Modeling nucleosome mediated mechanisms of gene regulation

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Evgeniy Andreevich Ozonov
aus Russland

Basel, 2013

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Erik van Nimwegen und Prof. Attila Becskei

Basel, den 13. November 2012

Prof. Dr. Jörg Schibler
Dekan

To my family and my friends.

Contents

Acknowledgements

I would like to express my very great appreciation to Prof. Erik van Nimwegen for his patient guidance, enthusiastic encouragement and useful advices of this research work.

I would also like to thank Piotr Balwierz, Peter Pemberton-Ross, Philipp Berninger, Mikhail Pachkov and all members of the van Nimwegen lab for their support and fruitful collaborations.

I also thank my friends, especially Erzhena Batorova, Anna Shagzhieva and Sergey Manzhuev, for being with me all along this long journey.

I especially thank my parents, Andrey and Lyudmila, and my brother Tumen. Without their warm support and encouragement I would not have made it this far.

Chapter 1

Introduction

In the post-genomic era the question of how the expression of genetic information is carried out is the central question of molecular biology. The fascinating fact that a complex multicellular organism originates from a single cell and the processes of cell differentiation are very reproducible and robust to environmental changes poses a very fundamental question of how the execution of the genetically encoded "program" is controlled. Although, nearly all cells have essentially the same genetic information there is a number of cell types with different function and morphological properties. This implies that different parts of the genome must be properly read and interpreted at very specific points in time and space during development.

Transcription is the first step of the genome readout. The concentration of mRNA is the key characteristic, despite all others, that defines cell identity. Although, there are post-transcriptional mechanisms that control mRNA levels in the cell, such as RNA decay and microRNA mediated RNA interference, it has been shown that transcription is the major process that determines mRNA abundance [108]. Scientists have made a great effort to investigate the process of transcription and remarkable achievements have taken place in the last few decades, nevertheless we are still far from full understanding of what determines transcription rate, and we are even further away from creating a computational model which could reliably predict mRNA levels in the cell.

In eukaryots the processes which preclude transcription elongation, such as binding of transcription factors and assembly of the preinitiation complex (PIC), occur in the context of chromatin. It has been shown that the role of chromatin extends far beyond only DNA compaction. In this chapter we briefly introduce the role of chromatin in gene regulation, methods which are used to study chromatin related effects and factors which determine chromatin configuration in promoters of genes.

1.1 Nucleosome - the basic unit of chromatin

The eukaryotic DNA is a long linear polymer. For instance, the human genome, containing about three billion base pairs which corresponds to length of approximately 2 meters, has to be folded in a nucleus of size of few micrometers. Moreover, the DNA is negatively charged polymer and electrostatic repulsion from neighboring phosphates does not allow it to fit within the small nucleus [45]. Solution to the packaging problem has appeared in the form of histone proteins that bind to DNA and neutralize the negative charges leading to compaction of the DNA. Five types of histones, i.e. H1, H2A,H2B, H3 and H4, have nearly perfect conservation across eukaryotic species. The lowest and the most fundamental level of DNA compaction, which is called nucleosome, was discovered in 1974 by Roger Kornberg [51]. The nucleosome is a complex of histone octamer, two copies of each type of histones H2A, H2B, H3 and H4, and a stretch of DNA wrapped around the histone octamer (Fig. 1.1 B and C). Although, there are higher levels of chromatin compaction, such as $30 - nm$ chromatin fibers, that allow up to 10000-fold compaction of the DNA, we focus on the most basic "beads-on-a-string" structure (Fig. 1.1 A) which is the most studied level of DNA compaction nowadays.

The biochemical analysis revealed roughly equal weights of histones and DNA in the cell which corresponds to about 200 bp of DNA per each histone octamer [51]. It implies that about 80% of the eukaryotic genome is packaged into nucleosomes. The crystal structure of the nucleosome core particle (Fig. 1.1 B) shows that the nucleosome consists of a DNA stretch with length 147 bp which is wrapped in approximately 1.65 super-helical turns around a histone octamer (Fig. 1.1 C) [29, 65]. The basic structure of chromatin comprises repeating nucleosomes separated by linkers of length 20-40bp.

About 25 years ago molecular biologists were skeptical about the role of chromatin in gene regulation [85]. It was thought that the only role of nucleosomes is the DNA compaction. However, later *in vitro* studies [49, 62, 121] showed that nucleosomes are barriers for both transcription initiation and elongation. The low copy number of the histone genes in *Saccharomyces cerevisiae* allowed researchers to carry out genetic studies with altered histone levels. The *in vivo* study of the PHO5 promoter [34] shows that under knock-down of the H4 histone the PHO5 promoter is activated even under normally repressing conditions. In general, these studies show repressive function of nucleosomes in transcription.

In addition, the histones are subjects to a number of posttranslational modifications, such as metylation, accetylation and ubiquitination. These histone marks, crucially affect transcription (reviewed in [56]) . The importance of the histone marks is further

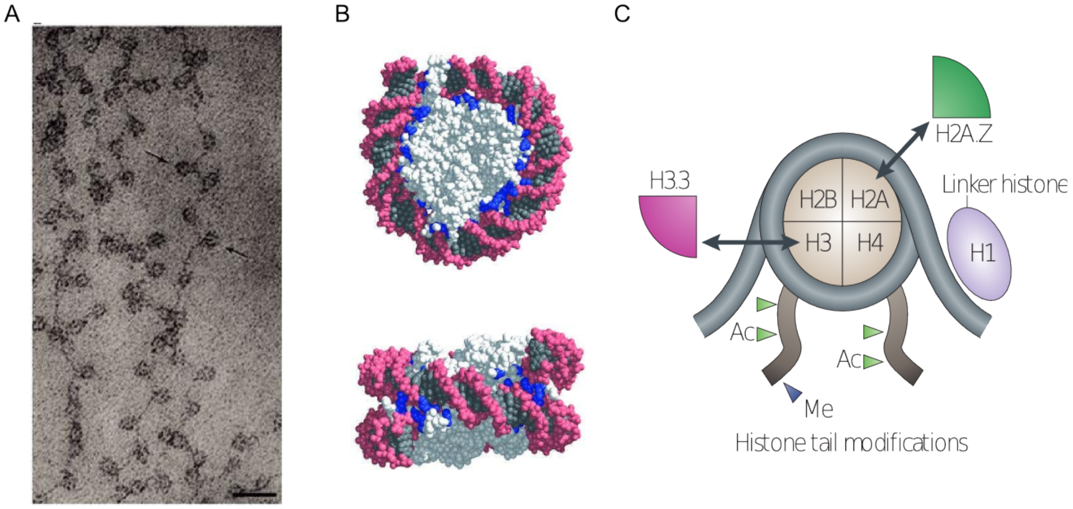

Figure 1.1: The nucleosome. **A:** Electron micrography of the "beads-on-a-string" structure of chromatin. Size marker: $30nm$. **B:** Crystal structure of a nucleosome core particle (front and side view). **C:** A scheme of the nucleosome core particle. The histone octamer comprising 4 types of histones (H3, H4, H2A and H2B) and a stretch of DNA wrapped around. Also the linker histone H1, examples of histone tail modifications and histone variants (H3.3 and H2A.Z) are shown. The **A** was adapted from reference [78] and **B**, **C** were adapted from reference [45] with permission of the Nature Publishing Group

supported by observations that disruptions in the epigenetic landscape are associated with diseases (reviewed in [81]). The variety of the histone marks led researches to a hypothesis of "histone code" as an extension of the genetic information, where different combination of histone modifications are read by other protein complexes and determine chromatin state of genes, for example silent or active [44].

Apart from canonical forms of the histones there are histone variants, such as H2A.Z and H3.3. The histone variants replace the canonical histones in the nucleosome core (Fig. 1.1 C) and may affect DNA-related metabolic processes (reviewed in [91]). Interestingly, the histone variant H2A.Z is found at 5' ends of nearly 2/3 of genes in *S.cereviciae* ([82], also see review [67]).

In general, it is now accepted that local chromatin configuration and epigenetic landscape affect almost all DNA-related metabolic processes, such as transcription, replication, DNA-repair and so forth. Therefore, elucidating the mechanisms which determine chromatin state is of great importance.

1.2 Genome-wide nucleosome mapping

10 years ago nucleosome configuration was known only for a few genomic loci, for instance for GAL1-10 [60], GAL80 [61] and PHO5 [4, 5] promoters. However, the technological breakthrough in the last decade allowed mapping of nucleosomes across the whole genome with unprecedented depth and accuracy. The first large scale experiments that measured nucleosome occupancy using microarray technology in promoters of genes revealed, despite their rather low resolution, that promoters of active genes are generally nucleosome depleted [15, 54]. The nucleosome mapping experiment with higher resolution (20 bp) confirmed this observation and showed that promoters of genes have distinct nucleosome pattern [125]. Later, the data from the high-resolution nucleosome experiment using high-throughput sequencing technology (ChIP-Seq experiment) showed that distinct nucleosome patterns occur not only in promoters but also at the 3' ends of genes [69].

All experimental methods for identifying nucleosome positions rely on the fact that nucleosomes protect DNA from exonuclease digestion, though, recently a new technique has appeared that uses chemically modified histones [17]. Nucleosome positions have been mapped both *in vivo* [28, 48, 55, 69, 100] and *in vitro* [48, 127, 128]. The *in vivo* studies aim to identify nucleosome positions in cells grown under certain condition, usually in rich media but there are datasets for different conditions, such as heat-shock [100] or cells grown in ethanol [48]. Usually, the histones are cross-linked to DNA using formaldehyde to fixate nucleosomes in their *in vivo* locations (Fig. 1.2 A).

The *in vitro* studies aim to measure nucleosome distribution which is governed solely by intrinsic sequence preferences of histones. The purified histones and DNA are assembled into nucleosomes (Fig. 1.2 A) using salt gradient dialysis (SGD) (methods for chromatin reconstitution are reviewed in [66]).

Once chromatin has been isolated *in vivo* or reconstituted *in vitro* it is sheared using micrococcal nuclease (MNase) (sometimes sonication is used) (Fig. 1.2 B). The MNase preferentially digests linker DNA, while nucleosomal DNA is protected from MNase digestion. Then nucleosome particles are isolated by immunoprecipitation using antibodies against histones or a certain histone modification and subjected to deproteinization to release nucleosomal DNA (Fig. 1.2 C). After the DNA was purified, the DNA fragments of length about 150 bp were selected using gel electrophoresis, and the positions in a reference genome from which these fragments originated were identified by microarray or high-throughput sequencing (Fig. 1.2 D). The development of the next-generation sequencing technologies allowed to identify nucleosome positions with up to 1 bp resolution in yeast [28, 48, 55, 69, 100, 120, 127] and other eukaryots [46, 70, 93, 113]

Although the described methods were able to generate nucleosome maps that are reproducible across different datasets, they have several experimental artifacts. Firstly, MNase have sequence preferences for cutting DNA at AT rich regions. Recently, control experiments carried out with naked DNA digested by MNase revealed quite strong correlation with nucleosome mapping datasets [22, 59]. This is particularly important for models of intrinsic sequence specificity of histones as the sequence biases which are introduced by MNase can lead to incorrect model. However, further analysis suggested that MNase bias doesn't significantly affect nucleosome maps [3]. Moreover, a nucleosome map generated by a new MNase free experimental technique shows very similar sequence features of the nucleosomal DNA [17]. This method uses engineered histone H4 with a unique cysteine introduced at position close to a nucleosome center. The introduced cysteine can attach a special label, and after addition of copper and hydrogen peroxide a short-lived radical created in a chemical reaction cleaves the DNA backbone at position of the introduced cysteine. After high-throughput sequencing and mapping to a reference genome a map of nucleosome centers with 1bp resolution can be created. Since this method was introduced not so long ago possible biases and limitations of it are not clear yet. Nevertheless, it is, perhaps, the most accurate method for nucleosome mapping nowadays.

Apart from the possible biases introduced by MNase digestion there are experimental artifacts related to microarray or high-throughput sequencing technologies. For example, it is well known that the nucleotide composition and propensity to form secondary

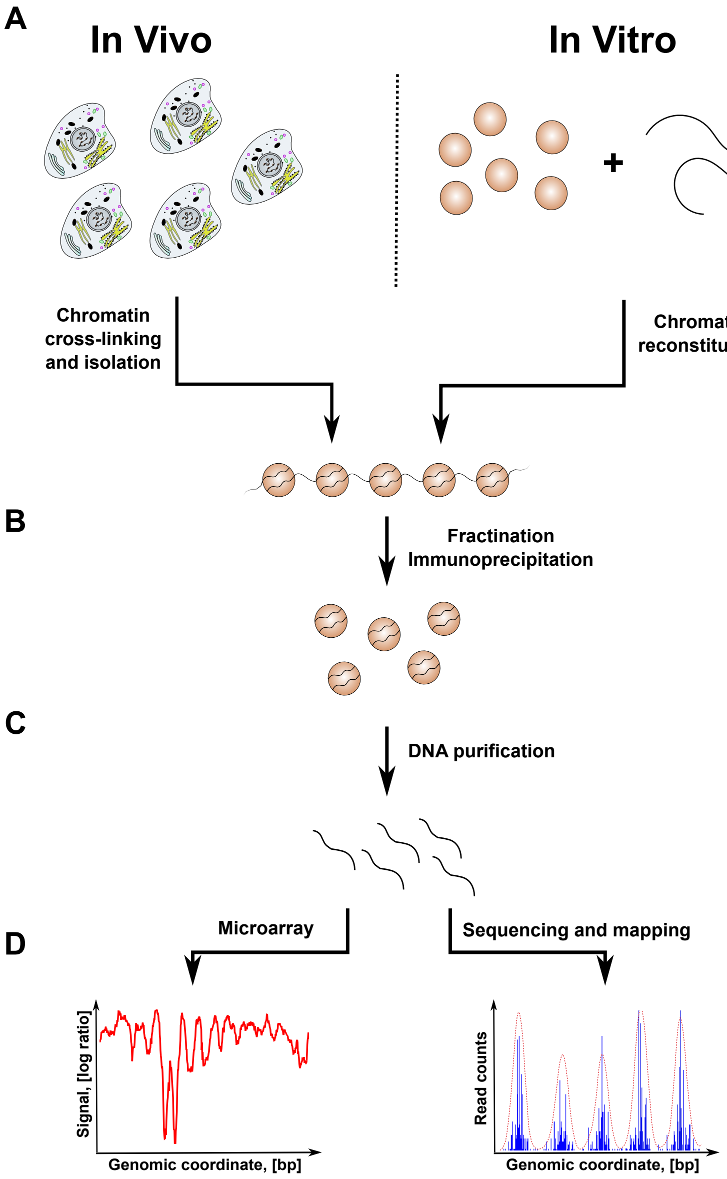

Figure 1.2: A general scheme of a nucleosome mapping experiment. (**A**) For *in vivo* experiment chromatin is cross-linked and isolated. For *in vitro* experiment chromatin is reconstituted using salt gradient dialysis or ATP-dependent chromatin remodelers. (**B**) After fractionation with MNase, which preferentially digests linker DNA, chromatin is immunoprecipitated using antibodies against a certain histone or epigenetic modification. (**C**) After deproteinization the DNA is purified and size-selected to get mononucleosomal DNA which is analyzed (**D**) using microarray or high-throughput sequencing technologies.

structures of the reads can systematically bias the read counts in ChIP-seq by more than 10-fold ([36, 104]). For example, comparison of nucleosome datasets generated in different studies shows that positions of nucleosomes are very reproducible across datasets [45]. However, actual signal values are poorly correlated. In other words, whereas the positions of peaks and troughs of the signal are consistent between datasets, the amplitude of the signal is not very well correlated. This point will be discussed in the section 2.1.

In summary, the methods for mapping nucleosomes and histone modification described above are extremely important and widely used in chromatin biology nowadays.

1.3 Nucleosome positioning

The experimental methods for nucleosome mapping described above have made very significant contribution to elucidating chromatin structure and its role in gene regulation. The experiments revealed that substantial amount of nucleosomes are not randomly distributed across the genome, but have distinct patterns, especially at genomic loci related to DNA-related metabolic processes, such as transcription, replication and so on.

The first nucleosome mapping experiment with high-resolution [125] allowed to discover remarkable nucleosome pattern at promoters of genes. Later experiments [55, 69] confirmed previous observations of nucleosome pattern at 5' end and revealed distinct nucleosome architecture at 3' ends of genes (Fig. 1.3).

The chromatin architecture at 5' end of genes is comprised of a region free of nucleosomes just upstream of the transcription start site (TSS), usually called nucleosome free region (NFR) or nucleosome depleted region (NDR), and a few well positioned nucleosomes (phased nucleosomes) up- and downstream of the NFR (+1, -1, +2, -2 nucleosomes) (Fig. 1.3 B and C). Importantly, the degree of positioning decreases further downstream of the TSS.

The nucleosome pattern at 3' end also comprised of an NFR just downstream of the transcription termination site (TTS) and nucleosomes surrounding the NFR, even though the nucleosomes at 3' end are much less positioned than at 5' end.

Interestingly, the +1 and -1 nucleosomes often contain the histone variant H2A.Z [82] and different histone modifications (reviewed in [56]).

Although, the nucleosome profile averaged across all genes helped to discover common chromatin features at 5' and 3' ends, the difference in nucleosome occupancy (Fig. 1.3 A) helped researchers to make a link between chromatin structure and transcriptional activity. Gene-by-gene analysis revealed two general classes of genes according to chro-

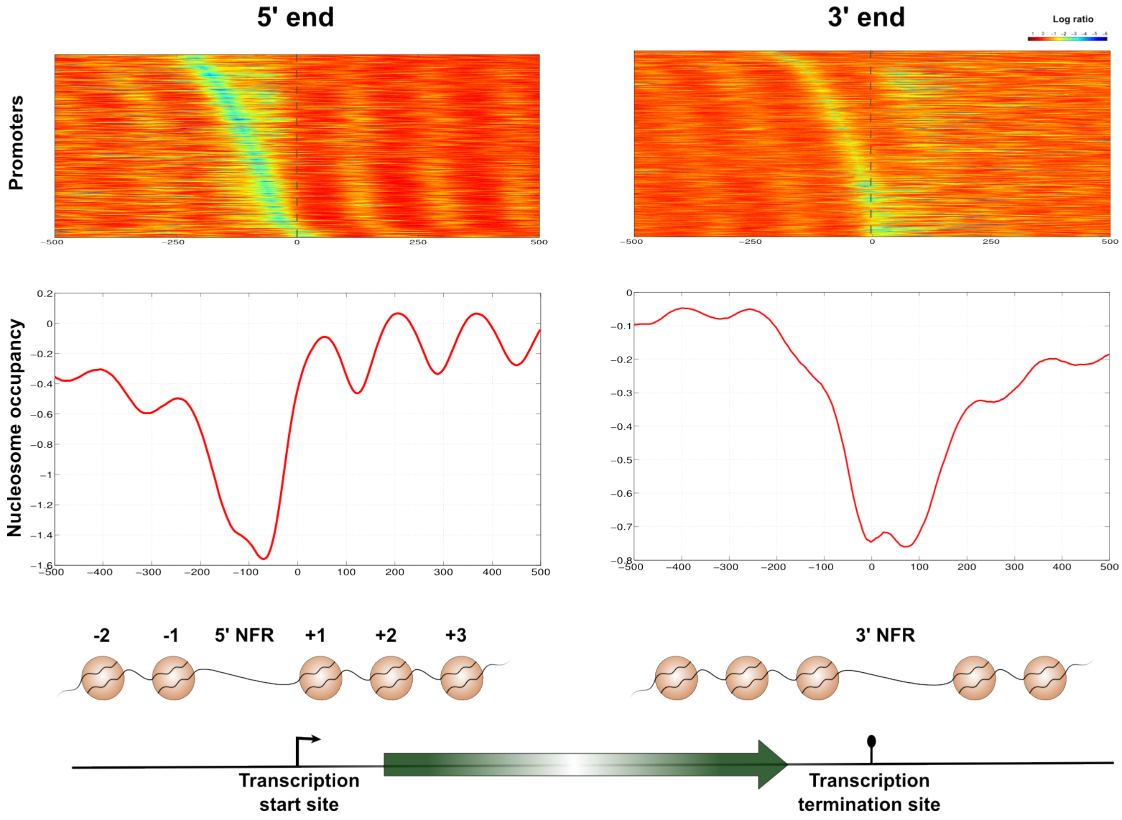

Figure 1.3: Nucleosome patterns at 5' and 3' ends of genes. **A:** Color coded nucleosome occupancy measured in [55] around every 5' and 3' end of genes ($+/-500bp$ around TSS or TTS acc.to [74]). **B:** Nucleosome occupancy averaged across all 5' and 3' ends of genes. **C:** Schematic representation of nucleosome patterns at 5' and 3' ends of genes. The nucleosome pattern at 5' ends comprised of a nucleosome free region (NFR) and few well positioned nucleosomes up- and downstream of the NFR. The nucleosome pattern 3' ends of genes contains an NFR as well. However, the NFR at 3' end is not as pronounced as at 5' end and nucleosomes around it are much less positioned.

matin structure and transcriptional activity: "growth" and "stress" genes ([55, 109] and reviewed in [85]).

The "growth" or housekeeping genes generally have very pronounced NFR and well-positioned +1/-1 nucleosomes in promoters [85]. Their expression is highest during rapid growth and often low during stress response. These genes usually regulated by TFIID rather than SAGA, lack TATA boxes, exhibit little noise in expression level and are not affected by deletion of most chromatin regulatory genes [12, 76].

The "stress" genes are almost silent in rich media and transcribed under some stress conditions. These genes are characterized by regulation by the SAGA complex, rather than TFIID, have TATA boxes, have high "transcriptional plasticity" and noisy or "bursty" expression [109]. Also, these genes are regulated by variety of chromatin-remodeling factors and exhibit more variable promoter architecture [85].

The importance of nucleosome architecture at promoters begs the question of what determines nucleosome positioning. The early studies suggested that underlying DNA can influence nucleosome formation [25, 41, 63]. Indeed, the structure of the nucleosome shows that DNA is bended around the histone octamer [29, 65] but bendability of underlying DNA depends of the nucleotide composition. The sequence determinants (or *cis*-determinants) of nucleosome positioning attracted great attention recently. Nucleosome maps generated *in vitro*, where nucleosomes are positioned solely by underlying DNA sequence, allowed the discovery of sequence features which favor nucleosome formation. Analysis of nucleosomal DNA showed periodic pattern of AA/TT/TA/AT dinucleotides spaced every ≈ 10 bp, which corresponds to one turn of the DNA helix, and similar periodic pattern of GC/CG/GG/CC dinucleotides but in anti-phase (shifted by 5-bp) with AA/TT/TA/AT pattern [2, 40, 92, 97]. The AA/TT/TA/AT and GC/CG/GG/CC dinucleotide periodical pattern determine so called *rotational* setting of nucleosomes, i.e. local orientation of DNA helix on the histone surface [45]. The observed positions of AA/TT/TA/AT and GC/CG/GG/CC dinucleotides determine energetically favorable configuration of DNA bending when AT rich dinucleotides face the histone surface and GC rich dinucleotides point away from the histone surface. Moreover analysis of linker DNA (nucleosome free DNA) showed that sequences which contain stretches of A or T (poly(dA:dT) elements) are less favorable for nucleosome formation [41, 48]. In general, it was shown that AT/GC content is highly correlated with nucleosome occupancy *in vitro* [107].

Despite clear evidence for the role of underlying DNA sequence in nucleosome positioning, the DNA sequence can't explain *translational* positioning of nucleosomes *in vivo*, i.e. nucleosome positioning relative to a chromosomal locus [127]. First of all, nucleosome maps generated *in vitro* do not reproduce the *in vivo* nucleosome pattern at promoters of genes. Even though, *in vitro* maps show nucleosome depletion at promoters (5' NFR) it is not as pronounced as *in vivo*, and nucleosome surrounding 5' NFR are not well-positioned [48, 128]. Moreover, the sequence determinants can't explain differences in nucleosome patterns in cells grown under different conditions [100].

It has been shown that a number of other protein complexes can affect nucleosome distribution. The *trans*-factors such as ATP-dependent chromatin remodelers and sequence specific DNA binding proteins can substantially affect nucleosome positioning. For instance, it was shown that upon loss of transcription factors ABF1, REB1 and RSC3 substantial amount of promoters become nucleosome occupied [8]. The other study of the CLN2 regulatory region showed that mutation of the binding sites for auxiliary proteins REB1, MCM1 and RSC30 leads to NFR loss and sporadic activation of CLN2 gene by SBF [9].

Interestingly, the study [128] showed that reconstitution of chromatin with ATP-dependent chromatin remodelers and yeast whole-cell extract allowed to reproduce *in vivo* nucleosome pattern at 5' end, i.e. 5' NFR and well-positioned nucleosomes surrounding the NFR. This study suggests that ATP-dependent chromatin remodelers and sequence specific DNA binding proteins work together to establish chromatin architecture at promoters of genes.

In 1988 Kornberg and Stryer suggested theoretical explanation of repeating nucleosome patterns by *statistical positioning* effect [52]. They showed that nucleosomes without sequence specificity become well-positioned against a barrier which prevents nucleosome formation (see section 2.4.2). Importantly, the degree of positioning decreases with distance from a barrier, which resembles the nucleosome pattern at 5' end of genes (Fig. 1.3 B). Originally, Kornberg and Stryer suggested transcription factors to play the role of barriers against which nucleosomes are positioned. Later, it was suggested that poly(dA:dT) elements or +1 well-positioned nucleosome may play the role of such barriers at promoters of genes [69, 72, 125].

The functional evolutionary approach introduced in the study [39] provided remarkable insights into mechanisms that control nucleosome positioning. The approach relies on the observation that there are species-specific differences in parameters of nucleosome positioning in a variety of yeast species [112]. The main idea of this approach is to compare native chromatin of a specie to chromatin reconstituted in the foreign context of another closely related specie. In other words, they took large genomic regions from *K.lactis*, *K. waltii* and *D. hansenii* and reassembled artificial chromosomes (YAC) in the context of *S.cereviciae*. In principle, features that change in the foreign context are determined by protein factors that are different in two species. On the other hand, features which are similar are due to either intrinsic DNA sequence or to conserved *trans*-acting factors.

This study showed that, even though many NFRs are maintained in the foreign context, the nucleosome depletion at NFRs are not as strong as in wild type. This suggests that, whereas poly(dA:dT) are important, other *trans*-acting factors play important role in NFR formation as well. The other conclusion which was drawn from this experiment was that position of +1 nucleosome is not determined by DNA sequence but linked to transcription initiation. Remarkably, the comparison between nucleosome maps in YAC and wild type revealed many NFRs which appeared in coding regions and not associated with poly(dA:dT) elements. These fortuitous NFRs are associated with intragenic transcripts and flanked by reasonably well-positioned nucleosomes. Authors, note that these NFRs are associated with TFIIB binding and most likely determined by fortuitous transcription factor binding sites that are recognized by transcription factors of *S. cereviciae.* Transcription factors bound to fortuitous binding sites recruit chromatin remodelers, which evict histones and generate NFR.

In general, it has been shown that chromatin architecture at promoters is tightly linked to processes of transcription initiation and elongation. Previous studies have made great achievements in elucidating mechanisms underlying nucleosome positioning. However, mechanistic quantitative explanation of nucleosome patterns *in vivo* is still missing.

1.4 Outline of the thesis

The content of the thesis is organized as follows: in chapter 2 we introduce thermodynamic biophysical model for calculating nucleosome and transcription factor occupancies. We also introduce statistical positioning effect and how it may affect binding of transcription factors. The chapter 2 mostly addresses a question of how competition with transcription factors can affect nucleosome positioning. We first examine nucleosome experimental data and address the question of reproducibility of the data across different experiments carried out in several labs. Then, we introduce a new method for quality assessment for prediction of the model and use it to optimize parameters of the model to fit experimental data. We focus on how transcription factors can explain observed *in vivo* nucleosome positioning and which transcription factors play crucial role in establishing nucleosome patterns at promoters of genes.

