Transcription Factor DNA Binding- and Nucleosome Formation Energies determined by High Performance Fluorescence Anisotropy

München 2020

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Parmazie
der Ludwig–Maximilians–Universität München

Transcription Factor DNA Binding- and Nucleosome Formation Energies determined by High Performance Fluorescence Anisotropy

Max Schnepf
aus Karlsruhe
31.01.2020

Erklärung:
Diese Dissertation wurde im Sinne von 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Roland Beckmann betreut.

Eidesstattliche Versicherung:
Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 26.06.2020

Max Schnepf

Dissertation eingereicht am 31.01.2020
1. Gutachter: Prof. Dr. Roland Beckmann
2. Gutachter: Prof. Dr. Nicolas Gompel
Mündliche Prüfung am: 22.05.2020

Contents

List of Figures

List of Tables

Publication list

Jung, C., Bandilla, P., von Reutern, M., **Schnepf, M.**, Rieder, S., Unnerstall, U. and Gaul, U. (2018). True equilibrium measurement of transcription factor-dna binding affinities using automated polarization microscopy, *Nature Communications* 9(1): 1605.

Jung, C., **Schnepf, M.**, Bandilla, P., Unnerstall, U. and Gaul, U. (2019). High sensitivity measurement of transcription factor-dna binding affinities by competitive titration using fluorescence microscopy., *Journal of visualized experiments : JoVE* .

Schnepf, M., Ludwig, C., Bandilla, P., Ceolin, S., Unnerstall, U., Jung, C. and Gaul, U. (2020). Sensitive automated measurement of histone-dna affinities in nucleosomes, *iScience* p. 100824. URL: http://www.sciencedirect.com/science/article/pii/S2589004220300079

Schnepf, M., v. Reutern,M., Ludwig,C., Unnerstall,U., Jung, C.* and Gaul,U., Non-linear interaction measurements and DNA shape readout analysis of transcription factors binding *(in preparation)*

Summary

Protein DNA binding is the core of transcriptional regulation, the process which controls the flow of information stored in an organism's genome to react to its environment and to maintain its functionality. The initial event of gene expression is the binding of a transcription factor (TF) to its target site. These binding events are integrated over several binding sites and TFs by which a fine tuned regulation can be achieved. The number, combination and strengths of the different binding sites encode the desired gene expression level and the plasticity of the regulated gene.

Efforts have been devoted with the goal of identifying the specific DNA sequences bound by different TFs. For more than two decades, it was thought that mutations at each position in this sequence independently contribute to the binding probability of a TF. This binding preference has therefore been described through position weight matrices (PWMs). PWMs describe the binding preference of a TF towards its target sites by assuming that each nucleotide position contributes independently to the total specificity (linearity assumption). However, current research has shown that this simplified view lacks a significant part of the information needed to precisely describe the binding preference of a TF. It was also shown that the most information missing in the PWM is encoded in dinucleotide mutations. Two questions are important in this regard: (1) Which information about TF-DNA interaction are we missing and are currently employed methods able to provide them? and (2) What is a comprehensive description of non-linearity that is based on biophysical properties rather then on abstract probabilities?

One important aspect is the three dimensional configuration of the DNA strand (DNA shape) which is known to affect TF binding to a varying degree. Through recent work by the group of Remo Rohs it is possible to predict shape parameters (features) from a DNA sequence and investigate to which degree they influence binding for any given set of measurements. The first aim of this thesis is therefore to determine non-linearity in TF-DNA interaction and investigate the influence of DNA shape on them.

Protein-DNA interactions were studied with a variety of methods using structural biology (NMR, crystallography, cryo EM) or quantitative Methods (EMSA, DNA binding arrays, ChIP-Seq, B1H, SELEX, MITOMI, Simile-Seq). Most of these quantitative methods to measure TF-DNA interactions, however, are not very sensitive to weak binders due to stringent washing steps or cut-offs they employ. Especially sequences with two positions differing from the consensus can be very weakly bound - therefore a sensitive method is needed to investigate non-linearity. The method called High Performance Fluorescence Anisotropy (HiP-FA, recently developed in our lab) provides the necessary sensitivity. Using HiP-FA, I determined the affinities of 13 TFs from the *Drosophila melanogaster* segmentation network and found most of them to contain a significant non-linearity in their specificity. The binding energies of the TFs correlated significantly with certain DNA shape features suggesting shape readout by the TFs. These results could be confirmed in existing structural biology data.

Besides the influence of information directly encoded in the DNA sequence, the binding of a TF in the genome is most influenced by the DNA accessibility. This property is a result of the genomic DNA being wrapped around histone octamers forming nucleosomes. Since the underlying sequence can also influence the binding of the histone complex to the DNA, a natural question to ask is which features of the DNA sequence are the major determinant of histone-DNA interaction.

Attempts to address this question used existing methods which were either MNase based and are therefore prone to the enzymes intrinsic cutting bias or based on dialysis and/or EMSA readout and have in consequence a low throughput and can only be automated to a small degree. This leads to a limited set of measurements which are usually only based on a single measurement point instead of a complete titration curve. The second aim of my thesis is therefore to develop an *in vitro* assay to determine free energies of nucleosome formation which improves on the limitations of existing methods.

Using the sensitive FA-microscopy setup, I developed an automated assay to determine the free energy of nucleosome formation in a competitive titration. In contrast to existing methods, the throughput of the assays allows for full competitor titration curves. By measuring the free binding energies of 42 sequences, I showed that GC-content is the factor most contributing to the free energy. The relationship between these quantities is non-monotonous with an optimal GC-content of 49 percent.

The results provided in this thesis give insight into the nature of non-linearity in TF-DNA interactions and highlight the DNA shape readout therein. Methodical advancements developed in this work can be used as a foundation to investigate other kinds of molecular interactions making use of the high sensitivity of FA-based microscopy.

Abbreviations

Abbreviation	meaning
bp	base pairs (length of DNA)
dH_2O	distilled water
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EGTA	ethylene glycol-bis(β-aminoethyl ether)-N,N,N,N-tetraacetic acid
EMSA	electrophoretic mobility shift assay
g	Gravitational acceleration, multiples of $9.80665\frac{m}{s^2}$
FA	Fluorescence anisotropy
FRET	Förster Resonance Energy Transfer
HiP-FA	High Performance Fluorescence Anisotropy (method name)
hydrogen bond	H-bonds
IC	Information content
K_D	Dissociation constant
M	molar [mole per liter]
MGW	Minor groove width (of DNA)
MITOMI	mechanically induced trapping of molecular interactions (method name)
PCR	Polymerase chain reaction
PFM	Position frequency matrix
PWM	Position weight matrix
rpm	rounds per minute
RT	room temperature (approximately 23 °C)
SELEX	Systematic evolution of ligands by exponential enrichment (method name)
SMiLE-Seq	Selective microfluidics-based ligand enrichment followed by sequencing(method name)
TF	transcription factor
V	volts
W	Watts

Chapter 1

Introduction

1.1 General introduction

All cells of a multicellular organism have the same primary genetic information encoded in their genomes. They, however, face different demands depending on their cell type and current state. Those need to be fulfilled by differential expression of the correct genes in the required amounts. The initial event in gene expression is the binding of a protein called transcription factor (TF) to its target site. The integration of several of these binding events determine the expression state of the controlled gene. The prediction of the resulting expression (Segal et al. (2008)) has been the subject of longstanding affords in the field of gene expression research. A major obstacle is the fact that the binding events need to be predicted correctly and precisely (Weirauch et al. (2013)). The reasons why this can be challenging are found on different levels: The first is the incomplete description of the TFs' binding sites.

It is nowadays commonly accepted that the binding site of a TF can't be described by a simple consensus sequence, but that a least a scoring matrix (Stormo et al. (1982)) needs to applied. Although the scoring matrices can be a good approximation of a TF's binding preferences, recent work has shown that the central linearity assumption (the readout of each position is statistically independent of other ones) doesn't hold true for many TFs. The prediction of binding strengths using the product of probabilities fails in these cases and will mostly be least accurate when looking at neighboring double mutations (first higher order). Siebert and Soding (2016) showed that most information of interdependencies between certain positions missing in the PWM is found in dinucleotide mutations in contrast to 3mer, 4mer or higher order mutations. It is therefore desirable to determine affinities between TFs and systematic dinicleotide mutations with high sensitivity to gain more insights into the degree non-linearity plays in TF-DNA interactions. This high sensitivity is essential, because double mutations are often significantly weaker than single mutations, which challenge many existing methods. By neglecting the influence of weak bindinders, they seem to overestimate the specificity of TFs. This is in particular true for Systematic evolution of ligands by exponential enrichment (SELEX) (Jolma et al. (2013)) or bacterial one hybrid (B1H) (Meng et al. (2005)) which contain a stringent selection step (Rastogi et al. (2018)) in their protocols.

The new method High performance Fluorescence Anisotropy (HiP-FA), recently developed in our group, is more sensitive to weak binders and was shown to perform better in both binding site and expression prediction (Jung et al. (2018)). To gain better mechanistic insights into the binding events it is of paramount importance to determine binding preferences of TFs accurately and in consequence also investigate their non-linearity.

An important feature often disregarded in models describing TF-DNA binding is the fact that TFs don't exclusively read out the DNA base sequence but also its three dimensional shape (Rohs et al. (2010)). Although known from structural biology studies for specific examples, this feature can be generally investigated since a simple and computationally inexpensive algorithm to determine

DNA shape from DNA sequence has been made available (Zhou et al. (2013)).

Besides the direct interaction of TFs and free DNA it is important to take the DNA-accessibility of TF binding sites into account. Only binding sites that are not hidden in a nucleosome complex can be bound and therefore this feature is important to consider when predicting binding sites in a genome. Although the positioning and binding strength of nucleosomes are influenced by many biological factors, it is of huge importance to know the contribution of the underlying sequence on nucleosome binding and thereby indirectly on TF binding behavior.

Existing methods either suffer from limitations like low throughput, low degree of automation and questionable accuracy and robustness (owed to the use of single measurement points instead of full competitor titration curves due the low throughput) - for salt titrations or or dialysis based methods or are limited by their design to include information from neighboring sequences, can be compromised by enzymes' DNA bias and are rather indirect when assigning sequencing reads - MNase-Seq, ATAC-Seq or other sequencing based methods.

To this end, the second major goal of this work is to develop an assay that can determine the binding strength of a defined sequence to a histone complex. This nucleosome formation assay should be easy to automate, have a higher throughput than existing method and should be sufficiently robust to draw conclusions about the influence of the DNA's general sequence features reported in literature to influence histone- DNA interactions like GC content (Tillo and Hughes (2009)), 10 bp periodicity of dinucleotides (Shrader and Crothers (1990)) and the presence of poly A:T stretches (Segal and Widom (2009a)).

1.2 Transcription factor binding

The expression of genes in all cells needs to be tightly regulated. The initial event to express any given gene in a genome is the binding of a TF -or a set of TFs. A TF is a protein that recognizes DNA and mediates transcription.

TFs can be divided into two groups based on their function. The first group is the one of basal TFs mediating the recruitment of the transcription machinery at the transcription start site. The second group of TFs bind to enhancers and thereby modulate the strength of one or several target genes. TFs investigated in this thesis are all part of the second group. Generally, TFs use chemical interactions of three types to bind DNA: Salt bridges, which are manly unspecific, hydrogen bonds and induced dipoles (Van-der-Waal forces). These are used to determine a binding site based on two different criteria to varying degrees: the correct sequence of the DNA's nucleobases ("base readout") and the spacial conformation of the DNA molecule -the DNA shape("shape readout"). To which degree this can be read out from binding data is an important part of this work. Subtle changes in DNA binding preference of TFs upon variation in DNA shape might be one, but are probably not the only reason, why weak binding sites are important in gene expression. The functionality of a TF has been invented several times during evolution (de Mendoza and Seb-Pedrs (2019)), which is why there are several families of TFs which all possess different modes of action when binding to their target sites. When describing the specificity of a TF, a PWM is generally a good approximation. Several studies have, however, shown that especially neighboring dinucleotide mutations still alter the binding strength significantly compared to the PWM's linearly expected binding weight. This work will therefore expand the existing protocols of HiP-FA to include systematic dinucleotide mutations and evaluate the influence of non-linearity.

1.2.1 Transcription factor families

TFs, like most other proteins consist of several domains. A domain all TFs have is a DNA binding domain. Based on the type of domain, TFs are classified into different families, sharing the same mode of DNA binding. In this work I investigated members of 7 different families. This section

explains those families occurring in the experimental work and being discussed in more extensively during the discussion in more detail, and mentions the remaining ones.

Zinc fingers Zinc finger proteins are the largest family of DNA binding proteins and also the most diverse. Some members bind RNA or other proteins instead or in addition to DNA. Their structure, as the name suggests, contains a zinc cation as central metal ion in a complex with cysteine and histidine residues. The most important sub-type of TFs are Cys2His2 (C2H2) zinc fingers (see also Fig.1.1). The second part of their name "finger" originates from the domain wrapping around the DNA contacting three to four bases (Choo and Klug (1994)), five in exceptional cases (Pavletich and Pabo (1993)). The wrapping occurs by contacting all but one nucleobase in the major groove of one DNA strand and the remaining one on the opposite strand. In this manner, several fingers can bind the DNA consecutively with always one position overlapping (Razin et al. (2012)). In contrast to other TF families, their DNA recognition motives don't show particular similarities (Kribelbauer et al. (2019)). Work by Najafabadi et al. (2017) showed that metazoan C2H2 TFs are making more contacts to DNA backbone, establishing binding independent of hydrogen bonds (H-bonds) to the nucleobases. This allowed a more diversified evolution, allowing metazoan TFs to almost bind any DNA triplet while other taxa are restricted to a much smaller set of triplets.

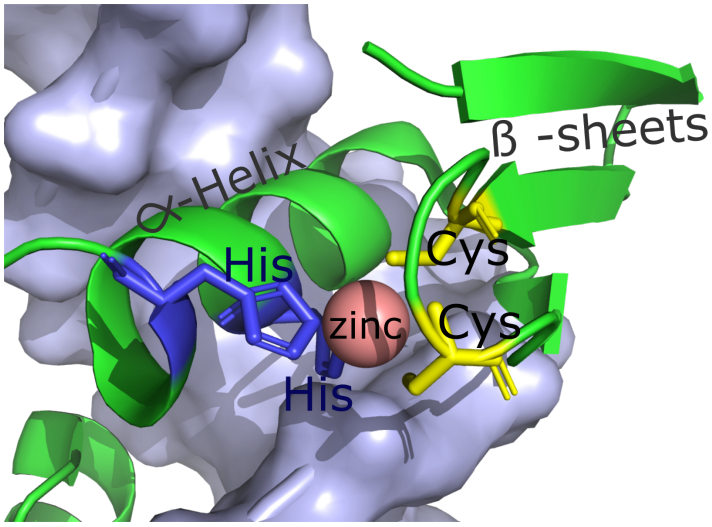


Figure 1.1: The figure shows an example of a C2H2 zinc finger protein. The central zinc ion is in a complex with two histidine (His) and two cysteine (Cys) residues. The general fold consists of two -sheets and an α-helix. (PDB-ID:2drp - Fairall et al. (1993))

Homeodomains Another large group of TFs are homeodomain proteins. This protein family is defined by an element in their genes called homeobox. This approximately 180 bp long sequence element is conserved among many eukaryotic TFs. The protein domain encoded by the homeobox, the homeodomain, consist of three α-helices with an unstructured N-terminal tail (Brglin and Affolter (2016), see also Fig. 1.2). An important subgroup of the homeodomain proteins are the hox proteins - Members of this group are TFs controlling the morphogenesis of animals resulting in a mirror symmetrical body plan (Rezsohazy et al. (2015)). They are highly similar between species up to a degree in which they can drive expression in distant different metazoan species (McGinnis et al. (1990)). In contrast to the very distinct functions these proteins can carry out, their binding specificities are often highly similar. They differ rather in low affinity binding sites, which might be a hint that the specific functions are not achieved by single Hox proteins but rather by cooperative binding (Affolter et al. (2008)). The affinities to these low binding sites might change when the

overall specificity of TF is altered upon hetero-dimerization with a second factor (e.g. a PCB protein, an atpyical member of the homeodomain family) (Rezsohazy et al. (2015))

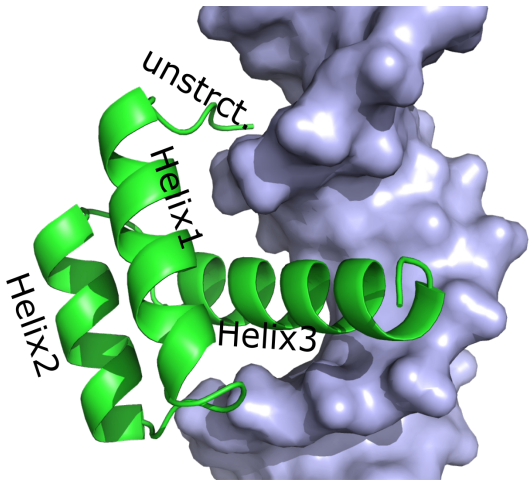


Figure 1.2: The figure shows an example of a homeodomain protein. The general fold consists of three α-helices, of which helix 3 contacts bases in the major groove, and an unstructured tail contacting the DNA minor groove. (PDB-ID:9ant -Fraenkel and Pabo (1998))

POU domains POU domains are closely related to homedomains, as they share a homedomain and in addition have a POU specific domain which alters the typical binding preference of homeodomains (Phillips and Luisi (2000)).

HMG domains Another family investigated in this study are nuclear hormone receptors which beside domains for regulation by the respective hormone have a very conserved domain consisting of two zinc fingers (Kumar and Thompson (1999)). High mobility group (HMG) factors prefer to bind unusual shapes of DNA and can widen the minor groove of their binding site (Štros et al. (2007)).

Helix-Turn-Helix proteins Helix-turn-Helix proteins harbor a domain consisting of two alpha helices binding the DNA major groove which are connected by a short kinked stretch of amino acids (Brennan and Matthews (1989)).

Winged HeliX proteins A subclass of Helix-turn-Helix proteins are members containing the winged-helix DNA binding domain. While their DNA recognition helix contacts the major groove, the "wings" (small beta sheets) make contact to the minor groove or the DNA backbone (Teichmann et al. (2012)).

B-Zip proteins The last family covered in this work are B-zips. With their two long dimerized alpha helices, they approach the DNA in a scissor-like manner, thereby mostly creating palindromic recognition sequences (Hurst (1995)).

1.2.2 DNA readout by Transcription factors

Base readout The most specific interactions between a TF and DNA are hydrogen bonds (Etheve et al. (2016)). In this interaction, a strongly polarized hydrogen atom (from the donor molecule) and a lone electron pair (from the acceptor molecule) substantially overlap (Steiner (2002)). With the

help of these bonds several positions at the nucleobases can be contacted and read out specifically (see 1.3). This readout happens in the major groove of DNA. After salt bridges (formed to the negatively charged phosphates in the DNA backbone), hydrogen bonds are the strongest bonds formed between proteins and DNA. To achieve specificity, the geometry of the hydrogen bonds matters: Depending on the number of H-Bonds, Donor and acceptor molecules, the specificity can change significantly. A single H-Bond for example is not able to provide specificity in general but is strongly dependent on its context to contribute to specificity. More specificity is created in bifurcated hydrogen bonds (two hydrogen bonds with different acceptors but to same hydrogen), while the highest specificity is created by bidentate hydrogen bonds (two hydrogen bonds within the same molecule but with each two different groups as donors and as acceptors) (Coulocheri et al. (2007))

Besides the discussed H-bonds, TFs can also contact bases via hydrophobic interactions. This is used when discriminating pyrimidines. Cytosin and thymin differ in the presence of a methyl group (green circle in Fig 1.3) - this difference can be read out by TFs (Harrison and Aggarwal (1990)).

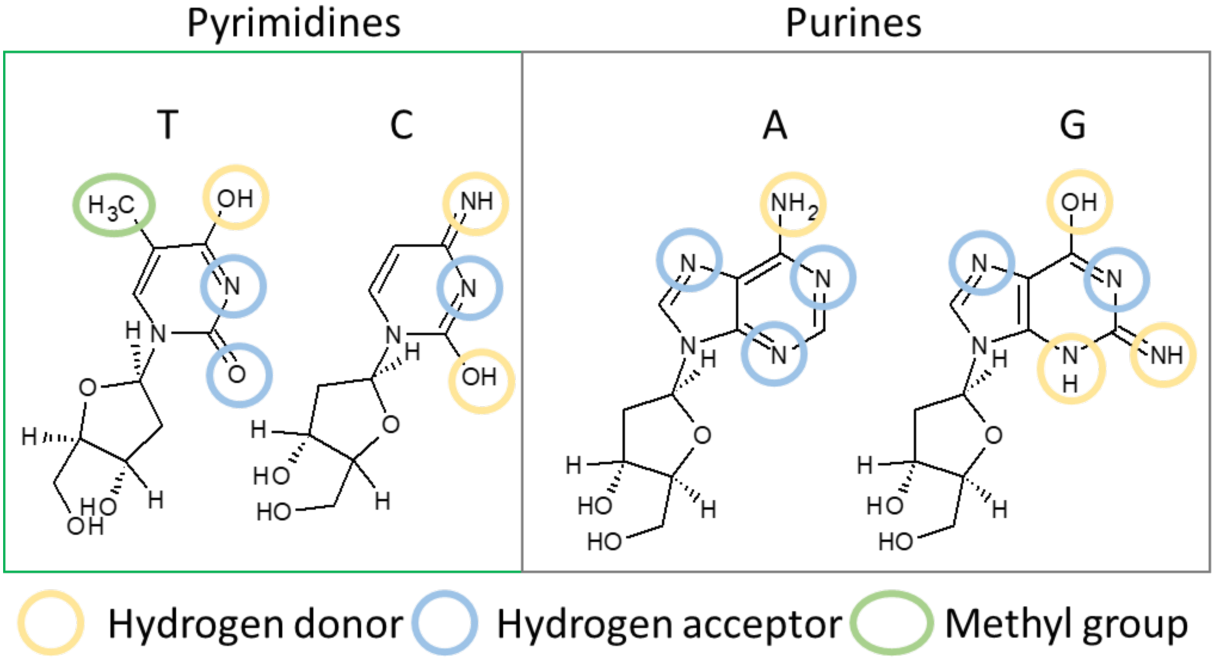


Figure 1.3: Readout of nucleobases. The figure shows the four nucleobases and their possible positions of hydrogen bond donors (yellow circles) and hydrogen acceptors (blue circles) which can be used by DNA binding proteins. The methyl group used to discriminate the pyrimidines from each other is marked in green. The Pyrimidines and Purines are grouped together to allow for a better comparison.

Shape readout The DNA double helix, like all biological macromolecules, can adopt different conformations in its spatial orientation. Like Fig. 1.4 shows, the configuration of DNA in space can be described by twelve parameters Dickerson (1989), six describing the orientation of the two opposing bases in the different DNA strands (intra features) and six describing the orientation of two consecutive base pairs towards each other (inter features). Both groups consist of three angles in which the bases could be turned and three space axes, defining the directions into which the bases could be shifted relative to each other. The possible orientations in combinations with

the three dimensional structure of a potential binding partner can lead to situations in which weak interactions like Van-der-Waals forces can differ as they are dependent on orientations and distances (Garrett and Grisham (2016)). This In addition to these it is known from structural biology that the geometry of the minor groove plays an important role when a TF reads out shape parameters (Rohs et al. (2010)), which is why it is informative to report this as a separate feature when talking about DNA shape geometry. Besides the prominent feature of minor groove width (MGW), it is known that TFs make more contacts to the DNA scanning several of the local features mentioned above as well as broader features like DNA kinks Rimini et al. (1995) , twists or DNA winding (Štros et al. (2007)). The degree to which bases or shape is read out varies from factor to factor (Slattery et al. (2014)).

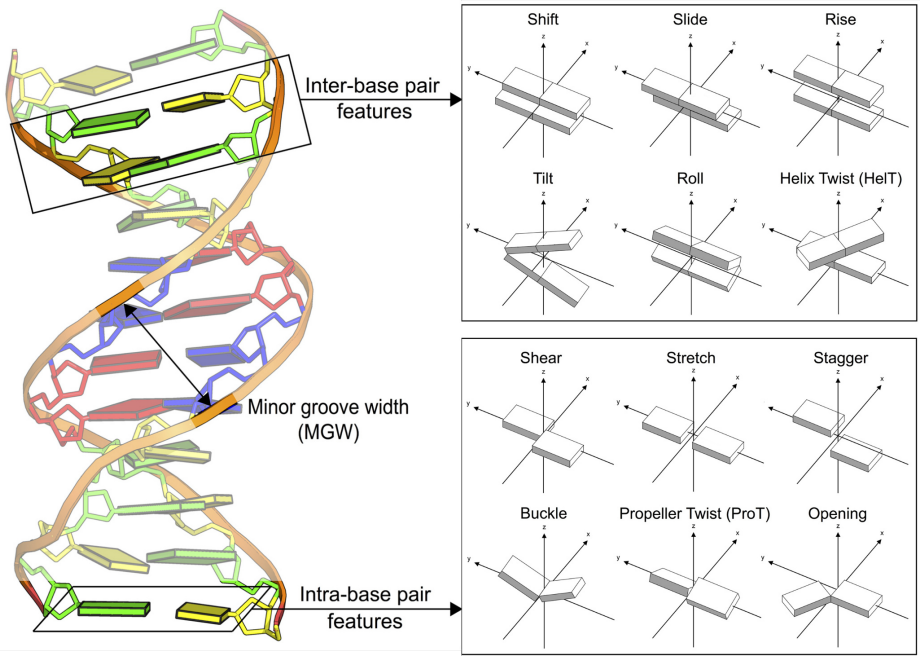


Figure 1.4: Illustration of DNA shape features. The left panel illustrates the difference between inter base pair features(the orientation of two consecutive base pairs) and intra base pair features (the orientation of two bases at opposing DNA strands forming a base pair), as well as the minor groove width. The right panel shows all twelve geometrical features with illustrations, angles and displacements grouped in rows. The figure is modified from Li et al. (2017), permission to reuse granted 05.11.2019.

The influence of low-affinity TF binding sites The example of homeodomains (see also section 1.2.1) shows that many TF binding sites are very similar which implies that their high-affinity binding sites are not sufficient to discriminate their target sites. Possible strategies to circumvent this problem could be, one the one hand, the combined binding of TFs in complexes or the exploitation of weak binding sites in different ways (Kribelbauer et al. (2019)). To understand how weak binding sites could influence the binding it is first of all important to understand how a TF finds its binding site. It uses facilitated diffusion - a combination of three dimensional diffusion and one dimensional sliding on the linear DNA (Li and Elf (2009)). Weak binding sites could therefore help to increase the residence time of a TF near them and thereby increase the local concentration of the TF, in turn increasing the probability of binding to its supposed target site (Ezer et al. (2014)).On the other hand, in addition to this concentrating effects, weak binding sites can also become relevant when the local TF concentration is increased by other means. This is for

example the case in transcriptional hubs originating from the high compartmentalization within the eukaryotic nucleus (Kribelbauer et al. (2019) and references therein). In cases of strongly increased local TF concentrations subtle differences in the binding preferences of TFs can play an important role, explaining how these otherwise very similar binding sites generate specificity.

1.2.3 Mathematical description of binding sites

Position weight matrices and information content Position weight matrices (PWMs) are a way to describe the binding preference of a TF under the assumption of linearity (Stormo et al. (1982)). Linearity means in this context that it is assumed that the overall preferences of a TF do not change if one position deviates from the consensus. It is generated from the a position frequency matrix which counts the occurrences of each base at any given position in the found binding sites:

$$PFM_{k,j} = \sum_{n=i}^{N} I(X_{i,j} == k)\frac{1}{N} \tag{1.1}$$

Where X is an aligned matrix of N sequences of length j. $I(X_{i,j} == k)$ equals 1 if the base at position i,j equals the base defined by k and is zero otherwise. By normalizing with $\frac{1}{N}$ the resulting values are frequencies normalized to a sum of one and can be interpreted as probabilities. A PWM is generated from a PFM by transforming it into log-likelihoods:

$$PWM_{k,j} = log_2(PFM(k,j)/(P_{background})) \tag{1.2}$$

in which $P_background$ is the background probability of the respective base. This value is dependent on the reference system in which the experiment is conducted. If the experiment is, for example, performed in a *Drosophila melanogaster* genome, the average probabilities of this organism are applied - with a GC content of 43%, the occurrence of a G or C is less expected by the background frequency and therefore more informative (higher information). In this work, I will, however, work with even background probabilities, although all TFs presented in this work are originated from *Drosophila melanogaster*. The reason is that all experiments are conducted in an *in vitro* environment - each mutation is therefore as informative as any other because there is no trend like in a genome of some bases occurring with a higher probability. Informativeness is quantified in this regard using information content (IC) (Werner (2008)):

$$IC(x) = 2 - log_2(\frac{1}{p_x}) \tag{1.3}$$

with log_2, the IC is calculated in bits, reflecting two binary choices of the PWM (either purines or pyrimidines, and afterwards which of the two members in the group), which is why the maximal information content can be 2. Despite of the definition of PWM, this work will display PWMs in the form of PFM. The distinction is not always clearly made in literature and the representation of PFMs is more common than the one of PWMs and the nomenclature is often using "PWM" if PFMs are depicted (Weirauch et al. (2013); Nitta et al. (2015); Isakova et al. (2017); Lambert et al. (2018); Jung et al. (2018)). To ensure consistency with previous publications (including those from our lab) I will use the term PWM and depict PFMs. The two representations still contain the same information and can be easily calculated from each other.

Non-linearity and mutual information The assumption of linearity (statistical independence) in PWMs doesn't necessary reflect biophysical reality. Both base readout by H-bonds (by changing their spatial orientation and thereby potentially the length of the bond) as well as shape readout can be highly influenced by mutations in their neighborhood. This is reflected in non-linear models often outperforming strictly linear ones Siebert (2016); Zhao et al. (2012). A possibility to describe this non-linearity between neighboring positions (dinucleotides) is the concept of mutual information,

based on A Kullback-Leibler divergence (Kullback and Leibler (1951))

$$Mutual\ information(X,Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)} \tag{1.4}$$

With X and Y being discrete random variables (the distribution of the two neighboring base positions in this case), p(x,y) the probability of the double mutation and p(x) and p(y) the respective probabilities of the single mutations. Both mutual information and the IC of PWMs are probabilities scaled using a 2 based logarithm to give a number that can easier be displayed in logos (see below)

Sequence logos To generate a less abstract representation of PFMs or PWMs, Schneider and Stephens (1990) developed the depiction via sequence logos. The letters are represented in colors and their height corresponds to their information content (see Figure 1.5).

$$PWM_{example} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.33 & 0.33 & 0.33 & 0.01 \\ 0.125 & 0.01 & 0.74 & 0.125 \\ 0.01 & 0.97 & 0.01 & 0.01 \\ 0.49 & 0.01 & 0.01 & 0.49 \end{pmatrix} \end{array} \tag{1.5}$$

The example PWM given in equation 1.5 leads to the sequence log depicted in Figure 1.5. Position 1 with an even distribution matching the one of the background doesn't have any information content while position 4 which is almost entirely occupied with the letter C has the highest IC.

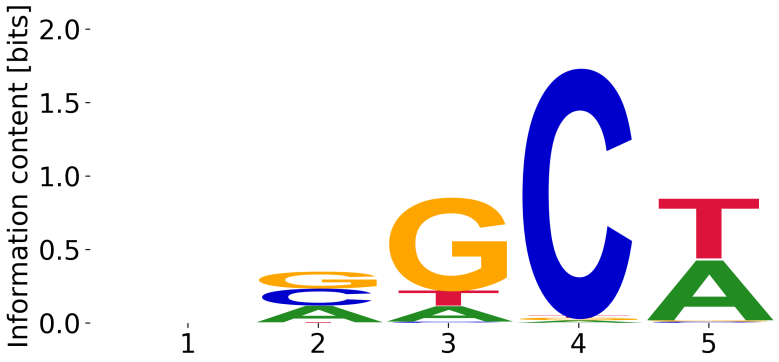


Figure 1.5: Example of a sequence logo illustrating the data given in equation 1.5. The logo shows an example of a position with evenly distributed probabilities for all letters (position 1) and position strongly dominated by one letter -C- (position 4) and other example combinations. The probabilities leading to this logo are given in Equation 1.5.

1.3 Segmentation in *Drosophila melanogaster* embryo development

The embryonic development of *Drosophila melanogaster* is well studied and some of its principles apply to all development of metazoans. The development results in a fully developed animal but the

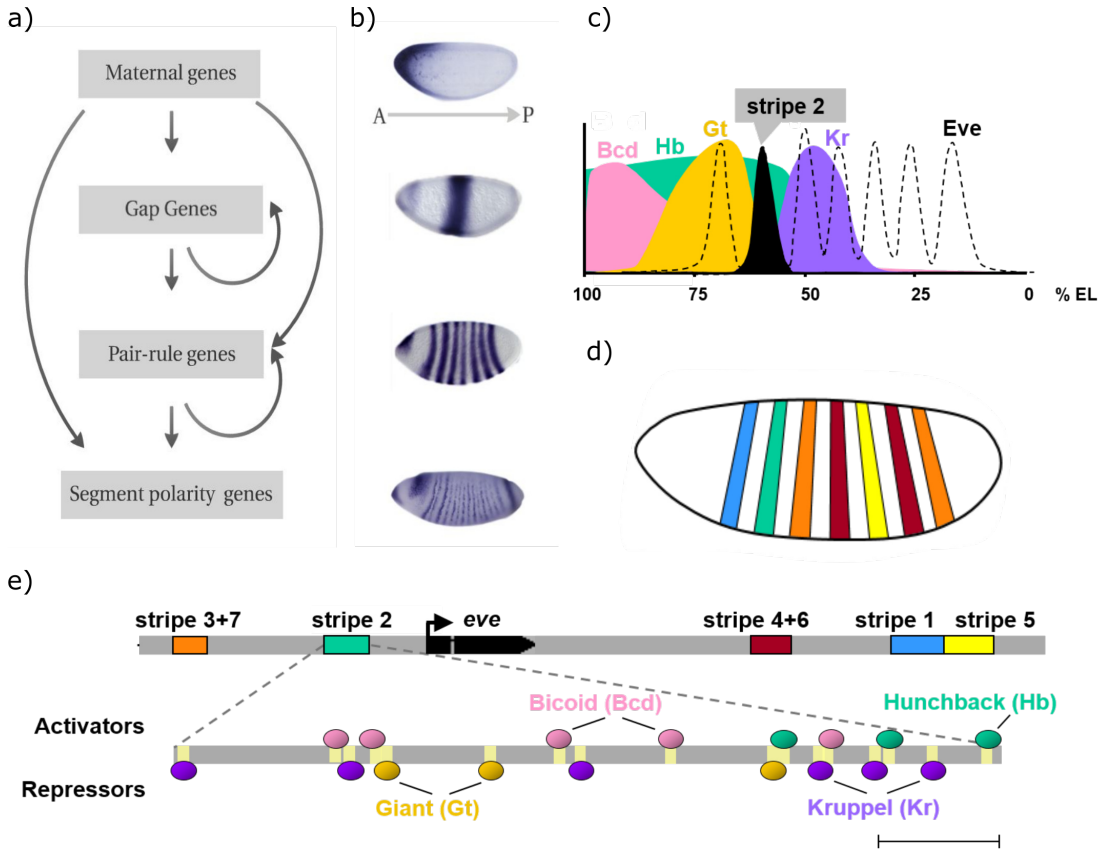


Figure 1.6: Segmentation in *Drosophila melanogaster*. a) Schematic representation of segmentation cascade in *Drosophila* embryos. Arrows symbolize how gene groups influence each other. b) Examples of lac-Z stainings indicating expression of one representative of each group of genes (compare also to a)). c) Expression patterns of different transcription factors over the embryo anterior-posterior axis. d) Colored illustration of the Eve stripes in the embryo, colors matching with e). e) *eve* gene locus with a zoom-in in the enhancer of stripe 2 (turquoise). The zoom-in illustrates the TF binding sites (yellow boxes) and the respective TFs (colored circles). The illustrations in this figure were created by members of the Gaul-lab for presentation purposes.

segmentation paradigm (the process of defining the body segments of the animal) also establishes the three body axes and establishes symmetry and asymmetry in the body plan. The fertilized *Drosphila* egg consist of a syncytium (several nucleii not separated by cell membranes) in which the products of several maternal genes (encoding the maternally provided TFs) are present, either in the from of mRNAs or proteins. Important for the anterior-posterior axis (head to tail, AP-axis) are *nanos, bicoid, hunchback* and *caudal*. The mRNAs encoding the first two TFs are located at opposing sites of the embryo while the latter two are uniformly present. The finer and finer definition of compartments in the egg is established by the interplay of gradients, cascades and the differential expression of TFs and other proteins under the control of the respective enhancers (see also 1.6 a-c). The gradients in maternal mRNA establish more gradual ones when their corresponding TFs influence the expression of each other. The next group of factors are the gap genes. Their enhancers integrate the information of the maternal gradients and the emerging gap gene expression patterns. To ensure the robustness of this process, several different enhancers beside the basic enhancers for each gap gene provide additional information in case of non-ideal conditions which would otherwise perturb the gradients (Perry et al. (2011)). The next finer level of expression

patterns are established by the body plan (body axes) pair-rule genes leading to the characteristic seven stripe pattern in *Drosophila* embryos (Figure 1.6 d). Information about this processes have been gained by screening experiments in which mutations in genes of the segmentation network lead to developmental defects (Nusslein-Volhard et al. (1985)). *in-sito* hybridizations were able to visualize and localize the expression patterns Hafen et al. (1984). Changing the enhancer sequence leads to ectopic expression (Kosman and Small (1997)), showing that the enhancers act in a relative autonomous fashion when reading out the gradients. The expression is thereby encoded in the number, binding strengths and relative positions of the respective binding sites in a given enhancer (see Figure 1.6e).

1.4 High Performance Fluorescence Anisotropy

This section explains the method High Performance Fluorescence Anisotropy (HiP-FA) which was first published by Jung et al. (2018). While this publication focused on insights gained by this method, in a second publication, Jung et al. (2019), we went more into detail on how the method is applied, including a video about it in the online material. The following section will use figures with high identity to some of the latter publication when they were created by me as the second author of this article.

1.4.1 Previously existing methods

This subsection describes alternative methods used to determine TF-DNA binding strengths and / or specificity.

ChIP-Seq Chromatin immunoprecipitation followed by sequencing (ChIPseq, Park (2009)) is a technique that, unlike the other ones mentioned in this section, is also able to determine the binding preferences of a TF *in vivo* in addition to *in vitro*. The method is based on chemically cross-linking TFs with DNA and thereby capturing binding events. The cross-link product is afterwards precipitated using antibodies against the TF. The pulled-down DNA is afterwards sequenced to deduce the TF binding preferences. Besides the possibility of also operating *in vivo* ChIP-seq also works *de novo*. The results need to be backed up with control samples to reduce the influence of fragmentation or sequencing bias. Besides challenges in data analysis (like mapping) ChIP-Seq also relies on functional antibodies to efficiently and specifically pull down the TF of choice. ChIP-Seq data are often regarded as the gold standard a method needs to be compared with (Weirauch et al. (2013)) as it is considered to be the most direct measurement of TF-DNA binding, although the aforementioned shortcomings can influence the data quality.

MITOMI Mechanically induced trapping of molecular interactions (MITOMI,Fordyce et al. (2010)) is micro-fluidics based and is based on the principle of mechanically trapping a mixture of a DNA library and an immobilized TF in the chamber. During a washing step, non-bound DNAs are removed and the remaining bound DNAs are read out via fluorescence intensity using a fluorescent label attached to each sequence. When operating with different concentrations, MITOMI is also able to estimate absolute affinities in addition to the relative ones provided by the measurements based on one concentration per sequence. MITOMI can measure sequences with a medium throughput for a single TF but is limited in the amount of concentrations per sequence it can measure, leading to potentially less accurate affinity measurements. Making use of the amount of sequences measurable, MITOMI can determine binding preferences *de novo*. Weak binders might be lost in the MITOMI washing steps, leading to potentially (over-) specific binding matrices. Although it measures equilibrium binding events, MITOMI needs to immobilize one of the binding partner (the TF), which might influence the obtained results.

