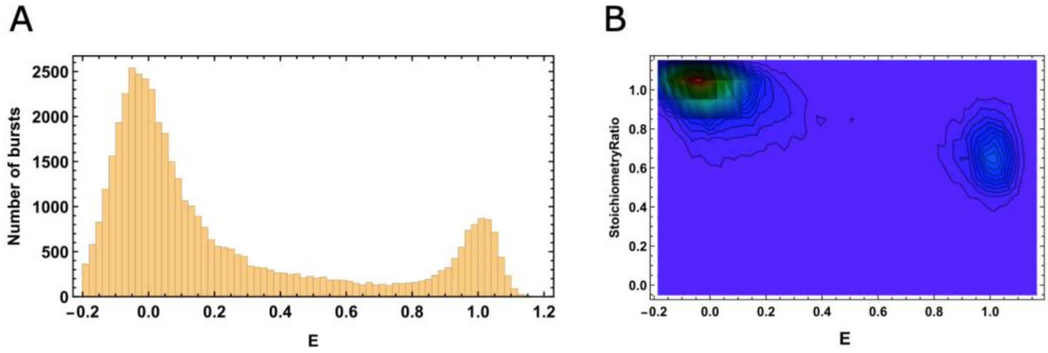

__Figure 23: smFRET results from "Sample 1". A)__ Shows a FRET efficiency histogram. FRET efficiency is determined by the amount of donor molecules that have excited the acceptor molecules[57]. The y-axis shows the number of photon bursts while the x-axis shows the E value (smFRET efficiency). __B)__ Shows a PIE stoichiometry graph, it shows the ratio of acceptor molecules compared to the total of donor and acceptor molecules. The y-axis is the stoichiometry ratio while the x-axis shows the E value.

Down below are the smFRET results of "Sample 2" from the labelled fraction 26-28 min. These results also include a smFRET time trace ***(Figure 24)***, a FRET efficiency histogram and a PIE stoichiometry graph. A stoichiometry of 1.0 with a FRET efficiency of zero tells us that this was most likely a single labelled population, only with the donor, Cy3b ***(Figure 25)***.

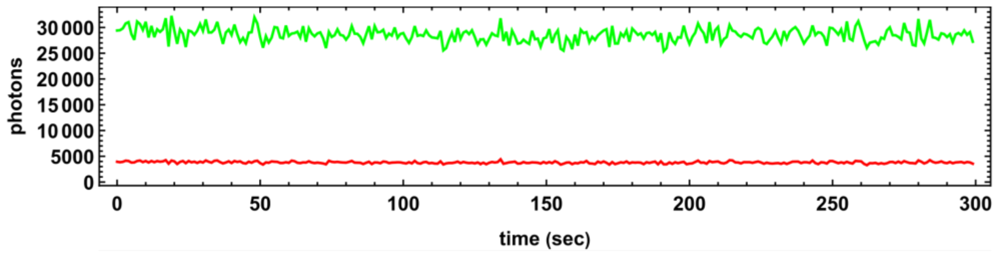

__Figure 24: smFRET time trace for "Sample 2".__ It shows how many donor (green) photons and acceptor (red) photons the microscope detects during the whole measurement, here 5 min. The y-axis shows the photon count, and the x-axis shows the measuring time in seconds.