In chapter 3 we address a question of how nucleosomes and promoter architecture affect binding of TFs. We model binding of TFs in the context of chromatin to a cluster of binding sites and investigate what factors determine main characteristics of TF binding. Finally, we study how TFBSs in the real genomes position relative to each other and show that there are certain biases in spacing between TFBSs, probably due to effects caused by competition with nucleosomes.

Chapter 2

Nucleosome free regions in yeast promoters result from competitive binding of transcription factors that interact with chromatin modifiers

Extended version of this chapter is accepted at PLOS Computational Biology

Evgeniy A. Ozonov[1] and Erik van Nimwegen[1,*]

1. Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland.

* Corresponding Author: Erik van Nimwegen, erik.vannimwegen@unibas.ch.

Because DNA packaging in nucleosomes modulates its accessibility to transcription factors (TFs), unraveling the causal determinants of nucleosome positioning is of great importance to understanding gene regulation. Although there is evidence that intrinsic sequence specificity contributes to nucleosome positioning, the extent to which other factors contribute to nucleosome positioning is currently highly debated. Here we obtained both in vivo and in vitro reference maps of positions that are either consistently covered or free of nucleosomes across multiple experimental data-sets in Saccharomyces cerevisiae. We then systematically quantified the contribution of TF binding to nucleosome positiong using a rigorous statistical mechanics model in which TFs compete

with nucleosomes for binding DNA. Our results reconcile previous seemingly conflicting results on the determinants of nucleosome positioning and provide a quantitative explanation for the difference between in vivo and in vitro positioning. On a genome-wide scale, nucleosome positioning is dominated by the phasing of nucleosome arrays over gene bodies, and their positioning is mainly determined by the intrinsic sequence preferences of nucleosomes. In contrast, larger nucleosome free regions in promoters, which likely have a much more significant impact on gene expression, are determined mainly by TF binding. Interestingly, of the 158 yeast TFs included in our modeling, we find that only 10-20 significantly contribute to inducing nucleosome-free regions, and these TFs are highly enriched for having direct interations with chromatin remodelers. Together our results imply that nucleosome free regions in yeast promoters results from the binding of a specific class of TFs that recruit chromatin remodelers.

2.1 Introduction

The genomes of all eukaryotic organisms are packaged into nucleosomes, which are the fundamental units of chromatin, each composed of approximately 147 base pairs (bp) of DNA wrapped around a histone octamer. Recent developments in technologies for measuring chromatin marks by chromatin immunoprecipitation (ChIP) on microarrays (ChIP-Chip) or by sequencing (ChIP-seq) have enabled the construction of genome-wide maps of nucleosome positions and modifications at high resolution across various conditions. These experimental data have revealed that nucleosomes are not uniformly distributed across the genome but rather that transcription start and termination sites are relatively depleted of nucleosomes [55, 69]. Furthermore, nucleosome positioning has been shown to vary across physiological conditions [100].

It has long been accepted that nucleosomes have intrinsic sequence preferences which influence nucleosome positioning, e.g. [64, 102, 105]. At the same time, it has also long been known that barriers in the DNA can cause nucleosomes to be 'statistically positioned' relative to such barriers, introducing a periodic pattern of nucleosome occupancy on both sides of the barrier [52]. Given the fact that nucleosomes may cover more than 80% of the genome [55], it is therefore also conceivable that a relatively small number of barriers on the DNA, in combination with statistical positioning relative to these barriers, determines most of the observed nucleosome positioning. For example, recent work suggests that nucleosome occupancy patterns around TSSs could at least partly be explained by such statistical positioning [72].

Probably the most obvious class of candidate molecules that could introduce condition-specific barriers on the DNA are sequence-specific transcription factors (TFs). Indeed, for some specific promoters in *S. cerevisiae* it has been established that binding of TFs is a major determinant of nucleosome positioning in the promoter region, e.g. [9, 30, 118]. Moreover, the resulting nucleosome positioning has major effects on gene regulation from these promoters. In addition, for a few TFs it has been established that their binding induces local nucleosome exclusion genome-wide [8, 31, 50, 55].

Although it is thus clear that both intrinsic sequence preferences of nucleosomes and competitive binding of other DNA binding factors play a role in nucleosome positioning, the relative importance of these factors have come under intense debate in recent years. For example, it has been proposed that the positioning of nucleosomes, in particular in *S. cerevisiae*, is mainly determined by intrinsic sequence preference of the nucleosomes, i.e. [95]. In this view, nucleosomes are mainly positioned by a 'code' in the DNA sequence and the accessibility of the DNA to TFs is downstream of this sequence-guided nucleosome positioning. However, these conclusions were challenged by several studies which suggested nucleosome sequence specificity can only explain a modest fraction of nucleosome positioning, and that statistical positioning likely also plays an important role [20, 55, 69, 80]. More recently, several groups have undertaken further experimental investigations into this question, in particular by experimentally comparing nucleosome positioning *in vivo* and *in vitro* [48, 127]. Although there is general agreement that these experimental studies confirmed that both intrinsic sequence preferences and the competitive binding of TFs play a role in nucleosome positioning, different authors came to strikingly different, and often seemingly contradictory conclusions regarding which of these factors play a dominant role [21, 47, 59, 96, 104]. It is thus clear that, rather than lacking sufficient experimental data, the current challenge in furthering our understanding of the determinants of nucleosome positioning lies in the quantitative interpretation of this data.

Here we show that, by analyzing existing experimental data in combination with rigorous computational modeling, important novel insights can be gained that reconcile previous seemingly contradictory observations, and that suggest a new picture of the mechanisms regulating nucleosome positions. In particular, we use a biophysical model to quantitatively assess the role of TFs in determining nucleosome positioning in *S. cerevisiae*, to assess which aspects of nucleosome positioning TFs contribute to most, and to identify whether there are subsets of TFs that play a predominant roles in this process. *S. cerevisiae* is a particularly attractive system for such an analysis because extensive nucleosome positioning data are available, and because it is essentially the only organism in which sequence-specificities are available for the very large majority of TFs.

Rather than assuming that intrinsic sequence preferences determine nucleosome positioning and that TF binding occurs preferentially at those regions not covered by nucleosomes, or vice versa, assuming that TF binding sets boundaries in the DNA against which nucleosomes are statistically positioned, in our model the TF binding and nucleosome positioning patterns are determined by a dynamic competition of all TFs and nucleosomes for binding to the DNA. Our model incorporates both the sequence preferences of the nucleosomes and of all TFs in a thermodynamic setting, and rigorously calculates the resulting equilibrium occupancies genome-wide as a function of the concentrations of all TFs and the nucleosomes.

Using this model in combination with experimental data we find that TF binding makes a substantial contribution to nucleosome positioning but only at a specific subset of genomic positions. In particular, the linker regions between nucleosomes can be clearly divided into two classes based on their size: the large majority of linkers is small (≈ 15 bp) and occurs within large nucleosome arrays in gene bodies, whereas a minority of linkers is large (> 80 bp) and occurs predominantly in promoters. Our results show that the phasing of the small linkers within nucleosome arrays, and thereby the majority of nucleosome positioning genome-wide, is mainly determined by sequence preferences of nucleosomes. In contrast, the larger nucleosome free regions in promoters, which are likely most relevant for effects on gene expression, are mainly determined by competitive binding of TFs. By applying our model to data on nucleosome positioning *in vitro* we also confirm that the ability of TFs to explain nucleosome positioning in promoters is restricted to *in vivo* data. Thus, our model provides a quantitative and mechanistic explanation for the observed discrepancies between *in vivo* and *in vitro* nucleosome positioning. Most strikingly, our results also show that, rather than all TFs contributing roughly equally to the competition with nucleosomes, the effect of TFs on nucleosome positioning is restricted to a relatively small set of about $10-20$ TFs. Although one might expect that these TFs are simply the highest expressed TFs with the largest number of TFBSs genome-wide in the conditions in which the experiments were performed, we find this not to be the case. Instead, we find that these TFs are highly enriched for having known protein-protein interactions with chromatin remodeling complexes, histones, and chromatin modification enzymes. Thus, the mechanistic picture suggested by our results is that there is a specific class of TFs who, upon binding to the DNA, recruit chromatin modifiers that then mediate local expulsion of nucleosomes.

2.2 Results

2.2.1 A biophysical model of TF and nucleosome binding to genomic DNA

To rigorously investigate the competition between TFs and nucleosomes for binding to DNA, and the role of TFs in nucleosome positioning, we take a statistical mechanics approach in which we explicitly consider all possible non-overlapping binding configurations to the genome for nucleosomes and a large set of TFs, assigning a probability to each configuration using standard Boltzmann-Gibbs statistics. The basic approach, which uses dynamic programming to efficiently sum over all possible binding configurations, has been used in computational methods for analysis of transcription regulation for over a decade, e.g. [18, 20, 83, 94, 114], and has been used more recently to specifically investigate the effect of competitive binding of nucleosomes and TFs [87, 119]. Here we use this approach to comprehensively investigate the role of TFs in determining nucleosome positioning. We employ an unprecendented complete set of 158 TF binding models, we investigate the dependence on the concentrations of these TFs, and we also introduce tunable sequence-specificities for all TFs and nucleosomes.

The model is explained in detail in the Materials and Methods. Briefly, each TF t is assumed to bind DNA segments of a fixed length l_t and, for any length-l_t DNA segment s, a binding energy $E(s|t)$ is determined. The energies $E(s|t)$ are calculated from a weight matrix representation of the TF's binding sites [14] and involve a tunable scale parameter γ_t which controls the sequence-specificity of the TF. To obtain energy matrices for the large majority of sequence-specific TFs in *S. cerevisiae* we used a collection of 158 WMs that we curated previously [19] and that are based on a combination of ChIP-chip and *in vitro* binding data. Notably, while the WMs allow us to determine how the binding energy (measured in units k_BT) varies across positions in the genome for each TF, the WMs do not allow us to determine the sequence-independent contribution to binding energy, i.e. the overall 'stickines' of each TF for DNA. To compare binding energies across TFs we set the sequence-independent contribution to the binding energy such that all TFs have equal overall affinity for the DNA (see Materials and Methods).

Of the computational work done on nucleosome positioning, probably most effort has been invested in developing models for nucleosome sequence-specificity based on data from both *in vivo* and *in vitro* nucleosome binding, e.g. [48, 95]. Exploiting analytical results from statistical mechanics, Locke et al. [59] rigorously inferred the energies of nucleosome binding from high-throughput data and used these to evaluate several models of different complexity for the sequence specificities of nucleosomes. The results from this study suggested that the sequence specificity of nucleosomes can be captured by fairly simple models. As we discuss below, our own analysis suggests that the performance of different models of nucleosome sequence specificity depends on the precise data-set and performance evaluation method used, but that all models make highly correlated predictions (Figure 2.1A). Of the models analyzed, the model of [48] gave robustly high performance across data-sets and we use this model in our study. In particular, we assume that nucleosomes bind to DNA segments of 147 nucleotides and determine an energy of binding $E(s|\text{nucl})$ for any length 147 segment s using a generalization of the model of [48], involving a scale parameter γ_{nucl} that controls the sequence specificity of the nucleosomes, analogous to the scale parameters γ_t for the TFs (see Materials and Methods). The parameter γ_{nucl} allows us to investigate the effect of enhancing or decreasing the nucleosome sequence specificity. For example, when setting $\gamma_{\text{nucl}} = 0.4$, the variation in nucleosome binding energies across different sequences is reduced to 40% of the energy variations predicted by the model of [48].

As mentioned above, the model assumes that any DNA segment can only be bound by a single TF or a nucleosome at a time. Although it is likely that there are exceptions to this simplification, it is generally accepted that TFs and nucleosomes compete for binding to DNA. In absence of specific information as to which TFs compete with nucleosomes and which can co-bind with nucleosomes, we make the simplifying assumption that all TFs compete with nucleosomes, as has been done previously by others [87, 119]. Like previous approaches, e.g. [72, 95, 96, 119], our model also assumes that the average occupancy profiles across a population of cells are well approximated by their thermodynamic equilibrium averages. Notably, given that there are many ATP-driven processes that cause nucleosome turnover and displacement by chromatin remodelers, it is not a priori clear that this equilibrium assumption holds. Ours and previous computational approaches thus essentially assume that these ATP-driven processes act mainly to affect kinetics, i.e. to allow nucleosomes to resample their positions, without systematically biasing their positioning. Some recent evidence appears to support this assumption [111].

The model considers all possible non-overlapping configurations C of TFs and nucleosomes bound along the genome. For each configuration C, a total energy $E(C|c, \gamma)$ is calculated. This energy depends on the concentrations of nucleosomes c_{nucl} and all TFs c_t, which we collectively denote as c, and also on all energy scale factors γ that determine sequence-specificity (Materials and Methods). The probability $P(C|c, \gamma)$ to find a cell in configuration C is then given by the standard Boltzmann-Gibbs formalism as

$$P(C|c,\gamma) = \frac{e^{-\beta E(C|c,\gamma)}}{Z}, \tag{2.1}$$

where $\beta = 1/(kT)$ is the inverse temperature, Z is the partition sum, and we have explicitly indicated that these probabilities depend on the concentrations c and scale factors γ. As explained in Materials and Methods, both the partition sum and the fractions of the time each TF t is bound at each genomic position can be calculated efficiently using standard dynamic programming techniques.

In summary, given a set of input concentrations c for all TFs and nucleosomes, the model efficiently calculates the equilibrium binding frequencies of all TFs and nucleosomes across the entire genome. Note that, because all TFs and nucleosomes are in competition for binding to the DNA, the occupancy of any factor to a sequence segment of the genome in principle depends, not only on the concentration of this factor and its affinity to the sequence segment, but on the concentrations of all other factors and their affinities to all other locations in the genome. Thus, the TF and nucleosome occupancy profiles across the genome can be changed by varying the concentrations c and scale factors γ. In particular, these parameters can be optimized to maximize the agreement with experimentally determined nucleosome occupancy profiles.

2.2.2 Comparing model predictions with experimental nucleosome position profiles

Many experimental studies have been carried out to map nucleosome positions in eukaryotic species, e.g. [46, 70, 93, 113], and in *Saccharomyces cerevisiae* in particular, e.g. [28, 48, 55, 69, 100, 120, 127], so that several data-sets of nucleosome positions in *S. cerevisiae* are available. In order to determine how to meaningfully compare computational predictions with these experimental data, we first performed a comparative analysis of several experimental data sets. Patterns of nucleosome positioning that are typically highlighted in publications, such as the nucleosome-depleted regions upstream of the transcription start sites (TSSs) and well-positioned nucleosomes immediately downstream of TSS, involve genome-wide averages of nucleosome occupancy across a class of positions. Such average patterns are robust to fluctuations and are shared by all data-sets.

Previous works have assessed the performance of models of nucleosome sequence specificity by determining both the predicted and experimentally observed nucleosome

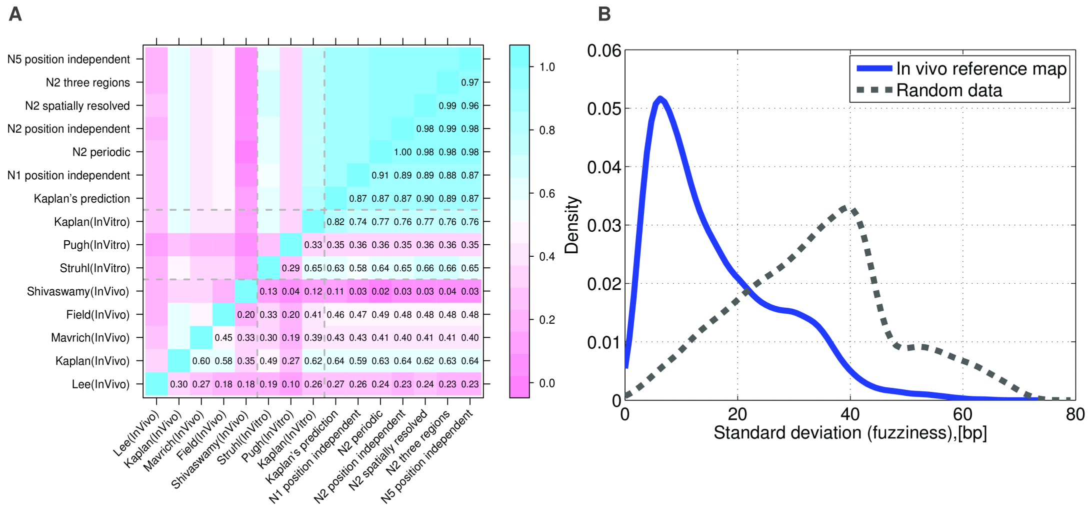

Figure 2.1: Reproducibility of *in vitro* and *in vivo* nucleosome data across different experiments and performance of nucleosome sequence-specificity models. **A:** Pearson correlation coefficients of the per-base nucleosome coverage between various experimental data-sets measuring nucleosome occupancy either *in vivo* [28, 48, 55, 69, 100] or *in vitro* [48, 127, 128], and predictions from a number of models of nucleosome sequence-specificity [48, 59]. **B:** Reproducibility of annotated nucleosome positions across the *in vivo* data-sets. For each annotated nucleosome in the reference map of [45], we calculated the standard deviation in the annotated positions of the corresponding nucleosomes across the 6 data-sets used to construct the map. The blue curve shows the distribution of standard deviations across nucleosomes. The grey dotted curve shows the analogous distribution that is obtained using randomized data (see Materials and Methods). The high reproducibility of nucleosome positions across different data-sets justifies the use of binary data, i.e. positions of "linkers" and "nucleosomes", instead of Pearson correlation for evaluation of the performance of computational models for predicting nucleosome positions.

occupancies across individual regions of the genome, and by calculating the Pearson correlation of these nucleosome occupancy profiles. To assess the validity of such an approach, we calculated Pearson correlations between observed occupancy profiles of several experimental data-sets (both *in vivo* and *in vitro*) as well as several models of nucleosome sequence specificity (Figure 2.1A). This shows that, unfortunately, the occupancy profiles correlate only weakly across different experimental data-sets, with Pearson correlation coefficients typically ranging from $r = 0.2$ to $r = 0.45$ for *in vivo* data-sets, and only marginally higher for *in vitro* data-sets. This large variability across data-sets may to some extent be due to biases of the technological platforms. For example, it is well known that the nucleotide composition and propensity to form secondary structures of the reads can systematically bias the read counts in ChIP-seq by more than 10-fold [36, 104]. Variations in details of the ChIP protocol are likely also responsible for some of the variation across data-sets, and previous studies have indicated that MNase digestion bias may also systematically affect nucleosome positioning data [21, 59]. Since all experiments were performed in YPD, true biological variation is likely only a minor source of variation in these data.

In contrast to the experimental data, the occupancy profiles predicted by the different computational models are all highly correlated. Moreover, the correlations across models for a given data-set vary much less than the correlations for a given method vary across data-sets. For example, all models consistently perform better on *in vitro* than on *in vivo* data. Among the *in vivo* data-sets, all methods perform by far best on the *in vivo* data of Kaplan et al.[48] (which is also far more correlated with *in vitro* data than any other *in vivo* data-set) and far worst on the *in vivo* data of Shivaswamy et al.[100]. Thus, comparison of different models with existing data supports the conclusions of [59] that different models of nucleosome-specificity perform similarly in explaining nucleosome positioning. Since the model of Kaplan et al.[48] exhibits highest performance for the majority of *in vivo* and *in vitro* data-sets, we chose to use this model in our analysis. However, the weak correlation of nucleosome occupancy profiles across data-sets shows that assessing the performance of computational predictions by directly comparing predicted and observed nucleosome occupancies is highly problematic. A meaningful comparison of computational models requires that one first extracts those features of the nucleosome positioning that are reproducible across experimental data-sets.

In contrast to the absolute value of the ChIP signal, we observed that the positions of local maxima and minima in nucleosome occupancy are much better reproduced across data-sets. This reproducibility of the 'peaks and troughs' in the nucleosome occupancy profile has been observed previously [45], and has been used to create a reference set of ‘nucleosome’ and ‘linker’ segments. In this procedure, local maxima and minima are used to annotate nucleosomes and linkers in each data-set. These annotations are then intersected, with reference nucleosomes placed at the consensus positions of regions annotated as nucleosomes in all data-sets, and reference linkers the regions free of nucleosomes in all annotations. That the positions of annotated nucleosomes are highly reproducible across data-sets, especially compared to raw coverage and compared to nucleosome maps based on randomized data, is illustrated in Figure 2.1B. The annotated positions of individual nucleosomes across different data-sets typically vary by less than 10 base pairs from the reference position (blue curve in Figure 2.1B) and the vast majority of annotated nucleosome positions vary by less than 20 bp from the reference position. In contrast, on randomized data positions of annotated nucleosomes typically vary by roughly 40 bp from the reference position (dotted curve in Figure 2.1B).

In summary, although ideally we would like to test whether computational models can predict relative nucleosome occupancies across the genome, it is not possible to meaningfully perform such an assessment given the variability observed in the experimental data. We thus evaluate the performance of different models by assessing their ability to predict nucleosome and linkers that occur consistently across different data-sets. We use the reference set annotated by [45] consisting of roughly $60'000$ annotated linker regions and $21'000$ annotated nucleosomes, that together cover about 50% of the genome, to assess the performance of the model in predicting *in vivo* nucleosome positioning. In addition, we have applied a similar annotation procedure (Materials and Methods) to produce a reference set of nucleosomes and linkers from 3 *in vitro* data-sets, which we use to assess the performance of the model in predicting nucleosome positioning *in vitro*.

To assess the model’s performance we compare the predicted nucleosome coverage at annotated linker and nucleosome segments. That is, instead of comparing the predicted and observed absolute occupancies, we assess the model’s ability to predict local maxima and minima in nucleosome occupancy, that occur consistently across data-sets. As described in Materials and Methods, based on the predicted nucleosome coverage, we classify each segment as either nucleosome or linker, and then calculate the *mutual information* I between the predicted and experimentally measured classification. Finally, we normalize this mutual information by the entropy H of the experimental classification to obtain the fraction $F = I/H$ of information that is captured by the model’s predictions, i.e. F runs from 0 (random predictions) to 1 (perfect predictions). An F value of 0.2 means that the model captures 20% of all the information needed to specificy which of the genomic segments correspond to nucleosomes and which to linkers. We will refer F as the ‘quality score’. As mutual information is the fundamental measure of dependence between two distributions[43, 98], we consider the quality score F the most rigorous quantification of model performance. However, as we show below, highly similar results are obtained with other performance measures that are popular in machine learning, such as area under the ROC curve (AUC).

2.2.3 Optimal fits to nucleosome positioning require weak nucleosome sequence specificity

We first tested what quality score can be obtained by the intrinsic sequence specificity of the nucleosomes, i.e. leaving all TFs out of the model, and how the quality of the fit depends on the sequence specificity of the nucleosomes. Figure 2.2A shows the quality scores F that are obtained for different scale factors γ_{nucl} on nucleosome sequence specificity (with 0 representing no sequence preference whatsoever and 1 representing the specificity used in Kaplan et al. [48]). The optimal fit is obtained for $\gamma_{\text{nucl}} \approx 0.47$, which corresponds to significantly lower nucleosome sequence specificity than those used in Kaplan et al. [48]. That is, for the model of [48], the standard deviation of nucleosome binding energies is approximately $1.64 k_B T$ across the genome (0.97kcal/mole), whereas we observe optimal fits for roughly 2-fold lower variations in binding energies (roughly $0.77 k_B T$). Moreover, the quality score depends weakly on γ_{nucl} and becomes small only for extremely small sequence specificities.

These results may seem contradictory, given that the sequence-specificity model of Kaplan et al. was developed specifically with the aim of explaining nucleosome positioning. However, Kaplan et al. optimized the overall Pearson correlation between predicted and observed nucleosome coverage, which depends strongly on the variation in absolute nucleosome occupancies. In contrast, the quality score F depends mainly on the locations of local maxima and minima in the occupancy, and much less on the absolute amount of variation in nucleosome occupancy. To investigate this further, we compared the distribution of nucleosome occupancies for the model with different values of γ_{nucl} with the distribution of nucleosome occupancies for the model of Kaplan et al. and the experimentally observed distribution of nucleosome occupancies for the data of Lee et al. [55] (Materials and Methods, and note that very similar distributions are obtained from other experimental data-sets; Figure A.2.10).

As shown in Figure 2.2B, the model of Kaplan et al. [48] predicts an overall nucleosome coverage that is dramatically lower than our fits, i.e. with a median nucleosome coverage of about 0.3. Such a coverage distribution is strongly at odds with the experimental data which shows that, rather than 30%, about 80% of the genome is covered

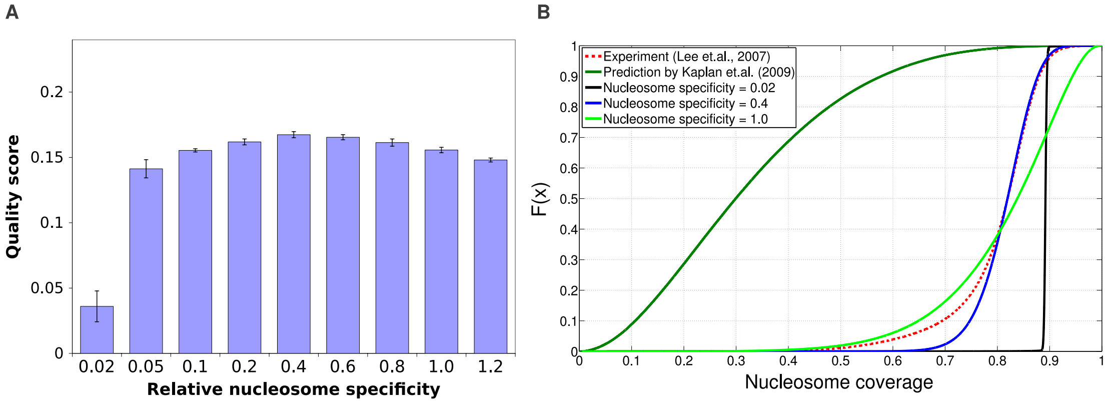

Figure 2.2: Performance of models that include only nucleosome sequence specificity. **A:** Fraction of information regarding experimentally annotated linker and nucleosome positions explained by the nucleosome-only model (quality score, vertical bars) as a function of relative nucleosome specificity. The relative nucleosome specificity is controlled by the scale factor γ_{nucl}, where $\gamma_{\text{nucl}} = 1.0$ corresponds to the sequence specificity of the model of Kaplan et al. [48], for which the binding energy of the nucleosomes has a standard-deviation of $1.64k_BT = 0.97$kcal/mole across the genome. The error-bars indicate standard-errors across 5 separate test sets. **B:** Experimentally observed cumulative distribution of nucleosome coverages (fraction of time a given genomic position is covered by a nucleosome) from [55] (red dotted line) and cumulative distributions of predicted nucleosome coverage of the models of [48] (dark green line) and our model using nucleosome specificity scale parameters of $\gamma_{\text{nucl}} = 0.02$ (black line), $\gamma_{\text{nucl}} = 0.4$ (blue line), and $\gamma_{\text{nucl}} = 1.0$ (light green line).

by nucleosomes, e.g. [42, 51, 55, 100]. It is likely that the unrealistically low nucleosome occupancy of Kaplan et al. [48] is an artefact of optimizing the Pearson correlation in nucleosome coverage, since this objective function favors high variance in predicted nucleosome coverage, and does not penalize the mismatch in the average nucleosome coverage.

For our model, the coverage distribution indeed strongly depends on the nucleosome specificity. Strikingly, by far the best fit between the observed and predicted coverage distribution occurs precisely at the specificity that maximizes our quality score (i.e. at $\gamma_{\text{nucl}} = 0.4$). This demonstrates that, in contrast to the predictions of Kaplan et al. [48], our fits produce realistic nucleosome coverage profiles, in spite of not specifically optimizing these coverage profiles. In fact, at the optimal nucleosome specificity, the predicted and experimentally observed nucleosome coverage distribution is virtually identical for the 70% of base pairs in the genome with highest nucleosome coverage (blue and red curves in Figure 2.2B). The main deviation between model and experimental data is that the model fails to predict regions with low nucleosome coverage that are observed experimentally. Indeed, as we will see below, whereas the model correctly predicts almost all nucleosomes, the model fails to correctly predict a substantial fraction of linker regions as nucleosome free.

