

Taking into account nucleosomes for predicting gene expression

Vladimir B. Teif^{1*}, Fabian Erdel¹, Daria A. Beshnova¹, Yevhen Vainshtein², Jan-Philipp Mallm¹ and Karsten Rippe¹

¹ Research Group Genome Organization & Function, and ² Division Theoretical Systems Biology, Deutsches Krebsforschungszentrum (DKFZ) & BioQuant, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

* Correspondence should be addressed to Vladimir B. Teif. Present address:

School of Biological Sciences, University of Essex, Wivenhoe Park, CO4 3SQ, UK.

E-mail: E-mail: vteif@essex.ac.uk

Running title: Predicting gene expression from DNA-protein binding

Keywords: gene regulation function; TF-nucleosome interference; protein binding map; combinatorial binding; chromatin.

Abstract

The eukaryotic genome is organized in a chain of nucleosomes that consist of 145-147 bp of DNA wrapped around a histone octamer protein core. Binding of transcription factors (TF) to nucleosomal DNA is frequently impeded, which makes it a challenging task to calculate TF occupancy at a given regulatory genomic site for predicting gene expression. Here, we review methods to calculate TF binding to DNA in the presence of nucleosomes. The main theoretical problems are (i) the computation speed that is becoming a bottleneck when partial unwrapping of DNA from the nucleosome is considered, (ii) the perturbation of the binding equilibrium by the activity of ATP-dependent chromatin remodelers, which translocate nucleosomes along the DNA, and (iii) the model parameterization from high-throughput sequencing data and fluorescence microscopy experiments in living cells. We discuss strategies that address these issues to efficiently compute transcription factor binding in chromatin.

1. Introduction

Predicting gene expression from mechanistic molecular considerations is a challenging subject, which currently has exact solutions only for a small number of mainly prokaryotic model systems [1-4]. However, this field is developing very fast, with many recent studies constructing bottom-up quantitative models of gene regulation [5-17]. Gene regulation in eukaryotes is much more complicated due to the dynamic organization of the DNA in chromatin, which modulates the accessibility of regulatory DNA regions to transcription factors (TFs) [18]. Furthermore, in a human organism tens of thousands of annotated genes exist whose expression levels depend on each other. The resulting large combinatorial number of possible expression patterns makes it impossible to determine these for all combinations of concentrations of all molecular players experimentally. However, the problem could be solved if one succeeds in constructing a model that predicts expression changes for individual genes as a function of TF concentrations and other input molecular parameters. This would be a highly valuable achievement for both basic research as well as medical systems biology. Accordingly, the field is rapidly expanding, and currently involves two large groups of approaches: One comprises descriptions based on biophysically formulated molecular binding models for protein arrangements along the DNA [5-17] and the other is based on bioinformatic strategies where the rules of gene expression are correlated to TF occupancies or histone modifications by learning from large datasets without knowing

the underlying molecular mechanisms [19-22]. Here we will focus on the first group of approaches, and specifically on one requirement that has to be accounted for in these types of models: the interference of TF-DNA binding with nucleosomes at regulatory genomic regions. We will review the main assumptions inherent to currently used approaches. Then we will describe a theoretical method to calculate transcription factor binding to regulatory DNA regions and the experimental methods to determine input parameters for such models. Finally, several examples of the implementation of this approach will be given.

2. Basic assumptions and concepts.

2.1. Gene expression rate is proportional to the probability of transcription initiation.

According to the classical central dogma of molecular biology, the genetic information encoded in the DNA is read by proteins to produce RNA, which is translated into proteins. This dogma has been revised multiple times during the last decades after the discoveries of reverse transcription of RNA into DNA, RNAs with enzymatic activities as well as non-coding regulatory RNAs and the identification of epigenetically determined gene expression implemented by the modifications of DNA and DNA-bound histone proteins. Thus, instead of a linear flow of information from the DNA to protein expression a complex regulatory network exists between DNA, RNA and proteins, which determines the readout of the DNA sequence [23-25]. Nevertheless, the main part of the dogma still holds true: the DNA is a carrier of genetic information, and it requires proteins to read, interpret and execute the information it encodes. The development of the final gene product depends on many regulatory events at all stages of transcription, processing and translation. In this sequence of events the first one is the initiation of transcription. Once it occurred, it still can be halted or modulated by a number of other downstream regulatory events. The critical assumption used in most theoretical works in this field is that the rate of expression is proportional to the probability of transcription initiation [2, 4, 26-31]. Although this is a significantly simplified view, it has proved to be a reasonable approach for many genes.

2.2. Transcription initiation depends on the promoter-enhancer interaction.

Transcription initiation is a complex process, with the main part being the assembly of the transcription machinery including RNA polymerase (RNAP) at the promoter [32]. The recruitment and activation of this complex depends on transcription factors (Figure 1A). TF binding is in many cases cooperative and/or involves competitive binding of several factors for the same DNA sequence [3, 33-35]. TF binding sites can be separated from each other or

clustered; they may be proximal to transcription initiation sites or at distal regulatory elements termed enhancers. In many cases promoter and enhancer regions come into contact through protein-assisted DNA looping [36], and the interaction between the pre-initiation complex at promoter and transcription factors assembled at the enhancer can be mediated by another large multiprotein complex called Mediator [37]. Thus, TF arrangement at enhancers is believed to determine transcription initiation through mechanistic interactions transmitted to RNAP. The commonly accepted view is that promoter-enhancer interactions are always by direct DNA-protein-DNA contacts (as opposed to other models that propose long-range information transmission e.g. through changes of the DNA conformation or quantum transfer [38]). In this paradigm, any advanced mathematical model of transcription initiation can be in principle supported by a corresponding mechanistic picture. The enhancer can be viewed either as a specific multicomponent structure that forms via cooperative binding of its components and is referred to as an “enhanceosome”, or a flexible “billboard”, which is a less defined structure arising from stochastic binding of a set of components [39]. In both cases its mathematical role is providing a single proxy for multiple TF signals [39-43].

2.3. The probability of transcription initiation is a function of TF bound states.

In many cases the exact mechanistic details of protein-DNA-protein complexes formed at a *cis*-regulatory module are not known. However, in principle molecular details can be determined as done for the beta-interferon enhanceosome as a prototypic example [44]. To compute the effect on gene expression, it is generally assumed that the expression of a given gene can be described by some mathematical function of TF occupancies at the enhancer and/or promoter. There have been several names for such functions in the literature, including “regulation factors” [45], “logic functions” [46], “input functions” [47], “*cis*-regulatory input functions” [48, 49] and “gene-regulation functions” (“GRF”) [4, 50], which is the term used here. Initially, GRFs were thought to be exclusively determined by the DNA sequence of the corresponding *cis*-regulatory modules [8]. However, recent studies have shown that GRFs are also strongly dependent on covalent histone modifications of nucleosomes covering the corresponding region [51]. In some cases GRFs can be defined in the form of Boolean functions of TF concentrations [47, 48], linear functions of TF occupancies at their binding sites [8] or mixed “analog” scenarios [52]. Recent studies of well-defined prokaryotic systems showed that in a general case GRFs are neither Boolean, nor linear [4, 48, 53]. For several classes of promoters where the relation of RNAP recruitment and TF binding is known it is possible to determine the nonlinear non-Boolean

gene regulation functions directly from TF binding maps [4]. The situation becomes much more complicated when one takes into account the nucleosomal organization of DNA in chromatin in eukaryotes (Figure 1B). In this case, some nucleosomes need to be removed or repositioned to allow transcription initiation complex assembly. Thus, the GRF becomes also dependent on the nucleosome states [54]. Even with this correction, several recent studies have challenged the classical assumption that expression of a gene is correlated to the corresponding TF occupancy [55]. Rather, it was proposed that in some cases GRFs are better correlated to the changes in histone modifications than to the changes in TF occupancies [17]. A special study devoted to the effect on gene expression of TF arrangement versus histone modifications has shown that TF occupancies are responsible for short-range effects (e.g. one gene) whereas histone modifications act more globally (genomic locus including several genes) [56]. In any case, histone modifications are thought to work predominantly by recruiting specific proteins, so that the GRF would be still determined by the protein-DNA binding state of a given regulatory module. Last but not the least, it is noted that the GRF concept assumes gene expression to be at least to some extent deterministic and not purely stochastic. The latter point might seem obvious at the macroscopic level since organisms develop according to a well-defined program. However, at the microscopic level this assumption is not strictly fulfilled, and relative contributions of stochastic/deterministic processes still have to be evaluated quantitatively [57].

2.4. Binding maps cannot be measured for all time points, and have to be calculated.

Current high-throughput techniques allow measuring genome-wide binding maps for a single protein in a given cell type and cell state. In general, the binding maps determined for different cell types do not coincide. For example, recent studies of genome-wide binding of an insulator protein CTCF in mouse embryonic stem cells and mouse embryonic fibroblasts have revealed that only ~ 30% of binding sites are identical between these two cell types [58]. The TF binding maps depend on the protein concentration, active nucleosome repositioning and changes in large-scale chromatin accessibility. In all three cases we have to account for TF competition with each other and with other molecules for DNA binding. Classical types of competitive binding may involve competition for overlapping and non-overlapping binding sites, formation of DNA loops and multilayer structures [59, 60]. In addition, molecular motor activities require the introduction of a non-equilibrium component, which still can be integrated in the frame of quasi-equilibrium thermodynamic models [61].

2.5. TF-DNA binding maps are calculated for equilibrium conditions.

The cell nucleus is a very crowded environment and equilibration times can be as large as hours [62]. Furthermore, many DNA binding proteins can undergo conformational changes that are driven by the hydrolysis of ATP and act as molecular motors against the thermal equilibrium [63, 64]. Nevertheless, most current methods for calculation of TF-DNA binding maps use the assumption that the binding map can be determined from the thermodynamically preferred protein-DNA contacts and the thermodynamic competition between different protein species [1-3, 65]. The use of this assumption is justified by the following considerations: (i) TF binding events frequently happen on a time scale of seconds [66, 67]. Thus, TF DNA occupancy is expected to be in a quasi-equilibrium although the cell's state might change on the hour scale, e.g. during progressing through the cell-cycle. (ii) Similarly, the ATP-dependent activity of chromatin remodelers that could translocate nucleosomes at promoters or enhancers would lead to a steady state of nucleosome positions at a given point of time in the cell that can be represented by equilibrium [61, 68]. It is also noteworthy that only a very small fraction of nucleosomes appears to be translocated in the absence of DNA replication or DNA repair [64]. (iii) One can also think of a collective equilibrium in an ensemble of many identical cells [4]. In this case, the binding map derived from the equilibrium assumption would represent an average pattern characteristic for many instances of the cell at different time points.

3. Calculation of TF binding maps in chromatin.

3.1. Integrating nucleosomes in thermodynamic TF binding models.

The nucleosome consists of 145-147 bp wrapped around the histone octamer core [69]. Without ATP-dependent remodelers, the nucleosome residence time is in the order of 1-2 hours, which is much larger than that for a typical transcription factor [70]. The energy of DNA-histone octamer interaction (~ 1 kT per bp) is also much larger than the energy of binding for a typical TF [71, 72]. From this perspective, the nucleosome can be viewed as almost immobile with respect to TF binding. Mathematically that would be described by a structure that always protects 147 base pairs from binding to other proteins. A given site on the DNA would be either nucleosome-free, or inside the nucleosome. However, the nucleosome is actually quite a dynamic structure [73-78] (Figure 2). Some of the four histone dimers can be lost leading to partial nucleosome disassembly [60, 79]. Alternatively, DNA can partially unwrap from the histone octamer due to a variable number of protein-histone bonds [72, 80, 81]. The nucleosome unwrapping model suggests two possible effects: first,

transcription factors can access the DNA inside the nucleosome, especially close to the nucleosome entry/exit site, and second, nucleosomes can invade the territories of each other. Both of these effects have been observed experimentally [75, 82]. Furthermore, this model was shown to be quantitatively consistent with *in vitro* measurements of DNA accessibility and DNA positioning [72].

3.2. Formulating one-dimensional DNA lattice models.

Genomic DNA is packaged into chromatin in a complex 3D structure, which is still poorly understood [83]. In particular, there are multiple contacts between distant genomic regions mediated by specific DNA-protein-DNA interactions [84, 85]. Mathematically speaking, this property can be described as a fractal dimension [86-88]. Nevertheless, for many problems involving a single genomic region it is useful to consider the DNA as a linear molecule, characterized by a single 1D coordinate numbering the nucleotides or base pairs (bp) along the genome. Each nucleotide can contribute to a potential protein-DNA contact [13, 89, 90]. Mutations changing distances between TF binding sites at *Drosophila* enhancers by several bp lead to different phenotypes [13], and sites of single-nucleotide polymorphism (SNP) affect differential TF binding at regulatory regions [91]. In one-dimensional models the DNA is considered as a lattice of base pair units numbered by index n (Figure 3A). Each DNA unit can be in one of several states determined by the reversible protein binding as is typical for Ising [92] and Markov chains [93]. We consider f types of proteins, which can competitively bind DNA depending on the protein type g , $g = (1, f)$. Macroscopic protein-DNA binding constants $K(n,g)$ determined by the energy of protein-DNA binding depend on the position of the binding site start along the DNA n and protein type g . For each protein-DNA complex, it is possible to enumerate base pairs within the binding site by index h with respect to the start of the canonical binding site n , and correspondingly distinguish microscopic binding constants $k(n,g,h)$ corresponding to individual protein-DNA bonds. The product of all microscopic binding constants $k(n,g,h)$ for a given complex gives the macroscopic binding constant $K(n,g)$. In principle, any DNA base pair in the sequence may be considered to represent the start of a potential binding site for a given protein. Proteins g_1 and g_2 can interact with each other depending on the distance j along the DNA with a potential $w = w(j, g_1, g_2)$. Proteins are characterized by their corresponding binding site sizes on the DNA, $m = m(g)$. It is frequently assumed that the binding site size for a given protein type is constant, e.g. a protein covers 10 bp upon binding to the DNA and protects these 10 bp from binding of other proteins. However, this is just a special case of a more general situation when each binding site is

characterized by h_1 unbound bp from the left and h_2 unbound bp from the right end of the binding site as shown in Figure 3. This model becomes particularly important for large protein-DNA complexes such as the nucleosome where it is known that partial unwrapping of DNA from the histone octamer occurs spontaneously [72, 94].