SELEX-Seq Systematic evolution of ligands by exponential enrichment combined with sequencing (SELEX-Seq, Riley et al. (2014)) is a method based on the several selection cycles, starting from a random DNA library and increasing sequences preferably bound by the TF of interest during each cycle of enrichment. The bound part of the library is amplified and used as the input for the next round of enrichment. The final readout is the next generation sequencing of the enriched sequences. The method shares several traits with MITOMI, as both methods can operate without prior knowledge about the binding preferences of the investigated TF, both have very high throughput (SELEX-Seq even more than MITOMI), both need to immobilize the TF and both have the risk of loosing weaker binders during their rather stringent washing steps. SELEX-Seq is in addition able to capture multiple binding events on a single DNA Nitta et al. (2015).

SMiLE-Seq Selective microfluidics-based ligand enrichment followed by sequencing (SMiLE-Seq,Isakova et al. (2017)) is a method that tries to combine the high statistical power of SELEX in combination with the the trapping of MITOMI to conserve weak binders. While it combines the strengths of both methods it still produces overly specific matrices in some cases, probably also due to partially loosing weak binders,

HiP-FA in comparison HiP-FA is one of the view methods that can accurately determine the affinity of TFs, especially the one towards weakly bound sequences and it does so without the potential artifacts caused by surface immobilization of one of the binding partners. Its full titration curves allow the investigation of each sequence individually. This eliminates the need for washing or thresholding, and thereby doesn't introduce artificial cut-offs to weak binders which still might be relevant in biological systems. It is therefore suited for questions in which both high affinity and weak binders matter. HiP-FA, however, unlike the competing methods mentioned above, requires prior knowledge about the studied DNA sequences bound by the TF. Its throughput is only moderate and it requires a special microscopy setup.

1.4.2 Fluorescence Anisotropy

To understand HiP-FA it is essential to understand the principle of fluorescence anisotropy (FA) (see also figure 1.7). Molecules in solution have a random rotation due to one degree of freedom of their thermal motion. When a fluorophore is transferred to its excited state it takes the lifetime τ of the fluorophore to return to its relaxed state while emitting a photon. τ is much larger than the rotational relaxation time of a small fluorophore. The emission angles will therefore be distributed randomly ("isotropically"). When observing macromolecules, however, the rotational speed becomes slower so that τ is small enough to determine their polarization (Weber (1952)). The normalized polarization is called anisotropy and can be determined when using a polarized excitation source and when determining the parallel and perpendicular parts of the polarized emission light. This is formalized in equation (1.6) with $I_{||}$ being the parallel and $I_{\perp}$ being the orthogonal part of the light and G the instrument's correction factor.

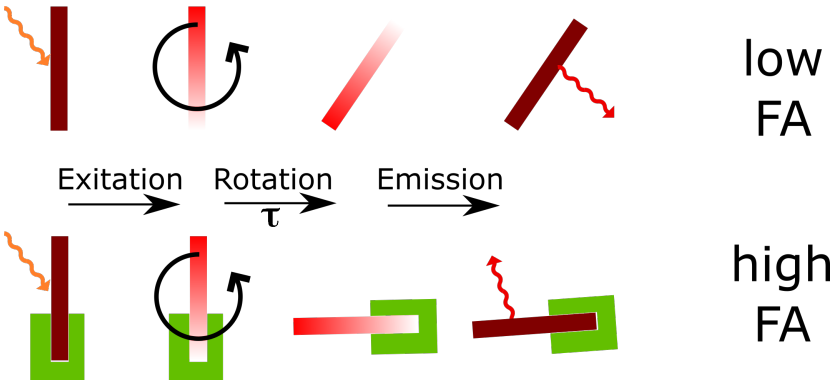


Figure 1.7: The figure shows the principle of FA. A fluorophore gets excited by light. Before the emission after the fluorescence lifetime τ it emits a photon. The direction of the emission is dependent on its rotational speed, being determined by the size of the fluorophore or the complex it is contained in. Fast rotation in a small complex leads to a low FA, a slow rotation in a big complex leads to a higher FA.

$$[htb]FA = \frac{I_{||} - G * I_{\perp}}{I_{||} + 2G * I_{\perp}} \tag{1.6}$$

FA can approximate the hydrodynamic volume of a molecule - its size (Gradinaru et al. (2010)). As the volume of a complex is lager than the individual binding partners it is possible to monitor molecular binding events by their change in FA (see also figure 1.7). We use this to monitor the displacement of a fluorescently labeled DNA oligomer by non-fluorescent competitors and therefore determine specificity and affinity of the protein to the competitor sequence (section 1.4.3).

1.4.3 HIP-FA principle

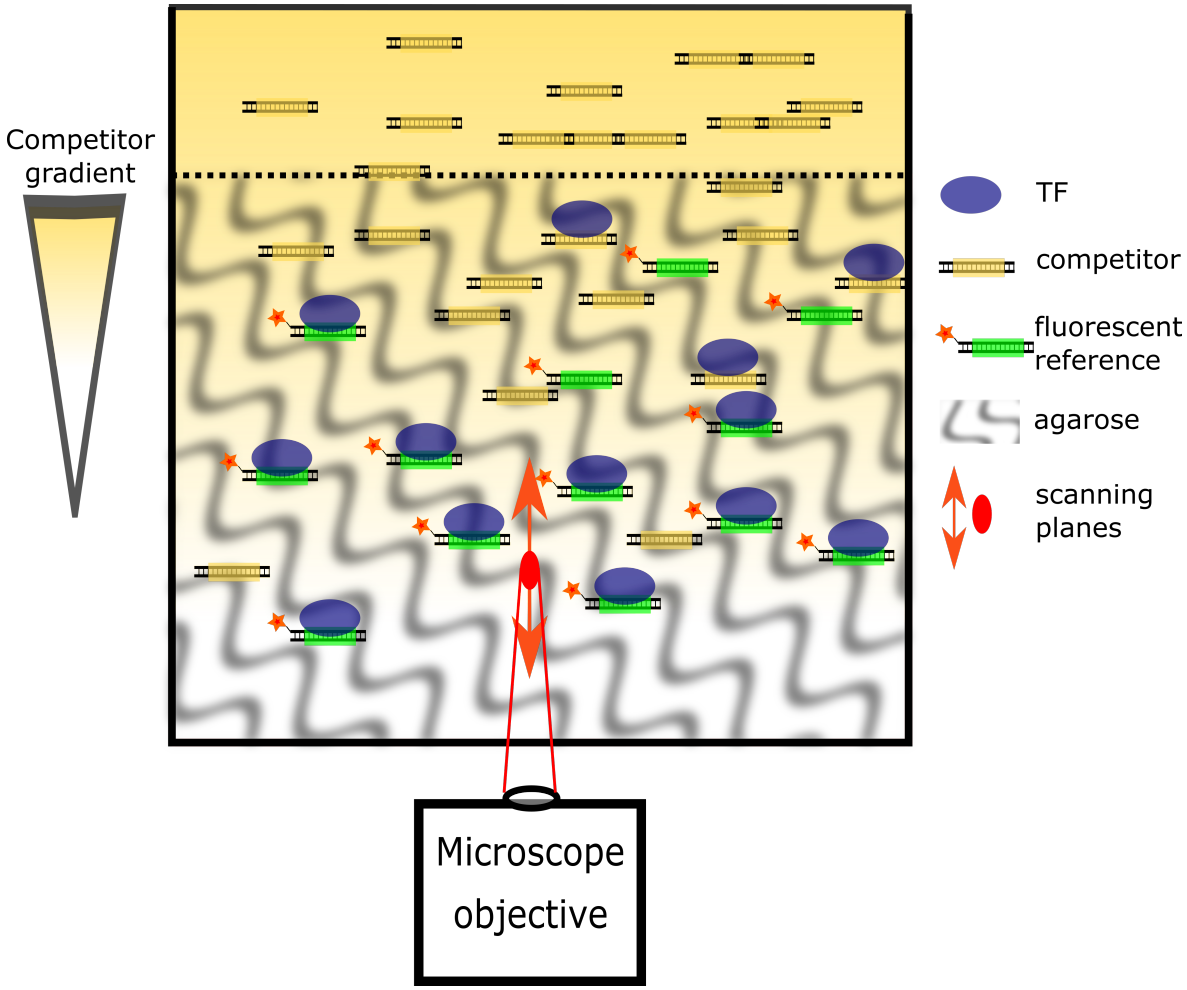


Figure 1.8: Sketch showing a single well during a HiP-FA assay. The well is partially filled with porous agarose gel (black strings), the upper boundary of the agarose is shown with a dotted line. In the gel, the fluorescently labelled reference DNA (binding site in green, fluorophore: orange star) and the TF (blue ellipse) are embedded already since before the experiment was started. The competitor sequences (binding sites in yellow) were put on top of the gel and already partially diffused into the gel. The competitor concentration gradient is symbolized by the yellow tone in the background. The microscope is depicted on the bottom of the well, the LASER beam and the focal planes are depicted in red and orange.

Controlled delivery system HiP-FA uses a competitive titration system and increases its performance by adding a "controlled delivery system" (see also Figure 1.8): The transcription factor

(TF) of interest and a fluorescently labeled reference DNA are cast into a porous agarose gel matrix. The pore size is big enough to allow for free diffusion but prevents convection. At the start of a measurement, unlabeled competitor DNA is added on top of the gel matrix. Over time, the competitor DNA diffuses into the gel, forming a gradient. By measuring the FA at different heights and time points, many different concentrations are measured and a whole titration series can be recorded in a single well. The resulting titration curves (see Figure 1.9) are steeper the stronger the competitor is bound by the TF.

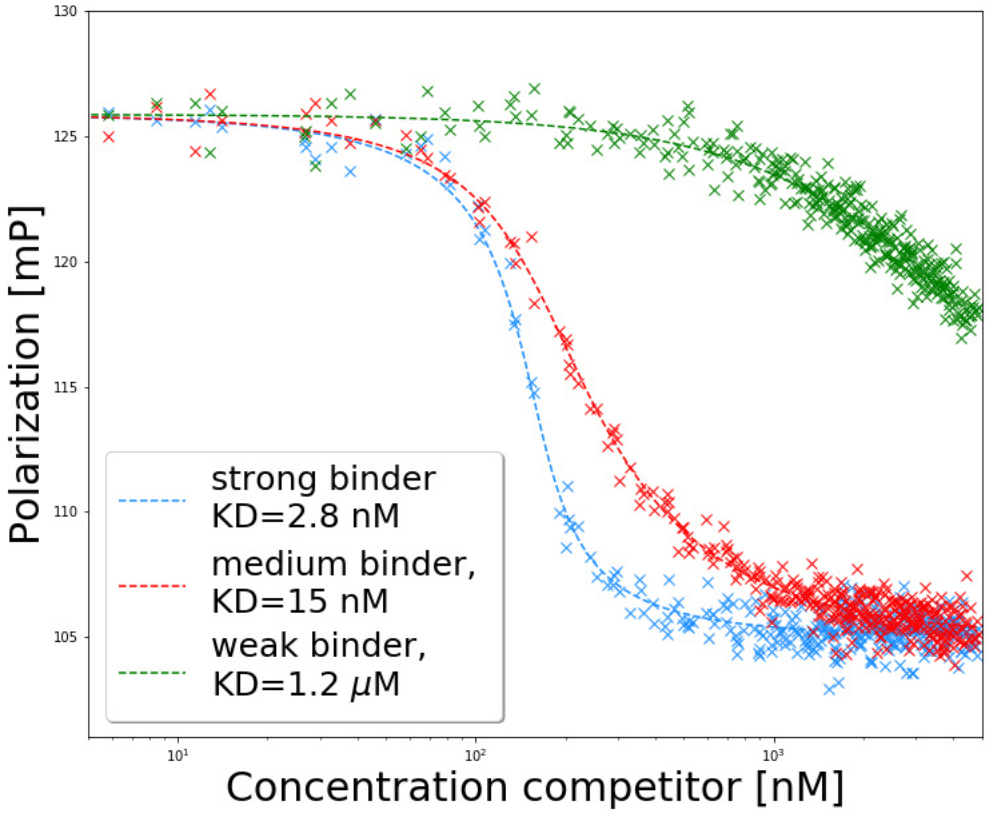


Figure 1.9: Examples of HiP-FA titration curves. The plot shows three different binding curves, a strong (blue), a medium (red) and a weak (green) binder as competitor sequence.

Concentration determination To calculate affinities in this assay, information about the concentration of competitor DNA at any given height and time is needed. To this end, one to two "calibration wells" are included in each row of sample. In these wells, Nile blue, a fluorescent dye intercalating into DNA is embedded into a gel matrix of the same agarose concentration. A reference DNA on top of these wells -with the same length and a corresponding diffusion coefficient- diffuses into these gels with the same pace as the competitor DNAs in the titration wells (Figure 1.10 a). With the help of a calibration before the experiment (Figure 1.10), the FA of the calibration wells can be converted into concentration at any given point and thereby the concentration in all wells at any given time or height can be deduced.

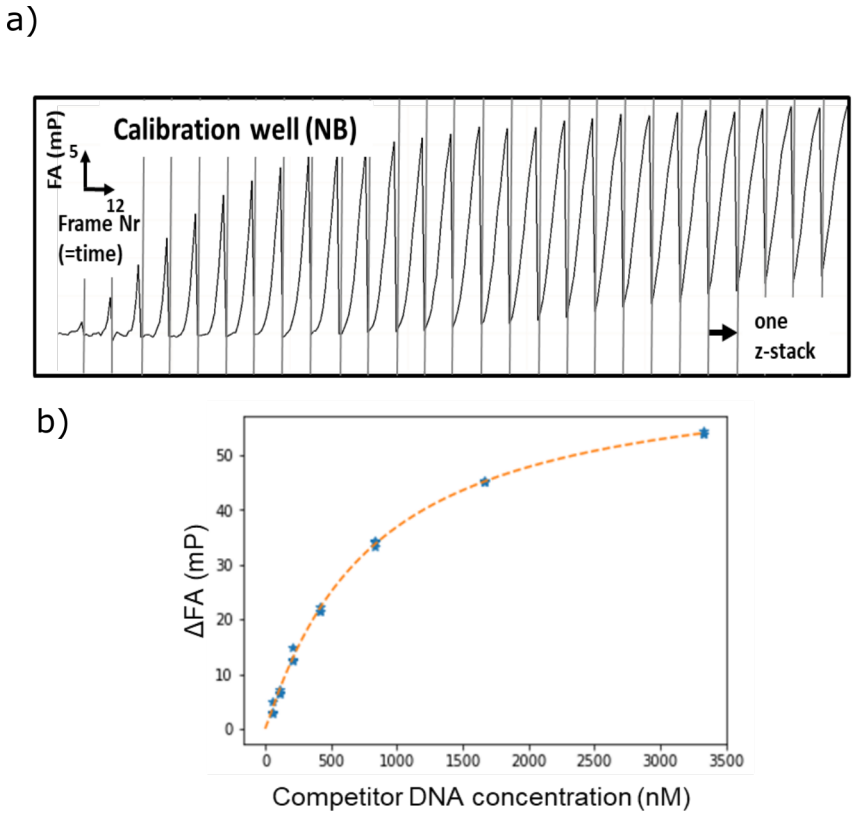


Figure 1.10: Nile blue calibration. a) example of the FA signal in a NB well over the course of an experiment. Between two vertical lines, one z-stack of usually 12 images is represented. With increasing concentration of competitor, the total FA increases over time. b) Calibration curve used to convert the FA in a NB well to concentration.

Microscopy setup Figure 1.11 shows a scheme of the microscopy setup used both in HiP-FA and in the nucleosome titration assay (section 1.5). The system is based on a commercial widefield microscope. The optical setup for both the excitation and the emission light paths are to a large degree mounted on an optical table. The sample is excited by polarized laser light. The resulting fluorescent emission with its respective shift in polarization is split by a polarizing beam splitter leading to two channels with orthogonal polarization. Both beams are projected onto an EM-CCD (electron multiplying charge coupled device) camera.

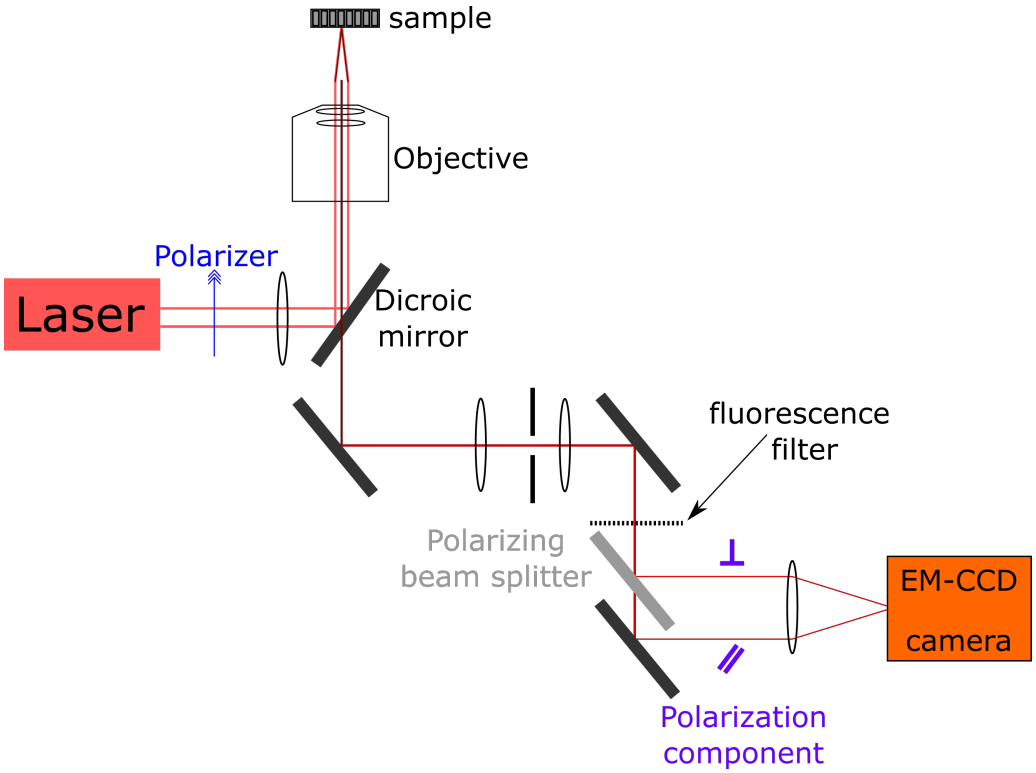


Figure 1.11: The figure is a depiction of the microscopy setup used for both HiP-FA and the nucleosome titration assay. The excitation beam is drawn in a brighter read than the emission.

1.5 Determination of histone-DNA binding energies in nucleosomes

1.5.1 DNA accessibility

An important aspect in the regulation of gene expression is the accessibility of DNA for the TFs and the transcriptional machinery. The main factor determining the availability of a given DNA sequence to other proteins is its interaction with the histone proteins. The basic function of histones is compacting the long DNA strands by wrapping it around them about 1.7 times, forming a nucleosome (see Fig. 1.12). This does not only reduce the extension of the DNA strands but also reduces interactions between possible binding sites incorporated into the nucleosome and TFs (Khorasanizadeh (2004)). The regulation can happen both on the level of "tightness" with which the complex is formed and on the relative positioning of the nucleosome along the linear DNA. How strong the interactions between a histone octamer and the DNA are mainly depends on post-translational modifications of the histones at their unstructured N-terminal tails with small

chemical groups like phosphates, acetylations and methylations. Generally, acetylations are usually weakening the interaction while methylations are mostly leading to a tighter binding (Tessarz and Kouzarides (2014)). Beside these purely charge based alterations of the interaction strength, the patterns of these modifications are in addition read out by nucleosome remodelling complex ("remodellers"), further changing the positioning and stability of nucleosomes (Lorch and Kornberg (2017)). The second mentioned regulation, the relative positioning, is influenced both by these remodellers and by features of the DNA sequence discussed during the specific introduction in chapter 3.

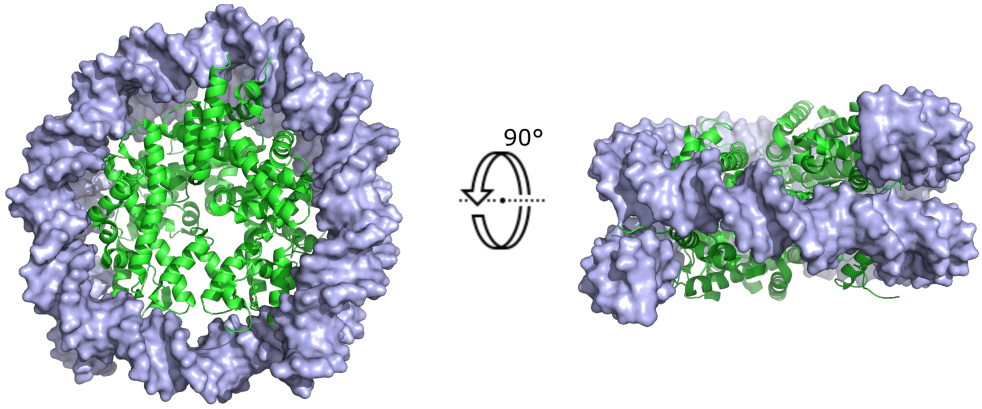


Figure 1.12: The figure shows the molecular view of a nucleosome (PDB-ID: 3KWQ, Watanabe et al. (2010)). The histones are depicted in green, the DNA in blue. Like can be seen in the two different view angels (top view on the right, front view on the right) the DNA wraps around the central histone octamer almost two turns.

1.5.2 Methods to determine histone-DNA interaction strength

In principle, the interaction strength between a histone octamer and a given DNA can be determined in two ways: Either by a salt-titration based assay or by an enzymatic one. At the beginning of a salt titration based assay, the binding energy between the DNA and the histone-octamer complex is weakened in a high salt environment Puhl and Behe (1993). By gradually reducing these concentration of salt these interactions are more and more permitted. The relative binding energy ($\Delta\Delta G$) can be determined if two different sequences compete for the histone octamer. The reference sequence is labelled (fluorescently or radioactively) and its ability to be incorporated into the nucleosome complex is monitored by comparing the ratio of incorporated and free labelled DNA. The unlabeled competitor occupies nucleosomes proportionally to its relative binding strength with respect to the reference sequence and is therefore never an absolute value. Figure 1.13 illustrates different conditions both what reference points and measurements are concerned using the example FA-readout to trace the amount of incorporated reference DNA in the nucleosome complex. The enzyme based Methods are directly based on the accessibility of the DNA to the protein sensitive DNA processing enzyme (MNase, ATAC). The binding strength is then afterwards determined by a high-throughput sequencing experiment and the mapping of (un-) occupied parts of the total DNA sequence.

Step wise titration In a step wise titration, the buffer conditions were traditionally changed in view steps (often 3 titration steps from 1 M NaCl to 0.1 M NaCl, like in Shrader and Crothers (1989)). This rather abrupt change in NaCl concentrations doesn't necessary allow for equilibration processes which might happen at certain ionic strengths during a titration process (like an internal

rearrangement in the forming nucleosome at 0.5 M NaCl (Oohara and Wada (1987))). This procedure might therefore favor kinetic stability over thermodynamic one.

Dialysis In a dialysis experiment, the same principles concerning the inhibitory effects of high salt are used to slowly form nucleosomes from a dissociated solution containing histones and labelled DNA (Thastrom et al. (1999)). The difference to a salt titration is the way the concentration of salt is reduced: Dialysis happens in a compartment which is separated from a reservoir by a semi-permeable membrane. The reservoir contains a buffer with low salt and by osmosis, the salt concentration in the compartment is slowly and constantly reduced over time. This is the strongest advantage of this method over a salt titration. Important disadvantages are the handling of the dialysis chambers (which prevents automation and higher throughput) as well as potential interactions of molecules with the semi-permeable membrane.

MNase-seq In an MNase-seq experiment (Segal et al. (2006)), the nucleosomes are assembled on DNA and in an MNase treatment, the DNA not protected by nucleosomes is cut by the protein sensitive nuclease. The protected DNA sequences are separated from the histones and are sequenced using next generation sequencing methods. By mapping the fragments back onto the whole DNA sequence, the positions and residing probabilities of nucleosomes can be determined. Using assumptions about total histone and nucleosome numbers affinities can be calculated. The advantage of this method or other sequencing based accessibility assays is their high throughput and the information about nucleosome positioning and occupancy. These methods, however, require careful analysis of their data and can suffer from enzyme sequence preferences (Jin et al. (2018)).

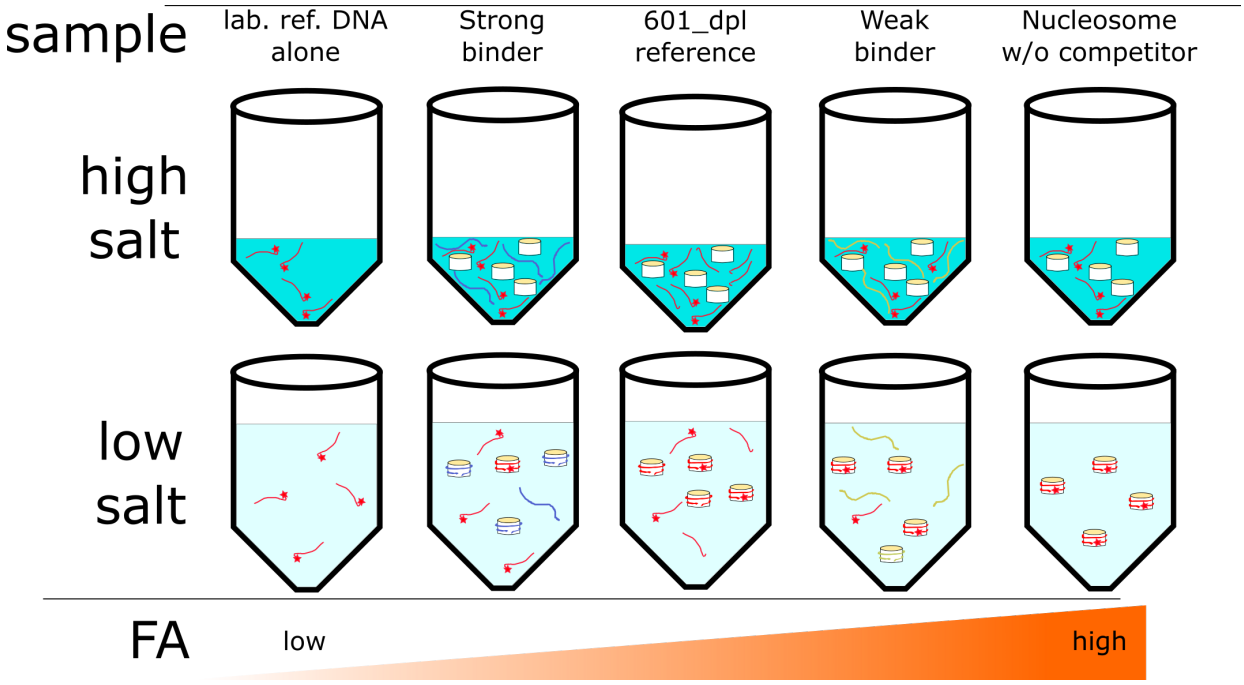


Figure 1.13: Schematic representation of different sample types and their corresponding fluorescence anisotropy levels. The schema depicts the composition of different samples mixed together with the fluorescently labeled reference DNA sequence and different competitor DNA sequences. The first sample (left) contains only the fluorescently labeled reference DNA, while all other samples also contain histone octamers. The 3 following samples contain in addition (from left to right): a strong competitor sequence, the unlabeled reference sequence, and a weak competitor sequence, respectively. The last sample (right) contains nucleosome and labelled reference DNA without competitor DNA sequence. The samples are ordered by their endpoint FA range.

Chapter 2

Transcription factor-DNA interactions

2.1 Introduction

The binding of transcription factors (TFs) to specific DNA sequences is essential for regulation of gene expression. The features defining a binding site have been the focus of several decades of research, starting from simple binding sites, later replaced by probabilistic models of TF binding. The so-called position-specific weight matrices (PWMs) (Stormo et al. (1982)) allow for different mutations in a TFs binding site under the assumption that each base contributes independently to the overall affinity. With the advent of high throughput methods, binding specificities have been available for thousands of TFs and it has become clear that more complex models for binding sites using non-independent nucleotide interactions lead to more accurate prediction than PWMs (Weirauch et al. (2013), Zhao and Stormo (2011)) Ruan and Stormo (2017). Numerous algorithms incorporating nucleotide dependencies have been developed and they proved to perform better than the PWM model that neglects higher-order interactions (Riley et al. (2015),Weirauch et al. (2013),). However, whereas determining precisely binding specificities - including non-linear dependencies - is crucial to predict accurately binding sites in the genome, such data are insufficient to fully describe TF-DNA binding interactions as they do not provide insights about the mechanism the TF employs to bind to different DNA sequences. To elucidate how the TF reads the DNA is of paramount importance not only to improve algorithms that predict binding sites, but also to refine our fundamental understanding of how the TFs are recruited to specific DNA regulatory sequences. To date, two distinct modes of protein-DNA recognition are known: base readout, that reflects the interplay at nucleobase-amino acid contacts resulting in the formation of hydrogen bonds and van der Waals (VdW) interactions, and shape readout that recognizes the 3D structure of the DNA double helix or the resulting electrostatic properties. Hence, if the TF uses the shape readout, then models incorporating DNA structural information should improve prediction of TF-DNA binding specificities. However, combining binding specificity information with DNA shape features remains challenging with existing methods. To help the development of such models it is highly desirable to determine as accurately as possible the TF-DNA binding specificities at the PWM (also called 0th [higher] order) and at the 1st order (nucleotide dependencies) of binding, but also the contribution to binding of the DNA shape readout. Whereas PWMs are available for numerous factors, the methods used to determine binding specificities have serious limitations. Despite the availability of high-throughput techniques able to measure protein-DNA interactions such as protein-binding microarray (PBM) (Berger et al. (2006)), SELEX-seq (Rastogi et al. (2018); Riley et al. (2014)) and SMILE-seq (Isakova et al. (2017)), the accurate measurement of their binding affinities remains problematic, which is critical especially to determine higher order matrices. In addition, most of

these methods use stringent protocols resulting in the loss of weak binders, which can lead to erroneously over-specific binding specificities. To prevent those, only few cycles of enrichment can be permitted and an elaborate algorithm like NoReadLeftBehind (NRLB) (compare Rastogi et al. (2018)) has to be employed. The determination of the shape readout contribution to binding also poses severe challenges. First, although it had been known for a long time from crystal structures that TFs read out the DNA shape, it is still experimentally impossible to determine at large scale the DNA shape features for a given DNA sequence. This would be necessary to quantitatively assess DNA shape influence on TF-DNA binding (Zhou et al. (2015); Yang et al. (2017)). This issue has been tackled by Zhou et al. (2013) who introduced DNAShape , an algorithm that predicts structural DNA features from a nucleotide sequence. The original set of four geometric shape features was completed by Li et al. (2017), who made tables available to calculate an expanded repertoire of 13 DNA shape features in total. Finally, Chiu et al. (2017a) added in a comparable fashion the electrostatic potential (EP), reflecting the charge density mean of the DNA backbone, sensed by positively charged amino acid residues of the binding protein. Another difficulty to analyze the influence of DNA shape to binding is that, in spite of all the advances made possible by DNAShape and the succeeding studies, it is still not clear to what degree apparent shape readout can be described as a function of the underling DNA sequence, i.e. is simply a more complex base readout. It is indeed difficult to tease apart whether a binding protein favors a given sequence because it recognizes certain nucleotides, or rather certain shapes features the DNA helix. An important step was made with homeodomain TFs by Abe et al. (2015), who were able to specifically remove the ability of the binding proteins to read a certain structural feature of DNA and to switch between different modes of DNA shape readouts. Another approach computationally dissects TF binding specificity in terms of base and shape readout (Rube et al. (2018)). Remarkably, the authors determined that 92-99% of the variance in the shape features can be explained with a model taking only dinucleotides dependencies into account. They also found that interactions were much stronger between neighboring nucleotides than for non-adjacent positions, indicating that these dinucleotide features are the most important for binding. Unfortunately, whereas these studies shed new lights on the role of DNA shape in TF-DNA recognition, they were limited to the analysis of only a few factors and to four different shape features. This was due to the lack of quantitative data on non-linear interactions, and to the unavailability of tables to calculate the remaining shape features at that time. Thus, a more comprehensive quantification of TF-DNA binding especially non-linear dependencies is urgently needed to deeper understand TF-DNA binding, in particular to what extend DNA shape features are recognized by TFs. Recently, we presented high-performance fluorescence anisotropy (HiP-FA) (Jung et al. (2018, 2019)), a method that determines TF-DNA binding energies in solution with high sensitivity and at large scale, and allows for measuring affinity of a TF to any given DNA sequence. These features predestinate HiP-FA to measure TF-DNA binding specificities, especially the non-linear dependencies since these interactions are intrinsically weak and their accurate measurement is both difficult and indispensable. In this study, we used HiP-FA to measure binding energies for 13 TFs of the Drosophila segmentation gene network and belonging to eight different binding domain families. We determined their 0th (PWMs) and 1st order (dinucleotide position weight matrices - DPWMs) binding specificities. Correlating our affinity data with the 13 known DNA shape features and the EP, we find that nearly all our factors extensively use shape readout for DNA recognition, independently from the binding domain family. Finally, we examined for five factors the correlations between their co-crystal structures and shape attributes obtained from our analysis, and ran a cluster analysis to test if certain shape features tend to co-occur in the DNA shape readout used by our TFs.

2.2 Results

Determination of the TF-DNA binding specificities and overall analysis strategy The PWMs of the 13 factors were already presented in Jung et al. (2018). We demonstrated that our

PWMs perform better than others obtained by Bacterial one hybrid (B1H) or DNase footprinting in predicting ChiP-seq data, and when used in a thermodynamic model for gene expression in Drosophila embryos (Jung et al. (2018)). Herein, we extended the binding preferences measurements to capture potential non-linear interactions. We measured binding affinities by HiP-FA again for all mononucleotide (0th order) and for all neighboring dinucleotide (1st order) mutations (in total 1600 individual titrations; Figure 2.1a) in the core of each TF's binding site (6 positions for the TF GATAe, 7 positions for all other TFs). We measured duplicates or triplicates for 6 factors. Two distinct analysis of the data were performed: first, we used our binding affinities to determine the PWMs and the DPWMs. In the analysis procedure we corrected for the energy contribution of off-target binding sites that might be created by chance in dinucleotide mutations (Figure 2.1b and Methods). Second, we assessed the influence of DNA shape on the binding strength over the core DNA binding sequence. The 13 shape features and the EP were calculated using the lookup tables provided by Zhou et al. (2013) and later expanded by Li et al. (2017), supplemented with the electrostatic potential (EP) (Chiu et al. (2017b)). We then applied robust linear regression (Methods) to determine the contribution of each shape feature by correlating its values with the binding energies of all possible mutations tested at a given position (Figure 2.1c; see below for details).

Consideration of off-target weights Figure 2.3 shows the predicted off-target weights for the already optimized consensus sequences. The figure illustrates the need for the off-target removal in a post-processing analysis. While most values are located above the threshold of zero (logarithm of the ratio 1), there are some factors that suffer from significant off-target binding due to the nature of their binding sites. The algorithm to optimize the flanking sequences minimized the metric depicted in blue, the off-target weight consisting of the sum of all off-target weights plus the strongest off target weight (effectively evaluated twice) . More relevant for practical purposed might be the ratio to the strongest off-target binder, depicted in red. If the assumption of linearity holds true, the binding affinity of sequences with a red dot under the black line are stronger influenced by the off-target binding site than by the actual (double-) mutation. This is mostly the case for Oc, Gsc, Hb and Fkh. Oc and Gsc were measured using the consensus site for Bcd since their PWMs are still quite similar. Hb suffers from its very monotonous PWM consisting manly of a T stretch. To avoid the influence of such off-target binding sites, Marc von Reutern developed an algorithm to construct a PWM and the resulting DPWM *de novo* taking all possible binding sites on the sequences into account. We present this algorithm in the method section.

Zeroth and first-first order binding specificities for the *Drosophila* TFs After having measured the binding affinities for all factors, we calculated their corresponding PWMs and DPWMs based on these data (Figure 2.4 and Methods). Overall, the PWMs are similar and largely share the same consensus than PWMs obtained by other methods, but they have generally a lower specificity (as measured by their information content IC), as already discussed in Jung et al. (2018). By contrast, our DPWMs show fewer but more preferred dinucleotides (as indicated by higher individual ICs) compared to computationally derived DPMMs (Siebert and Soding (2016)) or obtained using SMILE-seq data Rube et al. (2018). As an example, for Bcd (Figure 2.4) at position 5 in the DPWM (corresponding to the dinucleotide mutations between positions 4 and 5 in the PWM) the four pairs AT, AG, GT and CA have a cumulated IC of nearly 1, thereby predominating to the 11 other possible dinucleotides mutations. For all factors, we observe that the contribution to binding of the zeroth order predominates over the first order, as indicated by the higher ICs of the specificity logos (6.9 bits on average for the 0th order compared to 2.1 bits mutual information for the 1st order; Figure 2.4). This was expected as the simple PWM model has proven to capture most of the sequence preferences for numerous TFs (Stormo et al. (1982); Zhao and Stormo (2011)). Surprisingly, the DPWMs of nearly all our TFs (with the exceptions of GATAe and Gt) show a high contribution to the overall binding specificities, as indicated by their relatively high

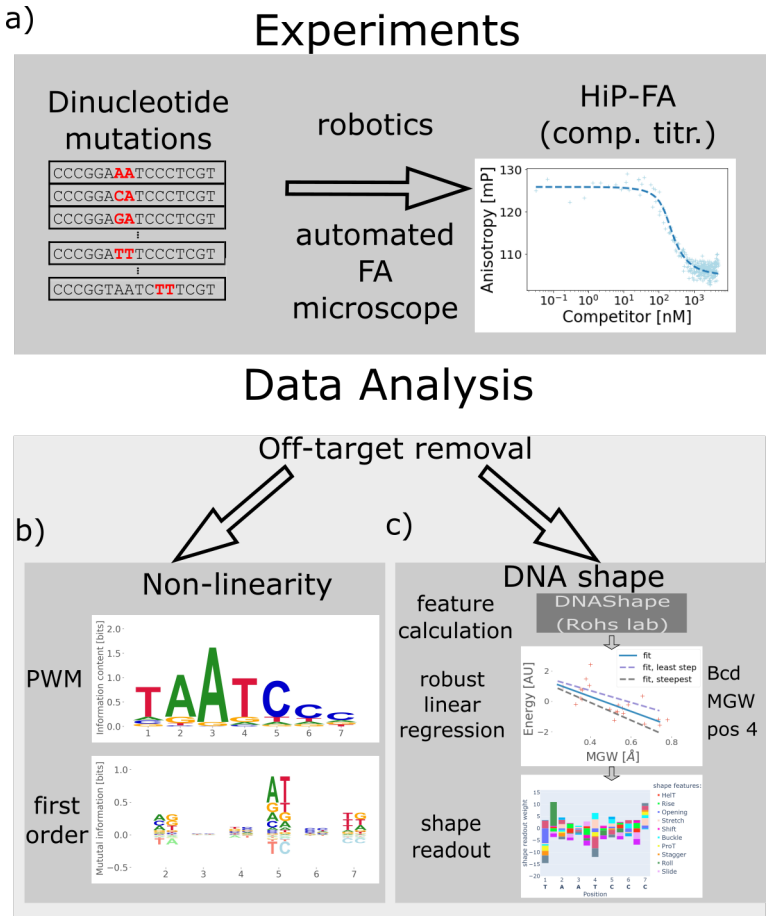


Figure 2.1: Experimental procedure. a) Experimental part. A consensus sequence is systematically mutated with all possible dinucleotide mutations. Using a robotics system and an automated FA microscope, the binding curves are determined using the HiP-FA method. b+c) Data analysis. After an off-target removal procedure, the data are analyzed for non-linearity (b) and DNA shape (c). b) Depicted are the Position weight matrix and the first higher order of the TF Bcd. The first order shows the mutual information not included in the linear PWM. c) The sensitivity to shape is analyzed following the subsequent steps: The DNA shape features are calculated using lookup tables provided in publications by the Rohs group. The resulting values per position are plotted against the binding energies and a robust linear regression is performed. Besides the fit (blue line) a steepest (grey dashed line) and a least steep fit (purple dashed line) are estimated using the confidence intervals of the robust linear regression. The least steep slope is the shape readout weight. Shape readout weights of all features and positions are depicted in the lower right for Bcd.