In summary, optimizing the quality score F produces much more realistic fits to the nucleosome coverage distribution than previous models, and shows that the best fits are obtained with only weak nucleosome sequence-specificity.

2.2.4 Transcription factor binding plays a major role in explaining nucleosome free regions at promoters

We next investigated to what extent competition with TFs improves the predicted nucleosome positioning. We first considered models in which, besides the nucleosomes, there is only a single TF. For each of these models we fitted the 4 parameters (i.e. the concentrations and sequence specificity of both nucleosomes and the TF) using simulated annealing, and calculated the quality score F obtained with this model using 80/20 cross-validation (Materials and Methods). We ranked TFs by the z-statistic they obtained in cross-validation (Materials and Methods), and then investigated what quality scores F can be obtained using the top 5, 10, 20 and top 30 TFs, refitting all concentrations and sequence specificity parameters. We find that adding the TFs clearly increases the quality of the predictions on the test-sets, although the improvement is relatively small, i.e. from $F \approx 0.17$ to $F \approx 0.2$, Figure 2.3A. Given this modest increase in F and the large number of parameters involved when including many TFs in parallel, one may wonder whether these results are affected by overfitting. However, as shown in Figure A.2.11, the observed F scores on train and test sets are essentially identical. In addition, adding the TFs to the model further improves the match between the observed and predicted nucleosome occupancy distribution (Figure A.2.10).

As already observed in [45], the length distribution of linkers is bimodal. The large majority of linkers is short, around on average 15 bps in length, corresponding to short linkers within arrays of nucleosomes. There is a second class, corresponding to roughly 25% of all annotated linkers, that are much longer, i.e. each more than 80 bps long. We will refer to these longer linkers as ‘nucleosome free regions’ (NFRs). We next asked whether TFs contribute more to explaining the positioning of the short linkers or the longer NFRs. Moreover, as TFs are expected to bind predominantly to promoter regions, we also investigated whether the contribution of the TFs to explaining nucleosome positioning is most significant in promoters (defined as running from 500 bp upstream

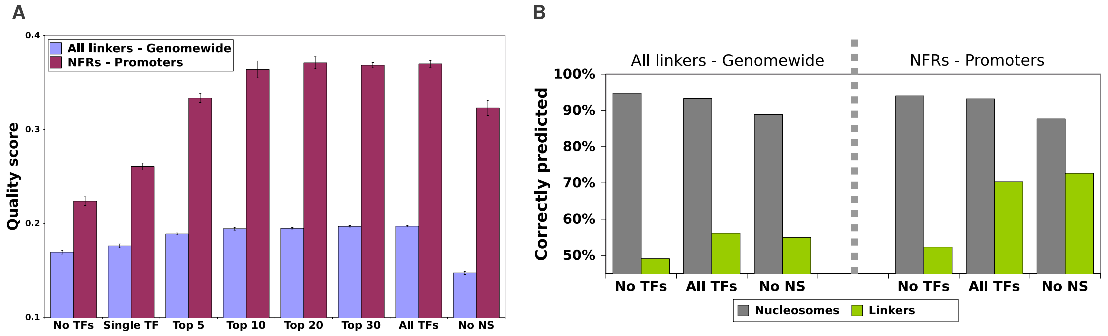

Figure 2.3: Incorporating competition with TFs improves predicted nucleosome positioning, particularly in promoter regions. **A**: Ability to predict nucleosome positioning as a function of the number of TFs used in the model. The bars show the fraction of all information regarding nucleosome positioning explained (quality score F) by each model. Results are shown for, from left to right, the model including only nucleosomes (no TFs), only the best TF, the top 5 TFs, top 10 TFs, etcetera. The rightmost pair of bars correspond to a model including all TFs but without any sequence specificity for the nucleosomes $\gamma_{\text{nucl}} = 0$. Blue bars correspond to quality scores for predicting all nucleosomes and linkers genome-wide and red bars correspond to quality scores for predicting nucleosomes and nucleosome free regions (long linkers) within promoters. The error bars show standard-error across 5 independent test-sets. **B**: Fractions of correctly predicted nucleosomes (grey bars) and linkers (green bars) for, from left to right, the model with nucleosome sequence specificity and no TFs, the model with all TFs, and the model with all TFs but no nucleosome sequence specificity. The left half of the figure shows results for predicting all linkers and nucleosome genome-wide, and the right half for predicting NFRs and nucleosomes in promoters.

to 500 bp downstream of TSS). We find that, generally, inclusion of the TFs leads to a substantially larger increase in performance for promoter regions, and TFs contribute much more to explaining NFRs than explaining small linkers (Figure A.2.12). In particular, considering NFRs and nucleosomes in promoter regions, inclusion of TFs almost doubles the quality score F, i.e. from 0.23 to 0.38, Figure 2.3A, red bars. As an aside, we note that these observations do not depend on assessing the model's performance by the quality score F. As shown in Figure A.2.13, we find essentially the same results when assessing the model's performance using ROC curves, and the area under the curve (AUC) is almost perfectly correlated ($r = 0.99$) with the quality score F. It is also noteworthy that, both when predicting all linkers genome-wide or NFRs in promoters, even though up to 158 TFs can be incorporated, the model essentially reaches its optimal performance after adding the first $10 - 20$ TFs. We investigate this in more detail below.

It thus appears that TFs contribute not so much to explaining positioned nucleosomes, but rather explain the location of longer NFRs, especially in promoters. Further supporting this observation, the rightmost pair of bars in Figure 2.3A shows the performance of the model including all TFs but with nucleosome sequence specificity removed, i.e. $\gamma_{\text{nucl}} = 0$. We see that removing nucleosome sequence specificity only modestly affects the ability of the model to predict NFRs in promoters. In contrast, the performance on predicting all linkers genome-wide drops significantly when nucleosome sequence specificity is removed, even falling clearly below the performance of the model without TFs. This is further confirmed by closer examination of the errors that the fitted models make (Figure 2.3B).

For all models, the large majority of nucleosomes is correctly predicted and the fraction of correctly predicted nucleosomes is most strongly affected by removing the sequence specificity of the nucleosomes, i.e. from 95% correct for the model with only nucleosome sequence specificity to 88% for the model with all TFs and no nucleosome specificity. The fraction of correctly predicted linkers is much smaller, e.g slightly below 50% for the model without TFs. Adding the TFs to the model consistently increases the fraction of correctly predicted linkers, and this increase does not require nucleosome sequence specificity. When considering all linkers genome-wide, the increase in correctly predicted linkers is relatively modest, i.e. from 50% to 56%. However, for NFRs in promoters the fraction of correctly predicted NFRs increases from 50% to around 70%. In summary, correctly predicting the phasing of nucleosome arrays over gene bodies crucially depends on nucleosome sequence specificity and is only weakly affected by including TFs, whereas correctly predicting NFRs is strongly dependent on inclusion of the TFs and is almost independent of nucleosome sequence specificity.

2.2.5 Characterization and additional validation of the fitted model

To characterize the biophysical properties of the fitted model we first determined the overall statistics of nucleosome and TF occupancies (Figure 2.4A). Nucleosomes cover more than 80% of the genome, and most of the remaining regions of the genome are uncovered, with all TFs combined covering less than 1% of the genome. The top 10 TFs with the highest genomic coverage occupy between 0.15% and 0.02% of the genome, corresponding to roughly 1500 and 200 binding sites genome-wide.

For the nucleosomes and the top 10 TFs with highest genomic coverage in the fitted model we also determined the mean and standard-deviation of the binding energies at their binding sites, and the entropy of the distribution of binding probabilities per site (Materials and Methods). The latter quantity is low whenever the TF's coverage results

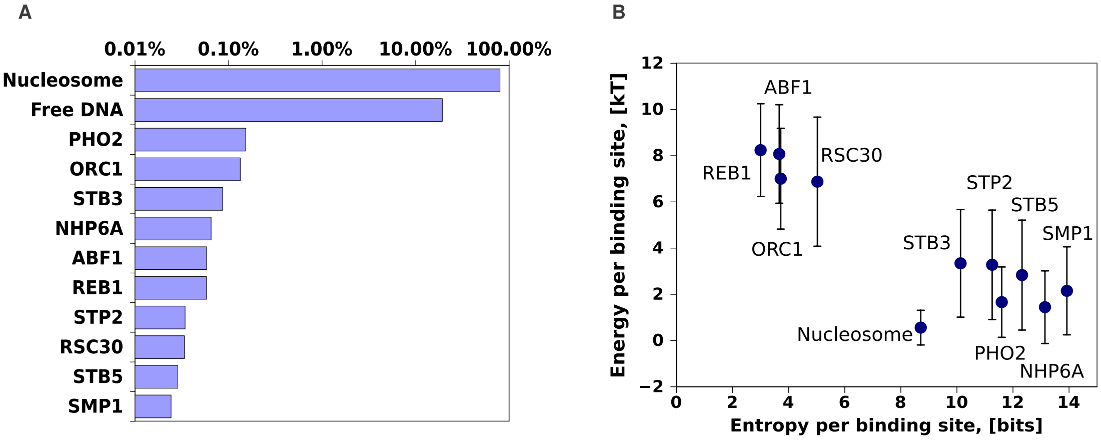

Figure 2.4: Biophysical properties of the fitted model. **A:** Average fraction of the genome covered by nucleosomes, free DNA, and the top 10 TFs with highest coverage. **B:** Average and standard-deviation of the binding energies (in units k_BT) at binding sites for nucleosomes and the top 10 TFs with highest coverage (vertical axis), against the average entropy per binding site of the distribution of binding probabilities for the corresponding TFs (horizontal axis).

from strong sites with high frequencies of binding, and is high when the TF's coverage comes from a large set of weak sites with lower binding frequencies. The results (Figure 2.4) show, first of all, that the binding sites of nucleosomes have both the lowest binding energy and the lowest variation in binding energies, i.e. they are the least sequence specific. Interestingly, the top 10 TFs clearly fall into 2 classes: a set of TFs (ABF1, REB1, ORC1, and RSC30) that are highly sequence specific and have strong binding sites, and a class of much less sequence specific TFs (PHO2, NHP6A, etcetera) that bind at a much larger number of weaker sites.

As has been observed previously, e.g. [55, 69], averaged nucleosome coverage profiles show a characteristic pattern relative to the starts of genes with a nucleosome depleted region immediately upstream of TSS, followed by a well-positioned nucleosome immediately downstream of TSS and a periodic pattern of nucleosome coverage downstream into the gene body. Although the nucleosome sequence specificity by itself, i.e. without including TFs, reproduces some of this pattern at the 5' end of genes (Figure 2.5A), the observed nucleosome depleted region and the oscillatory pattern into the gene body is much weaker than observed experimentally. As an additional test of the validity of our model, we checked whether inclusion of the TFs improves this average coverage profile relative to gene starts and ends.

We find that adding TFs to the model significantly improves the match between the

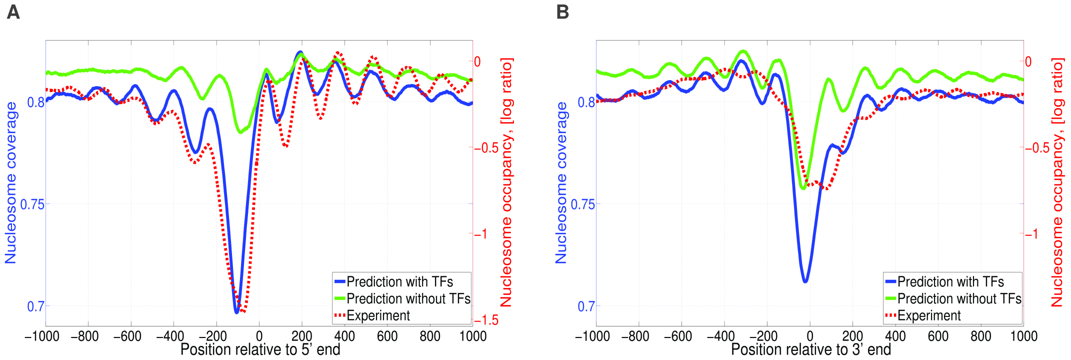

Figure 2.5: Predicted and observed nucleosome profiles around 5' and 3' ends of genes. **A**: Averaged nucleosome coverage near transcription starts. Each curve shows the average nucleosome coverage at different positions relative to transcription start averaged over all genes. Red dashed lines correspond to experimentally measured nucleosome coverage (data from [55], right vertical axis). The solid lines correspond to the predicted nucleosome coverage by the model including only nucleosomes (light green) and the model including all TFs (blue), left vertical axis. **B**: Averaged nucleosome coverage near transcription ends. Curves are as described for panel A.

theoretically predicted and experimentally observed nucleosome coverage pattern at the 5' ends of genes (Figure 2.5A). It is noteworthy that the nucleosome-depleted region immediately upstream of TSS coincides with a peak in the overall predicted binding of TFs (Figure A.2.14C), further illustrating the role of TFs in establishing nucleosome depletion in these regions. A local peak in TF binding is also predicted immediately downstream of the 3' ends of genes (Figure A.2.14D). Although at the 3' ends of genes, the inclusion of the TFs also improves the match between the theoretical predictions and the experimentally observed nucleosome coverage, the experimental data and predictions clearly disagree (Figure 2.5B). First, the width of the experimentally observed NFR is twice as big as the width of the predicted NFR. Second, the oscillations exhibited by the experimentally-determined distribution are not as pronounced as predicted by the model. This lack of a match can likely be attributed to the role of RNA polymerase. Our model considers only 158 TFs and, in particular, does not consider the effects of binding of general transcription factors and RNA polymerase. Experimental data on the positioning of the largest subunit of Pol II - Rpo21, and the general transcription factor Sua7 shows that these factors localize at 3' ends of genes [116], suggesting that they may contribute to the nucleosome free region observed at the 3' ends of genes (Figure A.2.15). This is further supported by the analysis in [27], which shows that rapid removal of Polymerase from 3' end regions increases local nucleosome occupancy.

As another validation of the model, we investigated whether the predicted TF binding matches experimental observations. For example, we compared the intergenic regions predicted to be targeted by the TFs Abf1, Reb1, and Sum1, with the observed target intergenic regions according ot the ChIP-chip data of [35]. This shows that, in spite of the fact that the model was only optimized to fit nucleosome positioning, the fitted model also accurately predicts which regions are targeted by these TFs (Figure A.2.16).

It is important to stress that, although we assess the model's performance by these global statistics, it predicts the precise locations of individual nucleosomes, NFRs, and TF binding sites. The full genome-wide nucleosome and TF coverage predictions obtained with the model including the TFs are made available through our SwissRegulon server **www.swissregulon.unibas.ch/ozonov**, allowing users to investigate in detail which NFRs at which promoters are explained by the binding of particular TFs. To illustrate the detailed comparison of the model's predictions and observed nucleosome occupancies Figure 2.6 shows the measured nucleosome coverage, the predictions of the model with and without TFs, and the predicted coverage of TFs, in two genomic regions. As the figure shows, whereas the locations of small peaks and troughs in occupancy across arrays of nucleosomes are reasonably well captured by nucleosome sequence specificity alone, competition with TF binding is needed to explain the occurrence of larger nucleosome free regions, which occur predominantly in promoters. Importantly, it is likely precisely this latter class of regions that are crucial for the effects of nucleosome positioning on gene expression.

However, this detailed comparison also reveals that, whereas the locations of TF binding typically matches the centers of observed NFRs, the predicted shape of these NFRs differs considerably between the model and the experimental observations. In particular, NFRs tend to be much narrower in the model's predictions than in the experimental data. This suggests that, although TF binding determines the genomic location where nucleosome depletion is observed, the observed nucleosome exclusion is more substantial than predicted from the steric hindrance between TFs and nucleosomes. This suggests that TF binding may recruit aditional factors involved in nucleosome exclusion. We return to this observation below.

2.2.6 Only a small subset of TFs, enriched for interacting with chromatin modifiers, crucially affects nucleosome positioning

Our model incorporates the role of TFs through a simple competition for binding DNA and one might thus naively expect that all TFs that are expressed in YPD would con-

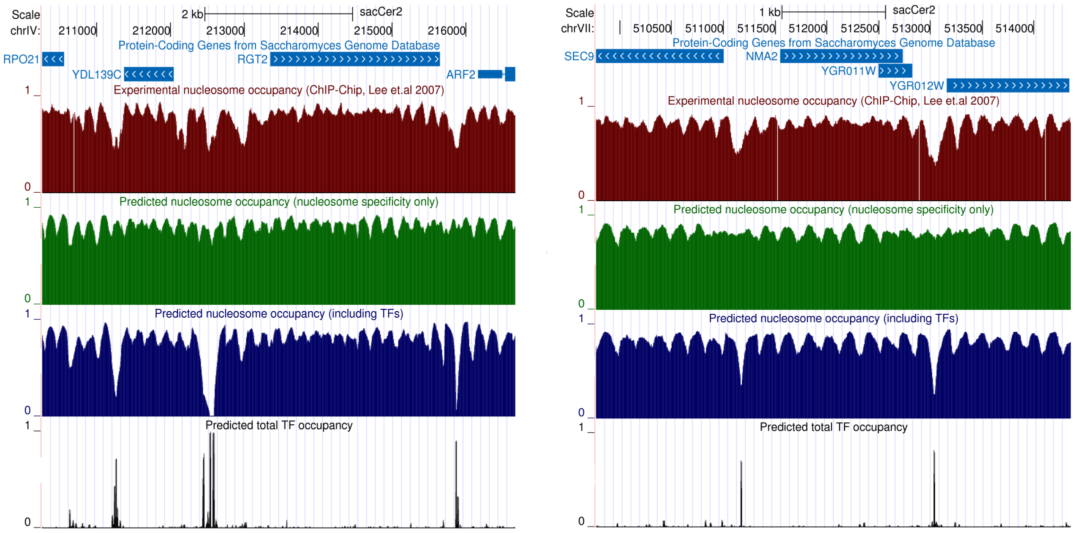

Figure 2.6: Illustration of the measured nucleosome occupancy and model predictions within individual genomic regions. Each panel shows a section of the yeast genome within our genome browser (swissregulon.unibas.ch/ozonov), with the tracks corresponding to, from top to bottom, chromosomal location, annotated genes, the measured nucleosome coverage based on the data from [55], the predicted nucleosome coverage using the model without TFs, the predicted nucleosome coverage using the model including TFs, and the total predicted TF coverage, i.e. summing over all TFs. Within the genome browser the coverage of individual TFs can also be displayed.

tribute similarly to explaining nucleosome positioning, maybe in proportion to the number of their binding sites in the genome. However, we observed above (Figure 2.3A) that when consecutively adding more TFs to the model, the performance already assymptotes after $10 - 20$ TFs. This could be due to redundancies in the contributions of the TFs, i.e. if sites for different TFs cluster in particular genomic regions, then binding by only a subset of the TFs will suffice to explain the occurrence of NFRs in these regions, and adding more TFs to the model would not further improve performance. Alternatively, it may be that there is a specific class of TFs that contribute much more to nucleosome positioning than other TFs.

To investigate this, we used 80/20 cross-validation on 5 independent training and test sets to assess, for each of the 158 TFs, whether a model containing only nucleosomes and the single TF statistically significantly outperforms the model with only nucleosome specificity, quantifying the significance by a z-statistic (Materials and Methods). Figure 2.7A shows the distribution of z-statistics obtained for the 158 TFs (blue dots), together with the distribution of z-statistics expected by chance (brown dotted curve). As the figure shows, only $15 - 20$ of the TFs significantly improve the predictions, indicating that there is indeed a specific class of TFs that dominate in explaining NFRs. Indeed, the large majority of all other TFs obtain quality scores on the test sets that are either the same or worse than the model without any TFs (Figure A.2.17).

As another validation, we checked whether the ability of this subset of TFs to explain nucleosome positioning is a specific property of the sequence specificities of yeast's TFs. That is, it is in principle conceivable that among *any* set of WMs with similar information content and sequence composition, a few will be able to help explain nucleosome positioning. To test this we constructed a set of synthetic WMs by randomly shuffling the columns of the original WMs, and fitted models with these 158 TFs in exact analogy to our fits with the original WMs. As shown in Figure 2.7A (green dots), none of the shuffled WMs perform better than expected by chance, confirming that the ability to explain nucleosome positioning is unique to the specific set of $15 - 20$ yeast WMs that we identified.

As a final test, we also evaluated whether the real WMs can explain the nucleosome positioning that is observed *in vitro* (Materials and Methods). On the one hand, since no TFs are present in the conditions at which the *in vitro* experiments are performed, the TFs should in principle not contribute to nucleosome positioning. On the other hand, as the raw *in vivo* and *in vitro* occupancies are significantly correlated (Figure 2.1A), one might expect that the TF WMs can still positively contribute to explaining *in vitro* nucleosome positioning. It is thus striking that none of the real yeast WMs

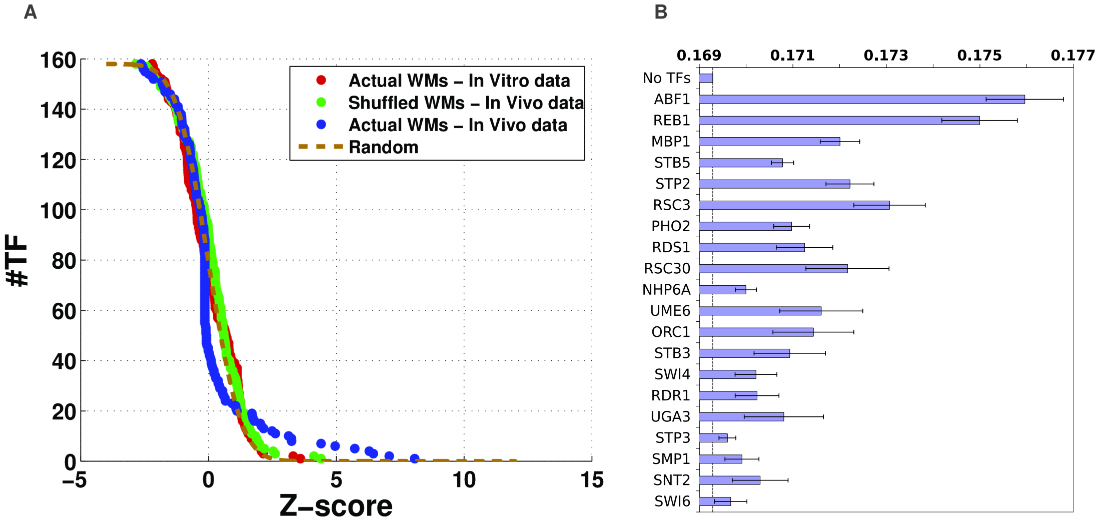

Figure 2.7: Only approximately 20 TFs contribute significantly to nucleosome positioning. **A**: For each TF an average quality score F across 5 test-sets was determined using the model containing nucleosomes and the corresponding TF. TFs were then ordered by the z-statistic $z = (F - F_{\text{noTFs}})/s_e$, with F_{noTFs} the quality score of the model without any TFs, and s_e the standard-error across the 5 test-sets (see Materials and Methods). The panel shows the reverse cumulative distribution of z-statistics observed across the 158 TFs (blue dots) together with the expected standard-normal distribution expected for random predictions (brown dotted curve). Note that about 20 TFs have z-statistics larger than expected by chance. The green dots show the reverse-cumulatives of z-statistics for the fits obtained with WMs in which the columns of each WM have been randomly shuffled. The red dots show the reverse-cumulatives of z-statistics obtained when fitting the original WMs to the *in vitro* map of nucleosome positions. Note that both the green and red dots closely follow the distribution expected by chance. **B**: The top 20 TFs that contribute most to *in vivo* nucleosome positioning sorted by their z-statistic. The bars show the average quality score F and standard-error s_e for each TF.

performs better than expected by chance in explaining *in vitro* nucleosome positioning (Figure 2.7A, red dots), i.e. including TFs does not help explaining *in vitro* nucleosome positioning. This shows that the actions of a specific set of $15 - 20$ TFs are crucial for explaining the differences between *in vivo* and *in vitro* nucleosome occupancies.

Figure 2.7B lists the top 20 TFs and shows their quality scores on the test sets. The fact that only around 20 TFs contribute significantly to nucleosome positioning raises the question of what distinguishes these TFs from the others and we investigated a number of hypotheses. One might hypothesize that the top TFs are simply those that are highest expressed in YPD, or those which occupy most sites genome-wide. However, expression data indicates that these TFs are not particularly highly expressed in YPD compared to other TFs (Figure A.2.18, data from [58]). Consistent with this, the genome-wide number of binding sites, as observed in genome-wide ChIP-chip experiments (Figure A.2.19), is not generally higher for these TFs. Thus, the role of these TFs in nucleosome positioning is not simply the result of increased binding or expression in YPD. Notably, for a considerable number of TFs our model predicts essentially no binding sites, and not all of these TFs are low expressed in YPD. It is conceivable that the low number of predicted sites for these TFs indicates that these TFs do not compete with nucleosomes but can bind to DNA which is wrapped around a nucleosome. We also investigated whether the top 20 TFs have particularly high or low information content and found that this is not the case (Figure A.2.20).

However, when we manually inspected the functional annotation of the top 20 TFs, we noticed that roughly half of these TFs are known to be involved in chromatin remodeling. Since, among our 158 TFs only 27 have been previously implicated in chromatin remodeling or nucleosome positioning, this amounts to a highly significant enrichment among our top 20 TFs (p-value 0.0016, see Materials and Methods). This suggested that the top 20 TFs may be characterized by interacting directly with chromatin modification machinery. To investigate this more systematically we investigated the occurrence of known direct protein-protein interactions between TFs and

1. Histones

2. Enzymes that modify histones

3. Proteins that are subunits of chromatin remodeling complexes

(see Materials and Methods). As detailed in Table 2.1, we find that our top 20 TFs are highly significantly enriched for direct protein-protein interactions with all 3 categories, showing the strongest enrichment for interacting directly with proteins in chromatin remodeling complexes. These results strongly suggest that our top 20 TFs are characterized by their ability to locally recruit chromatin modifiers.

Class	Total links	Links among top 20 TFs	p-value	Enrichment
Chromatin remodeler complexes	287	77	$9.2 * 10^{-11}$	3.26
Histone modification enzymes	369	74	$4.1 * 10^{-5}$	1.58
Histones	103	34	$7.3 * 10^{-8}$	2.6
All three classes	718	176	$4.1 * 10^{-18}$	1.94

Table 2.1: Statistical analysis of protein-protein interactions between TFs and chromatin remodeling complexes, histone modification enzymes, and histones.

The fact that only those TFs that interact directly with chromatin modifiers contribute significantly to explaining NFRs has interesting implications for the mechanisms of nucleosome positioning. It suggests that the creation of NFRs depends on the actions of chromatin modifiers whose activities lead to local expulsion of nucleosomes from the DNA. That is, the mechanistic picture that emerges is that, initially, the competition between TFs and nucleosomes for binding DNA, as implemented in our model, determines where TFs will end up binding DNA. Subsequently, in those places where TFs from the specific class that can recruit chromatin modifiers are bound, the recruitment of these modifiers will lead to local expulsion of the nucleosomes, leaving a larger region depleted of nucleosomes. This mechanistic picture also explains our previous observation that the predicted NFRs tend to be much narrower than those observed in the data.

2.3 Discussion

It is generally accepted that the packaging of DNA by nucleosomes in eukaryotes can modulate the accessibility of TFs to their cognate sites and thereby have major effects on gene regulation. In recent years there have been significant experimental efforts to determine nucleosome positioning patterns genome-wide, and to analyzing how these nucleosome-positioning patterns are established. As we discussed in the introduction, there has been a considerable debate as to whether nucleosome positioning in *Saccharomyces cerevisiae* is predominantly controlled by intrinsic sequence specificity of the nucleosomes, or that statistical positioning around barriers introduced by other DNA binding factors is more important for nucleosome positioning, and different researchers have presented seemingly contradictory results in this regard. We feel that these apparent contradictions may be reconciled by the results presented here.