3.3. Mathematical algorithms to solve 1D lattice models.

The aim of constructing 1D lattice models is to be able to predict probabilities of all bound protein-DNA configurations. Only few of these configurations are of particular interest, but to be able to calculate their probabilities one has to know the probabilities of the others. In the pseudo-equilibrium approximation, each bound configuration i can be given a weight which exponentially depends on its free energy, $\exp(-\Delta G_i / k_B T)$, where ΔG_i is the energy change corresponding to a given configuration of protein arrangement along the DNA, k_B is the Boltzmann constant, and T the absolute temperature in Kelvin. The sum of weights of all possible configurations is called the partition function. The straightforward way to calculate the partition function is via sampling through all possible states of the system. This can be done analytically for simplified systems e.g. assuming non-specific binding [95, 96] or numerically for realistic systems confined to short DNA lattices [97, 98], or systems with a small number of known discrete binding sites of a few transcription factors [99], such as the λ -switch [27, 28] or the Lac operon [29-31]. However, if both sequence-specific and nonspecific binding to overlapping DNA sites is taken into account, calculations for DNA regions longer than 30 bp are not feasible using this method with currently available computers [100], and special methods are needed to accelerate calculations [101]. These include the binary variable method, combinatorial method, generating function method, transfer matrix method and dynamic programming approach as reviewed elsewhere [60, 94]. Many currently used approaches are based on dynamic programming algorithms for historic reasons [94]. The dynamic programming algorithms were initially developed in the 1970s independently by DeLisi in USA and Gurskii and Zasedatelev in USSR [102-105] and for some time used only by specialists interested in theoretical aspects of such models [106-108]. Then they were almost forgotten, and recently have become very popular again in applied science, particularly in the nucleosome positioning and TF binding fields [8, 12, 109-117]. A first dynamic programming method to calculate TF-DNA binding taking into account the possibility of partial nucleosome unwrapping was developed in our recent publication [94]. In the dynamic programming approach, the partition function Z for a DNA of length N can be calculated recurrently if partition functions for smaller lattices are known using the

dynamic programming algorithm [94] (see Appendix). Alternatively, the transfer matrix method [72], or binary variable methods could be used for simpler systems where nucleosome unwrapping does not need to be taken into account [60]. As a result, one gets the partition function Z , which allows calculating the probability $P(n, g, h_1, h_2)$ that a protein of type g is bound starting at site n , leaving on the left and right sides correspondingly h_1 and h_2 unbound contacts with respect to its canonical binding length $m(g)$. The critical parameters in the modeling is $c_0(g)$, the free concentration of the protein of type g , the binding constant of the corresponding protein, and protein-protein interaction potentials $w(j, g_1, g_2)$. To account for the possibility of partial nucleosome unwrapping, the macroscopic binding constant K^* for the protein (or the histone octamer) whose first contact with DNA starts at position n is defined in dependence of the number of formed bonds as a function of $n, m(g), g, h_1, h_2$:

$$K^* = K(n, g, h_1, h_2) = \prod_{h=h_1+1}^{m(g)-h_2} k(n+h-h_1-1, g, h), \quad (1)$$

where k is the microscopic binding constant for the protein-DNA bond at position i with respect to the start of the completely bound protein binding site. In practice, it is impossible to determine all probabilities $P(n, g, h_1, h_2)$ experimentally. What is usually reported in the experiments is the occupancy of a given base pair by a given protein type. The probability that a specific DNA base pair is occupied by the protein of type g is

$$C(n, g) = \sum_{g=1}^f \sum_{h_1=0}^{m(g)-1} \sum_{h_2=0}^{m(g)-h_1-1} \sum_{i=n-m(g)+h_1+h_2+1}^n P(i, g, h_1, h_2) \quad (2)$$

Many experimental papers also report the value of the nucleosome dyad density as a function of position along the DNA. This is equivalent to the probability that a given DNA unit is bound by the middle of the protein according to

$$P_{center}(n, g) = \sum_{g=1}^f \sum_{h_1=0}^{m(g)-1} \sum_{h_2=0}^{m(g)-h_1-1} P(n - \text{Int}[(m(g) - h_1 - h_2) / 2], g, h_1, h_2), \quad (3)$$

where Int is the integer part of the corresponding expression. These equations allow calculating TF binding maps in the presence of nucleosomes and taking nucleosome

unwrapping into account. An implementation of this algorithm in a program called "TFnuc" will be made available online at <http://generegulation.info>. While this is a powerful method, its application requires the proper choice of input parameters (binding affinities, concentrations, interaction potentials) to yield meaningful results. The next chapters discuss strategies for obtaining these parameters.

4. Determining binding affinities from high-throughput sequencing experiments

4.1. Estimating relative TF binding constants.

Recent developments in high-throughput microarray-based and sequencing-based methods allow measuring protein binding maps for a complete genome in a single experiment [118]. Having such an experimental binding map, one can extract protein-DNA sequence preferences. Some proteins are more specific, recognizing just a single motif and some minor variation of it. For this class of proteins binding affinities are usually characterized by position weight matrices (PWM). It is assumed that each nucleotide within the binding site adds an independent contribution to the binding energy [119-121]. For many transcription factors, position weight matrices are available via databases such as FlyTF [122], JASPAR [123] and TRANSFAC [124]. Several methods exist to convert PWMs into protein binding affinities [116, 125-127]. However, some TFs can recognize many different motifs, and binding preferences can also be influenced by dependencies of neighboring nucleotides. In this case storing binding affinities in the form of weight matrices becomes an ineffective strategy. Until recently, TF binding motifs have been commonly determined *in vitro* by protein-binding microarrays, but this method is limited by the number of represented sequences on the microarray, which are usually not longer than 10 bp [128]. In general, microarray-based methods such as SELEX (systematic evolution of ligands by exponential enrichment) only return relative affinity values [129]. Such values, e.g. from the HTPSELEX database [130], can be technically stored in a genome-wide affinity profile file and sequentially read as input to feed them into Eq. 1 calculating the corresponding binding constant for each window of length $m(g)$ on the DNA. With high-throughput SELEX (HT-SELEX) [131], DNA fragments with a randomized 10 bp sequence are incubated with TFs and then protein-DNA complexes are purified and eventually sequenced using deep sequencing protocols [131]. When comparing the number of initial DNA fragments with TF-enriched sequences after sequencing, one can calculate the probability of binding to a particular 10 bp region. Using certain assumption, it is possible to calculate the free energy for each sequence depending on read statistics and estimations for the energy contribution of each nucleotide using the PMW energy model [131]. A recent study showed that absolute dissociation constants of fluorescently labeled TFs to immobilized DNA clusters can be

obtained from next generation sequencing data by plotting the signal intensity of the TFs with increasing TF concentrations [132]. This method can also resolve interdependencies of nucleotides for transcription factor binding, which is valuable information when calculating transcription factor binding probabilities. To identify binding sites that are functionally relevant *in vivo*, a ChIP-seq analysis of TF binding sites in a cell is informative (Fig. 4A). Binding maps derived by this method account for the chromatin organization, since nucleosome positions are implicitly taken into account [133]. When the resultant TF binding maps are compared to nucleosome binding maps determined by MNase-seq for the same cell type it becomes apparent that many TFs are preferentially bound to the linker DNA regions between nucleosomes. Importantly, extracting enriched binding sites with peak calling algorithms leads in many cases to the loss of information about the occupancy level of individual peaks. Therefore, the weighted sum of enriched fragments has been proposed as a measure for relative TF binding affinity and it has been shown that gene expression predictions can be made with improved precision when the weighted sum also depends on the proximity to a TSS [134]. Particularly accurate *in vivo* binding sites at single nucleotide resolution can also be obtained with the ChIP-exo method [135]. Applications of this method showed that not all consensus sequences previously described in *in vitro* experiments were occupied. Clusters of poor consensus sequences were also bound and utilized by TFs to initiate transcription [135]. It has to be noted that ChIP-seq based methods rely on the availability of good antibodies, and both direct binding of transcription factors to DNA and indirect association via other factors/complexes are captured. Importantly, the common assumption that ChIP-seq peak heights reflect relative binding affinity might not always be true, since other factors may influence the amount of immunoprecipitated DNA, including the formation of protein complexes that have different exposure of the epitope used in ChIP-seq, as well as the level of chromatin packing affecting the representation of a given genomic fragment in the input material after digestion [136]. Thus, a combination of *in vitro* and *in vivo* methods is needed to retrieve quantitative data of functionally relevant direct and indirect binding sites.

4.2. Experimental determination of cell-type dependent nucleosome occupancies.

A similar strategy as with TF-DNA affinities can also be applied to estimate the affinity of the histone octamer to any DNA sequence [111, 137-139]. Several web servers already exist for calculating affinities of the histone core particle to an arbitrary DNA sequence [111, 140-143]. Recent advancements in high-throughput sequencing methods allowed genome-wide mapping of individual nucleosomes at single base pair resolution [144, 145], with yeast serving as a model system for the initial pioneering studies [111, 138, 146]. Further studies showed that nucleosome positions in different cell types of the same organism differ [147-154]. As with any protein-DNA binding, nucleosome positioning is determined by several

contributions including the intrinsic histone-DNA preferences, competition with non-histone proteins for DNA binding, and the action of ATP-dependent molecular motors [61]. The relative roles of these contributions are still under discussions, but all of them seem to be relevant. A typical experiment for the determination of genome-wide nucleosome positions is currently based on chromatin extraction from the cell nucleus, digestion of chromatin with MNase (or alternatively, by sonication combined with exonucleases) to obtain mononucleosomes, followed by the purification from proteins and RNA, and subsequent submission of obtained segments of nucleosomal DNA for high-throughput sequencing (Figure 4B). Paired-end sequencing is the method of choice since it allows exact mapping of both ends of the nucleosomal DNA to the reference genome without any assumptions. The results of the MNase-seq experiment usually yield a somewhat fuzzy picture. A nucleosome is almost never strictly positioned at exactly the same position in all cells of the same type due to cell heterogeneity and due to the intrinsic nucleosome property to “breath” by unwrapping/rewrapping the DNA at the ends. Accordingly, the most informative parameter to describe such a nucleosome distribution is the nucleosome occupancy, i.e. the probability that a given DNA base pair is occupied by the nucleosome. Obtained nucleosome occupancy profiles strongly depend on the level of chromatin digestion, which can be used as a titration parameter [155]. Finally, nucleosome occupancy at position i can be determined from the experimental data according to Eq. 4 (Figure 4B):

$$C(i) = \sum_j I(n_j \leq i \leq s_j), \quad (4)$$

where I is an indicator function defined as follows: $I(\text{condition})=1$ if condition is satisfied, 0 otherwise; index j numbers individual nucleosome reads (current generation high-throughput experiments can provide up to two hundred millions of paired-end reads per sequencing run). The parameters n_j and s_j correspond to the mapped start and end of each individual nucleosome read after paired-end sequencing. Experimental data obtained with the help of Eq. 4 can be then normalized to the total number of reads per base pair and directly compared to the theoretical occupancy distribution calculated by Eq. 2.

If exact borders of nucleosomes are not known (e.g. due to the use of single-end sequencing which is less-expensive than paired-end sequencing) one usually determines the nucleosome dyad distribution, i.e. the probability that the nucleosome centre is at a given position along the genome. The latter can be directly compared to the theoretically calculated

distribution given by Eq. 3. In the absence of information about exact nucleosome boundaries, nucleosome start site maps (or dyad maps) can be converted to nucleosome occupancy maps assuming that the nucleosome consists of m base pairs and cannot unwrap, using the following approximation [61]:

$$n \leq m, C(n) = \sum_{k=1}^n P(k) \quad (5)$$

$$n < m, P(n) = C(n) - \sum_{k=1}^{n-1} P(k), P(1) = C(1) \quad (6)$$

$$m < n < N, C(n) = \sum_{k=n-m+1}^n P(k) \quad (7)$$

$$m - n \leq N - m + 1, P(n) = C(n) - \sum_{k=n-m+1}^{n-1} P(k), \quad (8)$$

where the probability that the DNA unit n is covered by a nucleosome is referred to as $C(n)$ and the probability that a nucleosome starts at a DNA unit n as $P(n)$. It is noted that the single-end sequencing approach introduces additional errors that arise because a certain length of the nucleosomal DNA has to be assumed. Although it is well established from crystal structure analysis that the nucleosome core particle contains 145-147 bp the fragment length obtained by MNase digestion is much more heterogeneous and typically ranges from 120 to 180 bp, depending on the digestion conditions. In addition, the linker histone H1 binds to the DNA at the entry-exit site of the nucleosome and protects an additional ~20 bp. Accordingly, the footprint of a nucleosome with bound H1 is typically larger than 160 bp.

5. Absolute chromatin binding affinities derived from fluorescence microscopy based methods in living cells.

As described above, high-throughput methods can yield genome-wide occupancy profiles for nucleosomes, TFs or other chromatin proteins. However, these profiles usually provide only relative binding affinities. Thus, the information on the competitive binding of two proteins for the same sequence cannot be derived from this type of data. Moreover, phenomena like cooperative binding or stabilization of chromatin loops that influence the GRF as described above depend on the interaction strength between chromatin-associated proteins, which cannot be derived from occupancy profiles. As discussed in several excellent reviews molecular details and interaction parameters can be obtained from *in vitro* studies [156-159].

These provide an approach to quantify the competitive binding of two proteins to DNA or reconstituted nucleosomes and to each other. However, although these experiments provide valuable information it is usually very difficult to relate these data to the situation in the highly crowded environment of the cell nucleus and a chromatin organization that is certainly different from that of an in vitro reconstituted nucleosome. Thus, the relevant binding parameters need to be determined in living cells. Strategies to accomplish this via quantitative fluorescence microscopy based techniques will be discussed in the following. As depicted in Figure 5, several complementary approaches exist to measure the interaction between fluorescently labeled proteins or labeled proteins and chromatin. Compared to biochemical methods, these approaches are non-invasive and can work without perturbing the cell. The caveat, however, is that they require expression of fluorescently tagged proteins, which in some instances might lead to interaction affinities that are different than those of the endogenous proteins. Thus, it needs to be confirmed that the fluorescent tag has no effect on the properties to be measured. The choice of the specific method depends on the mobility and the size of the proteins under study. To determine the affinity between a protein and a DNA sequence of interest one would ideally measure this interaction on a single-molecule level. However, this is on the one hand not always possible and on the other hand very laborious if lots of different DNA sequences are to be measured. Thus, an alternative strategy is to determine the affinity of the protein to an average DNA sequence by measuring at randomly chosen positions within the cell. This value together with the relative affinities from high-throughput experiments allows for estimating the absolute affinity profile of the protein. In addition, the interaction between transient chromatin-binders is relevant for calculating the binding cooperativity they exhibit on the DNA as well as for estimating their propensity to stabilize higher-order chromatin structures. Such structures may include chromatin loops at promoter regions or multilayer structures that have a direct impact on the GRF as mentioned above. The following methods are the most prominent ones used to determine such protein-DNA-protein interactions in living cells.