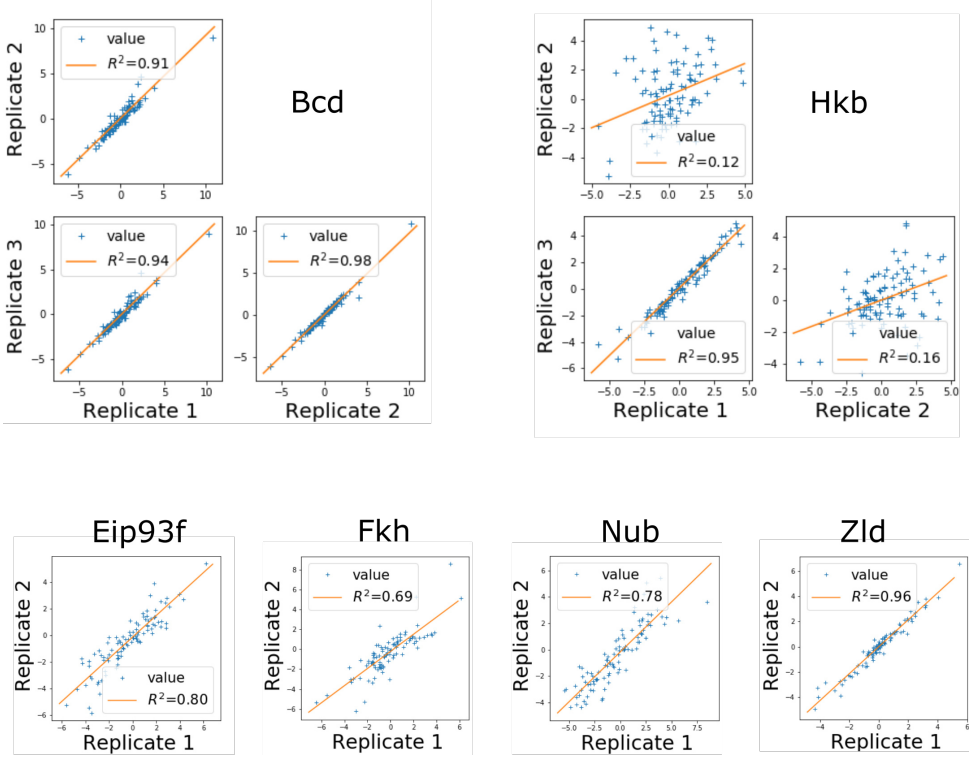


Figure 2.2: Reproducibility of shape readout weights between replicates. Triplicates plot each replicate against each other replicate. R^2 (squared Person-correlation-coefficient) is given for linear regression. Note that Replicate 2 for Hkb is in poor agreement with both replicates 1 and 3.

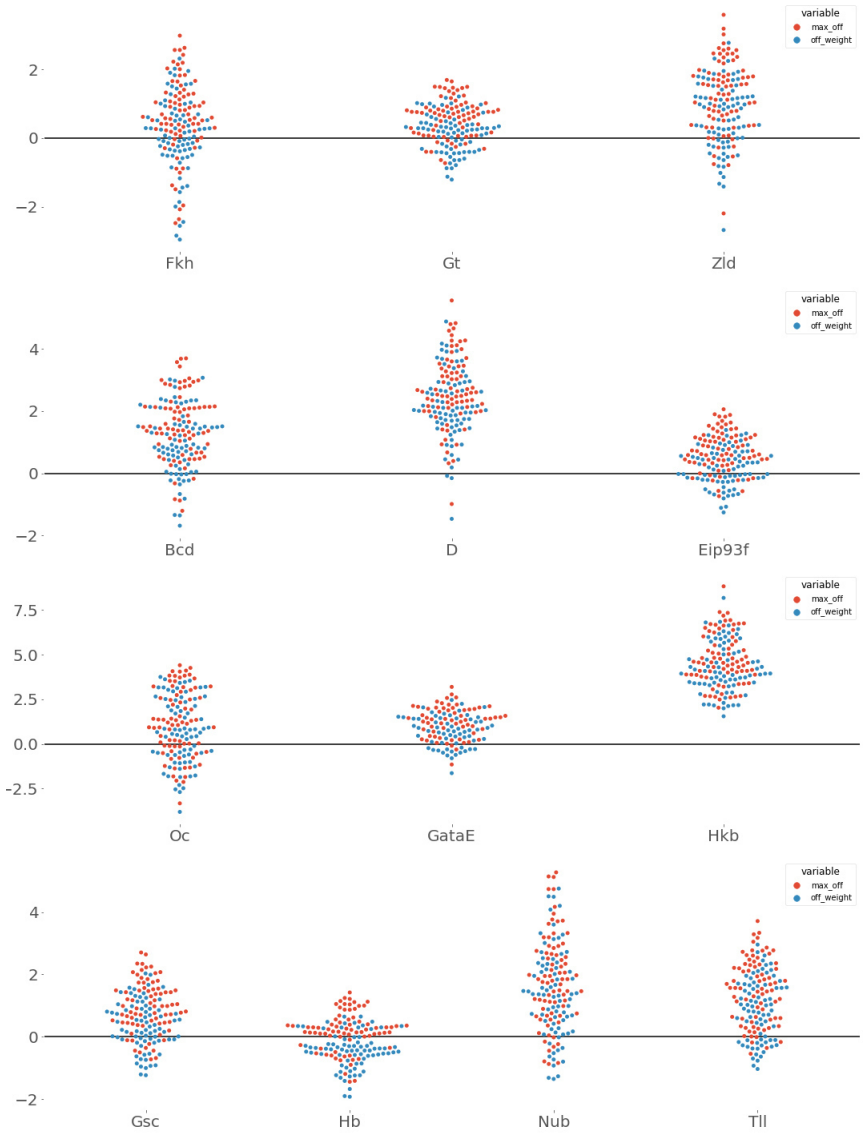


Figure 2.3: The figure shows the predicted off-target weights for all investigated dinucleotide mutations on an already optimized consensus sequence as 10 based logarithms of the ratio to the on target weight. The baseline of 0 (logarithm of 1) is depicted in black. Two metrics of off-target binding are displayed: (1) The ratio to the weight of the strongest off-target binder (in red), as this value is most important to judge the investigated K_D. (2) The off-weight as the ratio of all off-target binders plus the strongest off target. This value was used to optimize the flanking sequences.

mutual information (> 1 bit). Several studies already emphasized the importance of neighboring nucleotides in the prediction of TF binding (Siebert and Soding (2016); Nitta et al. (2015); Zhao et al. (2012)), but only for few factors. Our results suggest that the strong influence of non-linear interactions to binding is widespread. The sensitivity of HiP-FA enables us to accurately resolve weak but measurable binding events and their deviations from linearity, which are difficult to detect with other methods. Noteworthy, the three members of the Homeobox family (Bcd, Gsc and Oc; Figure 2.4) show similar PWMs and DPWMs, reflecting the high similarity of their binding domains. This observation is in line with previous works, showing the generally high similarity between homeodomain TFs zeroth order binding preferences (Affolter et al. (2008)). Our DPWMs show that this similarity holds at the first order. A closer inspection, however, reveals the presence of subtle differences in specificities. At position 5 of the DPWMs for example, although the preferred dinucleotides are very similar (AT has strongest positive non-linearity, TC has strongest repulsive one), their corresponding absolute mutual information differs substantially between the three TFs (for the positive mutual information: 0.76 bits for Bcd, 0.42 for Oc, and 0.27 for Gsc, respectively). In addition, Bcd differs at Position 2 in its DPWM from the two other factors with its relatively high mutual information (0.35 compared with 0.08 for Gsc and 0.04 for Oc, respectively). Although these differences are weak, their concerted effect might be important to allow these homeodomains to execute their distinct biological functions.

DNA shape correlation with binding preferences all investigated TFs The fact that most of the variance in DNA shape is encoded in dinucleotides (Rube et al. (2018)) encouraged us to tackle the question to which extend TF-DNA binding is driven by DNA shape. To this end, we calculated the 13 geometric shape features and the EP for all tested DNA sequences, and determined their influence on our binding energies. For a given factor, we evaluate whether the change in binding energies correlates with a feature of interest when a base at a certain position and/or at a neighboring position deviates from the consensus sequence. In the case of Bcd at position 4 for example (Figure 2.1c), the binding energies decrease over an amplitude of round 4 AU when the relative minor groove width (MGW) increases from approx. 0.2 to approx. 0.8 (Methods). The shape sensitivity is determined by a robust linear fitting procedure (Methods) to minimize the effect of extreme values (outliers) and to provide a confidence interval to the resulting fitting parameters. The slope of the robust linear fitting provides an estimate of how much the binding of the TF at the particular position could be influenced by the local DNA shape (termed shape sensitivity value in the following) - assuming that it is caused by the variance in shape influencing TF binding. For each TF, we applied this analysis for every shape feature and at different base positions along the DNA binding sequence. We encoded the significance levels obtained with the robust linear regression as different densities of hatches in the plots (Figure 2.5). The shape sensitivity values were standardized for better comparison (z/standard score; Methods). The reproducibility of the shape sensitivity values among replicates was high (mean squared Pearson coefficient (R^2)=0.76 for the 6 factors having duplicates or triplicates; Supplemental Fig. 1). Surprisingly, the shape sensitivity plots (Figure 2.5) reveal a widespread use of DNA shape readout for all our TFs, with strong differences in the shape feature values between factors, and at different base positions for a given factor. Remarkably, the members of the homeodomain family (blue box in Figure 2.5) show again a similar behavior what their shape readout values are concerned (discussed in details below), as already observed for the PWMs and DPWMs. This doesn't hold true for the zinc fingers family (green box) or for the other factors with different binding domains, for which the shape sensitivity plots exhibit various patterns along the DNA binding sequences. Other studies also reported that zinc fingers dont show similarities in their binding behavior (Kribelbauer et al. (2019) and references therein), in contrast to other TF families. Interestingly, we found that in the middle of the binding sites of GATAe and Zelda (positions 3 and 4 for each factor in the shape sensitivity plots) the shape readout values are both very low (discussed below in more details for GATAe). These are positions where the sequence logos have a high IC, as indicated by the prominent TC and GG bases in the PWMs of GATAe and Zelda, respectively (Figure 2.4). Conversely, shape features

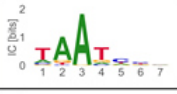
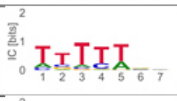
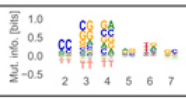
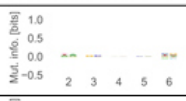
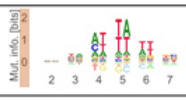
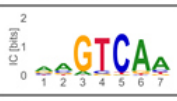
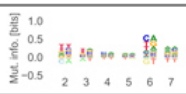
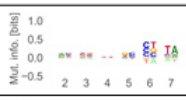
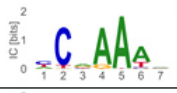
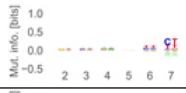
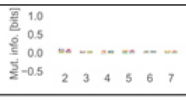
Factor (Family)	PWM (0th order)	IC	DPWM (1st order)	Mut. Info.
Bcd (Homedomain)		6.6		2.4
Gsc (homedomain		4.6		1.5
Oc (homedomain)		5.2		1.6
Hb (Zinc finger)		4.3		4.1
Hkb (Zinc finger)		11.8		2.1
GATAe (Zinc finger)		5.9		0.3
Zld (Zinc finger)		6.8		2.1
Nub (Pou domain)		7.8		6.5
Tll (NHR)		6.9		1.6
D (HMG Box)		8.4		2.5
Fkh (Winged helix)		8.4		1.6
Eip93f (Helix-turn-helix)		5.5		0.9
Gt (B-zip)		7.8		0.6

Figure 2.4: Overview of TF specificities. Depicted are the PWM and DPWMs of all TFs presented in this work. Plotted are the information content (IC) versus the position for the 0th order and the mutual information (Mut. info.) between two position for the 1st order. The total information is given in a separate column. Homeodomain factors and zinc fingers are grouped by color.

become important where sequence information is not well defined, as for GATAe at positions 5 and 6 and for Zelda at positions 1 and 7. This phenomenon has already been reported for other factors by Zhou et al. (2015), and can be generalized to the side chains of most of our other factors like for the three homeodomains, Hb, Tll, Fkh or Eip93f. In these cases, shape features contain more information than sequence alone.

Correlation between the DNA shape sensitivity of the TFs and their structural information We next wondered whether the predicted shape readout values can be confirmed in protein structures as interactions between the TF and its target DNA (figure 2.7). Among our TFs, only three have structural information available: Bcd has an NMR structure, Gt a crystal structure of a B-zip protein sharing the same recognition sequence (Pap1), and GATAe a crystal structure of a mouse homologue (GATA). Homeodomain proteins have been described reading out the minor groove at a local minimum (Baird-Titus et al. (2006); Dror et al. (2014); Yang et al. (2017)). Although the features ProT, Roll, HelT and MGW have been quantitatively investigated for Bcd by Rube et al. (2018), only the MGW had a significant shape-readout coefficient. We plotted a MGW profile of our homeodomain consensus sequence (figure 2.7 a, right panel) and identified the position of narrowing MGW to be located at the central T, corresponding to position 4 in Figures 2 and 3. Similarly to the aforementioned study, we found the MGW to have a significant correlation with the binding energy ($p < 5 \cdot 10^{-5}$) at this position (figure 2.7 a, left panel). For a more detailed comparison, we plotted our shape readout values for all positions of the MGW against the corresponding shape sensitivity coefficients determined by Rube et al. (2018). We found an excellent correlation (R^2= 0.99) for the subset of coefficients that Rube et al. found to be significant (Supplemental Figure 2.6), validating our approach (note that the MGW was the only significant shape feature in their data). Remarkably, we found in addition significant correlations for Stretch and the EP for all three homedomain proteins (figure 2.7 a, left panel), as well as ProT (for Gsc) and Buckle (for Gsc and Bcd), indicating that the TF reads multiple DNA shape features at this position. The reproducibility of the shape feature values is high among the different homeodomains (figure 2.7 a, left panel). As expected (Rohs et al. (2009)), all three proteins are sensitive to the MGW and to the EP (both a smaller MGW and a more negative EP enhance binding). It is noteworthy that a narrow minor groove is associated with a stronger EP, making these observations not completely independent. Since all three proteins contact the narrow minor groove with a their unstructured N-terminal tail (figure 2.7 a, middle panel), the strong similarity in the shape readout value of most features is not surprising. Another pertinent example is the TF Giant (Gt) belonging to the family of B-zip proteins (figure 2.7 b). Members of this family approach the DNA in a scissor like manor, with two alpha helices contacting the major groove from two opposing sites. Interestingly, the same mirror symmetry with a mirror plane between position 3 and 4 (C and G) is found in the PWM (figure 2.7 b, right panel), and partially in the shape sensitivity plot (figure 2.7 b, left panel). The shape readout values of both inter and intra features between positions 2 and 5 show a highly symmetrical pattern, in line with the binding mode of B-Zip proteins (figure 2.7 b, middle panel). This pattern, although conserved in the PWM, is not maintained at the side positions in the shape readout weight values, probably due to the fact that the DNA has more flexibility outside of the B-zips scissor and the TF has less contact to its minor groove and backbone. Finally, we examined the zinc finger protein GATAe (figure 2.7 c). Zinc fingers contact the DNA at two opposing strands with three contacts being at one strand (positions 4 to 2, ATC in the case of GATAe) and another at the opposing strand (position 1, T) Fedotova et al. (2017). There are multiple contacts at position 1 (blue circle in the middle panel of figure 2.7 c) between the TF and the DNA backbone, which matches the high shape readout values at this position (in total 14.1 AU (absolute sum)), blue circle in the left panel of figure 2.7 c). The contacts between TF and minor groove or DNA backbone decrease when going towards the central binding site, as seen in the crystal structure (blue circle in the middle panel of figure 2.7 c). The absolute values of the shape readout values show a similar behavior, a decreasing overall shape sensitivity going from position 1 to 4. At position 4 (red circles), one can observe contacts in the structure exclusively to

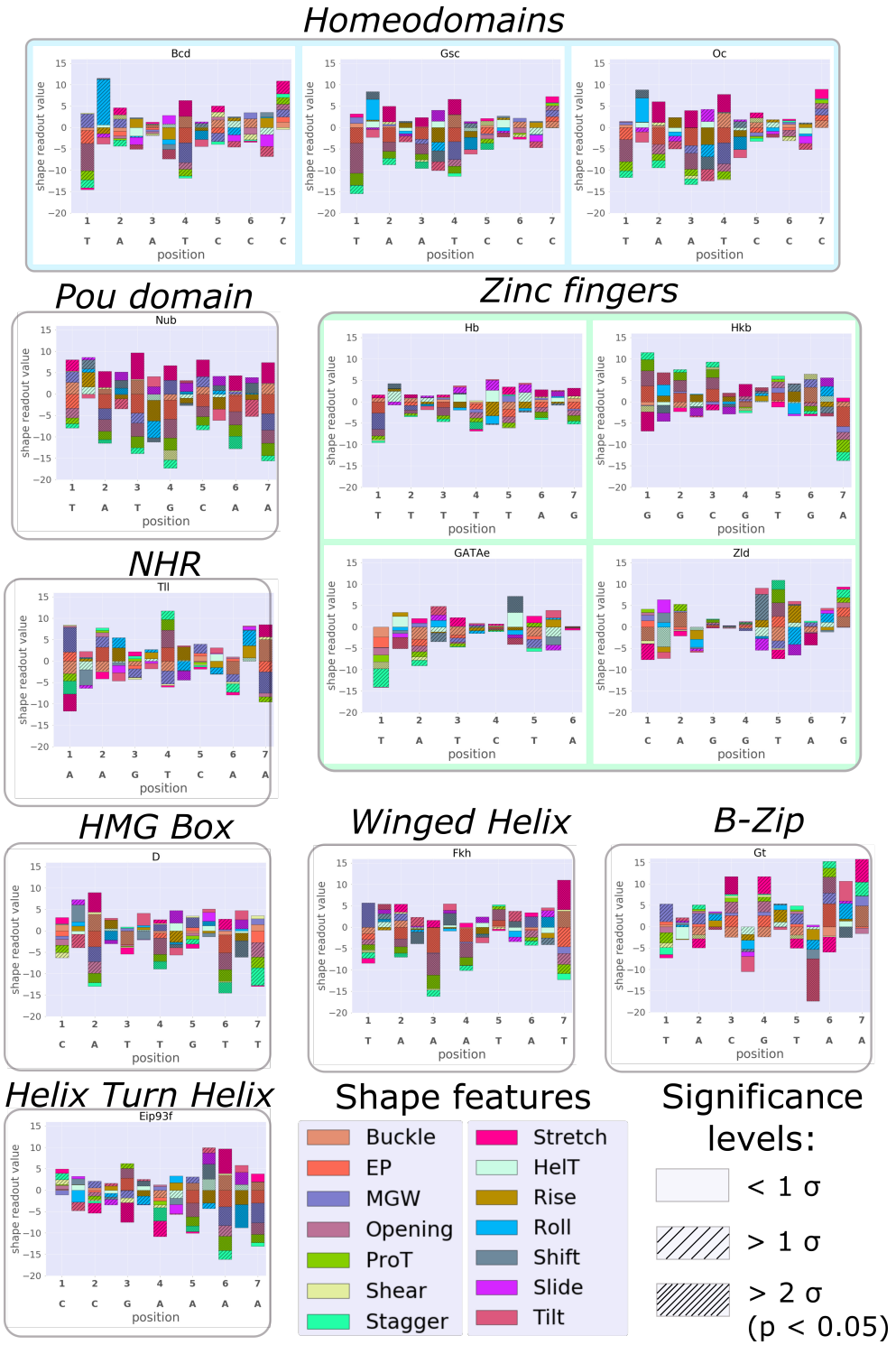


Figure 2.5: Overview of shape sensitivities for all TFs presented in this study. Plotted are the stacked shape readout weights for each feature at a position (intra features) or between two positions (inter features). To better draw easier comparisons to Figure 2.4, the positions are also labelled with their respective nucleobase at this position of the consensus sequence. The legend for the respective features is found in the lower right corner. Homeodomain TFs and zinc finger TFs are grouped together and indicated with the same colors as in Figure 2.4. The significance levels are indicated for each bars with a hashing code, indicated in the right bottom (see Methods for details)

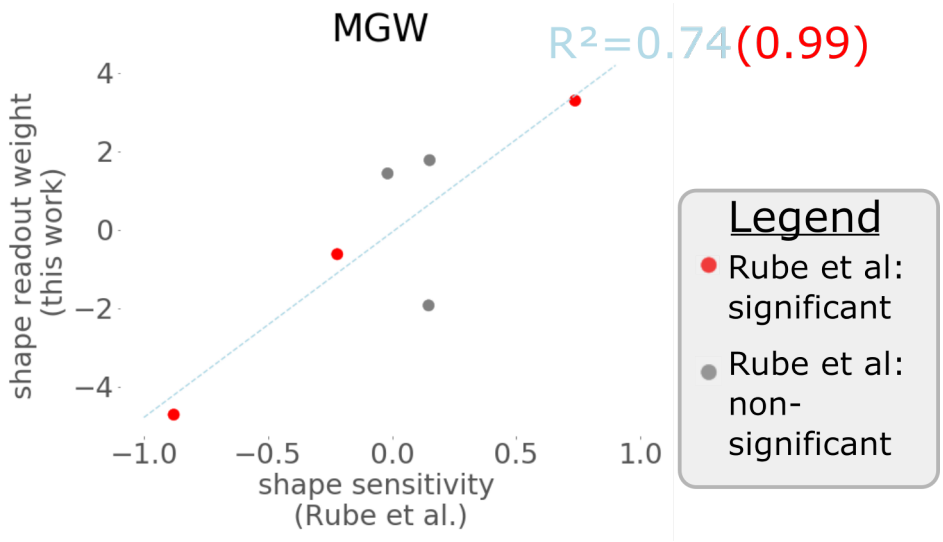


Figure 2.6: Comparison with MGW shape sensitivity from Rube et al. Plotted are the shape readout weights of the MGW for Bcd determined in this study against the shape sensitivity values reported by Rube et al. Rube et al. (2018). Values reported to be significant in their study are depicted in red, non-significant ones in grey. A linear regression (blue dashed line) has an R^2 of 0.74 for all values, 0.99 for only significant data points.

bases in the major groove, and the shape readout weights are reduced to a minimum (in total 1.8 AU). Intriguingly, we observe a similar behavior in the shape sensitivity plot of Zld (figure 2.5), with decreasing shape readout weights going from positions 1 to 3-4. Unfortunately, no structural information is available for this TF, but as a member of the zinc fingers the protein will contact the DNA in the three bases, one base on the opposite strand pattern. It is was recently reported that metazoan zinc fingers tend to establish several contacts to the DNA backbone (Najafabadi et al. (2017)), possibly permitting DNA shape readout at these positions. This binding behavior, partially based on unspecific backbone interactions, and the thereby promoted diversified evolution in metazoans might explain why this family uses an extremely diverse DNA shape readout (Yang et al. (2017)).

Shape sensitivity weights / TFs clustering Finally, we asked if the TFs use predominantly certain shape features to bind to DNA. To test whether shape features tend to co-occur in the shape readout, we performed two distinct cluster analysis of the shape sensitivity values matrix (Figure 2.8 and Methods): (1) the different features with respect to their feature readout by the TF (vertical lines in Figure 2.8), and (2) its converse a clustering of the TFs versus their shape readout of each feature as a matrix including all positions (horizontal lines in Figure 2.8). The TF clustering indicates that the different binding proteins show little similarity in their use of the shape features except for the homeodomains, which was expected (green tree on top of Figure 2.8). In contrast, the clustering of the shape features reflects structural dependencies between shape features. First, the heat map shows that EP is one of the features influencing TF-DNA binding energies the most, with strong means correlations for at least 7 factors (Gsc, Oc, Hkb, Eip93f, Fkh, Nub, and D; highlighted in green). The EP is sensitive to the interaction between positively charged residues and the minor groove, this strong impact on TF-DNA binding Chiu et al. (2017b) is therefore not surprising. Second, we observe three distinct clusters of shape features (cyan, red, and green trees on the right in Figure 2.8). These distinct groups may be related to biophysical properties of the DNA and its interplay with the binding protein (such as bends, kinks, A-/Z-DNA Rohs et al. (2010)). For instance, the first cluster (in cyan on the left) consists of slide, helix twist, roll and

MGW. These features were reported to correlate the most with each other both in unbound DNA El Hassan and Calladine (1996); Stella et al. (2010), and are read out concertedly in DNA-protein complexes Suzuki et al. (1997). This interdependency can explain their co-appearance in a cluster whithin our data. Moreover, the reported behavior of the features, that is decreasing values of HelT and Slide co-occurring with increasing Roll values Suzuki et al. (1997), is consistent with the observed subclusters: HelT and Slide are in the same subcluster, Roll clusters at the next higher level. It is also noteworthy that this cluster contains three inter-features (Slide, HelT and Roll) out of four, whereas the second cluster (in red) contains mainly intra-features (Stretch, Buckle and Shear). Thus, there seem to be a synergy between inter- and intra-features for the DNA shape readout. The third group (in green) is more heterogeneous and doesnt follow a simple pattern: it contains mixed shape features with, to the best of our knowledge, no known relationship to each other.

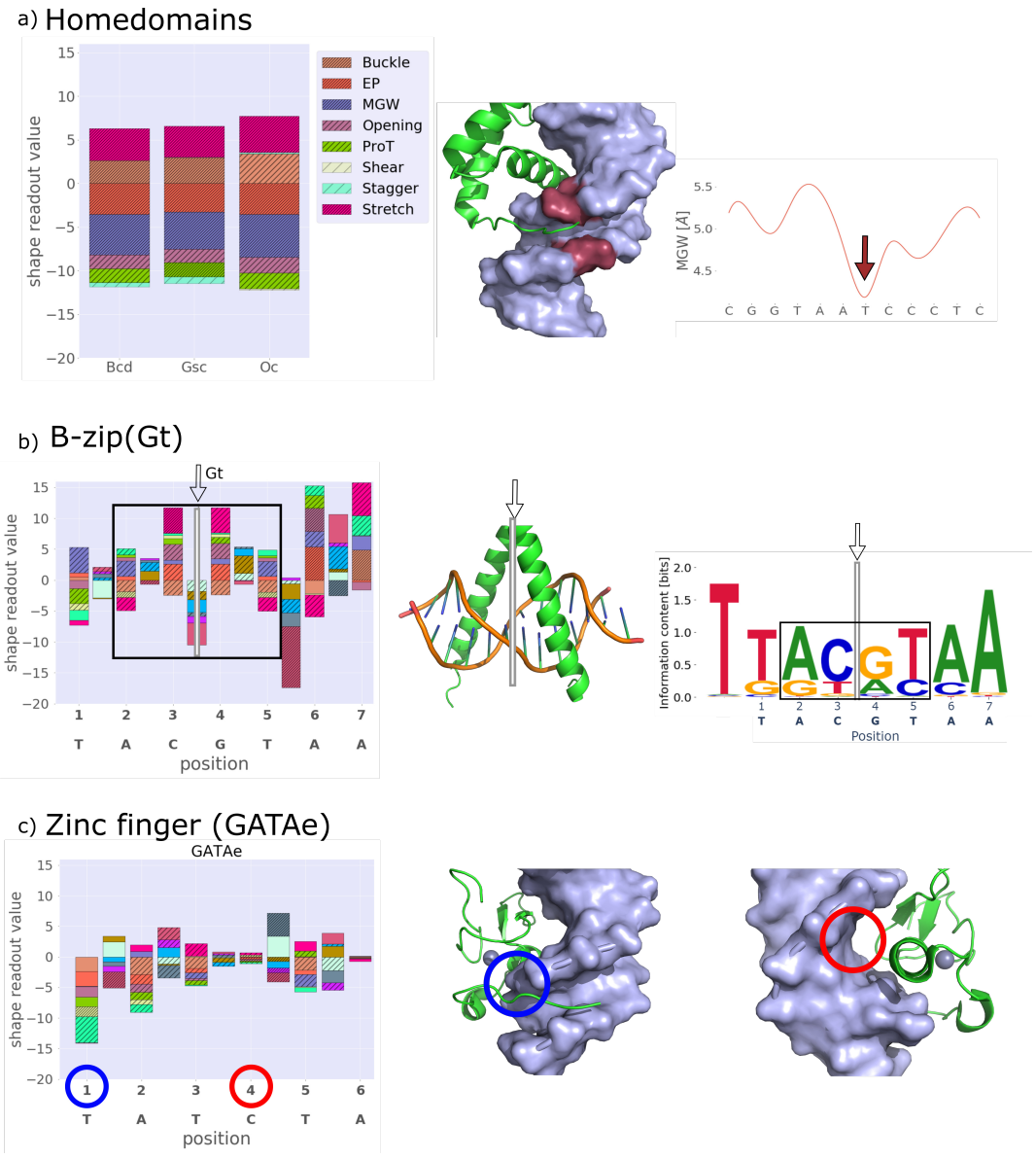


Figure 2.7: caption next page

Figure 2.7: (Previous page.)Detailed examples of shape readout. a) Minor groove contacts of Homeodomains. The left panel shows the shape readout weights of all three Homeodomain TFs at the same position. In addition to being very similar, all three show a strong readout of the minor groove at the discussed position. The middle image shows the crystal structure of Bcd (pdb-ID:1zq3) Baird-Titus et al. (2006). The bases at the above mentioned position are colored in red. The red arrow points at the position where the binding domain contacts the narrowing minor groove. The right panel shows the width of the minor groove width of the consensus sequence used for the Homeodomains. The position with the lowest minor groove is indicated with a red arrow. b) Symmetry in B-zip readout. The left panel shows a B-zip TF with the same core consensus sequence as Gt, the B-zip TF investigated. The middle panel shows the shape sensitivity of Gt. The black box indicates the region of high mirror symmetry around the grey mirror axis (added to all three panels of this row at the same position). The right panel shows the PWM of Gt, the first position augmented with data from Jung et al. (2018). The symmetry is distinct over the entire PWM. c) Differential shape readout by GATAe. The shape sensitivity of GATAe is depicted in the middle panel. Positions with strong (1, blue) and weak (4, red) shape readout weights are indicated at the x axis, as well their corresponding positions in a protein structure of a GATA TF at both sides. The left perspective shows a position with pronounced contacts to the DNA's phosphate backbone and minor groove. The right image shows a view emphasizing the strong dominance of major groove contacts at the second position.

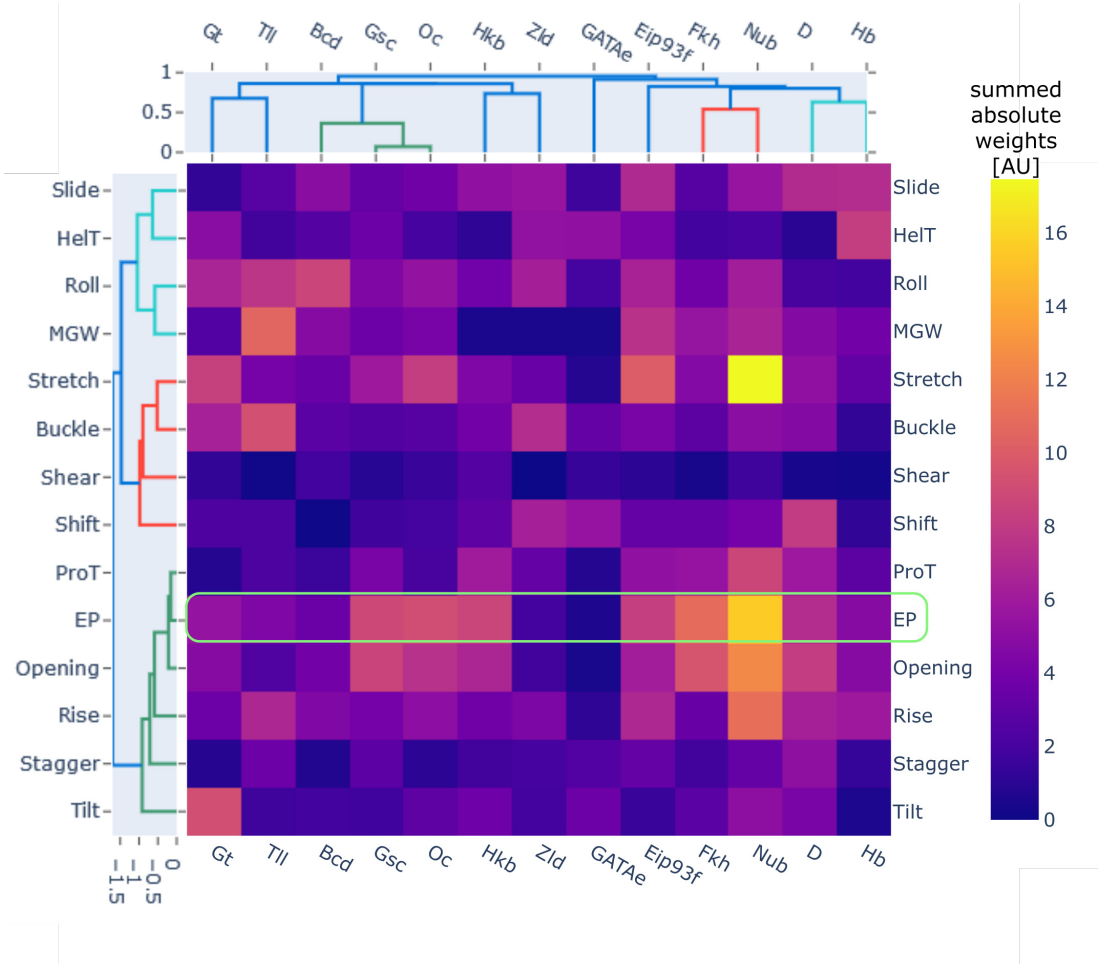


Figure 2.8: Heat map and clustering. The heat map shows the sums of the absolute shape readout weights over all positions of each of the TFs. In addition, both TFs and shape features are clustered by correlation distance. The clustering was performed on the non-aggregated data (with positions as a second dimension). The TFs cluster very little except for the members of the Homeodomains. In the clustering of the shape features, three distinct clusters are distinguishable.

2.3 Discussion

In this work, we expand the HiP-FA method to determine TF-DNA non-linear binding interactions. Other studies that characterized non-linear interactions (Yang et al. (2017); Chiu et al. (2017b); Rube et al. (2018)) relied on high throughput methods (Rastogi et al. (2018)), which often lack accuracy when determining affinities and are prompt to loosing weak binders during their stringent washing protocols. Although Rastogi et al. (2018) have developed an algorithm to limit this loss of very weak binders, only future applications of this method can tell how accurate it determines the binding strengths of weak binders. As HiP-FA can measure equally well strong and weak binders, the method enables distinguishing subtle variations in binding energies, and is therefore ideal to investigate non-linearity in TF-DNA binding. Taking advantage that most variation in the DNA shape features is already encoded in dinucleotides, we correlate our data with these features enabling us to draw conclusions about the interaction between TFs and the geometry of DNA. By combining information about binding specificities, the shape readout values, and structural information, we provide insights about the relationship between shape readout and non-linear interactions, a question often debated due to their intrinsic covariation. Importantly, our results suggest that DNA shape readout is widespread among our TFs. The extended use of DNA shape readout by TFs has become increasingly apparent over the past years (Zhou et al. (2015); Yang et al. (2017); Chiu et al. (2017b); Rube et al. (2018); Pal et al. (2019)), and is not surprising if one considers that the number of Van-der-Waals interactions enabling shape readout account for two-third of the protein-DNA interactions (Kribelbauer et al. (2019)). Our approach, which consists in measuring binding energies of a complete set of dinucleotide mutations, is more direct than the one used by Rube et al. (2018). The latter requires a prior analysis with the No Read Left Behind (NRLB) Rastogi et al. (2018) algorithm to derive affinities from high throughput data. In addition, our downstream analysis - the robust linear regression - uses fewer parameters and provides directly an interpretable characterization of shape sensitivity. However, our approach doesnt distinguish between base and shape readout. In the analysis procedure, we cannot exclude the possibility of energies changes due to base readout, leading to an incidental correlation between binding energies and shape features. We reason that this apparent contribution of shape features will average out in the linear fitting procedure and, as a consequence, will not lead to a significant correlation in the robust linear regression. This assumption is supported by the following: (1) our estimation of the shape sensitivity is in excellent agreement with the one obtained by the more elaborated algorithms of Rube et al. (Supplemental Figure 2.6), (2) the shape readout we determined is reflected in structural data for several factors (Figure 2.7), and (3) the clustering of the shape feature readout rediscovers already known interdependencies between shape features. Not all shape features are recognized independently by the TFs. The groups in our clustering might represent a specific DNA conformation which is read out by a TF, rather than the readout of several independent DNA features. These conformations might play an important role binding behavior of several TFs. In summary, our results give new insights on shape readout as a widespread DNA-readout mechanism by TFs. Our method could easily be extended to more factors and to different organisms to provide a refined catalog of the TF-DNA non-linear interactions and of the DNA shape readout landscapes.

2.4 Additional PWMs and assay improvements

In addition to the higher order PWM, I determined PWMs for the publication Jung et al. (2018). The investigated TFs belong to either the segmentation paradigm (see chapter 1.3) or are part of the ecdysone pathway (for review see Yamanaka et al. (2013)). They were either determined by myself or by Marc Nieveler as a Bachelor student under my supervision. This section additionally describes improvements to the HiP-FA assay which I developed, such as the modular reference oligo system and the optimization of the fluorescent dye.