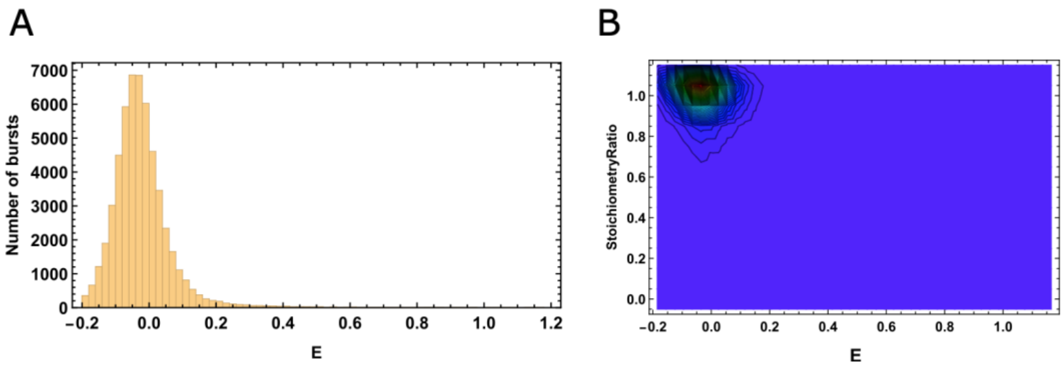

__Figure 25: smFRET results from "Sample 2". A)__ Shows a FRET efficiency histogram. FRET efficiency is determined by the amount of donor molecules that have excited the acceptor molecules[57]. The y-axis shows the number of photon bursts while the x-axis shows the E value (smFRET efficiency). __B)__ Shows a PIE stoichiometry graph, it shows the ratio of acceptor molecules compared to the total of donor and acceptor molecules. The y-axis is the stoichiometry ratio while the x-axis shows the E value.

Further experiments and tests on the microscope are needed to confirm for sure that we have a double-labelled HMGA1a protein, with both a donor and an acceptor.

5 Discussion

During this project we managed to both successfully express and purify the intrinsically disordered TF HMGA1a, and attempted to label it with fluorescent donor and acceptor dyes.

5.1 The successful expression and purification of HMGA1a

We were able to express the protein in the first attempt using BL21 pLysS cells as an expression host. Before the start of this project, my supervisor had successfully expressed the wild-type HMGA1a. We believed that the same expression protocol would work for all HMGA proteins, thus we tried that for our protein. Since the SUMO-tagged protein had an affinity for the nickel column, and we were able to digest the protein with the Ulp1 protease, it showed us that we managed to express the protein successfully. We were able to confirm this with the SDS-page gels we ran.

From the purification process, we were able to retrieve a good amount of purified protein for labelling. We used the same purification protocol as for the wild-type HMGA1a, which had previously been established. Only minor modifications were made to adjust it for our mutant. The most troublesome part during the purification/labelling process was when we had attempted to label the protein and needed to do one last RP-HPLC run, to make some final refinements. The elution peaks from that RP-HPLC run ***(Figure 19 and 20)*** were hard to read and the graph was chaotic, but we had no idea which peak corresponded to a hopefully double-labelled protein. That is why we collected such a large amount of the elution for the smFRET measurements.

5.2 The labelling process of HMGA1a

The labelling part of the process did not turn out to be as successful as we had hoped for, since we were unable to get a double-labelled protein. We came to this conclusion after we had put our protein samples under the microscope and performed several FRET measurements. The results from two peaks, which we called Sample 1 and 2 were presented in the previous chapter. These were elution peaks from the labelled fraction “26 min.” which we considered to be the most promising for labelling, based on the ratio between the absorbances at each wavelength. Unfortunately, the smFRET results revealed that more experiments are needed for the labelling of this protein, since our protocol didn’t work and we didn’t get a correctly double-labelled protein. For Sample 1 we got a stoichiometry ratio of 1.0 with a FRET efficiency of 0, which most likely means that we only managed to get a large population of donor-only labelled samples. In addition, we also received a stoichiometry ratio of ~0.5 against 1.0 FRET efficiency for Sample 1. Theoretically, a ratio of ~0.5 would be a good result for us but since the E value is 1.0, it isn’t. This ratio tells us that the acceptor molecules are most likely almost 50% of the donor/acceptor molecules population, and when the E value is 1.0 for that ratio, it can mean that almost all excited donors are exciting almost all the acceptors. In our case, this is implausible because the dyes are positioned at cysteine residues that are 65 a.a. apart. Since the FRET efficiency increases with decreasing intermolecular distance[57], this distance between the dyes is too long for them to give an E value of 1.0. For Sample 2 we got a stoichiometry ratio of 1.0 against a FRET efficiency of 0, which most likely means that again, this was a single labelled population, only with the donor, Cy3b.