The large majority of annotated nucleosomes and linkers genome-wide concern the phasing of short linkers within dense arrays of nucleosomes, mainly inside genes. We find that the positioning of these nucleosomes and short linkers crucially depends on the sequence specificity of the nucleosomes, and that TFs contribute relatively little to their positioning. Therefore, predicting all linkers and nucleosomes on a genome-wide scale, the sequence specificity of the nucleosomes provides the main contribution to explaining their positions. In contrast, we find that nucleosome specificity contributes little to explaining larger nucleosome free regions, especially those within promoter regions. As our modeling shows, NFRs in promoters are predominantly explained by the DNA binding of a specific class of $10-20$ transcription factors. Thus, while genome-wide locations of nucleosomes and short linkers are predominantly determined by nucleosome sequence-specificity, the large nucleosome free regions in promoters that likely contribute much more significantly to gene regulation, are determined mainly through the competitive binding of TFs. Importantly, the fact that competition with TFs can not help explain the *in vitro* nucleosome positioning shows that the contributions of the TFs is restricted to *in vivo* positioning. Thus, the competitive binding of TFs provides a quantitative and mechanistic explanation for the differences between *in vivo* and *in vitro* nucleosome occupancies.

That nucleosome free regions in promoters result from a competition between TF and nucleosome binding is supported by a number of recent studies of individual promoters, e.g. [9, 30, 53, 118]. In these studies the interplay of TF and nucleosome binding determines positions of NFRs and the resulting accessibility pattern has major consequences for gene expression. Our results suggest that this mechanism is not restricted to a few promoters, but is the typical situation genome-wide. Thus, whereas nucleosome sequence specificity does have a major impact on genome-wide nucleosome positioning, precisely those aspects of nucleosome positioning that have most impact on gene regulation are rather determined by the competition between nucleosomes and TF binding.

Another major result from our study is that less than 20 of the 158 TFs that we analyzed appear to have a significant effect on nucleosome positioning. As we have shown, these TFs are not characterized by particularly high expression or large numbers of binding sites in YPD, nor do they possess particular sequence specificities or DNA binding domains. Instead, our analysis suggests that these TFs engage in specific protein-protein interactions with chromatin remodelers, thereby effecting nucleosome eviction much more dramatically than other TFs.

Although the final predictions of our statistical mechanical model are quite competent, i.e. in promoters 96% of all nucleosomes and 70% of all NFRs are correctly identified, they are still far from perfect. This raises the question as to what additional elements are missing from the model. The main error the model makes is failing to identify roughly one third of nucleosome free regions as nucleosome free. This suggests that the model misses additional factors that promote displacement of nucleosomes. As most sequence-specific TFs in yeast are already represented in the model, and our results suggest that only a small fraction of these TFs significantly affect nucleosome positioning, it seems unlikely that the missing sequence-specific TFs play a major role in the overall quality of the results. In contrast, as shown in Figure A.2.15, general TFs including the RNA polymerase itself may play an important role in nucleosome positioning. In this context it has also been suggested [127] that the well-positioned nucleosome immediately downstream of TSS may result from a direct interaction between general transcription factors and the RNA polymerase with this nucleosome. Thus, including the recruitment and binding of general TFs and RNA polymerase will likely further improve the model.

In addition, TF binding can recruit chromatin modifying enzymes that displace nucleosomes and alter histone tails. The fact that experimentally observed NFRs are typically wider than the theoretically predicted ones suggest that the TF binding recruits chromatin modifiers which lead to a larger region of nucleosome exclusion than given by the TF binding itself. Thus, feed-back from TF binding to nucleosome modification and ejection as mediated by chromatin remodelers is a major feature that could improve the model's predictions. In summary, the picture that emerges from our study is that the binding of a specific class of $10-20$ TFs determines local recruitment of chromatin remodelers, which then mediate local expulsion of nucleosomes. The latter may further positively feed-back on TF binding and thereby expand and stabilize the nucleosome-free regions.

Although this work has focused on yeast, the competition between nucleosomes and TFs for binding DNA may even be more crucial for transcription regulation in higher eukaryotes. For example, in multi-cellular eukaryotes many gene regulatory elements occur in distal enhancers, i.e. local clusters of TF binding sites a few hundred base pairs in length, to which a combination of TFs binds to effect transcription at a promoter that can be hundreds of kilobases away. Recent mapping of enhancers based on chromatin marks has suggested that these enhancers are bound and activated in a highly tissue- and condition-specific manner [38, 117]. An attractive simplified model for such tissue-specific binding is that nucleosomes by default cause DNA to be inaccessible and that TF binding is too weak to access individual TF binding sites. Only in areas where a cluster with many binding sites for precisely that subset of TFs that is highly expressed in the condition will these TFs jointly outcompete the nucleosomes and create a region of DNA accessibility and TF binding, i.e. similar to the qualitative model presented in [71]. We believe that the statistical mechanics model that we have used here, might also be useful to quantitatively investigate such models of enhancer function.

2.4 Materials and Methods

2.4.1 A statistical mechanical model of competitive binding of proteins to the DNA

Based on a combination of ChIP-chip data, *in vitro* binding data, and computational analysis [8, 33, 101], we previously curated [19] a collection of 158 position specific weight matrices (WMs) representing the sequence-specificities of 158 *S. cerevisiae* TFs. We let $w_t(i, \alpha)$ denote the WM probability that position i in a binding site for TF t contains nucleotide α. Consequently, the probability that a binding site for TF t has sequence s is given by

$$P(s|t) = \prod_{i=1}^{l_t} w_t(i, s_i), \tag{2.2}$$

where l_t is the length of the WM for TF t and s_i is the nucleotide at position i in sequence segment s. For our statistical mechanical model we wish to determine energies $E(s|t)$ for the binding of sequence segment s to TF t. We make the standard assumption that the binding energy is a sum of individual contributions from different nucleotides in the site, i.e.

$$E(s|t) = E_t^c + \sum_{i=1}^{l_t} E_t(i, s_i), \tag{2.3}$$

where E_t^c is a sequence-independent contribution to the binding energy. Under this assumption, the sequence-specific energy components $E_t(i, \alpha)$ can be shown [14, 114] to be related to the WM components through

$$E_t(i, \alpha) = -\gamma_t \log[w_t(i, \alpha)], \tag{2.4}$$

where γ_t is a scale parameter, and the binding energy is expressed in units of $k_B T$.

There has been a significant amount of effort into modeling the sequence specificity of nucleosomes using data from both *in vivo* and *in vitro* experiments, e.g. [48, 55, 59, 95]. As shown in Figure 2.1A, different models of nucleosome sequence-specificity give

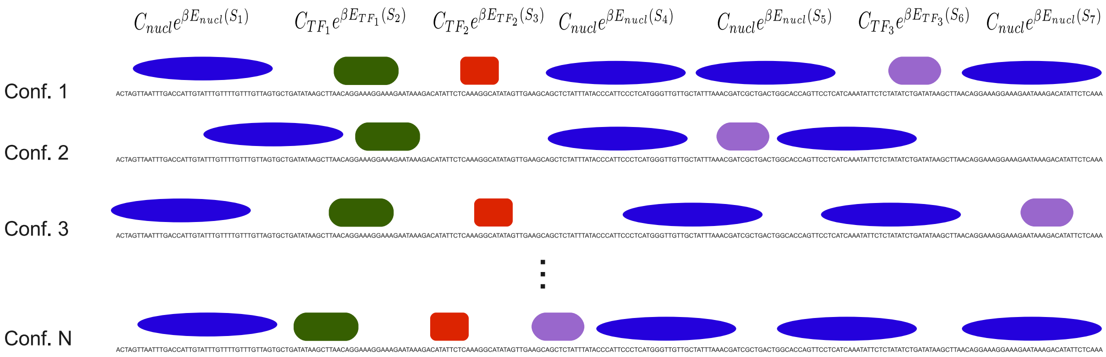
$C_{nucl}e^{\beta E_{nucl}(S_1)}$
$C_{TF_1}e^{\beta E_{TF_1}(S_2)}$
$C_{TF_2}e^{\beta E_{TF_2}(S_3)}$
$C_{nucl}e^{\beta E_{nucl}(S_4)}$
$C_{nucl}e^{\beta E_{nucl}(S_5)}$
$C_{TF_3}e^{\beta E_{TF_3}(S_6)}$
$C_{nucl}e^{\beta E_{nucl}(S_7)}$
Conf. 1
ACTAGTTAATTTGACCATTGTATTTGTTTTGTTTGTTAGTGCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAAGGCATATAGTTGAAGCAGCTCTATTTATACCCATTCCCTCATGGGTTGTTGCTATTTAAACGATCGCTGACTGGCACCAGTTCCTCATCAAATATTCTCTATATCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAA
Conf. 2
ACTAGTTAATTTGACCATTGTATTTGTTTTGTTTGTTAGTGCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAAGGCATATAGTTGAAGCAGCTCTATTTATACCCATTCCCTCATGGGTTGTTGCTATTTAAACGATCGCTGACTGGCACCAGTTCCTCATCAAATATTCTCTATATCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAA
Conf. 3
ACTAGTTAATTTGACCATTGTATTTGTTTTGTTTGTTAGTGCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAAGGCATATAGTTGAAGCAGCTCTATTTATACCCATTCCCTCATGGGTTGTTGCTATTTAAACGATCGCTGACTGGCACCAGTTCCTCATCAAATATTCTCTATATCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAA
Conf. N
ACTAGTTAATTTGACCATTGTATTTGTTTTGTTTGTTAGTGCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAAGGCATATAGTTGAAGCAGCTCTATTTATACCCATTCCCTCATGGGTTGTTGCTATTTAAACGATCGCTGACTGGCACCAGTTCCTCATCAAATATTCTCTATATCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAA

Figure 2.8: Illustration of example configurations of proteins bound to DNA. The top line indicates contributions from the individual binding sites to the overall probability of the configuration. Note that for illustration purposes, the sizes of TFs and nucleosomes are not shown to scale, e.g. the sizes of nucleosome footprints are much larger in reality.

predicted occupancies that are very highly correlated, and the model of [48] exhibits the most robustly high performance. We thus took the model of [48] as the basis for calculating binding energies $E(s|\text{nucl})$ of the nucleosome to each possible 147 bp stretch s. Specifically, the raw probability $P(s|\text{nucl})$ of a 147 bp long sequence segment s under Kaplan et al's model can be obtained using the "nucleosome_prediction.pl" script, that is provided by the authors on their website, with default parameters and using the option "raw_binding". Using this we define a binding energy under the Kaplan model as

$$E_{\text{kaplan}}(s) = -\log[P(s|\text{nucl})] + c, \tag{2.5}$$

In order to allow us to tune the sequence specificity of the nucleosomes, we introduce a similar scale parameter γ_{nucl} to obtain

$$E(s|\text{nucl}) = \gamma_{\text{nucl}} E_{\text{kaplan}}(s). \tag{2.6}$$

Note that, at $\gamma_{\text{nucl}} = 1$, the sequence-specificity of this model will be equal to that of Kaplan et al's model, whereas at $\gamma_{\text{nucl}} = 0$ nucleosomes will have no sequence preferences whatsoever. For notational simplicity, in the following we will consider the nucleosome as just another member of the set T of all DNA binding factors t.

Let C denote a (non-overlapping) configuration of TFs and nucleosomes bound to the genome and let S_t denote all segments in the genome where a binding site for factor t occurs. Using the standard Gibbs-Boltzmann approach, the probability of finding the cell in configuration C is given by

$$P(C|c,\gamma) = \frac{1}{Z} \prod_t \prod_{s \in S_t} c_t e^{-\beta E(s|t)}, \tag{2.7}$$

where c_t is the concentration of TF t, $\beta = 1/(k_B T)$ is the inverse temperature, and Z is the partition function

$$Z = \sum_C \prod_t \prod_{s \in S_t} c_t e^{-\beta E(s|t)}. \tag{2.8}$$

Note that the probability depends on the scale factors γ through the dependence of the binding energies $E(s|t)$ on the scale factors.

Note that, since we will be fitting the scale factors γ_t, we can define

$$\tilde{\gamma}_t = \beta \gamma_t \tag{2.9}$$

and fit the $\tilde{\gamma}_t$. For notational simplicity, we will drop the tilde and refer to these rescaled gammas as simply γ_t. Note that this is equivalent to measuring the energy in units of $k_B T$.

Using only information about known binding sites, i.e. the WM entries w^i_α, we cannot determine the sequence-independent contribution E^c_t for each TF, which essentially controls how generally 'sticky' the TF is to DNA. To allow the comparison of binding energies of different TFs on a common scale we set E^c_t such that, in the limit of low TF concentrations, each TF has equal binding to the yeast genome. Specifically, we set E^c_t such that the average $\langle e^{-E(s|t)} \rangle = 1$, when averaging over all sequence segments s in the genome.

Using this reparametrization the probability of a configuration becomes simply

$$P(C|c,\gamma) = \frac{1}{Z} \prod_t \prod_{s \in S_t} c_t e^{-\gamma_t E^c_t + \gamma_t \sum_i \log[w_t(i,s_i)]}. \tag{2.10}$$

Figure 2.8 shows a cartoon illustrating various configurations C and the factors contributing to their probabilities.

The partition function can be calculated efficiently using recursion relations variously known as transfer matrices or dynamic programming, and this has been routinely used in the field to sum over non-overlapping configurations of hypothesized binding sites, e.g. [18, 83, 87, 114, 119]. Let Z_n denote the partition sum for all configurations up to position n in a given chromosome. We then have

$$Z_n = Z_{n-1} + \sum_t Z_{n-l_t} c_t e^{-\gamma_t E_t^c + \gamma_t \sum_{i=1}^{l_t} \log[w_t(i, s_{n-l_t+i})]}. \tag{2.11}$$

Similarly, we can calculate the 'backward' partition sums B_n from position n to the end of the chromosome. Finally, the probability that a binding site for factor t covers positions $(n+1)$ through $(n+l_t)$ is given by

$$P(t, n|c, \gamma) = \frac{Z_n c_t e^{-\gamma_t E_t^c + \gamma_t \sum_{i=1}^{l_t} \log[w_t(i, s_{n+i})]} B_{n+l_t+1}}{Z_L}, \tag{2.12}$$

where L is the chromosome length. The occupancy of factor t to position n is then given by $O(t, n|c, \gamma) = \sum_{i=n-l_t+1}^{n} P(t, n|c, \gamma)$. Thus, given a set of scale factors γ and concentrations c, we can efficiently calculate the occupancies of all 158 TFs and the nucleosomes across the entire yeast genome.

2.4.2 Statistical positioning of nucleosomes

As first shown by Kornberg and Stryer [52], the repeating nucleosome pattern can be observed around a well-positioned barrier along the DNA which prevents binding of nucleosomes (Fig. 2.9**A**). The oscillating pattern of nucleosome occupancy occurs even assuming that nucleosomes lack any sequence specificity. The phenomena is usually called statistical positioning of nucleosomes. This phenomena arises from assumptions that nucleosome packaging is very tight ($\approx 80\%$ of the genome) and nucleosomes can't overlap. Since nucleosomes can't slide through the barrier, the configurations which contain nucleosomes just near the barrier are more likely. The positioned nucleosomes just near the barrier force positioning of all nucleosomes in the nucleosome array, since the tight packing restricts their lateral movement. Importantly, the positioning becomes weaker with the distance from the barrier. For example, when nucleosome coverage of the genome is 80% the effect disappears at distance corresponding to 4-5 nucleosomes (Fig. 2.9**A**).

Since transcription factors have to compete with nucleosomes for binding to DNA the statistical positioning phenomena have important effect on TF binding efficiency. In the model with a barrier and a binding site for a TF, the binding efficiency have oscillatory behaviour (Fig. 2.9**B**). In other words, at the same concentration of TF and the same binding energy of the binding site, the binding efficiency crucially depends on the position relative to the barrier.

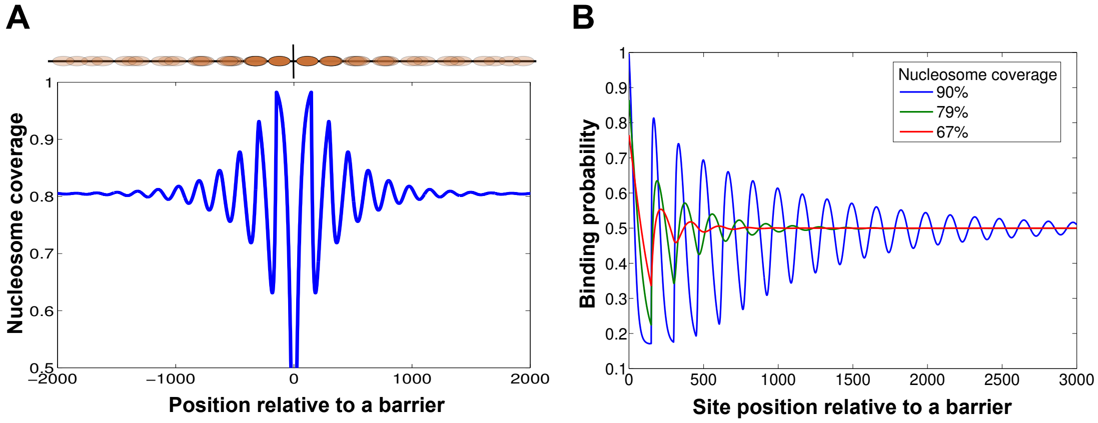

Figure 2.9: **A:** Statistical positioning of nucleosomes. A well-positioned barrier on the DNA leads to a periodic pattern of nucleosome positioning, even in the absence of sequence-specificity in nucleosome binding. **B:** Statistical positioning of nucleosome affects binding efficiency of TFs. TF binding to a single TFBS at different distances from a barrier shows a periodic pattern with period of roughly 1 nucleosome and amplitude increasing with overall nucleosome density.

Experimentally determined positions of nucleosomes and linkers

To compare the 'raw' occupancies as predicted by various models of nucleosome specificity and measured across several *in vivo* and *in vitro* experiments, we first downloaded the per base occupancy predictions provided by [48] and [59] and used these predicted occupancies directly. We also obtained raw data from the experiments [28, 48, 55, 69, 100]. To obtain per-base nucleosome occupancies we calculated, for the ChIP-seq data, the number of reads overlapping each position and log-transformed these read counts. For the ChIP-chip data we log-transformed the chip signal. We observed that there is a very small number of positions for which sometimes aberrantly high or low signals are reported. To avoid having these outliers skew the observed correlations we removed the 0.5% of genomic positions with highest signal and 0.5% with lowest signal. We then directly calculated Pearson correlation coefficients between all data-sets and all predictions.

For the *in vivo* data, we make use of the reference map of nucleosomes and linkers for *S. cerevisiae* growing in YPD that was constructed by combining 6 different experimental data-sets in [45]. We only retained nucleosomes that were observed in all 6 datasets and have occupancy bigger then 80% (according to the authors' annotation). This set contained 21′252 nucleosomes covering 26% of the *S. cerevisiae* genome, and covers approximately 90% of all annotated nucleosomes in [45]. Linkers were defined as regions lying in between segments that were annotated as nucleosomes in any of the 6 data-sets. This set contained 60′448 linkers covering 26% of the *S. cerevisiae* genome. As observed in [45] the distribution of linker lengths is bimodal and we separately considered 'short linkers' (less than 80 bps long) and 'nucleosome free regions' (longer than 80 bps) in our analysis. There were 45′981 short linkers and 14′467 nucleosome free regions, covering 9% and 17% of the genome, respectively. We also separately considered the quality of the predicted nucleosome positions in promoter regions, defined as running from 500 bps upstream to 500 bps downstream of the TSS for each gene. The TSS definitions, as well as the definitions of the 3' ends of genes, were taken from [74].

To assess the reproducibility of annotated nucleosome positions across the 6 experimental data-sets we calculated, for every nucleosome in the reference annotation, the standard-deviation in the positions of the associated annotated nucleosomes in each of the 6 data-sets. To compare the reproducibility of the annotated nucleosomes with what may be expected by chance, given the annotation procedure, we created randomized data-sets in which each sequencing read is mapped to a randomly chosen location in the genome. We then applied the same annotation procedure to this randomized data and calculated standard-deviations of the positions of annotated nucleosomes in the same way.

We constructed a reference map of *in vitro* nucleosome positioning using 3 independent data-sets from [48, 127, 128] using a procedure analogous to the one used in [45]. To annotate nucleosomes for every data-set we first run the GeneTrack software [1] using parameters $e = 294$ (width of the exclusion zone corresponding to configurations with non-overlapping nucleosomes), $s = 20$ (width of the smoothing gaussian kernel), $u = d = 73$ (half-width of the peak) and $F = 1$ (cut-off for peak height). The values of parameters e and u and d are dictated by the 147 bp width of the nucleosome footprint. Since the width $s = 20$ of the smoothing kernel is much smaller than the nucleosome width, the final nucleosome annotation is insensitive to the precise width of this kernel. Similarly, raising the cut-off F by 2-fold or 4-fold would only slightly reduce the number of called nucleosomes (i.e. 1% and 5% respectively) and not substantially affect the results presented in the paper. We use the annotated nucleosomes as input to GeneTrack (with the same settings), i.e. as if each annotated nucleosome were a read, to produce annotated reference nucleosomes. We retained the roughly 75% of annotated reference nucleosomes that occur in all 3 data-sets, leaving 18′867 reference nucleosome genome-wide. Reference linkers were defined as regions not covered by nucleosomes in any of the annotations. There were 30′824 such linkers genome-wide.

Assessing the match between predicted nucleosome coverage and experimental nucleosome positioning

To compare the experimentally annotated linker and nucleosome regions with the predicted nucleosome coverage we proceeded as follows. For a given set of parameters, i.e. concentrations c and scale parameters γ, we first calculate the median of the predicted nucleosome occupancy across each annotated linker and nucleosome region. Given a critical median occupancy level O_{crit}, we then classified each region as either 'nucleosome' n when its median occupancy was larger than O_{crit} and 'linker' l when its median occupancy was less than or equal to O_{crit}. We then determined the fraction of regions both predicted and annotated as nucleosome $P_{nn}(O_{\text{crit}})$, the fraction of regions predicted as nucleosome and annotated as linker $P_{nl}(O_{\text{crit}})$, the fraction of regions predicted as linker and annotated as nucleosome $P_{ln}(O_{\text{crit}})$, and the fraction both predicted and annotated as linkers $P_{ll}(O_{\text{crit}})$. Using these we determined the *mutual information* between the predictions and the annotations based on the experimental data:

$$I(O_{\text{crit}}, c, \gamma) = \sum_{i,j\in\{n,l\}} P_{ij}(O_{\text{crit}}) \log \left[\frac{P_{ij}(O_{\text{crit}})}{P_i(O_{\text{crit}})P_j^e} \right], \tag{2.13}$$

where $P_i(O_{\text{crit}})$ is the fraction of all regions predicted as i, P_j^e is the fraction of regions annotated as j, and we have explicitly indicated that this mutual information depends on the concentrations c and scale factors γ used in the predictions. We then define the mutual information $I(c, \gamma)$ as the maximal mutual information that can be obtained varying the critical occupancy O_{crit}, i.e.

$$I(c, \gamma) = \max_{O_{\text{crit}}} \left[I(O_{\text{crit}}, c, \gamma) \right]. \tag{2.14}$$

Finally, to normalize the mutual information on a more intuitive scale, we divide by the maximal possible mutual information, i.e. the entropy of the experimentally observed distribution:

$$H = -P_n^e \log[P_n^e] - P_l^e \log[P_l^e], \tag{2.15}$$

to obtain

$$F(c, \gamma) = \frac{I(c, \gamma)}{H}. \tag{2.16}$$

Thus, $F(c, \gamma)$ is the fraction of the information regarding nucleosome and linker positioning that is captured by the predictions, which we refer to as the *quality score.* We calculate the mutual informations I and quality score F in an entirely analogous manner when considering a particular subset of experimentally annotated nucleosomes and linkers, i.e. excluding short linkers and/or focusing only on promoter regions.

To obtain predicted nucleosome coverage distributions we simply calculate the predicted occupancy at each position in the genome as described above. To obtain nucleosome coverage distributions from different experimental data-sets we proceeded as follows. As has been observed previously [104], especially for ChIP-seq data-sets, the variance in read coverage along the genome is too large to be consistent with the known overall nucleosome coverage of roughly 80%. Consequently, a naive normalization in which one assumes read-coverage to be directly proportional to nucleosome occupancy would lead to unrealistically low overall nucleosome coverage. To address this, we normalize the data by rescaling log read-coverage, similar to the normalization procedure we developed previously for next-generation sequencing data [10].

Specifically, for ChIP-chip data (from a tiling array with 4 bp resolution) we obtain a signal x_i corresponding to the log-ratio of signal from the nucleosome and background sample for each probe i along the genome. Similarly, for ChIP-seq data we extend each read to length 147 bp and defined the 'signal' x_i at each genomic position i as the logarithm of the number of reads overlapping position i. We assume that the signal x_i is *proportional* to the logarithm of the probability P_i that a nucleosome is bound to the corresponding segment in the genome, i.e

$$x_i = \lambda \log(P_i) + c, \tag{2.17}$$

where λ and c are unknown constants. We determine c and λ by demanding that the *average* coverage probability matches the experimentally observed average nucleosome coverage of 0.8, and that all coverage probabilities P_i must lie in the interval $[0, 1]$. Finally, there is a small number of probes (0.1 percent of all probes) with an abnormally high signal x_i and we removed these outliers before fitting c and λ. As shown in Figure A.2.10, this procedure leads to highly similar coverage distributions for different data-sets.

Predicted average nucleosome coverage profiles around transcription starts and ends were obtained by simply averaging the predicted nucleosome coverage at different positions relative to TSS and transcription end over all genes. We similarly averaged the experimental coverage profiles relative to transcription starts and ends.

Model fitting

To optimize the concentration and specificity scaling parameters (c, γ) we used the Melder-Mead algorithm in combination with a simulated annealing algorithm that is implemented in the GNU Scientific Library (GSL). To avoid over-fitting when fitting different models with varying numbers of parameters we used a 80/20 cross-validation scheme for each model and data-set. That is, for each data-set and model, we randomly divide the data-set of annotated nucleosomes and linkers into 5 equally sized sub-sets. We then perform the parameter fitting 5 independent times, each time optimizing the parameters on 80% of the data and then evaluating the final quality score of the model on the 'test-set' containing the remaining 20% of the data. Whereever quality scores are shown we show the average quality score and its standard-error across the 5 test-sets.

For the *in vivo* reference set of nucleosomes and linkers, we first performed optimizations of the nucleosome-only model with different (fixed) values of the specificity scaling parameter γ_{nucl}, i.e. optimizing only the concentration c_{nucl}. For both the *in vivo* and *in vitro* reference sets we optimized the two-parameter nucleosome-only model (obtaining an optimal $\gamma_{\text{nucl}} = 0.47$ for the *in vivo* data, and $\gamma_{\text{nucl}} = 0.41$ for the *in vitro* data). After this we fixed the nucleosome specificity and concentration to their optimal values and, for the *in vivo* data, fitted the model with all TFs, fitting the concentrations and scale parameters for all TFs.