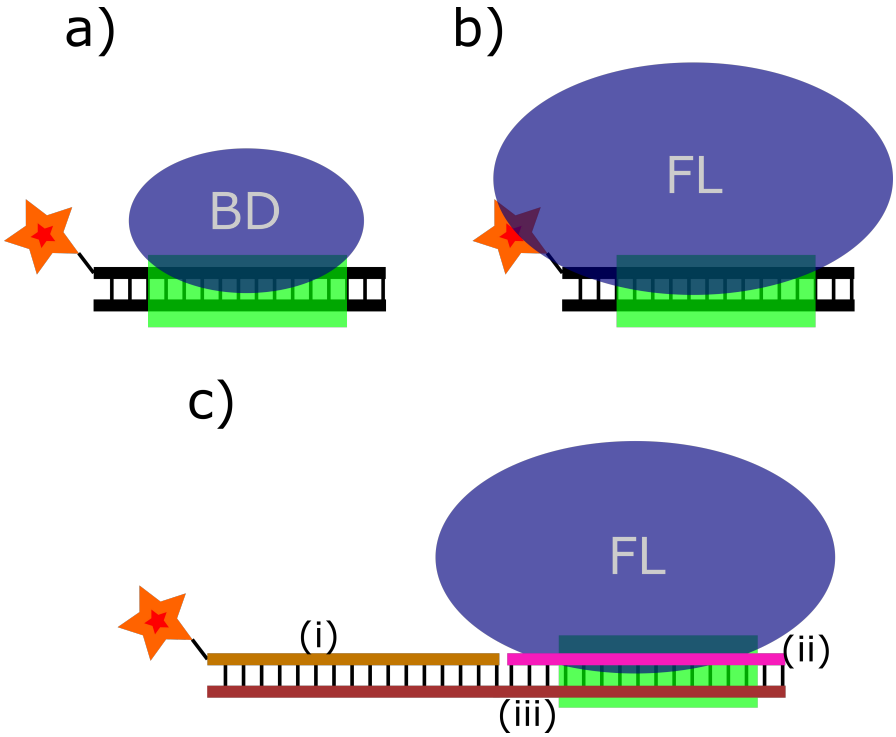


Figure 2.9: Schematic representation of the suggested mechanism of quenching and the modular longer oligomer to avoid interference of protein and dye. a) Situation with a small TF (BD: Binding domain), the fluorophore is not affected by the protein binding. b) Problematic binding of a bigger TF (FL: full length) to a small oligomer, the TF interferes with the dye upon physical proximity. c) Modular oligomer system to avoid interference between the TF and the fluorophore. The oligomer is assembled from three single stranded DNAs: (i) the fluorescently labelled DNA, depicted in light brown, (ii) the DNA specific for the TF of interest, depicted in magenta and (iii), the reverse DNA, complementary to DNAs i and ii.

2.4.1 Modular reference DNA system to determine full length TFs

In order to be able to measure certain FL proteins, the HiP-FA assay had to be modified. All data acquired speak for an FA quenching effect of the fluorescent dye at the end of the 16-mer DNA sequence (see Figure 2.9 a and b). This might be caused by changes in the electrostatic environment, potentially changing the fluorescence lifetime of the fluorophore which is inline with the observed loss in FA change along with a maintained comparable fluorescence intensity. The solution to avoid this was a modular DNA system (Figure 2.9 c). Here, three oligomers are allinged in a way that ensures distance between the TF's binding site and the fluorophore. This system allows for using the same fluorescently labelled DNA for many different TFs (excluding those with cryptic binding sites falling into the labelled DNA's sequence) at a lower cost (short fluorescently sequence, long unlabelled sequence). Besides the lower costs for the lablled oligomer, the synthesis time is also greatly reduced, making the system more convenient in general. In addition to increasing the length of the oligomer, the fluorescent dye was optimized, as well. The fluorescence properties of cyanine dyes, like Cy5 used in the original HiP-FA publication, depends on their electrostatic environment (Kretschy et al. (2016)) and can therefore be problematic when dealing with larger proteins. We therefore tested the influence of the dyes Dy649 and Bodypy650 and found the Bodipy dye to be superior to the previously used C5 (see Figure 2.10). This further increase in anisotropy change generally increases signal quality by a higher signal to noise ratio.

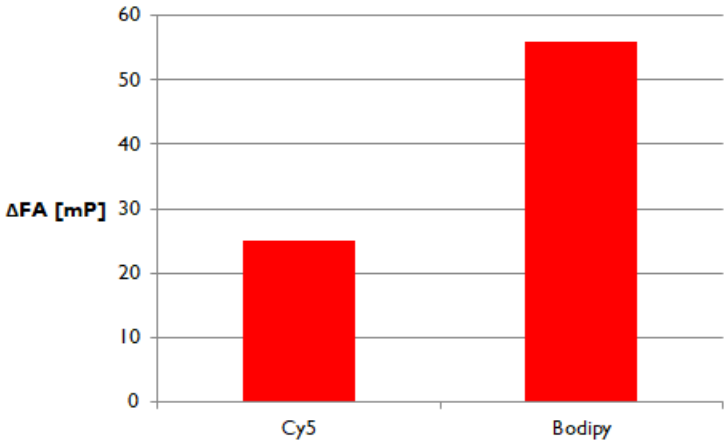


Figure 2.10: Comparison of the fluorescent dyes Cy5 and Bodipy 650. The figure shows the maximum change in FA for a longer modular oligo after the binding by the long construct of Zld. Biodypy leads to a stronger FA signal, improving signal to noise ratios.

2.4.2 Troubleshooting in determination of PWMs

Of the PWMs determined, Zld and Pan could not be measured using standard procedures. The special measures to determine their specificities are described in the following paragraphs.

Zld long The attempt was to express this protein as a full length version in *E. coli*, which is critical due to its high molecular weight of 170 kDa (̃200 inkluding GST). What encouraged us was the fact that the expressed protein showed binding activity in the HiP-FA assay together with the notion that the DNA binding domain is near the C-terminus of the protein (see Figure 2.12). It is therefore expected that any protein showing (specific) DNA binding activity was expressed at least until that point. Since the smallest fraction of the expressed and GST purified protein is of that size (Figure 2.14, lanes GST3 and 4), we wondered how to enrich for the (nearly) full length protein. We tried to enrich by binding activity using Zld's consensus DNA sequence as a bait. The principle of the developed functional purification is shown in Figure 2.13. The biotinylated DNA is bound to a streptavidin column which in turn can be bound by a functional TF. The whole complex can be eluted from the column by eluting the biotinylated DNA. Unexpectedly, the bands mostly enriched by this type of purification were of a molecular weight around 120 kDa, which is substantially smaller than the expected 200 kDa and not enough to include enough of the protein's N-terminal part. Since the input for this purification had been a GST purification (with the GST at the N-terminus of the protein) it is not possible that just complete C-terminal parts of the protein have been enriched. A possible explanation would be either the cloning of a splice variant or an internally, in frame truncated plasmid. Sequencing 3' (C-terminal equivalent) yields long reads into the coding sequence while sequencing from the 5' (after GST, N-terminal) produces only short ready, indicating that there might be a mixture of plasmids after purifing the plasmid from *E. coli* expression cells. To conclude, the measured protein was a truncated version of full length Zld, which might still contain useful information but is undefined. Further experiments on this matter should be conducted in a more complex expression system (insect *in vitro* or *in situ* expression).

Pan HiP-FA requires fast on- and off kinetics to produce reliable K_D values. This is necessary in order to have a binding equilibrium before the concentration of competitor changes significantly by diffusion. In the case of Pan-BD we measured binding saturation after approximately 15 min.

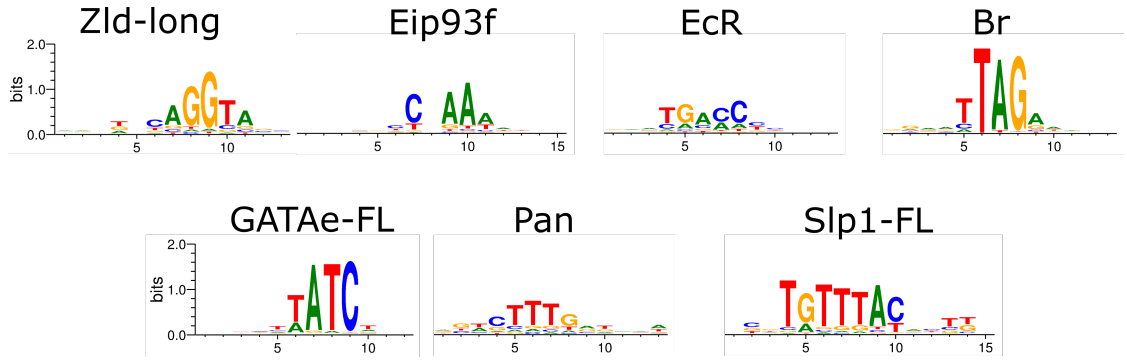


Figure 2.11: Sequence logos of PWMs measured in Jung et al. (2018), long: partial FL protein, FL: full length protein.

Figure 2.12: Coding sequence of Zld-FL with features annotated. One zinc finger (ZnF) is in the N-terminal half, the DNA binding domain consisting of four ZnFs are located near the C-terminal end, before a coiled coil domain.

To determine the PWM of Pan we therefore turned back to a classical titration approach, which we established in a 384 well plate on the robotics system in order to acquire accurate results under little protein consumption.

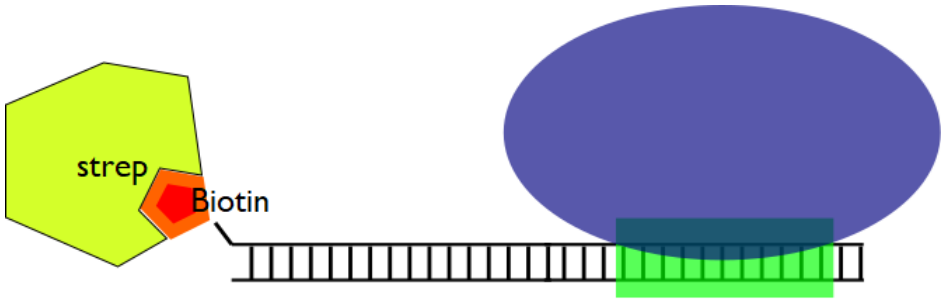


Figure 2.13: Sketch illustrating the principle of the functional purification. The biotinylated DNA is bound by a streptavidin-column (strep). Functional TF can bind to the DNA and be eluted together with it by weakening the interaction between biotin and steptavidin.

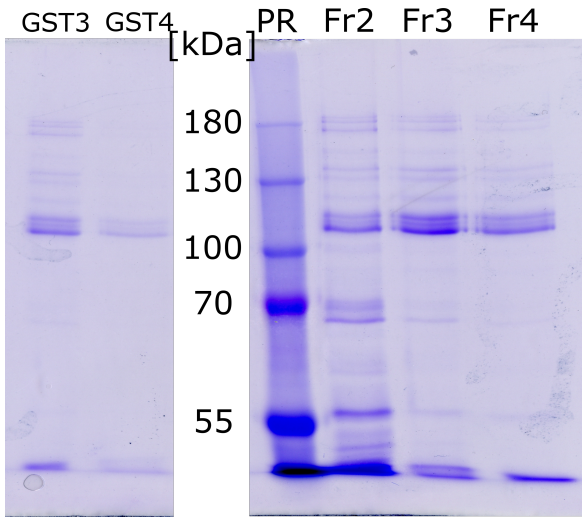


Figure 2.14: SDS-PAGE investigating the functional purification of Zld. Lanes: GST3&4, GST purification elution fractions, used as input for functional purification. PR: Page ruler, protein ladder, molecular weight in kDa at the left between the two halfs of the plot. Fr2-4: elution of the functional purification.

Chapter 3

Sensitive automated measurement of histone-DNA affinities in nucleosomes

The work presented in this chapter was published in *iScience* (see Schnepf et al. (2020)).

3.1 Introduction

Eukaryotes organize their genomes by wrapping their DNA around a protein complex of basic proteins called histones. Approximately 147 base pairs of DNA are wrapped 1.7 times around a histone octamer (composed of two copies each of H2A,H2B,H3 and H4) to form a nucleosome. The nucleosome core complex is the basic unit of chromatin, which can be separated into heterochromatin and euchromatin. While the former is tightly packed and biologically inactive to a large degree, the latter is strongly regulated by its positioning and binding strength to the DNA (Khorasanizadeh (2004)). Nucleosomes appear well positioned with respect to the DNA sequence especially around promoters in a regularly spaced pattern with a nucleosome depleted region at the transcription start site. The best positioned nucleosomes are upstream and downstream of this depleted region, commonly referred to as -1 and +1 nucleosomes, respectively (Lai and Pugh (2017)). It is commonly accepted that multiple factors like DNA sequence, nucleosome remodelers, transcription factors, and the RNA polymerase II transcription machinery play an important role in the positioning of nucleosomes in vivo. The degree to which the underlying sequence dictates nucleosome positioning is however still under debate (Zhang et al. (2009); Kaplan et al. (2010); Jin et al. (2018)). It has been shown that nucleosomes do have sequence preferences, which can be used to predict their positions within a genome (Segal et al. (2006) Tillo et al. (2010)). The best studied parameters determining nucleosome favored sequences include the GC content (Tillo and Hughes (2009) Fenouil et al. (2012)) and the base pairs (bp) periodicity of flexible nucleotides matching the contact sites frequency of the histone octamer with its nucleosomal DNA (Shrader and Crothers (1990); Jin et al. (2016) Klug and Lutter (1981) Drew and Calladine (1987)). Due to the double helical nature of DNA, the same face contacts the histones every ten to eleven base pairs which favors the periodical reoccurring of alternating dinucleotides that are easier or harder to deform - according to Wang et al. (2010), "deformation energy of DNA increases in the order $TA < CA < CG < GC < AC < AT$". In addition, it is known that the presence of short homopolymeric stretches of deoxyadenosine nucleotides referred as poly(dA:dT) or dA:dT tracks, intrinsically stiff, is inhibiting for nucleosome formation Segal and Widom (2009b) Raveh-Sadka et al. (2012)Jin et al. (2018). While many efforts have been devoted in the past years on characterizing nucleosome

sequence preferences in vitro (Krietenstein et al. (2012b), Segal et al. (2006)) and in vivo (Jin et al. (2018) Kaplan et al. (2010)), most studies were based on frequency counts of DNA sequences identified by deep-sequencing methods instead of true histone-DNA affinity measurements, mainly due to the lack of suitable experimental techniques.. The most widely used method for determining histone-DNA binding free energy in vitro was pioneered by Schrader, Crothers and Widom (Shrader and Crothers (1990),Lowary and Widom (1998)). In this assay, nucleosomes are typically reconstituted from purified core histones and mononucleosomal DNA by dialysis, or alternatively by a stepwise dilution method. A competition assay is used, which is based on reconstituting a mixture containing a DNA of interest in excess with (usually radio- or fluorescently-) labelled DNA, which serves as reference to compare the nucleosome-forming ability of different DNA sequences. Reconstituted samples are then analyzed on polyacrylamide or agarose gels by Electrophoretic Mobility Shift Assay (EMSA) (Thastrom et al. (1999)) to calculate the fraction of reference DNA that reconstitutes into nucleosomes in a given DNA fragment composition. This allows determining relative affinities (free energies) of histone octamer to differing DNA fragments. However, to the best of our knowledge, affinity data are only available for a relatively limited number of sequences (Shrader and Crothers (1990) Lowary and Widom (1998);Thastrom et al. (1999) Thastrom et al. (2004) Takasuka and Stein (2010) Cao et al. (1998) Filesi et al. (2000)). Furthermore, most studies focused on artificially designed nucleosomal DNA sequences with very strong non-physiological binding properties, whereas relevant genomic sequences have not been investigated extensively, probably because the binding strengths of genomic sequences are too similar to be well resolved with existing methods. Thus, there is a clear need for more accurate and comprehensive measurements of histone-DNA binding free energies. This lack of data can be explained by the limitations of the classical approach: both nucleosome reconstruction and EMSA readout steps are time consuming and difficult to parallelize. As a consequence, the histone free energies are usually determined using only a single concentration per sequence. The use of only one data point per affinity measurement rises accuracy and reproducibility issues for the obtained binding free energies. To overcome these limitations, we devised a method to measure histone-DNA free energy of nucleosome formation with high reproducibility and at large scale . Our technique recapitulates the classical approach, but integrates substantial improvements: (1) we carry out the competitive reconstitution of nucleosomes by small dilution steps using an automated system, which provides high reproducibility and parallelization; (2) we determine the fraction of bound DNA by means of Fluorescence Anisotropy (FA), which is measured with an epifluorescence microscope adapted as described in section 1.7 for High-Performance Fluorescence Anisotropy (HiP-FA), a new approach we recently reported for measuring transcription factor-DNA binding energies. The high parallelization of the nucleosome reconstruction and the fast and sensitive fluorescence readout allowed us to obtain full titration curves for each individual histone-DNA interaction instead of single concentration measurement, resulting in more accurate determination of the binding free energy. Here, we introduce this new method and demonstrate its applicability and capabilities by measuring histone-DNA binding free energies for 47 different DNA sequences (147 to 300 bp in length) from *Drosophila melanogaster* (*D. mel*) genomic nucleosomes (29 sequences; denoted Dmel01 - Dmel29), and from synthetic DNA constructs derived from D. mel. enhancers (11 sequences; Synt01 - Synt11) designed to test the effect of DNA structural features. To validate our assay, we additionally selected 7 nucleosomal DNA sequences measured in other works (601, TGGA-2, TAND-1, TG, Bombyx, 5S, and TG-T Shrader and Crothers (1990) Cao et al. (1998) Filesi et al. (2000) Thastrom et al. (1999)) and compared their binding energies to ours. We show that the free energies of nucleosome formation for all DNA sequences could be measured accurately and cover a wide dynamical range. Furthermore, we took advantage of the throughput of our method to explore how free energies correlate with DNA features such as the GC content, the 10 bp periodicity of flexible and stiff dinuleotides, and the number of short poly(dA:dT) stretches. We found GC content to be the most important feature in our data, alone explaining 30 percent of the variation of the free energies.

3.2 Results

3.2.1 Pre-experiments

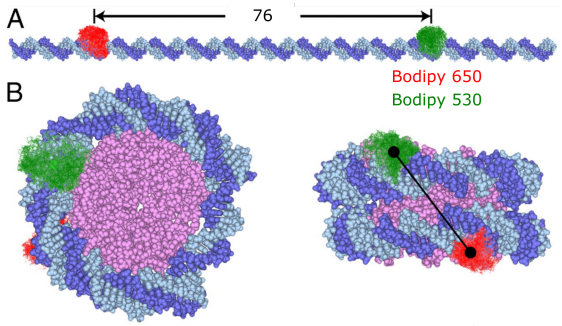


Figure 3.1: Sketch of FRET. A) shows the free DNA not incorporated into a nucleosome, the two Bodipy dyes are far appart. B) shows the formed nucleosome. The dyes come in closer proximity and allow for FRET to happen. The figure is modified from Gansen et al. (2009), distances are not to scale. Permission to reuse granted 09.01.2020

FRET measurements To ensure the identity of nucleosomes, a bulk measurement of a nucleosome reconstitution using a twice fluorescently labelled DNA and histones was performed determining Förster-Resonance-Energy-Transfer (FRET) efficiency using a commercial confocal microscope. FRET is radiation-free transfer of the excitation energy of a donor fluorophore to an acceptor fluorophore with the efficiency dependent on the distance between them. The the readout (FRET efficiency) is determined by the fluorescence intensity at the emission wavelength of the acceptor. The fluorescent dyes (BODIPY530 and BODIPY650) were introduced into the strong nucleosome forming sequence 601 (of the minimal length of 147 bp) at positions in which they are close (in FRET distance) when incorporated in a nucleosome and distant in the DNA free in solution (compare also Figure 3.1) (like performed in Lee et al. (2015)). Fig. 3.2 shows, that the FRET efficiency in the nucleosome formation is increased by a factor of 2.8 compared to the DNA in free solution. It is noteworthy that in this result that was obtained from early experiments, the ratio of nucleosome to free DNA was lower compared to the final nucleosome formation protocol according to EMSA data.

Low adhesion consumables An important factor for developing an automated nucleosome formation assay is the correct type of consumables. We determined the suitability in a semi-quantitative fashion scoring both fluorescence intensity loss and reproducibility comparing both replicates within a single experiment and between experiments. Tested were: non-coated reaction tubes (Sarstedt), protein low-binding reation tubes (Sarstedt), protein low-binding plates (Eurofins) and glass vials (Fisherbrand). The protein low-binding plates by Eurofins ans the Sarstedt protein low-binding tubes showed low overall adhesion but only the latter also provided a high reproducibility.

3.2.2 Automated assay to determine free energies of nucleosome formation

Assay presentation We determined the free energies of nucleosome formation by titration of unlabeled nucleosomal DNA sequences competing for nucleosome reconstruction with a fluorescently labeled DNA reference in low amount (Fig.3.3 a). Histones and competing DNAs were

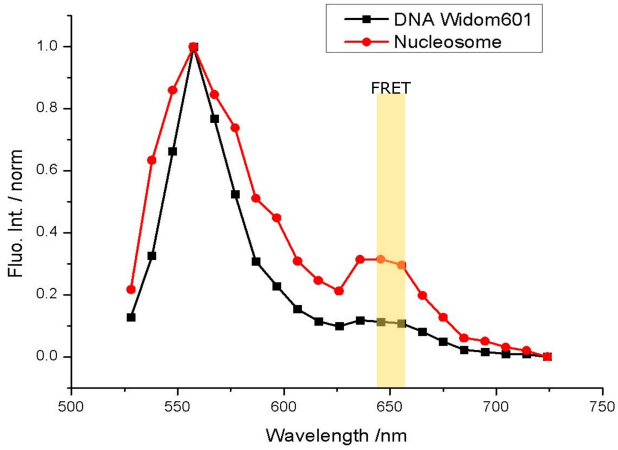


Figure 3.2: The figure shows two fluorescence profiles, one of the DNA alone and one of the nucleosome reconstitution product. The excitation is at 514 nm, if no FRET occurs, the emission is expected to happen at 551 nm. Excitation around 660 nm is an indicator of FRET (yellow bar).

initially mixed at a high salt concentration, and their interaction strength was progressively increased during a slow dilution process by small steps of buffer addition. The use of an automated system allowed carrying out nucleosome reconstruction over a long period of time (12 h, typically overnight), ensuring thermodynamic equilibrium, and thereby greatly improving reproducibility and throughput of our approach. It was also essential for the reproducibility of the reconstruction reaction to ensure limited evaporation (which typically occurs at borders and edges of well plates containers) and a stable temperature . To this purpose, we designed and fabricated a custom metal block accommodating up to 42 individual low protein binding tubes with a heated lid (Fig.3.3 b, appendix 4.1 and Materials and Methods). The metal block ensured temperature stability during the nucleosome formation process, while the heated lid prevented condensation.

The readout of the fraction of bound vs unbound fluorescently labeled reference DNA was carried out by using FA Roehrl et al. (2004) (Fig.3.3 c) instead of the typically used EMSA, providing the advantages of a fast and sensitive fluorescence readout. Different FA levels for DNA embeded in a nucleosome complex versus free DNA (Fig. 1.13) can be used to calculate their corresponding fractions in a titration series. After nucleosome reconstructions, we transferred the samples in 96 well microscopy plates, and measured FA in each well using the microscopy setup implemented in the HiP-FA method (Jung et al. (2018) and Fig.3.3 c). By performing salt titrations with different unlabeled competitor concentrations we obtained a full titration curve for each nucleosomal DNA sequence. The data could be fit using the Hill equation, as can be seen for a weak ($\Delta\Delta$G = 9.7 kJ$\cdot mol^{-1}$; DmeI08), a medium ($\Delta\Delta$G = 7.2 kJ$\cdot mol^{-1}$; DmeI28), and a strong ($\Delta\Delta$G= -2.2 kJ$\cdot mol^{-1}$; 601) competitor sequence (Fig.1d). To validate our assay, we included in our measurements 7 nucleosomal sequences measured in other works (Shrader and Crothers (1990) Cao et al. (1998) Filesi et al. (2000) Thastrom et al. (1999) - (Fig.3.3 e)). The free energies for nucleosome formation of these sequences were calculated relative to their respective reference sequences and were in good agreement with our measurements (free energies relative to their corresponding references are plotted in Fig.3.3 e for three different pairs of sequences). As an additional mean of validation of our FA approach, we also measured affinities by EMSA and observed a good agreement between these measurements and FA derived affinities. (3.4). The discrepancy between the absolute values could be explained as a result of the quenching effect happening in the nucleosome band of the EMSA gel (probably besides sensitivity a major reason why radioactivity is used in these EMSA experiments). Since anisotropy doesn't depend to the

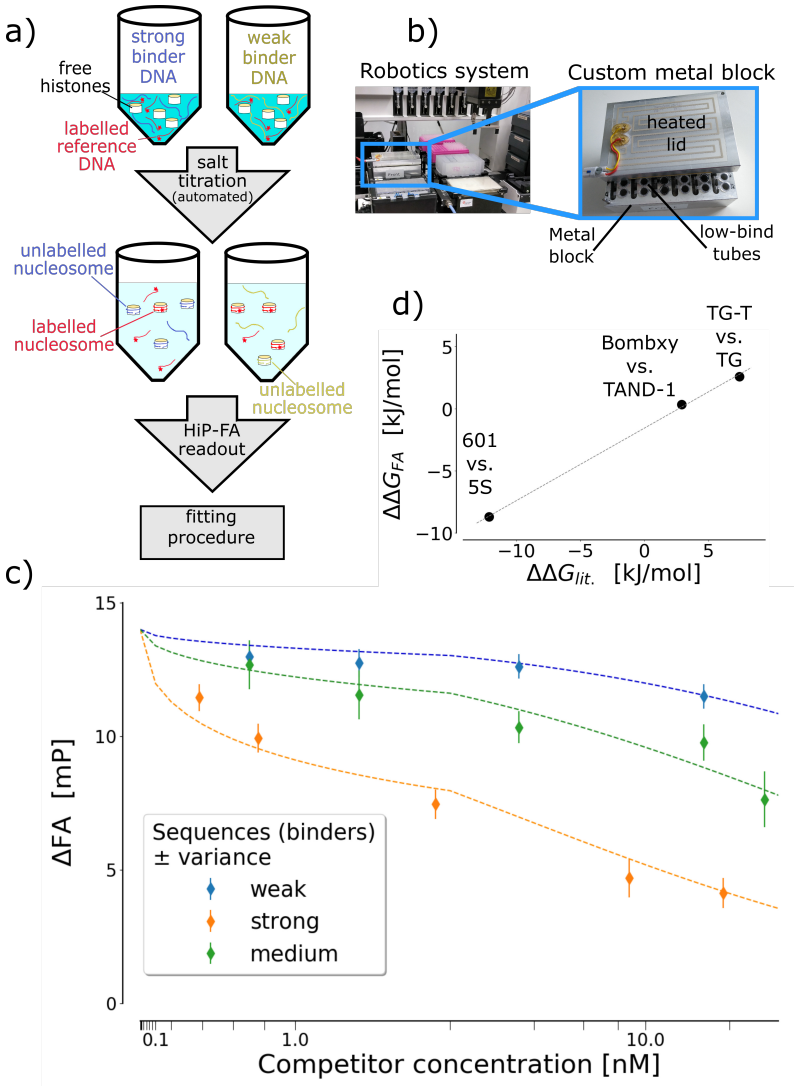


Figure 3.3: Competitive nucleosome reconstruction and affinity assay. (a) Schematic representation of the assays procedure. (b) Pictures showing the robotic system used for nucleosome reconstruction. The zoom-in shows the custom metal block with its heated lid. (c) Schematic drawing of the fluorescence microscopy setup used for the FA readout. (d) Exemplary histone-DNA affinity titration curves for 3 different competitor sequences together with their corresponding fits (in dashed lines) for a weak (Dmel08,in blue), a medium (Dmel28,in green), and a strong (601,in orange) binder, respectively. (e) Assay validation with DNA sequences measured in previous works Shrader and Crothers (1990) Cao et al. (1998) Filesi et al. (2000) Thastrom et al. (1999). The relative free energies determined in previous measurements are plotted against the corresponding values obtained in this study; dotted line shows linear regression; Pearson correlation coefficient R= 0.99.

same degree on the fluorescence intensity as the EMSA readout does, this is not an issue for the here established assay.

Applying the method for measuring affinities of genomic and synthetic nucleosomal DNA sequences The free energies of nucleosome formation obtained in most in vitro studies

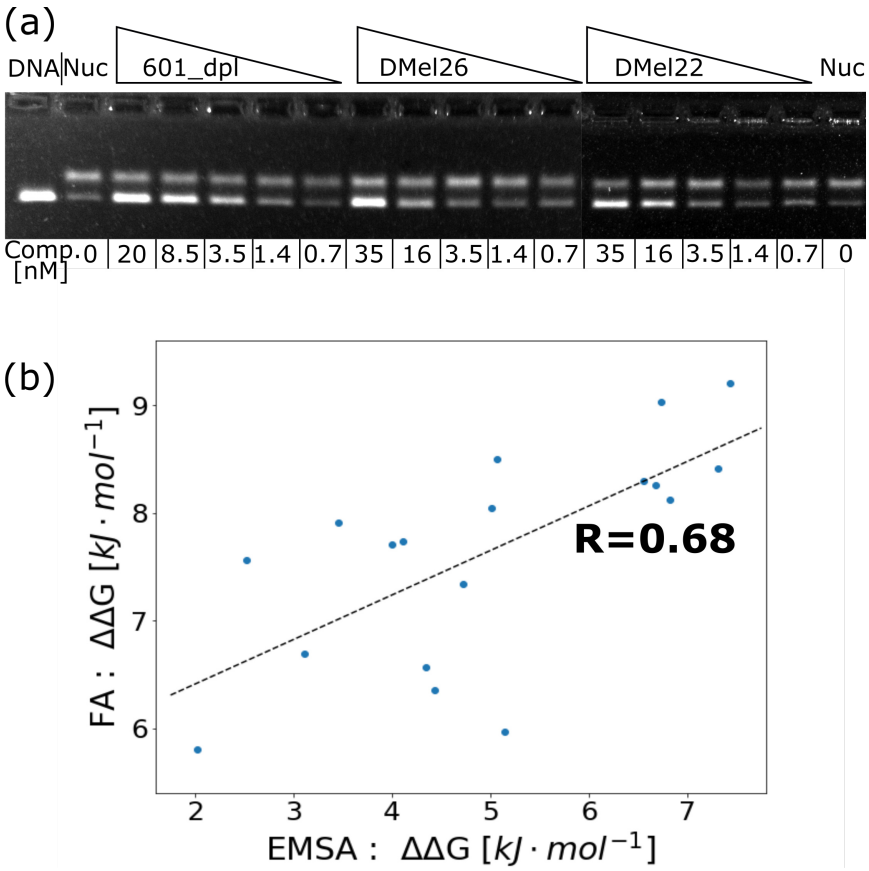


Figure 3.4: Competitive titrations using EMSA as readout. (a) Three competitor titration curves analyzed using EMSA. The lanes are labeled with the sample names, DNA being the fluorescently labeled 601_dpl DNA, Nuc being the reference sequence and histone octamer without competitor DNA sequence. 601_dpl, Dmel26 and Dmel22 denote the competitor sequences used for the titration. The triangles symbolize a decreasing concentration of competitor DNA, the concentration in nM is indicated at bottom. (b) Comparison of the nucleosome formation free energies obtained by EMSA and FA. The free energies obtained from 18 titrations with different competitor DNA sequences using FA as readout are plotted against the free energies obtained with the same samples, but using EMSA as readout. The dashed line shows a linear regression, R= 0.68.

focused on nucleosomal DNA sequences by essence selected or designed to cover a large range of affinities. The strongest known binders (like the well-known 601 sequence), however, are not found in endogenous genomic DNA. Although it is known that histone-DNA free energies of naturally occurring nucleosomes cover a much lower dynamical range than for synthetic sequences (Thastrom et al. (1999)), to our knowledge there is no exhaustive direct measurements of affinities for individual genomic nucleosomal sequences. Taking advantage of the throughput and accuracy of our technique, we turned to the *D. mel* model organism and determined the free energies in nucleosome formation for two groups of DNA sequences: the first group contains 29 endogenous nucleosomal DNA sequences (termed Dmel01-DMel29) determined by MNase-Seq (unpublished data). We focused on the -1 nucleosomes, since although the +1 and -1 nucleosomes are both enriched for positioning sequences, the +1 nucleosomes are however in addition enriched for cis-regulatory elements, suggesting that their positioning is more strongly influenced by biological activity (Mavrich et al. (2008)). The sequences were chosen randomly, provided that their GC contents span from 20 to 60 percent. The second group of tested sequences includes 11 synthetic constructs (termed

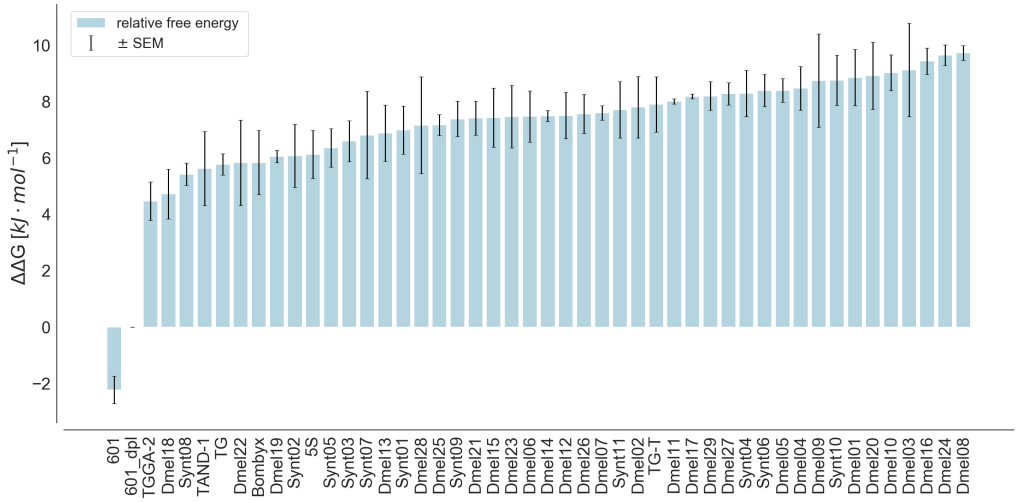


Figure 3.5: Overview of histone-DNA affinities. All affinities are given as free energy of nucleosome formation relative to the 601_dpl reference sequence and were calculated by the mean of, on average, three replicates. Error bars indicate standard error of the mean. The names are either kept from the original publications or indicate the different groups of investigated DNA sequences. Dmel: called -1 nucleosomes from *D. mel*, Synt: synthetic enhancers constructs driving expression during *D. mel* embryo development.

Synth01 Synth11) derived from enhancer sequences driving expression in *D. mel.* Embryos (used in another study currently in preparation). In brief, these synthetic sequences are highly similar (differing mostly only in less than 10 percent of their bases) (Fig.3.9 a). We however observed strong differences in their expression patterns in living embryos, that we partially attributed to the influence of binding to nucleosomes (unpublished data). These sequences were used in the present work to test the influence of short oligo dA:dT tracks on histone-DNA affinity by mutating (A/T rich) transcription factor sites. In total, we obtained -after quality control (see Materials and Methods)- 147 titrations of 47 different DNA sequences (three measurements per sequence on average; Fig. 3.5). All free energies ($\Delta\Delta$G) are determined relative to our reference sequence, a mutated version of the 601 sequence (601_dpl) that was designed to exhibit a lower affinity to histone-octamers. This sequence is well suited to measure accurately weaker competitor sequences (601_dpl maintains the same GC content as the 601 sequence, but with less pronounced 10 bp dinucleotide periodicity patterns). Overall, our measured free energies span over 12 kJmolmol^{-1} (Fig.3.3), a range similar to what was reported for comparable sequences (Thastrom et al. (1999) Thastrom et al. (1999)). The reproducibility is high with a mean coefficient of variation (CV) of 24 percent. Interestingly, except for the 601 and its weakened derivative 601_dpl, the free energies are all distributed over a much smaller free energy range of 5.3 kJ$\cdot mol^{-1}$. The synthetic enhancer sequences have generally slightly, but not significantly, lower free energies of nucleosome formation than the -1 nucleosome sequences.

We furthermore compared the free energies obtained in this study with predictions based on nucleosome PWMs. As the PWMs determined by Segal et al. (2006) were not available while writing this, we turned to the work by Heron (2017), who determined PWMs based both on chemical cleavage and DNaseI data. As Figure 3.8 shows, our data is in much better agreement with the prediction based on chemical cleavage (a, R^2=067) than the one based on MNase data (b, R^2 =0.12). Reasons for this relatively good agreement with the first PWM and the rather strong disagreement with the latter might be cutting bias in the enzymatic assay as well as genomic effects in the latter data set.

Dependency to GC-content and other DNA features To gain insights into the factors contributing to histone-DNA interactions, we correlated our data with several known sequence determinants of nucleosome formation, namely GC content, sequence features affecting bending: the 10 bp periodicities of flexible (WW where W is A or T) and stiff (SS where S is G or C) dinucleotides, and the presence of homopolymeric sequences poly (dA:dT). We first plotted the free

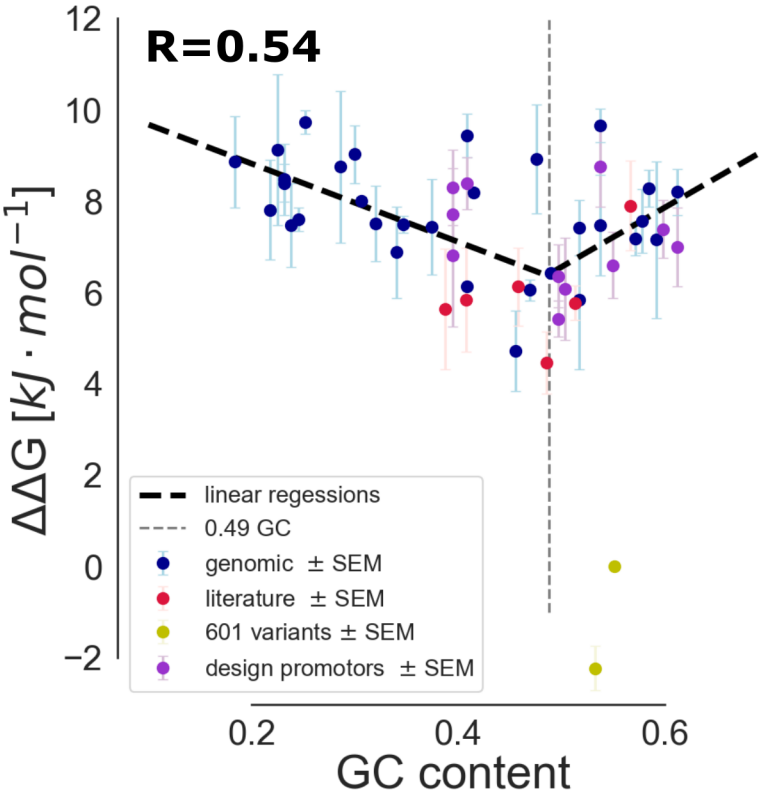


Figure 3.6: Correlation between free energy of nucleosome formation and GC content. The relative free energy is plotted against the GC content for all the investigated nucleosomal sequences. The data points are color-coded according to their sequence (the 601 variants appear as separated yellow dots). A combined linear regression was performed in sections for sequences with GC contents $<$ or $>$ 0.49, respectively. The correlation between free energy of nucleosome formation predicted by GC content and observes was 0.54 (Pearson correlation coefficient).

energy of nucleosome formation as a function of their GC content for all investigated sequences (Fig.3.6). We observed a decrease of $\Delta\Delta$G (i.e. increased affinity) with GC contents for GC contents values $<$ 0.5, and then an inverted trend at higher GC contents with elevated $\Delta\Delta$Gs. Hence, these data show that binding free energy is strongly influenced by GC content and there seems to be an optimal GC content value for binding of 0.5. Interestingly, sequences from all three groups (Fig.3.6) seem to follow this falling and rising pattern (with the exception of the two 601 variants; excluded from the following analysis). The simplest model to describe the behavior of $\Delta\Delta$G with respect to GC content is given by a segmented linear regression with two segments, intersecting at the optimal GC content. We fitted the data with this model, and found a relatively good correlation with Pearson correlation coefficient (R) of 0.54 and an optimal GC content value of 0.49 corresponding to the minimum $\Delta\Delta$G value of 6.4 kJ$\cdot mol^{-1}$. Hence, the GC content alone explains 30 percent (R) of the variance in the data. This value is in agreement with data

reported by another study (Fenouil et al. (2012)) examining nucleosome GC content preferences of +1 nucleosomes, with the presence of a maxima for nucleosome occupancy at GC contents of 0.44 (*in vivo*) and 0.58 (*in vitro*), respectively. Although this study have been performed in a different model system (human cell lines) and with different methods (ChIP-Seq), it confirms the presence of an optimal GC content, suggesting that the strongly positioned nucleosomes at the transcription start site are influenced by GC content to a large degree. Another study found the highest occupancy of nucleosomes in exons also to peak around a GC content of 0.5 Wang et al. (2014).

