BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 2. 187–201
Advance Access published on 6 July 2011
doi:10.1093/bib/bbr037

Calculating transcription factor binding maps for chromatin

Vladimir B. Teif and Karsten Rippe

Submitted: 8th March 2011; Received (in revised form): 25th May 2011

Abstract

Current high-throughput experiments already generate enough data for retrieving the DNA sequence-dependent binding affinities of transcription factors (TF) and other chromosomal proteins throughout the complete genome. However, the reverse task of calculating binding maps in a chromatin context for a given set of concentrations and TF affinities appears to be even more challenging and computationally demanding. The problem can be addressed by considering the DNA sequence as a one-dimensional lattice with units of one or more base pairs. To calculate protein occupancies in chromatin, one needs to consider the competition of TF and histone octamers for binding sites as well as the partial unwrapping of nucleosomal DNA. Here, we consider five different classes of algorithms to compute binding maps that include the binary variable, combinatorial, sequence generating function, transfer matrix and dynamic programming approaches. The calculation time of the binary variable algorithm scales exponentially with DNA length, which limits its use to the analysis of very small genomic regions. For regulatory regions with many overlapping binding sites, potentially applicable algorithms reduce either to the transfer matrix or dynamic programming approach. In addition to the recently proposed transfer matrix formalism for TF access to the nucleosomal organized DNA, we develop here a dynamic programming algorithm that accounts for this feature. In the absence of nucleosomes, dynamic programming outperforms the transfer matrix approach, but the latter is faster when nucleosome unwrapping has to be considered. Strategies are discussed that could further facilitate calculations to allow computing genome-wide TF binding maps.

Keywords: *DNA–protein binding; transcription factor; nucleosome; lattice model; dynamic programming; chromatin*

INTRODUCTION

One of the main goals in quantitative molecular and cell biology is it to predict gene expression from the assumptions of competitive and combinatorial binding of multiple proteins at genomic regulatory regions [1–3]. High-throughput genome-wide experiments already generate enough data to hope that this aim can be achieved. A number of recent studies approached the challenging problem of predicting gene expression from the DNA sequence and transcription factor (TF) concentrations [4–13]. For example, early stages of *Drosophila* embryonic development depend on ~40 TFs, which bind at a similar number of *cis*-regulatory regions, each spanning ~100–1000 bp and providing ~10–20 binding sites for two to five different TFs. Assuming that a typical TF covers ~10 bp, there are up to $4^{10} = 1\,048\,576$ possible combinations of 4 nt (A, T, G, C) at one binding site. This number of combinations can still be sampled experimentally using current high-throughput methods to obtain sequence dependent binding affinities in the form of so-called position weight matrices (PWM). Methods of learning PWMs are usually based on representing the DNA sequence as a 1-D lattice of units (base pairs, di-nucleotides, etc.) with each DNA unit contributing additively to the energy of protein binding to a given site depending on its position [14]. Current methods

Corresponding author. Vladimir B. Teif, BioQuant and German Cancer Research Center (DKFZ), Im Neuenheimer Feld 267, 69120 Heidelberg, Germany. Tel: +49-6221-54-51370; E-mail: vladimir.teif@bioquant.uni-heidelberg.de

Vladimir Teif is a BIOMS Fellow at the German Cancer Research Center and BioQuant. His current interests include modeling DNA–protein binding in chromatin and epigenetic regulation in general.

Karsten Rippe is a Group Leader at the German Cancer Research Center and BioQuant. In his work, he integrates molecular and cell biology with biophysics to relate chromatin organization and regulation of gene expression.

of learning DNA sequence dependent TF affinities have been recently reviewed and are not discussed here [15]. For many TFs, position weight matrices are available via databases such as FlyTF [16], JASPAR [17] and TRANSFAC [18]. For larger protein complexes, DNA binding affinities cannot be determined by direct sampling. For example, sampling over all possible 4^{147} nt combinations for the stretch of 147 DNA base pairs contacted by the histone octamer in the nucleosome is not possible. Therefore, the problem of getting histone–DNA affinities is more complex, but due to the intrinsic symmetry of the nucleosome, it also has experimental solutions [19–22]. Several web servers exist for calculating affinities of the histone octamer particle to an arbitrary DNA sequence [20, 23–26].

Learning TF- and histone–DNA affinities is only one part of the problem. Once protein binding affinities have been determined [27], one has to solve the reverse task: predicting how proteins at a given set of concentrations arrange in the genome where many TFs of different type and concentration compete for binding to overlapping DNA sites with more than three-fourth of DNA being occupied by histones [28]. Notably, the reconstruction of the complete binding map is algorithmically more difficult and computationally more expensive than the problem of obtaining binding affinities of individual proteins. In principle, if one assumes that binding happens at thermodynamic equilibrium, all binding probabilities can be calculated from a complete set of thermodynamic parameters. While this type of descriptions works well *in vitro* for dilute solutions, one has to be cautious in the very crowded environment of the cell with ATP-driven molecular motors acting against thermal equilibrium [29]. Nevertheless, in many cases studied so far, relative binding probabilities calculated under assumption of a quasi-equilibrium steady state reflect actual preferences for protein arrangement along the DNA *in vivo*. This approach can even be used to describe an obviously nonequilibrium process of sequence-specific nucleosome removal from gene promoters by chromatin remodelers [30]. Thus, although there is no true thermal equilibrium in the living cell, it is still reasonable to use the quasi-equilibrium assumption. Here, we will address how to perform these calculations. It is shown that single base pair accuracy of the binding maps can be currently achieved only when individual *cis*-regulatory modules are considered in chromatin. For a genome-wide analysis the computation time is becoming a bottleneck. Thus, the choice of the most efficient algorithm is crucial for increasing the range of applications that involve predicting gene expression from TF binding maps.

BIOPHYSICAL FORMULATION OF 1-D LATTICE MODELS

The majority of available data indicates that protein–DNA binding should be considered at a single base pair level to be biologically relevant for gene regulatory processes. Each base pair within the binding site contributes to the protein–DNA contact, and substitution of a single base pair may have large effects on TF–DNA binding affinity [31–33]. Furthermore, the distance between TF binding sites on the DNA matters for protein–protein interactions and is largely affected by adding or removing 1 DNA base pair [34, 35]. For example, changing the distances between TF binding sites at *Drosophila* enhancers by several base pairs lead to essentially different phenotypes [35]. In addition, recent studies emphasize the importance of single-nucleotide polymorphisms (SNP) for differential TF binding at regulatory regions by not only changing the protein coding sequence, but also by altering TF binding at promoters and enhancers [36].

Accordingly, we will focus here on one-dimensional DNA lattice models with single-base pair units numbered by index n (Figure 1A). Each DNA unit can be in one of several states determined by the reversible protein binding as is typical for Ising [37] and Markov chains [38]. 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 $K(n,g)$ depend on the position along the DNA n and protein type g. For each protein–DNA complex, it is possible to distinguish microscopic binding constants $k(n,g,h)$ corresponding to individual protein–DNA bonds. Here, index h numbers base pairs within the binding site, starting from the left-most base pair covered by the protein. The product of all microscopic binding constants for a given complex yields the macroscopic binding constant $K(n,g)$. In principle, any stretch of DNA nucleotides represents a potential binding site. Proteins g_1 and g_2 separated by l base pairs along the DNA can interact with each other with a potential $w=w(l,g_1,g_2)$. Proteins are characterized by the size of the

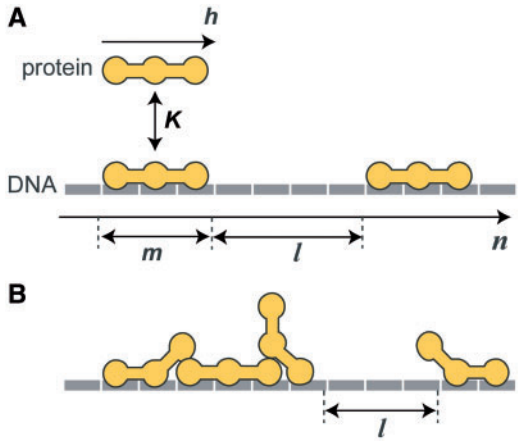

Figure 1: A lattice model for protein–DNA binding. The DNA is represented as a ID lattice of binding sites (e.g. base pairs). Proteins bind DNA with the association constant K depending on the protein type g and the position along the DNA n. Index h numbers DNA units at each protein–DNA binding interface. Protein covers up to m DNA units upon binding (in general, m depends on the protein type g; $m = 3$ in this example). DNA-bound proteins interact at distances $l \leq V$, which denotes the maximal range of interactions. (**A**) The all-or-none binding model where each protein binds completely to a binding site of m units. (**B**) A more general model where incomplete protein–DNA binding is allowed with binding sites less or equal m base pairs.

corresponding binding site $m(g)$, which is frequently assumed to be constant for a given protein type as shown in Figure 1A. However, in a more general description also partial binding to DNA is possible. In the example shown in Figure 1B, the protein can bind a maximum of 3 DNA units, but complexes with 1 and 2 bound units are also possible. This model becomes particularly important for extended protein–DNA complexes such as the nucleosome where it is known that partial unwrapping of DNA from the protein occurs spontaneously [39–47].