For the biophysical characterization of the fitted model, we first averaged the fitted concentrations c and scale parameters γ over the 5 training sets. We then calculated the predicted posterior binding probabilities $P(t, n|c, \gamma)$ for every factor t (i.e. the nucleosomes and all TFs) at every position n in the yeast genome. For each factor t, we then calculated the fraction of the genome f_t covered by this protein: $f_t = l_t \sum_n P(t, n|c, \gamma)/L_{genome}$, where l_t is the length of the footprint of protein t and L_{genome} is the length of the yeast genome. We also calculated the average binding energy $\langle E_t \rangle$ of the binding sites of each protein t, i.e. $\langle E_t \rangle = \sum_n E_{t,n} P(t, n|c, \gamma)/[\sum_n P(t, n|c, \gamma)]$, and its standard deviation $\sigma(E_t) = \sqrt{\langle E_t^2 \rangle - \langle E_t \rangle^2}$. Here $E_{t,n}$ is the binding energy of protein t at position n, measured in units $k_B T$. Finally, we calculated the average entropy H_t per binding site:

$$H_t = \frac{-\sum_n P(t, n|c, \gamma) \log_2[P(t, n|c, \gamma)] + (1 - P(t, n|c, \gamma)) \log_2[1 - P(t, n|c, \gamma)]}{\sum_n P(t, n|c, \gamma)}. \tag{2.18}$$

To calculate the information content for a TF t, as shown in Figure A.2.20, we used the standard formula

$$IC(t,\gamma_t) = \sum_{i=1}^{l_t} \sum_{\alpha \in \{A,C,G,T\}} \omega_t(i,\alpha) \log_2[\frac{\omega_t(i,\alpha)}{p_\alpha}], \tag{2.19}$$

where the $p_\alpha = 0.25$ are background probabilities (which we chose uniform) and the $\omega_t(i,\alpha)$ are the weight matrix entries. Note that, to incorporate the scaling parameter γ_t, the weight matrix entries are rescaled according to:

$$\omega_{\text{scaled}}(i,\alpha) = \frac{[\omega_{\text{unscaled}}(i,\alpha)]^{\gamma_t}}{\sum_{\alpha'} [\omega_{\text{unscaled}}(i,\alpha')]^{\gamma_t}}. \tag{2.20}$$

To assess the contribution of different TFs we fitted, for each TF, the model with nucleosomes and this single TF. For each TF we calculated, on each of the 5 test-sets, the difference dF between the quality score using only the nucleosome, and the quality score with the TF added, and determined the mean $\langle dF \rangle$ and standard error $SE = \sigma(dF)/\sqrt{5}$ over the 5 test-sets. We then ranked the TFs by the z-statistic $z = \langle dF \rangle / SE$. These fits and statistics were obtained separately for both the *in vivo* and the *in vitro* data. Finally, we also created a set of 158 randomized WMs by, for each WM, randomly shuffling the columns of the WM. Note that this randomization conserves both the sequence composition and the information scores of the WMs. We then performed the fitting with these 158 randomized WMs and obtained z-statistics in the precise same way.

For the *in vivo* data we then also fitted models including the top 5, 10, 20, and 30 TFs from the list ranked by their z-statistic, re-optimizing all parameters. Finally, to assess the contribution of the nucleosome specificity when TFs are added for the *in vivo* data, we fitted the model including all TFs, but without nucleosome sequence specificity, i.e. setting $\gamma_{\text{nucl}} = 0$.

Annotating chromatin related TFs

To annotate TFs with known roles in chromatin dynamics we used the Gene Ontology (GO) annotations available from the Saccharomyces cerevisiae genome database. We considered a TF ‘chromatin related’ when its GO annotation included any of the following categories:

- GO:0016568 chromatin modification.
- GO:0006338 chromatin remodeling.
- GO:0008301 DNA bending activity.

- GO:0031491 nucleosome binding.
- GO:0003682 chromatin binding.
- GO:0033698 Rpd3C(L) A histone deacetylase complex which deacetylates histones across gene coding regions.

Finally, we also added the TFs identified in [8] to this list. To calculate the over-representation of 'chromatin related' TFs among the top 20 TFs effecting nucleosome positioning, we performed a simple hypergeometric test.

Protein-protein interactions between TFs, histones, and chromatin remodelers

We first annotated yeast proteins that are either (1) part of chromatin remodeling complexes, (2) histone modification enzymes, or (3) histones themselves. Subunits of chromatin remodeler complexes were taken from [11, 103]. As subunits of histone modification enzymes we took genes that have GO annotation "covalent chromatin modification" and all children GO categories, i.e. histone methylation, acetylation etcera (108 genes in total). Information about protein-protein interactions were downloaded from the STRING database (http://www.string-db.org, file 'protein.links.detailed.v9.0.txt.gz'), using only experimental evidence with a cutoff of 400. After determining all known protein-protein interactions between the 158 TFs and the three classes of proteins (histones, histone modification enzymes, and subunits of chromatin remodeling complexes) we calculated enrichment of interactions between each class and the top 20 TFs that significantly explain nucleosome positioning. To assess the significance of the enrichment we used a simple hypergeometric test. The results are listed in Table 2.1.

2.5 Appendix

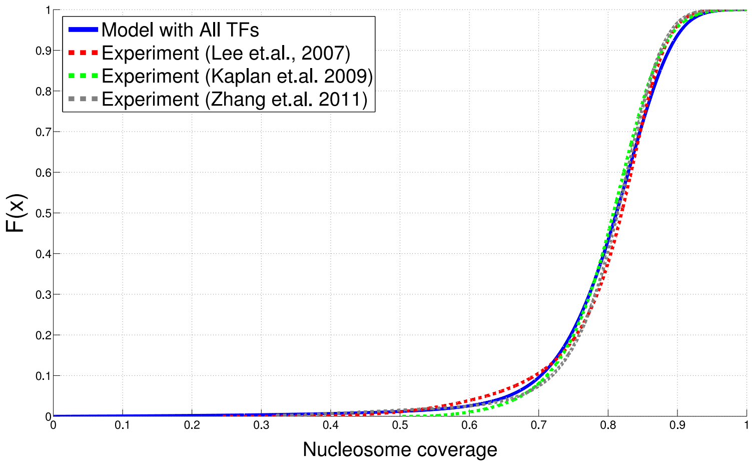

Figure A.2.10: Comparison of nucleosome occupancy distributions across three experimental data-sets and the model including all TFs. Cumulative distributions of nucleosome occupancies as measured in [55] (red dotted curve), as measured in [48] (green dotted curve), as measured in [128] (grey dotted curve), and as predicted by the model including all TFs (blue curve).

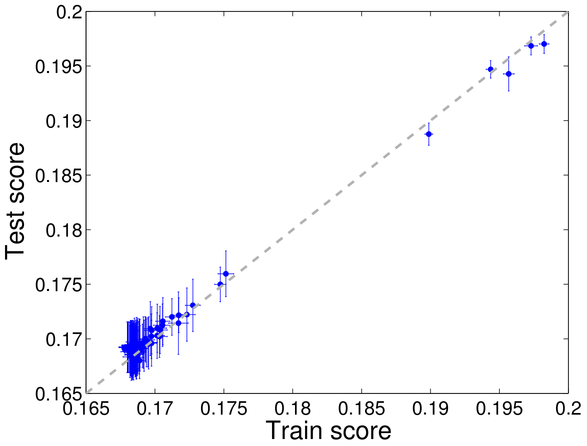

Figure A.2.11: Comparison of the model's training and test scores shows there is no over-fitting. For each of the 158 yeast TFs, we fitted a model containing a single TF plus nucleosomes to the *in vivo* reference map of nucleosomes and linkers genome-wide, optimizing the quality score F. We ranked all TFs by the z-statistic they attain and similarly fitted models that contain the nucleome plus the top 5, top 10, top 20, top 30, and all TFs. For each model we used 80/20 cross-validation, i.e. we fitted the model on 80% of the data and than evaluated it on the test-set of the remaining 20%. We performed this fitting 5 times and then calculated both the mean and standard-deviation of the quality score F obtained on both the training and test-sets. The figures shows a scatter of the quality scores on the training and test-sets for each of the models fitted. The error-bars denote the standard-error across 5 repeats. The figure shows that, although the test-set scores tend to vary more than the training set scores, the test scores are not consistently lower than the training scores, i.e. the model does not show any over-fitting.

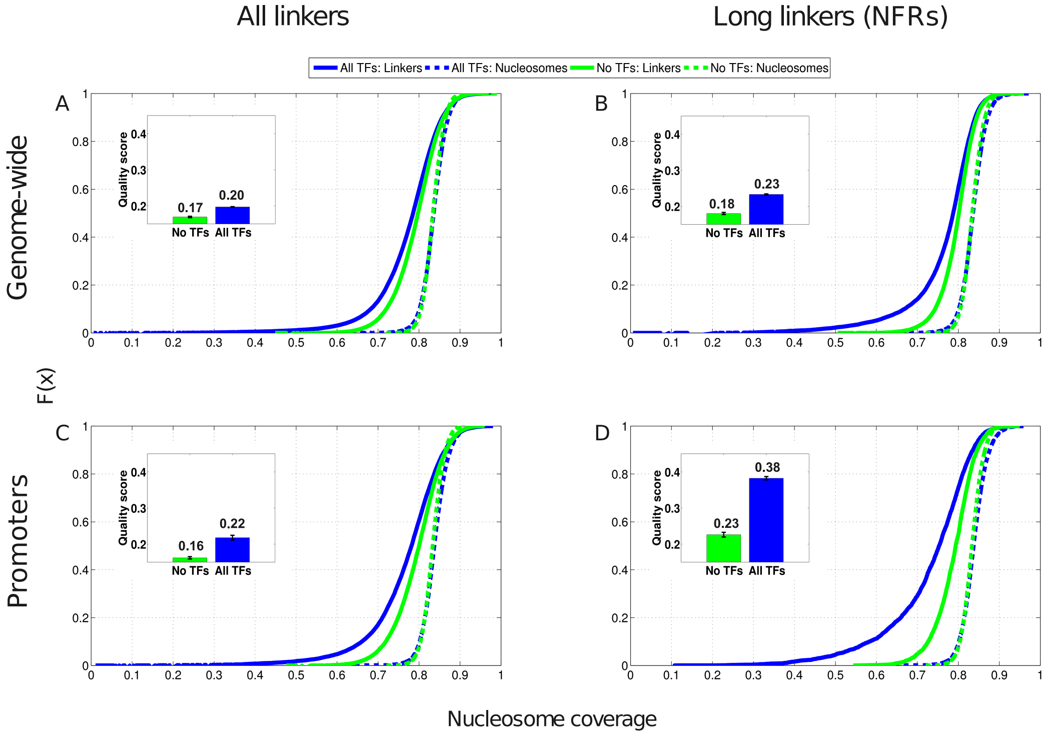

Figure A.2.12: Quality of the predicted nucleosome positioning profiles when including competition with TFs. The insets in each panel show the quality scores F of the model both including TFs (blue bar) and without TFs (green bar) in predicting annotated nucleosome and linker positions. The error-bars indicate the standard-error across 5 test sets. The curves in each panel show the cumulative distributions of predicted nucleosome coverage in annotated nucleosomes (dotted lines) and annotated linkers (solid lines) for the model using only nucleosomes (green) and the model including TFs (blue). **A**: Predicting all annotated linkers and nucleosome genome-wide. **B**: Predicting annotated nucleosomes and nucleosome free regions (long linkers) genome-wide. **C**: Predicting annotated nucleosomes and linkers in promoter regions. **D**: Predicting annotated nucleosomes and nucleosome free regions (long linkers) in promoter regions. Note that, for all 4 data-sets, inclusion of the TFs has very little effect on the coverage distribution observed at nucleosomes (i.e. annotated nucleosomes are generally predicted to be highly occupied) but that the TFs significantly lower the predicted coverage at annotated linkers, especially the long linkers in promoters.

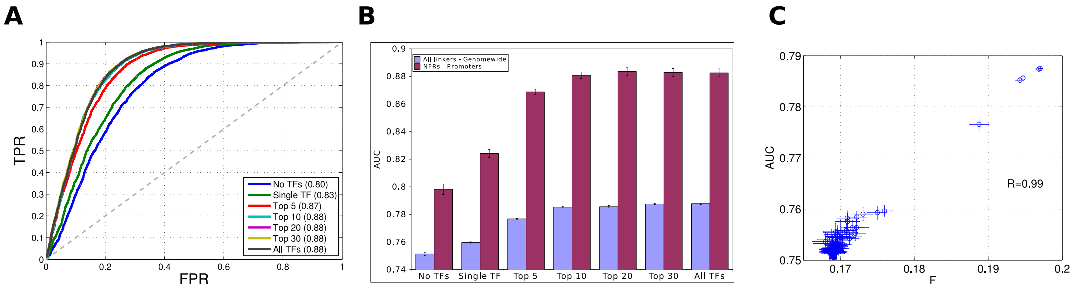

Figure A.2.13: Comparison of quality scores F with model assessment based on ROC curve analysis. **A:** For the models with no TFs, the top 1, 5, 10, 20, 30, and all TFs, we obtained a ROC curve for the classification accuracy of the model, i.e. by varying a cut-off in the predicted nucleosome coverage we calculated the rate of true-positive and true-negative prediction of nucleosomes/linkers. Similarly to the results obtained with the F quality score, the area under the curve (AUC) increases rapidly when the first few TFs are added and the performance saturates after $10 - 20$ TFs are added. **B:** Performance as measured by AUC for the models with increasing numbers of TFs, both for all linkers genome-wide (blue bars), as well as long linkers (NFRs) at promoters (red bars). Apart from a change in scale, the results look virtually identical to those obtained with the quality score F. **C:** A scatter of the quality score F against the AUC for all fitted models shows that the two measures of performance are very highly correlated.

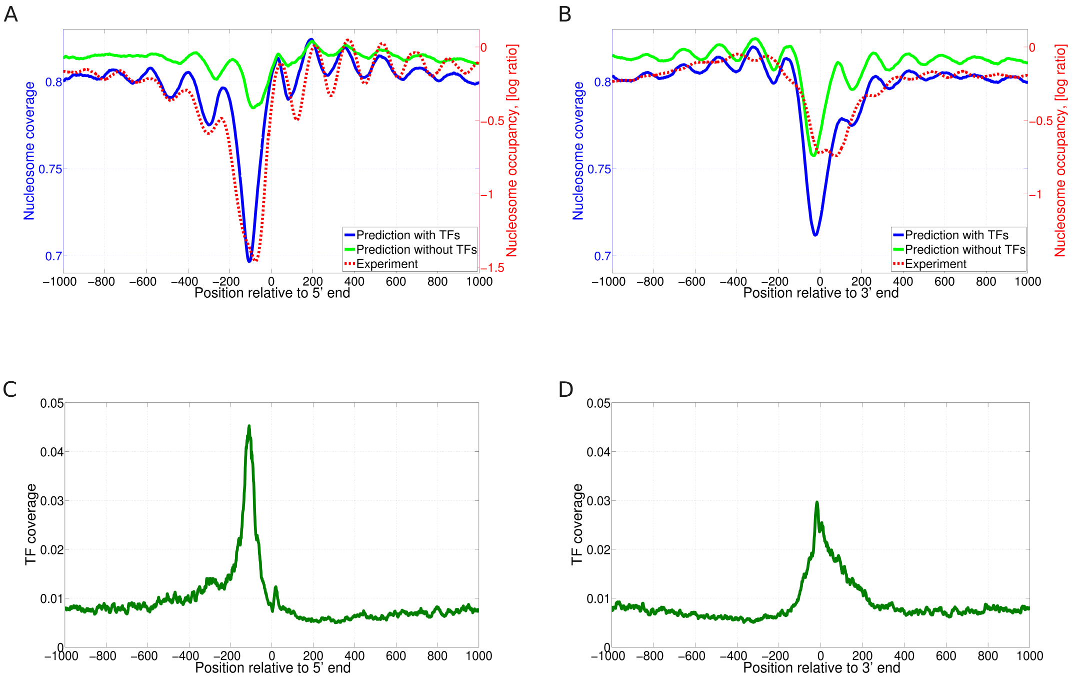

Figure A.2.14: Nucleosome and TF coverage profiles around starts and ends of genes. **A:** Averaged nucleosome coverage near the transcription starts. **B:** Average nucleosome coverage near the ends of genes. Each curve shows the average nucleosome coverage at different positions relative to transcription start or end averaged over all genes. Red dashed lines correspond to experimentally measured nucleosome coverage (data from [55], right vertical axis). The solid lines correspond to the predicted nucleosome coverage by the model including only nucleosomes (light green) and the model including all TFs (blue), left vertical axis. **C:** Averaged TF coverage (summed over all 158 TFs) relative to transcription start sites. **D:** Average TF coverage near transcription ends.

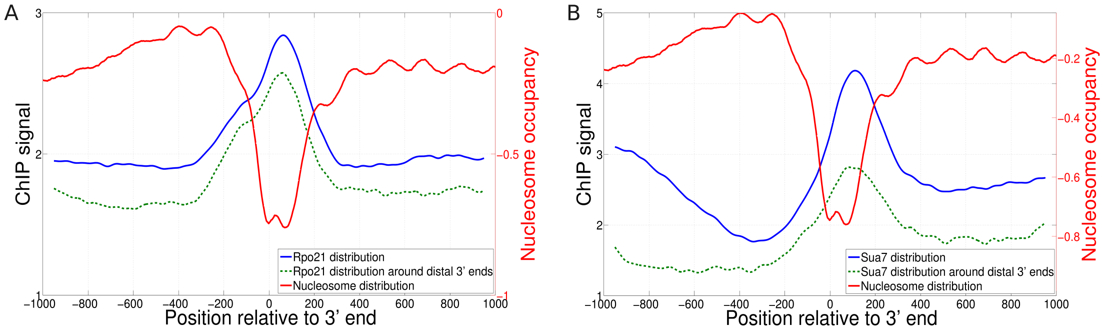

Figure A.2.15: Comparison of nucleosome coverage around ends of genes with binding profiles of RNA polymerase subunits. **A:** Average binding profile of the RNA polymerase II sub-unit Rpo21 around 3' ends of genes. **B:** Average binding profile of the general transcription factor Sua7 around 3' ends of genes. The blue curve corresponds to the average ChIP signal (log-ratio, left vertical axis) at each position from 1000 bps upstream to 1000 bps downstream of transcription end. The green dashed line shows the average ChIP signal when only genes whose ends are distal to the next transcription start are included. For reference, the red curves show the experimentally observed nucleosome coverage profiles (data from [55], right vertical axis). The results indicate that Rpo21 and Sua7 are observed to bind precisely in the region corresponding to the 3' nucleosome depleted region. The fact that the binding profiles look similar for 3' ends of genes that do not have a neighboring transcription start site nearby shows that the Rpo21 and Sua7 binding is not associated with a nearby promoter.

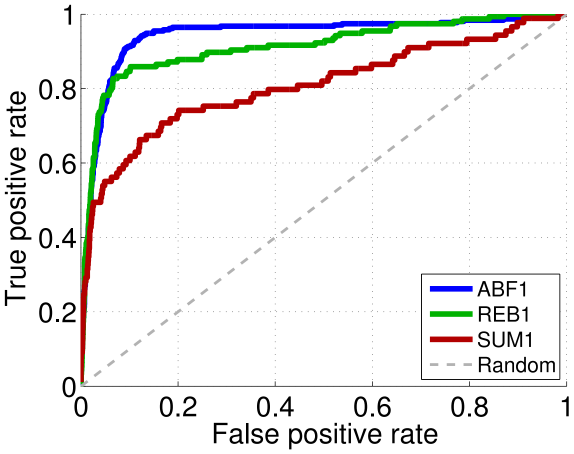

Figure A.2.16: Performance of the model with the nucleosome and all TFs in predicting the observed target promoters of the TFs Abf1, Reb1, and Sum1. The ChIP-chip binding data of [35] reports, for each TF, which promoter regions are bound by the factor and we used these as a reference set to compare with our predictions. For each promoter and each TF, we calculated a total 'target score' for the model by summing the predicted posterior probabilities of binding across all positions in the promoter. We then obtained ROC curves by varying a cut-off on this 'target score'. The figure shows the ROC curves of True positive and False positive rates obtained. Although our model was only optimized to fit observed nucleosome positioning, we see that it also accurately predicts the target promoters of these three TFs.

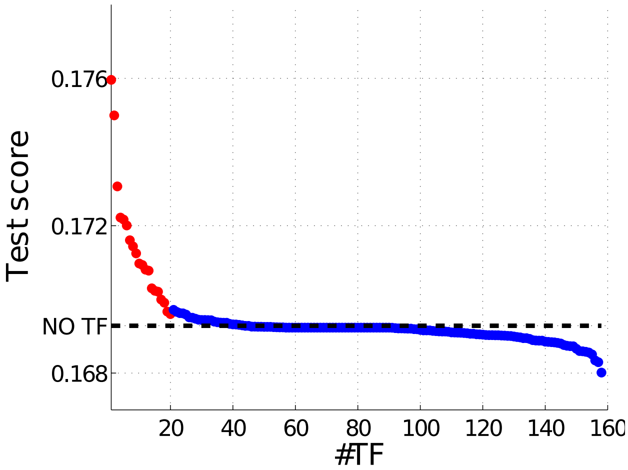

Figure A.2.17: Distribution of test scores for the models with nucleosomes and a single TF. For each TF we fitted the model including nucleosome specificity and the single TF on the training set and then determined the quality score (fraction of explained information) on the test set of annotated nucleosomes and linkers. We sorted TFs by their quality score and the figure shows the quality score as a function of TF number in this sorted list. For reference, the quality score obtained with the model without any TFs, i.e. nucleosome specificity only, is shown as a black dashed line. Note that the majority of TFs do not improve the quality score over the nucleosome-only model. The quality scores of the top 20 TFs are indicated in red.

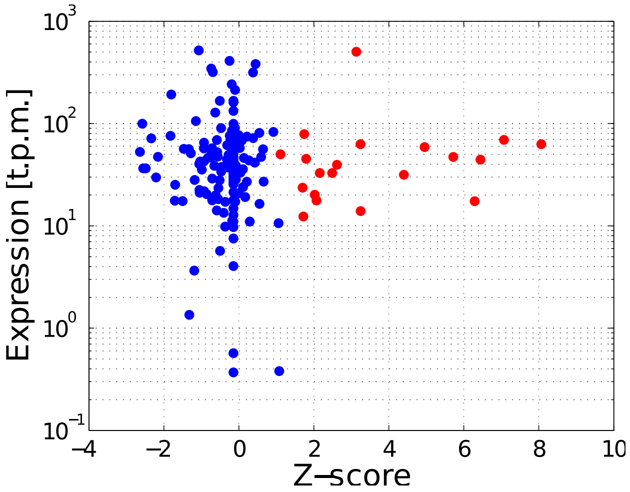

Figure A.2.18: Relation between TF significance for explaining nucleosome positioning and mRNA expression levels in YPD. For each TF a z-statistic was calculated (see Materials and Methods) that quantifies the extent to which the TF contributes to explaining nucleosome positioning genome wide. For each TF the z-statistic is shown on the horizontal axis against the TF's mRNA expression level in YPD expressed in tags per million (vertical axis, data from [58], note that the method smsDGE described in [58] does not require normalization by transcript length). Red dots correspond to the 20 TFs that most significantly contribute to nucleosome positioning. The figure shows that there is little correlation between expression level and the z-statistic.

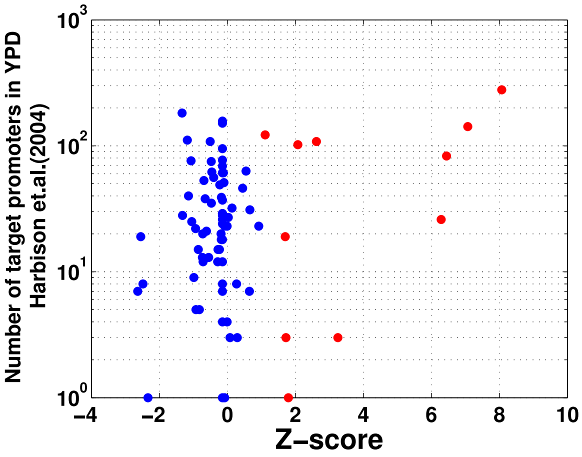

Figure A.2.19: Relation between the total number of promoters with binding in YPD and the significance in explaining nucleosome positioning of TFs. For each TF a z-statistic was calculated (see Materials and Methods) that quantifies the extent to which the TF contributes to explaining nucleosome positioning genome wide. For each TF the z-statistic is shown on the horizontal axis against the total number of promoters that have binding of the TF in YPD (p-value < 0.05) as measured by [35]. The top 20 TFs are indicated in red. There is no clear correlation between the total number of target promoters in YPD and the z-statistic.

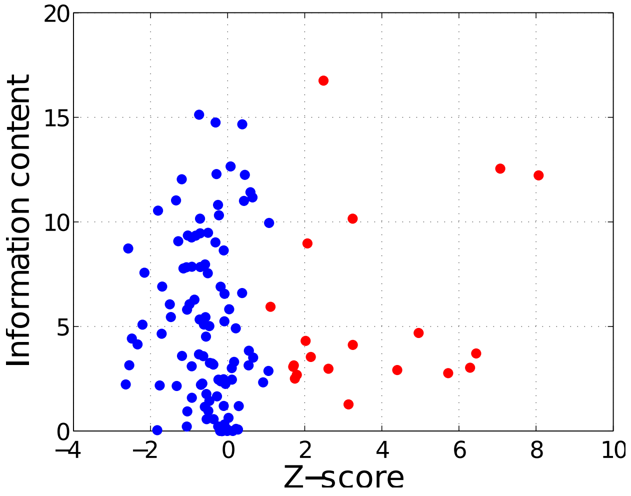

Figure A.2.20: Relation between the information content of each TF's binding motif and its significance in explaining nucleosome positioning. For each TF a z-statistic was calculated (see Materials and Methods) that quantifies the extent to which the TF contributes to explaining nucleosome positioning genome wide. For each TF the z-statistic is shown on the horizontal axis against the information content (in bits, vertical axis) of its binding motif (i.e. a position specific weight matrix). Note that the information content calculation takes into account the binding specificity factor γ_t that is fitted for each TF t. The top 20 most significant TFs are indicated in red. Note that there is no correlation between information content and z-statistic.

Chapter 3

Nucleosome mediated cooperativity between transcription factors

In chapter 2 we focused on the question of how transcription factors may affect nucleosome distribution and showed that competition between nucleosomes and transcription factors can significantly improve perfomance of the biophysical model by explaining nucleosome free regions in promoters of genes.

This chapter is devoted to the question of how competition with histones affect binding of transcription factors.

Biophysical modeling predicts that competition between nucleosomes and transcription factors (TF) for binding to nearby sites on the genome can induce both positive and negative cooperativity in TF binding. In particular, we show that the cooperative effect depends periodically on the distance between transcription factor binding sites (TFBSs), with positive cooperativity for sites less than 40 bp apart, negative cooperativity for larger distances up to one nucleosome length, and again positive cooperativity for distances just above one nucleosome length.

A comprehensive statistical analysis of TFBS positioning for 158 TFs of *Saccharomyces cerevisiae* and for 189 TFs of *Mus musculus* and *Homo sapiens* shows that many pairs of TFs have positioned their binding sites so as to optimize positive cooperativity of their binding. Moreover, this positioning is most significant for a number of TFs that have already been implicated in opening chromatin. In summary, our results show that the “grammar“ of the regulatory code in yeast promoters is shaped to a significant extent by nucleosome-mediated cooperativity of TFs.

3.1 Introduction

Binding of TFs to regulatory regions is a key event in regulation of transcription activity of genes. Most TFs recognize their cognate binding sites in sequence specific manner. Recent analysis of binding motifs reveals striking differences of gene regulation in prokaryots and eukaryots [122]. The analysis of binding motifs shows that, in contrast to prokaryots, the information content of TF's binding motifs in eukaryots is not high enough to recognize binding sites with sufficient specificity. In other words, whereas a typical binding motif in prokaryots provides enough information to recognize a unique and functional binding site, the eukaryotic TF's binding motifs are not so specific to distinguish functional binding sites from background DNA. Indeed, most of potential binding sites for a TFs in eukaryots are unoccupied and nucleosome occluded. Nevertheless, it has been suggested that the paradox of information deficiency of binding motifs in eukaryots can be resolved by organizing binding sites in clusters. For instances enhancers and promoters are natural examples of such clusters of TFBSs.

Due to development of experimental techniques, such as microarray and next-generation sequencing, it is now possible to map the occupancy of TFs across a huge number of cell types during development and differentiation in variety of organisms. The analysis of these maps revealed that regulatory regions are activated in highly cell type specific and time dependent manner[23, 38, 117, 123]. The mechanism of how the cell type specific activation of regulatory regions is achieved in multicellular organisms is still unclear. The "histone code" could serve as an explanation of this phenomena [44]. It has been suggested that the epigenetic histone marks can serve as code which define the local chromatin compaction in the genome, and therefore the accessibility of DNA to TFs. Indeed, the presence of certain types of posttranslational histone modifications is highly correlated with DNA accessibility, activity of regulatory modules and transcription rate of nearby genes [7, 56, 108]. However, the mechanisms by which the chromatin marks are targeted to specific genomic locations are not yet clear. It has been observed that some sequence specific TFs can recruit chromatin modifying complexes which in turn induce local chromatin modification and reorganization around TF binding sites [13, 75]. It is likely that there are multiple feedback loops between TFs binding, activity of chromatin modifying complexes and activity of ATP-dependent remodeling complexes. Thus, the hypothesis that TFs play a crucial role in targeting chromatin marks once again raises the question about what mechanisms control binding of TF to their cognate binding sites.