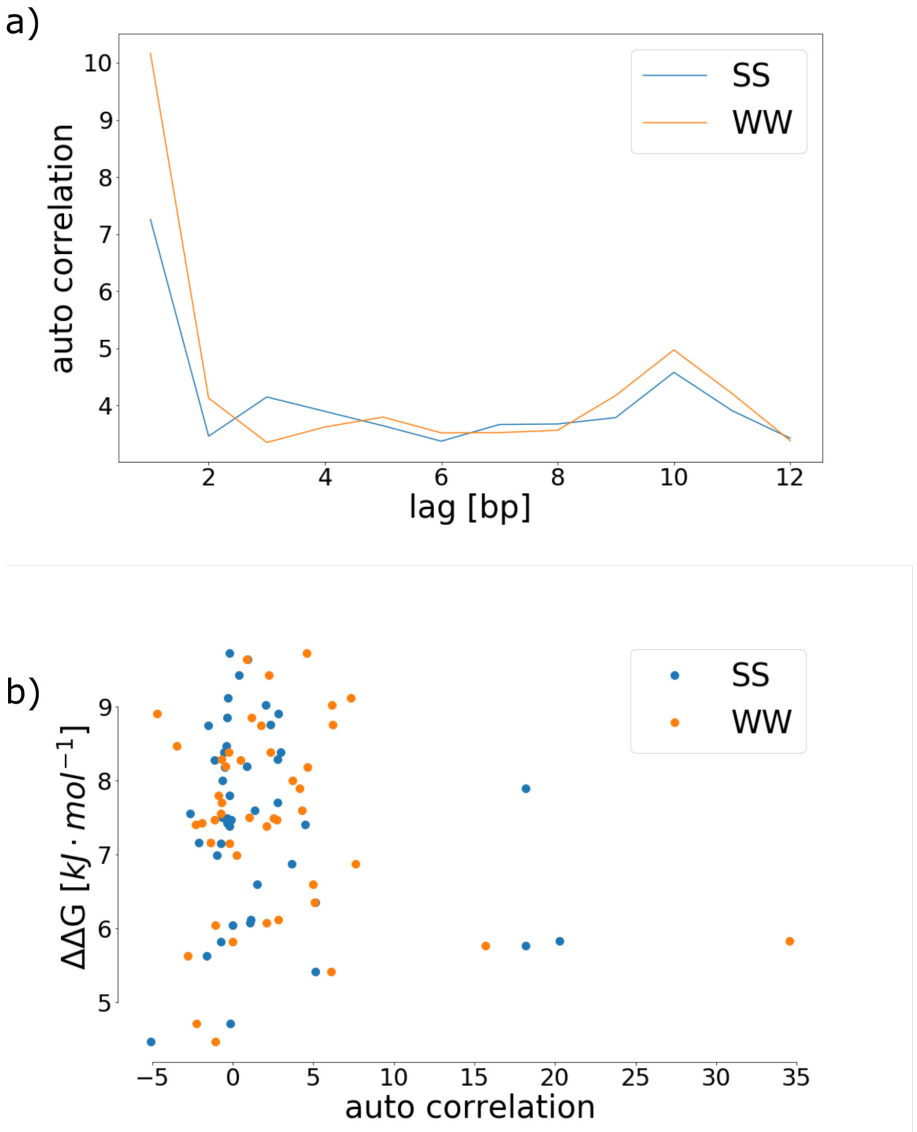


Figure 3.7: Autocorrelation analysis of dinucleotides. (a) Average autocorrelation of SS (S= G or C) or WW(W=A or T) dinucleotides by their lag (in bp). Both curves show a weak peak at 10 bp. (b) Relationship between the free energy of nucleosome formation and the corresponding autocorrelation at a shift of 10 bp. No significant correlation can be observed for our sequences (R= -0.15 and -0.16 for WW and SS, respectively). The extreme points (autocorrelation>15) originate from intrinsically strong periodic sequences (Bombyx, TG, TG-T).

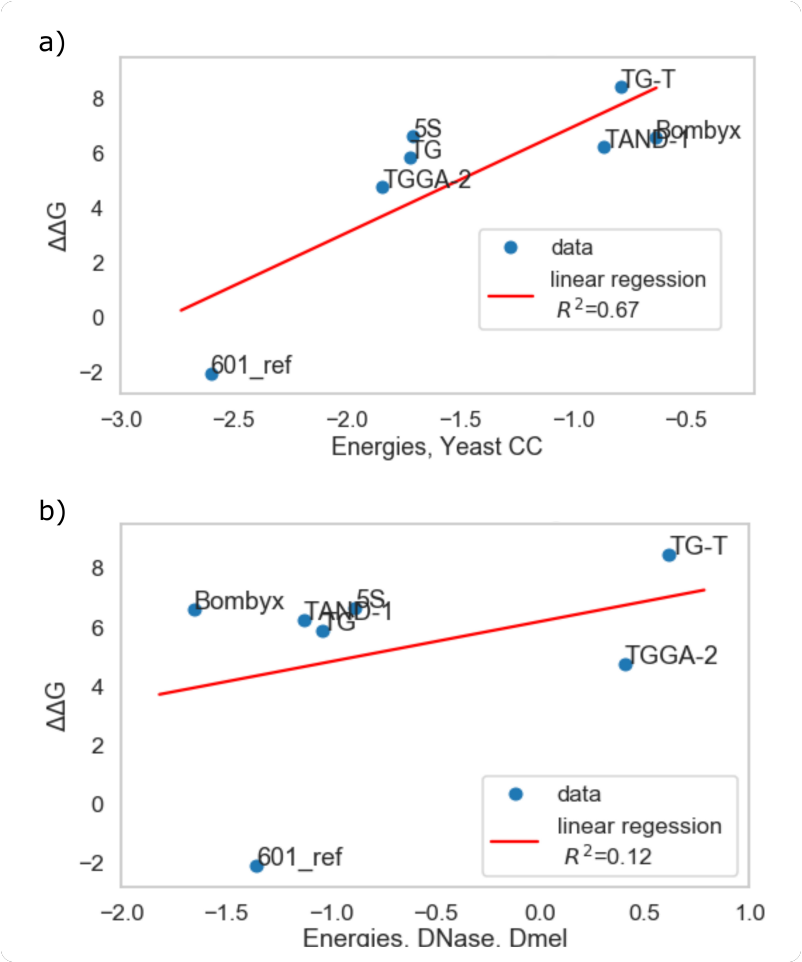


Figure 3.8: Comparison of obtained data with predicted data based on nucleosome PWMs determined in Heron (2017). Plotted are the free energies for artificial sequences obtained in this work against the energies predicted based on a chemical cleavage (a) or on a DNaseI experiment (b), respectively. Linear regression are depicted in red, all data points are labelled with sequence names. Squared pearson correlation (R^2) are in given in the figure legend.

As mentioned, the two variants of the 601 sequence do not follow the general trend we observed as function of GC content (Fig.3.6). However, this singular behavior is not surprising as the 601 clone was specifically engineered to harbor several sequence features increasing nucleosome binding (Lowary and Widom (1998)). A prominent feature that might underlie the singular behavior of these artificially strong binders is the presence of a clear periodicity in their sequence. In fact, the periodic occurrence of flexible and rigid sequence patterns, which align with the periodic changes in orientation of the DNA double strand facing the histones, is predicted to favor nucleosome binding in a the kink-and-slide model (Vasudevan et al. (2010)). We therefore wondered to what extend the dinucleotide periodicity influences the free energy of nucleosome formation for the pool of the other measured sequences (601 variants excluded). To this end, we started by computing the 10 bp periodicity of WW (W equals A or T) and SS (S equals C or G) dinucleotides using an autocorrelation function (Figs. 3.7) a+b and Materials and Methods , Cui and Zhurkin (2010)).Surprisingly, we found for both dinucleotide groups poor averages 10 bp autocorrelation factors derived for all sequences. We neither saw any strong 10 bp periodicity using Fourier transform , the alternative commonly used method. A closer inspection of the individual autocorrelation plots (Fig.3.7

b) revealed that the overall slight 10 bp autocorrelation observed was driven by few sequences, namely those that were already used in other studies (beside 601: Bombyx, TG, or TG-T, which by their nature are harboring strong 10 bp dinucleotide periodicities), whereas most of the genomic -1 nucleosome and synthetic enhancer sequences exhibited very low 10 bp autocorrelation factor. We analyzed the influence of this feature on free energy of nucleosome formation. To this end, we plotted $\Delta\Delta$G vs autocorrelation (Fig. 3.7 b), and observed no correlation in the data (R= -0.15 and -0.16 for WW and SS, respectively. Poly(dA:dT) tracks are believed to generally impede nucleosome formation (Segal and Widom (2009b), Raveh-Sadka et al. (2012) and Stanojevic et al. (1989)). The sequences investigated in this study show too few occurrences of sufficiently long dA:dT tracks (length>4 : only 53 percent of the sequences contain at least one 5 mer, 23 percent contain 6 mers) to make a reliable statement about their influence on nucleosome binding energy (Fig. 3.9). When varying the minimal track length (2-6), we find no significant correlations between the total number of nucleotides contained in the dA:dT tracks with measured G values (Fig. 3.9 a). However, four sequences (Synt02, Synt11, Synt04, and Synt06, in the following termed I, II, II and IV, respectively, for simplicity) from the group of synthetic enhancers constitute an important exception (Fig. 3.9 b+c). The synthetic enhancer sequences were specifically designed to investigate the influence of certain configurations of transcription factor binding sites in the *D. mel.* segmentation network, and, despite sharing most of their sequence, lead to strong differences in expression. One of the transcription factors of interest (Hunchback, Hb) possesses the consensus binding site AAAAAA (Fig. 3.9 c) (Stanojevic et al. (1989)). Whereas enhancer I contains no Hb binding sites, III contains three consensus sequences (in green in Fig. 3.9 b), and IV contains the same three sites, but with a different background sequence (yellow) and an additional AAA track at the beginning. Finally, II differs from III only by single point mutations in the 3 Hb binding sites (AAAAAA > AAATAA). For these four systematically mutated sequences, the free energy of nucleosome formation increases with the total number of nucleotides contained in the dA:dT tracks (Fig. 3.9c); R= 0.999 and R= 0.92 for minimal track lengths of 2 and 3, respectively (there were too few longer tracks for statistics). Although measurable, the influence of these features on $\Delta\Delta$G is rather weak (total range of 2 kJ$\cdot mol^{-1}$), and it might be completely obscured in cases where there are larger sequence differences, as in the majority of our investigated sequences (Fig. 3.9 a).

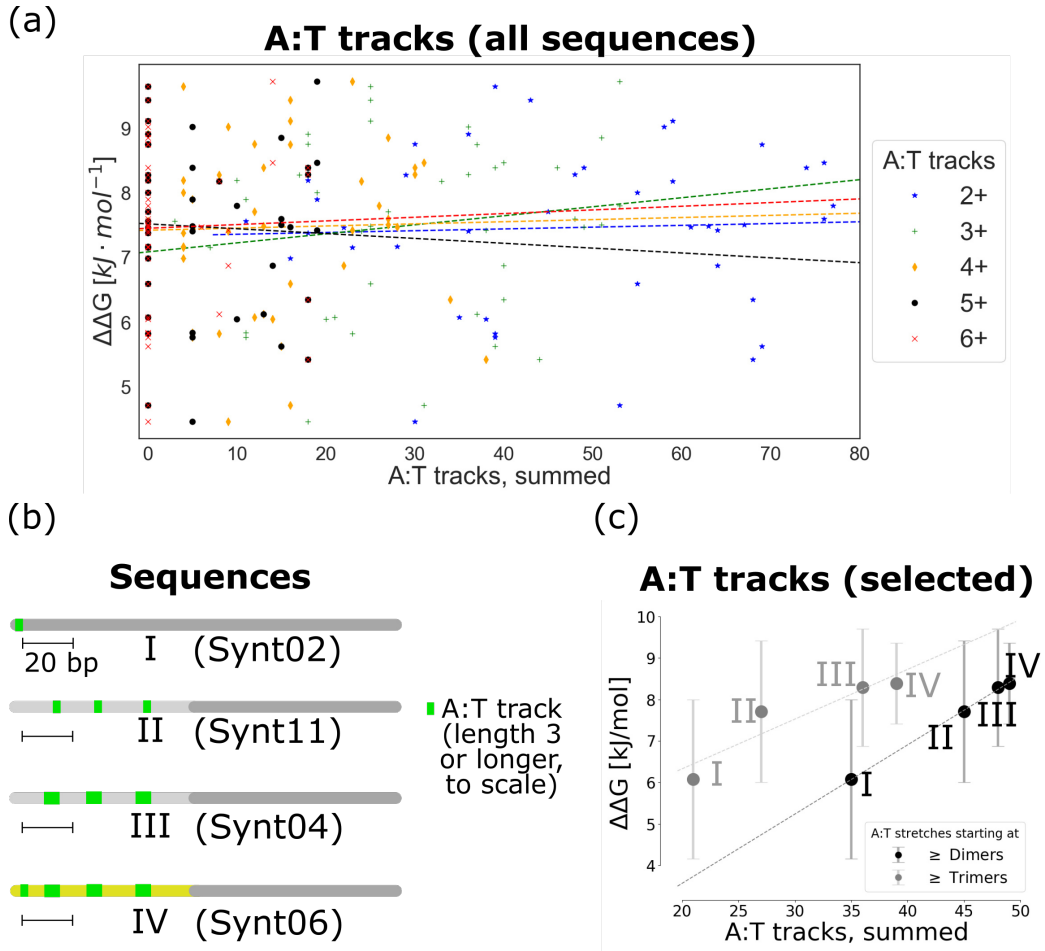


Figure 3.9: (a) Schematic representation of four synthetic enhancer sequences (Synth02,Synth11,Synth04 and Synth06) with systematically changed dA:dT tracks (dA:dT tracks differing between the sequences highlighted in green, length to scale). Yellow denotes a different background sequence. The sequence names are given in roman letters I to IV for reasons of simplicity. (b) Analysis of the influence of DA:DT tracks on free energy of nucleosome formation. G values are plotted against the summed dA:dT tracks length, weighted from a certain minimal length (starting at 2bp blue stars- up to tracks starting at length 6 - red x). The dashed lines represent linear regression with non significant Pearson correlation coefficients ranging from R =-0.04 to R=0.17. c) Free energy of nucleosome formation versus the total number of nucleotides in dA:dT tracks in selected synthetic enhancers (see a) for tracks of minimal length 2 (in black) and minimal length 3 (in grey); linear regression with dashed lines. All sequences would have expected to have very similar G values based on their similar GC content solely.

3.3 Discussion

We developed an automated assay to determine histone-DNA affinity using salt titration and fluorescence anisotropy microscopy. The assay aims at forming nucleosomes at conditions close to equilibrium by dilution over a long period of time and small changes in salt concentration per dilution step. The high degree of automation allows to carry out full titration series, a feature commonly absent in other competitive nucleosome formation assays. This provides a more robust determination of the free energies since the analysis relies on a fitting procedure over a larger dataset, and it allows introducing a quality control step for the detection of experimental errors (see Materials and Methods). In contrast to frequency count based assays, we directly measure the binding affinity of a given sequence without taking the effect of positioning, and thereby neighboring sequences, into account. Our measurements are therefore closer to nucleosome occupancy rather than positioning, however not necessarily identical. Determining the pure affinity of a histone-octamer binding to a given DNA sequence presents the advantage of being a defined measurement, whereas it is not totally clear which biophysical property is exactly captured when determining the occupancy in a frequency based nucleosome assay. However, compared to sequencing-based assays, our method requires some prior knowledge and selection for the investigated sequences, and has intrinsically a much more limited throughput. Additionally, genomic assays intrinsically analyze longer stretches of DNA and are more certain to capture an equilibrium state, although it might be less clear how strongly this state is determined by binding energy or other factors. Although our assay can capture affinities ranging over more than 2.5 orders of magnitude (in a non-log scale), we obtain a relatively limited dynamical range of only one order of magnitude for the free energy of nucleosome formation of the genomic and synthetic enhancer sequences (without taking the 601 variants into account; 3.5). This limited dynamical range is consistent with what was observed in previous studies Thastrom et al. (1999). Over the past years, it has become clear that nucleosome sequence preference is important for nucleosome positioning in vivo Kaplan et al. (2010), leading to substantial biological effects. Our results show that such effects could be caused by relatively small variations in absolute affinities. We found GC content to be the main feature determining the free energies of nucleosome formation for the DNA sequences investigated, namely for the -1 nucleosome genomic sequences randomly selected and explaining 30 percents of the variance in the data (Fig.3). Although this finding is in agreement with other studies (Fenouil et al. (2012) Tillo and Hughes (2009)), it was nevertheless debated and often attributed to GC content bias of MNase (Jin et al. (2018)). While some studies find a monotonous trend with GC content (Tillo et al. (2010)), others support the idea of an optimum driven GC content (Fenouil et al. (2012) and Wang et al. (2014)). The reduced nucleosome affinity for high GC content was also reported in CpG promoters in a *in vivo* study, and was attributed to biological activity (Tillo et al. (2010)). Since our method is not biased by GC content by design, we are able to confirm the strong influence of GC content on histone-DNA binding interaction. Numerous studies have focused on dinucleotide periodicity as a major determinant of nucleosome formation. While this feature, affecting bending, is certainly dominant for tuned sequences designed to exhibit strong 10 bp periodicities (the 601 clone being one of the best optimized sequence so far), we do not observe strong intrinsic periodicity in our random selection of genomic sequences and synthetic enhancers (Fig. 3.7 a), neither any significant correlation between our $\Delta\Delta$G and the autocorrelation factors of the corresponding DNA sequences (Fig. 3.7 b). Finally, we investigated the influence on nucleosome formation of the occurrence of poly (dA:dT) stretches, a feature that has been shown to deplete nucleosome at high frequency Tillo and Hughes (2009). Our sequences do not contain very few long dA:dT tracks and for the lengths that are present in our data, we dont find them to be correlated with $\Delta\Delta$G (Fig. 3.7 b). In contrast, for the sub-group of synthetic enhancer sequences whose background sequences are highly similar, we observe that histone-DNA affinity slightly but detectably decreases with growing number of nucleotides in very short dA:dT tracks (Fig. 3.7 c). Hence, this shows that even very short dA:dT tracks can lead to weakening of histone-DNA interactions. To conclude, we demonstrate with true affinity data that GC content is a major determinant of nucleosome formation for

our selection of DNA sequences. The high predictive value of GC content was already known and is not surprising since GC content can influence multiple attributes of the DNA structure such as favorable propeller twist and slide DNA shape features. High GC content will also tends to lack stiff poly (dA:dT) tracks. In this sense, GC content represents a relevant prediction parameter, since this single feature englobes many DNA properties. I am convinced that this assay will be useful to future works in characterizing comprehensively nucleosome-sequence binding specificities, and could be extended to investigate complex binding interactions like the interplay between pioneer transcription factors and nucleosome.

Chapter 4

Methods

4.1 Transcription factor DNA interactions

4.1.1 Protein purification

Protein expression *E.coli* BL21 DE3 calcium chemically competent bacteria were transfected with a pGEX -plasmid containing the coding sequence of the respective TF using a heat shock of 42 °C for 45 s. After 5 min incubation on ice the cells were recovered for at 37 °C for 60 min in SOC medium. 5 ml of LB medium were inoculated with 100 μl of the SOC culture and grown overnight at 37 °C and 150 rpm. 200 ml of Auto-induction medium were inoculated with the overnight preculture and incubated at 37 °C for 4 h and and 18 °C for 17 h. The bacteria were pelleted at 5000g for 15 min and frozen in liquid nitrogen for storage at - 80 °C.

Protein purification The bacteria pellets were resuspended in GST-Buffer A containing cOmplete protease inhibitor cocktail and lysed using a French pressure cell press. The cellular debris was pelleted at 18000g and 4 °C for 15 min. The protein was purified using an Äkta system with a 1 ml GST-column. The protein elution was performed with gluthation containing GST-Buffer B. The protein concentration was estimated spectro-photometrically using a nanodrop. Their identity and purity was checked on a 10 % SDS-PAGE gel.

4.1.2 Determination of affinities

The affinities towards the competitor sequences harboring the mutations of interest were determined using the previously described HiP-FA method (Jung et al. (2018, 2019) and introduction). In brief, the TF of interest and a fluorescently labelled reference DNA are embedded in a porous agarose gel matrix. The competitor solution is added on top of this gel at the start of the experiment. The spatial-temporal gradient of competitor solution allows to record a titration curve o to determine the affinity of the TF towards the respective competitor sequence. Separate wells containing the DNA intercalating dye Nile Blue are used to determine the competitor concentration at any given time and position. In contrast to the original HiP-FA protocol, neighboring double mutations are systematically introduced in the TFs consensus (i.e. strongest binder) sequence (at 7 positions in the center of the 16 bp DNA oligomer, 6 positions for GATAe). The flanking sequences of the binding site were optimized to reduce the occurrence of off-target binding, checking all possible sequences using a modified version of the Python application PySite (Von Reutern (2017)) with the PWMs from the HiP-FA publication as input. The binding sites for all permutations for a given sequence were tested and the off target was defined as a sum of the strongest off-target's weight and the sum off all off-target weights (effectively doubling the impact of the strongest off-target binder). This procedure was chosen in order to find a balance between the impact of many small

values and a single very strong value which is probably impacting the measurements stronger. For some longer PWMs, parts of the flanking sequences were fixed to the consensus sequences (IC $\geq$ 1.0) in order to ensure sufficient binding for all mutations.

4.1.3 Determination of binding weight and off-target removal

The following section was developed by Marc von Reutern but doesn't have a citable source yet which why it is included in this methods subsection:

From the perspective of the PWM model, every site is a binding site and two sequences differ only in their binding weight. For this reason, residual binding activity is even in regions of oligomers that were hand picked to contain nothing resembling a consensus binding site. These traces of binding activity can potentially influence the binding behaviour of the whole oligomer and should be incorporated in the calculation of binding weights. We developed a heuristic algorithm that constructs a PWM de-novo from a set of oligomers s_i with known binding affinity k_i. The binding weight is inverse proportional to the affinity. For simplicity, the ratio between affinity and inverse weight is set to 1. We assume that at most one protein binds to each oligomer at any time. Therefore, we can approximate the total binding weight w_j of oligomer j as $w_j = \sum_{i\ site\ on\ j} w_i$ where $w_i = \frac{1}{k_i}$ is the binding weight of site i. Our model searches the space of PWMs with an iterative approach. Goal of the algorithm is to find the PWM that best matches the measurements. We assess the quality of fit by a scaled sum of squared errors between estimated and measured binding weights. For the first iteration, we construct PWM_0 based on the target sites at the center of the oligomers. The following iterations carried out in three steps: First, find the binding sites and calculate the weights based on the PWM of the previous iteration. Second, assign heuristic binding weights to every called site. This is done by distributing the measured weight of an oligomer among all its sites, based on their calculated binding weight ratios from the last step. Third, construct a new PWM from the list of sites and their heuristically estimated binding weights. To transform the binding weights into an energy equivalent space, their natural logarithm was taken.

4.1.4 Representation of PWMs and DPWMs

PWMs are depicted as sequence logos according to the original publication (Schneider and Stephens (1990)) using a custom Python script. In the DPWMs, the information content (mutual information) is calculated using a KullbackLeibler divergence(Kullback and Leibler (1951)) using a logarithm with base 2.

4.1.5 Shape readout weight

To determine the influence of DNA shape on the binding weight of a TF, the energies were calculated from the binding weights as described above and normalized using a z-score to make the following analysis more robust against the range of the binding energies and thereby less influenced by a TFs specificity. The values of the DNA features were calculated using the lookup tables provided by the Rohs group (Zhou et al. (2013); Li et al. (2017); Chiu et al. (2017b)). The normalized Energies were plotted against the rescaled shape values of a feature (scaled from 0 to 1 for all possible values the feature can possibly take) for each position, see also Figure 1c. Per position, only those data points containing a mutation in this position or in the neighboring ones were taken. Since the features between two neighboring base pairs (inter features) are assigned to both base pairs, those features have four possible positions where mutations are included in the plot and following regression (two positions plus two neighbors). A robust linear regression Huber (2011) using Hubers T as M-estimator, with mean absolute deviation as scale factor (implemented in the rlm_model from the statsmodels package- v0.8.0-in Python) was used to estimate a linear regression of the data and an estimate of the confidence interval using an iteratively reweighted least squares approach. The statistics of the robust linear regression was used to define confidence intervals. Depending if

the value was more than one or two standard error different from zero, the significance level was determined. Only values more than two standard error different from the mean are regarded as statistically significant ($p<0.05$).

4.1.6 Clustering of features and TFs

The hierarchical clustering was performed using the figure_factory module from the plotly package v4.0.0- in Python. The distances were calculated based on correlations to avoid single features with high values to dominate the clustering and have a less scale variant distance function. The clustering was always performed on shape readout weights as defined above. To cluster the TFs, the features and positions were treated as two different dimensions for the clustering. To reduce the impact of the window chosen of the core sequence the distance function was modified to allow for up to two bases offset, against shape readout weights of zero when comparing with positions outside of the chosen window. The clustering for the features was performed on a transposed matrix with TFs and positions as dimensions of the data.

4.1.7 Software development

Titration curves fitting In the original HiP-FA paper, the program labview was used to fit both the nile blue curves and the titration curves of competitors against the labelled reference. As this fitting procedure was very manual, it was potentially less robust which is why I developed a program in Python 2.7 that determines the parameters automatically based on consensus curves and is less dependent on initial guess parameters. It has also the possibility of removing systematic plate effects based on water well (titration wells containing gel, TF and labelled reference, but no competitor in the top buffer). This can potentially guard against problems in protein stability, changes in the surrounding temperature or other potential influences leading to changes in anisotropy over time in all wells simultaneously. The program still takes the raw titration data from labview as an input, for which the concentrations are fitted based on anisotropy values, as I felt this fitting seems to be more robust and also more complicated to implement. The program is written in an interactive fashion so that a potential user can run it without the need of changing input variables within the code.

The main class (LabviewFitter) is initialized with the curves file which can be generated in the labview software. A polynomial fit on the water wells can optionally be performed and the changes from a stable curve are subtracted from all imported curves. The fitting parameters can either be determined based on all individual consensus curves or on one averaged curve of all consensus wells. For the following fit of all competitor titrations, the parameters Rt (the active protein concentration), K_D1 (the reference consensus affinity), Lst (the labelled reference DNA concentration, which is known) and C (the curve offset) are fixed. B (the bottom asymptote) is restrained to the consensus value ± a factor, 3% as default. The K_Ds of each curve are determined and with the help of an excel document assigning the respective sequences the results are written to a comma separated values table.

The code can be found in the appendix under A.2

4.2 Sensitive automated measurement of histone-DNA affinities in nucleosomes

4.2.1 DNA synthesis

Annealing and ligation Each DNA construct for the following PCR consists of six pieces, three constituting the forward strand and three constituting the reverse strand. To ensure a sufficient

annealing, there are overhangs of 14 bp at the complementary strands. The DNA pieces were synthesized by Eurofins, Germany, and were supplied at a concentration of 100 μM. For the annealing, 2 μl of each piece were mixed, the solution was heated to 70 °C in a standard PCR machine. The temperature was decreased to 25 °C at a rate of 0.1 K/s. The resulting annealed DNA was purified using a PCR purification kit (Qiagen) following the manufacturer's instructions. The cleaned DNA was eluted from the column using 30 μl of dH_2O. 10 μl of this eluate were ligated in 150 μl T4 DNA Ligase Buffer (NEB) using the T4 DNA ligase (NEB) in tenfold excess. The ligation was incubated for 30 min at RT. The reaction was cleaned up using the PCR purification kit (Qiagen) and eluting in 30 μ of H_2O.

Fluorescence labeling via PCR In order to generate the fluorescent labeled reference DNA, the aforementioned ligation product was amplified via a PCR in which one of the primers was 5' fluorescently labeled with Cy5 (Eurofins, Germany). The PCR program is shown in table 4.1. The product was again cleaned up using the PCR purification kit. The product was eluted using 30 μl of elution buffer (Qiagen).

Table 4.1: PCR program P.fu polymerase

Step	Temperature [°C]	duration [s]
1	95	120
2	95	30
3	55	30
4	72	60
5	go to step 2 for 29 times	
6	72	300

PCR of nucleosome sequences The unlabeled competitor sequences were amplified using a touchdown PCR (see table 4.2). As a template, genomic fly DNA (provided by M. Bozek) or a plasmid from gene synthesis (Synbio) was used. After the PCR, the product was inspected on an analytical 1.5 % agarose gel and cleaned up using a PCR cleanup kit. To ensure sufficient concentrations, eight PCR reactions of 50 μl each were pooled and eluted in 30 μl elution buffer.

4.2.2 Histone octamer purification

Embryo collection The histones used in this experiment were extracted from *Drosophila melanogaster*, Oregon^R, wild type flies. Approximately 2 10^4 flies per fly cage were supplemented with one apple juice agar plate per cage. The plates were changed every 12 h. Starting after 36 h, fly embryos were collected. To collect the embryos, they were rinsed off using water pressure and separated from other fly elements over three consecutive smaller sieves. The embryos were washed in cold embryo wash buffer. The chorion was removed by stirring in 200 ml embryo wash buffer with 60 ml hypochlorite solution for 3 min. The hypochlorite was removed by rinsing the dechorionated embryos in the smallest sieve with water for 5 min. The embryos were snap frozen in liquid nitrogen and stored at -80 °C.

Custom metal block To ensure temperature stability and minimized evaporation we designed a custom metal block (see Figure 4.1). The tightness of the tubes is ensured by the weight of the custom metal lid and the soft silicon pad at the bottom of the lid. Evaporation is minimized

Table 4.2: PCR program for competitor sequences (touchdown)

Step	Temperature [°C]	duration [s]
1	95	120
2	95	30
3	$T_{annealing}$+10 K -1 K per cycle	30
4	72	60
5	go to step 2 for 9 times	
6	95	30
7	$T_{annealing}$	30
8	72	70
9	go to step 6 for 19 times	
10	72	300

by a heating pad (like used for car exterior mirrors) attached to the top of the lid, operating at approximately 50 °C.

Nucleosome separation The following protocol was modified according to Krietenstein et al. (2012a). 30 g of embryos were thawed in 25 ml of embryo lysis buffer at 4 °C rotating. The suspension was mechanically lysed using a Yamato homogenizer (1000 rpm) passing it through six times.The lysate was centrifuged at 10 000 *g* at 4 °C for 15 min. The turbid supernatant was decanted, the softer, light brown phase above a blackish brown pellet was resuspended in 30 ml suc buffer using glass pipettes. The resulting suspension was centrifuged at 10 000 *g* and 4 °C for 15 min. The supernatant was removed by pipetting with a glass pipette, the light brown phase was again resuspended in 30 ml of suc buffer. The centrifugation and removal of supernatant was repeated like above, the light brown phase was resuspended in 18 ml of suc buffer. One tablet of cOmplete protease inhibitor cocktail (EDTA free) as dissolved in the suspension which was homogenized using a Dounce homogenizer (B-pestle). 54 μl of 1 M $CaCl_2$ were added, the solution was incubated at 26 °C for 5 min (water bath). 125 μl of 0.59 U/μl MNase were added, the digest was conducted for at 26 °C for 10 min, mixing the solution every 2 min by inversion. The reaction was stopped adding 360 μl of 0.5 M EDTA. The nuclei were centrifuged at 10 000 *g* and 4 °C for 30 min. The pellet ws resuspended in 6 ml TE and hypotonic lysis was conducted by rotating the sample at 4 °C for 45 min. The lysed nuclei were centrifuged at 15 000 *g* and 4 °C for 30 min. The KCl concentration was adjusted to 0.63 M using the embryo elution buffer containing 2 M KCl. The resulting solution was filtered through a 0.45 μm and a 0.22 μ filter.

Histone octamer isolation The filtrate of the previous paragraph was loaded onto the ÄKTA system pre-equilibrated with embryo running buffer. The histone octamers were loaded to a column consisting of 30 ml hydroxylapatite. After washing with 2 column volumes, the elution was performed with embryo elution buffer. The eluate was collected in fractions of 1 ml, the two fractions with the highest absorbance were concentrated using centrifugal filters with 10 000 MWCO. The centrifugal filters were pre-rinsed with embryo elution buffer and the concentration was carried out at 12 000 *g*, 4 °C. The resulting concentrate was mixed with glycerol to contain 50 % glycerol and stored at -20 °C.

4.2.3 Nucleosome reconstitution

Determination of histone to DNA ratio Before starting with titrations, the optimal ratio of histone octamers to fluorescently labeled DNA was determined. Each titration sample contained 2 μl of 20 ng/μl Cy5 labeled reference and 18 μl of nucleosome octamer dilution of various concentrations. A dilution series of the histone octamer stock solution in titration high salt buffer was performed. Depending on the histone octamer preparation, the necessary dilution can vary a lot and is best determined empirically. Optimal ratios were at 1:500 and 1:2000 for our octamer stock solutions.

Competitor titration When determining the affinity of a competitor sequence, in addition to labeled reference DNA and the histone octamer dillution at the optimal ratio, 2 μl of unlabeled competitor DNA of different concentrations are added.

Automated salt titration The samples for a titration are mixed in a 500 μ protein low binding reaction tube using low retention tips. These tubes are kept in a custom metal block with a heated lid to keep temperature as constant as possible and to minimize condensation at the lid. During the automated titration, the samples are kept at 30 °C, the lid is heated to 42 °C. Over the course of 12 h, the samples are diluted using titration low salt buffer; after each interval of 40 min a volume with 1 μl more than in the previous step is added (see also table 4.3). After a final incubation for 60 min the samples were measured using fluorescence anisotropy or EMSA.

Table 4.3: Automated titration for nucleosome reconstruction

added Volume	total Volume	NaCl concentation [mM]
0	22	909
1	23	869
2	25	800
3	28	714
4	32	625
5	37	540
6	43	465
7	50	400
8	58	344
9	67	298
10	77	259
11	88	227
12	100	200
13	113	176
14	127	157
15	142	140
16	158	126

4.2.4 Nucleosome measurement

All pipetting steps in the following section were carried out using low retention tips.

Electrophoretic Mobility Shift Assay The protocol was modified from Kim (2011). To analyze the samples via Electrophoretic Mobility Shift Assay (EMSA), 20 μl of each sample was mixed with 20 μl of EMSA sample buffer and 5 μl of 50% glycerol in a 500 μl protein low binding tube. The samples were loaded on a 0.5% agarose gel prepared with 2 fold tris glycine native running buffer. The gel was run with 50 V for 60 min at 4 °C with acquiring pictures every 30 min using a gel documentation system.

Fluorescence anisotropy measurement To analyze the samples via FA 100 μl of sample were mixed with 100 μl of FA buffer in a glass bottom microscopy plate. The samples were measured in the same microscopy setup presented in Jung et al. (2018) and - in more detail - in Jung et al. (2019). The laser power was 2.5 W at the end of the excitation part of the setup. The plate with all samples was measured threefold, doing height stacks with 12 pictures per well.

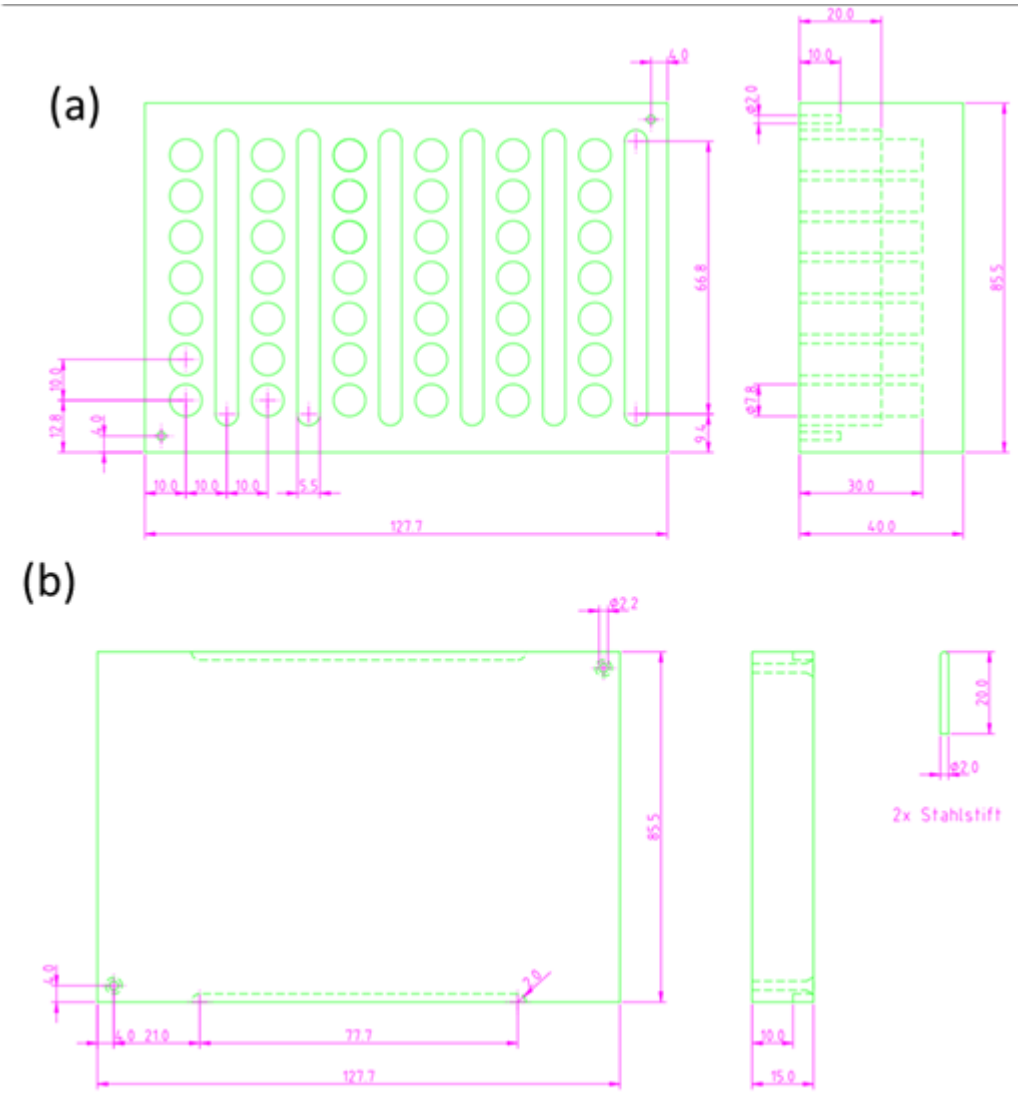


Figure 4.1: technical drawing for the custom metal block. Made from aluminum, tolerances according to ISO 8015 and ISO 2768-m, workpiece edge DIN 6784. (a) technical drawing for the metal block. (b) technical drawing for the lid, Stahlstift = metal pin

Chapter 5

Material

5.1 Consumables

Consumable	supplier
agarose	Biorad
Amicon Ultra-0.5 mL Centrifugal Filters	Merck
Calcium chloride ($CaCl_2$)	Roth
cOmplete,Mini, EDTA-free Protease Inhibitor Cocktail	Sigma-Aldrich
Di-hydrogen potassium phosphate	Merck
EDTA	Sigma Aldrich
EGTA	Sigma Aldrich
glycerol	Roth
Glycine	Roth
GSTrap HP Columns	GE Healthcare
Hepes	Merck
Hydrogen Di potassium phosphate	Merck
hydroxylapatite	Bio-Rad
hypochloite solution (DanKlorix Hygienereiniger)	Colgate-Palmolive
L-Glutathion reduziert	Roth
Magnesium Chloride	Merck
Micro tube 0.5ml protein LB	Sarstedt
nile blue	Merck
Nonident-P40 (NP40)	Sigma
Nuclease micrococcal from *Staphylococcusaureus* (MNase)	Sigma-Aldrich
PMSF	Sigma-Aldrich

Potassium Chloride	Merck
QIAquick PCR Purification Kit	Qiagen
senso plate	Greiner
Sodium Chloride	Merck
Sodiummetabisulfite	Sigma-Aldrich
StrepTrap HP	GE Healthcase
Succrose	Roth
T4 DNA Ligase	NEB
T4 DNA Ligase Buffer (10X)	NEB
Tris	Roth
Tween20	Roth

5.2 equipment

device	manufacturer
dounce homogenizer 30ml	vwr
rotation weel rotator sb3	stuart
super loop 50 ml	GE Healthcare
water bath	memmert

5.3 devices

device	manufacturer
Äkta Prime plus	GE Healthcare
French pressure cell press: HC-5000	Microfluidics
geldoc	biorad
Homogenizer	Yamato
centrifuge 5415R	eppendorf
centrifuge Allegra X12R	Beckman Coulter
centrifuge Avanti J26XP	Beckman Coulter
Confocal LSM710	Zeiss
Incubator innova44	New Brunswick Scientific
microscope DMI6000B	Leica
NanoDrop1000	Peqlab
PCR cycler Mastercycler gradient	eppendorf
Red Laser PhoxX638-40	Micron Laserage
Robots Biomek NXP	Beckman Coulter
Sonicator UP200	Hielscher

5.4 Oligos

The following table gives an overview of the sequences used for the systematic dinucleotide mutations. The part of sequences which were systematically mutated with neighboring dinucleotides is depicted in red in Table 5.2, the constant flanks are depicted in black.