6 Conclusions

The original aim of this project can be split into two parts. The first one was to express, purify and label a double-cysteine mutant HMGA1a with a donor and an acceptor fluorescent dye. The second part was to study the labelled HMGA1a with the smFRET approach, to monitor shifts in intra-molecular distances, in complexes with reconstituted nucleosomes or DNA.

The first part of the project had an overall positive result. The expression and purification of the protein were successful, but we were able to get a decent amount of purified protein, plenty enough for the labelling. The labelling part did not go as planned, but further experiments and tests on the microscope are needed to confirm for sure that we have a double-labelled protein with both a donor and an acceptor. One idea to solve the labelling problem could be using a labelling buffer that has an extreme basic pH. The pI of the HMGA1a protein is high (~9.22), thus using a labelling buffer with a pH above 8 would perhaps help to provide the optimum charges in the protein for the labelling. Hopefully, when the protein labelling turns out to be successful, it will be exciting to continue with this project to further investigate the activity of HMGA1a. As previously mentioned, the dream would be to characterize its potential interactions between DNA or histone proteins, as well as the subsequent modulation of the nucleosome structure, by using the smFRET method.

In a wider context, outside the scope of this research, studies have shown that the overexpression of HMGA1 proteins, along with other HMG proteins, can be linked to multiple cancer diseases found in humans. Therefore it´s clear that more research on this protein family will only be beneficial and useful. At least for now, our initial experiments have provided a successful protocol for the expression and purification of the intrinsically disordered transcription factor HMGA1a.

References

1 Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biology* **1**, reviews001.001 (2000).

2 Yesudhas, D., Batool, M., Anwar, M. A., Panneerselvam, S. & Choi, S. Proteins Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors. *Genes (Basel)* **8** (2017).

3 Zheng, Z. & Wang, Y. DNA binding proteins: outline of functional classification. *Biomol Concepts* **2**, 293-303 (2011).

4 Cherstvy, A. G. Positively Charged Residues in DNA-Binding Domains of Structural Proteins Follow Sequence-specific Positions of DNA Phosphate Groups. *The Journal of Physical Chemistry B* **113**, 4242-4247 (2009).

5 Már, M., Nitsenko, K. & Heidarsson, P. O. Multifunctional Intrinsically Disordered Regions in Transcription Factors. *Chemistry – A European Journal* **29**, e202203369 (2023).

6 Adcock, I. M. & Caramori, G. in *Asthma and COPD (Second Edition)* (eds Peter J. Barnes, Jeffrey M. Drazen, Stephen I. Rennard, & Neil C. Thomson) 373-380 (Academic Press, 2009).

7 Sammak, S. & Zinzalla, G. Targeting protein–protein interactions (PPIs) of transcription factors: Challenges of intrinsically disordered proteins (IDPs) and regions (IDRs). *Progress in Biophysics and Molecular Biology* **119**, 41-46 (2015).

8 Minezaki, Y., Homma, K., Kinjo, A. R. & Nishikawa, K. Human Transcription Factors Contain a High Fraction of Intrinsically Disordered Regions Essential for Transcriptional Regulation. *Journal of Molecular Biology* **359**, 1137-1149 (2006).

9 Tsafou, K., Tiwari, P. B., Forman-Kay, J. D., Metallo, S. J. & Toretsky, J. A. Targeting Intrinsically Disordered Transcription Factors: Changing the Paradigm. *J Mol Biol* **430**, 2321-2341 (2018).

10 Hibino, E. & Hoshino, M. A novel mode of interaction between intrinsically disordered proteins. *Biophys Physicobiol* **17**, 86-93 (2020).