5.1. Fluorescence Resonance Energy Transfer (FRET)

Fluorescence Resonance Energy Transfer (FRET) is a convenient tool to test whether two proteins interact directly with each other. Non-radiative energy transfer in FRET occurs between two spectrally suitable fluorophores with distinct fluorescence excitation and emission characteristics. Upon excitation of the “donor”, the absorbed energy can be emitted either via fluorescence emission or via non-radiative FRET if a suitable “acceptor” is present within a distance of up to 10 nm. In this case, the acceptor is excited by the donor and emits light according to its characteristic emission spectrum. The efficiency of this process is inversely related to the 6th power of the distance between the two fluorescently labeled

proteins so that interactions on the molecular level can be detected. If no FRET is observed, the interpretation is not straightforward since negative results can occur for different reasons. These include a too large spatial distance or a mostly perpendicular orientation of the two fluorophores' transition dipole moments. To extract binding affinities using FRET, the size of the interacting and non-interacting pools as well as the FRET efficiency have to be determined. Since the latter parameter depends on the orientation of the two fluorophores, i.e. the composition and geometry of a complex, it is difficult to construct appropriate controls for quantitative interpretations. However, estimates for relative affinities can be readily obtained both in live cells and in vitro [160, 161]. Moreover, FRET measurements in conjunction with microinjection provide access to additional protein interaction parameters like the association rate [162]. Importantly, FRET can be used to study the interaction of a transcription factor or another mobile protein with chromatin. To this end, chromatin is labeled via incorporation of a fluorescent histone [163] or by labeling the DNA with a fluorescent dye [164]. As discussed above, high FRET efficiencies are indicative of strong interactions and allow for the calculation of relative binding affinities to "average" chromatin. There are conceptually two possibilities to assess the interaction with a particular DNA sequence or chromatin region in living cells. If the sequence can be easily located in microscopy images, e.g. the murine pericentric sequences in the dense chromocenters of mouse cells, the measurement can simply be performed at the desired location. Otherwise, i.e. for all non-repetitive and non-macroscopic sequences, a sensor protein binding to the sequence/region of interest could be used as FRET counterpart. A critical requirement for such an experiment would be that the sensor does not directly bind to the protein of interest.

5.2. Fluorescence Recovery After Photobleaching (FRAP)

FRAP is a method to measure the mobility of a protein moving in the cell. Since binding interactions with the rather immobile chromatin network reduce a protein's mobility, mobility and binding strength are inversely correlated. There are sophisticated reaction-diffusion models to extract pseudo-association and dissociation rates from FRAP recovery curves [165, 166], which can be used to estimate the binding affinity if the substrate concentration is known. Due to the inherently limited spatial resolution, such experiments typically yield the interaction behavior with an average site on chromatin. However, if macroscopic amounts of repetitive binding sites are used FRAP can also measure the interaction between a transcription factor and a distinct DNA sequence [167, 168]. Both types of measurements can be useful to convert occupancy profiles into absolute affinities. For example, FRAP was used to determine the dissociation constant of the glucocorticoid receptor at a tandem array of mouse mammary tumor virus promoter sites, yielding a value of about 100 nM [169].

5.3. Fluorescent Two-Hybrid Assay (F2H)

A convenient way to study the interaction of two nuclear proteins is the Fluorescent Two-Hybrid Assay [170] that has been applied in a recent study to dissect protein interactions at telomeres [171]. One of the two proteins tagged with Green Fluorescent Protein (GFP) is recruited to a macroscopic array of *lacO* sites on the DNA that can readily be identified within a microscopy image. The interaction with a second protein tagged with Red Fluorescent Protein (RFP) is read-out by testing for colocalization at the array in both color channels. In the presence of an interaction, both proteins are at the array; in the absence of an interaction (or for very weak interaction) only the recruited protein is detected. The main advantage of the assay is that it can easily be implemented using standard microscopy hardware and that it typically does not give false-positive results if spectral cross talk is avoided. However, *lacO* arrays are typically constructed with a high density of binding sites, accommodating large numbers of recruited proteins in direct proximity. If one of the proteins of interest is incorporated into large complexes, it has to be ensured that the steric requirements are compatible with the local constraints of the protein-bound *lacO* array. To obtain semi-quantitative information from F2H experiments, the fluorescence intensity at the array and in the rest of the nucleus can be used to estimate the size of the free and bound fractions, from which the affinity can be deduced if the endogenous protein concentrations are known. A prerequisite for such an analysis is to exclude quenching or saturation effects, which might not always be trivial. Although F2H has to our knowledge not been used for quantitative studies so far, it should be suitable for determining the binding affinity between two proteins if the conditions mentioned above are met.

5.4. Fluorescence Correlation Spectroscopy

An elegant approach to detect the interaction between proteins is Fluorescence (Cross) Correlation Spectroscopy with a one or a two-color label (FCS/FCCS). In particular the ability to detect low concentrations of multimeric complexes in solution makes FCCS an attractive method. It does not require recruitment to an artificial array but works with two species of fluorescently proteins only [172]. As opposed to FRET, the orientation of the two proteins in their complex(es) has no impact on the measurement. FCCS relies on the correlation of the presence of the two labeled proteins in the microscope's focus over time, i.e. it measures if both proteins enter or leave the focus together or independently. Since both the total and the interacting protein species are measured with single molecule sensitivity, conclusions about the affinity of the complex can be made if the endogenous concentrations of the proteins are known. This is a major advantage with respect to many other methods since quantitative evaluation typically requires an extensive calibration procedure, which is not the case here. However, appropriate controls have to be used to account for aberrations in the optical setup

or maturation problems of the fluorophores [173]. As an example, the affinity between the small Rho-GTPase Cdc42 and the actin-binding scaffolding protein IQGAP1 was successfully measured in living zebrafish embryos and cultured mammalian cells using FCCS [174].

5.5. Single Particle Tracking (SPT)

Another way to exploit the connection between the mobility and the binding affinity of a transcription factor is single particle tracking. Here, the protein of interest is expressed in very low concentrations and individual molecules are imaged and followed over time. Under certain conditions, one can deduce the binding affinity to the target site from such time-series, as shown for the lac repressor binding to its lac operator target site [175]. This approach works best for low concentrations of transcription factors and binding sites since the spatial resolution is inherently limited by the microscope.

6. Insight in molecular details of transcription regulation from model systems.

6.1. Lessons learned from the prokaryotic world.

Many paradigms in this field have been developed using a limited number of relatively well-understood model systems. In the studies of prokaryotes, the phage λ -switch [27, 28] and the Lac operon [29-31] served as model systems for many years. The λ -switch model revealed that direct competition between the two transcription factors CI and Cro and RNAP can regulate activation/repression of two neighboring promoters P_R and P_{RM} that determine the fate of the *E. coli* bacteria invaded by bacteriophage λ . Most energetic parameters characterizing protein-DNA and protein-protein binding at the λ -switch promoter region have been determined experimentally, which allowed constructing many quantitative models. Interestingly, our understanding of this well-defined system is being constantly refined. One refinement was the discovery of a DNA loop between the P_R - P_{RM} region and a distant P_L region. The loop energy was measured and the structure of CI multimer holding the loop was characterized, allowing the quantitative agreement with the experiment of an updated model [176, 177]. Another refinement to the classical λ -switch model came when the significant role of nonspecific protein binding was pointed out in addition to the specific binding to their recognition sites [178]. An additional refinement to the model was the introduction of the distant-dependent interference between RNAPs bound to adjacent promoters. This interaction was quantified with the help of the long-range interaction potential $w(j, RNAP, RNAP)$, and it appeared that this interaction shapes the GRF of a given cis-regulatory module to make it more digital-like [4, 59]. Yet, studies of the λ -switch indicate that we are

perhaps still missing some details in the complete understanding of this system [179]. Similarly to the λ -switch, the classical Lac repressor system also taught as important lessons about regulation of transcription initiation through TF competition and cooperativity. Recent conceptual insight was obtained from this system with respect to long-range cooperativity between DNA-bound TFs due to protein-induced DNA looping [180]. It was known for a long time that protein binding sites separated by $n \times 10$ bp along the DNA appear on one side of the double helix and therefore exhibit higher cooperativity in protein binding. In addition, this study also characterized the intermediate distances so that it can now be quantified with a continuous distance-dependent potential $w(j, g_1, g_2)$.

6.2. TF interference with nucleosomes at eukaryotic regulatory regions.

Similar to the prokaryotic studies, in eukaryotes there are also several systems that have been well-defined at the molecular level and studied for a long time, e.g. the IFN- β enhanceosome formation [181], Epstein-Barr virus promoter activation [182] and yeast PHO5 promoter [54]. The Epstein-Barr virus promoter activation was perhaps the first theoretical model that addressed transcription initiation in a eukaryotic system explicitly considering multiprotein combinatorial assembly [182], but this model did not take nucleosomes into account. The difference from the prokaryotic analogues was simply in the number of TF binding sites that are involved in the cooperative interaction with the pre-initiation complex. The introduction of nucleosomes in this type of models is illustrated by the recent study of the yeast PHO5 promoter [54]. In this case, binding site occupancy by the nucleosome was considered in a binary way: as occupied or not occupied. This allowed getting quantitative agreement with the experimentally measured expression for this system. However, as noted in many molecular studies, nucleosome removal is usually not a binary process, with nucleosomes either being continuously unwrapped [72, 80, 81] or moved by a remodeler along the DNA in small steps such as 10 bp [61, 183], so that a binary description is in general not adequate. Accordingly, several theoretical models have been proposed to include continuous nucleosome competition with TFs in the description of transcription initiation [59-61, 72, 80, 81, 184, 185]. The practical use of such models is currently limited by the absence of suitable biological systems, which are characterized well enough to set input values for Equations 1-4 with the affinities and concentrations for histone and non-histone proteins. Mathematical approaches including fitting of the missing values can help [8, 12], but should be used with care due to avoid over-fitting (the more unknown parameters are in the model, the easier it is to find a satisfactory parameter set, but it is much more

difficult to derive a biologically relevant physical model). On the other hand, the continuous increase of high-throughput genome-wide datasets for a limited set of model systems (e.g. *Drosophila* embryonic development, human T-cell activation, mouse embryonic stem cell differentiation) promises more well-defined genomic modules suitable for the complete bottom-up description. Meanwhile, it is instructive to decipher some general mechanisms that are characteristic for eukaryotes and have not been observed previously in the prokaryotic studies. One such mechanism is the cooperativity between transcription factors mediated by nucleosomes.

In 2008, Segal and coauthors concluded their computational analysis of the experimental *Drosophila* development data with the following statement: “We do not know how [transcription factor binding] cooperativity is achieved mechanistically – by homotypic protein–protein interactions, transcriptional synergy, or perhaps competition with nucleosomes – but the similar narrow range within which the clustering occurs for most factors suggests a general common mechanism” [8]. The source of this cooperativity still has not been identified. The characteristic distance for such interactions is ~50 bp, and therefore several computational models just use the corresponding interaction potential, derived from fitting the experimental data [12]. An attractive possibility suggested by this characteristic length is that this cooperativity is mediated by the nucleosomes [186]. We recently developed a quantitative model for TF-nucleosome interference using the concept of partial nucleosome unwrapping that is described by equations 1-4, and tested it on the experimental dataset by Fakhouri et al [13]. This dataset is particularly interesting because the authors have looked at enhancer-promoter cross-talk involved in *Drosophila* embryonic development by varying the distance between binding sites for a repressor/activator transcription regulation module. The binding sites themselves were not altered, and the promoter regions remained intact. The authors observed that gene expression followed a complex nonlinear dependence as a function of the distance between the repressor and activator binding sites: the repressor efficiency was high at small separations ~5 bp, low around 30 bp, reached a maximum at 50-60 bp, and decreased at larger distances. Such distances are large enough to rule out direct protein-protein interactions. On the other hand, these distances are too small to be accounted for by usual DNA looping, which has a characteristic length of ~500 bp [187]. Moreover, the experimental dependence did not reveal a 10-bp periodicity characteristic for prokaryotes [180]. Therefore, to explain distance-dependent cooperativity at these distances we are only left with nucleosomes or other complexes geometrically resembling the

nucleosome such as the enhanceosome. Indeed, our calculations showed that the nonlinear distance-dependent behavior can be quantitatively explained when TF-nucleosome competition is considered and the nucleosome unwrapping is taken into account [81]. This mechanism would explain evolutionary clustering of TF binding sites at the regulatory regions with characteristic 60-80 bp distances [188, 189]. Of course there is still no direct proof that this mechanism is really in operation and more direct molecular experiments are required to solve this issue [190].

6.3. ATP-dependent nucleosome repositioning.

Another important feature specific to eukaryotic systems is the contribution of ATP-dependent chromatin remodelers to gene regulation. Pioneering high-throughput experiments in yeast showed that genomic nucleosome positions are highly correlated with preferred nucleosome positions on the same DNA sequences *in vitro* [111]. This suggested that nucleosome arrangement *in vivo* might be primarily governed by intrinsic preferences of histone octamers to DNA at a thermodynamic equilibrium. Subsequently, it was shown that an ATP dependent activity, most likely that of chromatin remodelers, is needed to establish the nucleosome positioning pattern found in the cell and that this can override DNA intrinsic positioning [191]. Nucleosome occupancy profiles around genomic barriers such as the insulator CTCF proteins or transcription start sites have pronounced oscillatory patterns [192-194]. These are very typical for reversibly binding ligands equilibrated in the presence of a boundary on the DNA [97, 195]. Using the assumption of reversible equilibrium binding of histone octamers allowed quantitatively explaining oscillatory nucleosome patterns around genomic barriers without the need of introducing ATP-dependent chromatin remodelers [196, 197]. This has led to the view that the oscillatory nucleosome occupancy patterns around genomic barriers arise simply due to statistical positioning [195]. However, as we have demonstrated theoretically, very similar periodic oscillatory of nucleosome occupancy around a boundary can also be the result of the activity of nonspecific nucleosome translocations due to chromatin remodeling activity [61]. By taking into account both the sequence-specific histone preferences and ATP-dependent remodeler activities it is predicted that one role of nonspecific chromatin remodelers is to distribute nucleosomes with equal spaces. Several other models have been developed that account for the remodeler's ability to evenly space nucleosomes [198]. It is noted that the oscillatory nucleosome patterns around transcription start sites observed *in vivo* were absent for the same DNA sequences *in vitro* in the absence of chromatin remodelers or the absence of ATP, while addition of remodelers plus ATP re-

established the oscillatory nucleosome pattern [191]. Thus, remodeler activity appears to be essential for nucleosome positioning in the cell. Methodologically, remodeler activity can be taken into account either (i) as a dynamic redistribution according to the remodeler rules of the equilibrated binding map determined by the nucleosome/TF competition, or (ii) as a cell-type dependent refinement to the intrinsic histone-DNA affinities, followed by the equilibration of histone octamers with competitively binding TFs [61]. The first option is biophysically better defined, but it requires the definition of remodeler activity rules for different classes of remodelers. This is difficult to do even with sophisticated experiments like knock-out or recruitment of specific remodelers [152, 153, 199]. The second option provides less mechanistic insight, but allows effectively characterizing different cell states by histone-DNA preferences, which already take into account remodeler action, and then proceed with the calculation of TF binding maps according to Equations 1-7. Future studies will show which of these methods is more suited for the quantitative description of eukaryotic gene regulation.