It has been suggested that mechanism of nucleosome-mediated cooperativity may be the ground of so called “combinatorial gene regulation“, where the combination of binding sites, their local arrangement relative to each other, TF's concentrations and local nucleosome occupancy determine activity of a regulatory module [71]. In other words, whether or not a certain binding site is occupied by a TF depends not solely on a sequence of the binding site and TF concentration but also on the local context of nearby TFBSs and nucleosomes.

The comprehensive biophysical model considered in chapter 2.4.1 and in other studies (e.g. [73, 88]) takes into account nucleosome sequence specificity along with competition between nucleosomes and TFs. It predicts that competition between nucleosomes and TFs for binding to nearby sites on the genome can induce both positive and negative cooperativity between TFs. The theoretical results suggest that distance between adjacent binding sites is the major determinant of the cooperativity effect, and the strength of the cooperativity has oscillatory behavior with a period of roughly one nucleosome length.

In this work we study how architecture of a cluster of TFBSs affect binding of TFs. Considering the simplest model of two binding sites next to each other we summarize previously known theoretical results [88], namely 1) spacing between binding sites is the major factor which determines the cooperativity effect, 2) the cooperativity of TFs with coordinated binding activity can reduce noise in TF binding. However, the main result of this is study is investigation of spacing between binding sites in real genomes. For the first time we show that some TFs in yeast, mouse and human have positioned their binding sites to account for nucleosome mediated cooperativity. Remarkably, this positioning is most significant for a number of TFs that have already been implicated in opening chromatin. Our results support the idea of a special class of TFs, often reffered as "pioneer" TFs, which bind to DNA prior to transcription initiation and help in organizing local chromatin configuration.

3.2 Results

3.2.1 Nucleosome mediated cooperativity between transcription factors

The biophysical model described in Chapter 2.4.1 allows us to rigorously calculate the probability of binding as a function of concentration and binding energy of TFs. To investigate effects induced by competition between nucleosomes and TFs we consider the most trivial architecture of a cluster of TFBSs. The toy promoter comprises only two binding sites for two TFs , and nucleosomes without sequence specificity (see the inset in the Fig. 3.1B).

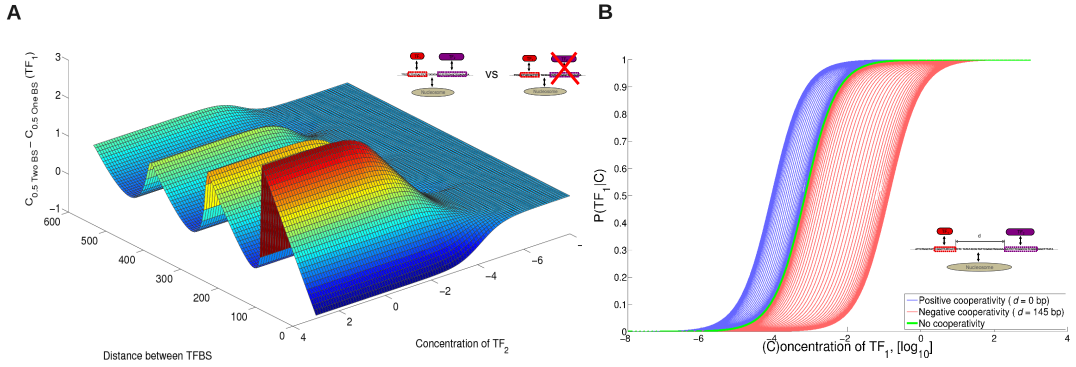

Figure 3.1: The cooperativity effect depends periodically on the spacing between TFBS. **A:** The cooperative effect as a function of concentration of TF_2 and distance between binding sites. The axes X and Y represent the distance between binding sites and concentration of TF_2 correspondigly. The Z-axis represents the difference $\delta C_{0.5} = C_{0.5TwoBS} - C_{0.5OneBS}$ for TF_1, namely how the $C_{0.5}$ for TF_1 changes after addition of TF_2 into the system (see the inset). **B:** The inset illustrates the toy model that we investigate, namely a cluster of two TFBS (promoter) and two TFs which have to compete with nucleosomes that have no sequence specificity. The curves at the plot illustrate binding curves for different promoter architectures. The green curve is a binding curve for TF_1 without cooperativity effect, i.e. at promoter without a binding site for TF_2 (No cooperativity), set of blue curves are binding curves for TF_1 at promoter with positively cooperating TFs at different concentrations of TF_2 (Positive cooperativity ($d = 0$ bp)), set of red curves are binding curves for TF_1 at promoter with negatively cooperating TFs at different concentrations of TF_2 (Negative cooperativity ($d = 145$ bp)).

To address the question of how TFBSs in a cluster affect binding of each other we study the binding curve, i.e. binding probability as a function of TF concentration $P(“on“|C)$ (Fig. 3.8A). The first important characteristics of the binding curve is $C_{0.5}$, i.e. the concentrations of 0.5 probability level ($P(“on“|C_{0.5}) = 0.5$). In other words the $C_{0.5}$ is the point when a binding site changes its state. It is also important to note that the $C_{0.5}$ is proportional to the effective affinity of a binding site (see Materials and Methods).

We first tested how the $C_{0.5}$ of the binding curve depends on the promoter architecture. For different distances d between binding sites (Fig. 3.1B) we sweep concentrations of TFs and calculate $C_{0.5}$ for one of the binding site. We found that the $C_{0.5}$ depends periodically on the distance between binding sites (Fig. 3.1A).

For distances between binding sites up to 40 bp TFs help each other to outcompete nucleosomes (positive cooperativity), for distances larger then 40 bp and up to one nucleosome length (147 bp) TFs hinder each other, and for distances larger then 147 bp up to roughly 210 bp again help each others binding.

The mechanism by which TFs cooperate or interfere is clear. Since our toy model assumes that nucleosomes do not have sequence specificity they are positioned solely by statistical positioning effect arisen by TFs bound to their binding sites. As was described in 2.4.2 the nucleosome occupancy profile for nucleosomes has periodic pattern relative to a barrier. Therefore, the promoter architectures for which the statistical positioning effect reduces average nucleosome occupancy of TFBSs are more favorable for TFs binding, due to higher effective affinity of TFBSs (see Fig. 2.9B). In other words, we observe effect of positive cooperativity when binding to nearby sites reduces nucleosome occupancy of TFBSs by statistical positioning effect and, therefore, increases effective affinity of TFBSs. And opposite, we observe effect of negative cooperativity when binding to nearby sites increases nucleosome occupancy of TFBSs and decreases effective affinity of TFBSs.

3.2.2 Noise minimization by cooperativity between transcription factors

Recently, cell-to-cell variability, often reffered as noise, in a clonal population has received great attention, e.g. [16, 26]. It has been suggested that noise in gene expression may be an evolvable trait that can be optimized [86]. One important source of noise in gene expression is the stochastic nature of gene regulation. Indeed, events of TFs binding and assembly of preinitiation complex (PIC) are purely probabilistic and depend on concentrations of TF and local chromatin configuration of a promoter.

Intuitively, the high level of noise in TF binding is associated with intermediate state of a binding site, i.e. when the most heterogeneity in the cell population with respect to a binding site is observed. Indeed, the very low or very high probability of TF binding should correspond to low level of noise as most cells in the population have this binding site free or occupied. And opposite, the intermediate state of the binding site, namely when the probability is about 0.5, leads to the largest heterogeneity in the cell population, since about half of the population have the binding site free and the other half have it occupied. Thus, the average noise can be thought as a measure of how likely it is to find the TFBS in its uncertain state during TF activation. Therefore, the noise in TF binding is reflected in the steepness of the binding curve, namely the steep binding curve implies low level of average noise in TF binding since the transition between “on“/“off“ state happens faster.

We quantify the noise level in TF binding as

$$Noise = \int_{-\infty}^{\infty} P(“on“|C) \cdot (1 - P(“on“|C))dC \qquad (3.1)$$

which reflects the steepness of the binding curve (see Matherials and Methods).

It is noteworthy that the noise for the simplest system with only one binding site in promoter always equals 1 and does not depend on the affinity of a binding site (see Materials and Methods). Therefore, the affinity of the binding site affects $C_{0.5}$ of the binding curve, i.e. shifts the binding curve left or right along the concentration axis (Fig. 3.8A), but the noise of the binding curve remains constant and does not depend on the affinity of the binding site (see Matherials and Methods).

However, when TFBSs are organized in a cluster and TFs have to compete with nucleosomes the steepness of the binding curve depends on the arrangement of TFBSs in the cluster. For instance, in Fig. 3.2A the noise for two identical TFBSs for the same TF periodically depends on the spacing between TFBSs. The case of two identical binding sites for the same TF can be thought as an example of perfectly coexpressed TFs, namely when concentrations of TF_1 and TF_2 are linearly dependent. In this case positive cooperativity can reduce noise and negative cooperativity can induce noise (Fig 3.2B).

We next investigated how concentrations, or binding activities, of TFs should be coordinated to minimize noise by positive or negative cooperativity between TFs.

The Fig. 3.3A illustrates the noise curves for positively cooperating TFs. In principal, for certain concentration of TF_1 there is a range of noise levels that can be achieved by modulating concentration of TF_2 (blue area in Fig. 3.3A). Therefore, possible scenario which could lead to noise reduction by positive cooperativity would be rapid increase of the binding activity of TF_2 when the binding site for TF_1 in its intermediate state (Fig. 3.3C). This scenario could lead to a rapid transition of the binding site from “off“ to “on“ state. In principle, binding activity of TFs can be changed in two ways, either by changing concentration or by changing binding energy. If the fast increase of the TF concentration seems unrealistic, the fast increase of binding energy seems possible, for instance by changing phosphorylation state of a TF.

In case of negative cooperativity, the noise reduction can be even more dramatic (Fig. 3.4). In this case TF_2 could play a role of a trigger, that keeps the binding site for TF_1 occluded by nucleosomes until the concentration of TF_1 is high enough to rapidly move

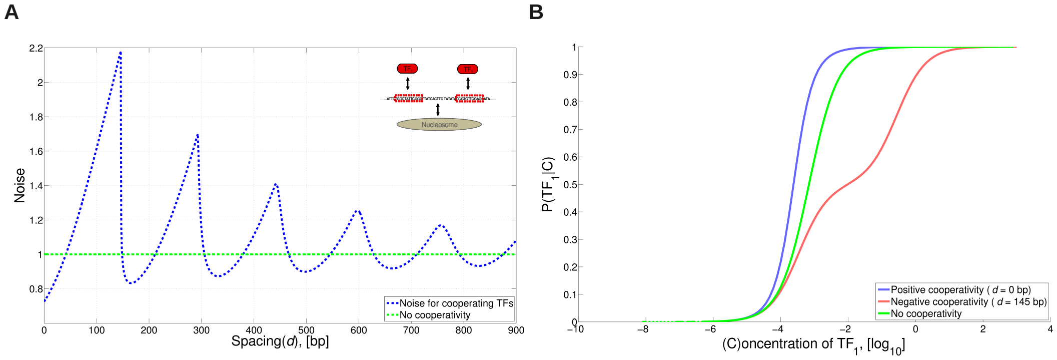

Figure 3.2: Cooperativity between TFs can reduce noise. **A:** Noise as a function of spacing between two identical binding sites for the same TF (see the inset). **B:** Binding curves for positively (blue curve, spacing(d)=0 bp) and negatively (red curve, spacing(d)=145 bp) cooperating TFs, and non-cooperating TFs (green curve).

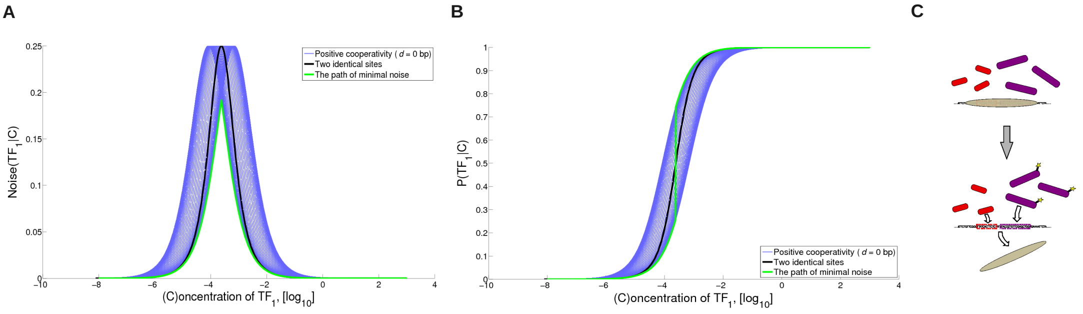

Figure 3.3: Noise reduction by positive cooperativity between TFs. **A** and **B** illustrate noise curves $Noise(TF_1|C_1)$ and binding curves $P(TF_1|C_1)$ for promoter architecture with binding sites for TF_1 and TF_2 next to each other (blue curves, Positive cooperativity (d=0 bp)). The green curves correspond to the path of minimal average noise in binding of TF_1. The black curves correspond to promoter architecture where only two identical binding sites for TF_1 are present (the case of linear coexpression, see Fig. 3.2**A**). **C:** Possible scenario for noise minimization by positive cooperativity. The path of minimal noise in **A** and **B** corresponds to scenario where TF_2 acquires binding activity at the point when concentration of TF_1 is high enough to evict a nucleosome together with TF_2.

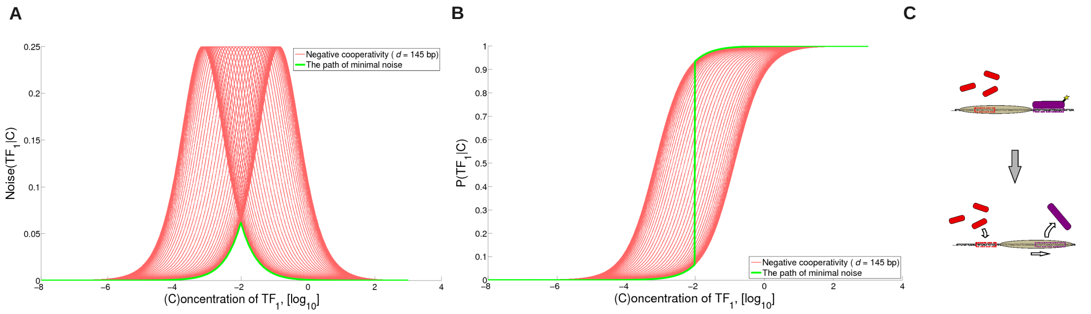

Figure 3.4: Noise reduction by negative cooperativity between TFs. **A** and **B** illustrate noise curves $Noise(TF_1|C_1)$ and binding curves $P(TF_1|C_1)$ for promoter architecture with binding sites for TF_1 and TF_2 spaced by 145 bp (red curves, Negative cooperativity (d=145 bp)). The green curves correspond to the path of minimal average noise in binding of TF_1. **C:** Possible scenario for noise minimization by negative cooperativity. The path of minimal noise in **A** and **B** corresponds to scenario where TF_2 loses its binding activity at the point when concentration of TF_1 is high enough to displace a nucleosome aside.

the nucleosome aside.

In summary, the theoretical analysis of plausible scenarios that could reduce noise in TF binding make us hypothesize that in a cluster of TFBSs for TFs with well coordinated binding activity the noise can be substantially reduced by positive or negative nucleosome mediated cooperativity between TFs. The required coordination of binding activity of TFs could be achieved by regulation of phosphorylation state of TFs.

3.2.3 Spacing between binding sites is biased so as to optimize positive cooperativity of binding

Theoreticaly, the nucleosome mediated cooperativity between TFs has substantial effect of the effective affinity of binding sites and on average noise in TF binding. One would expect that this effects may cause biases in the arrangement of TFBSs in the real genomes.

Next we investigated whether TFBSs in real genomes are positioned relative to each other so as to account for cooperativity between TFs. Using computational method for TFBS prediction which takes into account sequence specificities (using position-specific weight matrices (WM)) along with evolutionary conservation [6] we predicted binding sites for 158 WMs in *S.cerevisiae* genome, and for 189 WMs in *M.musculus* and *H.sapiens* genomes. Given the TFBS prediction we investigated the distributions of spacing between binding sites for every possible pair $TF_i : TF_j$.

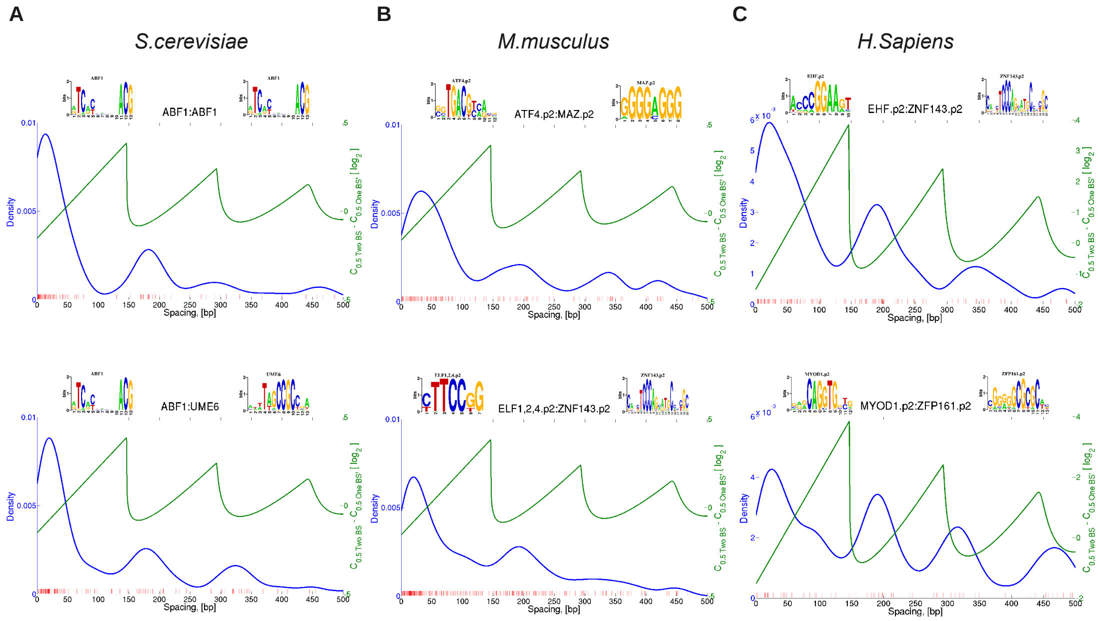

Figure 3.5: Spacing between TFBSs is biased so as to optimize positive cooperativity of binding. Examples of TF pairs with optimized spacing distribution for positive cooperativity in yeast (**A**), mouse (**B**) and human (**C**).

Remarkably, we find that spacing distribution for some TFs pairs has periodic pattern with peaks falling precisely to the areas of positive cooperativity between TFs (Fig. 3.5). It suggests that spacing between binding sites of those TFs have evolved so as to account for positive cooperativity of their binding.

To disentangle real spacing biases between TFBSs from possible biases due to TFBS positioning with respect to transcription start sites (TSS) we carried out comprehensive statistical analysis where we randomly shuffle TFBSs keeping position relative to TSSs constant. If a spacing distribution (Fig. 3.5) was due to TFBS positioning relative to TSS the biases would not disappear after the shuffling. For every pair $TF_i : TF_j$ we calculate Z statistic

$$Z(TF_i : TF_j) = \frac{< \delta C_{0.5}^{actual} > - \mu(< \delta C_{0.5}^{random} >)}{\sigma(< \delta C_{0.5}^{random} >)} \tag{3.2}$$

where $< \delta C_{0.5}^{actual} >$ is expected value of $\delta C_{0.5} = C_{0.5TwoBS} - C_{0.5OneBS}$ for actual spacing distribution (product of blue and green curves in Fig. 3.5), $\mu(< \delta C_{0.5}^{random} >)$ and $\sigma(< \delta C_{0.5}^{random} >)$ are mean and standard deviation of $< \delta C_{0.5}^{random} >$ distribution obtained by random permutations (see Materials and Methods). Negative Z statistic implies that actual spacing distribution is optimized for positive cooperativity, i.e. actual $\delta C_{0.5}$ is significantly lower than expected by chance.

Statistical analysis shows that many pairs of TFs have positioned their binding sites to take positive cooperativity between TFs into account, namely have significantly negative Z scores (Suppl. Tables A.3.1 and A.3.2).

Strikingly, pairs of TFs with optimized spacing distribution, i.e. with significantly negative Z scores, seem to be conserved between human and mouse (Fig. 3.6).

Next, we addressed a question of what distinguishes TFs with optimized spacing from others. It turns out that for *S.cerevisiae* the TFs pairs with optimized spacing between TFBSs are enriched for TFs that have already been implicated in altering chromatin ($p = 3 \cdot 10^{-5}$, Suppl. Table A.3.1). It suggests that TFs that affect local chromatin state tend to have spacing distribution optimized for positive nucleosome mediated cooperativity.

The hypothesis that TF pairs with optimized spacing can affect local chromatin architecture is further supported by analysis of nucleosome patterns around TFBSs spaced for positive cooperativity (nucleosome data taken from [55]). Indeed, the experimental nucleosome profiles in *S.cerevisiae* (Suppl. Fig. A.3.9) show that positively spaced binding sites of some TFs fall into NFRs. Interestingly, the nucleosome profiles for binding sites with spacing from 150 to 211 bp reveal a nucleosome which is trapped between two binding sites (Suppl. Fig. A.3.9 B).

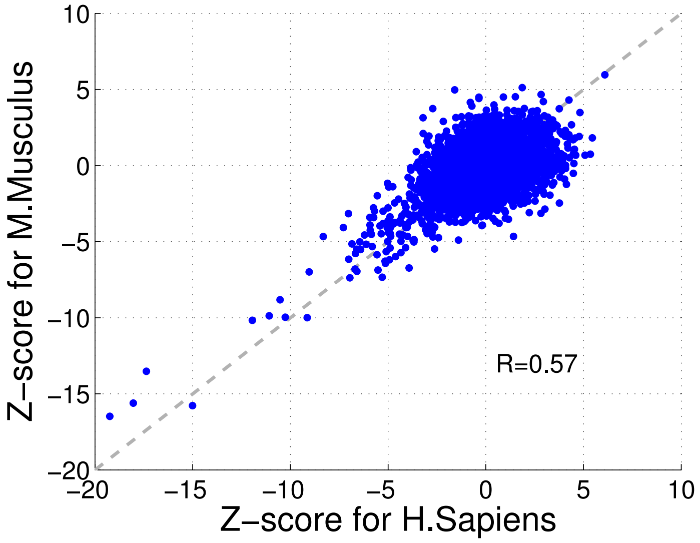

Figure 3.6: Optimized spacing distribution is conserved between *H.Sapiens* and *M.Musculus*. There are 27 TF pairs with Z score less then -5 (see Suppl. Table A.3.2) and only one pair with Z score bigger then 5,(*NRF1* : *TLX2*, $Z = 6$ both in human and mouse)

In summary, we show that binding sites for some TFs are positioned in a way to account for nucleosome mediated cooperativity between TFs.

3.3 Discussion

Promoter architecture plays important role in regulation of expression [110], in particular, chromatin architecture at promoter crucially affects important characteristics of gene regulation, such as expression rate, expression noise and transcriptional placticity [85, 109]. Even though, it has been known about the cooperative mode of TFs binding which do not include protein-protein interaction [115], the study presented here helps in mechanistic understanding of how arrangement of binding sites can affect gene regulation.

The biophysical modeling presented in this study shows that architecture of a simplest TFBS cluster affects how well TFs compete with nucleosomes. The competition between TFs and nucleosomes changes both efficiency and noise of TFs binding, and the distance between binding sites is the main factor that defines the cooperative effect.

The analysis of spacing between transcription factor binding sites in the real genomes supports the idea that nucleosome mediated cooperativity between transcription factor

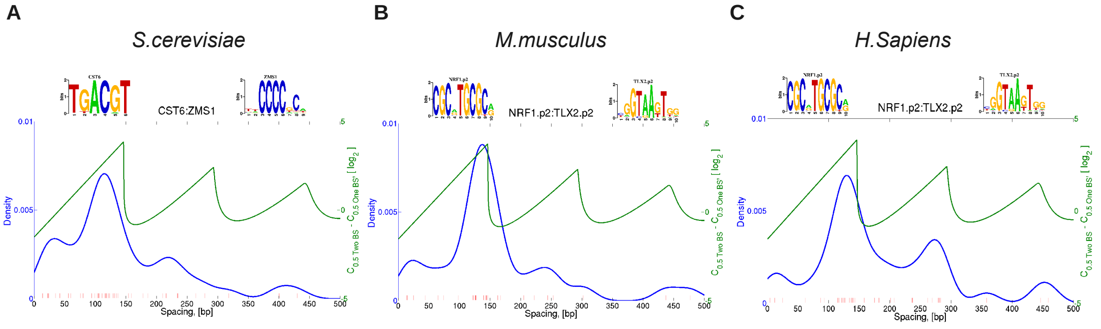

Figure 3.7: Examples of TF pairs with optimized spacing distribution for negative cooperativity in yeast (**A**), mouse (**B**) and human (**C**).

may really happen in the cell. The remarkable fact that peaks in spacing distribution of some TF pairs coincide with spacing constraints for positive cooperativity predicted by the model suggests that the "grammar" of the regulatory code in eukaryots may be indeed shaped to account for the nucleosome mediated cooperativity (Fig. 3.5). Interestingly, the spacing biases for positive cooperativity is most significant for transcription factors that have been implicated in chromatin related activity (Suppl. Table A.3.1). For example, factors ABF1, REB1 and RAP1 are transcription factors which are well-known for their ability to change local chromatin structure [8, 85]. It supports the idea of a special class of TFs, called "pioneering TFs" [126]. It has been speculated that these TFs factors bind to DNA and open up chromatin by evicting/shifting nucleosomes or by recruiting ATP-dependent chromatin remodeling enzymes which destabilize local chromatin and create nucleosome free region (NFR). After, establishment of an NFR the transcriptional machinery is able to bind to promoter and initiate transcription. Our study suggests that such factors need to have binding sites arranged in a way to help one another to outcompete nucleosomes.

We suggested plausible scenarios of how the noise could be minimized by positive or negative cooperativity. We show that cooperativity between TFs with coordinated binding activity could substantially reduce noise in TF binding. Moreover, in case of the negative cooperativity this reduction is most dramatic. However, we find not so many TFs pairs with optimized spacing for negative cooperativity, even though there are few examples (Fig. 3.7). It may be that the negative cooperativity is not important for gene regulation, or our modeling is missing some important details. For example, we use the assumption that nucleosomes can not be in partially unwrapped state, however, there is experimental evidence that nucleosomes can partially unwrap [30, 32, 77].

This work provides significant insight into the question of how promoter architecture control main characteristics of gene regulations. However, the hypothesis that have been made in this study need experimental validation. The direct experimental validation would be measuring TF and nucleosome occupancy at synthetic promoters with different architecture using experimental techniques with high-resolution, for example ChIP-exo presented in [89]. It would also be interesting to investigate how dynamics of TF binding depends on relative positioning of binding site using methods similar to presented in [57].

3.4 Materials and methods

3.4.1 Definitions of $C_{0.5}$ and $Noise$ in TF binding

We investigate characteristics of a binding curve, i.e. binding probability as a function of TF concentrations, namely $P(“on“|C)$.

For the simplest system with one binding site in promoter the probability is expressed as

$$P(“on“|C) = \frac{e^{\beta E+C}}{1+e^{\beta E+C}}, \tag{3.3}$$

where $\beta = 1/kT$, E is binding energy and C is log-concentration. Further, for simplicity we call the log-concentration just concentration if not mentioned otherwise.