The nucleosome consists of 145–147 bp of DNA wrapped around the histone octamer protein core, and we will focus here on the 147 bp high-resolution structure [48]. This structure can be represented in different types of lattice models: (i) the histone octamer is considered as a single ligand that covers 147 bp, treated exactly as with other proteins [49]; (ii) the octamer is described as being formed by four histone dimers and so that partial nucleosome disassembly is possible [50]; and (iii) the histone octamer is treated as a single entity, which can form a variable number of protein–histone bonds (Figure 2A) [39]. The latter model allows for a good representation of the physiologically relevant partial nucleosome unwrapping. It immediately suggests two possible effects: first, TFs can access the DNA inside the nucleosome, especially close to its nucleosome entry/exit site (Figure 2B) and second, nucleosomes can invade the territories of each other (Figure 2C). Both of these effects have been observed experimentally [40–47]. Furthermore, this model was shown to be quantitatively consistent with *in vitro* measurements of DNA accessibility and nucleosome positioning [39]. Since it takes into account early all-or-none binding models (Figure 1A) and cooperative competitive binding of multiple proteins to overlapping DNA sites, we will use this nucleosome representation in the following.

At thermodynamic equilibrium, each bound state i is characterized by its statistical weight $\exp(-\Delta G_i/k_BT)$, 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 partition function Z is defined as a sum of weights of all configurations of the system:

$$Z = \sum_i e^{-\Delta G_i/k_B T} \quad (1)$$

The summation in Equation (1) is done over all possible configurations keeping the number of each type of molecules in the system constant. Usually the molecule numbers are given by their molar concentrations, and the partition function is considered to be a function of protein concentrations, while the DNA sequence, binding energies and stoichiometric parameters are considered as constants. However, as we will see below, in some cases it is useful redefining the partition function, e.g. as a function of the DNA sequence.

The knowledge of the partition function allows calculating any quantitative characteristic of protein–DNA binding. In particular, the probability P_i that a given binding configuration is realized is a ratio of the weight of a given state and the partition function Z:

$$P_i = \frac{e^{-\Delta G_i/k_B T}}{Z} \quad (2)$$

The partition function in Equation (1) depends linearly on individual binding constants and concentrations. Thus, the numerator of Equation (2) can be defined for any binding event simply by

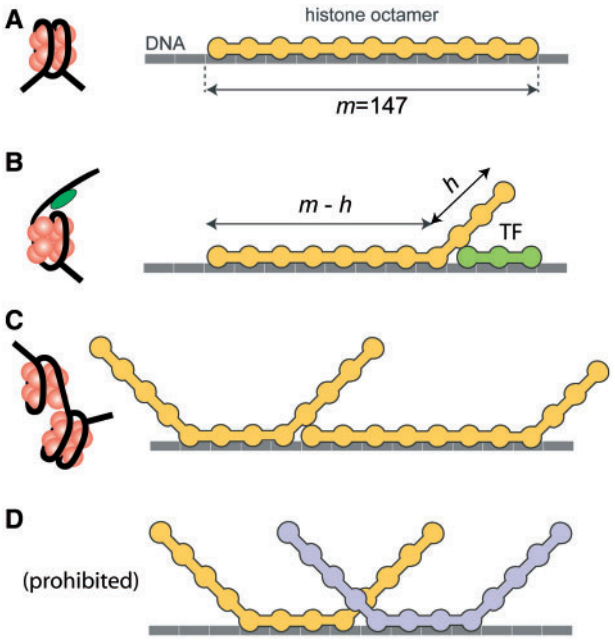

Figure 2: A lattice model for TF access to nucleosomal DNA. Schematic 3D representations on the left panel and the corresponding ID representations on the right. (**A**) The nucleosome is represented as a ligand covering up to $m = 147$ DNA base pairs. (**B**) TFs can access nucleosomal DNA partially unwrapped from the histone octamer. (**C**) Nucleosomes are allowed to invade territories of each other so that a dinucleosome structure consisting of two histone octamers would protect less than 2×147 DNA base pairs. (**D**) Due to steric exclusion, the situation when both nucleosomes are partially unwrapped at the point of contact is prohibited in the implementation of the matrix formalism used here. For the case of overlapping of two partially unwrapped nucleosomes, only configurations that have the DNA interacting regions of both nucleosomes separated by one or more free DNA units are allowed.

using the corresponding derivative of the partition function [51].

It is usually assumed that each base pair cannot be bound by more than one protein. However, one can also construct more complex binding models within the same mathematical framework, e.g. by including explicitly multilayer protein assembly and DNA looping [51]. In chromatin, some proteins bind to the DNA and the protein component of the nucleosome while others recognize an interface consisting of two nucleosomes, etc. The lattice binding approach can cover these situations as described in our recent work [50]. Here, we will not explicitly consider these more complex binding scenarios but the conclusions derived below about the underlying mathematics will apply to these systems as well.

Once the physical model is defined, mathematical models may be formulated and solved using different algorithms, which are more or less efficient. The main difference between computational algorithms is in the way they calculate the partition function, which determines the computation time. Here, we will consider the five most relevant classes of mathematical algorithms for this task.

Binary variable method

The simplest way to calculate the partition function is via sampling through all possible states of the system straight as defined by Equation (1). However, even in the case of a single type of protein binding to DNA, the number of possible configurations increases as $\sim m^N$ where m is the number of possible states of the DNA unit, and N is the DNA length. For example, 2^{100} summations would be an already too large a number for current computers. Therefore this method can be applied only to short DNA lattices [52, 53], or to a small number of known discrete binding sites of a few TFs [54]. Many quantitative models constructed for simple prokaryotic systems such as the λ-switch [55, 56] or the Lac operon [57–59] are based on this method. This is reasonable if just several binding sites with experimentally measured thermodynamic parameters are included. If any position along the DNA can be considered as a potential binding site, calculations for DNA regions >30 bp are not feasible using this method with currently available computers [60].

It was noted that the binary variable formulation of protein–DNA binding is equivalent to the neural

network models [59, 61–63]. Neural network algorithms can handle multiple nonoverlapping *cis*-regulatory modules. However, calculating protein–DNA binding at single base pair resolution, characterized by zillions of overlapping sites, goes beyond the capabilities of neural networks [64, 65]. Recently, Mjolsness [63] attempted to solve the problem of overlapping binding sites at *cis*-regulatory modules using binary variables, which required invoking the transfer matrix formalism discussed below. With respect to the pure binary variable method, its main limitation is the assumption that the partition function Z needs to be calculated by sampling through all the states. Other methods that we will consider use mathematical tricks to avoid sampling through all the states and still get the exact partition function.