7. Conclusions

In higher eukaryotes, specific cell types and tissues are established from the same DNA via different protein-DNA binding patterns that determine gene expression. These patterns correspond to distinct chromatin states that are maintained via a complex epigenetic network that includes DNA methylation and histone modifications and can be transmitted through cell division. Accordingly, it is essential to consider the chromatin state for the computation of TF binding maps at regulatory elements like enhancers and promoters to predict gene expression. As discussed here, an essential step towards this goal is to include the nucleosome in the calculation of TF binding maps at thermodynamic equilibrium conditions. Protein concentrations, binding affinities and long-range interaction potentials are needed as input parameters for such calculations (Figure 6). Here, we have discussed how these parameters can be obtained using high-throughput sequencing experiments in combination with fluorescence microscopy in living cells. Being able to calculate TF occupancy in the presence of nucleosomes more accurately is an important advancement. However, it is also clear that in general it is not possible to reliably predict gene expression from molecular binding events with the current experimental datasets and theoretical methods. Taking into account the ever-growing amount of experimental data, it seems that the bottleneck will be on the theoretical side. A crucial step is the computation speed when considering TF-nucleosome competition and partial nucleosome unwrapping. We showed how this issue can be addressed with novel faster algorithms [94]. A second challenge is the incorporation of

ATP-dependent remodeler activities. As discussed in section 6.3, this problem also has conceptual solutions that can help constructing quantitative models for gene regulation at new levels. Finally, nucleosome-dependent gene regulation is realized not only through nucleosome translocations and dissociation/unwrapping, but also through covalent histone modifications. These are accounted for by equation 1 through the use of the microscopic binding constants $k(n, g, h)$. Thus, introducing histone modifications changes the histone octamer binding constant through the change of the nucleosome type g , and through the change of the unwrapping potential (dependence of k on the unwrapping length h). Furthermore, the interaction energy of the corresponding nucleosome with nucleosome-binding proteins also changes depending on g . Given that there are several dozens of known histone modifications and histone variants and a huge number of their combinations [200], the amount of nucleosome types is tremendous. As with protein binding it is impossible to determine maps for all histone modification states experimentally for all cell states. Thus, one of the current challenges is it to identify a manageable subset of histone modifications that needs to be taken into account, derive a method of predicting its changes from the protein arrangements and to combine calculation of protein binding maps with the calculation of the corresponding changes in histone modifications.

Acknowledgements

We thank David Arnosti for valuable comments on the manuscript. VT acknowledges the support by a fellowship from the Heidelberg Center for Modeling and Simulation in the Biosciences (BIOMS) and the DKFZ Intramural Program. This work was funded within project EpiGenSys by the BMBF as a partner of the ERASysBio+ initiative supported under the EU ERA-NET Plus scheme in FP7.

Figure Legends

Figure 1. Different levels of understanding of cis-regulatory module functioning. A) Enhancer and promoter regions can be connected by a DNA loop bridged by transcription factors. One of the bound proteins is RNA polymerase (RNAP), whose binding determines the probability of transcription initiation. B) In chromatin, both enhancer and promoter regions might be covered by nucleosomes; some nucleosomes need to be removed or repositioned to be compatible with TF binding, which becomes an additional layer of regulation of the initiation of transcription.

Figure 2. Understanding TF-nucleosome competition. A) The DNA site is either bound by a transcription factor, or a nucleosome is formed and TFs cannot bind to this site. B) The nucleosome is viewed as a dynamic structure that can allow partial DNA unwrapping or partial disassembly of histone dimers. TFs bind to the released sites on the DNA inside the canonical nucleosome.

Figure 3. 1D lattice models for TF-DNA binding in the context of nucleosomes and other protein complexes. A) The histone octamer is represented as a single ligand covering up to 147 bp when completely bound. Partial bonding of histone octamer and DNA results in unwrapping of the DNA from the nucleosome entry/exit. Transcription factors can bind the unwrapped DNA. The forward partial partition function is calculated left to right (B). The reverse partial partition function is calculated right to left (C). A partial partition function of the system with the bound pink protein in the middle is given by the product of the corresponding forward and reverse partition functions, divided by the weight of this bound protein.

Figure 4. A schematic representation of the ChIP-Seq and MNase-Seq workflow to determine genome-wide nucleosome occupancy profiles. (A) ChIP-Seq. The chromatin is extracted from the cell nucleus, sonicated into short fragments and immunoprecipitated using antibodies specific to the chromatin protein of interest. The resulting DNA segments associated with the target protein are mapped, which results in sharp peaks for proteins that realize specific binding to well-defined binding sites. The peaks are identified with peak-calling software and used for TF binding DNA motif discovery. (B) MNase-Seq. The chromatin is extracted from the cell nucleus and digested by MNase to remove the linker DNA between nucleosomes. Subsequently, the remaining nucleosomal DNA is sequenced

and mapped to the reference genome. In the paired-end sequencing setup exact positions of individual nucleosome instances are determined, but they are overlapping due to some sample heterogeneity, resulting in a smooth distribution of the nucleosome occupancy. The nucleosome occupancy is defined as a normalized number of individual nucleosome reads covering a given DNA position.

Figure 5. Fluorescence fluctuation microscopy methods used to measure protein-chromatin and protein-protein interactions in living cells.

The principles of the most prominent living cell imaging techniques such as Single Particle Tracking (SPT), Fluorescence Recovery After Photobleaching (FRAP), Fluorescence (Cross) Correlation Spectroscopy (FCS/FCCS) and Fluorescence Resonance Energy Transfer (FRET) are illustrated. In SPT, a single particle (with high enough contrast) is followed within a sequence of microscopy images to determine its mobility. While the particle is bound to chromatin, it exhibits lower mobility, enabling to distinguish bound and free states as well as the kinetic rate constants. In FRAP, fluorescent particles within a large region are bleached, and the recovery of non-bleached molecules from the surrounding is recorded over time. The shape of the recovery curve encodes information about the chromatin interactions of the protein. In FCCS, the presence of two differently labeled proteins is assessed over time, and the correlated presence of the two proteins (corresponding to the amplitude of the correlation function) is used as readout for their interaction. Finally, the presence of FRET between two proteins labeled in two colors indicates that they are in close proximity, which can be interpreted as interaction between the two.

Figure 6. Flow chart for the integrative analysis to predict gene expression in the presence of nucleosomes.

TF binding affinities, interaction potentials and concentrations, which are the input parameters for the calculation of the binding maps, can be obtained from two interdependent sources: experiments and bioinformatic predictions based on the sequence analysis. The experimental part consists on high-throughput sequencing based methods detailed in Figure 4 and imaging techniques in living cells detailed in Figure 5. After the model parameterization, gene regulation functions are being calculated which allows predicting gene expression.

References

- [1] H.G. Garcia, A. Sanchez, T. Kuhlman, J. Kondev, R. Phillips, Transcription by the numbers redux: experiments and calculations that surprise, *Trends Cell Biol*, 20 (2010) 723-733.
- [2] L. Saiz, The physics of protein-DNA interaction networks in the control of gene expression, *J Phys Condens Matter*, 24 (2012) 193102.
- [3] E. Segal, J. Widom, From DNA sequence to transcriptional behaviour: a quantitative approach, *Nat Rev Genet*, 10 (2009) 443-456.
- [4] V.B. Teif, Predicting gene-regulation functions: lessons from temperate bacteriophages, *Biophys J*, 98 (2010) 1247-1256.
- [5] C.H. Yuh, H. Bolouri, E.H. Davidson, Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene, *Science*, 279 (1998) 1896-1902.
- [6] J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K.N. Kozlov, Manu, E. Myasnikova, C.E. Vanario-Alonso, M. Samsonova, D.H. Sharp, J. Reinitz, Dynamic control of positional information in the early *Drosophila* embryo, *Nature*, 430 (2004) 368-371.
- [7] H. Janssens, S. Hou, J. Jaeger, A.R. Kim, E. Myasnikova, D. Sharp, J. Reinitz, Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster* even skipped gene, *Nat Genet*, 38 (2006) 1159-1165.
- [8] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, U. Gaul, Predicting expression patterns from regulatory sequence in *Drosophila* segmentation, *Nature*, 451 (2008) 535-540.
- [9] J. Gertz, E.D. Siggia, B.A. Cohen, Analysis of combinatorial cis-regulation in synthetic and genomic promoters, *Nature*, 457 (2009) 215-218.
- [10] Y. Yuan, L. Guo, L. Shen, J.S. Liu, Predicting gene expression from sequence: a reexamination, *PLoS Comput Biol*, 3 (2007) e243.
- [11] M.A. Beer, S. Tavazoie, Predicting gene expression from sequence, *Cell*, 117 (2004) 185-198.
- [12] X. He, M.A. Samee, C. Blatti, S. Sinha, Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression, *PLoS Comput Biol*, 6 (2010) e1000935.
- [13] W.D. Fakhouri, A. Ay, R. Sayal, J. Dresch, E. Dayringer, D.N. Arnosti, Deciphering a transcriptional regulatory code: modeling short-range repression in the *Drosophila* embryo, *Mol Syst Biol*, 6 (2010) 341.
- [14] T. Kaplan, X.Y. Li, P.J. Sabo, S. Thomas, J.A. Stamatoyannopoulos, M.D. Biggin, M.B. Eisen, Quantitative models of the mechanisms that control genome-wide patterns of

- transcription factor binding during early *Drosophila* development, *PLoS Genet*, 7 (2011) e1001290.
- [15] R.P. Zinzen, C. Girardot, J. Gagneur, M. Braun, E.E. Furlong, Combinatorial binding predicts spatio-temporal cis-regulatory activity, *Nature*, 462 (2009) 65-70.
- [16] T. Irie, S.J. Park, R. Yamashita, M. Seki, T. Yada, S. Sugano, K. Nakai, Y. Suzuki, Predicting promoter activities of primary human DNA sequences, *Nucleic Acids Res*, 39 (2011) e75.
- [17] I.G. Costa, H.G. Roider, T.G. do Rego, A. de Carvalho Fde, Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models, *BMC Bioinformatics*, 12 Suppl 1 (2011) S29.
- [18] M. Ptashne, Regulation of transcription: from lambda to eukaryotes, *Trends Biochem Sci*, 30 (2005) 275-279.
- [19] Z. Ouyang, Q. Zhou, W.H. Wong, ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells, *Proc Natl Acad Sci U S A*, 106 (2009) 21521-21526.
- [20] J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J.J. Collins, T.S. Gardner, Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles, *PLoS Biol*, 5 (2007) e8.
- [21] S.K. Kurdistani, S. Tavazoie, M. Grunstein, Mapping Global Histone Acetylation Patterns to Gene Expression, *Cell*, 117 (2004) 721-733.
- [22] E. Segal, N. Friedman, N. Kaminski, A. Regev, D. Koller, From signatures to models: understanding cancer using microarrays, *Nat Genet*.
- [23] R. Sadeh, C.D. Allis, Genome-wide "re"-modeling of nucleosome positions, *Cell*, 147 (2011) 263-266.
- [24] B.D. Strahl, C.D. Allis, The language of covalent histone modifications, *Nature*, 403 (2000) 41-45.
- [25] M. Ptashne, Binding reactions: epigenetic switches, signal transduction and cancer, *Curr Biol*, 19 (2009) R234-241.
- [26] M.S. Sherman, B.A. Cohen, Thermodynamic State Ensemble Models of cis-Regulation, *PLoS Comput Biol*, 8 (2012) e1002407.
- [27] A. Bakk, R. Metzler, K. Sneppen, Sensitivity of OR in phage lambda, *Biophys J*, 86 (2004) 58-66.
- [28] G.K. Ackers, A.D. Johnson, M.A. Shea, Quantitative model for gene regulation by lambda phage repressor, *Proc Natl Acad Sci U S A*, 79 (1982) 1129-1133.
- [29] J.M. Vilar, S. Leibler, DNA looping and physical constraints on transcription regulation, *J Mol Biol*, 331 (2003) 981-989.

- [30] P.H. von Hippel, A. Revzin, C.A. Gross, A.C. Wang, Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects, *Proc Natl Acad Sci U S A*, 71 (1974) 4808-4812.
- [31] N.E. Buchler, U. Gerland, T. Hwa, On schemes of combinatorial transcription logic, *Proc Natl Acad Sci U S A*, 100 (2003) 5136-5141.
- [32] H.S. Rhee, B.F. Pugh, Genome-wide structure and organization of eukaryotic pre-initiation complexes, *Nature*, 483 (2012) 295-301.
- [33] N. Banerjee, M.Q. Zhang, Identifying cooperativity among transcription factors controlling the cell cycle in yeast, *Nucleic Acids Res*, 31 (2003) 7024-7031.
- [34] P.H. von Hippel, From "simple" DNA-protein interactions to the macromolecular machines of gene expression, *Annu. Rev. Biophys. Biomol. Struct.*, 36 (2007) 79-105.
- [35] D. Lebrecht, M. Foehr, E. Smith, F.J. Lopes, C.E. Vanario-Alonso, J. Reinitz, D.S. Burz, S.D. Hanes, Bicoid cooperative DNA binding is critical for embryonic patterning in *Drosophila*, *Proc Natl Acad Sci U S A*, 102 (2005) 13176-13181.
- [36] K. Rippe, P.H. von Hippel, J. Langowski, Action at a distance: DNA-looping and initiation of transcription, *Trends Biochem Sci*, 20 (1995) 500-506.
- [37] C.J. Tsai, R. Nussinov, Gene-specific transcription activation via long-range allosteric shape-shifting, *Biochem J*, 439 (2011) 15-25.
- [38] V.V. Ogryzko, Erwin Schroedinger, Francis Crick and epigenetic stability, *Biol Direct*, 3 (2008) 15.
- [39] D.N. Arnosti, M.M. Kulkarni, Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?, *J Cell Biochem*, 94 (2005) 890-898.
- [40] C.T. Ong, V.G. Corces, Enhancer function: new insights into the regulation of tissue-specific gene expression, *Nat Rev Genet*, 12 (2011) 283-293.
- [41] J. Rister, C. Desplan, Deciphering the genome's regulatory code: the many languages of DNA, *Bioessays*, 32 (2010) 381-384.
- [42] Y. Pan, C.J. Tsai, B. Ma, R. Nussinov, Mechanisms of transcription factor selectivity, *Trends Genet*, 26 (2010) 75-83.
- [43] D. Michel, Hierarchical cooperativity mediated by chromatin remodeling; the model of the MMTV transcription regulation, *J Theor Biol*, 287 (2011) 74-81.
- [44] Y. Pan, R. Nussinov, The Role of Response Elements Organization in Transcription Factor Selectivity: The IFN- β Enhanceosome Example, *PLoS Comput Biol*, 7 (2011) e1002077.
- [45] L. Bintu, N.E. Buchler, H.G. Garcia, U. Gerland, T. Hwa, J. Kondev, R. Phillips, Transcriptional regulation by the numbers: models, *Curr Opin Genet Dev*, 15 (2005) 116-124.
- [46] S. Istrail, E.H. Davidson, Logic functions of the genomic *cis*-regulatory code, *Proc Natl Acad Sci U S A*, 102 (2005) 4954-4959.