An important feature of the binding curve is $C_{0.5}$,i.e. concentration of a TF at which probability of binding equals 0.5 (Fig. 3.8A). In other words, the $C_{0.5}$ is the concentration when a binding site changes it state from free to occupied.

Given the Equation 3.3 it is easy to calculte $C_{0.5}$ for the simplest case: $C_{0.5} = -\beta E$. In other words, the $C_{0.5}$ is proportional to effective affinity (binding energy) of a binding site, namely the higher affinity of a binding site the earlier the site becomes occupied.

The noise in TF binding at certain TF concentration can be thought as variance of the binomial distribution $B(1,p)$. If probability that a binding site occupied is $P(“on“|C)$ and probability that the binding site is free is $1 - P(“on“|C)$ then the noise is

$$Noise(C) = Var[B(1, P(“on“|C))] = P(“on“|C) \cdot (1 - P(“on“|C)) \tag{3.4}$$

The bell-shaped function $Noise(C)$ reflects what level of heterogeneity in a large population of identical cells one can expect with respect to a certain binding site (Fig. 3.8 B). Therefore the highest level of noise is associated with uncertain state of the binding, namely when $P(“on“|C) = 0.5$.

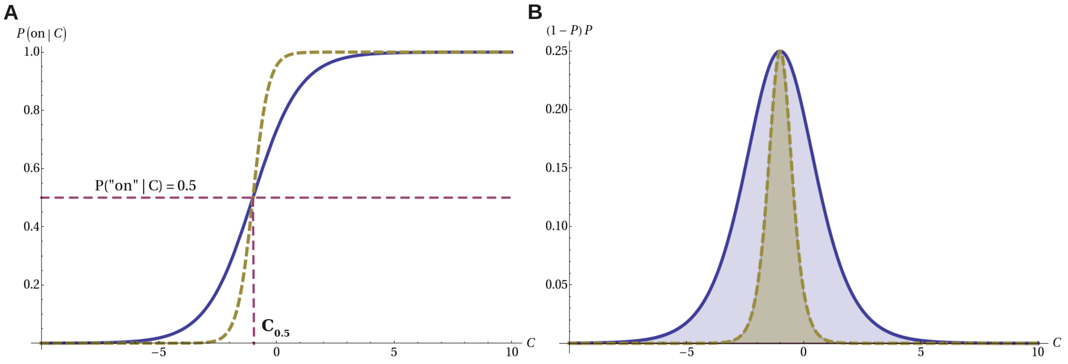

Figure 3.8: The definitions of $C_{0.5}$ and $Noise(C)$ of TF binding curve. The binding curves in **A** correspond to noise curves in **B**. The steeper binding curve (light green dashed curve in **A**) implies lower level of noise. The average noise during activation is the surface under the noise curve (filled areas in **B**).

Thus, the noise averaged across all concentrations (assuming uniform prior distribution of concentrations) is

$$Noise = \int_{-\infty}^{\infty} Noise(C)dC = \int_{-\infty}^{\infty} P(\text{“}on\text{“}|C) \cdot (1 - P(\text{“}on\text{“}|C))dC \qquad (3.5)$$

Note that for the simplest system noise in TF binding does not depend on affinity of a binding site. Indeed, given the Eq. 3.3 and the Eq. 3.5 it is easy to calculate

$$Noise = \int_{-\infty}^{\infty} P_{bound}(C) \cdot (1 - P_{bound}(C))dC = \int_{-\infty}^{\infty} \frac{e^{\beta E + C}}{(1 + e^{\beta E + C})^2} = 1 \qquad (3.6)$$

3.4.2 Modeling

We consider an artificial promoter comprised of only two TFBSs spaced distance d apart (Fig. 3.1 B). We assume that nucleosomes bind to DNA without sequence specificity, and set the nucleosome concentration such that average nucleosome coverage is about 80%, which is consistent with experimental data, e.g. [42, 55, 100].

For different distances d we sweep concentrations of TFs and use the biophysical model described in Chapter 2.4.1 to calculate the binding curve $P(\text{“}on\text{“}|C)$. Given the binding curve we investigate how the characteristics $C_{0.5}$ and $Noise$ depend on promoter architecture (i.e. distance d).

3.4.3 TFBS prediction

For transcription factor binding site (TFBS) prediction we used computational algorithm *MotEvo* [6]. The MotEvo algorithm uses position-specific weight matrix (WM) together with evolutionary conservation for calculating posterior probability of every position on a sequence to be a true binding site.

We predicted TFBSs for 158 WMs on intergenic regions of *S.cervisiae* (assembly sacCer2) using multiple allignments for 5 species: *S.cerevisiae*, *S.paradoxus*, *S.kudriavzevii*, *S.mikatae*, *S.bayanus*.

The set of 189 WMs (which cover 340 TFs) for *H.sapiens* and *M.musculus* was obtained by manual curation of WMs from JASPAR [90] and TRANSFAC [68] databases. For promoters (+/- 500bp relative to transcription start clusters) constructed from deepCAGE data [10] we predicted TFBSs on multiple allignments of 7 species: *H.sapiens (hg18)*, *R.macaca (rheMac2)*, *M.musculus (mm9)*, *C.familiaris (canFam2)*, *E.caballus (equCab1)*, *M.domestica (minDom4)* and *B.taurus (bosTau3)*. We consider only TFBSs with posterior probability $P \geq 0.75$. We also eliminate TFBSs falling into repeat regions since they can cause artificial spacing biases.

Given the TFBS prediction for every pair $TF_i : TF_j$ we searched for all pairs of binding sites and analyzed the distribution of distances between binding sites. We consider only those TF pairs which have at least 25 TFBS pairs spaced shorter then 900 bp. The density curve in Fig. 3.5 was plotted for distances weighted by the product of MotEvo posterior probabilities $P_{BS_1} \cdot P_{BS_2}$ using MatLab function *ksdensity* with gaussian kernel and window length 20 bp.

3.4.4 Statistical analysis of spacing between TFBSs

To disentangle spacing biases between TFBSs from possible spacing biases due to positioning relative to transcription start sites (TSS) we carried out a statistical test as following. To quantify the strength of cooperativity between TFs $TF_i : TF_j$ we used the value

$$< \delta C_{0.5} >= \sum_s P(s) \cdot \delta C_{0.5}(s) \tag{3.7}$$

where $P(s)$ is the probability to find a spacing s between TFBSs of TF_i and TF_j, and $\delta C_{0.5}(s) = C_{0.5TwoBS}(s) - C_{0.5OneBS}$, i.e. the difference in $C_{0.5}$ after addition of a second identical binding site to a promoter.

As the $C_{0.5}$ is closely related to effective binding energy of a binding site, i.e. $C_{0.5} =$

$-\beta E_{eff}$, the $< \delta C_{0.5} >$ can be thought as expected energy contribution $-\delta E_{eff}$ to a binding site of TF_i from nearby binding of TF_j.

For each pair of TFBSs $(TF_i : TF_j)_k$ and TSS t we construct a vector $v_{kt} = (d_{ikt}, d_{jkt}, T_k, P_{ik}, P_{jk})$, where d_{ikt} and d_{jkt} are distances from binding sites BS_{ik} and BS_{jk} to a TSS t, T_k is total number of TSSs which are near the TFBS pair and P_{ik}, P_{jk} are MotEvo posterior probabilities of binding sites. It is clear that the actual spacing distribution is

$$P_{actual}(s) = \frac{\sum_{v_{kt}} \delta_{s|d_{ikt}-d_{jkt}|} \frac{P_{ik}P_{jk}}{T_k}}{\sum_{v_{kt}} \frac{P_{ik}P_{jk}}{T_k}} \tag{3.8}$$

where $\delta_{s|d_{ikt}-d_{jkt}|}$ is the Kronecker delta.

To obtain $P_{random}(s)$ we randomly shuffle distances d_{ikt} among all vectors v_{kt} which is equivalent to random permutation of binding sites for TF_i among all promoters keeping constant TFBS distribution with respect to TSSs. Let vector $v^*_{kt} = (d^*_{ikt}, d_{jkt}, T_k, P_{ik}, P_{jk})$ be the vector with shuffled distances d^*_{ikt}. Then the random spacing distribution is

$$P_{random}(s) = \frac{\sum_{v^*_{kt}} \delta_{s|d^*_{ikt}-d_{jkt}|} \frac{P_{ik}P_{jk}}{T_k}}{\sum_{v^*_{kt}} \frac{P_{ik}P_{jk}}{T_k}} \tag{3.9}$$

After 10000 random permutations for every pair $TF_i : TF_j$ we obtain distribution of $< \delta C^{random}_{0.5} >$ and calculate Z statistic

$$Z(TF_i : TF_j) = \frac{< \delta C^{actual}_{0.5} > -\mu(< \delta C^{random}_{0.5} >)}{\sigma(< \delta C^{random}_{0.5} >)} \tag{3.10}$$

where $\mu(< \delta C^{random}_{0.5} >)$ and $\sigma(< \delta C^{random}_{0.5} >)$ are mean and standart deviation of $< \delta C^{random}_{0.5} >$.

The negative $Z(TF_i : TF_j)$ implies that the actual $\delta C_{0.5}$ is lower then expected by chance which means that $TF_i : TF_j$ positioned their binding sites so as to optimize positive cooperativity. And opposite, TF pairs with positive Z tend to negatively cooperate with each other.

The TSS positions in *S.cerevisiae* were taken from [74]. In human and mouse the TSS positions were defined as representative positions of transcription start clusters (TSC) constructed from the deepCAGE data [10], namely positions of the maximum number of CAGE tags in TSCs.

3.5 Appendix

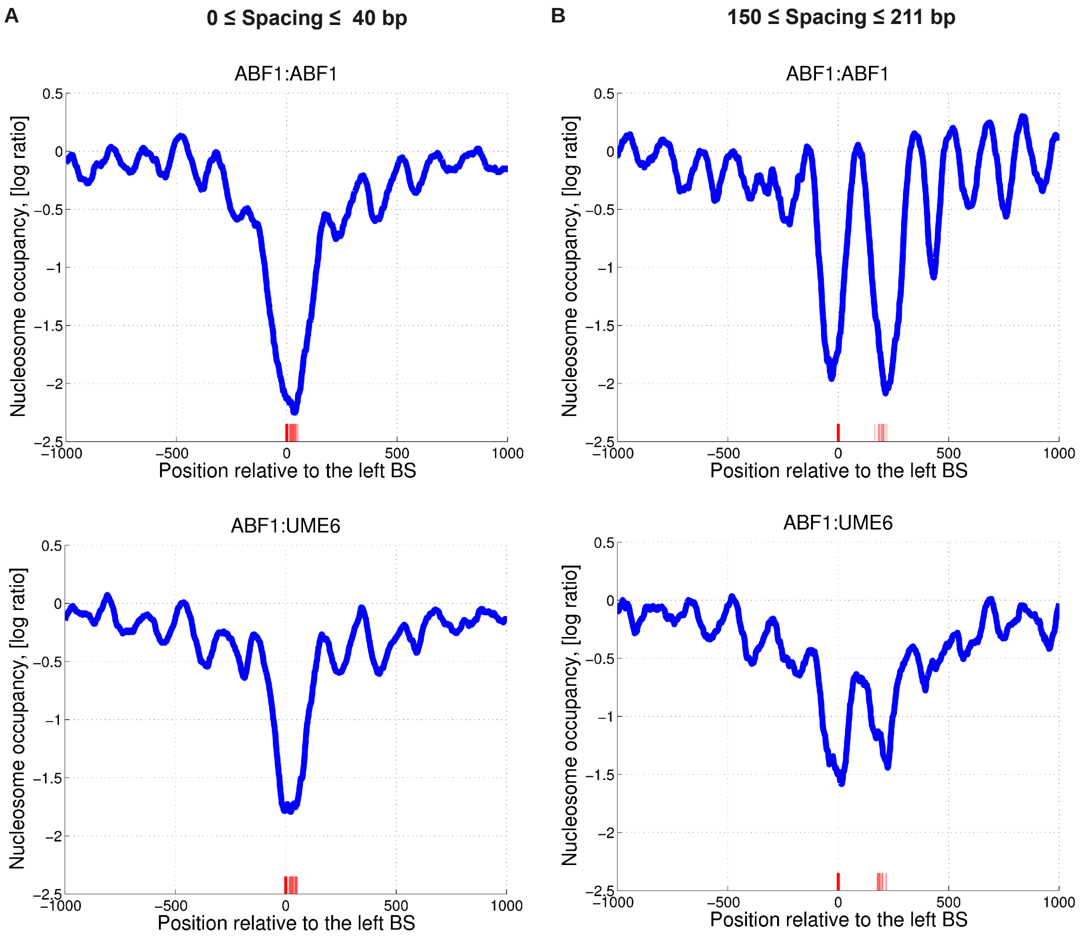

Figure A.3.9: Experimental nucleosome occupancy around positively spaced pairs of TFBSs. Experimental nucleosome occupancy (taken from [55]) alligned relative to the leftmost BS for TFBS pairs with spacing from 0 to 40 bp in **A** and from 150 to 211 bp in **B**.

TF_i	TF_j	Z-score	Number of pairs	Chromatin related TFs
PBF1	STB3	-11.73	181	PBF1
PBF2	STB3	-11.02	162	PBF2
RAP1	RAP1	-10.37	80	RAP1, RAP1
ABF1	REB1	-10.06	135	ABF1, REB1
STB3	YBL054W	-8.42	126	
ABF1	ABF1	-7.57	80	ABF1, ABF1
PBF1	SFP1	-7.19	62	PBF1
ABF1	YDR026c	-7.18	97	ABF1
PBF2	SFP1	-6.80	61	PBF2
ABF1	UME6	-6.24	96	ABF1, UME6
ABF1	RTG3	-6.17	52	ABF1
ABF1	CBF1	-5.98	44	ABF1
REB1	REB1	-5.78	73	REB1, REB1
TBF1	TBF1	-5.75	82	TBF1, TBF1
YRR1	YRR1	-5.46	35	
REB1	STB3	-5.45	155	REB1
NHP10	YNR063W	-5.45	37	NHP10
ABF1	UGA3	-5.36	91	ABF1
ABF1	STB3	-5.27	208	ABF1
NHP10	RDS2	-5.14	32	NHP10
ABF1	TYE7	-5.04	27	ABF1
REB1	UME6	-5.03	88	REB1, UME6
STB3	STB3	-4.63	89	
YDR026c	YDR026c	-4.57	45	
ABF1	RPN4	-4.45	104	ABF1
MBP1	REB1	-4.43	52	REB1
REB1	YDR026c	-4.36	57	REB1
SFP1	YBL054W	-4.34	44	
PDR3	YLR278C	-4.25	31	
DAL80	DAL80	-4.11	54	

Table A.3.1: Top 30 TFs with most significant spacing biases in *S.cerevisiae.* Among TF pairs with optimized spacing the chromatin related TFs are overrepresented ($p = 3{\cdot}10^{-5}$).

TF_i	TF_j	Z-score for $H.Sapiens$	Z-score for $M.Musculus$	#TF pairs in $H.Sapiens$	#TF pairs in $M.Musculus$
SP1	SP1	-19.24	-16.48	4096	3027
KLF4	SP1	-18.04	-15.61	2761	2372
NFY{A,B,C}	SP1	-17.37	-13.52	1307	1035
KLF4	KLF4	-15.01	-15.78	1859	1800
KLF4	NFY{A,B,C}	-11.95	-10.17	890	717
ELK1,4 GABP{A,B1}	ELK1,4 GABP{A,B1}	-11.08	-9.87	527	407
ELF1,2,4	ELK1,4 GABP{A,B1}	-10.53	-8.82	569	451
ELF1,2,4	ELF1,2,4	-10.26	-9.96	667	562
PATZ1	SP1	-9.14	-10.00	2303	1724
KLF4	PATZ1	-9.04	-6.99	1799	1426
bHLH_family	bHLH_family	-7.01	-6.15	92	78
CREB1	NFY{A,B,C}	-6.96	-7.38	102	78
CREB1	SP1	-6.85	-5.15	394	268
ATF5_CREB3	SP1	-6.69	-6.81	686	576
ARNT ARNT2 BHLHB2 MAX MYC USF1	ARNT ARNT2 BHLHB2 MAX MYC USF1	-6.66	-5.78	79	56
ELF1,2,4	FEV	-6.61	-5.50	207	180
EHF	ELK1,4 GABP{A,B1}	-6.60	-6.95	259	228
SP1	SREBF1,2	-6.45	-5.02	504	383
ARNT ARNT2 BHLHB2 MAX MYC USF1	SREBF1,2	-6.40	-5.52	72	60
ELF1,2,4	ETS1,2	-6.12	-5.17	257	175
ZNF143	ZNF143	-5.81	-5.32	276	186
ARNT ARNT2 BHLHB2 MAX MYC USF1	SP1	-5.57	-5.86	544	400
SP1	TFDP1	-5.52	-6.87	2013	1740
EHF	ELF1,2,4	-5.30	-7.34	268	224
MAZ	MAZ	-5.15	-6.28	1914	1526
CREB1	KLF4	-5.14	-6.42	274	186
ATF4	SP1	-5.10	-5.65	337	217

Table A.3.2: Pairs of TFs with $Z \leq -5$ both in human and mouse.

Chapter 4

Conclusions and Discussion

This thesis is devoted to a quantitative investigation of the effects which are induced by competition between nucleosomes and transcription factors. We use a rigorous statistical mechanics model that calculates occupancy profiles of nucleosomes and TFs to study binding processes and address two key questions: 1) how competition with transcription factors affects nucleosome positioning and 2) how nucleosomes and promoter architecture affect binding of transcription factors.

The second chapter was devoted to the question of what determines nucleosome positioning in promoters of genes. First we investigated how well different datasets which were generated in different labs correlate with each other. It turns out that measured nucleosome occupancies does not correlate very well and can vary from lab to lab. For instance, Pearson correlation between *in vivo* nucleosome occupancies can range from $r = 0.18$ to $r = 0.65$ (Fig. 2.1 A). Presumably, this large variability across datasets to some extent may be due to biases of the technological platforms, i.e. microarray and high-throughput sequencing. Even though the raw nucleosome occupancy is not very well correlated, the positions of nucleosomes and linkers are quite consistent between datasets. In other words we show that whereas amplitude of nucleosome occupancy signal can vary strongly from dataset to dataset, the positions of peaks (nucleosomes) and troughs (linkers) are very reproducible (Fig. 2.1 B). We show that comparison of the positions of nucleosomes and linkers is more informative than comparison of the raw occupancy signal. Therefore, we introduced a new method for assessment of performance of the biophysical model which is based on mutual information between experimental annotation of "nucleosomes" and "linkers" and prediction. The F score introduced in the chapter 2 was used as an objective function to fit the model to nucleosome data.

Rigorous modeling helped us to estimate how well the competition between nucleosomes and transcription factors can explain observed nucleosome patterns *in vivo*. We show that adding transcription factors into the model improves its performance in predicting nucleosome positions, especially at promoters of genes. Importantly, the model of competitive binding of transcription factors can reproduce the remarkable nucleosome pattern at the 5' end of genes (Fig. 2.5). These results reconcile the previous seemingly conflicting results on the determinants of nucleosome positioning, and provide a quantitative explanation for the difference between *in vivo* and *in vitro* positioning. We the draw conclusion that whereas nucleosomes in gene bodies are positioned due to intrinsic sequence specificity of nucleosomes, the nucleosome free regions (NFRs) result from competitive binding of transcription factors.

Importantly, we show that only a small subset of TFs contribute to NFR formation (Fig. 2.7). The test with shuffled WMs proved that the ability of TFs in this small subset to explain nucleosome positioning is a specific property of the sequence specificities of yeast's TFs. The other test involving fitting of *in vitro* data, showed that this small subset of TFs is specific to *in vivo* nucleosome positioning. Remarkably, many TFs in this small subset tend to be involved in the processes of chromatin remodeling.

In summary, the results presented in chapter 2 support the hypothesis that general regulatory factors play a major role in nucleosome organization in promoters of genes, in particular in the formation of nucleosome free regions.

In chapter 3 we studied how nucleosomes may affect binding of transcription factors. We theoretically investigated processes of TF binding to a toy cluster of TFBSs in the context of chromatin. We show that competition between nucleosomes and TFs for binding to DNA can induce interesting effects, such as cooperativity between TFs. Biophysical modeling carried out in chapter 3 shows that the ability of TFs to outcompete nucleosomes crucially depends on the architecture of a cluster of TFBSs. Due to the statistical positioning of nucleosomes, the cooperativity between TFs periodically depends on distance between binding sites (Fig. 3.1). Moreover, we demonstrate that cooperativity between TFs can significantly reduce noise in TF binding.

Strikingly, the investigation of binding site arrangement in real genomes shows that some TFs have their binding sites optimized for positive cooperativity (Fig. 3.5, Tables A.3.1 and A.3.2). The observed biases in spacing between TFBSs strongly support the hypothesis that nucleosome mediated cooperativity may play an important role in gene regulation.

Interestingly, many TFs which have been suggested in chapter 2 to participate in NFR formation also have pronounced spacing biases for positive nucleosome mediated cooperativity. It implies that 1) positions of binding sites for those TFs explain positions of nucleosome free regions and 2) binding sites for those TFs are arranged so as to outcompete nucleosomes in the most efficient way. It suggests that those TFs have special functions in chromatin related processes.

Even though, the simple analysis of binding sites in real genomes presented in this thesis reveals interesting features of TFBS clusters, it still unclear how the actual architecture of binding sites affect transcription. It would be very interesting to investigate whether there are distinct motifs or patterns in TFBS clusters across different promoters and, if so, how the presence of such motifs correlates with characteristics of gene expression, such as "noise", transcriptional plasticity and so on. In other words, it would be very helpful to develop an algorithm which finds patterns in the relative positioning of binding sites given a set of TFBS clusters. For instance, in this thesis we discovered simple patterns for some pairs of TFs, e.g. $ABF1 : L_{0-40,150-200} : UME6$, which stands for an ABF1 binding site separated by a linker of length 0-40bp or 150-200bp from a binding site of UME6 (see Fig. 3.5). However, there might be more complex patterns in arrangement of binding sites, e.g. $ABF1 : L_{15-40} : ABF1 : L_{160-180} : REB1$. In order to identify such TFBS patterns several problems need to be resolved. Firstly, patterns which could appear in TFBS clusters are "fuzzy", namely there is not much difference between linkers with length 15bp and 16bp. Secondly, it is not clear how to deal with cases when different binding sites in a cluster are functional only under some particular conditions. For example, it might be that in a TFBS cluster $BS_1 : L_1 : BS_2 : L2 : BS_3 : L_3 : BS_4$, binding sites BS_1 and BS_2 are only used under heatshock condition, and sites BS_3 and BS_4 only in rapid growth of a cell.

In spite of the simplicity of the model introduced in this thesis, it is able to reproduce observed nucleosome patterns and predict effect which may influence gene regulation. Nevertheless, there are still many factors which need to be taken into account in order to get more realistic model. For example, it is now unclear whether the assumption holds that the system is in its thermodynamic equilibrium state. It is also accepted nowadays that the posttranslational modifications of histones can greatly affect binding energy of nucleosomes [56]. This could be taken into account by introducing into the model different kinds of nucleosomes which have different energy scaling factors γ. However, the variety of histone modifications might lead to huge number of parameters in the model. Nevertheless, by restricting the number of histone modifications under consideration, e.g. only "no modification", "accetylation" and methylation" similar to [24], it seems possible to take it into account.

In addition, nucleosomes have been observed in a partially unwrapped or "loose" state [32, 124], which also could be taken into account by introducing different kinds of nucleosomes with variable footprint, e.g. normal nucleosomes which cover 147bp and partially unwrapped nucleosomes which cover smaller regions, for instance from 130 till 146 bp.

Chromatin biology is a rapidly evolving field nowadays. Due to technological breakthroughs and original experiments significant insights have been made into the role of chromatin in gene regulation. Nevertheless, there are many open questions that require further investigation. For example, it is now unclear what mechanisms underlie the formation of the NFR at 3' ends of genes. It has been suggested that the special spatial configuration of genes called a "gene loop" may be the reason for the 3' end NFR [69]. The gene loop conformation, when a promoter is juxtaposed with a terminator, has been observed for FMP27 and SEN1 genes in *S.cerevisiae* and was hypothesized to help in maintaining high rate of transcription [79]. Another study suggests that the gene loop conformation controls transcriptional directionality [106]. However, whether 3' end NFRs are linked with transcription termination and how they are related to gene regulation is not fully understood yet.

Since all experimental techniques for mapping locations of nucleosomes and transcription factors deal with large populations of cells, the data that we obtain from such experiments represent average across the whole population. Therefore, the question of cell-to-cell variability in transcription factor and nucleosome configurations remains open. Although it is now possible to measure mRNA abundance on the single cell level, e.g. [84], experimental techniques which are able to reveal chromatin and TF distribution for a single cell are still missing.

In general, understanding of how information encoded in regulatory sequences of the genome is interpreted and converted into gene expression is a key challenge of modern molecular biology. Experiments, similar to [99], where artificially designed promoters drive expression of a reporter protein would be very helpful in unraveling the enigma of gene regulation. These experiments make it possible to measure the activity of thousands of synthetic promoters by measuring expression of a reporter protein. I believe that such experimental data, together with mathematical modeling, similar to [37, 88] and presented in this thesis, will lead to the construction of a comprehensive model of gene regulation.