Combinatorial methods

The algorithms in the next class of approaches use binomial or multinomial expansions to derive analytical expressions for the numbers of possible rearrangements of proteins along the DNA [29, 66–77]. For example, a protein, which covers m base pairs upon binding to the DNA, may adopt $(N-m+1)$ positions along the lattice of length N. This yields $\{[N-k\times(m-1)]!/(N-m\times k)!/k!\}$ possible rearrangements of k proteins. If binding is not sequence-specific and proteins do not interact, each of these individual configurations is characterized by the same energy, and one can just multiply the weight of one conformation by the number of conformations with this energy. With this approach, the classical McGhee–von Hippel [73] model was derived for a protein binding nonspecifically to infinitely long DNA. Other features can be also included using this method, such as polarity of protein–protein interactions [72, 78], competitions between a single specific site and nonspecific sites on a short DNA oligomer [68], electrostatic interactions [69], binding of flexible branched oligopolymers [66, 67] and two-state models such as DNA condensation coupled to ligand binding [29, 70, 71, 79]. Recently, Mirny [80] obtained an equilibrium combinatorial solution for the model of TF access to the nucleosomal-organized DNA taking into account nucleosome unwrapping. Previously, Chou [81] applied combinatorial formulations to the kinetic aspects of nucleosome unwrapping. However, when protein–DNA binding is sequence-specific, the description of binding to the lattice via combinatorial coefficients is not sufficient since each bound protein binds with a different energy.

Generating function method

Another classical approach is based on the idea to characterize the system by a mathematical expression called the generating function defined in terms of infinite series of the partition function [76, 82–89]. The physical meaning of the generating function can be understood as the partition function of the grand canonical ensemble [88]. The grand canonical ensemble is defined as the open system where the numbers of proteins are not fixed, while the canonical ensemble is defined as a closed system where the numbers of proteins are fixed. The generating function method is a powerful tool, which has been applied to a number of sophisticated protein–DNA binding models including for example the 'piggy-back' binding of proteins on the backs of other proteins already bound to DNA [85] and the problem of long-range interactions between DNA-bound proteins [88, 90]. The classical generating functions method [83] fails if more than one type of large proteins with long-range interactions exists in the system [87], but a recent modification of this method allows treating multiple-protein binding [82, 88]. The updated version of the generating function method by Di Cera and Kong [82, 88, 89] is based on recurrent relations as well as the class of dynamic programming algorithms. Thus, our conclusions about the computational time complexity of the dynamic programming algorithms made below will apply also to the generating function method.

Transfer matrix method

The transfer matrix approach constructs the partition function by sequentially multiplying the transfer matrices (weight matrices) assigned to each DNA unit [39, 50, 86, 89, 91–101]. First matrix models were made for oligonucleotide–DNA binding motivated by the success of previous matrix descriptions of the DNA helix-coil transition. Matrices were constructed both for the case of the all-or-none binding as in Figure 1A [99] and for partial binding as in Figure 1B [93]. Although the partial binding of two oligonucleotides and partial unwrapping of nucleosomal DNA are very different processes, they can be described by similar mathematical formalisms [39]. The transfer matrices are constructed so that each matrix element $Q_n(i,j)$ contains the weights assigned to the combination of states where the

lattice unit n is in a state i followed by the unit $n+1$ in state j. Prohibited combinations of states are characterized by zero weights. The partition function is given by sequential multiplication of all transfer matrices enclosed between two unit vectors:

$$Z = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \times \prod_{n=1}^{N} Q_n \times \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix} \quad (3)$$

The algorithm of transfer matrix construction depends on the model. A detailed description was given previously and includes multi-protein competitive binding with long-range interactions in the presence of nucleosomes which can partially unwrap [39]. For example, for the case of noncooperative binding of a single ligand with length $m=3$, we have six possible states for each DNA unit: (i) bound to 1st protein unit; (ii) bound to 2nd protein unit; (iii) bound to 3rd protein unit; (iv) free unit not at the DNA ends; (v) free unit at the left DNA end; and (vi) free unit at the right DNA end. If all-or-none protein binding is considered as in Figure 1A, the following transfer matrix can be constructed for DNA units far from the boundaries:

$$\begin{array}{l} \text{Bound, 1st unit} \\ \text{Bound, 2nd unit} \\ \text{Bound, 3rd unit} \\ \text{Free, not at ends} \\ \text{Free, left end} \\ \text{Free, right end} \end{array} \begin{pmatrix} 0 & K(n)c_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (4)$$

Here, the number of the row corresponds to the state of the unit n, and the number of the column corresponds to the state of the next unit $n+1$. $K(n)$ is the binding constant for a site beginning at position n, and c_0 is the protein concentration. This matrix is highly sparse because state 1 can be followed only by state 2, and state 2 can be followed only by state 3. If incomplete binding is allowed (Figure 1B), these constraints are lifted and the matrix changes to the following:

Here, instead of the macroscopic binding constants $K(n)$, microscopic binding constants $k(n, h)$ are assigned to each bond h within the binding site $(n, n+m-1)$. Each unit preceding the leftmost position of a bound protein is assigned a weight c_0 for the entropy of removing the protein from the solution to the DNA. Matrix element $Q(1,1)$ is multiplied by c_0 twice to account for bringing from solution to DNA two proteins contacting at this point. $Q(2,2)=0$ since two proteins are not allowed to be both unwrapped at the point of contact as shown in Figure 2D. The latter condition is introduced to disallow ambiguities for the elements $Q(1,2)$ and $Q(2,3)$, which otherwise could have been interpreted both as a continuation of the bound protein and the contact of two partially unwrapped proteins. The product of all microscopic binding constants gives the macroscopic binding constant:

$$\prod_{h=1}^{m} k(n,h) = K(n) \quad (6)$$

From this, microscopic constants can be estimated if relative strengths of interactions within the binding site are known, which can be derived for example by molecular dynamics simulations [39]. If all bonds within a binding site are equivalent, one writes:

$$k(n,h) = \sqrt[m]{K(n)} \quad (7)$$

In the incomplete binding model [Equation (5)] the number of nonzero matrix elements increases, and the matrix is not sparse any more in contrast to the all-or-none description [Equation (4)]. This is also true for the general case of f protein types if the maximal range of interactions V is comparable with the typical length m of the DNA binding site. Therefore, the computation time required to multiply the matrices increases. The computation time of the matrix approach scales linearly with N, since it relies in the sequential multiplication of N matrices. The number of matrix elements scales as $O(m^2 f^2 V)$.

$$\begin{array}{l} \text{Bound, 1st unit} \\ \text{Bound, 2nd unit} \\ \text{Bound, 3rd unit} \\ \text{Free, not at ends} \\ \text{Free, left end} \\ \text{Free, right end} \end{array} \begin{pmatrix} k(n,1)c_0^2 & k(n,1)c_0 & k(n,1)c_0 & k(n,1)c_0 & 0 & k(n,1)c_0 \\ k(n,2)c_0 & 0 & k(n,2) & k(n,2) & 0 & k(n,2) \\ k(n,3)c_0 & k(n,3)c_0 & k(n,3)c_0 & k(n,3) & 0 & k(n,3) \\ c_0 & c_0 & c_0 & 1 & 0 & 0 \\ c_0 & c_0 & c_0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (5)$$

Recurrent relations a.k.a. dynamic programming

In 1974, when the transfer matrix and generating function approach already existed, DeLisi [102] proposed a new method to calculate the partition function recurrently, which he called the renewal theory. This method is based on the idea that the partition function for a DNA of length N can be calculated recurrently if partition functions for smaller lattices are known [7, 12, 20, 89, 102–113]. Conceptually, this approach belongs to the class of dynamic programming algorithms defined as recurrent solving complex problems by breaking them down into simpler overlapping sub-problems. The term 'dynamic programming' was coined in 1950 by Richard Bellman [114], but as far as we know, dynamic programming algorithms for calculation of protein–DNA binding did not exist until 1974. DeLisi applied this method to the problem of oligomer–polymer binding without long-range interactions [106]. Examples considered in his original publication could be also solved by other existing approaches, and therefore the power of this method was not noticed at that time. In 1978, Gurskii and Zasedatelev [115] developed a recurrent theory which considered protein–DNA binding with arbitrary long-range interactions. This case is already quite complex and recurrent relations decreased its computational time complexity from exponential to linear (as a function of DNA length N). However, the method was again not widely accepted by the scientific community for two other reasons. The first reason was that back in 1978, there was no real need for calculating such complex systems because typical *in vitro* experiments conducted at that time could be described with the simple McGhee–von Hippel equation [73]. The second reason was that Gurskii and Zasedatelev published the first derivation of their equations in Russian. Although several follow-up publications with simplified equation appeared in English [107, 112], the authors themselves mostly cited their original Russian works not accessible for the majority of scientists. In 1990s, Di Cera and co-authors [89, 103, 104] also approached the idea of the use of recurrent equations, now departing from the generating function method. They derived systematic mathematical theorems justifying the use of recurrent equations, and applied this method to several DNA–ligand examples. However, the recurrent relation method became widely used only a decade later, when high-throughput genome-wide binding studies became feasible [7, 12, 20, 105, 107–111, 113, 116]. The latter publications considered the problem of competitive and cooperative binding of multiple proteins assuming the all-or-none binding without the possibility of incomplete protein–DNA binding.