11 DeForte, S. & Uversky, V. N. Order, Disorder, and Everything in Between. *Molecules* **21** (2016).

12 Hosoya, Y. & Ohkanda, J. Intrinsically Disordered Proteins as Regulators of Transient Biological Processes and as Untapped Drug Targets. *Molecules* **26** (2021).

13 Frege, T. & Uversky, V. N. Intrinsically disordered proteins in the nucleus of human cells. *Biochem Biophys Rep* **1**, 33-51 (2015).

14 Bjarnason, S., Ruidiaz, S. F., McIvor, J., Mercadante, D. & Heidarsson, P. O. Protein intrinsic disorder on a dynamic nucleosomal landscape. *Prog Mol Biol Transl Sci* **183**, 295-354 (2021).

15 Habchi, J., Tompa, P., Longhi, S. & Uversky, V. N. Introducing protein intrinsic disorder. *Chem Rev* **114**, 6561-6588 (2014).

16 van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews* **114**, 6589-6631 (2014).

17 Musselman, C. A. & Kutateladze, T. G. Characterization of functional disordered regions within chromatin-associated proteins. *iScience* **24**, 102070 (2021).

18 Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* **16**, 18-29 (2015).

19 Mitić, N. S., Malkov, S. N., Kovačević, J. J., Pavlović-Lažetić, G. M. & Beljanski, M. V. Structural disorder of plasmid-encoded proteins in Bacteria and Archaea. *BMC Bioinformatics* **19**, 158 (2018).
20 Liu, Z. & Huang, Y. Advantages of proteins being disordered. *Protein Sci* **23**, 539-550 (2014).
21 Jeffery, C. J. Protein moonlighting: what is it, and why is it important? *Philos Trans R Soc Lond B Biol Sci* **373** (2018).
22 Bah, A. & Forman-Kay, J. Modulation of Intrinsically Disordered Protein Function by Post-Translational Modifications. *Journal of Biological Chemistry* **291**, jbc.R115.695056 (2016).
23 Darling, A. L. & Uversky, V. N. Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter. *Front Genet* **9**, 158 (2018).
24 Newcombe, E. A., Delaforge, E., Hartmann-Petersen, R., Skriver, K. & Kragelund, B. B. How phosphorylation impacts intrinsically disordered proteins and their function. *Essays Biochem* **66**, 901-913 (2022).
25 Midic, U., Oldfield, C., Dunker, A., Obradovic, Z. & Uversky, V. Protein disorder in the human diseasome: Unfoldomics of human genetic diseases. *BMC genomics* **10 Suppl 1**, S12 (2009).
26 Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* **37**, 215-246 (2008).
27 Mavromoustakos, T. *et al.* Strategies in the rational drug design. *Curr Med Chem* **18**, 2517-2530 (2011).
28 Reeves, R. Nuclear functions of the HMG proteins. *Biochim Biophys Acta* **1799**, 3-14 (2010).
29 Cleynen, I. & Van de Ven, W. J. The HMGA proteins: a myriad of functions (Review). *Int J Oncol* **32**, 289-305 (2008).
30 Ozturk, N., Singh, I., Mehta, A., Braun, T. & Barreto, G. HMGA proteins as modulators of chromatin structure during transcriptional activation. *Frontiers in Cell and Developmental Biology* **2** (2014).
31 Zhang, Q. & Wang, Y. HMG modifications and nuclear function. *Biochim Biophys Acta* **1799**, 28-36 (2010).
32 Mansoori, B. *et al.* HMGA2 as a Critical Regulator in Cancer Development. *Genes (Basel)* **12** (2021).
33 Aiello, A. *et al.* HMGA1 protein is a positive regulator of the insulin-like growth factor-I receptor gene. *Eur J Cancer* **46**, 1919-1926 (2010).
34 Watanabe, M. *et al.* Characterization of the Stoichiometry of HMGA1/DNA Complexes. *Open Biochem J* **7**, 73-81 (2013).
35 Li, H. D., Menon, R., Omenn, G. S. & Guan, Y. Revisiting the identification of canonical splice isoforms through integration of functional genomics and proteomics evidence. *Proteomics* **14**, 2709-2718 (2014).
36 The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 51:D523–D531 (2023). (2023).
37 Wei, T. *et al.* Phosphorylation-regulated HMGA1a-P53 interaction unveils the function of HMGA1a acidic tail phosphorylations via synthetic proteins. *Cell Chem Biol* **28**, 722-732.e728 (2021).
38 Hergeth, S. P. & Schneider, R. The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO Rep* **16**, 1439-1453 (2015).
39 Gordon, J. A. R. *et al.* in *Brenner's Encyclopedia of Genetics (Second Edition)* (eds Stanley Maloy & Kelly Hughes) 538-541 (Academic Press, 2013).
40 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).
41 Reeves, R. Molecular biology of HMGA proteins: hubs of nuclear function. *Gene* **277**, 63-81 (2001).