Bibliography

[1] I. Albert, S. Wachi, C. Jiang, and B. F. Pugh. GeneTrack–a genomic data processing and visualization framework. *Bioinformatics*, 24(10):1305–1306, May 2008. 45

[2] Istvan Albert, Travis N. Mavrich, Lynn P. Tomsho, Ji Qi, Sara J. Zanton, Stephan C. Schuster, and B Franklin Pugh. Translational and rotational settings of h2a.z nucleosomes across the saccharomyces cerevisiae genome. *Nature*, 446(7135):572–576, Mar 2007. 11

[3] James Allan, Ross M. Fraser, Tom Owen-Hughes, and David Keszenman-Pereyra. Micrococcal nuclease does not substantially bias nucleosome mapping. *J Mol Biol*, 417(3):152–164, Mar 2012. 7

[4] A. Almer and W. Hörz. Nuclease hypersensitive regions with adjacent positioned nucleosomes mark the gene boundaries of the pho5/pho3 locus in yeast. *EMBO J*, 5(10):2681–2687, Oct 1986. 6

[5] A. Almer, H. Rudolph, A. Hinnen, and W. Hörz. Removal of positioned nucleosomes from the yeast pho5 promoter upon pho5 induction releases additional upstream activating dna elements. *EMBO J*, 5(10):2689–2696, Oct 1986. 6

[6] Phil Arnold, Ionas Erb, Mikhail Pachkov, Nacho Molina, and Erik van Nimwegen. Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. *Bioinformatics*, 28(4):487–494, Feb 2012. 70, 77

[7] Phil Arnold, Anne Schöler, Mikhail Pachkov, Piotr Balwierz, Helle Jørgensen, Michael B. Stadler, Erik van Nimwegen, and Dirk Schübeler. Modeling of epigenome dynamics identifies transcription factors that mediate polycomb targeting. *Genome Res*, Sep 2012. 64

[8] Gwenael Badis, Esther T Chan, Harm van Bakel, Lourdes Pena-Castillo, Desiree Tillo, Kyle Tsui, Clayton D Carlson, Andrea J Gossett, Michael J Hasinoff, Christopher L Warren, Marinella Gebbia, Shaheynoor Talukder, Ally Yang, Sanie Mnaimneh, Dimitri Terterov, David Coburn, Ai Li Yeo, Zhen Xuan Yeo, Neil D Clarke, Jason D Lieb, Aseem Z Ansari, Corey Nislow, and Timothy R Hughes. A library of yeast transcription factor motifs reveals a widespread function for rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell*, 32(6):878–887, Dec 2008. 12, 17, 40, 50, 74

[9] L. Bai, A. Ondracka, and F. R. Cross. Multiple sequence-specific factors generate the nucleosome-depleted region on CLN2 promoter. *Mol. Cell*, 42:465–476, May 2011. 12, 17, 38

[10] Piotr J Balwierz, Piero Carninci, Carsten O Daub, Jun Kawai, Yoshihide Hayashizaki, Werner Van Belle, Christian Beisel, and Erik van Nimwegen. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepcage data. *Genome Biol*, 10(7):R79, 2009. 47, 77, 78

[11] Y. Bao and X. Shen. SnapShot: chromatin remodeling complexes. *Cell*, 129(3):632, May 2007. 50

[12] Andrew D. Basehoar, Sara J. Zanton, and B Franklin Pugh. Identification and distinct regulation of yeast tata box-containing genes. *Cell*, 116(5):699–709, Mar 2004. 11

[13] Christian Beisel and Renato Paro. Silencing chromatin: comparing modes and mechanisms. *Nat Rev Genet*, 12(2):123–135, Feb 2011. 64

[14] O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193:723–750, 1987. 19, 40

[15] Bradley E. Bernstein, Chih Long Liu, Emily L. Humphrey, Ethan O. Perlstein, and Stuart L. Schreiber. Global nucleosome occupancy in yeast. *Genome Biol*, 5(9):R62, 2004. 6

[16] William J Blake, Mads KAErn, Charles R Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, Apr 2003. 67

[17] Kristin Brogaard, Liqun Xi, Ji-Ping Wang, and Jonathan Widom. A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486(7404):496–501, Jun 2012. 6, 7

[18] H. J. Bussemaker, H. Li, and E. D. Siggia. Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 97:10096–100, 2000. 19, 42

[19] K. Chen, E. van Nimwegen, N. Rajewsky, and M. L. Siegal. Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae. *Genome Biol Evol*, 2:697–707, 2010. 19, 40

[20] G. Chevereau, L. Palmeira, C. Thermes, A. Arneodo, and C. Vaillant. Thermodynamics of intragenic nucleosome ordering. *Phys Rev Lett*, 103(18):188103, Oct 2009. 17, 19

[21] H. R. Chung, I. Dunkel, F. Heise, C. Linke, S. Krobitsch, A. E. Ehrenhofer-Murray, S. R. Sperling, and M. Vingron. The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS ONE*, 5:e15754, 2010. 17, 23

[22] Ho-Ryun Chung, Ilona Dunkel, Franziska Heise, Christian Linke, Sylvia Krobitsch, Ann E Ehrenhofer-Murray, Silke R Sperling, and Martin Vingron. The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One*, 5(12):e15754, 2010. 7

[23] Sara J Cooper, Nathan D Trinklein, Elizabeth D Anton, Loan Nguyen, and Richard M Myers. Comprehensive analysis of transcriptional promoter structure and function in 1 *Genome Res*, 16(1):1–10, Jan 2006. 64

[24] Ian B. Dodd, Mille A. Micheelsen, Kim Sneppen, and Geneviève Thon. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell*, 129(4):813–822, May 2007. 85

[25] H. R. Drew and A. A. Travers. Dna bending and its relation to nucleosome positioning. *J Mol Biol*, 186(4):773–790, Dec 1985. 11

[26] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, Aug 2002. 67

[27] X. Fan, Z. Moqtaderi, Y. Jin, Y. Zhang, X. S. Liu, and K. Struhl. Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. *Proc. Natl. Acad. Sci. U.S.A.*, 107(42):17945–17950, Oct 2010. 31

[28] Yair Field, Noam Kaplan, Yvonne Fondufe-Mittendorf, Irene K Moore, Eilon Sharon, Yaniv Lubling, Jonathan Widom, and Eran Segal. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol*, 4(11):e1000216, Nov 2008. 6, 7, 21, 22, 44

[29] J. T. Finch, L. C. Lutter, D. Rhodes, R. S. Brown, B. Rushton, M. Levitt, and A. Klug. Structure of nucleosome core particles of chromatin. *Nature*, 269(5623):29–36, Sep 1977. 4, 11

[30] M. Floer, X. Wang, V. Prabhu, G. Berrozpe, S. Narayan, D. Spagna, D. Alvarez, J. Kendall, A. Krasnitz, A. Stepansky, J. Hicks, G. O. Bryant, and M. Ptashne. A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. *Cell*, 141:407–418, Apr 2010. 17, 38, 75

[31] M. Ganapathi, M. J. Palumbo, S. A. Ansari, Q. He, K. Tsui, C. Nislow, and R. H. Morse. Extensive role of the general regulatory factors, Abf1 and Rap1, in determining genome-wide chromatin structure in budding yeast. *Nucleic Acids Res.*, 39:2032–2044, Mar 2011. 17

[32] D. S. Geraghty, H. B. Sucic, J. Chen, and D. S. Pederson. Evidence that partial unwrapping of dna from nucleosomes facilitates the binding of heat shock factor following dna replication in yeast. *J Biol Chem*, 273(32):20463–20472, Aug 1998. 75, 85

[33] Raluca Gordan, Alexander J Hartemink, and Martha L Bulyk. Distinguishing direct versus indirect transcription factor-dna interactions. *Genome Res*, 19(11):2090–2100, Nov 2009. 40

[34] M. Han and M. Grunstein. Nucleosome loss activates yeast downstream promoters in vivo. *Cell*, 55(6):1137–1145, Dec 1988. 4

[35] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok DK, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S.

Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004. 32, 57, 60

[36] Olivier Harismendy, Pauline C Ng, Robert L Strausberg, Xiaoyun Wang, Timothy B Stockwell, Karen Y Beeson, Nicholas J Schork, Sarah S Murray, Eric J Topol, Samuel Levy, and Kelly A Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*, 10(3):R32, 2009. 9, 23

[37] Xin He, Md Abul Hassan Samee, Charles Blatti, and Saurabh Sinha. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol*, 6(9), 2010. 86

[38] Nathaniel D Heintzman, Gary C Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, Keith A Ching, Jessica E Antosiewicz-Bourget, Hui Liu, Xinmin Zhang, Roland D Green, Victor V Lobanenkov, Ron Stewart, James A Thomson, Gregory E Crawford, Manolis Kellis, and Bing Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, May 2009. 39, 64

[39] Amanda L. Hughes, Yi Jin, Oliver J. Rando, and Kevin Struhl. A functional evolutionary approach to identify determinants of nucleosome positioning: A unifying model for establishing the genome-wide pattern. *Mol Cell*, Aug 2012. 12

[40] I. Ioshikhes, A. Bolshoy, K. Derenshteyn, M. Borodovsky, and E. N. Trifonov. Nucleosome dna sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J Mol Biol*, 262(2):129–139, Sep 1996. 11

[41] V. Iyer and K. Struhl. Poly(da:dt), a ubiquitous promoter element that stimulates transcription via its intrinsic dna structure. *EMBO J*, 14(11):2570–2579, Jun 1995. 11

[42] A. Jansen and K. J. Verstrepen. Nucleosome positioning in Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.*, 75:301–320, Jun 2011. 26, 76

[43] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. 25

[44] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, Aug 2001. 6, 64

[45] Cizhong Jiang and B. Franklin Pugh. A compiled and systematic reference map of nucleosome positions across the saccharomyces cerevisiae genome. *Genome Biol*, 10(10):R109, 2009. 4, 5, 9, 11, 22, 23, 24, 27, 44, 45

[46] Steven M Johnson, Frederick J Tan, Heather L McCullough, Daniel P Riordan, and Andrew Z Fire. Flexibility and constraint in the nucleosome core landscape of caenorhabditis elegans chromatin. *Genome Res*, 16(12):1505–1516, Dec 2006. 7, 21

[47] N. Kaplan, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal. Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome Biol.*, 11:140, 2010. 17

[48] Noam Kaplan, Irene K Moore, Yvonne Fondufe-Mittendorf, Andrea J Gossett, Desiree Tillo, Yair Field, Emily M LeProust, Timothy R Hughes, Jason D Lieb, Jonathan Widom, and Eran Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, Mar 2009. 6, 7, 11, 12, 17, 19, 20, 21, 22, 23, 25, 26, 40, 41, 44, 45, 51

[49] J. A. Knezetic and D. S. Luse. The presence of nucleosomes on a dna template prevents initiation by rna polymerase ii in vitro. *Cell*, 45(1):95–104, Apr 1986. 4

[50] R. Thomas Koerber, Ho Sung Rhee, Cizhong Jiang, and B. Franklin Pugh. Interaction of transcriptional regulators with specific nucleosomes across the saccharomyces genome. *Mol Cell*, 35(6):889–902, Sep 2009. 17

[51] R. D. Kornberg. Chromatin structure: a repeating unit of histones and dna. *Science*, 184(4139):868–871, May 1974. 4, 26

[52] R. D. Kornberg and L. Stryer. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res*, 16(14A):6677–6690, Jul 1988. 12, 16, 43

[53] F. H. Lam, D. J. Steger, and E. K. O'Shea. Chromatin decouples promoter threshold from dynamic range. *Nature*, 453:246–250, May 2008. 38

[54] Cheol-Koo Lee, Yoichiro Shibata, Bhargavi Rao, Brian D. Strahl, and Jason D. Lieb. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet*, 36(8):900–905, Aug 2004. 6

[55] William Lee, Desiree Tillo, Nicolas Bray, Randall H Morse, Ronald W Davis, Timothy R Hughes, and Corey Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet*, 39(10):1235–1244, Oct 2007. 6, 7, 9, 10, 11, 16, 17, 21, 22, 25, 26, 30, 31, 33, 40, 44, 51, 55, 56, 72, 76, 79

[56] Bing Li, Michael Carey, and Jerry L Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, Feb 2007. 4, 9, 64, 85

[57] Colin R. Lickwar, Florian Mueller, Sean E. Hanlon, James G. McNally, and Jason D. Lieb. Genome-wide protein-dna binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, 484(7393):251–255, Apr 2012. 75

[58] Doron Lipson, Tal Raz, Alix Kieu, Daniel R Jones, Eldar Giladi, Edward Thayer, John F Thompson, Stan Letovsky, Patrice Milos, and Marie Causey. Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol*, 27(7):652–658, Jul 2009. 36, 59

[59] George Locke, Denis Tolkunov, Zarmik Moqtaderi, Kevin Struhl, and Alexandre V Morozov. High-throughput sequencing reveals a simple model of nucleosome energetics. *Proc Natl Acad Sci U S A*, 107(49):20998–21003, Dec 2010. 7, 17, 19, 22, 23, 40, 44

[60] D. Lohr. Organization of the gal1-gal10 intergenic control region chromatin. *Nucleic Acids Res*, 12(22):8457–8474, Nov 1984. 6

[61] D. Lohr. Chromatin structure and regulation of the eukaryotic regulatory gene gal80. *Proc Natl Acad Sci U S A*, 90(22):10628–10632, Nov 1993. 6

[62] Y. Lorch, J. W. LaPointe, and R. D. Kornberg. On the displacement of histones from dna by transcription. *Cell*, 55(5):743–744, Dec 1988. 4

[63] P. T. Lowary and J. Widom. New dna sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol*, 276(1):19–42, Feb 1998. 11

[64] P. T. Lowary and J. Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.*, 276(1):19–42, Feb 1998. 16

[65] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–260, Sep 1997. 4, 11

[66] Alexandra Lusser and James T. Kadonaga. Strategies for the reconstitution of chromatin. *Nat Methods*, 1(1):19–26, Oct 2004. 7

[67] Maud Marques, Liette Laflamme, Alain L. Gervais, and Luc Gaudreau. Reconciling the positive and negative roles of histone h2a.z in gene transcription. *Epigenetics*, 5(4):267–272, May 2010. 6

[68] V. Matys, E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–378, Jan 2003. 77

[69] Travis N Mavrich, Ilya P Ioshikhes, Bryan J Venters, Cizhong Jiang, Lynn P Tomsho, Ji Qi, Stephan C Schuster, Istvan Albert, and B. Franklin Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res*, 18(7):1073–1083, Jul 2008. 6, 7, 9, 12, 16, 17, 21, 22, 30, 44, 86

[70] Travis N Mavrich, Cizhong Jiang, Ilya P Ioshikhes, Xiaoyong Li, Bryan J Venters, Sara J Zanton, Lynn P Tomsho, Ji Qi, Robert L Glaser, Stephan C Schuster, David S Gilmour, Istvan Albert, and B. Franklin Pugh. Nucleosome organization in the drosophila genome. *Nature*, 453(7193):358–362, May 2008. 7, 21

[71] Leonid A Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci U S A*, 107(52):22534–22539, Dec 2010. 40, 65

[72] W. Mobius and U. Gerland. Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites. *PLoS Comput. Biol.*, 6, 2010. 12, 16, 20

[73] Alexandre V Morozov, Karissa Fortney, Daria A Gaykalova, Vasily M Studitsky, Jonathan Widom, and Eric D Siggia. Using dna mechanics to predict in vitro

nucleosome positions and formation energies. *Nucleic Acids Res*, 37(14):4707–4722, Aug 2009. 65

[74] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349, Jun 2008. 10, 45, 78

[75] Geeta J Narlikar, Hua-Ying Fan, and Robert E Kingston. Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, 108(4):475–487, Feb 2002. 64

[76] John R S. Newman, Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. DeRisi, and Jonathan S. Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, Jun 2006. 11

[77] Justin A. North, John C. Shimko, Sarah Javaid, Alex M. Mooney, Matthew A. Shoffner, Sean D. Rose, Ralf Bundschuh, Richard Fishel, Jennifer J. Ottesen, and Michael G. Poirier. Regulation of the nucleosome unwrapping rate controls dna accessibility. *Nucleic Acids Res*, Sep 2012. 75

[78] Donald E. Olins and Ada L. Olins. Chromatin history: our view from the bridge. *Nat Rev Mol Cell Biol*, 4(10):809–814, Oct 2003. 5

[79] JM O'Sullivan, SM Tan-Wong, A Morillon, B Lee, J Coles J Mellor, and NJ Proudfoot. Gene loops juxtapose promoters and terminators in yeast. *Nat Genet*, 36(9):1014–1018, 2004. 86

[80] H. E. Peckham, R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng. Nucleosome positioning signals in genomic DNA. *Genome Res.*, 17(8):1170–1177, Aug 2007. 17

[81] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nat Biotechnol*, 28(10):1057–1068, Oct 2010. 6

[82] Ryan M. Raisner, Paul D. Hartley, Marc D. Meneghini, Marie Z. Bao, Chih Long Liu, Stuart L. Schreiber, Oliver J. Rando, and Hiten D. Madhani. Histone variant h2a.z marks the 5' ends of both active and inactive genes in euchromatin. *Cell*, 123(2):233–248, Oct 2005. 6, 9

[83] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia. Computational detection of genomic cis-regulatory modules, applied to body patterning in the early drosophila embryo. *BMC Bioinformatics*, 3(30), 2002. 19, 42

[84] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R. Faridani, Gregory A. Daniels, Irina Khrebtukova, Jeanne F. Loring, Louise C. Laurent, Gary P. Schroth, and Rickard Sandberg. Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. *Nat Biotechnol*, 30(8):777–782, Aug 2012. 86

[85] Oliver J. Rando and Fred Winston. Chromatin and transcription in yeast. *Genetics*, 190(2):351–387, Feb 2012. 4, 11, 73, 74

[86] Jonathan M Raser and Erin K O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–1814, Jun 2004. 67

[87] T. Raveh-Sadka, M. Levo, and E. Segal. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res.*, 19(8):1480–1496, Aug 2009. 19, 20, 42

[88] Tali Raveh-Sadka, Michal Levo, and Eran Segal. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res*, 19(8):1480–1496, Aug 2009. 65, 86

[89] Ho Sung Rhee and B Franklin Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, Dec 2011. 75

[90] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, Jan 2004. 77

[91] Kavitha Sarma and Danny Reinberg. Histone variants meet their match. *Nat Rev Mol Cell Biol*, 6(2):139–149, Feb 2005. 6

[92] S. C. Satchwell, H. R. Drew, and A. A. Travers. Sequence periodicities in chicken nucleosome core dna. *J Mol Biol*, 191(4):659–675, Oct 1986. 11

[93] Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, Mar 2008. 7, 21

[94] David J. Schwab, Robijn F. Bruinsma, Joseph Rudnick, and Jonathan Widom. Nucleosome switches. *Phys Rev Lett*, 100(22):228105, Jun 2008. 19

[95] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442:772–778, Aug 2006. 17, 19, 20, 40

[96] E. Segal and J. Widom. What controls nucleosome positions? *Trends Genet.*, 25:335–343, Aug 2009. 17, 20

[97] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K. Moore, Ji-Ping Z. Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, Aug 2006. 11

[98] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27, 1948. 25

[99] Eilon Sharon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*, 30(6):521–530, Jun 2012. 86

[100] Sushma Shivaswamy, Akshay Bhinge, Yongjun Zhao, Steven Jones, Martin Hirst, and Vishwanath R Iyer. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol*, 6(3):e65, Mar 2008. 6, 7, 12, 16, 21, 22, 23, 26, 44, 76

[101] R. Siddharthan, E. D. Siggia, and E. van Nimwegen. Phylogibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, 2005. 40

[102] R. T. Simpson and D. W. Stafford. Structural features of a phased nucleosome core particle. *Proc. Natl. Acad. Sci. U.S.A.*, 80(1):51–55, Jan 1983. 16

[103] C. L. Smith, R. Horowitz-Scherer, J. F. Flanagan, C. L. Woodcock, and C. L. Peterson. Structural analysis of the yeast SWI/SNF chromatin remodeling complex. *Nat. Struct. Biol.*, 10(2):141–145, Feb 2003. 50

[104] Arnold Stein, Taichi E Takasuka, and Clayton K Collings. Are nucleosome positions in vivo primarily determined by histone-dna sequence preferences? *Nucleic Acids Res*, 38(3):709–719, Jan 2010. 9, 17, 23, 47

[105] K. Struhl. Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, 82(24):8419–8423, Dec 1985. 16

[106] Sue Mei Tan-Wong, Judith B. Zaugg, Jurgi Camblong, Zhenyu Xu, David W. Zhang, Hannah E. Mischo, Aseem Z. Ansari, Nicholas M. Luscombe, Lars M. Steinmetz, and Nick J. Proudfoot. Gene loops enhance transcriptional directionality. *Science*, Sep 2012. 86

[107] Desiree Tillo and Timothy R. Hughes. G+c content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, 10:442, 2009. 11

[108] Sylvia C. Tippmann, Robert Ivanek, Dimos Gaidatzis, Anne Schöler, Leslie Hoerner, Erik van Nimwegen, Peter F. Stadler, Michael B. Stadler, and Dirk Schübeler. Chromatin measurements reveal contributions of synthesis and decay to steady-state mrna levels. *Mol Syst Biol*, 8:593, 2012. 3, 64

[109] I Tirosh and N Barkai. Two strategies for gene regulation by promoter nucleosomes. *Genome Res*, 18:1084–1091, 2008. 11, 73

[110] Itay Tirosh, Naama Barkai, and Kevin J. Verstrepen. Promoter architecture and the evolvability of gene expression. *J Biol*, 8(11):95, 2009. 73

[111] D. Tolkunov, K. A. Zawadzki, C. Singer, N. Elfving, A. V. Morozov, and J. R. Broach. Chromatin remodelers clear nucleosomes from intrinsically unfavorable sites to establish nucleosome-depleted regions at promoters. *Mol. Biol. Cell*, 22(12):2106–2118, Jun 2011. 20

[112] Alexander M. Tsankov, Dawn Anne Thompson, Amanda Socha, Aviv Regev, and Oliver J. Rando. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol*, 8(7):e1000414, 2010. 12

[113] Anton Valouev, Jeffrey Ichikawa, Thaisan Tonthat, Jeremy Stuart, Swati Ranade, Heather Peckham, Kathy Zeng, Joel A Malek, Gina Costa, Kevin McKernan, Arend Sidow, Andrew Fire, and Steven M Johnson. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res*, 18(7):1051–1063, Jul 2008. 7, 21

[114] Erik van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, 8 Suppl 6:S4, 2007. 19, 40, 42

[115] S. Vashee, K. Melcher, W. V. Ding, S. A. Johnston, and T. Kodadek. Evidence for two modes of cooperative dna binding in vivo that do not involve direct protein-protein interactions. *Curr Biol*, 8(8):452–458, Apr 1998. 73

[116] Bryan J Venters and B. Franklin Pugh. A canonical promoter organization of the transcription machinery and its regulators in the saccharomyces genome. *Genome Res*, 19(3):360–371, Mar 2009. 31

[117] Axel Visel, Matthew J Blow, Zirong Li, Tao Zhang, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M Rubin, and Len A Pennacchio. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, Feb 2009. 39, 64

[118] X. Wang, L. Bai, G. O. Bryant, and M. Ptashne. Nucleosomes and the accessibility problem. *Trends Genet.*, 27:487–492, Dec 2011. 17, 38

[119] T. Wasson and A. J. Hartemink. An ensemble model of competitive multi-factor binding of the genome. *Genome Res.*, 19:2101–2112, Nov 2009. 19, 20, 42

[120] Iestyn Whitehouse, Oliver J Rando, Jeff Delrow, and Toshio Tsukiyama. Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, 450(7172):1031–1035, Dec 2007. 7, 21

[121] J. L. Workman and R. G. Roeder. Binding of transcription factor tfiid to the major late promoter during in vitro nucleosome assembly potentiates subsequent initiation by rna polymerase ii. *Cell*, 51(4):613–622, Nov 1987. 4

[122] Zeba Wunderlich and Leonid A Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet*, 25(10):434–440, Oct 2009. 64

[123] Hualin Xi, Hennady P Shulha, Jane M Lin, Teresa R Vales, Yutao Fu, David M Bodine, Ronald D G McKay, Josh G Chenoweth, Paul J Tesar, Terrence S Furey, Bing Ren, Zhiping Weng, and Gregory E Crawford. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet*, 3(8):e136, Aug 2007. 64

[124] Yuanxin Xi, Jianhui Yao, Rui Chen, Wei Li, and Xiangwei He. Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. *Genome Res*, 21(5):718–724, May 2011. 85

[125] Guo-Cheng Yuan, Yuen-Jong Liu, Michael F. Dion, Michael D. Slack, Lani F. Wu, Steven J. Altschuler, and Oliver J. Rando. Genome-scale identification of nucleosome positions in s. cerevisiae. *Science*, 309(5734):626–630, Jul 2005. 6, 9, 12

[126] Kenneth S Zaret and Jason S Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*, 25(21):2227–2241, Nov 2011. 74

[127] Yong Zhang, Zarmik Moqtaderi, Barbara P Rattner, Ghia Euskirchen, Michael Snyder, James T Kadonaga, X. Shirley Liu, and Kevin Struhl. Intrinsic histone-dna interactions are not the major determinant of nucleosome positions in vivo. *Nat Struct Mol Biol*, 16(8):847–852, Aug 2009. 6, 7, 11, 17, 21, 22, 39, 45

[128] Z. Zhang, C. J. Wippo, M. Wal, E. Ward, P. Korber, and B. F. Pugh. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science*, 332(6032):977–980, May 2011. 6, 12, 22, 45, 51

Evgeniy A. Ozonov

PERSONAL INFORMATION

Nationality	Russian
E-mail	evgeniy.ozonov@gmail.com
Languages:	Russian (native speaker) English (professional working proficiency) German (basic proficiency)

SKILLS

Computer Languages	C/C++, Perl, Matlab, R, Mathematica
Operating systems	Windows, Linux
Expertise	Analysis of next-generation sequencing data, Mathematical modeling and simulation, Functional Genomics, Epigenetics, Data analysis

EXPERIENCE

Biozentrum, Swiss Institute of Bioinformatics December 2012 - Present
PostDoc *Basel, Switzerland*

Biozentrum, Swiss Institute of Bioinformatics May 2008 - November 2012
PhD candidate *Basel, Switzerland*

Schlumberger Nov 2007 - May 2008
Intern at R&D department *Novosibirsk, Russia*

Gene Networks Ltd. Sep 2007 - May 2008
Junior marketing assistant *Novosibirsk, Russia*

Institute of Cytology and Genetics SB RAS Oct 2005 - May 2008
Research assistant *Novosibirsk, Russia*

EDUCATION

PhD in Computational biology *2008 - 2012*
Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Switzerland
Title: ”Modeling nucleosome mediated mechanisms of gene regulation”
Supervisor: Prof. Erik van Nimwegen

B.Sc. in Management (second degree) *2005 - 2008*
Novosibirsk State University, Russia
Title: ”Promotion of the innovation product. The case study of a computer system for the network analysis of staff”
Supervisor: PhD Svetlana A. Kuznecova

M.Sc. in Biophysics *2005 - 2007*
Novosibirsk State University
Title: ”Modeling of morphogenesis of the Arabidopsis Thaliana embryo using cellular automata formalism”
Supervisor: D.Sc. Vitaliy A. Likhoshvay

B.Sc. in Physics *2001 - 2005*
Novosibirsk State University

PUBLICATIONS

- Evgeniy A. Ozonov and Erik van Nimwegen. **Computational analysis reveals a bias in spacing between TFBSs induced by competition between transcription factors and nucleosomes.** *In preparation.*
- Evgeniy A. Ozonov and Erik van Nimwegen. **Nucleosome free regions in yeast promoters result from competitive binding of transcription factors.** *Accepted at PLOS Computational Biology.*
- Pachkov Mikhail, Balwierz Piotr, Arnold Phil, Ozonov Evgeniy, Van Nimwegen Erik. **SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates.** *Nucleic Acids Research, 2012*
- Omelyanchuk N.A., Mironova V.V., Zalevsky E.M., Podkolodny N.L., Ponomarev D.K., Nikolaev S.V., Akberdin I.R., Ozonov E.A., Likhoshvay V.A., Fadeev S.I., Penenko A.V., Lavrekha V.V., Zubairova U.S., Kolchanov N.A. **Plant Morphogenesis: reconstruction in databases and modelling.** *Computational systems biology, publishing house of the SB RAS, pp. 539-587, 2008 (in Russian)*
- I.R.Akberdin, E.A. Ozonov, V.V. Mironova, N.A. Omelyanchuk, V.A. Likhoshvay, D.N. Gorpinchenko and N.A. Kolchanov. **A cellular automaton to model the development of primary shoot meristems of Arabidopsis thaliana.** *Journal of Bioinformatics and Computational Biology (JBCB), Vol. 5, 02B, pp. 641-650, 2007*
- Komarov A.V, Akberdin I.R, Ozonov E.A, Evdokimov A.A. **On the reconstruction of a genetic automaton on the basis of Boolean dynamic data.** *Proceedings of the fifth international conference on bioinformatics of genome regulation and structure. pp. 69-73, 2006*

CONFERENCE PRESENTATIONS

- Nucleosome mediated cooperativity between transcription factors. Basel Computational Biology Seminar, Basel 2012.
- Nucleosome mediated cooperativity between transcription factors. SIB days, Biel/Bien 2012.
- A biophysical model of genome-wide nucleosome and transcription factor binding. Biozentrum PhD symposium, Interlaken 2012.
- A biophysical model of genome-wide nucleosome and transcription factor binding. Summer school Quantitative Imaging and Modeling of Biological Processes, Amsterdam 2010.

SUMMER SCHOOLS AND SPECIALIZED COURSES

- From Data to Models in Biological Systems. *SIB/SystemX Summer school, Kandersteg 2011.*
- Analysis of Differential Gene Expression. *SIB course, Lausanne 2011.*
- Computational Analysis of Ultra-High-Throughput (UHT) Sequencing Data. *SIB course, Lausanne 2010.*
- Quantitative Imaging and Modeling of Biological Processes. *NBIC/SIB Summer school, Amsterdam 2010.*
- Machine Learning. Prof. Thomas Vetter. *University of Basel 2010.*

- Quantitative Reasoning with Biological Data. *Prof. Erik van Nimwegen, University of Basel 2009.*
- Computational Modeling and Simulation. *Prof. Mihaela Zavolan, University of Basel 2009.*
- Regulatory (Epi-) Genomics. *Otto Warburg International Summer School. Berlin 2009*
- Determinism, Stochasticity and Robustness in Biological Processes. *SIB Summer School, Lugano 2009.*
- Evolution, Systems Biology and High Performance Computing Bioinformatics. *International School of Young Scientists, Novosibirsk 2008.*