For comparison with the transfer matrix formalism described above, we will derive here an updated dynamic programming algorithm that takes into account TF access to partially unwrapped nucleosomal DNA. We will first follow the reasoning of Gurskii and Zasedatelev. Consider the case of the all-or-none binding of a single protein type ($g=1$) with long-range interactions. Let the DNA lattice of N units be characterized by the partition function Z_N. The last lattice unit N is either free or bound by the protein. If Nth unit is free, such a lattice is equivalent to the lattice of length $N-1$ characterized by the partition function Z_{N-1}. Therefore, the difference between partition functions Z_N and Z_{N-1} corresponds to all states where unit N is bound by the protein. For all-or-none binding, each protein covers m DNA units. Therefore, if unit N is bound by a protein, units $[N-m+1, N]$ are also bound. Thus, two possibilities exist: this protein either interacts with a preceding protein separated by distance $l \le V$, or it does not interact if it is preceded by $l > V$ free DNA units. The weight of a bound protein which does not interact with other proteins is given as $K(N-m+1)\times c_0$. The weight of the lattice of length N, containing a protein bound at units $[N-m+1, N]$ separated by l units ($l \le$ V) from the previous protein is given as $K(N-m+1)w(l)c_0(Z_{N-m-l}-Z_{N-m-l-1})$. Here the term $(Z_{N-m-l}-Z_{N-m-l-1})$ corresponds to the weight of the states with bound unit $N-m-l$. Therefore Z_n can be expressed according to the Gurskii–Zasedatelev equation:

$$Z_N = Z_{N-1} + K(N-m+1)c_0 Z_{N-m-V} + \sum_{l=0}^{V} w(l)K(N-m+1)c_0(Z_{N-m-l}-Z_{N-m-l-1}) \quad (8)$$

This equation has the following starting conditions:

$$Z_n = 1 \text{ if } n < m;\ K(n) = 0 \text{ if } n < 1 \quad (9)$$

Equations 8 and 9 allow calculating Z_N recurrently for any N. One can see that the calculation time of this algorithm scales as $O(NV)$ and does not depend

on m. This is the fastest existing algorithm for the given task. If there are $f > 1$ protein types denoted by index $g = (1, f)$, the recurrent algorithm transforms to the following:

$$Z_N = Z_{N-1} + \sum_{g=1}^{f} K(N - m(g) + 1, g) c_0(g) Z_{N-m(g)-V} + \sum_{g=1}^{f} \sum_{g'=1}^{f} \sum_{l=0}^{V} w(l, g', g) K(N - m(g) + 1, g) c_0(g) \times (Z_{(N-m(g)-l)} - Z_{N-m(g)-l-1}) \quad (10)$$

Equation (10) with boundary conditions (9) holds for the case of the all-or-none binding of multiple competing interacting proteins with long-range interactions. The time complexity of this algorithm is $O(Nf^2V)$. In a specific case $w(l) = \text{const}$ for $l \leq V$, the dependence on V vanishes [111, 112].

Now let us consider the case of the partial binding that accounts for nucleosome unwrapping. Here, we will follow the reasoning of DeLisi [106] to change boundary conditions. In our initial considerations above having a protein hanging out from the DNA ends was prohibited. Now it will be allowed that a protein starts at a unit $n < N$ and protrudes beyond the end of the DNA lattice. Since partial unwrapping is possible, the knowledge that a given DNA unit is bound does not define a precise binding site of length $m(g)$ as before. Instead, several possibilities exist for the protein to have h_1 units unwrapped at its left end and h_2 units unwrapped at its right end. Correspondingly, the binding constant K^* for a protein, which covers the Nth base pair depends on the number of formed bonds as a function of N, g, h_1 and h_2:

$$K^* = K(N, g, h_1, h_2) = \prod_{h=h_1+1}^{m(g)-h_2} k(N - m(g) + h, g, h), \quad (11)$$

where $k(n, g, h)$ is the miscroscopic binding constant as introduced previously. Index $n = N - m(g) + h$ numbers base pairs starting from the left DNA end, and index h numbers base pairs starting from the left-most position of the binding site for a completely bound protein of type g (Figure 1). When counting the states of protein–protein interactions for the proteins separated by $l \leq V$ DNA units, one has to keep in mind that now both interacting proteins can be partially unbound. This leads to the following recurrent algorithm to calculate the partition function for the case of partial binding of multiple proteins with long-range interactions:

$$Z_N = Z_{N-1} + \sum_{g=1}^{f} \sum_{h_1=0}^{m(g)-1} \sum_{h_2=0}^{m(g)-h_1-1} c_0(g) Z_{N-m(g)+h_1+h_2-V} K^* + \sum_{l=0}^{V} \sum_{g'=1}^{f} \sum_{g=1}^{f} \sum_{h_1=0}^{m(g)} \sum_{h_2=0}^{m(g)-h_1-1} w(l, g', g) c_0(g) \times \left(Z_{N-m(g)+h_1+h_2-l} - Z_{N-m(g)+h_1+h_2-l-1}\right) K^* \quad (12)$$

According to Equations 11 and 12, the computation time of this algorithm scales as $O(Nf^2m^3V)$. The linear dependence of computation time on the DNA length N both, in the dynamic programming and transfer matrix algorithms most likely cannot be improved. For example, in the conceptually related problem of DNA melting, researchers are satisfied with the computation time scaling as $O(N^2)$, which allows calculating the double helix stability for complete chromosomes [117]. In our case, most of the complexity arises from the dependence on f^2 and m^3 (or m^2 in the case of a homogenous unwrapping potential given by Equation 7 instead of Equation 11).

Test calculations for a *Drosophila* enhancer

In order to compare the performance of the transfer matrix and dynamic programming algorithms, both methods were applied to a prototypic biological system using the same computer setup. The test case is derived from a recent experimental study of *Drosophila* embryonic development, in which fly mutants with different distances between TF binding sites in a particular enhancer region were investigated [35]. To describe this system, three types of proteins can assemble on the DNA: the TFs denoted 'TF_1' (e.g. representing 'Giant') and 'TF_2' (e.g. representing 'Twist') and the histone octamer forming the nucleosome denoted 'N' in Figure 3. The 560 bp DNA region under consideration contains two 6-bp binding sites for TF_1 acting as a short-term transcription repressor and four 6-bp binding sites for TF_2 acting as a transcription activator. Repressor and activator do not touch each other physically but may indirectly interact through a nucleosome [118]. In this case, the possibility of partial nucleosome unwrapping largely changes the geometry of the complex and needs to be taken