- [47] S. Kaplan, A. Bren, A. Zaslaver, E. Dekel, U. Alon, Diverse two-dimensional input functions control bacterial sugar genes, *Mol Cell*, 29 (2008) 786-792.
- [48] A.E. Mayo, Y. Setty, S. Shavit, A. Zaslaver, U. Alon, Plasticity of the cis-regulatory input function of a gene, *PLoS Biol*, 4 (2006) e45.
- [49] Y. Setty, A.E. Mayo, M.G. Surette, U. Alon, Detailed map of a cis-regulatory input function, *Proc Natl Acad Sci U S A*, 100 (2003) 7702-7707.
- [50] H.D. Kim, E.K. O'Shea, A quantitative model of transcription factor-activated gene expression, *Nat Struct Mol Biol*, 15 (2008) 1192-1198.
- [51] R. Karlic, H.R. Chung, J. Lasserre, K. Vlahovicek, M. Vingron, Histone modification levels are predictive for gene expression, *Proc Natl Acad Sci U S A*, 107 (2010) 2926-2931.
- [52] S. Payankulam, D.N. Arnosti, Gene regulation: boundaries within limits, *Curr Biol*, 18 (2008) R653-R655.
- [53] M.J. Schilstra, C.L. Nehaniv, Bio-logic: gene expression and the laws of combinatorial logic, *Artif Life*, 14 (2008) 121-133.
- [54] H.D. Kim, E.K. O'Shea, A quantitative model of transcription factor-activated gene expression, *Nat Struct Mol Biol*, 15 (2008) 1192-1198.
- [55] T.L. Lenstra, F.C.P. Holstege, The discrepancy between chromatin factor location and effect, *Nucleus*, 3 (2012) 1-7.
- [56] C. Cheng, M. Gerstein, Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells, *Nucleic Acids Res*, 40 (2012) 553-568.
- [57] B. Munsky, G. Neuert, A. van Oudenaarden, Using gene expression noise to understand gene regulation, *Science*, 336 (2012) 183-187.
- [58] D. Martin, C. Pantoja, A. Fernandez Minan, C. Valdes-Quezada, E. Molto, F. Matesanz, O. Bogdanovic, E. de la Calle-Mustienes, O. Dominguez, L. Taher, M. Furlan-Magaril, A. Alcina, S. Canon, M. Fedetz, M.A. Blasco, P.S. Pereira, I. Ovcharenko, F. Recillas-Targa, L. Montoliu, M. Manzanares, R. Guigo, M. Serrano, F. Casares, J.L. Gomez-Skarmeta, Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes, *Nat Struct Mol Biol*, 18 (2011) 708-714.
- [59] V.B. Teif, General transfer matrix formalism to calculate DNA-protein-drug binding in gene regulation: application to O_R operator of phage λ , *Nucleic Acids Res*, 35 (2007) e80.
- [60] V.B. Teif, K. Rippe, Statistical-mechanical lattice models for protein-DNA binding in chromatin, *J. Phys.: Condens. Matter*, 22 (2010) 414105.

- [61] V.B. Teif, K. Rippe, Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities, *Nucleic Acids Res*, 37 (2009) 5641-5655.
- [62] V.B. Teif, K. Bohinc, Condensed DNA: condensing the concepts, *Progr Biophys Mol Biol* 105 (2011) 199-213.
- [63] F. Erdel, K. Rippe, Chromatin remodelling in mammalian cells by ISWI-type complexes--where, when and why?, *FEBS J*, 278 (2011) 3608-3618.
- [64] F. Erdel, T. Schubert, C. Marth, G. Langst, K. Rippe, Human ISWI chromatin-remodeling complexes sample nucleosomes via transient binding reactions and become immobilized at active sites, *Proc Natl Acad Sci U S A*, 107 (2010) 19873-19878.
- [65] H.G. Garcia, J. Kondev, N. Orme, J.A. Theriot, R. Phillips, Thermodynamics of biological processes, *Methods Enzymol*, 492 (2011) 27-59.
- [66] C.R. Lickwar, F. Mueller, S.E. Hanlon, J.G. McNally, J.D. Lieb, Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function, *Nature*, 484 (2012) 251-255.
- [67] T.C. Voss, R.L. Schiltz, M.H. Sung, P.M. Yen, J.A. Stamatoyannopoulos, S.C. Biddie, T.A. Johnson, T.B. Miranda, S. John, G.L. Hager, Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism, *Cell*, 146 (2011) 544-554.
- [68] X. Wang, G.O. Bryant, M. Floer, D. Spagna, M. Ptashne, An effect of DNA sequence on nucleosome occupancy and removal, *Nat Struct Mol Biol*, 18 (2011) 507-509.
- [69] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, T.J. Richmond, Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution, *J Mol Biol*, 319 (2002) 1097-1113.
- [70] J. Svaren, E. Klebanow, L. Sealy, R. Chalkley, Analysis of the competition between nucleosome formation and transcription factor binding, *J Biol Chem*, 269 (1994) 9335-9344.
- [71] I.M. Kulic, H. Schiessel, DNA spools under tension, *Phys Rev Lett*, 92 (2004) 228101.
- [72] V.B. Teif, R. Ettig, K. Rippe, A lattice model for transcription factor access to nucleosomal DNA, *Biophys J*, 99 (2010) 2597-2607.
- [73] W. Mobius, R.A. Neher, U. Gerland, Kinetic accessibility of buried DNA sites in nucleosomes, *Phys Rev Lett*, 97 (2006) 208102.
- [74] J.D. Anderson, A. Thastrom, J. Widom, Spontaneous access of proteins to buried nucleosomal DNA target sites occurs via a mechanism that is distinct from nucleosome translocation, *Mol Cell Biol*, 22 (2002) 7147-7157.
- [75] M.G. Poirier, E. Oh, H.S. Tims, J. Widom, Dynamics and function of compact nucleosome arrays, *Nat Struct Mol Biol*, 16 (2009) 938-944.

- [76] J.L. Killian, M. Li, M.Y. Sheinin, M.D. Wang, Recent advances in single molecule studies of nucleosomes, *Curr Opin Struct Biol*, 22 (2012) 80-87.
- [77] K. Voltz, J. Trylska, N. Calimet, J.C. Smith, J. Langowski, Unwrapping of nucleosomal DNA ends: a multiscale molecular dynamics study, *Biophys J*, 102 (2012) 849-858.
- [78] R.A. Forties, J.A. North, S. Javaid, O.P. Tabbaa, R. Fishel, M.G. Poirier, R. Bundschuh, A quantitative model of nucleosome dynamics, *Nucleic Acids Res*, 39 (2011) 8306-8313.
- [79] V. Bohm, A.R. Hieb, A.J. Andrews, A. Gansen, A. Rocker, K. Toth, K. Luger, J. Langowski, Nucleosome accessibility governed by the dimer/tetramer interface, *Nucleic Acids Res*, 39 (2011) 3093-3102.
- [80] L.A. Mirny, Nucleosome-mediated cooperativity between transcription factors, *Proc Natl Acad Sci U S A*, 107 (2010) 22534-22539.
- [81] V.B. Teif, K. Rippe, Nucleosome mediated crosstalk between transcription factors at eukaryotic enhancers, *Phys Biol*, 8 (2011) 044001.
- [82] M. Engholm, M. de Jager, A. Flaus, R. Brenk, J. van Noort, T. Owen-Hughes, Nucleosomes can invade DNA territories occupied by their neighbors, *Nat Struct Mol Biol*, 16 (2009) 151-158.
- [83] G. Li, D. Reinberg, Chromatin higher-order structures and gene regulation, *Curr Opin Genet Dev*, 21 (2011) 175-186.
- [84] Y. Zhang, R.P. McCord, Y.J. Ho, B.R. Lajoie, D.G. Hildebrand, A.C. Simon, M.S. Becker, F.W. Alt, J. Dekker, Spatial organization of the mouse genome and its role in recurrent chromosomal translocations, *Cell*, 148 (2012) 908-921.
- [85] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, G. Cavalli, Three-dimensional folding and functional organization principles of the *Drosophila* genome, *Cell*, 148 (2012) 458-472.
- [86] E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, R. Sandstrom, B. Bernstein, M.A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L.A. Mirny, E.S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science*, 326 (2009) 289-293.
- [87] A. Bancaud, S. Huet, N. Daigle, J. Mozziconacci, J. Beaudouin, J. Ellenberg, Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin, *EMBO J*, 28 (2009) 3785-3798.
- [88] J.G. McNally, D. Mazza, Fractal geometry in the nucleus, *EMBO J*, 29 (2010) 2-3.
- [89] A. Sarai, H. Kono, Protein-DNA recognition patterns and predictions, *Annu Rev Biophys Biomol Struct*, 34 (2005) 379-398.

- [90] M.G. Strainic, Jr., J.J. Sullivan, J. Collado-Vides, P.L. deHaseth, Promoter interference in a bacteriophage lambda control region: effects of a range of interpromoter distances, *J Bacteriol*, 182 (2000) 216-220.
- [91] O.J. Bandele, X. Wang, M.R. Campbell, G.S. Pittman, D.A. Bell, Human single-nucleotide polymorphisms alter p53 sequence-specific binding at gene regulatory elements, *Nucleic Acids Res*, 39 (2011) 178-189.
- [92] E. Ising, Beitrag zur Theorie des Ferromagnetismus, *Z. Phys.*, 31 (1925) 253-258.
- [93] A.A. Markov, Investigation of a specific case of dependent observations, *Izv. Imper. Akad. Nauk (St.-Petersburg)*, 3 (1907) 61-80.
- [94] V.B. Teif, K. Rippe, Calculating transcription factor binding maps for chromatin, *Brief Bioinform*, 13 (2012) 187-201.
- [95] D.Y. Lando, V.B. Teif, Modeling of DNA condensation and decondensation caused by ligand binding, *Journal of Biomolecular Structure & Dynamics*, 20 (2002) 215-222.
- [96] D.Y. Lando, V.B. Teif, Long-range interactions between ligands bound to a DNA molecule give rise to adsorption with the character of phase transition of the first kind, *J Biomol Struct Dyn*, 17 (2000) 903-911.
- [97] I.R. Epstein, Cooperative and noncooperative binding of large ligands to a finite one-dimensional lattice. A model for ligand-oligonucleotide interactions., *Biophys. Chem.*, 8 (1978) 327-339.
- [98] V.B. Teif, S.G. Haroutiunian, V.I. Vorob'ev, D.Y. Lando, Short-range interactions and size of ligands bound to DNA strongly influence adsorptive phase transition caused by long-range interactions, *J Biomol Struct Dyn*, 19 (2002) 1093-1100.
- [99] J.M.G. Vilar, L. Saiz, CplexA: a Mathematica package to study macromolecular-assembly control of gene expression, *Bioinformatics*, 26 (2010) 2060-2061.
- [100] D.A. Beshnova, E.G. Bereznyak, A.V. Shestopalova, M.P. Evstigneev, A novel computational approach 'BP-STOCH' to study ligand binding to finite lattice, *Biopolymers*, 95 (2011) 208-216.
- [101] E. Mjolsness, On cooperative quasi-equilibrium models of transcriptional regulation, *J Bioinform Comput Biol*, 5 (2007) 467-490.
- [102] C. DeLisi, Cooperative phenomena in homopolymers. An alternative formulation of the partition function, *Biopolymers*, 13 (1974) 1511-1512.
- [103] C. DeLisi, Statistical thermodynamics of oligomer-polymer interactions, *Biopolymers*, 13 (1974) 2305-2314.
- [104] G.V. Gurskii, A.S. Zasedatelev, Precise relationships for calculating the binding of regulatory proteins and other lattice ligands in double-stranded polynucleotides, *Biofizika*, 23 (1978) 932-946.
- [105] A.S. Krylov, S.L. Grokhovsky, A.S. Zasedatelev, A.L. Zhuze, G.V. Gursky, B.P. Gottikh, Quantitative estimation of the contribution of pyrrolcarboxamide groups of the

- antibiotic distamycin A into specificity of its binding to DNA AT pairs, *Nucleic Acids Res*, 6 (1979) 289-304.
- [106] E. Di Cera, Y. Kong, Theory of multivalent binding in one and two-dimensional lattices, *Biophys Chem*, 61 (1996) 107-124.
- [107] E. Di Cera, Thermodynamics of local linkage effects. Contracted partition functions and the analysis of site-specific energetics, *Biophys Chem*, 37 (1990) 147-164.
- [108] E. Di Cera, S. Keating, Site-Specific thermodynamics of ising networks: A theorem for linearly connected subsystems, *Biopolymers*, 34 (1994) 673-678.
- [109] A.V. Morozov, K. Fortney, D.A. Gaykalova, V.M. Studitsky, J. Widom, E.D. Siggia, Using DNA mechanics to predict in vitro nucleosome positions and formation energies, *Nucleic Acids Research*, 37 (2009) 4707-4722.
- [110] Y.D. Nechipurenko, B. Jovanovic, V.F. Riabokon, G.V. Gursky, Quantitative methods of analysis of footprinting diagrams for the complexes formed by a ligand with a DNA fragment of known sequence, *Annals of the New York Academy of Sciences*, 1048 (2005) 206-214.
- [111] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore, J.P. Wang, J. Widom, A genomic code for nucleosome positioning, *Nature*, 442 (2006) 772-778.
- [112] T. Wasson, A.J. Hartemink, An ensemble model of competitive multi-factor binding of the genome, *Genome Res*, 19 (2009) 2101-2112.
- [113] J.A. Granek, N.D. Clarke, Explicit equilibrium modeling of transcription-factor binding and gene regulation, *Genome Biol*, 6 (2005) R87.
- [114] K. Laurila, O. Yli-Harja, H. Lahdesmaki, A protein-protein interaction guided method for competitive transcription factor binding improves target predictions, *Nucl. Acids Res.*, 37 (2009) e146.
- [115] R. Hermsen, S. Tans, P.R. ten Wolde, Transcriptional regulation by competing transcription factor modules, *PLoS Comput Biol*, 2 (2006) e164.
- [116] X. He, C.C. Chen, F. Hong, F. Fang, S. Sinha, H.H. Ng, S. Zhong, A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data, *PLoS ONE*, 4 (2009) e8155.
- [117] R. Hermsen, B. Ursem, P.R. ten Wolde, Combinatorial gene regulation using auto-regulation, *PLoS Comput Biol*, 6 (2010) e1000813.
- [118] P.J. Farnham, Insights from genomic profiling of transcription factors, *Nat Rev Genet*, 10 (2009) 605-616.
- [119] M. Annala, K. Laurila, H. Lahdesmaki, M. Nykter, A linear model for transcription factor binding affinity prediction in protein binding microarrays, *PLoS ONE*, 6 (2011) e20059.