Table 5.2: Overview unlabelled competitor sequences. Systematically mutated part of the sequence is depicted in red, the constant part in black

TF name	consensus sequence
Fkh	CTTAGATAAATATCGC
Zld	GAACTCAGGTAGCCCG
Gt	GACATTACGTAACCTC
D	CTGCCATTGTTCCGGG
Bcd	CCCGGTAATCCCTCGT
Eip93f	AGCAGCCGAAAATGGG
GATAe	TGCATGTTATCTACGT
Oc	CCCGGTAATCCCTCGT
Hb	GGACTTTTTTAGAAGG
Gsc	CCCGGTAATCCCTCGT
Hkb	TCGGGGCGTGAAAATT
Nub	ACGCCTATGCAAAGGG
Tll	CCCCAAGTCAAGGGGG

5.5 Plasmids

All plasmids used for protein expression are based on pGEX6p1 (see Figure 5.1).

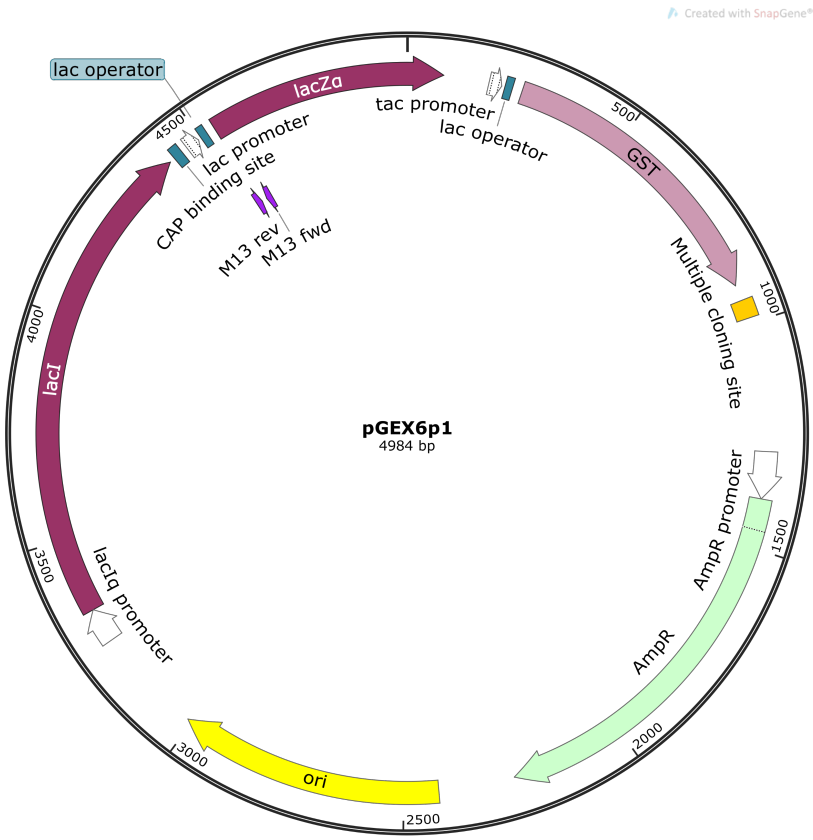


Figure 5.1: Plasmid map of pGEX6p1. Shown are the features of the plasmid. AmpR:ampicillin resistance, ori: origin of replication, lacI: lac repressor, GST: glytathion-S-transferase. The plasmid map was generated using snape gene viewer.

5.6 Buffers

Table 5.3: Recipe: Embryo lysis buffer

Component	Concentration
Hepes pH 7.5	15 mM
KCl	10 mM
$MgCl_2$	5 mM
EDTA	0.1 mM
EGTA	0.5 mM
Sucrose	350 M
DTT	1 mM
PMSF	0.2 mM
Sodiummetabisulfite	1 mM

Table 5.4: Recipe: Embryo Suc buffer

Component	Concentration
Hepes pH 7.5	15 mM
KCl	10 mM
$MgCl_2$	5 mM
EDTA	0.05 mM
EGTA	0.25 mM
Sucrose	350 mM
DTT	1 mM
PMSF	0.1 mM

Table 5.5: Recipe: Embryo running buffer (Äkta wash buffer)

Component	Concentration
$KHPO_4$ pH 7.2	0.1 mM
KCl	630 mM

Table 5.6: Recipe: Embryo elution buffer

Component	Concentration
$KHPO_4$ pH 7.2	0.1 mM
KCl	2 M

Table 5.7: Recipe: Titration high salt buffer (nucleosome assay)

Component	**Concentration**
NaCl	1 M
Tris pH 8.0	10 mM

Table 5.8: Recipe: Titration low salt buffer (nucleosome assay)

Component	**Concentration**
Tris pH 8.0	10 mM
EDTA	1 mM

Table 5.9: Recipe: EMSA sample buffer (nucleosome assay)

Component	**Concentration**
Tris pH 8.0	25 mM
Glycerol	50 % (v/v)

Table 5.10: Recipe: tris glycine native running buffer (nucleosome assay)

Component	**Concentration**
Tris pH 8.0	25 mM
Glycine	192 mM

Table 5.11: Recipe: Binding buffer (HiP-FA)

Component	**Concentration**
KH_2PO_4	13.9 mM
K_2HPO_4	19.1 mM
Tween 20	1 mM
0.01 % (v/v)	

Chapter 6

Conclusion

This work deals with with two different topics , the determination of non-linearity in TF-DNA binding and the development of a new assay to determine histone-DNA affinities in nucleosomes, which are connected on several distinctive levels. On the biological level, the specificities determined in the chapter about non-linearity can only be applied in a holistic model under consideration of the occupancy with nucleosomes in the genomic context, partially linked via their affinity to DNA, the other topic of this thesis. Additionally, the topics are connected on the methodical level, since both methods apply an *in vitro* measurement of fluorescence anisotropy to gain insights into the interaction strengths between a protein (complex) and a DNA. Finally, the two topics are also connected on the level of experimental principle since both apply a competitive titration of a labelled reference with an unlabelled competitor.

The resulting affinity (or binding energy) data are very valuable for the interpretation and modeling of biological processes. Only when both concentration and affinity of a protein are known, a modeling of there interaction is possible. It is in this regard important to mention that most assays investigating interactions between proteins and DNA (be it nucleosomes or TFs) do only determine specificity and not affinity with a few exceptions (MITOMI and SMILE-Seq for TFs and the traditional, manual salt titrations for nucleosomes). The specificity of the TFs is in this work only calculated as an additional, secondary quantity using the individual affinities determined from titrations of each competitor in a separate titration. This direct titration is in deed an important feature of the assays presented in this work. By this procedure it is ensured that there is information about each single sequence provided and no data is lost as opposed to some method employing bulk measurements. This in turn provides valuable knowledge especially about weak binders that are often lost in alternative methods, which is, in my opinion, the reason why the PWMs provided in the original HiP-FA publication (Jung et al. (2018)) outperformed the ones generated with other approaches.

The biological relevance of weak binding sites has been discussed above (see section 1.2.2) and is getting more and more attention in the community (Kribelbauer et al. (2019)). If we are able to better recognize relevant weak sites, we might also learn about so far discovered interaction between TFs which can't be determined using current technologies to determine TFs' binding preferences. In this regard, the impact of the non-linear interactions is most relevant.

Most of these affinities have been determined to sequences deviating two bases from the consensus sequence and are (very) weak binders in many cases. The information provided in addition to the PWMs in the original publication is therefore valuable in itself. In addition, we evaluated the DNA shape readout of the TFs and found it to be very widespread, a finding in line with several publication in recent years (see discussion under 2.3). The rather simple and direct evaluation using a robust linear regression (being in excellent agreement with a more elaborated one - Figure 2.6, Rube et al. (2018)) is in my opinion only possible because of the sensitive and accurate measurements provided by HiP-FA using titration curves each containing more than one hundred

points.

The work presented here is in addition one of the first looking at all geometrical DNA features provided by Li et al. (2017). The following clustering of feature readout resulted in three clusters of which one is in line with several publications (2.3), opening the possibility to find evidence explaining the nature of the other two in future works.

Like mentioned above, all data about TFs have ultimately to be evaluated in the light of DNA occupancy by nucleosomes. This thesis contributes to the question of which features are important for a nucleosome to bind DNA. The direct measurements of minimal sequences exclude the influence of positioning effects. It also sheds light on the spread of naturally occurring sequences, which seems to be rather low, again questioning the biological relevance of sequences designed to be artificially strong. The observed non-monotonous influence of GC content on nucleosome binding has been described in (to my knowledge) only two works so far (see 3.3). The low range of affinities for naturally occuring sequences might be a hint that positioning plays a much larger role for the probability of a sequence being occupied by a nucleosome *in vivo* than the actual differences in affinity towards different sequences.

The achievements reached in terms of methods development in this work might be relevant to others developing interaction studies bases on FA. These are by no means restricted to protein-DNA interactions: Binding events of protein-protein, protein-drug, protein-RNA binding could be monitored; even biological interactions of ribozymes could be investigated - the FA principle is not limited to biological interactions with proteins.

The results provided herein could be used in modeling, especially given the possibility to apply the calculated shape readouts on the DNA shape profiles of full genomes, possible by the works of the Rohs lab. It would also be interesting to see to which degree the shape readout weights determined in this study can be confirmed in structural data or alternatively molecular dynamics simulations.

Appendix A

Appendix

A.1 Data

A.1.1 PWMs and DPWMs

This section gives the dinucleotide weight matrices (displayed as absolute frequencies) for all measurements of TF non-linearity .

$$PWM_{Bcd\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.10 & 0.11 & 0.11 & 0.69 \\ 0.75 & 0.03 & 0.15 & 0.07 \\ 0.92 & 0.01 & 0.07 & 0.00 \\ 0.09 & 0.05 & 0.14 & 0.72 \\ 0.07 & 0.76 & 0.05 & 0.12 \\ 0.13 & 0.57 & 0.12 & 0.18 \\ 0.19 & 0.49 & 0.17 & 0.15 \end{pmatrix} \end{array} \tag{A.1}$$

$$DPWM_{Bcd\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 6.38e-2 & 6.98e-1 & 8.25e-2 & 2.16e-2 & 4.39e-3 & 2.51e-2 \\ 6.56e-3 & 4.12e-3 & 4.26e-2 & 5.09e-2 & 4.35e-2 & 5.22e-2 \\ 6.44e-2 & 5.28e-2 & 1.20e-1 & 5.24e-2 & 1.87e-3 & 4.10e-2 \\ 3.76e-2 & 5.72e-5 & 6.42e-1 & 1.12e-1 & 1.34e-3 & 2.55e-2 \\ 6.76e-2 & 2.44e-2 & 4.75e-3 & 3.20e-2 & 1.04e-1 & 8.86e-2 \\ 4.27e-3 & 1.98e-3 & 4.65e-3 & 2.63e-2 & 4.67e-1 & 2.33e-1 \\ 5.53e-2 & 7.81e-4 & 3.49e-3 & 8.48e-3 & 1.00e-1 & 8.03e-2 \\ 8.11e-3 & 4.07e-5 & 3.79e-3 & 1.58e-2 & 1.45e-1 & 7.21e-2 \\ 6.73e-2 & 1.36e-1 & 9.07e-3 & 1.27e-2 & 4.36e-3 & 1.62e-3 \\ 5.58e-3 & 1.16e-3 & 1.05e-2 & 7.42e-2 & 2.82e-2 & 5.00e-2 \\ 2.24e-2 & 7.41e-3 & 1.86e-2 & 8.72e-3 & 2.87e-3 & 1.84e-2 \\ 1.05e-2 & 1.64e-4 & 4.85e-2 & 6.55e-2 & 1.31e-3 & 2.99e-2 \\ 4.42e-1 & 6.80e-2 & 9.34e-4 & 3.69e-2 & 1.60e-2 & 5.82e-2 \\ 1.54e-2 & 1.85e-4 & 6.38e-3 & 3.96e-1 & 7.34e-2 & 7.25e-2 \\ 8.63e-2 & 4.06e-3 & 1.82e-3 & 2.40e-2 & 1.61e-3 & 7.77e-2 \\ 4.30e-2 & 1.84e-5 & 5.26e-5 & 6.23e-2 & 4.46e-3 & 7.39e-2 \end{pmatrix} \end{array} \tag{A.2}$$

$$PWM_{Bcd\_rep2} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.08 & 0.09 & 0.05 & 0.78 \\ 0.89 & 0.01 & 0.06 & 0.03 \\ 0.96 & 0.00 & 0.03 & 0.00 \\ 0.04 & 0.02 & 0.09 & 0.85 \\ 0.07 & 0.77 & 0.04 & 0.12 \\ 0.13 & 0.60 & 0.11 & 0.17 \\ 0.18 & 0.45 & 0.20 & 0.17 \end{pmatrix} \end{array} \tag{A.3}$$

$$DPWM_{Bcd\_rep2} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 6.07e-2 & 8.50e-1 & 3.82e-2 & 1.44e-2 & 5.58e-3 & 9.59e-3 \\ 6.49e-3 & 2.78e-3 & 2.08e-2 & 2.33e-2 & 4.43e-2 & 4.97e-2 \\ 4.59e-2 & 2.80e-2 & 8.02e-2 & 2.74e-2 & 3.76e-3 & 2.87e-2 \\ 2.79e-2 & 2.22e-4 & 7.94e-1 & 9.16e-2 & 8.22e-3 & 1.62e-2 \\ 6.65e-2 & 1.24e-2 & 3.46e-3 & 2.87e-2 & 1.04e-1 & 9.34e-2 \\ 3.83e-3 & 6.58e-4 & 3.05e-3 & 1.27e-2 & 4.88e-1 & 2.34e-1 \\ 3.51e-2 & 1.06e-3 & 2.65e-3 & 7.48e-3 & 8.79e-2 & 1.03e-1 \\ 8.62e-3 & 7.35e-5 & 2.60e-3 & 1.50e-2 & 1.36e-1 & 9.05e-2 \\ 3.99e-2 & 6.10e-2 & 6.10e-3 & 1.23e-2 & 4.62e-3 & 1.09e-3 \\ 3.85e-3 & 7.65e-4 & 6.98e-3 & 4.89e-2 & 2.29e-2 & 4.21e-2 \\ 1.31e-2 & 3.79e-3 & 1.23e-2 & 1.13e-2 & 2.13e-3 & 1.04e-2 \\ 7.45e-3 & 5.76e-4 & 2.62e-2 & 7.76e-2 & 1.31e-3 & 2.73e-2 \\ 6.06e-1 & 3.24e-2 & 1.37e-3 & 4.40e-2 & 8.38e-3 & 8.38e-2 \\ 8.80e-3 & 1.58e-3 & 9.78e-4 & 4.85e-1 & 7.85e-2 & 6.51e-2 \\ 4.35e-2 & 3.88e-3 & 1.02e-3 & 2.27e-2 & 1.48e-3 & 8.13e-2 \\ 2.31e-2 & 4.35e-4 & 2.07e-4 & 7.80e-2 & 2.89e-3 & 6.35e-2 \end{pmatrix} \end{array} \tag{A.4}$$

$$PWM_{Bcd\_rep3} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.12 & 0.09 & 0.08 & 0.72 \\ 0.69 & 0.02 & 0.24 & 0.05 \\ 0.90 & 0.01 & 0.09 & 0.00 \\ 0.08 & 0.05 & 0.14 & 0.73 \\ 0.06 & 0.79 & 0.04 & 0.11 \\ 0.15 & 0.57 & 0.10 & 0.18 \\ 0.17 & 0.54 & 0.13 & 0.15 \end{pmatrix} \end{array} \tag{A.5}$$

$$DPWM_{Bcd\_rep3} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 7.39e-2 & 6.26e-1 & 7.10e-2 & 2.13e-2 & 6.43e-3 & 1.72e-2 \\ 6.34e-3 & 4.61e-3 & 4.14e-2 & 4.47e-2 & 3.73e-2 & 6.80e-2 \\ 3.31e-2 & 6.52e-2 & 1.21e-1 & 4.12e-2 & 1.91e-3 & 2.89e-2 \\ 2.55e-2 & 9.84e-5 & 6.43e-1 & 1.47e-1 & 3.35e-3 & 2.12e-2 \\ 5.59e-2 & 2.12e-2 & 2.97e-3 & 2.69e-2 & 1.21e-1 & 8.58e-2 \\ 3.44e-3 & 2.22e-3 & 2.11e-3 & 2.61e-2 & 4.73e-1 & 2.67e-1 \\ 3.16e-2 & 1.50e-3 & 2.97e-3 & 8.36e-3 & 8.67e-2 & 6.52e-2 \\ 9.89e-3 & 6.71e-5 & 4.74e-3 & 1.87e-2 & 1.47e-1 & 7.41e-2 \\ 5.03e-2 & 2.16e-1 & 9.26e-3 & 1.37e-2 & 7.72e-3 & 1.13e-2 \\ 5.11e-3 & 1.19e-3 & 1.08e-2 & 7.63e-2 & 2.49e-2 & 4.89e-2 \\ 3.12e-2 & 7.24e-3 & 1.86e-2 & 9.48e-3 & 3.02e-3 & 2.34e-2 \\ 1.18e-2 & 3.28e-4 & 6.69e-2 & 5.25e-2 & 1.42e-3 & 2.88e-2 \\ 4.54e-1 & 4.88e-2 & 7.12e-4 & 3.18e-2 & 1.66e-2 & 5.69e-2 \\ 1.54e-2 & 4.39e-4 & 2.88e-3 & 4.04e-1 & 6.54e-2 & 8.27e-2 \\ 1.57e-1 & 5.75e-3 & 1.31e-3 & 2.12e-2 & 1.31e-3 & 6.14e-2 \\ 3.54e-2 & 1.82e-5 & 1.01e-4 & 5.59e-2 & 3.52e-3 & 5.96e-2 \end{pmatrix} \end{array} \tag{A.6}$$

$$PWM_{D\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.07 & 0.77 & 0.08 & 0.08 \\ 0.81 & 0.03 & 0.04 & 0.13 \\ 0.01 & 0.03 & 0.03 & 0.93 \\ 0.04 & 0.01 & 0.02 & 0.94 \\ 0.07 & 0.06 & 0.84 & 0.03 \\ 0.05 & 0.02 & 0.02 & 0.92 \\ 0.09 & 0.10 & 0.08 & 0.73 \end{pmatrix} \end{array} \tag{A.7}$$

$$DPWM_{D\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 5.02e-2 & 7.55e-3 & 5.52e-3 & 1.53e-2 & 1.79e-2 & 2.65e-2 \\ 1.50e-2 & 2.43e-2 & 2.03e-3 & 8.25e-3 & 1.08e-2 & 2.34e-2 \\ 1.59e-2 & 1.97e-2 & 2.21e-3 & 3.04e-2 & 9.89e-3 & 2.15e-2 \\ 9.16e-3 & 6.93e-1 & 9.16e-3 & 4.80e-3 & 5.77e-2 & 3.22e-2 \\ 5.37e-1 & 2.99e-3 & 1.37e-2 & 1.99e-3 & 4.37e-2 & 7.88e-3 \\ 2.00e-2 & 7.26e-3 & 3.27e-3 & 3.86e-3 & 5.99e-3 & 7.27e-3 \\ 2.60e-2 & 5.64e-3 & 4.11e-3 & 7.78e-3 & 1.12e-2 & 7.04e-3 \\ 8.36e-2 & 2.58e-2 & 2.95e-2 & 3.62e-3 & 4.63e-2 & 1.03e-2 \\ 5.33e-2 & 4.14e-3 & 3.92e-3 & 6.52e-3 & 3.51e-2 & 9.82e-3 \\ 1.55e-2 & 9.36e-3 & 3.78e-3 & 4.01e-3 & 1.12e-2 & 7.88e-3 \\ 9.69e-3 & 7.41e-3 & 3.30e-3 & 1.24e-2 & 1.18e-2 & 8.91e-3 \\ 2.21e-2 & 3.35e-2 & 2.39e-2 & 1.11e-3 & 6.60e-1 & 1.08e-2 \\ 5.92e-2 & 1.42e-2 & 3.37e-2 & 6.61e-2 & 2.49e-2 & 7.07e-2 \\ 1.67e-2 & 1.94e-2 & 8.63e-3 & 5.31e-2 & 8.01e-3 & 8.07e-2 \\ 1.78e-2 & 1.82e-2 & 1.37e-2 & 7.57e-1 & 2.43e-2 & 6.93e-2 \\ 4.85e-2 & 1.08e-1 & 8.40e-1 & 2.37e-2 & 2.07e-2 & 6.06e-1 \end{pmatrix} \end{array} \tag{A.8}$$

$$PWM_{Eip93f\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.23 & 0.38 & 0.01 & 0.38 \\ 0.00 & 0.98 & 0.00 & 0.02 \\ 0.44 & 0.16 & 0.36 & 0.03 \\ 0.89 & 0.00 & 0.10 & 0.00 \\ 0.99 & 0.00 & 0.00 & 0.01 \\ 0.79 & 0.03 & 0.02 & 0.17 \\ 0.30 & 0.26 & 0.26 & 0.17 \end{pmatrix} \end{array} \tag{A.9}$$

$$DPWM_{Eip93f\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 1.53e-6 & 4.81e-7 & 4.22e-1 & 8.88e-1 & 7.64e-1 & 2.10e-1 \\ 2.28e-1 & 2.29e-18 & 2.02e-18 & 1.59e-5 & 2.78e-2 & 1.80e-1 \\ 2.94e-5 & 7.02e-6 & 7.04e-3 & 1.12e-3 & 1.57e-2 & 1.80e-1 \\ 1.10e-3 & 8.28e-9 & 2.31e-4 & 4.70e-3 & 1.65e-1 & 1.18e-1 \\ 7.60e-6 & 4.30e-1 & 1.56e-1 & 9.79e-8 & 1.37e-5 & 7.63e-3 \\ 3.80e-1 & 1.59e-1 & 1.68e-16 & 3.76e-7 & 3.73e-3 & 7.45e-3 \\ 7.37e-7 & 3.51e-1 & 1.62e-6 & 3.68e-25 & 1.37e-3 & 3.26e-3 \\ 9.49e-3 & 3.01e-2 & 1.36e-5 & 1.29e-5 & 2.66e-3 & 7.49e-2 \\ 1.89e-8 & 2.46e-5 & 3.44e-1 & 1.04e-1 & 9.61e-4 & 4.32e-3 \\ 8.53e-3 & 8.71e-8 & 3.80e-8 & 6.02e-6 & 1.84e-4 & 2.31e-3 \\ 5.17e-6 & 6.80e-7 & 4.04e-2 & 1.83e-3 & 5.05e-5 & 7.51e-6 \\ 2.06e-32 & 2.64e-4 & 4.47e-5 & 2.55e-5 & 6.07e-6 & 2.00e-9 \\ 2.80e-6 & 1.40e-2 & 2.96e-2 & 1.15e-4 & 4.04e-3 & 4.54e-2 \\ 3.73e-1 & 1.19e-18 & 3.25e-16 & 1.59e-7 & 1.79e-11 & 1.54e-2 \\ 2.41e-5 & 8.77e-3 & 6.47e-4 & 4.63e-5 & 9.64e-4 & 7.81e-2 \\ 7.40e-4 & 7.95e-3 & 8.14e-6 & 2.88e-4 & 1.38e-2 & 7.29e-2 \end{pmatrix} \end{array} \tag{A.10}$$

$$PWM_{Eip93f\_rep2} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.19 & 0.45 & 0.08 & 0.27 \\ 0.02 & 0.83 & 0.01 & 0.14 \\ 0.34 & 0.14 & 0.32 & 0.20 \\ 0.89 & 0.00 & 0.09 & 0.02 \\ 0.98 & 0.00 & 0.00 & 0.02 \\ 0.60 & 0.07 & 0.03 & 0.30 \\ 0.44 & 0.21 & 0.16 & 0.20 \end{pmatrix} \end{array} \tag{A.11}$$

$$DPWM_{Eip93f\_rep2} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 2.22e-3 & 7.00e-6 & 3.14e-1 & 8.77e-1 & 5.66e-1 & 2.24e-1 \\ 1.56e-1 & 1.37e-8 & 3.67e-9 & 1.92e-3 & 6.52e-2 & 1.04e-1 \\ 6.52e-3 & 5.23e-3 & 2.82e-2 & 1.27e-4 & 2.94e-2 & 8.06e-2 \\ 2.63e-2 & 4.01e-3 & 4.68e-3 & 1.57e-2 & 2.81e-1 & 9.98e-2 \\ 6.68e-3 & 2.98e-1 & 1.28e-1 & 1.78e-7 & 1.24e-3 & 2.58e-2 \\ 3.64e-1 & 1.21e-1 & 2.52e-7 & 5.18e-9 & 3.00e-3 & 3.08e-2 \\ 5.65e-3 & 2.84e-1 & 1.76e-6 & 1.43e-10 & 1.40e-4 & 1.83e-2 \\ 6.10e-2 & 1.80e-1 & 8.24e-7 & 1.00e-10 & 4.32e-3 & 4.66e-2 \\ 4.27e-3 & 2.54e-5 & 3.00e-1 & 8.98e-2 & 8.18e-5 & 1.16e-2 \\ 6.68e-2 & 2.77e-7 & 6.07e-8 & 5.67e-5 & 1.84e-3 & 1.12e-2 \\ 6.93e-3 & 4.42e-3 & 3.07e-2 & 1.10e-6 & 9.46e-7 & 8.44e-3 \\ 3.90e-2 & 5.37e-3 & 5.18e-3 & 3.56e-6 & 1.22e-4 & 8.62e-3 \\ 4.95e-3 & 1.08e-2 & 1.89e-1 & 1.52e-2 & 1.02e-2 & 1.11e-1 \\ 2.20e-1 & 7.12e-8 & 1.35e-9 & 7.86e-6 & 8.62e-3 & 8.03e-2 \\ 2.04e-3 & 4.78e-2 & 5.30e-5 & 1.70e-8 & 4.76e-4 & 5.04e-2 \\ 2.83e-2 & 3.86e-2 & 5.02e-4 & 1.84e-9 & 2.81e-2 & 8.82e-2 \end{pmatrix} \end{array} \tag{A.12}$$

$$PWM_{Fkh\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.00 & 0.08 & 0.00 & 0.91 \\ 0.97 & 0.03 & 0.00 & 0.00 \\ 0.98 & 0.01 & 0.00 & 0.00 \\ 0.99 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.06 & 0.01 & 0.93 \\ 0.98 & 0.01 & 0.01 & 0.01 \\ 0.21 & 0.34 & 0.04 & 0.41 \end{pmatrix} \end{array} \tag{A.13}$$

$$DPWM_{Fkh\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 3.21e-3 & 9.35e-1 & 9.69e-1 & 2.13e-4 & 2.03e-4 & 2.00e-1 \\ 5.68e-3 & 1.38e-2 & 1.96e-3 & 5.95e-2 & 1.43e-4 & 3.22e-1 \\ 8.88e-4 & 1.38e-3 & 2.08e-3 & 6.19e-3 & 6.94e-6 & 3.91e-2 \\ 2.81e-3 & 1.78e-3 & 2.91e-3 & 9.08e-1 & 4.13e-5 & 3.86e-1 \\ 7.78e-2 & 2.56e-2 & 1.43e-2 & 1.85e-3 & 5.67e-2 & 3.04e-3 \\ 6.13e-3 & 3.25e-3 & 2.30e-4 & 2.50e-3 & 1.33e-2 & 3.26e-3 \\ 8.26e-4 & 1.09e-3 & 1.29e-6 & 3.65e-4 & 1.60e-2 & 2.02e-3 \\ 4.04e-3 & 3.71e-3 & 6.97e-4 & 1.84e-3 & 1.88e-2 & 2.61e-3 \\ 3.85e-3 & 1.14e-3 & 1.42e-3 & 1.36e-13 & 5.89e-3 & 2.72e-3 \\ 2.68e-3 & 9.89e-4 & 8.83e-4 & 7.55e-3 & 3.42e-3 & 3.40e-3 \\ 1.11e-3 & 1.29e-3 & 6.96e-4 & 8.99e-7 & 9.93e-4 & 3.09e-3 \\ 1.64e-3 & 1.91e-3 & 9.10e-4 & 1.95e-3 & 1.84e-4 & 1.99e-3 \\ 8.62e-1 & 2.89e-3 & 1.84e-3 & 2.99e-10 & 8.65e-1 & 1.62e-2 \\ 2.36e-2 & 2.30e-3 & 1.09e-7 & 6.88e-3 & 5.84e-3 & 7.82e-3 \\ 1.05e-3 & 2.54e-3 & 2.07e-6 & 3.61e-4 & 4.45e-3 & 3.21e-3 \\ 2.67e-3 & 1.43e-3 & 3.34e-3 & 2.73e-3 & 9.02e-3 & 4.03e-3 \end{pmatrix} \end{array} \tag{A.14}$$

$$PWM_{Fkh\_rep2} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.01 & 0.42 & 0.01 & 0.57 \\ 0.88 & 0.11 & 0.00 & 0.01 \\ 0.93 & 0.06 & 0.00 & 0.01 \\ 0.98 & 0.01 & 0.01 & 0.01 \\ 0.00 & 0.22 & 0.02 & 0.75 \\ 0.89 & 0.03 & 0.03 & 0.06 \\ 0.03 & 0.39 & 0.19 & 0.39 \end{pmatrix} \end{array} \tag{A.15}$$

$$DPWM_{Fkh\_rep2} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 6.36e-3 & 7.85e-1 & 9.02e-1 & 5.39e-4 & 4.22e-4 & 2.52e-2 \\ 1.03e-2 & 5.32e-2 & 5.18e-3 & 2.08e-1 & 7.94e-5 & 3.03e-1 \\ 9.13e-4 & 4.20e-3 & 9.40e-3 & 1.97e-2 & 1.00e-4 & 1.51e-1 \\ 5.76e-3 & 5.54e-3 & 8.30e-3 & 7.01e-1 & 1.00e-4 & 3.03e-1 \\ 3.72e-1 & 1.03e-1 & 6.11e-2 & 3.41e-4 & 1.63e-1 & 1.73e-2 \\ 9.66e-3 & 8.86e-3 & 4.29e-4 & 5.35e-3 & 4.63e-2 & 8.80e-3 \\ 6.44e-4 & 1.49e-3 & 1.05e-4 & 9.65e-4 & 4.70e-2 & 5.76e-3 \\ 7.11e-3 & 9.33e-3 & 8.31e-4 & 4.03e-3 & 9.13e-2 & 8.77e-3 \\ 4.98e-3 & 1.62e-3 & 4.82e-3 & 1.38e-5 & 1.54e-2 & 1.54e-2 \\ 5.18e-3 & 5.15e-3 & 1.25e-5 & 2.65e-2 & 8.18e-3 & 1.08e-2 \\ 1.34e-3 & 1.42e-3 & 5.03e-6 & 1.44e-4 & 4.02e-3 & 8.59e-3 \\ 2.51e-3 & 1.75e-3 & 1.86e-5 & 7.31e-3 & 6.10e-3 & 9.04e-3 \\ 5.03e-1 & 6.72e-3 & 6.37e-3 & 5.59e-5 & 5.50e-1 & 8.00e-2 \\ 6.57e-2 & 5.29e-3 & 3.78e-5 & 1.84e-2 & 1.59e-2 & 2.32e-2 \\ 1.03e-3 & 5.77e-3 & 1.00e-4 & 1.08e-3 & 1.64e-2 & 1.01e-2 \\ 4.30e-3 & 1.52e-3 & 1.53e-3 & 6.45e-3 & 3.62e-2 & 2.00e-2 \end{pmatrix} \end{array} \tag{A.16}$$

$$PWM_{GATAe\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.47 & 0.01 & 0.04 & 0.48 \\ 0.89 & 0.00 & 0.11 & 0.00 \\ 0.04 & 0.00 & 0.00 & 0.96 \\ 0.00 & 0.99 & 0.00 & 0.01 \\ 0.30 & 0.14 & 0.27 & 0.29 \\ 0.22 & 0.29 & 0.25 & 0.24 \end{pmatrix} \end{array} \tag{A.17}$$

$$DPWM_{GATAe\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{ccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 \end{array} \\ \begin{pmatrix} 4.34e-1 & 3.11e-2 & 2.86e-4 & 1.06e-3 & 7.26e-2 \\ 6.70e-7 & 7.07e-4 & 3.47e-2 & 2.41e-4 & 7.25e-2 \\ 9.22e-3 & 1.58e-4 & 8.63e-5 & 5.36e-5 & 8.54e-2 \\ 6.12e-5 & 8.51e-1 & 4.25e-4 & 2.75e-4 & 6.56e-2 \\ 1.39e-2 & 3.65e-4 & 6.78e-6 & 3.03e-1 & 3.41e-2 \\ 1.39e-6 & 2.59e-5 & 7.89e-4 & 1.42e-1 & 3.05e-2 \\ 7.98e-5 & 2.46e-5 & 4.01e-6 & 2.64e-1 & 4.53e-2 \\ 2.95e-6 & 3.74e-3 & 8.37e-5 & 2.85e-1 & 3.22e-2 \\ 3.67e-2 & 1.56e-3 & 9.65e-5 & 6.43e-5 & 6.32e-2 \\ 8.47e-4 & 5.02e-3 & 1.76e-4 & 1.15e-4 & 6.77e-2 \\ 1.81e-3 & 1.10e-3 & 3.42e-6 & 2.47e-4 & 6.06e-2 \\ 6.65e-5 & 1.05e-1 & 1.46e-4 & 8.67e-5 & 5.96e-2 \\ 4.46e-1 & 6.86e-6 & 9.18e-4 & 4.99e-4 & 6.83e-2 \\ 1.96e-3 & 7.42e-7 & 9.50e-1 & 1.58e-4 & 8.88e-2 \\ 5.49e-2 & 4.08e-8 & 2.89e-4 & 2.55e-4 & 7.88e-2 \\ 7.09e-7 & 1.35e-6 & 1.18e-2 & 3.54e-3 & 7.48e-2 \end{pmatrix} \end{array} \tag{A.18}$$

$$PWM_{Gsc\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.17 & 0.09 & 0.06 & 0.68 \\ 0.82 & 0.03 & 0.08 & 0.07 \\ 0.99 & 0.00 & 0.01 & 0.00 \\ 0.11 & 0.06 & 0.13 & 0.70 \\ 0.19 & 0.50 & 0.10 & 0.20 \\ 0.11 & 0.39 & 0.22 & 0.28 \\ 0.24 & 0.30 & 0.29 & 0.17 \end{pmatrix} \end{array} \tag{A.19}$$

$$DPWM_{Gsc\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 1.33e-1 & 8.08e-1 & 1.00e-1 & 1.90e-2 & 1.40e-2 & 1.89e-2 \\ 8.52e-3 & 5.23e-4 & 6.01e-2 & 4.53e-2 & 9.38e-2 & 3.31e-2 \\ 2.03e-2 & 1.09e-2 & 1.25e-1 & 1.40e-2 & 1.30e-2 & 2.69e-2 \\ 2.18e-2 & 9.05e-5 & 6.67e-1 & 6.64e-2 & 1.65e-2 & 1.45e-2 \\ 7.01e-2 & 2.84e-2 & 3.49e-3 & 2.75e-2 & 6.81e-2 & 9.48e-2 \\ 4.45e-3 & 6.42e-4 & 1.94e-3 & 2.72e-2 & 2.43e-1 & 1.18e-1 \\ 1.24e-2 & 1.56e-4 & 3.08e-3 & 1.52e-2 & 1.41e-1 & 1.17e-1 \\ 1.10e-2 & 1.59e-5 & 4.31e-4 & 2.93e-2 & 1.73e-1 & 6.67e-2 \\ 4.61e-2 & 8.09e-2 & 3.52e-3 & 3.35e-2 & 1.85e-2 & 1.35e-2 \\ 1.90e-3 & 5.05e-6 & 6.30e-3 & 5.66e-2 & 4.82e-2 & 6.84e-2 \\ 5.68e-3 & 7.95e-7 & 8.11e-3 & 1.29e-2 & 4.32e-3 & 3.66e-2 \\ 1.17e-2 & 2.45e-5 & 9.01e-3 & 5.54e-2 & 1.36e-2 & 4.71e-2 \\ 5.34e-1 & 7.03e-2 & 8.62e-4 & 1.16e-1 & 2.45e-2 & 8.01e-2 \\ 1.88e-2 & 2.35e-5 & 1.03e-2 & 3.01e-1 & 9.66e-2 & 8.43e-2 \\ 5.35e-2 & 3.61e-5 & 7.43e-4 & 5.98e-2 & 9.20e-3 & 9.00e-2 \\ 4.65e-2 & 3.31e-5 & 7.47e-5 & 1.20e-1 & 2.31e-2 & 9.05e-2 \end{pmatrix} \end{array} \tag{A.20}$$

$$PWM_{Gt\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.00 & 0.03 & 0.20 & 0.76 \\ 0.77 & 0.01 & 0.22 & 0.01 \\ 0.01 & 0.73 & 0.05 & 0.21 \\ 0.29 & 0.05 & 0.65 & 0.01 \\ 0.01 & 0.24 & 0.00 & 0.75 \\ 0.80 & 0.17 & 0.02 & 0.01 \\ 0.96 & 0.01 & 0.02 & 0.01 \end{pmatrix} \end{array} \tag{A.21}$$