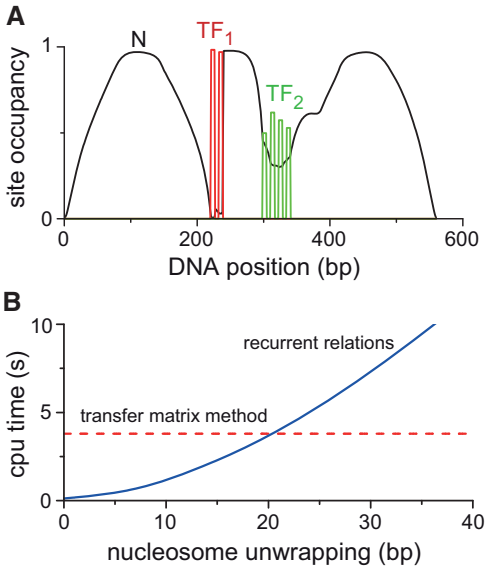

Figure 3: Calculation for TF binding in the presence of nucleosomes for a *Drosophila* enhancer. (**A**) An example of a binding map calculated with the transfer matrix formalism for a DNA region of 560 bp with two 6 bp binding sites for protein TF_1 and four 6 bp binding sites for protein TF_2. In the model, nucleosomes can be present nonspecifically at any location, the TF_1-nucleosome contact is cooperative and TF_1 and TF_2 exclude each other at distances <6 bp. (**B**) Calculation time for one point of the binding map from panel A as a function of the maximum allowed nucleosome unwrapping length. Solid line: dynamic programming method; dashed line: transfer matrix method. Computations of all points of the binding map can be performed in parallel. Thus, this plot also gives an estimate for the total time needed to calculate the binding map.

into account. In the current calculations it was assumed that TF_1–nucleosome contact is cooperative; TF_1 and TF_2 physically exclude each other within a distance of 6 bp; the nucleosome can form at any position nonspecifically and can partially unwrap characterized by the homogeneous unwrapping potential [39]. The set of binding parameters chosen for the calculations is described elsewhere [118]. A regular laptop computer (Intel Core2 Duo, 2.53 GHz, 3 GB RAM) was used for the calculations.

Figure 3A shows an example of the binding map calculated for this system using the transfer matrix method, which required 3.8 s for one point of the map. The map shows probabilities of protein binding for each DNA base pair for each of the three protein species. In the absence of nucleosome unwrapping, the dynamic programming Equation 10 is much faster, requiring ~0.1 s to calculate the whole binding map shown in Figure 3A. However, in the case of unwrapping of the nucleosomal DNA, the dynamic programming algorithm (Equation 12) performs slower than the equivalent matrix algorithm. The reason is that physical models for the mathematical representation of binding in both methods are different. In the case of the matrix approach, the situation when both neighboring nucleosomes are partially unwrapped at the point of their contact is prohibited (Figure 2D). Only the situation when one of the two 'colliding' nucleosomes is unwrapped at the point of the nucleosome–nucleosome contact is allowed (Figure 2C). This limitation follows naturally from the mathematics of the matrix algorithm implemented here. It is justified in terms of the biological system to consider only the case when the partially unwrapped regions of two adjacent nucleosomes are separated by at least one free base pair. These constraints do not exist in the dynamic programming algorithm as it is formulated in Equation 12, which increases the computation time. Nevertheless, the dynamic programming algorithm can be considered a good choice in a situation when nucleosome unwrapping can be restricted to a limited number of base pairs close to the nucleosome entry/exit. That would change the upper summation limit for indexes h_1 and h_2 in Equation 12, significantly decreasing the computation time. Figure 3B shows calculation times for one point of the binding map as a function of the maximum allowed length of DNA unwrapping from the nucleosome. It is apparent that the dynamic programming method is a better choice when the maximum unwrapping length is <20 bp, but the transfer matrix method outperforms for larger unwrapping lengths. Our previous MD simulations showed that unwrapping of at least 30 bp needs to be considered in the lattice model [39]. Experimental data support the view that TFs can access the DNA even deeper inside the nucleosome [41–47]. Whether the possibility of complete nucleosome unwrapping needs to be considered remains to be tested in future experiments. For the middle-range unwrapping lengths the transfer matrix and dynamic programming algorithms give similar calculation times (Figure 3B).

It should be noted that only exact solutions with single base pair precision were considered here. Using a more coarse-grained DNA lattice model [50] or introducing thresholds neglecting weak

binding [27, 108] would make calculations faster but less precise. Such possibilities are similarly applicable to each of the algorithms and thus do not change the relative differences in computing time. However, given the well-established dependency of transcription binding on the exact DNA sequence, it remains questionable whether meaningful binding maps can be obtained if the resolution is reduced above the single base pair level.

Strategies for accelerating the calculations

As discussed above, methods to calculate protein binding maps for chromatin do exist. However, the required computational time prevents their application to calculating genome-wide TF binding as required for predicting gene expression. This is due to the higher complexity of the problem as, for example compared to predicting nucleosome positions from the sequence. The calculation of binding of two proteins in chromatin with the transfer matrix formalism for one *Drosophila* enhancer region shown in Figure 3A is already ~100 times slower than predicting the arrangement of histone octamers (without partial unwrapping) on a DNA of this size using existing algorithms [20, 24–26]. Due to the linear dependence of the calculation time on N, calculations of genome-wide chromatin TF maps would be millions of times slower. Thus, the search for new more powerful algorithms is highly relevant. However, it is also worth to consider the underlying biology in more detail. The calculation times estimated above are much larger than typical equilibration times in the cell. Obviously, the cell does not really 'compute' anything. Nevertheless, it can be estimated that it cannot sample on the fly through all possible configurations of all its TFs and nucleosomes to derive the equivalent of a partition function. What could be mechanisms that operate in the cell to solve this problem? And can these be adapted to the DNA lattice based calculations discussed here? To address these questions we propose two approaches to accelerate the sampling, namely parallelization and sequential ordering of binding events (Figure 4).

It is usually argued that a high speed of biological processes can be achieved by the parallelization of 'calculations' through many molecules. However, in the case of our problem, the one-dimensional connectivity of the DNA chain needs to be accounted for. Therefore, parallelization to derive

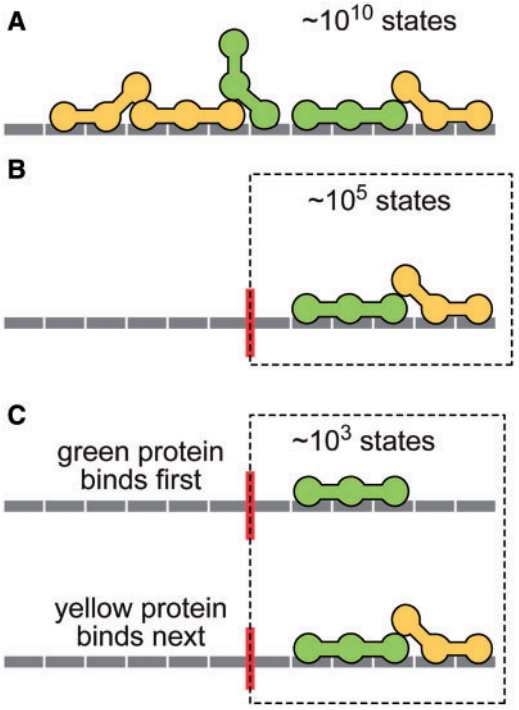

Figure 4: Strategies for calculating genome-wide TF binding maps. In this example, two different protein types (shown in green and yellow) bind DNA. Each protein can cover up to three DNA lattice units upon binding. Each DNA unit can be in seven states (free, or bound to one of the three segments of two proteins). (**A**) The DNA lattice consisting of 12 U can be in up to 12^7 ($\sim 10^{10}$) states. (**B**) Boundary element set in the middle splits the DNA into two independent lattices. Each lattice of 6 U has up to 6^7 ($\sim 10^5$) states. (**C**) The sequential binding model. First, only binding of green proteins is considered, characterized by up to 4^7 ($\sim 10^3$) states. Then the same number of states is added when binding of yellow proteins to the DNA lattice with fixed green proteins is considered.

the protein–DNA configurations is strictly speaking not possible, neither on a computer, nor in the cell. The latter statement holds assuming that all potential binding sites overlap and no boundaries between different chromatin regions exist. However, the situation in the cell is very different in a number of aspects: (i) different chromosomes can be calculated in parallel, decreasing the computation time for a complete human genome by an order of magnitude; (ii) chromatin is further organized into domains. One well-established structure are domains of ~1 million base pairs, which are readily apparent in high-resolution microscopy images and by *in situ*-cross linking [119, 120]. These 1 Mb genomic subcompartments (containing ~5000 nucleosomes)

could represent independent DNA modules. Considering these individual chromatin domains in the lattice models would allow decreasing the computation time by at least two more orders magnitude; (iii) molecular insulators such as CTCF proteins can provide boundaries for nucleosome and TF binding to DNA [121, 122]. Since one boundary positions up to 20 nucleosomes, this can reduce the number of 'meaningful' nucleosome positions by an order of magnitude assuming that boundaries are distributed more or less homogenously in the genome [50, 100] (Fig. 4B); and (iv) large-scale formation of more compacted and biologically inactive heterochromatin on the scale of megabases could reduce TF access to these regions [123, 124]. This could significantly reduce the size of the actively transcribed genome where TF binding needs to be considered.