42 LeBlanc, S. J., Kulkarni, P. & Weninger, K. R. Single Molecule FRET: A Powerful Tool to Study Intrinsically Disordered Proteins. *Biomolecules* **8** (2018).
43 Metskas, L. A. & Rhoades, E. Single-Molecule FRET of Intrinsically Disordered Proteins. *Annu Rev Phys Chem* **71**, 391-414 (2020).
44 Na, J. H., Lee, W. K. & Yu, Y. G. How Do We Study the Dynamic Structure of Unstructured Proteins: A Case Study on Nopp140 as an Example of a Large, Intrinsically Disordered Protein. *Int J Mol Sci* **19** (2018).
45 Bjarnason, S., Ruidiaz, S. F., McIvor, J., Mercadante, D. & Heidarsson, P. O. in *Progress in Molecular Biology and Translational Science* Vol. 183 (ed Vladimir N. Uversky) 295-354 (Academic Press, 2021).
46 Hellenkamp, B. *et al.* Precision and accuracy of single-molecule FRET measurements—a multi-laboratory benchmark study. *Nature Methods* **15**, 669-676 (2018).
47 Götz, M., Wortmann, P., Schmid, S. & Hugel, T. in *Methods in Enzymology* Vol. 581 (eds Maria Spies & Yann R. Chemla) 487-516 (Academic Press, 2016).
48 Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat Methods* **5**, 507-516 (2008).
49 Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61-66 (2018).
50 Asher, W. B. *et al.* GPCR-mediated β-arrestin activation deconvoluted with single-molecule precision. *Cell* **185**, 1661-1675.e1616 (2022).
51 Asher, W. B. *et al.* Single-molecule FRET imaging of GPCR dimers in living cells. *Nature Methods* **18**, 397-405 (2021).
52 Sasmal, D. K., Pulido, L. E., Kasal, S. & Huang, J. Single-molecule fluorescence resonance energy transfer in molecular biology. *Nanoscale* **8**, 19928-19944 (2016).
53 Paudyal, N., Bhatia, N. K. & Jayaraman, V. in *Methods in Enzymology* Vol. 652 (eds Daniel L. Minor & Henry M. Colecraft) 193-212 (Academic Press, 2021).
54 Zosel, F., Holla, A. & Schuler, B. Labeling of Proteins for Single-Molecule Fluorescence Spectroscopy. *Methods Mol Biol* **2376**, 207-233 (2022).
55 Shekhawat, D. & Kirthivasan, N. in *Encyclopedia of Separation Science* (ed Ian D. Wilson) 2732-2739 (Academic Press, 2000).
56 Yan, Y., Orcutt, S. & Strickler, J. The use of SUMO as a fusion system for protein expression and purification. *Chem Today* **27** (2009).
57 Chen, H., Puhl, H. L., 3rd, Koushik, S. V., Vogel, S. S. & Ikeda, S. R. Measurement of FRET efficiency and ratio of donor to acceptor concentration in living cells. *Biophys J* **91**, L39-41 (2006).