$$DPWM_{Gt\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 3.92e-4 & 9.58e-3 & 3.25e-3 & 1.82e-3 & 1.05e-2 & 7.58e-1 \\ 4.54e-4 & 5.52e-1 & 1.17e-3 & 3.02e-2 & 8.15e-3 & 5.35e-3 \\ 2.13e-3 & 3.56e-2 & 7.93e-3 & 1.94e-3 & 2.13e-3 & 1.74e-2 \\ 9.32e-4 & 1.60e-1 & 2.32e-4 & 2.25e-1 & 8.78e-4 & 8.26e-3 \\ 2.63e-2 & 1.44e-3 & 2.04e-1 & 9.35e-4 & 1.93e-1 & 1.58e-1 \\ 3.35e-3 & 4.09e-3 & 3.33e-2 & 7.60e-3 & 1.18e-2 & 4.22e-3 \\ 8.95e-3 & 3.65e-3 & 4.57e-1 & 2.63e-3 & 4.91e-3 & 4.67e-3 \\ 6.51e-3 & 2.29e-3 & 8.60e-3 & 3.67e-2 & 2.15e-3 & 3.51e-3 \\ 1.57e-1 & 1.01e-2 & 1.64e-2 & 8.64e-3 & 3.40e-3 & 2.05e-2 \\ 5.67e-3 & 1.57e-1 & 8.50e-3 & 1.59e-1 & 3.12e-3 & 1.81e-3 \\ 1.30e-2 & 1.02e-2 & 2.94e-2 & 2.80e-3 & 4.08e-4 & 4.79e-3 \\ 1.07e-2 & 4.06e-2 & 2.22e-3 & 5.03e-1 & 1.69e-4 & 5.04e-3 \\ 5.86e-1 & 1.44e-3 & 7.77e-2 & 2.51e-4 & 6.10e-1 & 6.77e-3 \\ 4.34e-3 & 6.96e-3 & 1.54e-2 & 9.70e-3 & 1.27e-1 & 2.48e-4 \\ 1.67e-1 & 2.27e-3 & 1.32e-1 & 2.17e-4 & 1.65e-2 & 1.14e-3 \\ 7.39e-3 & 2.37e-3 & 2.62e-3 & 9.46e-3 & 5.45e-3 & 4.89e-4 \end{pmatrix} \end{array} \tag{A.22}$$

$$PWM_{Hb\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.18 & 0.11 & 0.00 & 0.71 \\ 0.05 & 0.14 & 0.13 & 0.68 \\ 0.09 & 0.06 & 0.05 & 0.80 \\ 0.03 & 0.22 & 0.04 & 0.71 \\ 0.31 & 0.01 & 0.02 & 0.66 \\ 0.32 & 0.22 & 0.33 & 0.14 \\ 0.32 & 0.26 & 0.17 & 0.25 \end{pmatrix} \end{array} \tag{A.23}$$

$$DPWM_{Hb\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 1.98e-2 & 7.83e-3 & 8.62e-3 & 1.14e-2 & 1.27e-1 & 1.08e-1 \\ 6.57e-2 & 1.78e-2 & 2.84e-2 & 5.58e-3 & 1.08e-3 & 8.84e-2 \\ 4.02e-2 & 2.11e-2 & 4.32e-2 & 1.49e-2 & 4.22e-4 & 5.85e-2 \\ 1.02e-1 & 3.01e-2 & 4.80e-2 & 2.06e-2 & 2.62e-4 & 8.60e-2 \\ 1.15e-2 & 3.29e-2 & 3.62e-2 & 4.65e-2 & 3.00e-3 & 6.64e-2 \\ 1.04e-1 & 3.27e-2 & 7.56e-2 & 1.18e-2 & 6.25e-3 & 5.44e-2 \\ 1.75e-3 & 7.62e-2 & 5.10e-3 & 2.88e-3 & 6.67e-3 & 3.92e-2 \\ 5.96e-2 & 7.74e-2 & 3.09e-2 & 1.44e-1 & 5.85e-4 & 5.48e-2 \\ 2.50e-4 & 5.67e-2 & 4.80e-2 & 9.22e-4 & 7.79e-3 & 2.26e-2 \\ 5.02e-4 & 3.66e-2 & 1.79e-2 & 3.01e-3 & 5.64e-4 & 1.41e-1 \\ 2.79e-4 & 7.25e-2 & 3.78e-2 & 1.45e-2 & 1.41e-3 & 5.92e-2 \\ 8.07e-7 & 7.26e-2 & 2.43e-2 & 2.41e-2 & 5.48e-5 & 1.02e-1 \\ 3.23e-2 & 4.24e-2 & 1.88e-2 & 2.16e-1 & 2.72e-1 & 3.09e-2 \\ 8.31e-2 & 2.73e-2 & 1.32e-1 & 5.12e-3 & 1.83e-1 & 2.88e-2 \\ 7.79e-2 & 2.15e-2 & 2.19e-2 & 1.33e-2 & 2.75e-1 & 2.48e-2 \\ 4.02e-1 & 3.74e-1 & 4.24e-1 & 4.65e-1 & 1.15e-1 & 3.56e-2 \end{pmatrix} \end{array} \tag{A.24}$$

$$PWM_{Hkb\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 \\ 0.02 & 0.96 & 0.02 & 0.01 \\ 0.01 & 0.01 & 0.97 & 0.02 \\ 0.00 & 0.02 & 0.02 & 0.96 \\ 0.01 & 0.00 & 0.98 & 0.01 \\ 0.81 & 0.07 & 0.05 & 0.07 \end{pmatrix} \end{array} \tag{A.25}$$

$$DPWM_{Hkb\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 4.58e-4 & 3.17e-4 & 1.11e-3 & 2.12e-3 & 4.84e-4 & 9.26e-3 \\ 4.36e-4 & 1.17e-3 & 1.41e-3 & 7.99e-3 & 5.68e-5 & 8.37e-3 \\ 1.85e-4 & 4.48e-3 & 1.51e-2 & 2.90e-3 & 1.58e-3 & 7.24e-3 \\ 2.69e-4 & 1.18e-4 & 1.65e-3 & 6.66e-3 & 3.05e-4 & 1.13e-2 \\ 2.56e-5 & 5.06e-5 & 6.48e-3 & 2.58e-3 & 2.65e-3 & 1.34e-3 \\ 1.88e-6 & 4.16e-4 & 5.30e-3 & 1.43e-3 & 8.14e-4 & 1.38e-3 \\ 6.93e-5 & 2.35e-5 & 8.74e-1 & 2.41e-3 & 1.58e-2 & 9.94e-4 \\ 1.96e-7 & 2.10e-5 & 1.40e-2 & 5.44e-3 & 2.27e-3 & 1.46e-3 \\ 1.22e-3 & 1.65e-2 & 4.88e-3 & 1.54e-3 & 6.95e-4 & 7.52e-1 \\ 4.33e-4 & 9.54e-1 & 3.31e-2 & 1.54e-2 & 2.50e-3 & 6.52e-2 \\ 9.94e-1 & 1.50e-2 & 1.37e-2 & 2.20e-2 & 2.26e-2 & 4.81e-2 \\ 4.06e-4 & 5.45e-3 & 1.36e-2 & 8.97e-1 & 8.07e-3 & 6.27e-2 \\ 4.37e-4 & 8.96e-4 & 1.93e-3 & 6.26e-3 & 1.14e-2 & 5.89e-3 \\ 1.68e-4 & 3.90e-4 & 7.06e-4 & 7.96e-3 & 1.64e-3 & 4.41e-3 \\ 1.83e-3 & 7.39e-4 & 5.00e-3 & 3.55e-3 & 9.22e-1 & 6.91e-3 \\ 4.80e-4 & 1.73e-5 & 7.90e-3 & 1.43e-2 & 7.23e-3 & 1.32e-2 \end{pmatrix} \end{array} \tag{A.26}$$

$$PWM_{Hkb\_rep2} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 \\ 0.02 & 0.95 & 0.02 & 0.01 \\ 0.01 & 0.01 & 0.97 & 0.01 \\ 0.00 & 0.03 & 0.05 & 0.92 \\ 0.01 & 0.01 & 0.96 & 0.02 \\ 0.80 & 0.04 & 0.06 & 0.10 \end{pmatrix} \end{array} \tag{A.27}$$

$$DPWM_{Hkb\_rep2} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 5.57e-4 & 4.11e-4 & 2.88e-3 & 4.46e-3 & 7.91e-4 & 9.96e-3 \\ 8.06e-4 & 1.57e-3 & 3.32e-3 & 1.42e-2 & 3.39e-4 & 1.09e-2 \\ 1.58e-3 & 6.82e-3 & 2.00e-2 & 5.82e-3 & 2.82e-3 & 1.18e-2 \\ 7.77e-4 & 5.54e-4 & 4.07e-3 & 9.94e-3 & 6.28e-4 & 1.45e-2 \\ 4.98e-5 & 6.71e-4 & 1.02e-2 & 4.57e-3 & 5.75e-3 & 4.47e-3 \\ 5.66e-5 & 8.57e-4 & 7.36e-3 & 2.56e-3 & 2.29e-3 & 3.97e-3 \\ 2.73e-4 & 3.21e-4 & 8.56e-1 & 4.41e-3 & 2.97e-2 & 3.28e-3 \\ 1.54e-5 & 3.25e-4 & 1.16e-2 & 7.15e-3 & 4.55e-3 & 4.02e-3 \\ 1.66e-3 & 2.19e-2 & 8.27e-3 & 2.73e-3 & 4.00e-3 & 7.08e-1 \\ 9.03e-4 & 9.38e-1 & 2.74e-2 & 2.87e-2 & 6.69e-3 & 3.54e-2 \\ 9.88e-1 & 1.78e-2 & 1.62e-2 & 4.11e-2 & 4.25e-2 & 4.86e-2 \\ 9.77e-4 & 6.48e-3 & 1.45e-2 & 8.31e-1 & 8.24e-3 & 9.20e-2 \\ 6.19e-4 & 1.70e-3 & 3.11e-3 & 1.13e-2 & 1.21e-2 & 1.15e-2 \\ 5.63e-4 & 9.28e-4 & 1.60e-3 & 1.40e-2 & 5.44e-3 & 1.17e-2 \\ 2.72e-3 & 1.52e-3 & 5.91e-3 & 6.90e-3 & 8.60e-1 & 1.24e-2 \\ 7.31e-4 & 3.60e-4 & 7.98e-3 & 1.12e-2 & 1.39e-2 & 1.79e-2 \end{pmatrix} \end{array} \tag{A.28}$$

$$PWM_{Hkb\_rep3} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.01 & 0.00 & 0.99 & 0.00 \\ 0.01 & 0.00 & 0.99 & 0.00 \\ 0.03 & 0.93 & 0.03 & 0.02 \\ 0.00 & 0.00 & 0.99 & 0.00 \\ 0.01 & 0.01 & 0.03 & 0.95 \\ 0.01 & 0.00 & 0.97 & 0.02 \\ 0.84 & 0.04 & 0.03 & 0.08 \end{pmatrix} \end{array} \tag{A.29}$$

$$DPWM_{Hkb\_rep3} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 5.63e-3 & 7.15e-3 & 6.27e-3 & 5.33e-3 & 8.74e-4 & 8.48e-3 \\ 7.34e-3 & 6.25e-3 & 1.28e-2 & 9.71e-3 & 9.42e-4 & 8.98e-3 \\ 1.06e-2 & 1.74e-2 & 2.27e-2 & 1.18e-2 & 6.13e-3 & 8.99e-3 \\ 9.45e-3 & 4.85e-3 & 6.64e-3 & 3.29e-3 & 2.10e-3 & 6.74e-3 \\ 5.51e-3 & 1.10e-3 & 3.13e-3 & 6.02e-3 & 7.21e-3 & 3.91e-3 \\ 3.99e-3 & 3.08e-3 & 2.63e-3 & 2.85e-3 & 5.51e-3 & 2.42e-3 \\ 3.71e-3 & 4.09e-3 & 8.36e-1 & 9.40e-3 & 1.14e-2 & 3.95e-3 \\ 5.12e-5 & 7.40e-3 & 5.86e-4 & 2.76e-3 & 8.71e-3 & 2.67e-3 \\ 6.88e-3 & 2.33e-2 & 8.65e-3 & 6.15e-3 & 4.61e-3 & 7.73e-1 \\ 3.39e-3 & 8.57e-1 & 2.96e-2 & 1.14e-2 & 8.61e-3 & 3.66e-2 \\ 9.43e-1 & 2.46e-2 & 2.40e-2 & 2.81e-2 & 2.80e-2 & 3.15e-2 \\ 4.87e-4 & 1.85e-2 & 1.46e-2 & 8.78e-1 & 1.29e-2 & 7.46e-2 \\ 3.34e-6 & 1.49e-2 & 3.30e-3 & 9.56e-3 & 9.59e-3 & 1.31e-2 \\ 2.58e-7 & 4.43e-4 & 1.32e-4 & 9.96e-3 & 4.43e-3 & 1.05e-2 \\ 8.26e-17 & 1.02e-2 & 1.81e-2 & 5.18e-3 & 8.74e-1 & 8.16e-4 \\ 8.48e-9 & 1.10e-11 & 1.07e-2 & 6.15e-4 & 1.48e-2 & 1.41e-2 \end{pmatrix} \end{array} \tag{A.30}$$

$$PWM_{Nub\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.09 & 0.05 & 0.06 & 0.80 \\ 0.87 & 0.04 & 0.04 & 0.05 \\ 0.13 & 0.06 & 0.05 & 0.76 \\ 0.07 & 0.05 & 0.67 & 0.21 \\ 0.07 & 0.72 & 0.06 & 0.15 \\ 0.76 & 0.06 & 0.08 & 0.11 \\ 0.64 & 0.09 & 0.09 & 0.18 \end{pmatrix} \end{array} \tag{A.31}$$

$$DPWM_{Nub\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 7.66e-2 & 8.97e-2 & 6.06e-2 & 4.19e-2 & 3.56e-2 & 3.97e-1 \\ 1.53e-2 & 4.29e-2 & 2.40e-2 & 2.55e-2 & 2.17e-2 & 5.42e-2 \\ 1.40e-2 & 3.79e-2 & 5.41e-2 & 1.32e-2 & 3.09e-2 & 5.53e-2 \\ 7.52e-3 & 5.37e-1 & 1.71e-1 & 3.78e-2 & 6.25e-2 & 1.12e-1 \\ 4.28e-2 & 2.34e-2 & 1.53e-2 & 2.12e-2 & 3.78e-1 & 2.96e-2 \\ 1.02e-2 & 2.43e-2 & 1.21e-2 & 1.68e-2 & 2.82e-2 & 2.10e-2 \\ 1.11e-2 & 2.27e-2 & 2.59e-2 & 1.10e-2 & 3.75e-2 & 2.43e-2 \\ 4.29e-3 & 2.55e-2 & 6.95e-2 & 2.02e-2 & 5.37e-2 & 2.07e-2 \\ 4.59e-2 & 3.12e-2 & 1.37e-2 & 2.23e-2 & 3.24e-2 & 3.94e-2 \\ 1.01e-2 & 2.11e-2 & 1.10e-2 & 2.37e-1 & 1.23e-2 & 2.83e-2 \\ 9.62e-3 & 2.26e-2 & 2.29e-2 & 2.03e-2 & 2.10e-2 & 3.03e-2 \\ 3.89e-3 & 2.32e-2 & 3.67e-2 & 4.91e-2 & 2.65e-2 & 2.50e-2 \\ 6.49e-1 & 3.35e-2 & 3.49e-2 & 1.64e-1 & 7.84e-2 & 5.64e-2 \\ 3.08e-2 & 1.46e-2 & 2.31e-2 & 7.38e-2 & 3.01e-2 & 3.52e-2 \\ 2.80e-2 & 1.66e-2 & 3.24e-1 & 3.29e-2 & 2.93e-2 & 3.12e-2 \\ 4.08e-2 & 3.38e-2 & 1.01e-1 & 2.13e-1 & 1.22e-1 & 4.04e-2 \end{pmatrix} \end{array} \tag{A.32}$$

$$PWM_{Nub\_rep2} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.07 & 0.01 & 0.03 & 0.89 \\ 0.99 & 0.00 & 0.00 & 0.01 \\ 0.12 & 0.01 & 0.02 & 0.85 \\ 0.06 & 0.00 & 0.75 & 0.19 \\ 0.03 & 0.74 & 0.06 & 0.16 \\ 0.76 & 0.06 & 0.05 & 0.13 \\ 0.69 & 0.07 & 0.07 & 0.17 \end{pmatrix} \end{array} \tag{A.33}$$

$$DPWM_{Nub\_rep2} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 6.84e-2 & 1.18e-1 & 9.33e-2 & 4.48e-2 & 1.93e-2 & 4.92e-1 \\ 1.33e-7 & 8.34e-3 & 7.32e-3 & 1.96e-2 & 8.63e-3 & 4.82e-2 \\ 3.15e-7 & 2.10e-2 & 5.63e-2 & 1.02e-3 & 2.07e-2 & 5.33e-2 \\ 2.11e-3 & 8.08e-1 & 1.91e-1 & 4.10e-2 & 8.67e-2 & 1.19e-1 \\ 1.26e-2 & 4.13e-8 & 8.33e-3 & 6.49e-3 & 4.36e-1 & 3.90e-2 \\ 4.47e-8 & 8.26e-9 & 5.44e-4 & 1.01e-3 & 3.46e-2 & 1.01e-2 \\ 3.56e-9 & 4.71e-7 & 4.00e-3 & 2.02e-3 & 2.96e-2 & 3.83e-5 \\ 1.28e-6 & 3.38e-7 & 7.59e-2 & 9.07e-3 & 7.69e-2 & 2.27e-2 \\ 2.77e-2 & 2.51e-4 & 8.21e-3 & 1.15e-2 & 3.39e-2 & 3.34e-2 \\ 2.56e-8 & 7.51e-5 & 5.97e-4 & 2.59e-1 & 3.69e-3 & 2.08e-2 \\ 3.37e-8 & 2.29e-5 & 1.01e-2 & 2.01e-2 & 2.09e-2 & 9.67e-4 \\ 4.91e-6 & 1.69e-6 & 2.96e-2 & 5.74e-2 & 2.45e-3 & 2.06e-2 \\ 8.76e-1 & 3.25e-2 & 2.93e-2 & 1.95e-1 & 9.66e-2 & 8.67e-2 \\ 3.67e-7 & 6.93e-4 & 1.50e-3 & 6.50e-2 & 2.07e-3 & 1.11e-2 \\ 1.84e-6 & 1.50e-6 & 3.87e-1 & 3.97e-2 & 1.07e-4 & 1.55e-2 \\ 1.28e-2 & 1.18e-2 & 9.72e-2 & 2.27e-1 & 1.28e-1 & 2.70e-2 \end{pmatrix} \end{array} \tag{A.34}$$

$$PWM_{Oc\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.12 & 0.09 & 0.05 & 0.74 \\ 0.78 & 0.02 & 0.09 & 0.11 \\ 0.99 & 0.00 & 0.01 & 0.00 \\ 0.09 & 0.03 & 0.19 & 0.70 \\ 0.17 & 0.58 & 0.07 & 0.17 \\ 0.08 & 0.41 & 0.10 & 0.40 \\ 0.28 & 0.28 & 0.24 & 0.20 \end{pmatrix} \end{array} \tag{A.35}$$

$$DPWM_{Oc\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 9.11e-2 & 7.81e-1 & 8.51e-2 & 2.17e-2 & 1.38e-2 & 2.72e-3 \\ 2.82e-3 & 9.78e-40 & 2.47e-2 & 4.13e-2 & 8.75e-2 & 2.68e-2 \\ 2.08e-2 & 4.59e-3 & 1.89e-1 & 2.19e-2 & 7.36e-3 & 1.61e-2 \\ 1.95e-2 & 8.78e-13 & 6.88e-1 & 8.75e-2 & 7.87e-3 & 7.56e-3 \\ 7.35e-2 & 2.30e-2 & 4.24e-41 & 1.37e-2 & 5.72e-2 & 1.38e-1 \\ 6.83e-4 & 9.82e-36 & 5.05e-40 & 1.20e-2 & 3.01e-1 & 1.41e-1 \\ 6.47e-3 & 7.20e-29 & 3.29e-37 & 5.44e-3 & 7.51e-2 & 1.23e-1 \\ 3.35e-3 & 4.05e-25 & 8.62e-40 & 1.01e-2 & 2.94e-1 & 9.94e-2 \\ 3.65e-2 & 8.47e-2 & 1.34e-46 & 4.21e-2 & 8.86e-3 & 1.02e-3 \\ 1.06e-4 & 1.15e-43 & 1.67e-4 & 9.16e-2 & 3.84e-2 & 3.52e-2 \\ 8.04e-3 & 1.19e-28 & 3.04e-3 & 7.61e-3 & 7.60e-8 & 1.72e-2 \\ 4.25e-3 & 4.71e-47 & 4.04e-3 & 7.32e-2 & 2.07e-3 & 2.25e-2 \\ 5.75e-1 & 1.07e-1 & 5.18e-9 & 9.70e-2 & 1.38e-2 & 5.76e-2 \\ 1.69e-2 & 3.77e-50 & 5.74e-3 & 3.34e-1 & 8.82e-2 & 1.38e-1 \\ 6.24e-2 & 2.58e-45 & 1.07e-39 & 4.26e-2 & 5.98e-6 & 9.92e-2 \\ 7.87e-2 & 7.71e-18 & 7.74e-13 & 9.79e-2 & 4.36e-3 & 7.46e-2 \end{pmatrix} \end{array} \tag{A.36}$$

$$PWM_{Tll\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.53 & 0.15 & 0.24 & 0.09 \\ 0.61 & 0.17 & 0.13 & 0.09 \\ 0.10 & 0.00 & 0.89 & 0.01 \\ 0.00 & 0.07 & 0.05 & 0.88 \\ 0.01 & 0.91 & 0.02 & 0.06 \\ 0.88 & 0.02 & 0.07 & 0.04 \\ 0.71 & 0.12 & 0.08 & 0.10 \end{pmatrix} \end{array} \tag{A.37}$$

$$DPWM_{Tll\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 3.05e-1 & 5.78e-2 & 1.36e-3 & 9.71e-5 & 7.54e-3 & 5.27e-1 \\ 8.57e-2 & 2.74e-3 & 5.49e-3 & 9.76e-5 & 2.96e-3 & 8.70e-2 \\ 6.69e-2 & 5.20e-1 & 2.09e-2 & 2.26e-4 & 2.83e-3 & 5.75e-2 \\ 4.38e-2 & 5.42e-3 & 8.48e-2 & 4.45e-4 & 1.94e-3 & 7.59e-2 \\ 8.36e-2 & 1.56e-2 & 2.78e-6 & 3.63e-3 & 7.64e-1 & 9.58e-3 \\ 1.43e-2 & 2.02e-3 & 6.82e-4 & 6.26e-2 & 1.39e-2 & 1.35e-2 \\ 3.66e-2 & 1.46e-1 & 8.55e-5 & 5.80e-3 & 6.32e-2 & 8.36e-3 \\ 4.87e-3 & 1.88e-3 & 4.02e-3 & 3.29e-2 & 3.08e-2 & 1.19e-2 \\ 1.36e-1 & 3.72e-2 & 9.96e-5 & 1.52e-3 & 1.67e-2 & 4.36e-2 \\ 5.35e-2 & 1.76e-3 & 6.39e-2 & 4.26e-2 & 4.52e-3 & 3.98e-2 \\ 4.91e-2 & 1.14e-1 & 4.35e-2 & 2.47e-3 & 4.23e-3 & 3.15e-2 \\ 3.06e-2 & 8.70e-3 & 7.64e-1 & 2.94e-2 & 2.39e-3 & 3.49e-2 \\ 4.91e-2 & 8.94e-3 & 1.25e-4 & 7.38e-3 & 4.73e-2 & 2.12e-2 \\ 2.22e-2 & 1.06e-3 & 1.03e-3 & 7.48e-1 & 9.59e-3 & 1.65e-2 \\ 1.14e-2 & 7.47e-2 & 2.32e-3 & 1.64e-2 & 2.20e-2 & 9.43e-3 \\ 7.39e-3 & 1.28e-3 & 7.96e-3 & 4.63e-2 & 6.01e-3 & 1.26e-2 \end{pmatrix} \end{array} \tag{A.38}$$

$$PWM_{Zld\_rep1} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.13 & 0.44 & 0.17 & 0.26 \\ 0.65 & 0.07 & 0.12 & 0.15 \\ 0.01 & 0.01 & 0.96 & 0.02 \\ 0.01 & 0.01 & 0.98 & 0.01 \\ 0.02 & 0.19 & 0.10 & 0.69 \\ 0.57 & 0.18 & 0.11 & 0.14 \\ 0.24 & 0.28 & 0.30 & 0.19 \end{pmatrix} \end{array} \tag{A.39}$$

$$DPWM_{Zld\_rep1} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 7.56e-2 & 5.13e-3 & 4.07e-3 & 6.64e-3 & 1.17e-2 & 1.07e-1 \\ 1.41e-2 & 7.07e-3 & 2.85e-3 & 1.90e-3 & 1.87e-2 & 1.27e-1 \\ 4.25e-2 & 6.07e-1 & 7.52e-3 & 3.56e-3 & 1.73e-3 & 1.35e-1 \\ 1.94e-2 & 1.30e-2 & 2.52e-3 & 4.62e-3 & 1.17e-2 & 8.43e-2 \\ 2.63e-1 & 3.75e-3 & 5.28e-3 & 4.13e-4 & 8.85e-2 & 6.91e-2 \\ 2.75e-2 & 3.90e-3 & 3.29e-2 & 9.97e-3 & 8.34e-2 & 6.11e-2 \\ 5.00e-2 & 6.35e-2 & 1.04e-2 & 3.55e-3 & 3.80e-2 & 4.21e-2 \\ 6.21e-2 & 6.19e-3 & 3.39e-3 & 3.92e-3 & 7.94e-2 & 6.93e-2 \\ 1.04e-1 & 3.67e-3 & 6.25e-3 & 2.34e-2 & 4.59e-2 & 2.53e-2 \\ 3.57e-2 & 3.67e-3 & 5.30e-3 & 1.77e-1 & 2.54e-2 & 5.63e-2 \\ 4.14e-2 & 1.16e-1 & 8.90e-1 & 9.20e-2 & 6.65e-3 & 2.58e-2 \\ 2.45e-2 & 6.62e-3 & 5.20e-3 & 6.58e-1 & 1.32e-2 & 4.10e-2 \\ 1.55e-1 & 6.20e-3 & 2.66e-3 & 1.70e-3 & 3.28e-1 & 4.26e-2 \\ 1.58e-2 & 4.82e-3 & 8.09e-4 & 4.51e-3 & 1.02e-1 & 4.34e-2 \\ 2.41e-2 & 1.43e-1 & 1.91e-2 & 5.12e-3 & 6.29e-2 & 3.38e-2 \\ 4.54e-2 & 6.92e-3 & 2.21e-3 & 3.84e-3 & 8.22e-2 & 3.76e-2 \end{pmatrix} \end{array} \tag{A.40}$$

$$PWM_{Zld\_rep2} = \begin{array}{c} \\ pos\,1 \\ pos\,2 \\ pos\,3 \\ pos\,4 \\ pos\,5 \\ pos\,6 \\ pos\,7 \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.07 & 0.51 & 0.10 & 0.32 \\ 0.78 & 0.05 & 0.08 & 0.09 \\ 0.00 & 0.01 & 0.96 & 0.02 \\ 0.00 & 0.00 & 0.99 & 0.00 \\ 0.02 & 0.11 & 0.06 & 0.80 \\ 0.73 & 0.13 & 0.08 & 0.07 \\ 0.22 & 0.27 & 0.36 & 0.15 \end{pmatrix} \end{array} \tag{A.41}$$

$$DPWM_{Zld\_rep2} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ CC \\ CG \\ CT \\ GA \\ GC \\ GG \\ GT \\ TA \\ TC \\ TG \\ TT \end{array} \begin{array}{c} \begin{array}{cccccc} pos\,1 & pos\,2 & pos\,3 & pos\,4 & pos\,5 & pos\,6 \end{array} \\ \begin{pmatrix} 4.98e-2 & 3.62e-3 & 1.31e-3 & 1.52e-3 & 1.37e-2 & 1.37e-1 \\ 1.62e-2 & 9.04e-3 & 1.37e-3 & 3.34e-4 & 1.74e-2 & 1.67e-1 \\ 3.42e-2 & 7.32e-1 & 4.58e-3 & 1.72e-5 & 6.45e-3 & 2.21e-1 \\ 1.79e-2 & 1.58e-2 & 1.83e-4 & 1.97e-3 & 1.34e-2 & 9.28e-2 \\ 3.38e-1 & 4.73e-3 & 8.41e-3 & 7.04e-6 & 6.64e-2 & 4.91e-2 \\ 2.13e-2 & 3.91e-3 & 2.04e-2 & 1.17e-2 & 5.48e-2 & 3.89e-2 \\ 3.50e-2 & 4.62e-2 & 1.14e-2 & 1.66e-5 & 3.47e-2 & 3.80e-2 \\ 3.93e-2 & 4.38e-3 & 4.35e-4 & 1.88e-3 & 4.88e-2 & 4.16e-2 \\ 6.34e-2 & 2.69e-3 & 2.32e-3 & 2.30e-2 & 3.41e-2 & 1.79e-2 \\ 2.66e-2 & 1.56e-3 & 2.21e-3 & 1.12e-1 & 3.05e-2 & 4.14e-2 \\ 3.39e-2 & 7.59e-2 & 9.27e-1 & 5.75e-2 & 1.49e-2 & 2.34e-2 \\ 2.50e-2 & 4.63e-3 & 3.90e-4 & 7.89e-1 & 2.21e-2 & 2.96e-2 \\ 2.16e-1 & 5.47e-3 & 1.26e-5 & 1.25e-6 & 4.67e-1 & 2.67e-2 \\ 2.09e-2 & 1.62e-3 & 1.14e-5 & 9.06e-4 & 8.05e-2 & 2.85e-2 \\ 2.66e-2 & 8.53e-2 & 2.00e-2 & 9.71e-6 & 4.96e-2 & 2.13e-2 \\ 3.59e-2 & 3.29e-3 & 4.07e-6 & 3.32e-4 & 4.52e-2 & 2.58e-2 \end{pmatrix} \end{array} \tag{A.42}$$

A.2 Python code

The following code can be used to automatically fit titration curves in a HiP-FA experiment. For details see 4.1.7

```
"""
Created on Fri Oct 26 14:40:08 2018

@author: schnepf
"""

from scipy.optimize import least_squares
import matplotlib.pyplot as plt
import numpy as np
import os
from shutil import copyfile

#######
Version = 1.2011


#######


class LabviewFitter:
def __init__(self, filename, consensus_pos, conc_lab_ref=[1.4], plotting=False,
↪  fit_individual=False,
subtract_water=False, water=None, degree_fit=3, water_dest_file=None):
"""
initialize the class with data and do intial fitting
:param filename: path to the text file containing the titration curves produced
↪  by the labview program(str)
:param consensus_pos: positions of consensus sequences located on the 96 well
↪  plate (list of int)
:param conc_lab_ref: concentration of the labelled reference sequence in nM
↪  (float)
:param plotting: should the plots be displayed (bool)
:param fit_individual:  fit the consensus curves individually or  fit averaged
↪  values (bool)
:param subtract_water: subtract a polynomial fit of the water curves from all
↪  curves (normalization) (bool)
:param water: positions of water wells (list of int)
:param degree_fit: degree of polynomial to smooth the water curves (int)
:param water_dest_file: path to save the plot of  the water fit (str)
"""
self.consensus_pos = consensus_pos
self.Lst = conc_lab_ref
# read in the data
self.concentrations, self.FA_vals = read_curves(filename)
self.subtract_water = subtract_water
# Do polynomial fitting to smooth water curves
if subtract_water:
```

```
self.poly = np.polyfit(self.concentrations, self.FA_vals[water, :].mean(axis=0),
↪  degree_fit)
self.p1 = np.poly1d(self.poly)
# plot the curves for control
for w in water:
plt.plot(self.concentrations, self.FA_vals[w], '+')
plt.plot(self.concentrations, self.p1(self.concentrations), '--', linewidth=3)
plt.title('water control fit')
plt.show()
# save plot if desired
if water_dest_file:
plt.savefig(water_dest_file)

# define fitting functions and initial parameters
fitfun_FA = lambda p, x, fix: FA_fit(x, Kd2=p[0], B=p[1], C=p[2], Rt=p[3],
↪  Lst=fix[0], Kd1=p[4])
bounds_default = ([0, 30, 5, 1, .01], [1e4, 300, 100, 500, 100])
errfun_FA = lambda p, x, y, fix: FA_fit(x, Kd2=p[0], B=p[1], C=p[2], Rt=p[3],
↪  Lst=fix[0],
Kd1=p[4]) - y
p_init = [50, 130, 20, 50, 2]

self.Lst = conc_lab_ref[0]
fix = conc_lab_ref
# do the fitting for each consensus curve individually
if fit_individual:
consensus_results = []
for n in consensus_pos:
if subtract_water:
# subtract the fit but add the offset again
FA_vals = self.FA_vals[n] - self.p1(self.concentrations) + self.poly[-1]
else:
FA_vals = self.FA_vals[n]
out = least_squares(errfun_FA, p_init, args=(self.concentrations, FA_vals, fix),
↪  bounds=bounds_default)
if plotting:
plt.plot(self.concentrations, fitfun_FA(out['x'], self.concentrations, fix),
↪  '--')
plt.plot(self.concentrations, self.FA_vals[n], '+')
plt.xscale('symlog')
print (out['x'])
consensus_results.append(out['x'])
self.consensus_results = np.array(consensus_results)
self.consensus_outliers = np.ndarray(self.consensus_results.shape, bool)
# define outliers based on meas absolute deviation. One failed curve would have a
↪  too strong impact
for n, line in enumerate(self.consensus_results.T):
self.consensus_outliers[:, n] = mad_based_outlier(line)
self.mean_cons =
↪  self.consensus_results[np.where(self.consensus_outliers.sum(axis=1) <
↪  2)].mean(axis=0)
else:
```

```
# do a fit on the averaged values, single outlier points are averaged out
FA_cons_mean = self.FA_vals[self.consensus_pos, :].mean(axis=0)
out = least_squares(errfun_FA, p_init, args=(self.concentrations, FA_cons_mean,
↪  fix))
self.mean_cons = out['x']

def save_kds(self, results, destname, water=(82, 83, 84, 85, 86),
↪  default_len_cons=8):
"""
function to save the results
:param results: array with KDs and B values fitted (np.array)
:param destname: path to the destination (str)
:param water: water samples to exclude from output (list/iterable of ints)
:return: nothing (None)
"""
# check if the length of the consensus wells differs, ask for water wells
↪  (deviation from default)
if len(self.consensus_pos) != default_len_cons:
try:
import Tkinter, tkSimpleDialog
water_update = tkSimpleDialog.askstring('Change water positions?',
'If water positions are not ' + str(water) + '\n\n '
'insert them here (separator: , ) otherwise press cancel to continue')
if water_update is not None:
water = [int(x) for x in water_update.split(',')]
except ImportError:
print('Warning: length of Consensus and water positions might not fit')
results_out = results[[x for x in range(len(results)) if x not in water]]
results_out = np.c_[
results_out[:, 1] + self.mean_cons[2], np.repeat(self.mean_cons[3],
↪  len(results_out)), results_out[:,
0], results_out[:,
1]]
np.savetxt(destname, results_out, fmt='%5.3f', delimiter='\t')

def fit_curves_to_cons(self, tolerance, Lst=None, plotting=False, n_cols=8,
↪  pic_path=None, fig_size=(160, 100),
y_lim=(0, 0), kde_weighted=False, subtract_water=False):
"""
function to fit the titration curves produced by labview, parameters estimated on
↪  consensus curves
:param tolerance: relative deviation of B compared to the consensus values
↪  (float)
:param Lst: concentration of labeled oligomer in nM (float)
:param plotting: should the fits be plotted? (bool)
:param n_cols: number of columns in the multifit plot (int)
:param pic_path: path to save the plot (str)
:param fig_size: size of the plot in inches (tuple of ints)
:param y_lim: limits of the y axis in mili polarization (tuple of ints)
:return: fit values KD, B (np.array)
"""
# check input parameters and process them
```

```
if np.sum(y_lim) == 0:
if subtract_water:
y_lim = (
np.nanmin(self.FA_vals[[self.consensus_pos], :] -
↪  self.p1(np.max(self.concentrations)) + self.poly[-1]),
np.nanmax(self.FA_vals[[self.consensus_pos], :] +
↪  self.p1(np.max(self.concentrations)) - self.poly[-1]))
else:
y_lim = (
np.nanmin(self.FA_vals[[self.consensus_pos], :]),
↪  np.nanmax(self.FA_vals[[self.consensus_pos], :]))

fix = [Lst, self.mean_cons[2], self.mean_cons[3], self.mean_cons[4]]

# define the fitting and error functions
fitfun2_FA = lambda p, x, fix: FA_fit(x, Kd2=p[0], B=p[1], C=fix[1], Rt=fix[2],
↪  Lst=fix[0], Kd1=fix[3])
if not kde_weighted:
errfun2_FA = lambda p, x, y, fix: FA_fit(x, Kd2=p[0], B=p[1], C=fix[1],
↪  Rt=fix[2], Lst=fix[0],
Kd1=fix[3]) - y
else:
errfun2_FA = lambda p, x, y, fix: (FA_fit(x, Kd2=p[0], B=p[1], C=fix[1],
↪  Rt=fix[2], Lst=fix[0],
Kd1=fix[3]) - y) / self.kde.kde_weight(x)
# initial parameters: kd 5x consensus fited, B value of consensus
p_init = [self.mean_cons[0] * 5, self.mean_cons[1]]
# loop over all curves and fit the ones being a sample
results = []
if plotting:
fig = plt.figure(figsize=fig_size, dpi=50)
n_plot = 0
n_line = 1
for n in range(len(self.FA_vals)):
# check with variable sample if the pattern of the data points looks like a nile
↪  blue curve-->no sample
sample = np.nanmedian(self.FA_vals[n][:int(len(self.FA_vals[n]) / 15)]) > \
(np.nanmedian(self.FA_vals[n][int(len(self.FA_vals[n]) / 15):])) * .9
if plotting:
n_plot += 1
ax = fig.add_subplot(n_line, n_cols, n_plot % n_cols + 1)
# based on the assigned type, process the data accordingly
if sample:
if subtract_water:  # subtract water baseline based on polynomial fit
FA_vals = self.FA_vals[n] - self.p1(self.concentrations) + self.poly[-1]
else:
FA_vals = self.FA_vals[n]
try:
out2 = least_squares(errfun2_FA, p_init, args=(self.concentrations, FA_vals,
↪  fix), bounds=(
[0, self.mean_cons[1] * (1 - tolerance)], [np.inf, self.mean_cons[1] * (1 +
↪  tolerance)]))
```

```
except ValueError:  # ValueError raised if NAs are in the FA_vals --> remove NAs
conc_temp = \
list(zip(*filter(lambda x: not np.isnan(x[0]), zip(FA_vals,
↪  self.concentrations))))[1]
try:
out2 = least_squares(errfun2_FA, p_init,
args=(conc_temp, FA_vals[~np.isnan(FA_vals)], fix),
bounds=([0, self.mean_cons[1] * (1 - tolerance)],
[np.inf, self.mean_cons[1] * (1 + tolerance)]))
except ValueError:
out2 = {'x': [np.nan for x in p_init]}
print ('Problem in curve number ' + n)
if plotting:
_ = ax.plot(self.concentrations, FA_vals, '+', color='orange')
_ = ax.plot(self.concentrations, fitfun2_FA(out2['x'], self.concentrations, fix),
↪  '--', linewidth=3)
_ = ax.set_title(str(n))
_ = ax.set_ylim(y_lim)

results.append(out2['x'])
else:
if plotting:
_ = ax.plot(self.concentrations, self.FA_vals[n], '*', color='blue')
# _=ax.set_ylim(y_lim)
if plotting:
_ = ax.set_xscale('log')
n_plot += 1
if n_plot % n_cols == 0:
n_line += 1
n_ax = len(fig.axes)
for i in range(n_ax):
fig.axes[i].change_geometry(n_line, n_cols, i + 1)
if plotting:
_ = plt.tight_layout()
# plt.show()
if pic_path is not None:
plt.savefig(pic_path)
return np.array(results)


def mad_based_outlier(points, thresh=3.5):
"""
# functions found @
https://stackoverflow.com/questions/22354094/
pythonic-way-of-detecting-outliers-in-one-dimensional-observation-data
# used to classify data as outliers which are [thresh] z-scores away from the
↪  median
:param points: data points to apply outlier test to (np.array)
:param thresh: z-score equivalent, threshold for classification as outlier
:return:
"""
if len(points.shape) == 1:
```

```
points = points[:, None]
median = np.median(points, axis=0)
diff = np.sum((points - median) ** 2, axis=-1)
diff = np.sqrt(diff)
med_abs_deviation = np.median(diff)

modified_z_score = 0.6745 * diff / med_abs_deviation

return modified_z_score > thresh

def longest(list_list):
"""
function to pick concentrations not containing NA values
:param list_list: read in list of labview concentrations (list of lists of
↪  floats)
:return: concentrations to use (np.array)
"""
list_list = np.array(list_list)
lengths = [len(x) for x in list_list]
return list_list[np.where(lengths == np.max(lengths))[0][0]]

def read_curves(filename, min_conc=None, max_conc=None):
"""
function to read curves produced by the labview program (Protein-Binding Assay)
:param filename: file containing the labview curves
:param min_conc: minimal concentration to consider (if early cycles are
↪  problematic)
:param max_conc: maximal concentration to consider (if late cycles are
↪  problematic)
:return: arrays of concentration and the FA values for all curves
"""
# read the data and extract the relevant curves
with open(filename, 'r') as f:
lines = f.readlines()
concentrations = lines[1::4]

FA = lines[2::4]
# check for formatting
try:
concentrations = [[float(x) for x in y.split('\t')] for y in concentrations]
except ValueError:
concentrations = [[float(x.replace(',', '.')) for x in y.split('\t')] for y in
↪  concentrations]
concentrations = longest(concentrations)  # avoid picking a concentrations line
↪  containing NAs
try:
FA = [[float(x) for x in y.split('\t')] for y in FA]
except ValueError:
FA = [[float(x.replace(',', '.')) for x in y.split('\t')] for y in FA]
max_len = np.max([len(x) for x in FA])
# extract FA values
for i, line in enumerate(FA):
```

```
if len(line) < max_len:
line += [np.nan] * (max_len - len(line))
FA[i] = np.array(line)
else:
FA[i] = np.array(line)
# filter for critical concentrations
if min_conc is not None:
for n, sample in enumerate(FA):
FA[n] = [x[0] for x in zip(sample, concentrations) if x[1] > min_conc]
concentrations = filter(lambda x: x > min_conc, concentrations)
if max_conc is not None:
for n, sample in enumerate(FA):
FA[n] = [x[0] for x in zip(sample, concentrations) if x[1] < max_conc]
concentrations = filter(lambda x: x < max_conc, concentrations)
FA = np.array(FA)
return concentrations, FA

# define function to do the competitor fits for the FA curves
# break down the function into sub-functions
def d(Kd1, Kd2, Lst, Lt, Rt):
d_val = Kd1 + Kd2 + Lst + Lt - Rt
return d_val


def e(Lt, Rt, Kd1, Lst, Kd2):
e_val = (Lt - Rt) * Kd1 + (Lst - Rt) * Kd2 + Kd1 * Kd2
return e_val


def f(Kd1, Kd2, Rt):
f_val = -1 * Kd1 * Kd2 * Rt
return f_val


def theta(d, e, f):
argum = (-2 * d ** 3 + 9 * d * e - 27 * f) / (2 * np.sqrt((d ** 2 - 3 * e) ** 3))
theta_val = np.arccos(argum)
return theta_val


def FA(B, C, d, e, f, theta, Kd1):
numerator = 2 * np.sqrt(d ** 2 - 3 * e) * np.cos(theta / 3) - d
denominator = 3 * Kd1 + 2 * np.sqrt(d ** 2 - 3 * e) * np.cos(theta / 3) - d
FA_val = B + C * (numerator / denominator)
return FA_val


def FA_fit(conc, Kd2, B, C, Rt, Lst, Kd1):
Lt = conc
e_tmp = e(Lt, Rt, Kd1, Lst, Kd2)
d_tmp = d(Kd1, Kd2, Lst, Lt, Rt)
f_tmp = f(Kd1, Kd2, Rt)
```

```
theta_tmp = theta(d_tmp, e_tmp, f_tmp)
FA_tmp = FA(B, C, d_tmp, e_tmp, f_tmp, theta_tmp, Kd1)
return FA_tmp

def main(return_obj=False):
"""
main function, running the fitting in an interactive way
:param return_obj: return the fitted data (bool)
:return: instance of LabviewFitter object if requested, saves result to file of
↪  choice
"""
# interactive pop-up windows to check for parameters or confirm default
↪  parameters
try:
import Tkinter, tkFileDialog, tkSimpleDialog, tkMessageBox
root = Tkinter.Tk()
root.withdraw()
file_path =
↪  tkFileDialog.askopenfilename(initialdir="P:\\TF-DNA-Binding\\FA\\Data\\",
title='select FA curves file',
filetypes=[('text files', '*.txt')])
consensus_pos = tkSimpleDialog.askstring('Change consensus positions?',
'If consensus positions are not [11, 12, 35, 36, 48, 71, 72, 95]\n\n '
'insert them here (separator: , ) otherwise press cancel to continue')
tolerance = tkSimpleDialog.askfloat('Set tolerance for B', 'Enter the maximal
↪  factor between \n\n the '
'fitted consensus B value and B values for sequences')
if tolerance is None:
tolerance = 100
plotting = tkMessageBox.askyesno('Plot the curves?')
data_path = '\\'.join(file_path.split('/')[:-1]) + '\\'
data_file = file_path.split('/')[-1]
if not os.path.exists(data_path + '\\Python_fit'):
os.mkdir(data_path + '\\Python_fit')
dest_dir = data_path + '\\Python_fit\\'
copyfile(file_path, dest_dir + data_file)
tf_name = tkSimpleDialog.askstring('TF - name', 'enter name of transcription
↪  factor')
subtract_water = tkMessageBox.askyesno('subtract polynomial fit of the water
↪  curves?')
if subtract_water:
water = tkSimpleDialog.askstring('Change water positions?',
'If water positions are not [90,91,92,93,94]\n\n insert  '
'them here (separator: , ) otherwise press cancel to continue')
water_dest_file = dest_dir + '\\water_control.jpg'
if water is None:
water = [90, 91, 92, 93, 94]
else:
try:
water = [int(x) for x in water.split(',')]
except ValueError:
water = [90, 91, 92, 93, 94]
```

```
else:
water = None
water_dest_file = None
if consensus_pos is None:
consensus_pos = [11, 12, 35, 36, 48, 71, 72, 95]
else:
consensus_pos = [int(x) for x in consensus_pos.split(',')]
root = Tkinter.Tk()
root.withdraw()
except ImportError:
file_path =
↪  'P:\\TF-DNA-Binding\\FA\Data\\181023_Eip93_part1\\strong_binders_curves.txt'
consensus_pos = [11, 12, 35, 36, 48, 71, 72, 95]
consensus_pos += [90]
tf_name = 'Eip93'
plotting = False
dest_dir = 'P:\\TF-DNA-Binding\\FA\Data\\181023_Eip93_part1\\'
tolerance = 100
# initialize the class the do the consensus fits
HIP_FA_fitter = LabviewFitter(filename=file_path, consensus_pos=consensus_pos,
↪  conc_lab_ref=[1.4],
fit_individual=True, subtract_water=subtract_water,
water=water, water_dest_file=water_dest_file)
# fit all curves
results_fix_c = HIP_FA_fitter.fit_curves_to_cons(tolerance=tolerance, Lst=1.4,
↪  plotting=plotting, n_cols=12,
pic_path=dest_dir + tf_name + '_fits.jpg',
fig_size=(130, 50))
# save the results
HIP_FA_fitter.save_kds(results_fix_c, dest_dir + tf_name + '_python_fits.txt')
# write parameter file for documentation
with open(dest_dir + 'parameters.txt', 'w') as f:
f.write('Version :' + str(Version))
f.write('\n### Parameters/Fit values:\n\ntolerance = ')
f.write(str(tolerance))
f.write('\nKd consensus = ' + str(HIP_FA_fitter.mean_cons[0]))
f.write('\nB = ' + str(HIP_FA_fitter.mean_cons[1]) + '+/- ' + str(tolerance *
↪  HIP_FA_fitter.mean_cons[1]))
f.write('\nC = ' + str(HIP_FA_fitter.mean_cons[2]))
f.write('\nRt = ' + str(HIP_FA_fitter.mean_cons[3]))
f.write('\nKd1 = ' + str(HIP_FA_fitter.mean_cons[4]))

f.write('\nLst = ' + str(HIP_FA_fitter.Lst))
if return_obj:
return HIP_FA_fitter

# run the file
if __name__ == '__main__':
main()
```