The second concept that is realized in the cell and could be transferred to computations is to introduce a sequential ordering of binding events (Figure 4C). When the sequence of assembly of f types of proteins is unknown, the computational complexity scales with f^2 (Equations 10 or 12). On the other hand, if protein assembly would follow a known sequence of binding events, one would be able to calculate binding maps for each protein one by one and just repeat this f times (because previously bound proteins change the binding interface for the next bound proteins). This would decrease the computation time from f^2 to f. In reality, protein binding in the cell seems to operate via a mixture of stochaistic and sequential binding events [125, 126], the latter being more effective in the case of cooperative binding [127]. Experimental details on this are currently available for only a few well-studied systems such as the interferon-β enhanceosome assembly [128]. However, it is clear that in many instances the binding of 'pioneering' protein factors is required before other proteins can bind to a given locus. This is particularly relevant for the binding to nucleosomal DNA [129]. A number of these pioneering factors have been identified that are thought to displace the histone octamer from a given DNA target sequence to allow for the subsequent binding of other factors [28]. Accordingly, a complete decrease of the computation time from f^2 to f is not possible, but a significant acceleration of calculations appears feasible.

In summary, a computational approach that makes use of additional information on the biological system with respect to genome compartmentalization and sequential order of binding events appears to be a promising strategy (Figure 4). It would involve first splitting each chromosome into functional chromatin domains, for example, according to structural data [130] and/or the known positions of boundary elements such as CTCF [121, 122] and defining regions that oppose binding of a given TF due to the underlying chromatin state. In addition, the sequential order of certain binding cascades would be introduced as additional constraints of the type of 'A binds before B binds before C' (Figure 4C). The successful implementation of these two approaches could significantly reduce the computation time so that calculating genome-wide TF binding maps in humans would become feasible.

Key Points

Predicting gene expression from the competitive and combinatorial binding of multiple proteins at genomic regulatory regions is one of the ultimate goals in quantitative biology. Here, we have considered five classes of algorithms for calculating such protein–DNA binding maps using DNA lattice models in a chromatin context. The following conclusions are drawn:

- Competitions for binding sites with histone octamers as well as the partial unwrapping of nucleosomal DNA need to be considered for meaningful calculation of TF DNA occupancy.
- The combinatorial approach is not suited for the problem of sequence-specific binding. In the limit of large genomic regions with overlapping binding sites, the binary variable approach reduces to the transfer matrix approach, and the generating function approach reduces to the dynamic programming approach.
- The transfer matrix and dynamic programming algorithms are the only two principal methods suited for calculating TF binding maps in chromatin. These approaches represent different ways of calculating the partition function and cannot be reduced to each other.
- The computation times for a genome-wide analysis of TF binding are prohibitively long if one considers nucleosome unwrapping in the absence of any further constraints of the system.
- Calculations can be accelerated by including additional biological information with respect to genome compartmentalization, heterochromatin formation and a sequential order of binding events.

Acknowledgements

Discussions with Peter H. von Hippel, Enrico Di Cera, Yurii Nechipurenko and Rutger Hermsen helped to clarify the historical timeline explained here. We thank the anonymous reviewers for their suggestions.

FUNDING

The Bundesministerium für Bildung und Forschung funded EpiGenSys project within the EraSysBioPlus

program; the Belarus National Foundation for Fundamental Investigations (grant B10M-060).

References

1. Garcia HG, Sanchez A, Kuhlman T, *et al*. Transcription by the numbers redux: experiments and calculations that surprise. *Trends Cell Biol* 2010;**20**:723–33.
2. Bintu L, Buchler NE, Garcia HG, *et al*. Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev* 2005;**15**:125–35.
3. Bintu L, Buchler NE, Garcia HG, *et al*. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* 2005;**15**:116–24.
4. Yuh CH, Bolouri H, Davidson EH. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 1998;**279**:1896–902.
5. Jaeger J, Surkova S, Blagov M, *et al*. Dynamic control of positional information in the early Drosophila embryo. *Nature* 2004;**430**:368–71.
6. Janssens H, Hou S, Jaeger J, *et al*. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. *Nat Genet* 2006;**38**: 1159–65.
7. Segal E, Raveh-Sadka T, Schroeder M, *et al*. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* 2008;**451**:535–40.
8. Gertz J, Siggia ED, Cohen BA. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 2009;**457**:215–8.
9. Yuan Y, Guo L, Shen L, *et al*. Predicting gene expression from sequence: a re-examination. *PLoS Comput Biol* 2007;**3**: e243.
10. Beer MA, Tavazoie S. Predicting gene expression from sequence. *Cell* 2004;**117**:185–98.
11. Zinzen RP, Senger K, Levine M, *et al*. Computational models for neurogenic gene expression in the Drosophila embryo. *Curr Biol* 2006;**16**:1358–65.
12. He X, Samee MA, Blatti C, *et al*. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol* 2010;**6**:e1000935.
13. Kaplan T, Li XY, Sabo PJ, *et al*. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genet* 2011;**7**:e1001290.
14. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 1987; **193**:723–50.
15. Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* 2010;**11**:751–60.
16. Pfreundt U, James DP, Tweedie S, *et al*. FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database. *Nucleic Acids Res* 2010;**38**: D443–7.
17. Portales-Casamar E, Thongjuea S, Kwon AT, *et al*. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2010; **38**:D105–10.
18. Wingender E, Dietze P, Karas H, *et al*. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 1996;**24**:238–41.
19. Trifonov EN. Cracking the chromatin code: Precise rule of nucleosome positioning. *Phys Life Rev* 2011;**8**:39–50.
20. Segal E, Fondufe-Mittendorf Y, Chen L, *et al*. A genomic code for nucleosome positioning. *Nature* 2006; **442**:772–8.
21. Yuan GC, Liu YJ, Dion MF, *et al*. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* 2005; **309**:626–30.
22. Peckham HE, Thurman RE, Fu Y, *et al*. Nucleosome positioning signals in genomic DNA. *Genome Res* 2007;**17**: 1170–7.
23. Levitsky VG. RECON: a program for prediction of nucleosome formation potential. *Nucleic Acids Res* 2004;**32**: W346–9.
24. Xi L, Fondufe-Mittendorf Y, Xia L, *et al*. Predicting nucleosome positioning using a duration hidden Markov model. *BMC Bioinformatics* 2010;**11**:346.
25. Gabdank I, Barash D, Trifonov EN. FineStr: a web server for single-base-resolution nucleosome positioning. *Bioinformatics* 2010;**26**:845–6.
26. Locke G, Tolkunov D, Moqtaderi Z, *et al*. High-throughput sequencing reveals a simple model of nucleosome energetics. *Proc Natl Acad Sci USA* 2010;**107**: 20998–1003.
27. Roider HG, Kanhere A, Manke T, *et al*. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 2007;**23**:134–41.
28. Längst G, Teif VB, Rippe K. Chromatin remodeling and nucleosome positioning. In: Rippe K, (ed.). *Genome Organization and Function in the Cell Nucleus*. Weinheim: Wiley-VCH, 2011, in press.
29. Teif VB, Bohinc K. Condensed DNA: condensing the concepts. *Progr Biophys Mol Biol* 2011;**105**:208–22.
30. Wang X, Bryant GO, Floer M, *et al*. An effect of DNA sequence on nucleosome occupancy and removal. *Nat Struct Mol Biol* 2011;**18**:507–9.
31. Sarai A, Kono H. Protein-DNA recognition patterns and predictions. *Annu Rev Biophys Biomol Struct* 2005;**34**:379–98.
32. Hoopes BC, LeBlanc JF, Hawley DK. Contributions of the TATA box sequence to rate-limiting steps in transcription initiation by RNA polymerase II. *J Mol Biol* 1998;**277**: 1015–31.
33. Verrijzer CP, Alkema MJ, van Weperen WW, *et al*. The DNA binding specificity of the bipartite POU domain and its subdomains. *EMBO J* 1992;**11**:4993–5003.
34. Strainic MG Jr, Sullivan JJ, Collado-Vides J, *et al*. Promoter interference in a bacteriophage lambda control region: effects of a range of interpromoter distances. *J Bacteriol* 2000; **182**:216–20.
35. Fakhouri WD, Ay A, Sayal R, *et al*. Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. *Mol Syst Biol* 2010;**6**:341.
36. Bandele OJ, Wang X, Campbell MR, *et al*. Human single-nucleotide polymorphisms alter p53 sequence-specific binding at gene regulatory elements. *Nucleic Acids Res* 2011;**39**:178–89.
37. Ising E. Beitrag zur Theorie des Ferromagnetismus. *Z Phys* 1925;**31**:253–8.