- [120] O.G. Berg, 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 (1987) 723-750.
- [121] G.D. Stormo, Y. Zhao, Determining the specificity of protein-DNA interactions, *Nat Rev Genet*, 11 (2010) 751-760.
- [122] U. Pfreundt, D.P. James, S. Tweedie, D. Wilson, S.A. Teichmann, B. Adryan, FlyTF: improved annotation and enhanced functionality of the *Drosophila* transcription factor database, *Nucleic Acids Res*, 38 (2010) D443-447.
- [123] E. Portales-Casamar, S. Thongjuea, A.T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W.W. Wasserman, A. Sandelin, JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles, *Nucleic Acids Res*, 38 (2010) D105-110.
- [124] E. Wingender, P. Dietze, H. Karas, R. Knuppel, TRANSFAC: a database on transcription factors and their DNA binding sites, *Nucleic Acids Res*, 24 (1996) 238-241.
- [125] H.G. Roider, A. Kanhere, T. Manke, M. Vingron, Predicting transcription factor affinities to DNA from a biophysical model, *Bioinformatics*, 23 (2007) 134-141.
- [126] B.C. Foat, A.V. Morozov, H.J. Bussemaker, Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE, *Bioinformatics*, 22 (2006) e141-149.
- [127] M. Djordjevic, A.M. Sengupta, B.I. Shraiman, A biophysical approach to transcription factor binding site discovery, *Genome Res*, 13 (2003) 2381-2390.
- [128] M.F. Berger, M.L. Bulyk, Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors, *Nature protocols*, 4 (2009) 393-411.
- [129] E. Roulet, S. Busso, A.A. Camargo, A.J. Simpson, N. Mermod, P. Bucher, High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites, *Nature biotechnology*, 20 (2002) 831-835.
- [130] V. Jagannathan, E. Roulet, M. Delorenzi, P. Bucher, HTPSELEX--a database of high-throughput SELEX libraries for transcription factor binding sites, *Nucleic acids research*, 34 (2006) D90-94.
- [131] Y. Zhao, D. Granas, G.D. Stormo, Inferring binding energies from selected binding sites, *PLoS computational biology*, 5 (2009) e1000590.
- [132] R. Nutiu, R.C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G.P. Schroth, C.B. Burge, Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument, *Nature biotechnology*, 29 (2011) 659-664.
- [133] R. Gordan, A.J. Hartemink, M.L. Bulyk, Distinguishing direct versus indirect transcription factor-DNA interactions, *Genome Res*, (2009).

- [134] Z. Ouyang, Q. Zhou, W.H. Wong, ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells, *Proceedings of the National Academy of Sciences of the United States of America*, 106 (2009) 21521-21526.
- [135] H.S. Rhee, B.F. Pugh, Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution, *Cell*, 147 (2011) 1408-1419.
- [136] L. Teytelman, B. Ozaydin, O. Zill, P. Lefrancois, M. Snyder, J. Rine, M.B. Eisen, Impact of chromatin structures on DNA processing for genomic analyses, *PLoS ONE*, 4 (2009) e6700.
- [137] E.N. Trifonov, Cracking the chromatin code: Precise rule of nucleosome positioning, *Phys Life Rev*, 8 (2011) 39-50.
- [138] G.C. Yuan, Y.J. Liu, M.F. Dion, M.D. Slack, L.F. Wu, S.J. Altschuler, O.J. Rando, Genome-scale identification of nucleosome positions in *S. cerevisiae*, *Science*, 309 (2005) 626-630.
- [139] H.E. Peckham, R.E. Thurman, Y. Fu, J.A. Stamatoyannopoulos, W.S. Noble, K. Struhl, Z. Weng, Nucleosome positioning signals in genomic DNA, *Genome Res*, 17 (2007) 1170-1177.
- [140] V.G. Levitsky, RECON: a program for prediction of nucleosome formation potential, *Nucleic Acids Res*, 32 (2004) W346-349.
- [141] L. Xi, Y. Fondufe-Mittendorf, L. Xia, J. Flatow, J. Widom, J.P. Wang, Predicting nucleosome positioning using a duration Hidden Markov Model, *BMC Bioinformatics*, 11 (2010) 346.
- [142] I. Gabdank, D. Barash, E.N. Trifonov, FineStr: a web server for single-base-resolution nucleosome positioning, *Bioinformatics*, 26 (2010) 845-846.
- [143] G. Locke, D. Tolkunov, Z. Moqtaderi, K. Struhl, A.V. Morozov, High-throughput sequencing reveals a simple model of nucleosome energetics, *Proc Natl Acad Sci U S A*, 107 (2010) 20998-21003.
- [144] Z. Zhang, B.F. Pugh, High-resolution genome-wide mapping of the primary structure of chromatin, *Cell*, 144 (2011) 175-186.
- [145] K. Cui, K. Zhao, Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq, *Methods Mol Biol*, 833 (2012) 413-419.
- [146] I.P. Ioshikhes, I. Albert, S.J. Zanton, B.F. Pugh, Nucleosome positions predicted through comparative genomics, *Nat Genet*, 38 (2006) 1210-1215.
- [147] G. Hu, D.E. Schones, K. Cui, R. Ybarra, D. Northrup, Q. Tang, L. Gattinoni, N.P. Restifo, S. Huang, K. Zhao, Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1, *Genome Res*, 21 (2011) 1650-1658.
- [148] A. Valouev, S.M. Johnson, S.D. Boyd, C.L. Smith, A.Z. Fire, A. Sidow, Determinants of nucleosome organization in primary human cells, *Nature*, 474 (2011) 516-520.

- [149] C.J. Ott, J.M. Bischof, K.M. Unti, A.E. Gillen, S.H. Leir, A. Harris, Nucleosome occupancy reveals regulatory elements of the CFTR promoter, *Nucleic Acids Res*, 40 (2012) 625-637.
- [150] D.E. Schones, K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, K. Zhao, Dynamic regulation of nucleosome positioning in the human genome, *Cell*, 132 (2008) 887-898.
- [151] L. Zhang, H. Ma, B.F. Pugh, Stable and dynamic nucleosome states during a meiotic developmental process, *Genome Res*, 21 (2011) 875-884.
- [152] I. Tirosh, N. Sigal, N. Barkai, Widespread remodeling of mid-coding sequence nucleosomes by *Isw1*, *Genome Biol*, 11 (2010) R49.
- [153] Y.M. Moshkin, G.E. Chalkley, T.W. Kan, B.A. Reddy, Z. Ozgur, W.F. van Ijcken, D.H. Dekkers, J.A. Demmers, A.A. Travers, C.P. Verrijzer, Remodelers organize cellular chromatin by counteracting intrinsic histone-DNA sequence preferences in a class-specific manner, *Mol Cell Biol*, 32 (2012) 675-688.
- [154] Z. Li, J. Schug, G. Tuteja, P. White, K.H. Kaestner, The nucleosome map of the mammalian liver, *Nat Struct Mol Biol*, 18 (2011) 742-746.
- [155] G.O. Bryant, Measuring nucleosome occupancy in vivo by micrococcal nuclease, *Methods Mol Biol*, 833 (2012) 47-61.
- [156] W. Bujalowski, Thermodynamic and kinetic methods of analyses of protein-nucleic acid interactions. From simpler to more complex systems, *Chem Rev*, 106 (2006) 556-606.
- [157] M.T. Record, Jr., J.H. Ha, M.A. Fisher, Analysis of equilibrium and kinetic measurements to determine thermodynamic origins of stability and specificity and mechanism of formation of site-specific complexes between proteins and helical DNA, *Methods Enzymol*, 208 (1991) 291-343.
- [158] P.H. von Hippel, From "simple" DNA-protein interactions to the macromolecular machines of gene expression, *Annu Rev Biophys Biomol Struct*, 36 (2007) 79-105.
- [159] R. Helwa, J.D. Hoheisel, Analysis of DNA-protein interactions: from nitrocellulose filter binding assays to microarray studies, *Anal Bioanal Chem*, 398 (2010) 2551-2561.
- [160] G.W. Gordon, G. Berry, X.H. Liang, B. Levine, B. Herman, Quantitative fluorescence resonance energy transfer measurements using fluorescence microscopy, *Biophys. J.*, 74 (1998) 2702-2713.
- [161] A.R. Hieb, S. D'Arcy, M.A. Kramer, A.E. White, K. Luger, Fluorescence strategies for high-throughput quantification of protein interactions, *Nucleic Acids Res*, 40 (2012) e33.
- [162] Y. Phillip, V. Kiss, G. Schreiber, Protein-binding dynamics imaged in a living cell, *Proc. Natl. Acad. Sci. USA*, 109 (2012) 1461-1466.

- [163] T. Kanno, Y. Kanno, R.M. Siegel, M.K. Jang, M.J. Lenardo, K. Ozato, Selective recognition of acetylated histones by bromodomain proteins visualized in living cells, *Mol. Cell*, 13 (2004) 33-43.
- [164] F.G. Cremazy, E.M. Manders, P.I. Bastiaens, G. Kramer, G.L. Hager, E.B. van Munster, P.J. Verschure, T.J. Gadella, Jr., R. van Driel, Imaging in situ protein-DNA interactions in the cell nucleus using FRET-FLIM, *Exp. Cell Res.*, 309 (2005) 390-396.
- [165] B.L. Sprague, R.L. Pego, D.A. Stavreva, J.G. McNally, Analysis of binding reactions by fluorescence recovery after photobleaching, *Biophys. J.*, 86 (2004) 3473-3495.
- [166] F. Mueller, P. Wach, J.G. McNally, Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching, *Biophys. J.*, 94 (2008) 3323-3339.
- [167] T.S. Karpova, M.J. Kim, C. Spriet, K. Nalley, T.J. Stasevich, Z. Kherrouche, L. Heliot, J.G. McNally, Concurrent fast and slow cycling of a transcriptional activator at an endogenous promoter, *Science*, 319 (2008) 466-469.
- [168] J.G. McNally, W.G. Muller, D. Walker, R. Wolford, G.L. Hager, The glucocorticoid receptor: rapid exchange with regulatory sites in living cells, *Science*, 287 (2000) 1262-1265.
- [169] B.L. Sprague, F. Muller, R.L. Pego, P.M. Bungay, D.A. Stavreva, J.G. McNally, Analysis of binding at a single spatially localized cluster of binding sites by fluorescence recovery after photobleaching, *Biophys J*, 91 (2006) 1169-1191.
- [170] K. Zolghadr, O. Mortusewicz, U. Rothbauer, R. Kleinhans, H. Goehler, E.E. Wanker, M.C. Cardoso, H. Leonhardt, A fluorescent two-hybrid assay for direct visualization of protein interactions in living cells, *Mol. Cell Proteomics*, 7 (2008) 2279-2287.
- [171] I. Chung, H. Leonhardt, K. Rippe, De novo assembly of a PML nuclear subcompartment occurs through multiple pathways and induces telomere elongation, *J Cell Sci*, 124 (2011) 3603-3618.
- [172] K. Bacia, S.A. Kim, P. Schuille, Fluorescence cross-correlation spectroscopy in living cells, *Nat Methods*, 3 (2006) 83-89.
- [173] Y.H. Foo, N. Naredi-Rainer, D.C. Lamb, S. Ahmed, T. Wohland, Factors affecting the quantification of biomolecular interactions by fluorescence cross-correlation spectroscopy, *Biophys. J.*, 102 (2012) 1174-1183.
- [174] X. Shi, Y.H. Foo, T. Sudhakaran, S.W. Chong, V. Korzh, S. Ahmed, T. Wohland, Determination of dissociation constants in living zebrafish embryos with single wavelength fluorescence cross-correlation spectroscopy, *Biophys. J.*, 97 (2009) 678-686.
- [175] J. Elf, G.W. Li, X.S. Xie, Probing transcription factor dynamics at the single-molecule level in a living cell, *Science*, 316 (2007) 1191-1194.

- [176] K. Sneppen, G. Zocchi, *Physics in Molecular Biology*, Cambridge University Press, Cambridge, 2005.
- [177] M. Ptashne, *A Genetic Switch, Third Edition: Phage Lambda Revisited*, Cold Spring Harbor Laboratory Press, New York, 2004.
- [178] A. Bakk, R. Metzler, In vivo non-specific binding of lambda CI and Cro repressors is significant, *FEBS Lett*, 563 (2004) 66-68.
- [179] M. Werner, E. Aurell, A computational study of lambda-lac mutants, *Phys Biol*, 6 (2009) 046007.
- [180] L. Saiz, J.M. Rubi, J.M. Vilar, Inferring the in vivo looping properties of DNA, *Proc Natl Acad Sci U S A*, 102 (2005) 17642-17645.
- [181] E. Ford, D. Thanos, The transcriptional code of human IFN-beta gene expression, *Biochim Biophys Acta*, 1799 (2010) 328-336.
- [182] J. Wang, K. Ellwood, A. Lehman, M.F. Carey, Z.S. She, A mathematical model for synergistic eukaryotic gene activation, *J Mol Biol*, 286 (1999) 315-325.
- [183] G. Längst, V.B. Teif, K. Rippe, Chromatin remodeling and nucleosome positioning, in: K. Rippe (Ed.) *Genome organization and function in the cell nucleus*, Wiley-VCH, Weinheim, 2011, pp. 111-139.
- [184] T. Raveh-Sadka, M. Levo, E. Segal, Incorporating nucleosomes into thermodynamic models of transcription regulation, *Genome Res*, 19 (2009) 1480-1496.
- [185] A.V. Morozov, K. Fortney, D.A. Gaykalova, V.M. Studitsky, J. Widom, E.D. Siggia, Using DNA mechanics to predict in vitro nucleosome positions and formation energies, *Nucleic Acids Res*, 37 (2009) 4707-4722.
- [186] K.J. Polach, J. Widom, A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites *J Mol Biol*, 258 (1996) 800-812.
- [187] K. Rippe, Making contacts on a nucleic acid polymer, *Trends Biochem Sci*, 26 (2001) 733-740.
- [188] V.J. Makeev, A.P. Lifanov, A.G. Nazina, D.A. Papatsenko, Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information, *Nucleic Acids Res*, 31 (2003) 6016-6026.
- [189] D. Papatsenko, Y. Goltsev, M. Levine, Organization of developmental enhancers in the *Drosophila* embryo., *Nucleic Acids Res*, 37 (2009) 5665-5677.
- [190] G. Moyle-Heyrman, H.S. Tims, J. Widom, Structural constraints in collaborative competition of transcription factors against the nucleosome, *J Mol Biol*, 412 (2011) 634-646.
- [191] Z. Zhang, C.J. Wippo, M. Wal, E. Ward, P. Korber, B.F. Pugh, A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome, *Science*, 332 (2011) 977-980.