A.3 Sequences

A.3.1 Amino acid sequences of TFs

This subseqction provides the amino acid sequences encoding the TFs used in this study.

Bcd-BD MPKPEELPDSLVMRRPRRTRTTFTSSQIAELEQHFLQGRYLTAPRL ADLSAK-
LALGTAQVKIWFKNRRRRHKIQSDQHKDQSYEGMPLSP

Hb-BD NIRMPIYNSHGKMKNYKCKTCGVVAITKVDFWAHTRTHMKPDKIL QCP-
KCPFVTEFKHHLEYHIRKHKNQKPFQCDKCSYTCVNKSMLNS HRKSHSSVYQYRCAD-
CDYATKYCHSFKLHLRKYGHKPGMVLDEDGTPNPS

Gt-BD ATAANSGISSGSQVKDAAYYERRRKNNAAAKKSRDRRRIKEDEIA
IRAAYLERQNIELLCQIDALKVQLAAFTSAKVTTA

Hkb-BD QLKALNSRKQRPKKFKCPNCDVAFSNNGQLKGHIRIHTGERPFKCD VNTCGK-
TFTRNEELTRHKRIHTGLRPYPCSACGKKFGRRDHLKKHM KTHMPQERQLGPSIFVPMY-
SYLYG

Fkh-BD AKPPYSYISLITMAIQNNPTRMLTLSEIYQFIMDLFPFYRQNQQRW QN-
SIRHSLSFNDCFVKIPRTPDKPGKGSFWTLHPDSGNMFENGCYL

D-BD AGMHSLATSPGQEGHIKRPMNAFMVWSRLQRRQIAKDNPKMHNSEI SKRL-
GAEWKLLAESEKRPFIDEAKRLRALHMKEHPDYKYRPRRKPKNPLTAGPQGGL

Oc-BD AVGFSQGMWGVNTRKQRRERTTFTRAQLDVLEALFGKTRYPDIFMR EE-
VALKINLPESRVQVWFKNRRAKCRQQLQQQQQSNSLSSSKNA

Gsc-Bd QHHLSHLGHGPPPKRKRRHRTIFTEEQLEQLEATFDKTHYPDVVL RE-
QLALKVDLKEERVEVWFKNRRAKWRKQKREEQERLRKLQEEQC

Pdm2-BD MTSTLSSTPESILGRRRKKRTSIETTVRTTLEKAFLMNCKPTSEE ISQLSERL-
NMDKEVIRVWFCNRRQKEKRINPSLDLDSPTGTPLSS

Nub-BD AALQATVSTPEIIGRRRKKRTSIETTIRGALEKAFLANQKPTSEE
ITQLADRLSMEKEVVRVWFCNRRQKEKRINPSLDSPTGADDDESS

Zelda-BD TTLPSGRIKCLECDKEFTKNCYLTQHNKSFHSGEYPFRCQKCGKR
FQSEDVYTTHLGRHRTQDKPHKCELCPKQFHHKTDLRRHVEAIHT GLKQHM-
CDICEKGFCRKDHLRKHLETHNRPRVVGKKSAA

Tll-BD SPAASSRILYHVPCKVCRDHSSGKHYGIYACDGCAGFFKRSIRRS RQYVCK-
SQKQGLCVVDKTHRNQCRACRLRKCFEVGMNKDAVQHER GPRNSTLR

Eip93f-BD AQEALGKGTRPKRGKYRNYDRDSLVEAVKAVQRGEMSVHRAGSYY GVPH-
STLEYKVKERHLMRPRKREPKPQPDLVGLT

GATAe MVCKTISPSVNMQLKMEQQTTQQQQQQQQQQQQQQLQQQQHQAL TKQQLQLLDKIKLESSNGADQLAQQTANNLDEQQEQQQQHQQQAA TSVGVVVQT-GQAGVSEPEEQYVVVPRNQRRILTTAGTLELNEARE GEPSTNASNASSGSASD-SHIEYQRSAHQSPGATHYVQMAPRNAEV TEQVGAAAGAPPGTIFAYPIICNGDDVAAIKI-ETLEKGEATGESQ QQQQQLQQHQHQQQQQCPTPNGASYGETIVISSEAEALQHH-HQQQ QQHQQQQHQQQHHQHQAAAAASAAAQTVHIATSSHGGTVRFVTED VRFT-TAGPETSASNMYYDVPVVDGSVHANESKTYADLGNAYAPFP PSSSFSSNSYAATLQQGN-TIYSVPGTGQFLAKSESGLNQTGLLRQ TGPATFQTISFEGGNGIEPLWASPAPPEYQSVQF-SNFHPQVIDEY GSGNMSTSHWPPASSIGQYDGSLVTASSTSSPNHELKCENCHGPF LRKGSEYFCPNCPAFMRMAPRITQRQAKPKAAAAPNNRRNGVTCA NCQTNSTTLWR-RNNEGNPVCNACGLYYKLHNMNRPLSMKKEGIQK RKRKPKNNGGAPMHRAPLPSM-SQGVNLMANSPLYPSQVPVSMLNS QLNSQQNSSPELHDMSTTGQAGGQRVVSISLNAT-APPTPDGTLNM SARHHVTGESHSPYSQQSTPQSQSPHLPGTVPINRQIVQPVPTIE SSRSS-NTELTPSVITRTGLPERSSNN

Bibliography

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., Rohs, R. and Mann, R. S. (2015). Deconvolving the recognition of dna shape from sequence, *Cell* **161**(2): 307–18.

Affolter, M., Slattery, M. and Mann, R. S. (2008). A lexicon for homeodomain-dna recognition., *Cell* **133**: 1133–5.

Baird-Titus, J. M., Clark-Baldwin, K., Dave, V., Caperelli, C. A., Ma, J. and Rance, M. (2006). The solution structure of the native k50 bicoid homeodomain bound to the consensus taatcc dna-binding site, *J Mol Biol* **356**(5): 1137–51.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep Iii, P. W. and Bulyk, M. L. (2006). Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities, *Nat Biotech* **24**(11): 1429–1435.

Brennan, R. G. and Matthews, B. W. (1989). The helix-turn-helix dna binding motif., *The Journal of biological chemistry* **264**: 1903–6.

Brglin, T. R. and Affolter, M. (2016). Homeodomain proteins: an update, *Chromosoma* **125**(26464018): 497–521.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4901127/*

Cao, H., Widlund, H. R., Simonsson, T. and Kubista, M. (1998). Tgga repeats impair nucleosome formation, *J Mol Biol* **281**(2): 253–60.

Chiu, T. P., Rao, S., Mann, R. S., Honig, B. and Rohs, R. (2017a). Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-DNA binding, *Nucleic Acids Res.* **45**(21): 12565–12576.

Chiu, T. P., Rao, S., Mann, R. S., Honig, B. and Rohs, R. (2017b). Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-dna binding, *Nucleic Acids Res* **45**(21): 12565–12576.

Choo, Y. and Klug, A. (1994). Selection of dna binding sites for zinc fingers using rationally randomized dna reveals coded interactions., *Proceedings of the National Academy of Sciences of the United States of America* **91**: 11168–72.

Coulocheri, S. A., Pigis, D. G., Papavassiliou, K. A. and Papavassiliou, A. G. (2007). Hydrogen bonds in protein-dna complexes: where geometry meets plasticity., *Biochimie* **89**: 1291–303.

Cui, F. and Zhurkin, V. B. (2010). Structure-based analysis of dna sequence patterns guiding nucleosome positioning in vitro, *Journal of biomolecular structure & dynamics* **27**(6): 821–841.

de Mendoza, A. and Seb-Pedrs, A. (2019). Origin and evolution of eukaryotic transcription factors, *Current Opinion in Genetics & Development* **58-59**: 25 – 32. Evolutionary genetics.
URL: *http://www.sciencedirect.com/science/article/pii/S0959437X1830128X*

Dickerson, R. E. (1989). Definitions and nomenclature of nucleic acid structure parameters., *The EMBO journal* **8**: 1–4.

Drew, H. R. and Calladine, C. R. (1987). Sequence-specific positioning of core histones on an 860 base-pair dna. experiment and theory, *J Mol Biol* **195**(1): 143–73.

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 Research* **42**(1): 430–441.

El Hassan, M. A. and Calladine, C. R. (1996). Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in dna, *Journal of Molecular Biology* **259**(1): 95–103.

Etheve, L., Martin, J. and Lavery, R. (2016). Protein-dna interfaces: a molecular dynamics analysis of time-dependent recognition processes for three transcription factors., *Nucleic acids research* **44**: 9990–10002.

Ezer, D., Zabet, N. R. and Adryan, B. (2014). Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression, *Computational and Structural Biotechnology Journal* **10**(17): 63 – 69.
URL: *http://www.sciencedirect.com/science/article/pii/S2001037014000142*

Fairall, L., Schwabe, J. W., Chapman, L., Finch, J. T. and Rhodes, D. (1993). The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/dna recognition., *Nature* **366**: 483–7.

Fedotova, A. A., Bonchuk, A. N., Mogila, V. A. and Georgiev, P. G. (2017). C2h2 zinc finger proteins: The largest but poorly explored family of higher eukaryotic transcription factors, *Acta naturae* **9**(2): 47–58.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J. Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I. and Andrau, J.-C. (2012). Cpg islands and gc content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters, *Genome research* **22**(12): 2399–2408.

Filesi, I., Cacchione, S., De Santis, P., Rossetti, L. and Savino, M. (2000). The main role of the sequence-dependent dna elasticity in determining the free energy of nucleosome formation on telomeric dnas, *Biophys Chem* **83**(3): 223–37.

Fordyce, P. M., Gerber, D., Tran, D., Zheng, J. S., Li, H., DeRisi, J. L. and Quake, S. R. (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis, *Nature Biotechnology* **28**(9): 970–976.

Fraenkel, E. and Pabo, C. O. (1998). Comparison of x-ray and nmr structures for the antennapedia homeodomain-dna complex., *Nature structural biology* **5**: 692–7.

Gansen, A., Valeri, A., Hauger, F., Felekyan, S., Kalinin, S., Tth, K., Langowski, J. and Seidel, C. A. M. (2009). Nucleosome disassembly intermediates characterized by single-molecule fret, *Proc Natl Acad Sci USA* **106**(36): 15308.
URL: *http://www.pnas.org/content/106/36/15308.abstract*

Garrett, R. and Grisham, C. (2016). *Biochemistry*, Cengage Learning.
URL: *https://books.google.de/books?id=RWBzCwAAQBAJ*

Gradinaru, C. C., Marushchak, D. O., Samim, M. and Krull, U. J. (2010). Fluorescence anisotropy: from single molecules to live cells, *Analyst* **135**(3): 452–9.

Hafen, E., Kuroiwa, A. and Gehring, W. J. (1984). Spatial distribution of transcripts from the segmentation gene fushi tarazu during Drosophila embryonic development, *Cell* **37**(3): 833–841.

Harrison, S. C. and Aggarwal, A. K. (1990). Dna recognition by proteins with the helix-turn-helix motif., *Annual review of biochemistry* **59**: 933–69.

Heron, M. E. L. (2017). *Analysing and quantitatively modelling nucleosome binding preferences*, Dissertation, lmu munich.

Huber, P. J. (2011). *Robust Statistics*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1248–1251.

Hurst, H. C. (1995). Transcription factors 1: bzip proteins., *Protein profile* **2**: 101–68.

Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R., Ambrosini, G., Trono, D., Bucher, P. and Deplancke, B. (2017). Smile-seq identifies binding motifs of single and dimeric transcription factors, *Nat Methods* .

Jin, H., Finnegan, A. I. and Song, J. S. (2018). A unified computational framework for modeling genome-wide nucleosome landscape, *Phys Biol* **15**(6): 066011.

Jin, H., Rube, H. T. and Song, J. S. (2016). Categorical spectral analysis of periodicity in nucleosomal dna, *Nucleic Acids Res* **44**(5): 2047–57.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J., Vincentelli, R., Luscombe, N., Hughes, T., Lemaire, P., Ukkonen, E., Kivioja, T. and Taipale, J. (2013). Dna-binding specificities of human transcription factors, *Cell* **152**(12): 327–339.

Jung, C., Bandilla, P., von Reutern, M., Schnepf, M., Rieder, S., Unnerstall, U. and Gaul, U. (2018). True equilibrium measurement of transcription factor-dna binding affinities using automated polarization microscopy, *Nature Communications* **9**(1): 1605.

Jung, C., Schnepf, M., Bandilla, P., Unnerstall, U. and Gaul, U. (2019). High sensitivity measurement of transcription factor-dna binding affinities by competitive titration using fluorescence microscopy., *Journal of visualized experiments : JoVE* .

Kaplan, N., Moore, I., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., Hughes, T. R., Lieb, J. D., Widom, J. and Segal, E. (2010). Nucleosome sequence preferences influence in vivo nucleosome organization, *Nat Struct Mol Biol* **17**(8): 918–20.

Khorasanizadeh, S. (2004). The nucleosome: From genomic organization to genomic regulation, *Cell* **116**(2): 259–272.

Kim, R. (2011). Native agarose gel electrophoresis of multiprotein complexes., *Cold Spring Harbor protocols* **2011**: 884–7.

Klug, A. and Lutter, L. C. (1981). The helical periodicity of dna on the nucleosome, *Nucleic Acids Res* **9**(17): 4267–83.

Kosman, D. and Small, S. (1997). Concentration-dependent patterning by an ectopic expression domain of the drosophila gap gene knirps., *Development (Cambridge, England)* **124**: 1343–54.

Kretschy, N., Sack, M. and Somoza, M. M. (2016). Sequence-dependent fluorescence of cy3- and cy5-labeled double-stranded dna., *Bioconjugate chemistry* **27**: 840–8.

Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. and Mann, R. S. (2019). Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes., *Annual review of cell and developmental biology* **35**: 357–379.

Krietenstein, N., Wippo, C. J., Lieleg, C. and Korber, P. (2012a). Chapter nine - genome-wide in vitro reconstitution of yeast chromatin with in vivo-like nucleosome positioning, *in* C. Wu and C. D. Allis (eds), *Nucleosomes, Histones & Chromatin Part B*, Vol. 513 of *Methods in Enzymology*, Academic Press, pp. 205 – 232.

Krietenstein, N., Wippo, C. J., Lieleg, C. and Korber, P. (2012b). Genome-wide in vitro reconstitution of yeast chromatin with in vivo-like nucleosome positioning, *Methods Enzymol* **513**: 205–32.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *The Annals of Mathematical Statistics* **22**(1): 79–86.
URL: *http://www.jstor.org/stable/2236703*

Kumar, R. and Thompson, E. B. (1999). The structure of the nuclear hormone receptors., *Steroids* **64**: 310–9.

Lai, W. K. M. and Pugh, B. F. (2017). Understanding nucleosome dynamics and their links to gene expression and dna replication., *Nature reviews. Molecular cell biology* **18**: 548–562.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. and Weirauch, M. T. (2018). The human transcription factors., *Cell* **172**: 650–665.

Lee, J. Y., Lee, J., Yue, H. and Lee, T. H. (2015). Dynamics of nucleosome assembly and effects of dna methylation, *J Biol Chem* **290**(7): 4291–303.

Li, G.-W. and Elf, J. (2009). Single molecule approaches to transcription factor kinetics in living cells., *FEBS letters* **583**: 3979–83.

Li, J., Sagendorf, J. M., Chiu, T. P., Pasi, M., Perez, A. and Rohs, R. (2017). Expanding the repertoire of dna shape features for genome-scale studies of transcription factor binding, *Nucleic Acids Res* **45**(22): 12877–12887.

Lorch, Y. and Kornberg, R. D. (2017). Chromatin-remodeling for transcription., *Quarterly reviews of biophysics* **50**: e5.

Lowary, P. T. and Widom, J. (1998). New dna sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning, *J Mol Biol* **276**(1): 19–42.

Mavrich, T. N., Jiang, C., Ioshikhes, I. P., Li, X., Venters, B. J., Zanton, S. J., Tomsho, L. P., Qi, J., Glaser, R. L., Schuster, S. C., Gilmour, D. S., Albert, I. and Pugh, B. F. (2008). Nucleosome organization in the drosophila genome, *Nature* **453**(7193): 358–62.

McGinnis, N., Kuziora, M. A. and McGinnis, W. (1990). Human hox-4.2 and drosophila deformed encode similar regulatory specificities in drosophila embryos and larvae., *Cell* **63**: 969–76.

Meng, X., Brodsky, M. H. and Wolfe, S. A. (2005). A bacterial one-hybrid system for determining the dna-binding specificity of transcription factors., *Nature biotechnology* **23**: 988–94.

Najafabadi, H. S., Garton, M., Weirauch, M. T., Mnaimneh, S., Yang, A., Kim, P. M. and Hughes, T. R. (2017). Non-base-contacting residues enable kaleidoscopic evolution of metazoan c2h2 zinc finger dna binding, *Genome Biology* **18**(1): 167.
URL: *https://doi.org/10.1186/s13059-017-1287-y*

Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E. E. and Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution, *Elife* **4**.

Nusslein-Volhard, C., Kluding, H. and Jurgens, G. (1985). Genes affecting the segmental subdivision of the drosophila embryo., *Cold Spring Harbor symposia on quantitative biology* **50**: 145–54.

Oohara, I. and Wada, A. (1987). Spectroscopic studies on histone-dna interactions, *Journal of Molecular Biology* (2): 399–411.
URL: *https://dx.doi.org/10.1016/0022-2836(87)90700-5*

Pal, S., Hoinka, J. and Przytycka, T. M. (2019). Co-select reveals sequence non-specific contribution of dna shape to transcription factor binding in vitro., *Nucleic acids research* **47**: 6632–6641.

Park, P. J. (2009). Chip-seq: advantages and challenges of a maturing technology., *Nature reviews. Genetics* **10**: 669–80.

Pavletich, N. P. and Pabo, C. O. (1993). Crystal structure of a five-finger gli-dna complex: new perspectives on zinc fingers., *Science (New York, N.Y.)* **261**: 1701–7.

Perry, M. W., Boettiger, A. N. and Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the drosophila embryo., *Proceedings of the National Academy of Sciences of the United States of America* **108**: 13570–5.

Phillips, K. and Luisi, B. (2000). The virtuoso of versatility: Pou proteins that flex to fit11edited by p. wright, *Journal of Molecular Biology* **302**(5): 1023 – 1039.
URL: *http://www.sciencedirect.com/science/article/pii/S002228360094107X*

Puhl, H. L. and Behe, M. J. (1993). Structure of nucleosomal dna at high salt concentration as probed by hydroxyl radical., *Journal of molecular biology* **229**: 827–32.

Rastogi, C., Rube, H. T., Kribelbauer, J. F., Crocker, J., Loker, R. E., Martini, G. D., Laptenko, O., Freed-Pastor, W. A., Prives, C., Stern, D. L., Mann, R. S. and Bussemaker, H. J. (2018). Accurate and sensitive quantification of protein-dna binding affinity., *Proceedings of the National Academy of Sciences of the United States of America* **115**: E3692–E3701.

Raveh-Sadka, T., Levo, M., Shabi, U., Shany, B., Keren, L., Lotan-Pompan, M., Zeevi, D., Sharon, E., Weinberger, A. and Segal, E. (2012). Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast., *Nature genetics* **44**: 743–50.

Razin, S. V., Borunova, V. V., Maksimenko, O. G. and Kantidze, O. L. (2012). Cys2his2 zinc finger protein family: classification, functions, and major members., *Biochemistry. Biokhimiia* **77**: 217–26.

Rezsohazy, R., Saurin, A. J., Maurel-Zaffran, C. and Graba, Y. (2015). Cellular and molecular insights into hox protein action., *Development (Cambridge, England)* **142**: 1212–27.

Riley, T. R., Lazarovici, A., Mann, R. S. and Bussemaker, H. J. (2015). Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE, *Elife* **4**.

Riley, T. R., Slattery, M., Abe, N., Rastogi, C., Liu, D., Mann, R. S. and Bussemaker, H. J. (2014). Selex-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes., *Methods in molecular biology (Clifton, N.J.)* **1196**: 255–78.

Rimini, R., Pontiggia, A., Spada, F., Ferrari, S., Harley, V. R., Goodfellow, P. N. and Bianchi, M. E. (1995). Interaction of normal and mutant sry proteins with dna., *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **350**: 215–20.

Roehrl, M. H., Wang, J. Y. and Wagner, G. (2004). A general framework for development and data analysis of competitive high-throughput screens for small-molecule inhibitors of protein-protein interactions by fluorescence polarization, *Biochemistry* **43**(51): 16056–66.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B. and Mann, R. S. (2010). Origins of specificity in protein-dna recognition., *Annual review of biochemistry* **79**: 233–69.

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S. and Honig, B. (2009). The role of dna shape in protein-dna recognition., *Nature* **461**: 1248–53.

Ruan, S. and Stormo, G. D. (2017). Inherent limitations of probabilistic models for protein-dna binding specificity, *PLoS Comput Biol* **13**(7): e1005638.

Rube, H. T., Rastogi, C., Kribelbauer, J. F. and Bussemaker, H. J. (2018). A unified approach for quantifying and interpreting dna shape readout by transcription factors, *Molecular Systems Biology* **14**(2): e7902.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences, *Nucleic Acids Res* **18**(20): 6097–100.

Schnepf, M., Ludwig, C., Bandilla, P., Ceolin, S., Unnerstall, U., Jung, C. and Gaul, U. (2020). Sensitive automated measurement of histone-dna affinities in nucleosomes, *iScience* p. 100824. **URL:** *http://www.sciencedirect.com/science/article/pii/S2589004220300079*

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J. P. and Widom, J. (2006). A genomic code for nucleosome positioning, *Nature* **442**(7104): 772–8.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in drosophila segmentation, *Nature* **451**(7178): 535–40.

Segal, E. and Widom, J. (2009a). Poly(da:dt) tracts: major determinants of nucleosome organization, *Curr Opin Struct Biol* **19**(1): 65–71.

Segal, E. and Widom, J. (2009b). What controls nucleosome positions?, *Trends Genet* **25**(8): 335–43.

Shrader, T. E. and Crothers, D. M. (1989). Artificial nucleosome positioning sequences, *Proceedings of the National Academy of Sciences of the United States of America* **86**(2798415): 7418–7422. **URL:** *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC298075/*

Shrader, T. E. and Crothers, D. M. (1990). Effects of dna sequence and histone-histone interactions on nucleosome placement, *J Mol Biol* **216**(1): 69–84.

Siebert, M. (2016). *Quantitative modeling and statistical analysis of protein-DNA binding sites*, PhD thesis, LMU Munich.

Siebert, M. and Soding, J. (2016). Bayesian markov models consistently outperform pwms at predicting motifs in nucleotide sequences, *Nucleic Acids Res* **44**(13): 6055–69.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordan, R. and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome., *Trends in biochemical sciences* **39**: 381–99.

Stanojevic, D., Hoey, T. and Levine, M. (1989). Sequence-specific dna-binding activities of the gap proteins encoded by hunchback and kruppel in drosophila., *Nature* **341**: 331–5.

Steiner, T. (2002). The hydrogen bond in the solid state., *Angewandte Chemie (International ed. in English)* **41**: 49–76.

Stella, S., Cascio, D. and Johnson, R. C. (2010). The shape of the dna minor groove directs binding by the dna-bending protein fis., *Genes & development* **24**: 814–26.

Stormo, G. D., Schneider, T. D., Gold, L. and Ehrenfeucht, A. (1982). Use of the perceptron algorithm to distinguish translational initiation sites in e. coli, *Nucleic Acids Research* **10**(9): 2997–3011.

Štros, M., Launholt, D. and Grasser, K. D. (2007). The hmg-box: a versatile protein domain occurring in a wide variety of dna-binding proteins, *Cellular and Molecular Life Sciences* **64**(19): 2590.
URL: *https://doi.org/10.1007/s00018-007-7162-3*

Suzuki, M., Amano, N., Kakinuma, J. and Tateno, M. (1997). Use of a 3d structure data base for understanding sequence-dependent conformational aspects of dna., *Journal of molecular biology* **274**: 421–35.

Takasuka, T. E. and Stein, A. (2010). Direct measurements of the nucleosome-forming preferences of periodic dna motifs challenge established models, *Nucleic acids research* **38**(17): 5672–5680.

Teichmann, M., Dumay-Odelot, H. and Fribourg, S. (2012). Structural and functional aspects of winged-helix domains at the core of transcription initiation complexes., *Transcription* **3**: 2–7.

Tessarz, P. and Kouzarides, T. (2014). Histone core modifications regulating nucleosome structure and dynamics., *Nature reviews. Molecular cell biology* **15**: 703–8.

Thastrom, A., Gottesfeld, J. M., Luger, K. and Widom, J. (2004). Histone-dna binding free energy cannot be measured in dilution-driven dissociation experiments, *Biochemistry* **43**(3): 736–41.

Thastrom, A., Lowary, P. T., Widlund, H. R., Cao, H., Kubista, M. and Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning dna sequences, *J Mol Biol* **288**(2): 213–29.

Tillo, D. and Hughes, T. R. (2009). G+c content dominates intrinsic nucleosome occupancy, *BMC Bioinformatics* **10**: 442.

Tillo, D., Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Field, Y., Lieb, J. D., Widom, J., Segal, E. and Hughes, T. R. (2010). High nucleosome occupancy is encoded at human regulatory sequences, *PloS one* **5**(2): e9129–e9129.

Vasudevan, D., Chua, E. Y. and Davey, C. A. (2010). Crystal structures of nucleosome core particles containing the '601' strong positioning sequence, *J Mol Biol* **403**(1): 1–10.

Von Reutern, M. (2017). Pysite, Github.
URL: *https://github.com/Reutern/PySite*

Wang, D., Ulyanov, N. B. and Zhurkin, V. B. (2010). Sequence-dependent Kink-and-Slide deformations of nucleosomal DNA facilitated by histone arginines bound in the minor groove, *J. Biomol. Struct. Dyn.* **27**(6): 843–859.

Wang, L., Stein, L. and Ware, D. (2014). The relationships among GC content, nucleosome occupancy, and exon size, *ArXive* .

Watanabe, S., Resch, M., Lilyestrom, W., Clark, N., Hansen, J. C., Peterson, C. and Luger, K. (2010). Structural characterization of h3k56q nucleosomes and nucleosomal arrays., *Biochimica et biophysica acta* **1799**: 480–6.

Weber, G. (1952). Polarization of the fluorescence of macromolecules. i. theory and experimental method., *The Biochemical journal* **51**: 145–55.

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., Bussemaker, H. J., Morris, Q. D., Bulyk, M. L., Stolovitzky, G., Hughes, T. R. and Consortium, D. (2013). Evaluation of methods for modeling transcription factor sequence specificity, *Nature Biotechnology* **31**(2): 126–134.

Werner, M. (2008). *Information und Codierung - Grundlagen und Anwendungen*, 2. vollst. berarb. u. erw. aufl. 2009 edn, Springer-Verlag, Berlin Heidelberg New York.

Yamanaka, N., Rewitz, K. F. and O'Connor, M. B. (2013). Ecdysone control of developmental transitions: lessons from drosophila research, *Annual review of entomology* **58**(23072462): 497–516.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060523/*

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R. and Rohs, R. (2017). Transcription factor family-specific dna shape readout revealed by quantitative specificity models, *Mol Syst Biol* **13**(2): 910.

Zhang, Y., Moqtaderi, Z., Rattner, B. P., Euskirchen, G., Snyder, M., Kadonaga, J. T., Liu, X. S. and Struhl, K. (2009). Intrinsic histone-dna interactions are not the major determinant of nucleosome positions in vivo, *Nat Struct Mol Biol* **16**(8): 847–52.

Zhao, Y., Ruan, S., Pandey, M. and Stormo, G. D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions, *Genetics* **191**(3): 781–90.

Zhao, Y. and Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity, *Nat Biotechnol* **29**(6): 480–3.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., Bussemaker, H. J., Gordan, R. and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using dna shape, *Proc Natl Acad Sci U S A* **112**(15): 4654–9.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A. C., Ghane, T., Di Felice, R. and Rohs, R. (2013). Dnashape: a method for the high-throughput prediction of dna structural features on a genomic scale, *Nucleic Acids Res* **41**(Web Server issue): W56–62.

Acknowledgement

Es gibt viele Menschen, denen ich Dank schulde und die ihren Teil dazu beigetragen haben, dass ich diese Arbeit erfolgreich abschließen konnte. Als erstes gebührt mein Dank meiner "Doktormutter" Ulrike Gaul, die großes Vertrauen in mich gesetzt hat, als sie mir mein Projekt gegeben hat und die mich immer soweit unterstützt hat, wie es ihre Gesundheit zugelassen hat. Auch meinem "Doktorvater", Roland Beckmann, möchte ich danken. Er hat mir in der Stunde der Not seine Unterstützung zugesagt und mir immer die Freiheit gegeben, die nötig war, um mit Beharrlichkeit und vielleicht auch etwas Trotz gegen die Wahrscheinlichkeit dieses Projekt zu einem guten Abschluss zu führen. Ich möchte allen Mitgliedern der Gaul-Gruppe für die tolle Atmosphäre und jederzeitige Hilfsbereitschaft danken, im Besonderen aber Peter Bandilla für seine fast schon magische Fähighkeiten beim Lösen technischer Probleme, Ulrich Unnerstall für intensive Diskussionen und das Aufzeigen von "natural questions", Marc von Reutern für seine Dienste als Bioinformatiker und seine Hilfe, damit ich mich selbst zu einem entwickeln konnte, Sabine Bergelt für ihre Rolle as Fremdenführer im Bürokratiedschungel, Stefano Ceolin für excellente konstruktive Kritik und Freundschaft in den letzen Jahren der Gruppe und nicht zuletzt Claudia Ludwig für ihre unschätzbare Hilfe allgemein und speziell beim Nukleosomen Projekt. Der Stingele Gruppe möchte ich für ihre Unterstützung und Gesellschaft in einem immer leerer werdenden Labor danken und dass sie mich in den letzten zwei Jahren adoptiert haben. Meiner Graduiertenschule QBM möchte ich für die finanzielle Unterstützung, aber auch für die tolle Umgebung aus exzellenten Wissenschaftlern und daraus resultierende Freundschaften. Ich möchte Markus Hohle, Filiz Civril, Mara Kiecke, Julia Schlehe und Dietmar Martin dafür danken, dass sie es geschafft haben, eine wunderbare interdisziplinäre Mischung aus Experimentatoren und Theoretikern verschiedenster Fachrichtungen zusammenzubringen und uns beigebracht haben, einander besser zu verstehen.

Ich möchte mich für die Hilfe bei der Histonaufreinigung aus dem Becker Labor (vor allem Peter Becker und Sandro Baldi) und dem Korber Labor (Philipp Korber, Elisa Oberbeckmann und Iris Langstein) bedanken und ganz besonders für die Probe an Histonen, nachdem die Reinigung in unserem Labor nicht mehr funktionieren wollte.

Es gibt noch zwei Menschen, für die ich den Dank bis zum Schluss aufgespart habe, weil sie mir die wichtigsten waren. Zum einen Mal Christophe Jung, der mir als Mentor, Betreuer, Lunch Partner und Freund über die gesammte Zeit der Promotion zur Seite stand, der sich über gute Nachrichten gefreut und bei schlechten stehts bereit war, mit mir eine Lösung zu finden. Der letze Dank geht an meine Frau, die mich über die gesamte Zeit der Doktorarbeit unterstützt hat und die mir die Kraft gegeben hat, weiterzumachen, wenn ich selbst nicht mehr daran geglaubt habe. An alle, die ich nun vergessen habe: Auch euch einen herzlichen Dank und ein Entschuldigung!