38. Markov AA. Investigation of a specific case of dependent observations. *Izv Imper Akad Nauk* 1907;**3**:61–80.
39. Teif VB, Ettig R, Rippe K. A lattice model for transcription factor access to nucleosomal DNA. *Biophys J* 2010;**99**: 2597–607.
40. Engeholm M, de Jager M, Flaus A, *et al*. Nucleosomes can invade DNA territories occupied by their neighbors. *Nat Struct Mol Biol* 2009;**16**:151–8.
41. Poirier MG, Oh E, Tims HS, *et al*. Dynamics and function of compact nucleosome arrays. *Nat Struct Mol Biol* 2009;**16**: 938–44.
42. Poirier MG, Bussiek M, Langowski J, *et al*. Spontaneous access to DNA target sites in folded chromatin fibers. *J Mol Biol* 2008;**379**:772–86.
43. Gansen A, Valeri A, Hauger F, *et al*. Nucleosome disassembly intermediates characterized by single-molecule FRET. *Proc Natl Acad Sci USA* 2009;**106**:15308–13.
44. Bucceri A, Kapitza K, Thoma F. Rapid accessibility of nucleosomal DNA in yeast on a second time scale. *EMBO J* 2006;**25**:3123–32.
45. Li G, Levitus M, Bustamante C, *et al*. Rapid spontaneous accessibility of nucleosomal DNA. *Nat Struct Mol Biol* 2005; **12**:46–53.
46. Koopmans WJ, Buning R, Schmidt T, *et al*. spFRET using alternating excitation and FCS reveals progressive DNA unwrapping in nucleosomes. *Biophys J* 2009;**97**: 195–204.
47. Anderson JD, Thastrom A, Widom J. Spontaneous access of proteins to buried nucleosomal DNA target sites occurs via a mechanism that is distinct from nucleosome translocation. *Mol Cell Biol* 2002;**22**:7147–57.
48. Davey CA, Sargent DF, Luger K, *et al*. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J Mol Biol* 2002;**319**:1097–113.
49. Raveh-Sadka T, Levo M, Segal E. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res* 2009;**19**:1480–96.
50. Teif VB, Rippe K. Statistical-mechanical lattice models for protein-DNA binding in chromatin. *J Phys Condens Matter* 2010;**22**:414105.
51. Teif VB. General transfer matrix formalism to calculate DNA-protein-drug binding in gene regulation: application to O_R operator of phage λ. *Nucleic Acids Res* 2007;**35**:e80.
52. Epstein IR. Cooperative and noncooperative binding of large ligands to a finite one-dimensional lattice. A model for ligand-oligonucleotide interactions. *Biophys Chem* 1978; **8**:327–39.
53. Teif VB, Haroutiunian SG, Vorob'ev VI, *et al*. Short-range interactions and size of ligands bound to DNA strongly influence adsorptive phase transition caused by long-range interactions. *J Biomol Struct Dyn* 2002;**19**:1093–100.
54. Vilar JMG, Saiz L. CplexA: a Mathematical package to study macromolecular-assembly control of gene expression. *Bioinformatics* 2010;**26**:2060–1.
55. Bakk A, Metzler R, Sneppen K. Sensitivity of OR in phage lambda. *Biophys J* 2004;**86**:58–66.
56. Ackers GK, Johnson AD, Shea MA. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci USA* 1982;**79**:1129–33.
57. Vilar JM, Leibler S. DNA looping and physical constraints on transcription regulation. *J Mol Biol* 2003;**331**:981–9.
58. von Hippel PH, Revzin A, Gross CA, *et al*. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects. *Proc Natl Acad Sci USA* 1974;**71**:4808–12.
59. Buchler NE, Gerland U, Hwa T. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci USA* 2003;**100**: 5136–41.
60. Beshnova DA, Bereznyak EG, Shestopalova AV, *et al*. A novel computational approach 'BP-STOCH' to study ligand binding to finite lattice. *Biopolymers* 2011;**95**:208–16.
61. Mjolsness E, Sharp DH, Reinitz J. A connectionist model of development. *J Theor Biol* 1991;**152**:429–53.
62. Mjolsness E. Towards a calculus of biomolecular complexes at equilibrium. *Brief Bioinform* 2007;**8**:226–33.
63. Mjolsness E. On cooperative quasi-equilibrium models of transcriptional regulation. *J Bioinform Comput Biol* 2007;**5**: 467–90.
64. Bishop CM. *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995.
65. Hertz J, Krogh A, Palmer RG. *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley, 1991.
66. Horsky J. Semiflexible oligomer-polymer binding: Combinatorial and conditional probability analyses and stochastic simulation. *Macromolecules* 2008;**41**:5014–23.
67. Nishio T, Shimizu T. Model analysis of surfactant–polymer interaction as cooperative ligand binding to linear lattice. *Biophys Chem* 2005;**117**:19–25.
68. Tsodikov OV, Holbrook JA, Shkel IA, *et al*. Analytic binding isotherms describing competitive interactions of a protein ligand with specific and nonspecific sites on the same DNA oligomer. *Biophys J* 2001;**81**:1960–9.
69. Rouzina I, Bloomfield VA. Competitive electrostatic binding of charged ligands to polyelectrolytes: practical approach using the non-linear Poisson-Boltzmann equation. *Biophys Chem* 1997;**64**:139–55.
70. Maltsev E, Wattis JA, Byrne HM. DNA charge neutralization by linear polymers. II. Reversible binding. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006;**74**:041918.
71. Teif VB. Ligand-induced DNA condensation: choosing the model. *Biophys J* 2005;**89**:2574–87.
72. Wolfe AR, Meehan T. Use of binding site neighbor-effect parameters to evaluate the interactions between adjacent ligands on a linear lattice: effects on ligand-lattice association. *J Mol Biol* 1992;**223**:1063–87.
73. McGhee JD, von Hippel PH. Theoretical aspects of DNA-protein interactions: co-operative and non-co-operative binding of large ligands to a one-dimensional homogeneous lattice. *J Mol Biol* 1974;**86**:469–89.
74. Tsuchiya T, Szabo A. Cooperative binding of n-mers with steric hindrance to finite and infinite one-dimensional lattices. *Biopolymers* 1982;**21**:979–84.
75. Zasedatelev AS, Gurskii GV, Vol'kenshtein MV. Theory of one-dimensional adsorption. I. Adsorption of small molecules on a homopolymer. *Mol Biol* 1971;**5**:194–8.
76. Latt SA, Sober HA. Protein-nucleic acid interactions. II. Oligopeptide-polyribonucleotide binding studies. *Biochemistry* 1967;**6**:3293–306.
77. Nechipurenko YD, Gursky GV. Cooperative effects on binding of proteins to DNA. *Biophys Chem* 1986;**24**: 195–209.