- [192] Y. Fu, M. Sinha, C.L. Peterson, Z. Weng, The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome, *PLoS Genetics*, 4 (2008) e1000138.
- [193] T.N. Mavrich, I.P. Ioshikhes, B.J. Venters, C. Jiang, L.P. Tomsho, J. Qi, S.C. Schuster, I. Albert, B.F. Pugh, A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome, *Genome Res*, 18 (2008) 1073-1083.
- [194] V.B. Teif, E. Vainstein, K. Marth, J.-P. Mallm, M. Caudron-Herger, T. Höfer, K. Rippe, Genome-wide nucleosome positioning during embryonic stem cell development, *Nat Struct Mol Biol*, 19 (2012) 1185-1192.
- [195] R.D. Kornberg, L. Stryer, Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism, *Nucleic Acids Res*, 16 (1988) 6677–6690.
- [196] W. Mobius, 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).
- [197] J. Riposo, J. Mozziconacci, Nucleosome positioning and nucleosome stacking: two faces of the same coin, *Molecular bioSystems*, 8 (2012) 1172-1178.
- [198] R. Blossey, H. Schiessel, The dynamics of the nucleosome: thermal effects, external forces and ATP, *FEBS J*, 278 (2011) 3619-3632.
- [199] E.-L. Mathieu, F. Finkernagel, M. Murawska, M. Scharfe, M. Jarek, A. Brehm, Recruitment of the ATP-dependent chromatin remodeler dMi-2 to the transcribed region of active heat shock genes, *Nucleic Acids Research*, (2012).
- [200] S.J. Prohaska, P.F. Stadler, D.C. Krakauer, Innovation in gene regulation: the case of chromatin computation, *J Theor Biol*, 265 (2010) 27-44.

Appendix. A dynamic programming algorithm to calculate TF-DNA binding probabilities for chromatin taking into account partial nucleosome unwrapping.

In order to calculate probabilities of TF-DNA binding in Eq. 1, one needs to know the partition function of the system. This calculation becomes non-trivial when partial nucleosome unwrapping is considered. The corresponding calculation strategy using the transfer matrix formalism has been described elsewhere [72]. An equivalent approach is also available in the frame of the dynamic programming approach [94]. In the latter study, the algorithm was derived only in the case of homotypic interactions between DNA-bound TFs (when the TF-TF interaction potential depends on the distance but does not depend on the TF type). Here, the extension of this algorithm is described that allows calculations for the general case of heterotypic TF-TF interactions.

Let us consider the genomic region of length N , with index n numbering the first bp covered by a protein of type g , and index s numbering the last bp covered by a protein of type g ($s = n + m(g) - h_1 - h_2 - 1$) (Figure 3). Then the partition function Z for a DNA of length s can be calculated recurrently according to Eq. A1:

$$Z_s = Z_{s-1} + \sum_{g=1}^f \sum_{h_1=0}^{m(g)-1} \sum_{h_2=0}^{m(g)-h_1-1} c_0(g) Z_{s-m(g)+h_1+h_2-1} K^* + \sum_{j=0}^V \sum_{g'=1}^f \sum_{g=1}^f \sum_{h_1=0}^{m(g)-1} \sum_{h_2=0}^{m(g)-h_1-1} \sum_{h_1'=0}^{m(g')-1} \sum_{h_2'=0}^{m(g')-h_1'-1} \omega(j, g', g) \cdot c_0(g) \left(Z_{s-m(g)+h_1+h_2-j}^{+} (n - m(g') + h_1' + h_2' - j, g', h_1', h_2') \right) K^* \quad (\text{A1})$$

With the following boundary conditions:

$$Z_s = 1 \text{ for } s < m(g) - h_1 - h_2, \quad (\text{A2})$$

Here $c_0(g)$ is the free concentration of the protein of type g , and the macroscopic binding constant $K^* = K(n, g, h_1, h_2)$ for the protein whose first contact with DNA starts at position n is defined by Eq. 1 in the main text with the following boundary conditions:

$$K(n, g, h_1, h_2) = 0 \text{ for } n < 1 \text{ or } s > N \quad (\text{A3})$$

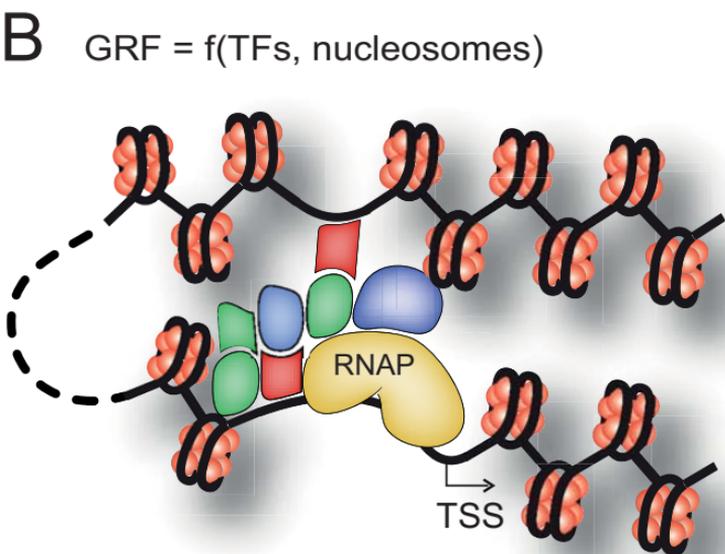
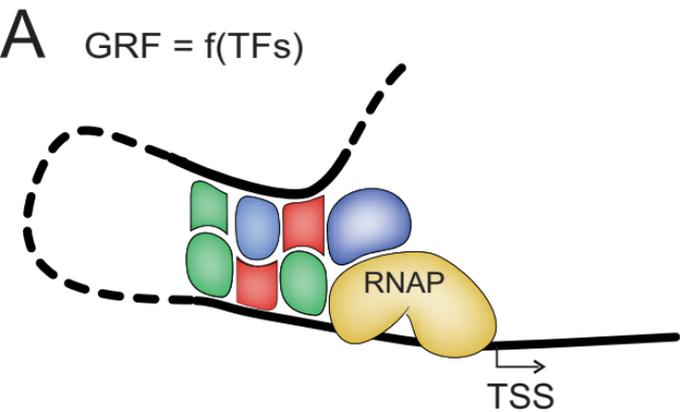
A given configuration with DNA positions $[n, s]$ covered by a bound protein of type g with unbound h_1 and h_2 bp from its left and right ends, respectively, is described by the following partial partition function:

$$Z_s^+(n, g, h_1, h_2) = c_0(g) Z_{s-m(g)+h_1+h_2-V-1} K^* + \sum_{j=0}^V \sum_{g'=1}^f \sum_{h_1'=0}^{m(g')-1} \sum_{h_2'=0}^{m(g')-h_1'-1} [Z_{s-m(g)+h_1+h_2-j}^+(n-m(g')+h_1'+h_2'-j, g', h_1', h_2')] w(j, g', g) c_0(g) K^* \quad (\text{A4})$$

Equation A4 is based on the recurrent calculation of the partition function in the forward direction (left to right in Figure 3B). Analogously, we can calculate the partial partition function backwards (right to left from N to n in Figure 3C) for the situation when the protein of type g with unwrapped h_1 and h_2 bp covers region $[n, s]$ on the DNA. This partial partition function is denoted as $Z_n^-(n, g, h_1, h_2)$. Then the product of partial partition functions

$Z_s^+(n, g, h_1, h_2) \cdot Z_n^-(n, g, h_1, h_2)$ gives the sum of all states of the system where the protein of type g with unwrapped h_1 and h_2 bp covers region $[n, s]$ on the DNA. This expression has to be divided by $c_0(g) \cdot K(n, g, h_1, h_2)$ because the forward and reverse partition functions take into account our protein of interest twice. Finally, in order to find the probability of TF binding event we have to divide this expression by the total partition function Z_n of the system. Then the probability that the protein of type g with unwrapped h_1 and h_2 bp starts at position n is given by the following expression:

$$P(n, g, h_1, h_2) = \frac{Z_s^+(n, g, h_1, h_2) \cdot Z_n^-(n, g, h_1, h_2)}{Z_n \cdot c_0(g) \cdot K(n, g, h_1, h_2)} \quad (\text{A5})$$



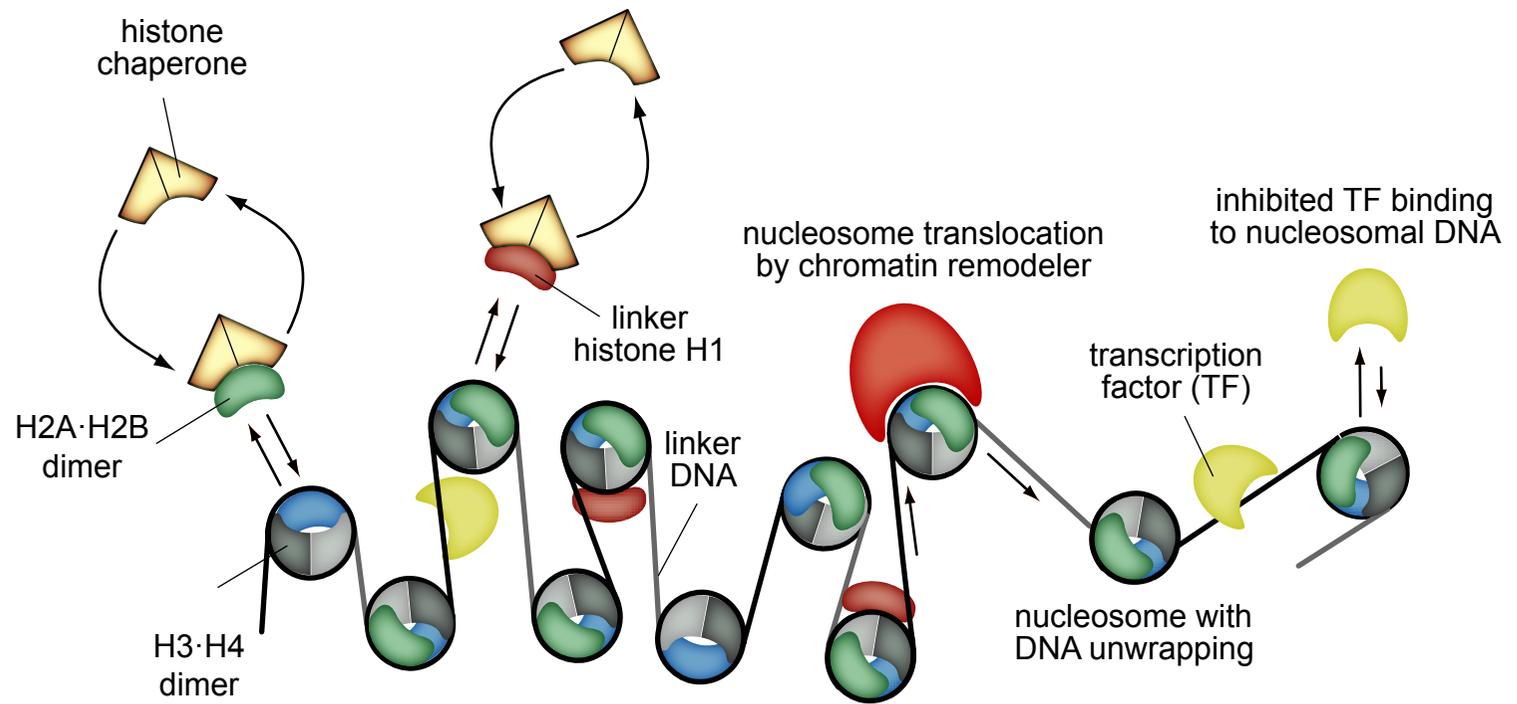


Figure 2, Teif et al.

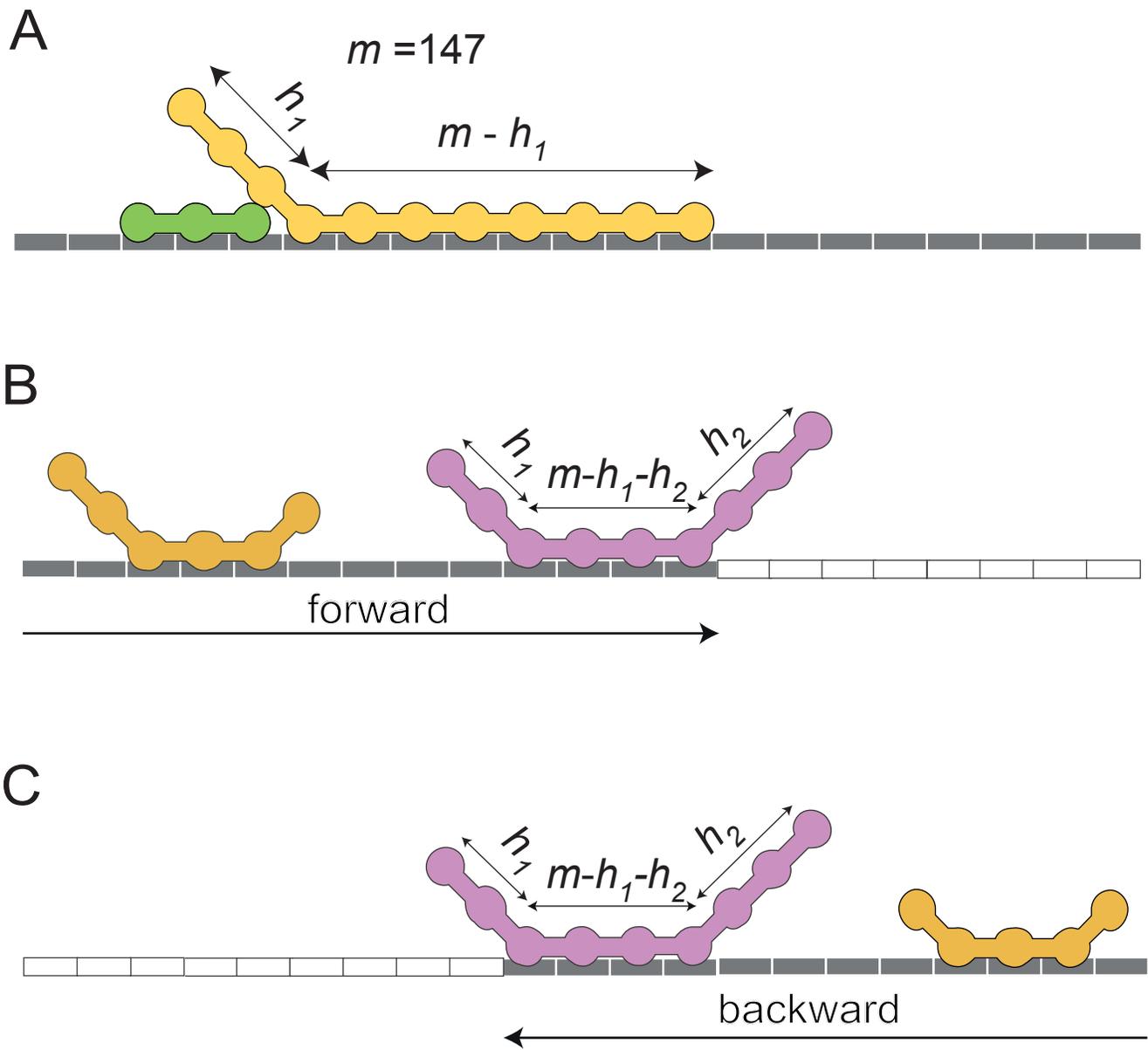
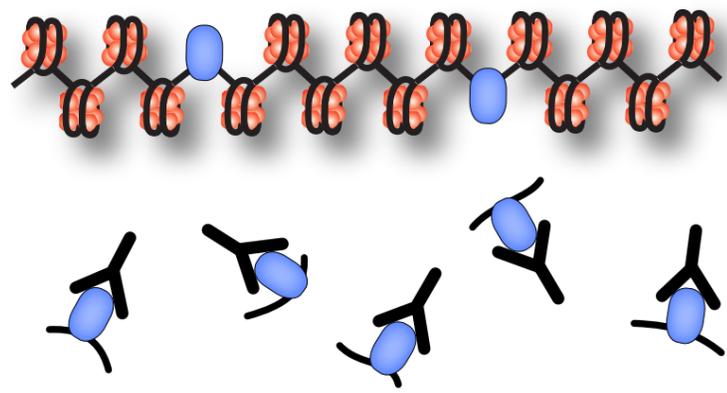