78. Nechipurenko YD, Zasedatelev AS, Gurskii GV. Theory of unidimensional adsorption onto homopolymers. Calculation of different ligand molecule orientations. *Biofizika* 1979;**24**:351–61.
79. Teif VB, Lando DY. Calculations of DNA condensation caused by ligand adsorption. *Molecular Biology* 2001;**35**: 106–7.
80. Mirny LA. Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci USA* 2010;**107**: 22534–9.
81. Chou T. Peeling and sliding in nucleosome repositioning. *Phys Rev Lett* 2007;**99**:058105.
82. Kong Y. A simple method for evaluating partition functions of linear polymers. *J Phys Chem B* 2001;**105**: 10111–4.
83. Lifson S. Partition functions of linear-chain molecules. *J Chem Phys* 1964;**40**:3705–10.
84. Schellman JA. Cooperative multisite binding to DNA. *Isr J Chem* 1974;**12**:219–38.
85. Chen Y, Maxwell A, Westerhoff HV. Co-operativity and enzymatic activity in polymer-activated enzymes. A one-dimensional piggy-back binding model and its application to the DNA-dependent ATPase of DNA gyrase. *J Mol Biol* 1986;**190**:201–14.
86. Chen Y. Binding of n-mers to one-dimensional lattices with longer than close-contact interactions. *Biophys Chem* 1987;**27**:59–65.
87. Chen YD. A general secular equation for cooperative binding of n-mer ligands to a one-dimensional lattice. *Biopolymers* 1990;**30**:1113–21.
88. Kong Y. Distribution of runs and longest runs. A new generating function approach. *J Am Stat Assoc* 2006;**101**: 1253–63.
89. Di Cera E, Kong Y. Theory of multivalent binding in one and two-dimensional lattices. *Biophys Chem* 1996;**61**: 107–24.
90. Lando DY, Nechipurenko YD. Distribution of unselectively bound ligands along DNA. *J Biomol Struct Dyn* 2008; **26**:187–96.
91. Woodbury CP Jr. Direct product-matrix method treatment of macromolecular binding. *Biopolymers* 1988;**27**:1305–17.
92. Teif VB. Predicting gene-regulation functions: lessons from temperate bacteriophages. *Biophys J* 2010;**98**: 1247–56.
93. Magee WS Jr, Gibbs JH, Newell GF. Statistical thermodynamic theory for helix-coil transitions involving poly- and oligonucleotides. II. The case of partial binding. *J Chem Phys* 1965;**43**:2115–23.
94. Hill TL. Some statistical problems concerning linear macromolecules. *J Polym Sci* 1957;**23**:549–62.
95. Gurskii GV, Zasedatelev AS, Vol'kenshtein MV. Theory of one-dimensional absorption. II. Adsorption of small molecules on a heteropolymer. *Mol Biol* 1972;**6**:385–93.
96. Crothers DM. Calculation of binding isotherms for heterogeneous polymers. *Biopolymers* 1968;**6**:575–84.
97. Chen YD. Multiple binding of ligands to a linear biopolymer. *Methods Enzymol* 2004;**379**:145–52.
98. Teif VB. General transfer matrix formalism to calculate DNA-protein-drug binding in gene regulation: application to O_R operator of phage lambda. *Nucleic Acids Res* 2007;**35**:e80.
99. Magee WS, Gibbs JH, Zimm BH. Theory of helix–coil transitions involving complementary poly- and oligo-nucleotides. I. The complete binding case. *Biopolymers* 1963;**1**:133–43.
100. Teif VB, Rippe K. Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities. *Nucleic Acids Res* 2009;**37**:5641–55.
101. Akhrem AA, Fridman AS, Lando DY. Effect of ligand interaction with the boundaries between different forms of DNA on intramolecular transitions. *Dokl Akad Nauk SSSR* 1985;**284**:212–6.
102. DeLisi C. Cooperative phenomena in homopolymers. An alternative formulation of the partition function. *Biopolymers* 1974;**13**:1511–2.
103. Di Cera E, Keating S. Site-specific thermodynamics of ising networks: a theorem for linearly connected subsystems. *Biopolymers* 1994;**34**:673–8.
104. Di Cera E. Thermodynamics of local linkage effects. Contracted partition functions and the analysis of site-specific energetics. *Biophys Chem* 1990;**37**:147–64.
105. Morozov AV, Fortney K, Gaykalova DA, *et al*. Using DNA mechanics to predict in vitro nucleosome positions and formation energies. *Nucleic Acids Res* 2009;**37**:4707–22.
106. DeLisi C. Statistical thermodynamics of oligomer-polymer interactions. *Biopolymers* 1974;**13**:2305–14.
107. Nechipurenko YD, Jovanovic B, Riabokon VF, *et al*. Quantitative methods of analysis of footprinting diagrams for the complexes formed by a ligand with a DNA fragment of known sequence. *Ann N Y Acad Sci* 2005;**1048**: 206–14.
108. Wasson T, Hartemink AJ. An ensemble model of competitive multi-factor binding of the genome. *Genome Res* 2009; **19**:2101–12.
109. Granek JA, Clarke ND. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol* 2005;**6**:R87.
110. Laurila K, Yli-Harja O, Lahdesmaki H. A protein-protein interaction guided method for competitive transcription factor binding improves target predictions. *Nucleic Acids Res* 2009;**37**:e146.
111. Hermsen R, Tans S, ten Wolde PR. Transcriptional regulation by competing transcription factor modules. *PLoS Comput Biol* 2006;**2**:e164.
112. Krylov AS, Grokhovsky SL, Zasedatelev AS, *et al*. 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* 1979;**6**:289–304.
113. He X, Chen CC, Hong F, *et al*. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS One* 2009;**4**:e8155.
114. Dreyfus S. Richard Bellman on the birth of dynamic programming. *Oper Res* 2002;**50**:48–51.
115. Gurskii GV, Zasedatelev AS. Precise relationships for calculating the binding of regulatory proteins and other lattice ligands in double-stranded polynucleotides. *Biofizika* 1978;**23**:932–46.
116. Hermsen R, Ursem B, ten Wolde PR. Combinatorial gene regulation using auto-regulation. *PLoS Comput Biol* 2010;**6**: e1000813.

117. Yeramian E. Genes and the physics of the DNA double-helix. *Gene* 2000;**255**:139–50.

118. Teif VB, Rippe K. Nucleosome mediated crosstalk between transcription factors at eukaryotic enhancers. *Phys. Biol.* 2011;**8**:04400.

119. Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* 2001;**2**:292–301.

120. Muller WG, Rieder D, Kreth G, *et al*. Generic features of tertiary chromatin structure as detected in natural chromosomes. *Mol Cell Biol* 2004;**24**:9359–70.

121. Cuddapah S, Jothi R, Schones DE, *et al*. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 2009;**19**:24–32.

122. Fu Y, Sinha M, Peterson CL, *et al*. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* 2008;**4**: e1000138.

123. Grewal SI, Jia S. Heterochromatin revisited. *Nat Rev Genet* 2007;**8**:35–46.

124. Maison C, Quivy JP, Probst AV, *et al*. Heterochromatin at mouse pericentromeres: a model for de novo heterochromatin formation and duplication during replication. *Cold Spring Harb Symp Quant Biol* 2010;**75**:155–65.

125. Wachsmuth M, Caudron-Herger M, Rippe K. Genome organization: balancing stability and plasticity. *Biochim Biophys Acta* 2008;**1783**:2061–79.

126. Luijsterburg MS, von Bornstaedt G, Gourdin AM, *et al*. Stochaistic and reversible assembly of a multiprotein DNA repair complex ensures accurate target site recognition and efficient repair. *J Cell Biol* 2010;**189**:445–63.

127. D'Orsogna MR, Chou T. First passage and cooperativity of queuing kinetics. *Phys Rev Lett* 2005;**95**:170603.

128. Ford E, Thanos D. The transcriptional code of human IFN-beta gene expression. *Biochim Biophys Acta* 2010;**1799**: 328–36.

129. Sekiya T, Muthurajan UM, Luger K, *et al*. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev* 2009;**23**:804–9.

130. Lieberman-Aiden E, van Berkum NL, Williams L, *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**: 289–93.