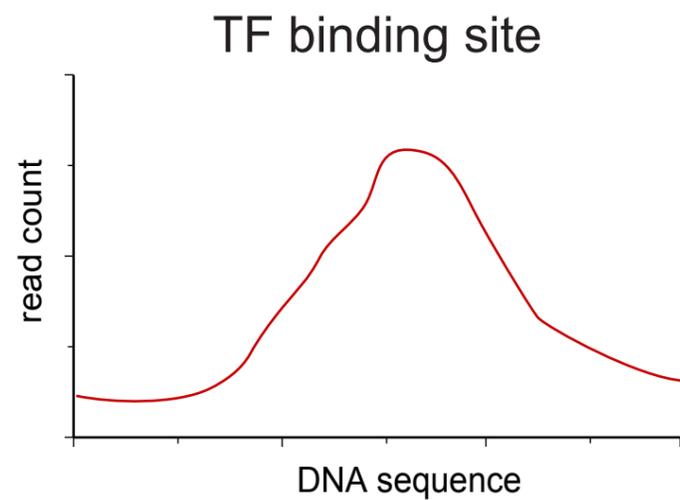
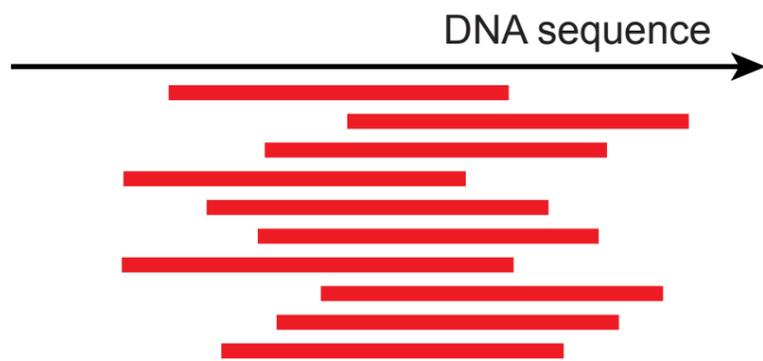
Figure 3, Teif et al.

A CHIP-seq

Immunoprecipitation of cross-linked TF after DNA sonication into small fragments

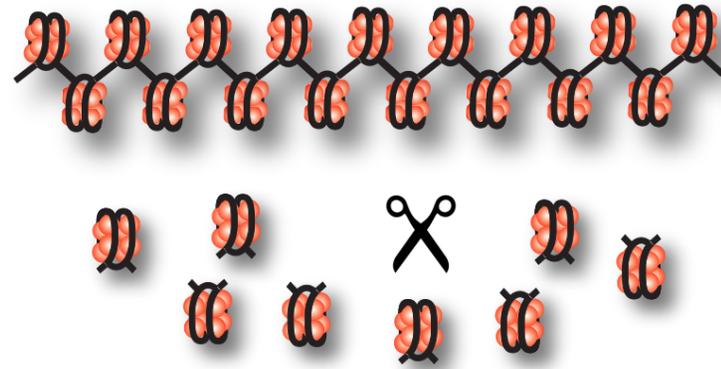


DNA sequencing and peak calling



B MNase-seq

MNase digestion of chromatin



DNA sequencing and mapping

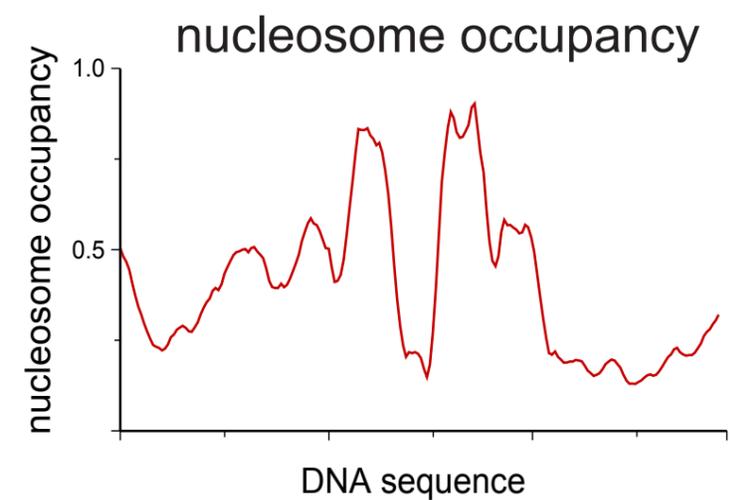
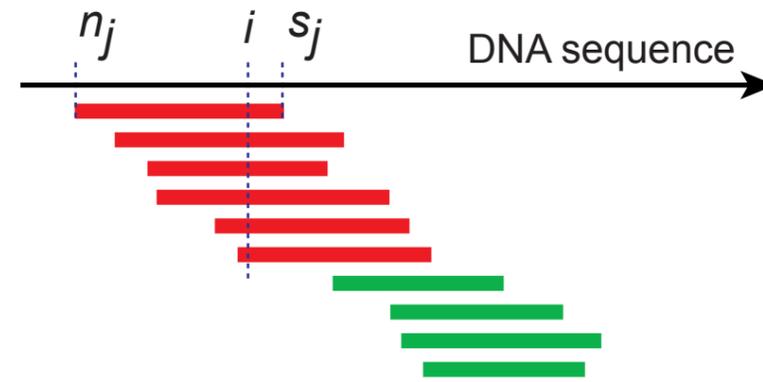


Figure 4, Teif et al.

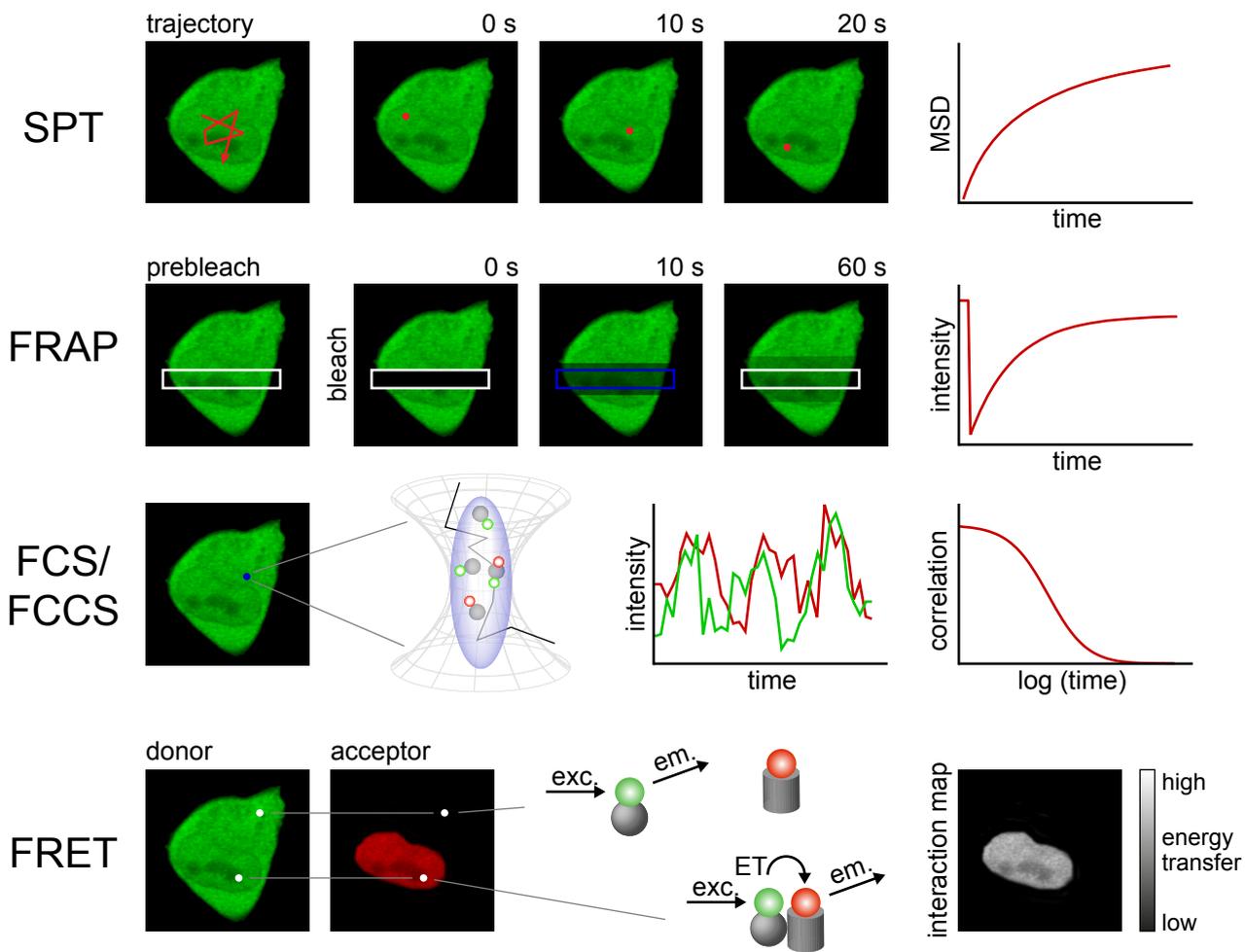


Figure 5

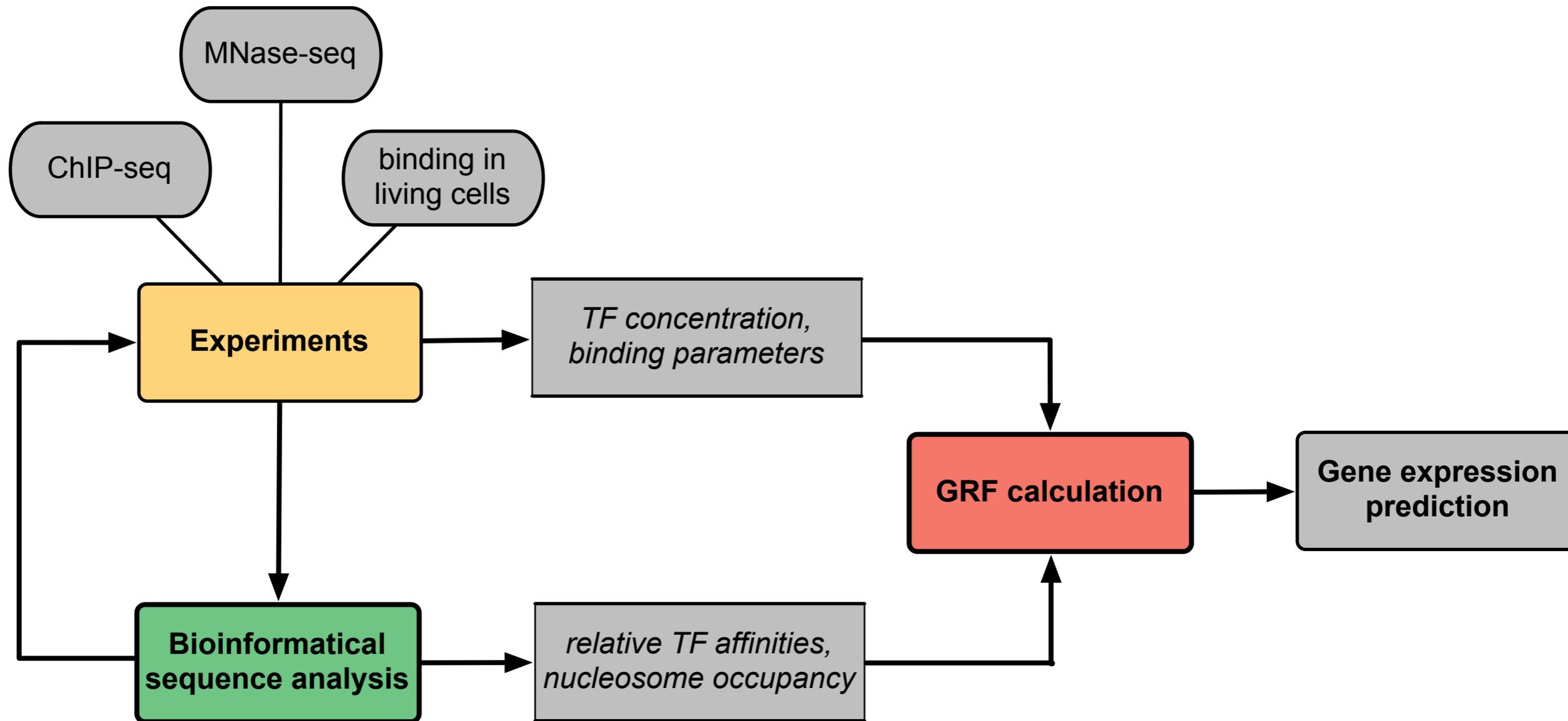


Figure 6, Teif et al.